

Data-driven stochastic modelling of gene regulation

Dimitris Volteras

Doctor of Philosophy in Mathematics

Imperial College London
2024

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Statement of Originality

I declare that the work contained in this thesis is my own, except where it has formed part of a jointly-authored publication stated below and where contributions of others are acknowledged.

This thesis contains work from previously published research:

- Dimitris Volterras, Vahid Shahrezaei, and Philipp Thomas. “Global transcription regulation revealed from dynamical correlations in time-resolved single-cell RNA sequencing”. *Cell Systems* 15 (2024), pp. 694-708. DOI: [10.1016/j.cels.2024.07.002](https://doi.org/10.1016/j.cels.2024.07.002)

Γρήγορα καμωμένη, στο κατάστρωμα του πλοίου · ένα μαγευτικό απόγευμα.
Το Ιόνιον Πέλαγος ολόγυρά μας.

Κ. Π. Καβάφης

Dedicated to Roula and Giorgos, who inspired me to pursue an academic path.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Dr Vahid Shahrezaei and Dr Philipp Thomas, for their constant support and for always being available to offer invaluable advice. Our countless meetings, full of engaging and insightful conversations, were the true highlight of my PhD experience and played a fundamental role in shaping me as a scientist.

I would also like to thank friends and colleagues in the Shahrezaei's group, Wenhao, Jack, Andreas and Alicja, and the Thomas' group, Fern, Francesco, Paul, Johannes, Zekai, Hannah, and Alasdair, for stimulating and fun, scientific and non-scientific, discussions and for all the unforgettable moments in conferences. Also, thank you to friends and colleagues in the 3rd floor of SEC building and the Biomathematics community, for making the endless PhD days more enjoyable.

I would like to thank my friends in London, who have been with me from the very start and made this journey truly exciting. I am grateful to my friends in Greece, Isidoros, Haris, Stefanos, Nikoletta, Adonis, Fanis and Kostis, who have always believed in me and have supported me unconditionally. A special thank you to my cousin, Christina, who has been my most enthusiastic supporter and a limitless source of inspiration.

An enormous thank you to Mar, who understands better than anyone what this PhD means to me and has been by my side through all the challenges and the best moments. We did this together.

Finally, I am indebted to my parents, Olga and Vangelis, and my brother, Haris. Their constant belief in my potential and their support have been the foundation of everything I have achieved. None of this would have been possible without them.

Abstract

Gene expression varies significantly from cell to cell and over time, giving rise to phenotypic diversity within populations of genetically identical cells. This variation has a crucial role in gene regulatory networks and cell fate decisions. The recent rise in single-cell genomics technologies promises to exploit cellular variability to elucidate mechanisms of gene regulation. However, little emphasis is given on developing stochastic models that can effectively harness single-cell data to make sense of this variability. Moreover, single-cell sequencing data only reveals static snapshots of gene expression and is corrupted by technical noise, hindering biologically meaningful variation and posing challenges in uncovering gene regulatory mechanisms. The first part of this thesis introduces a stochastic modelling framework that leverages time-resolved single-cell transcriptomics data to tackle these challenges. The method delineates biological and technical noise sources and employs Bayesian inference to reveal global mechanisms of transcriptional regulation. The second part extends the framework to analyse gene co-expression relationships in time-resolved single-cell data, using stochastic models of gene co-regulation. The approach quantifies biological and technical contributions to transcriptional covariation and utilises machine learning techniques to identify modes of gene interaction. The final part attempts to address challenges arising from the complexity of modelling gene regulatory networks. It presents a rigorous theoretical method for approximating stochastic biochemical networks under timescale separation conditions, with applications to models of gene regulation.

Contents

Abstract	ix
1 Introduction	1
1.1 Aims and Outline	5
2 Bayesian inference of transcription regulation from time-resolved single-cell transcriptomics	8
2.1 Introduction	8
2.2 Stochastic modelling of time-resolved single-cell RNA sequencing	11
2.2.1 Time-resolved scRNA-seq data	11
2.2.2 Stochastic modelling framework	17
2.2.3 Capture efficiency estimation and downsampling of moments	26
2.2.4 Time-dependent summary statistics of gene expression	29
2.2.5 Model-based noise decomposition	31
2.3 Bayesian inference methodology	34
2.3.1 Approximate Bayesian Computation framework	34
2.3.2 Testing the model performance on synthetic data	39
2.4 Applications: Revealing global properties of transcriptional regulation	40

2.4.1	Stochastic modelling captures distinct patterns of bursty transcription dynamics along the cell cycle	40
2.4.2	Approximate Bayesian Computation reveals global patterns of transcriptional regulation	42
2.4.3	The inference method robustly identifies modes of regulation	47
2.4.4	Modelling capture and labelling efficiency allows to uncouple technical and biological noise	50
2.4.5	Bayesian inference reconstructs the transcriptional noise landscape	53
2.4.6	The model identifies scaling properties of gene expression with cell size .	56
2.4.7	The model predicts patterns of cell cycle-dependent kinetics regulation .	58
2.5	Discussion	61
3	Identifying modes of gene regulation in time-resolved single-cell transcriptomics	65
3.1	Introduction	65
3.2	Stochastic modelling of gene regulatory interactions	68
3.2.1	Stochastic reaction kinetics of interacting genes	68
3.2.2	Decomposition of gene-gene transcriptional covariation	74
3.2.3	Time-dependent summary statistics of gene co-expression	78
3.3	Synergising machine learning and mechanistic modelling	79
3.3.1	Generative-discriminative modelling for classification	79
3.3.2	Synthetic data generation	81
3.3.3	Neural network classifiers	82
3.4	Applications: Identifying modes of gene regulatory interaction	87

3.4.1	Independent gene models produce gene-gene correlations comparable to data	87
3.4.2	Mechanistic modelling and machine learning distinguish regulation from independent models	89
3.4.3	Dynamical gene-gene correlations identify regulatory modes of gene interaction	93
3.4.4	The model predicts modes of regulation from time-resolved scRNA-seq data	101
3.5	Discussion	105
4	Approximation of stochastic reaction networks under timescale separation	108
4.1	Introduction	108
4.2	A reduced diffusion approximation of stochastic reaction networks	112
4.2.1	Exact and approximate descriptions of stochastic reaction networks . . .	112
4.2.2	A heuristic stochastic approximation	116
4.2.3	Projection of the system to the slow timescale	117
4.2.4	Conditional linear noise approximation of the fast sub-system	123
4.2.5	The slow-scale chemical Langevin equation	128
4.3	Application: Michaelis-Menten enzyme kinetics	131
4.4	Discussion	138
5	Discussion	141
5.1	Summary of contributions	141
5.2	Conclusions and future directions	143

5.2.1	Inferring complex modes of gene regulation from time-resolved transcrip- tomics	144
5.2.2	Mechanistic modelling, Bayesian inference and machine learning for single- cell multi-omics	145

Bibliography	147
---------------------	------------

Chapter 1

Introduction

Biological functions inside living cells are precisely controlled by genes that are expressed and interact through complex regulatory networks. According to the central dogma of molecular biology, genetic information encoded in DNA is translated into cellular function via gene expression, a complex process involving *transcription* (RNA synthesis) and *translation* (protein synthesis) [1]. The molecular components responsible for this process, such as RNA polymerases for transcription and ribosomes for translation [2], are typically present at low copy numbers, giving rise to stochastic fluctuations in genetic activity and interaction [3–6]. As a consequence of stochasticity, clonal populations of cells that share the same genetic content can exhibit substantial variability in behaviours and phenotypes [7, 8]. Yet, despite the presence of noise, cells make decisions in a highly robust and coordinated way [9]. This remarkable phenomenon is of central interest in biological research. Importantly, noise is not merely a perturbative force that cells attempt to suppress in order to achieve precise biological function. Evidence suggests that cells exploit noise to diversify their phenotypes and respond to environmental signals, thereby optimising their behaviour [10–14]. This makes stochasticity a crucial factor in gene regulation and cellular decision making [15–18].

The standard approach for studying cellular noise involves distinguishing between intrinsic and extrinsic contributions to the overall gene expression variation [19, 20]. *Intrinsic noise* originates from the random timing of biochemical interactions that occur during gene expression, causing temporal fluctuations in the activity of a gene in a single cell [21]. *Extrinsic noise* arises from variation in cellular states, such as cell cycle stage, cell size and levels of transcription

factors, or differences in the cellular environment, causing fluctuations in the expression of a gene from cell to cell [11, 22]. The development of dual-reporter assays provided the first experimental evidence of intrinsic and extrinsic noise effects, by measuring the fluorescence of two protein reporters controlled by identical gene promoters, in clonal populations of prokaryotic and eukaryotic cells [3, 8, 23, 24]. These experiments demonstrated that extrinsic noise causes cell-to-cell variation in the expression of a single gene, but also leads to correlated fluctuations in the expression of two gene copies in the same cell. Intrinsic noise is observed when two gene copies in the same intracellular environment exhibit uncorrelated expression fluctuations. These observations made it apparent that detecting and quantifying noise is key for understanding gene regulation [21, 25–30]. Importantly, they pioneered a new research direction towards a quantitative characterisation of gene expression stochasticity and its consequences.

A step along this direction came with novel technologies that have enabled precise and high-throughput measurement of gene expression and cell-to-cell variability. Various methods enable real-time monitoring of gene expression dynamics in single cells with the use of live fluorescent microscopy and time-lapse imaging [31–34]. Visualising and tracking temporal expression patterns has offered fundamental insights into gene regulation at single-molecule resolution [35–38]. An important example is the observation that transcription occurs in short-lived, intermittent pulses of activity followed by periods of inactivity, a phenomenon known as *transcriptional bursting* [3, 39, 40]. Nonetheless, a limitation of these methods is that they can track the fluorescence of only a few genes at a time, thus prohibiting genome-wide analyses of gene expression.

Genome-wide quantification of transcription in individual cells has been made possible with *single-cell RNA sequencing* (scRNA-seq) [41], that enables whole-transcriptome profiling across thousands to millions of cells simultaneously [42, 43]. This method has provided unprecedented insights into cell-to-cell transcriptional heterogeneity and its role in gene regulation, with numerous applications including developmental and cancer biology, neuroscience and biomedicine [44–50]. A major limitation of single-cell sequencing is that it disrupts cellular physiology and only reveals a snapshot of transcriptome profiles in time, obscuring information about the underlying transcription dynamics [51].

The advent of single-cell technologies has boosted the development of mathematical and computational methods for modelling and analysing gene regulation. These methods utilise *biochem-*

ical reaction network descriptions to capture key steps in the gene expression process, such as transcriptional activation, mRNA synthesis, processing and degradation, and regulatory events such as protein-DNA interactions [52]. These networks can be mathematically formulated in a deterministic manner using ordinary differential equations (ODEs) to describe the average dynamics of gene expression in a cell population. Alternatively, stochastic formulations account for cell-to-cell variability by characterising the probability distribution of gene expression across cells [52–59]. In stochastic analyses, gene expression dynamics is typically modelled as a discrete-state, continuous-time Markov process, that tracks the abundance of molecular species comprising the reaction network [60–62]. The time-evolution of the probability mass function of this process is governed by the *chemical master equation* (CME), that describes the probability fluxes among all possible states [63]. As the CME is analytically intractable for most detailed models of gene regulation, modelling often relies on the *Gillespie stochastic simulation algorithm* (SSA) to sample trajectories of the stochastic process and approximate the exact probability distribution [64–66]. Various approximations of the CME [67–70] and approximate stochastic simulation algorithms [66, 71, 72] provide alternative means for computing probability distributions or statistics of gene expression. Consequently, the CME, the SSA and their approximations are the standard tools for modelling stochastic gene expression.

The rise in single-cell genomics technologies [73] has also prompted advances in computational methods for single-cell data analysis [74, 75]. A wide range of computational tools has been developed, with applications including the discovery of novel cell types [45, 76–78], the reconstruction of gene regulatory networks [79–82] and the prediction of cell state transitions and developmental fate decisions [83–85] from single-cell data. Yet, most existing methods primarily focus on descriptive statistical analyses to explain gene expression patterns, while little emphasis is given on identifying mechanisms of gene regulation.

The intersection of single-cell technologies and advanced mathematical models presents a unique opportunity for the mechanistic investigation of gene expression through data-driven modelling. Weighing quantitative models of gene expression against single-cell data with statistical inference [71, 86] allows to approximate the true generative processes underlying observed expression profiles and gain insights into regulatory mechanisms [80, 87–91]. For instance, fitting models of transcription dynamics to scRNA-seq data enables the estimation of kinetic parameters that govern the mRNA life cycle, thereby shedding light on mechanisms of transcriptional regu-

lation [92–97]. A wide variety of inference methods are available for estimating parameters of biochemical reaction network models. Commonly used methods are maximum likelihood estimation approaches [98], moment-based inference [99–101], likelihood-based Bayesian inference [102] and simulation-based Bayesian inference [103], including Markov chain Monte Carlo (MCMC) [104, 105], *approximate Bayesian computation* (ABC) [106, 107] and machine learning approaches [108].

Despite the unprecedented access to large-scale, high-resolution biological measurements and advanced computational methods, several challenges remain for modelling gene regulation using single-cell data. One of the most critical challenges is that standard single-cell RNA sequencing only provides an instantaneous snapshot of gene expression, posing constraints on dynamic modelling and inference [51, 109]. Quantitative models are typically restricted to steady-state descriptions, as it is hard to identify transient gene expression changes in the data. Also, since kinetic parameters driving gene expression are measured in units of time, it is only possible to infer their relative values across genes [88]. Inference of gene regulatory networks from snapshot data is particularly challenging, as it is hard to detect time-shifted expression changes caused by direct gene interaction from simultaneous transcriptomic profiles [82, 91, 110].

Attempts to address these challenges include the development of trajectory inference methods, that aim to computationally map cells along pseudotemporal trajectories [83, 111, 112]. However, these methods rely on arbitrary metrics of gene expression similarity and have known limitations [113, 114]. An alternative approach to dynamically resolving snapshot data is through RNA velocity methods, that leverage bimodal measurements available from RNA sequencing (nascent and mature mRNA) to infer the directionality of gene expression states [115–118]. Although widely applied to investigate gene regulation dynamics with scRNA-seq [119–125], these methods depend on arbitrary data preprocessing assumptions [126] and lack a rigorous mechanistic foundation [127, 128]. An alternative experimental approach for temporally resolving single-cell sequencing is through *metabolic RNA labelling* [129]. This technique enables tracking the timescales of mRNA production and degradation by chemically modifying mRNA and separately measuring pre-existing and newly synthesised transcripts [130–134]. As a result, it provides a natural means of experimentally accessing genome-wide expression dynamics without relying on computational predictions.

Another important challenge for quantitative modelling is the introduction of confounding

factors of technical variation by single-cell sequencing protocols. Transcriptomic measurements are corrupted by technical noise due to the low detection efficiency of mRNA transcripts in each cell [51, 135]. This noise complicates the accurate quantification of expression patterns and makes it difficult to distinguish between technical errors and true biological variation [136, 137]. To address these issues, several studies have proposed normalisation methods to account for technical biases [137–142], and others have integrated measurement noise models with mechanistic models of gene expression to deconvolute technical and biological sources of stochasticity [94, 136, 143–148]. Another technical source of noise in single-cell sequencing data is batch effects, that arise from measuring cells under multiple experimental conditions [51, 149]. Accounting for such effects typically involves statistical integration approaches that rely on arbitrary distributional assumptions [140, 150, 151].

1.1 Aims and Outline

The primary aim of this PhD thesis is to develop rigorous methodologies for the mechanistic interrogation of gene regulation, and address some of the challenges associated with modelling single-cell gene expression data. We employ mathematical descriptions of biochemical reaction networks, using the chemical master equation and its approximations, to construct mechanistic models of stochastic gene expression and characterise its properties. Building on this mathematical framework, we integrate information from single-cell RNA sequencing data, and account for sources of technical variability. We utilise statistical inference and machine learning techniques to inform our models and gain insights into gene regulation from the data. Additionally, we focus on the complexity associated with models of stochastic reaction networks and develop a method that reduces this complexity. Overall, the methods presented in this thesis can be utilised and further extended to study stochastic gene regulation in a data-driven manner that harnesses the power and scale of single-cell genomics technologies.

The Chapter 2 of this thesis focuses on modelling gene expression dynamics using time-resolved scRNA-seq data. It is known that the global control of gene expression is closely linked to the cell cycle [152, 153], a fundamental biological process during which cells grow, replicate their genome and divide [154]. Yet, it is unclear how noisy transcription dynamics is coordinated throughout the cell cycle. In this chapter, we develop a model of gene expression in growing

and dividing cells to investigate properties of cell cycle-dependent transcription regulation. To inform our model, we leverage data from a time-resolved scRNA-seq protocol, scEU-seq [155], that enriches single-cell transcriptomes with metabolic RNA labelling and cell cycle ordering information. In section 2.2, we describe the scEU-seq dataset and present a model of gene expression that integrates its temporal features, including RNA labelling timepoints and cell cycle stages. We develop a methodology to estimate and account for technical noise effects in the data, enabling reliable quantification of biological contributions to transcriptional noise. In section 2.3, we introduce a Bayesian inference framework, that allows testing multiple hypotheses describing the dependence of bursty transcription and mRNA degradation on the cell cycle and cell size. This involves inference of individual model parameters and model selection to weigh the competing models against the data. Additionally, we present an approach for testing the robustness of our inference using synthetic data. In section 2.4, we apply our method to reveal global patterns of transcriptional regulation from scEU-seq data. We demonstrate how our model exploits dynamical patterns in time-resolved data to identify distinct regulatory mechanisms. This allows us to pinpoint groups of genes associated with different modes of cell cycle- and cell size-dependent regulation and analyse their properties.

Chapter 3 expands beyond the gene-specific modelling approach of Chapter 2, focusing on modelling gene regulatory interactions. Genes are functional units encoding cellular functions and operate as part of regulatory networks to orchestrate these functions [1, 156]. Gene regulatory networks are core determinants of cell fate decisions and cellular diversity [157, 158], and decoding these networks is a fundamental question in biological research. It is a well-established belief that gene interactions contribute significantly to gene expression stochasticity [4, 6] and this stochasticity has been proposed as a tool to investigate gene regulation [27, 91, 159]. In this chapter, we utilise cell-to-cell transcriptional variation revealed by single-cell RNA sequencing data to explore pairwise gene relationships. In section 3.2, we develop a stochastic model to describe dynamics of gene co-expression arising from different modes of regulatory interaction. Similar to Chapter 2, we leverage the time-resolved scEU-seq dataset [155] to distinguish temporal components of gene-gene covariation. This allows us to define dynamical correlations of gene pairs across RNA labelling timepoints. In section 3.3, we introduce a machine learning framework that exploits these correlations to identify links between observed statistics and underlying regulatory models. This framework effectively combines generative modelling with supervised learning using synthetic data, to associate gene pairs with potential regulatory re-

relationships. In section 3.4, we demonstrate that our framework can identify modes of gene regulation in synthetic data. We provide interpretability results to explain the model’s predictions and highlight conditions under which the different modes of regulation are distinguishable. Finally, we apply our model to scEU-seq data and predict gene regulatory relationships from observed gene-pair statistics.

Chapter 4 shifts the focus from modelling gene regulation to more broadly modelling stochastic biochemical networks. At the molecular level, all cellular processes, including gene regulation, are governed by biochemical reactions that are inherently stochastic [21]. Formulating and simulating stochastic models to describe the dynamics of these networks relies on the chemical master equation (CME) and the Gillespie simulation algorithm (SSA) [63, 66]. The analytical intractability of the CME and the computational inefficiency of the SSA for complex biochemical networks, have motivated the development of approximation methods aimed at reducing model complexity and increasing efficiency [86, 160]. Some of these methods involve continuous approximations of the CME, such as the linear noise approximation (LNA) [161] and the chemical Langevin equation (CLE) [68], that enable tractability of biochemical network statistics. Other approximation methods exploit timescale separation conditions to reduce model dimensionality and derive approximate simulation algorithms [162, 163]. In this chapter, we develop a novel approximation method based on the chemical Langevin equation, for stochastic reaction networks characterised by well-separated timescales. In section 4.2, we provide the full derivation of this method. We show that the resulting reduced descriptions can reliably capture the stochasticity revealed by the exact CME description, improving upon existing heuristic stochastic approximations that exhibit limitations. In section 4.3, we apply our method to a reaction network of enzyme kinetics. We demonstrate errors between the exact SSA statistics and statistics obtained with a heuristic approximation, and highlight that our method corrects for these errors, increasing approximation accuracy.

Chapter 5 summarises the main findings of this thesis and highlights the most important contributions. It provides links between the different chapters and discusses how the presented methodologies can be combined for studying gene regulation through a mathematical and data-driven modelling approach. Finally, it discusses ongoing work and points to future research directions.

Chapter 2

Bayesian inference of transcription regulation from time-resolved single-cell transcriptomics

2.1 Introduction

Transcription occurs in episodic bursts of activity, causing fluctuations in the level of transcripts produced by a single gene [3, 40, 164]. The frequency and size of these bursts lead to diverse temporal patterns of gene activity [165, 166], often referred to as cellular noise, that can have a functional role in gene regulatory networks and cell fate decisions [10, 13–16, 18, 21, 167, 168]. The cell cycle has been shown to contribute significantly to transcript variability [169–173], either through random partitioning of molecules at cell division [174–176], modulation of transcription rates [58, 93, 155, 177, 178] or through coupling to cell size as a mechanism of transcript homeostasis [179–182]. Yet, it remains elusive what the global drivers of transcription dynamics are on a transcriptome-wide scale.

Single-cell RNA sequencing (scRNA-seq) has become the de facto standard in profiling transcription patterns [44]. The method allows quantifying distributions of transcript levels across cells that can be harnessed to learn about gene regulation [47]. A few studies have started to investigate how transcriptional processes are regulated through synergising stochastic modelling and inference [88, 92, 94, 96, 183]. To this end, stochastic models of reaction kinetics inside

living cells are simulated using the Gillespie algorithm or solved analytically and matched with experimental mRNA counts using statistical inference [54, 57, 58, 99, 184, 185]. However, a challenge is that not one model fits it all at a genome-wide scale, and identifying transcriptional patterns requires weighing competing model hypotheses against each other that need to be tested against experiments.

In principle, this aim can be achieved using Bayesian inference by computing the posterior distribution over models and their corresponding parameters. The posterior model probability weighs the model likelihood by its prior distribution. Several problems with this approach appear in practice: (i) the likelihood of mechanistic models is intractable or is computationally expensive to optimise [99], (ii) scRNA-seq data is heavily corrupted by technical noise including dropout [141], (iii) absolute kinetic rates are experimentally inaccessible from static snapshots [51].

In the literature, these problems are typically dealt with separately. The first issue is addressed by replacing likelihoods with moment- or simulation-based methods [92, 94, 96, 183, 186]. As an example, the authors of [96] present an approximate Bayesian computation framework to infer transcriptional burst kinetics from scRNA-seq data. Yet, the above methods either neglect technical noise of sequencing protocols or remain computationally challenging as they need to be performed for several competing models and many genes. The second issue has been addressed by rescaling transcript counts [94, 137, 141, 183], but Bayesian normalisation has been shown to capture distributions more truthfully [145, 146]. To this end, a recent study [94] demonstrated that modelling technical noise can improve inference of transcriptional burst kinetics, yet it does not explore multiple modelling hypotheses. The third issue remains a major obstacle for single-cell analysis, as traditional sequencing protocols can only capture a static snapshot of cells in time, thus prohibiting absolute quantification of transcription and mRNA turnover rates.

Recently, metabolic RNA labelling has become available to temporally resolve conventional single-cell transcriptomics [129]. These protocols allow monitoring of mRNA synthesis and degradation in single cells, thus capturing temporal transcription dynamics [155, 178, 187–193]. This provides new possibilities for modelling as it enables absolute rate quantification, as demonstrated recently for the inference of transcriptional burst kinetics through maximum likelihood methods [194] or MCMC sampling from the posterior distribution [95]. These ap-

proaches have either neglected technical noise or have been limited to relatively simple models and have not explored the landscape of the model posterior. It thus remains elusive which models describe gene expression on a transcriptome-wide scale.

In this chapter, we develop a scalable approach to Bayesian inference, normalisation and model selection of bursty gene expression dynamics along the cell cycle. A stochastic model enables absolute quantification of transcription dynamics along the cell cycle, resolving mRNA synthesis, bursting and degradation processes from scRNA-seq data with metabolic labelling and cell cycle reporters [155]. We demonstrate that correlations between unlabelled and labelled mRNA levels can delineate different modes of transcriptional regulation and quantify the contributions from the cell cycle and mRNA synthesis, bursting, and degradation kinetics to the overall cell-to-cell variation. The framework allows us to discover which of these processes are modulated by the cell cycle and identify genes with distinct transcription regulation. Our study thus reveals, for the first time, how genome-wide single-cell transcriptional dynamics is orchestrated with global factors such as cell size and cell cycle. The findings of this chapter are also reported in a jointly-authored publication [195].

2.2 Stochastic modelling of time-resolved single-cell RNA sequencing

Our first aim is to develop a stochastic model of gene expression that harnesses the rich information of time-resolved scRNA-seq data. We start with a detailed description of the scEU-seq dataset [155] and its temporal features, as well as we specify data pre-processing steps and summary statistics extracted from the data. We then define a stochastic model that effectively integrates the data features and provides a versatile framework for quantifying transcriptional variation.

2.2.1 Time-resolved scRNA-seq data

Unlabelled and labelled mRNA count matrices

A standard scRNA-seq dataset can be summarised in a single matrix $\mathbf{X} \in \mathbb{N}^{n \times d}$ that contains the mRNA counts of d genes in n cells. Time-resolved scRNA-seq protocols with metabolic RNA labelling are multi-modal, providing with at least 2 mRNA count measurements for each gene in each single cell. Specifically, we are focusing on the scEU-seq dataset that includes 4 mRNA count matrices for $d = 11848$ genes across $n = 5422$ cells: unlabelled unspliced (**UU**), unlabelled spliced (**US**), labelled unspliced (**LU**) and labelled spliced (**LS**) mRNA counts. Unspliced mRNA refers to nascent or pre-mature transcripts that include both exons and introns, while spliced mRNA refers to mature transcripts that have been processed, with all introns removed - a process that typically occurs on the timescale of minutes [115]. It is important to note that here we use the count matrices reported in the original study and that the scEU-seq protocol for transcript quantification only distinguishes between unspliced and spliced reads and does not account for ambiguous reads (see for example [196]). Additionally, as our study is focused on unlabelled and labelled mRNA and does not account for RNA splicing, we aggregate the unspliced and spliced count matrices. This aggregation gives us an estimate of the total unlabelled and total labelled mRNA counts, summarised in the matrices $\mathbf{U} = \mathbf{UU} + \mathbf{US}$ and $\mathbf{L} = \mathbf{LU} + \mathbf{LS}$, respectively.

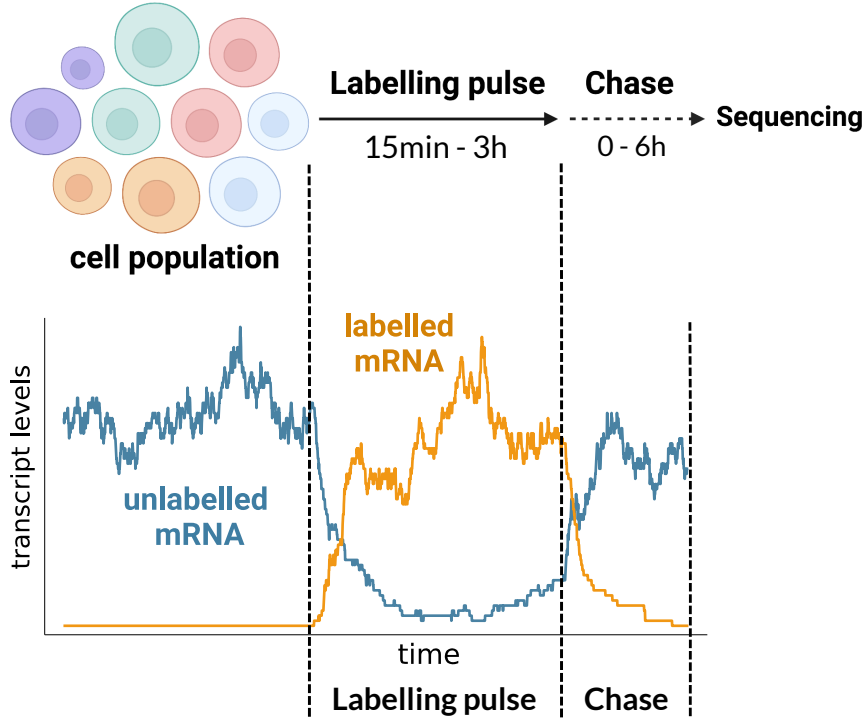


Figure 2.1: **Time-resolved single-cell RNA sequencing.** Illustration of the single-cell EU RNA sequencing (scEU-seq) protocol, developed by [155]. A heterogeneous cell population is pulse-chase labelled prior to sequencing. An example of stochastic transcription of a gene in a single cell is shown below, where labelled mRNA is produced and degraded following the timing of pulse-chase labelling. Varying the duration of the pulse period (15 minutes to 3 hours) and the chase period (0 to 6 hours) is informative of the timescales of mRNA turnover.

Pulse-chase metabolic labelling

The scEU-seq protocol labels transcripts with a pulse of a certain duration, in which cells are exposed to a chemical analogue of uridine (EU) that is incorporated into nascent RNA, causing nucleotide conversion [129, 155]. Pulse labelling is followed by a chase period, in which cells are treated with uridine (U) to chemically "invert" the labelling process. After the pulse-chase experiment, transcripts are sequenced (Figure 2.1). This procedure is performed several times, by varying the durations of pulse and chase treatment. The different combinations of pulse and chase labelling allow to investigate diverse timescales of mRNA synthesis and degradation.

The overall protocol results in a dataset of distinct subpopulations of cells that were treated with different metabolic labelling conditions. These include the *pulse* conditions (c_{pulse}) that involve pulse labelling for time t_{pulse} and zero chase period, as well as the *chase* conditions (c_{chase}) that involve pulse labelling for $t_{pulse} = 22$ hours followed by a chase experiment for time t_{chase} . The different combinations of (t_{pulse}, t_{chase}) are outlined in the following tables (time in

hours):

$$c_{pulse} = \begin{bmatrix} t_{pulse} : & 0.25 & 0.5 & 0.75 & 1 & 2 & 3 \\ t_{chase} : & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (2.1a)$$

$$c_{chase} = \begin{bmatrix} t_{pulse} : & 22 & 22 & 22 & 22 & 22 \\ t_{chase} : & 0 & 1 & 2 & 4 & 6 \end{bmatrix} \quad (2.1b)$$

Note that the above conditions are ordered with respect to the total pulse-chase treatment time $t_{pulse} + t_{chase}$. By denoting with t the particular labelling condition (t_{pulse}, t_{chase}) , we can define temporal trajectories indexed by t , which we will refer to as "labelling time", for simplicity. Temporal trajectories across labelling time are the key feature distinguishing the scEU-seq protocol [155] from other metabolic labelling protocols and conventional scRNA-seq technologies, and enable a fully dynamical analysis of gene expression.

Cell cycle progression information

The original scEU-seq study provides an estimated cell cycle ordering that assigns each cell to a relative cell cycle position along the interval $(0, 1]$ based on the fluorescence of cell cycle reporters. We rely on these estimates and divide the 5422 cells into 5 groups by partitioning this ordering trajectory into 5 sub-intervals $(0, 1/5], (1/5, 2/5], \dots, (4/5, 1]$. According to the fluorescence of cell cycle reporters (Figure 2.2), we assume that the first two intervals correspond to G1-G1/S phase, the next two intervals correspond to S-S/G2 phase and the last interval corresponds to G2/M phase. We then use the proportion of cells across cell cycle phases to define an empirical *age distribution*:

$$\text{age distribution} = \left\{ \frac{N_\theta}{N} : \theta = \text{G1, G1/S, S, S/G2, G2/M} \right\} \quad (2.2)$$

where N_θ denotes the number of cells in cell cycle phase θ and $N = 5422$. Additionally, each cell cycle phase includes cells from different labelling conditions (as described in the previous paragraph) and therefore we define age distributions per labelling time, $\left\{ \frac{N_{\theta,t}}{N_t} \right\}$, where N_t is the number of cells in condition t and $N_{\theta,t}$ is the number of cells in condition t and cell cycle phase θ .

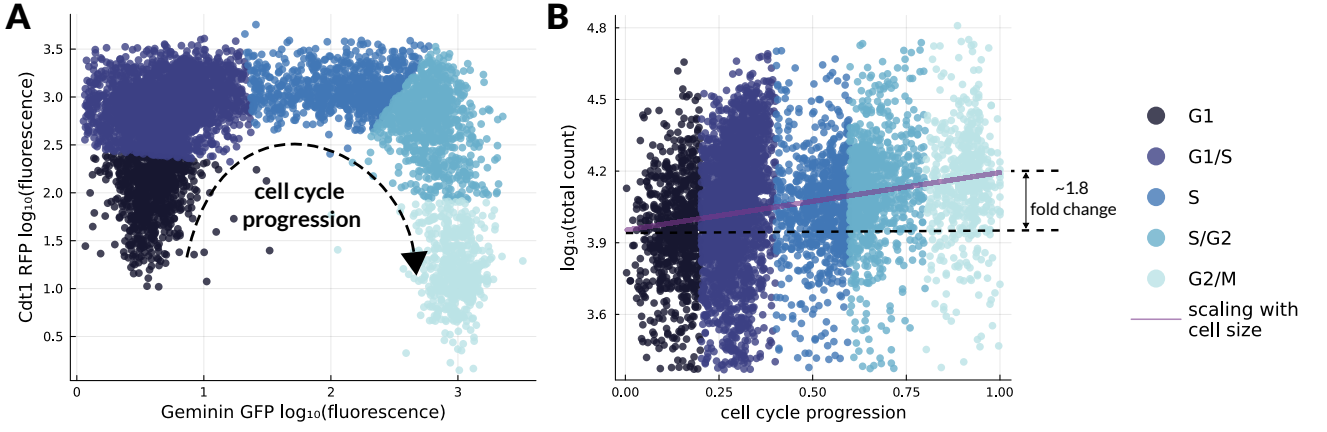


Figure 2.2: **Definition of cell cycle phases and observed scaling of total count with cell size.** (A) Scatter plot of log-fluorescence of the cell cycle markers Geminin-GFP and Cdt1-RFP as reported in the original scEU-seq study [155]. The colours show the assignment of cells to different cell cycle phases. (B) Observed cell-specific total mRNA counts along cell cycle progression, with respect to the cell cycle trajectory reported in [155]. The purple line shows a linear regression fit and the almost 2-fold increase indicates the scaling of total count with cell size.

Gene filtering

We impose certain criteria to select genes for our downstream analyses. Starting from 11848 genes of the scEU-seq dataset, we first select genes with at least 500 labelled mRNA counts across all cells (similarly to the original study [155]), which gives us a subset of 6087 genes. Then, we impose an additional criterion to filter out genes with very low expression. We require that each gene has at least 1 count on average in at least one cell cycle phase. That is, if $\mathbf{X} = (x_{ij}) = \mathbf{U} + \mathbf{L}$ is the total (unlabelled + labelled) mRNA count matrix (where cells are indexed with i and genes are indexed with j), then the gene-specific condition is that

$$\frac{1}{N_\theta} \sum_{i \in \theta} x_{ij} \geq 1 \text{ for at least one cell cycle phase } \theta \in \{\text{G1, G1/S, S, S/G2, G2/M}\} \quad (2.3)$$

This criterion gives us the final set of 3419 genes used for further analysis.

Summary statistics

We now describe how we compute summary statistics from the data. We consider the count matrices $\mathbf{U} = (u_{ij})$, $\mathbf{L} = (l_{ij})$ of unlabelled and labelled mRNA, as well as the count matrix $\mathbf{X} = (x_{ij}) = \mathbf{U} + \mathbf{L}$ of total mRNA. The summary statistics of a gene j in the data depend

on the cell cycle phase $\theta \in \{\text{G1}, \text{G1/S}, \text{S}, \text{S/G2}, \text{G2/M}\}$ or on the labelling condition $c \in \{c_{\text{pulse}}, c_{\text{chase}}\}$. The definitions of gene-specific summary statistics are outlined in Table 2.1. In these definitions, the notation $i \in c$ refers to either all *pulse*-treated cells or all *chase*-treated cells, while the notation $i \in t$ refers to a specific labelling condition $(t_{\text{pulse}}, t_{\text{chase}})$, where $t = t_{\text{pulse}} + t_{\text{chase}}$ denotes the total duration of an experiment (see paragraph 2.2.1).

Standard errors of summary statistics

Aside from computing summary statistics from data, we also estimate standard errors for each statistic to quantify the measurement bias. If $s_{\text{obs}} = s_{\text{obs}}(y)$ is a summary statistic of the data y , we estimate the standard error of s_{obs} with

$$\widehat{\text{se}}(s_{\text{obs}}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B \left(s_{\text{obs}}^{(b)} - \frac{1}{B} \sum_{j=1}^B s_{\text{obs}}^{(j)} \right)^2} \quad (2.4)$$

where $s_{\text{obs}}^{(b)}$ is the value of the summary statistic that we obtain from each bootstrapped dataset $y^{(b)}$, $b = 1, \dots, B = 100$. These standard errors are used during parameter inference to normalise the magnitudes of different summary statistics, as it will be described later in 2.3.1.

Table 2.1: **Definitions of summary statistics.** (A) Theoretical model-based definitions. Dependence on the cell cycle is denoted by τ and dependence on labelling condition is denoted by t . Unlabelled, labelled and total mRNA levels of a single gene are denoted by u , l and $x = u + l$, respectively. (B) Data-based definitions. Notation is explained in paragraph 2.2.1.

	A. Model	B. Data
(1)	$\mathbb{E}(x \mid \tau)$	$\bar{x}_j(\theta, c) = \frac{1}{N_{\theta, c}} \sum_{i \in \theta \cap c} x_{ij}$
(2)	$\frac{\text{Var}(x \mid \tau)}{\mathbb{E}(x \mid \tau)}$	$\frac{\frac{1}{N_{\theta, c}-1} \sum_{i \in \theta \cap c} (x_{ij} - \bar{x}_j(\theta, c))^2}{\bar{x}_j(\theta, c)}$
(3)	$\frac{\mathbb{E}_\tau(\mathbb{E}(l \mid \tau, t))}{\mathbb{E}_\tau(\mathbb{E}(u \mid \tau, t) + \mathbb{E}(l \mid \tau, t))}$	$\frac{\frac{1}{N_t} \sum_{i \in t} l_{ij}}{\frac{1}{N_t} \sum_{i \in t} x_{ij}}$
(4)	$\frac{\mathbb{E}_\tau(\text{Cov}(u, l \mid \tau, t))}{\sqrt{\text{Var}(u \mid t) \text{Var}(l \mid t)}}$	$\frac{\sum_\theta \frac{N_{\theta, t}}{N_t} \sum_{i \in \theta \cap t} (u_{ij} - \bar{u}_j(\theta, t))(l_{ij} - \bar{l}_j(\theta, t))}{\left(\frac{1}{N_t} \sum_{i \in t} (u_{ij} - \bar{u}_j(t))^2 \frac{1}{N_t} \sum_{i \in t} (l_{ij} - \bar{l}_j(t))^2 \right)^{1/2}}$
(5)	$\frac{\text{Cov}_\tau(\mathbb{E}(u \mid \tau, t), \mathbb{E}(l \mid \tau, t))}{\sqrt{\text{Var}(u \mid t) \text{Var}(l \mid t)}}$	$\frac{\sum_\theta \frac{N_{\theta, t}}{N_t} \left(\bar{u}_j(\theta, t) - \sum_\theta \frac{N_{\theta, t}}{N_t} \bar{u}_j(\theta, t) \right) \left(\bar{l}_j(\theta, t) - \sum_\theta \frac{N_{\theta, t}}{N_t} \bar{l}_j(\theta, t) \right)}{\left(\frac{1}{N_t} \sum_{i \in t} (u_{ij} - \bar{u}_j(t))^2 \frac{1}{N_t} \sum_{i \in t} (l_{ij} - \bar{l}_j(t))^2 \right)^{1/2}}$

Table 2.2: **Interpretation of summary statistics.**

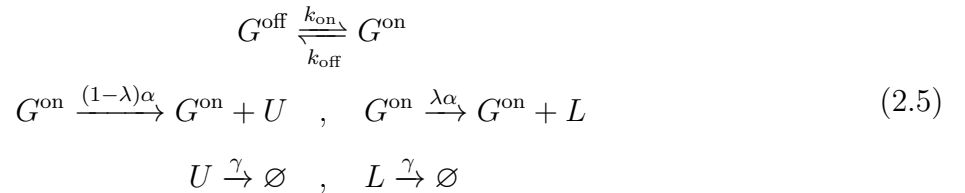
	Interpretation
(1)	cell cycle-dependent mean of mRNA levels
(2)	cell cycle-dependent Fano factor of mRNA levels
(3)	ratio of mean labelled vs total mRNA levels over labelling time
(4)	cell cycle-averaged correlation of labelled-unlabelled mRNA levels over labelling time
(5)	correlation of cell cycle-dependent labelled & unlabelled means over labelling time

2.2.2 Stochastic modelling framework

In this section, we develop a stochastic model of cell cycle-dependent transcription dynamics in growing and dividing cells, that additionally captures the effects of metabolic RNA labelling. This provides with an effective description of the underlying generative process of mRNA count data under the scEU-seq protocol. We assume reaction kinetics based on a two-stage telegraph model of bursty synthesis and degradation of labelled and unlabelled mRNA, with cell cycle and cell size-dependent rates. Our model postulates labelling occurs on nascent mRNA [197] and that labelled and unlabelled mRNA are produced with rates proportional to a gene-specific labelling efficiency. We then formulate moment equations that describe the temporal dependence of mean and covariance of labelled and unlabelled mRNA on cell cycle progression and cell division dynamics, as well as on pulse-chase labelling conditions. As we will describe in detail in paragraph 2.2.2, these equations track mRNA statistics over time, capturing the random partitioning of molecules at cell division as well as the effect of pulse-chase labelling. The cell cycle-dependent changes of mRNA statistics and kinetic rates are described in five distinct cell cycle phases (as shown in Figure 2.2 and paragraph 2.2.1).

Model of bursty transcription of unlabelled and labelled mRNA

We assume that mRNA molecules are produced according to a telegraph model of gene expression [60], capturing random switching of the gene promoter between active and inactive states with rates k_{on} and k_{off} respectively, mRNA synthesis with rate α and mRNA decay with rate γ . During pulse-chase labelling, labelled and unlabelled mRNA molecules are synthesised with rates $\lambda\alpha$ and $(1-\lambda)\alpha$ respectively, where λ denotes the gene-specific labelling efficiency. These reaction events are encoded in the following scheme and illustrated in Figure 2.3:



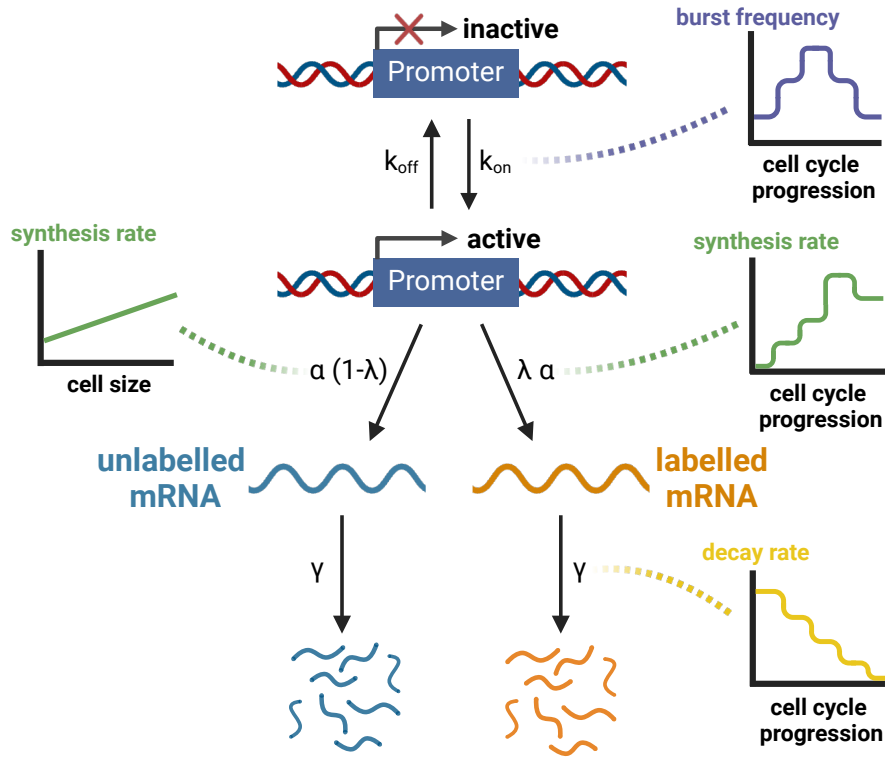


Figure 2.3: **Telegraph model of transcription with cell cycle- and cell size-dependent rates.** Illustration of the reaction kinetics model (equation (2.5)) describing bursty transcription of unlabelled and labelled mRNA with cell cycle-dependent kinetic rates. The promoter activation rate (burst frequency), synthesis rate and decay rate are cell cycle-dependent.

Cell cycle and cell size-dependent kinetic rates

We further allow for three of the reaction rates to vary over the cell cycle (Figure 2.3). As a result, they are time-dependent periodic functions $k_{\text{on}} = k_{\text{on}}(t), \alpha = \alpha(t), \gamma = \gamma(t)$, for $t_{\text{birth}}^{(n)} \leq t \leq t_{\text{division}}^{(n)}$, $n = 1, 2, \dots$ (with n representing the n^{th} cell generation in a single-cell lineage). As they are periodic functions, their value does not depend on n and their cell cycle-dependence can be equivalently written as

$$k_{\text{on}} = k_{\text{on}}(\tau), \alpha = \alpha(\tau), \gamma = \gamma(\tau), \quad 0 \leq \tau \leq T_d \quad (2.6)$$

where τ is the cell age (time along cell cycle progression) and $T_d = t_{\text{division}}^{(n)} - t_{\text{birth}}^{(n)}$ is the interdivision time (cell cycle length). The rate k_{off} is always assumed to be constant. Another assumption on the kinetics is that the synthesis rate α can scale linearly with cell size [179, 198, 199] (Figure 2.3). This is in accordance with observations from the data (Figure 2.2), which allows us to assume that cells follow linear growth. This implies that the cell size s at age τ is:

$$s(\tau) = s_0 \left(1 + \frac{\tau}{T_d} \right) \quad (2.7)$$

where s_0 is the cell size at birth. Accordingly, we model the dependence of synthesis rate on cell size as:

$$\alpha(s(\tau), \tau) = \alpha(\tau) \left(1 + \frac{\tau}{T_d} \right) \quad (2.8)$$

where $\alpha(\tau)$ describes the dependence of the synthesis rate on cell cycle progression without size-scaling. Additionally, we define the burst kinetics as follows:

$$\begin{aligned} \text{burst frequency} &= k_{\text{on}} \\ \text{burst size} &= \frac{\alpha}{k_{\text{off}}} \end{aligned} \quad (2.9)$$

As a consequence, both burst frequency and burst size can be cell cycle-dependent and additionally the burst size could be size-dependent (except for the constant non-scaling model), in accordance to the synthesis rate. The unit of all kinetic parameters is h^{-1} , apart from burst size which is unitless. Finally, we define the mRNA half-life $t_{1/2}$ (measured in hours) that

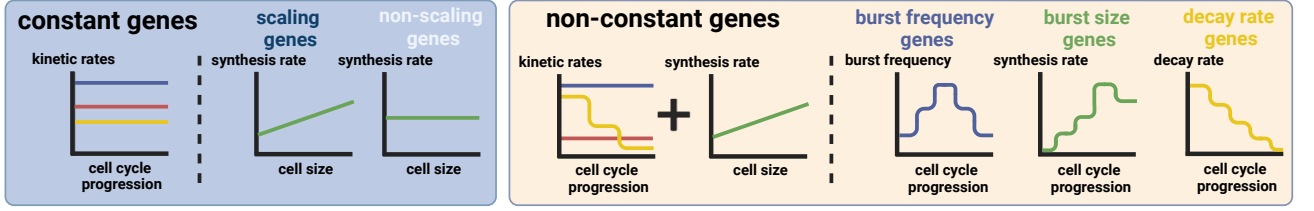


Figure 2.4: **Cartoon overview of the different model - gene classes.** Constant models (blue box) assume that synthesis rate either scales with cell size (scaling genes) or stays constant (non-scaling genes), while all other kinetic rates stay constant along the cell cycle. Non-constant models (yellow box) assume that synthesis rate scales with cell size and additionally one kinetic rate varies with cell cycle progression, while the rest of the kinetic rates stay constant. The cartoon plots illustrate that varying kinetic rates are piece-wise constant over the five cell cycle stages (see paragraph 2.2.1).

corresponds to a constant degradation rate γ as:

$$t_{1/2} = \frac{\ln(2)}{\gamma} \quad (2.10)$$

Defining multiple modes of transcriptional regulation

Using our model of transcription with time-dependent rates, we define five distinct model classes representing hypotheses about specific cell size- and cell cycle-dependent regulation of kinetic rates (Figure 2.4). Two of the model classes, which we refer to as *constant* models, assume that the synthesis rate either scales linearly with cell size or stays constant, while the rest of the kinetic rates are assumed to be constant along the cell cycle. The other three model classes, which we refer to as *non-constant* models, assume scaling of the synthesis rate with cell size and additionally that one of the kinetic rates, either burst frequency or synthesis rate or decay rate, varies with cell cycle progression. We refer to the three non-constant models as the *burst frequency* model, the *burst size* model and the *decay rate* model, respectively. The two constant models identify genes for which the kinetics are conserved through the cell cycle and distinguish between size-scaling and size-independent expression. The three non-constant models identify genes for which there is a kinetic rate being dominantly modulated through the cell cycle. As we will see in following sections, our inference and downstream analyses aim to highlight distinctive biological properties of the five different models and the groups of genes that we will associate with those models.

Moment equations with time-dependent kinetic rates

The time-dependent joint probability distribution $P(\mathbf{x}, t)$ of molecule numbers $\mathbf{x} = (x_1, x_2, x_3)^T$ of the respective species $\{G^{\text{on}}, U, L\}$ in the reaction scheme (2.5) obeys the Chemical Master Equation (CME):

$$\frac{d}{dt}P(\mathbf{x}, t) = \sum_{r=1}^6 \left(w_r(\mathbf{x} - \mathbf{S}_r, t) P(\mathbf{x} - \mathbf{S}_r, t) - w_r(\mathbf{x}, t) P(\mathbf{x}, t) \right) \quad (2.11)$$

where $\mathbf{S}_r = (\mathbf{S}_{1r}, \mathbf{S}_{2r}, \mathbf{S}_{3r})^T$ denotes the r -th column of the stoichiometric matrix

$$\mathbf{S} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix} \quad (2.12)$$

of the reaction system (with $\mathbf{S}_{ir}$ being the change in molecule number of species i by reaction r) and $w_r(\mathbf{x}, t)$ denotes the propensity of reaction r , which is the probability per unit time for the r -th reaction to occur. The propensities of the network are

$$\begin{aligned} w_1(\mathbf{x}, t) &= k_{\text{on}}(t)(1 - x_1), \quad w_2(\mathbf{x}, t) = k_{\text{off}}(t)x_1, \quad w_3(\mathbf{x}, t) = (1 - \lambda(t))\alpha(t)x_1, \\ w_4(\mathbf{x}, t) &= \lambda(t)\alpha(t)x_1, \quad w_5(\mathbf{x}, t) = \gamma(t)x_2, \quad w_6(\mathbf{x}, t) = \gamma(t)x_3 \end{aligned} \quad (2.13)$$

Note that the gene promoter state x_1 is either 0 or 1 so it is sufficient to only describe the state of G^{on} in the equations. In our model, we are interested in the time-evolution of the 1st and 2nd order central moments of the molecule abundances, which can be derived from the CME [99]:

$$\frac{d}{dt}\mu_i(t) = \sum_{r=1}^6 \mathbf{S}_{ir} w_r(\boldsymbol{\mu}, t) \quad (2.14)$$

$$\frac{d}{dt}\Sigma_{ij}(t) = \sum_{\substack{r=1, \dots, 6 \\ k=1, \dots, 4}} \left(\mathbf{S}_{ir} \frac{\partial w_r(\boldsymbol{\mu}, t)}{\partial \mu_k} \Sigma_{kj}(t) + \mathbf{S}_{jr} \frac{\partial w_r(\boldsymbol{\mu}, t)}{\partial \mu_k} \Sigma_{ki}(t) \right) + \sum_{r=1}^6 \mathbf{S}_{ir} \mathbf{S}_{jr} w_r(\boldsymbol{\mu}, t) \quad (2.15)$$

where $\mu_i = \mathbb{E}(x_i)$ is the mean molecule number of species i and $\boldsymbol{\Sigma} = (\Sigma_{ij})$ is the covariance matrix with $\Sigma_{ij} = \mathbb{E}((x_i - \mu_i)(x_j - \mu_j))$ being the covariance of molecule numbers of species i and j and the diagonal elements $\Sigma_{ii} = \text{Var}(x_i)$ being the variance of molecule number of species i . Note that since the reactions in (2.5) are linear, the propensities are linear functions of the abundances x_i , that is $\mathbb{E}(w_r(\mathbf{x}, t)) = w_r(\mathbb{E}(\mathbf{x}), t) = w_r(\boldsymbol{\mu}, t)$, and therefore the system of

moment equations is closed (does not depend on higher order moments). Given our assumption about cell cycle-dependence of the reaction rate constants, the propensities are periodic functions of time. As a consequence, the moment equations are constrained by periodic boundary conditions.

Coupling to cell division dynamics

To determine the boundary conditions, we model the random partitioning of molecules at cell division as a binomial process, which gives us the following conditions for the mean, variance and covariance of unlabelled and labelled mRNA at the n -th cell division:

$$\mu_i(t_{\text{birth}}^{(n)}) = \frac{\mu_i(t_{\text{division}}^{(n-1)})}{2}, \quad i = 2, 3 \quad (2.16)$$

$$\Sigma_{ij}(t_{\text{birth}}^{(n)}) = \frac{\Sigma_{ij}(t_{\text{division}}^{(n-1)})}{4} + \delta_{ij} \frac{\mu_i(t_{\text{division}}^{(n-1)})}{4}, \quad i, j = 2, 3 \quad (2.17)$$

The gene promoter state $g^{(\text{birth})}$ in a daughter cell at birth is identical to the promoter state $g^{(\text{division})}$ in the mother cell at division, and therefore the mean and variance of the promoter state remain constant before and after division:

$$\mu_1(t_{\text{birth}}^{(n)}) = \mu_1(t_{\text{division}}^{(n-1)}), \quad \Sigma_{11}(t_{\text{birth}}^{(n)}) = \Sigma_{11}(t_{\text{division}}^{(n-1)}) \quad (2.18)$$

However, the covariance of the promoter state with the mRNA decreases due to partitioning of mRNA molecules:

$$\Sigma_{12}(t_{\text{birth}}^{(n)}) = \frac{\Sigma_{12}(t_{\text{division}}^{(n-1)})}{2}, \quad \Sigma_{13}(t_{\text{birth}}^{(n)}) = \frac{\Sigma_{13}(t_{\text{division}}^{(n-1)})}{2} \quad (2.19)$$

To prove the above, for notation simplicity, m denotes mRNA (unlabelled or labelled) and g denotes the gene promoter state. Also, m_0, g_0 denote the mRNA and gene states in a daughter cell at birth and m_T, g_T denote the mRNA and gene states in the mother cell at cell division. At birth, a daughter cell inherits m_0 mRNA molecules out of total m_T molecules of the mother

cell according to a Binomial distribution:

$$P(m_0 \mid m_T) = \binom{m_T}{m_0} p^{m_0} (1-p)^{m_T-m_0} \quad (2.20)$$

with $p = 1/2$. Also, the gene promoter state is not altered at cell division and therefore:

$$g_0 \mid g_T \sim \delta(g_T) \quad (2.21)$$

Using the law of total expectation, the expected number of mRNA at cell birth is:

$$\mathbb{E}(m_0) = \mathbb{E}(\mathbb{E}(m_0 \mid m_T)) = \frac{\mathbb{E}(m_T)}{2} \quad (2.22)$$

Using the law of total variance, the variance of mRNA molecules at birth is:

$$\text{Var}(m_0) = \mathbb{E}(\text{Var}(m_0 \mid m_T)) + \text{Var}(\mathbb{E}(m_0 \mid m_T)) \quad (2.23a)$$

$$= \mathbb{E}\left(\frac{m_T}{4}\right) + \text{Var}\left(\frac{m_T}{2}\right) \quad (2.23b)$$

$$= \frac{1}{4}(\mathbb{E}(m_T) + \text{Var}(m_T)) \quad (2.23c)$$

Denoting with u, l the unlabelled and labelled mRNA and using the law of total covariance, we have that the covariance of unlabelled and labelled mRNA at birth is:

$$\text{Cov}(u_0, l_0) = \mathbb{E}(\text{Cov}(u_0, l_0 \mid u_T, l_T)) + \text{Cov}(\mathbb{E}(u_0 \mid u_T, l_T), \mathbb{E}(l_0 \mid u_T, l_T)) \quad (2.24a)$$

$$= 0 + \text{Cov}(\mathbb{E}(u_0 \mid u_T), \mathbb{E}(l_0 \mid l_T)) \quad (2.24b)$$

$$= \text{Cov}\left(\frac{u_T}{2}, \frac{l_T}{2}\right) \quad (2.24c)$$

$$= \frac{1}{4} \text{Cov}(u_T, l_T) \quad (2.24d)$$

where in (2.24b) we use that the conditional variables $u_0 \mid u_T$ and $l_0 \mid l_T$ are independent.

The covariance of the mRNA and gene states at birth is given by:

$$\text{Cov}(m_0, g_0) = \mathbb{E}(\text{Cov}(m_0, g_0 \mid m_T, g_T)) + \text{Cov}(\mathbb{E}(m_0 \mid m_T, g_T), \mathbb{E}(g_0 \mid m_T, g_T)) \quad (2.25a)$$

$$= 0 + \text{Cov}(\mathbb{E}(m_0 \mid m_T), \mathbb{E}(g_0 \mid g_T)) \quad (2.25b)$$

$$= \text{Cov}\left(\frac{m_T}{2}, g_T\right) \quad (2.25c)$$

$$= \mathbb{E}\left(\frac{m_T}{2} g_T\right) - \mathbb{E}\left(\frac{m_T}{2}\right) \mathbb{E}(g_T) \quad (2.25d)$$

$$= \frac{1}{2} \text{Cov}(m_T, g_T) \quad (2.25e)$$

where in (2.25b) we use that the conditional variables $m_0 \mid m_T$ and $g_0 \mid g_T$ are independent.

Partitioning of 3rd order moments

In the case where we want to extend the moment equations of the model up to 3rd order, we need additional formulas for partitioning the 3rd order central moments. Starting from the 3rd central moment of the mRNA, we use the law of total cumulance to get:

$$\mu_3(m_0) = \mathbb{E}(\mu_3(m_0 \mid m_T)) + \mu_3(\mathbb{E}(m_0 \mid m_T)) + 3 \text{Cov}(\mathbb{E}(m_0 \mid m_T), \text{Var}(m_0 \mid m_T)) \quad (2.26a)$$

$$= 0 + \mu_3\left(\frac{m_T}{2}\right) + 3 \text{Cov}\left(\frac{m_T}{2}, \frac{m_T}{4}\right) \quad (2.26b)$$

$$= \mu_3\left(\frac{m_T}{2}\right) + \frac{3}{8} \text{Var}(m_T) \quad (2.26c)$$

$$= \frac{1}{8} (\mu_3(m_T) + 3 \text{Var}(m_T)) \quad (2.26d)$$

Similarly, for the 3rd central moment of the gene promoter:

$$\mu_3(g_0) = \mathbb{E}(\mu_3(g_0 \mid g_T)) + \mu_3(\mathbb{E}(g_0 \mid g_T)) + 3 \text{Cov}(\mathbb{E}(g_0 \mid g_T), \text{Var}(g_0 \mid g_T)) \quad (2.27a)$$

$$= 0 + \mu_3(g_T) + 3 \text{Cov}(g_T, 0) \quad (2.27b)$$

$$= \mu_3(g_T) \quad (2.27c)$$

Finally, we derive the partitioning for the mixed 3rd central moments of the gene promoter and the mRNA. We get that:

$$\mathbb{E}\left((g_0 - \mathbb{E}(g_0))(m_0 - \mathbb{E}(m_0))^2\right) = \frac{1}{4} \left(\mathbb{E}\left((g_T - \mathbb{E}(g_T))(m_T - \mathbb{E}(m_T))^2\right) + \text{Cov}(g_T, m_T)\right) \quad (2.28)$$

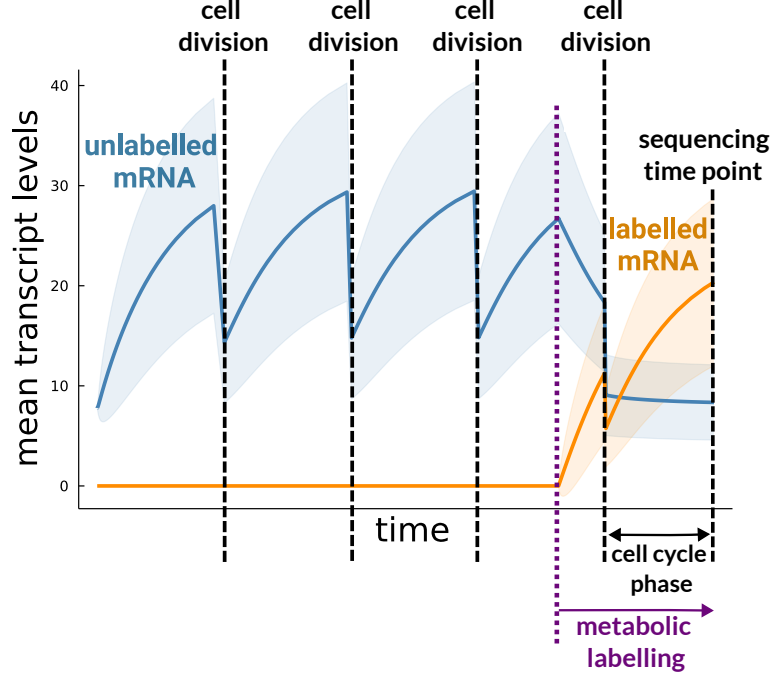


Figure 2.5: **Example realisation of the stochastic model.** The mean (blue and orange lines) and noise (shaded blue and orange regions) of unlabelled and labelled mRNA levels are tracked over time along a single-cell lineage, following consecutive cell divisions, until the sequencing time point.

and

$$\mathbb{E} \left((g_0 - \mathbb{E}(g_0))^2 (m_0 - \mathbb{E}(m_0)) \right) = \frac{1}{2} \mathbb{E} \left((g_T - \mathbb{E}(g_T))^2 (m_T - \mathbb{E}(m_T)) \right) \quad (2.29)$$

where we are using the following formulas:

$$\mathbb{E}(g_0 m_0^2) = \mathbb{E}(\mathbb{E}(g_0 | g_T) \mathbb{E}(m_0^2 | m_T)) = \frac{1}{4} \mathbb{E} \left(g_T (m_T + m_T^2) \right) \quad (2.30a)$$

$$\mathbb{E}(g_0^2 m_0) = \mathbb{E}(\mathbb{E}(g_0^2 | g_T) \mathbb{E}(m_0 | m_T)) = \frac{1}{2} \mathbb{E} \left(g_T^2 m_T \right) \quad (2.30b)$$

$$\mathbb{E}(g_0 m_0) = \mathbb{E}(\mathbb{E}(g_0 | g_T) \mathbb{E}(m_0 | m_T)) = \frac{1}{2} \mathbb{E} \left(g_T m_T \right) \quad (2.30c)$$

Description of a single realisation of the model

A single model realisation depends on a gene-specific parameter set $\{k_{\text{off}}, k_{\text{on}}(t), \alpha(t), \gamma(t), \lambda\}$, a cell age τ and a labelling condition $(t_{\text{pulse}}, t_{\text{chase}})$. We solve the moment equations numerically over time for multiple consecutive cell divisions, following a single-cell lineage up until the sequencing time point t_{seq} (Figure 2.5). We additionally assume that cells have a deterministic

cell cycle length, $T_d = 20$ hours. First, the equations are solved for multiple cell cycles until convergence to a steady state. The convergence criterion is that the relative change of all trajectories at the end of two consecutive cell cycles needs to be smaller than 0.01%. After reaching convergence, we solve the model for additional time equal to $3T_d + \tau$, up until the final time point t_{seq} (where sequencing occurs). To model the effect of pulse labelling on mRNA synthesis, the labelling efficiency $\lambda(t)$ is set to its constant parameter value λ for time t_{pulse} :

$$\lambda(t) = \begin{cases} \lambda & \text{if } t_{seq} - t_{pulse} - t_{chase} \leq t < t_{seq} - t_{chase} \\ 0 & \text{otherwise} \end{cases} \quad (2.31)$$

The endpoints $\mathbb{E}(u), \mathbb{E}(l), \text{Var}(u), \text{Var}(l), \text{Cov}(u, l)$ of the trajectories at time t_{seq} comprise the model estimates, which are conditional on cell age τ and pulse-chase labelling time (t_{pulse}, t_{chase}) .

2.2.3 Capture efficiency estimation and downsampling of moments

To take into account technical effects of sequencing in our modelling framework, we correct the model outputs using estimates of cell-specific capture efficiencies. Consider the count matrices $\mathbf{U} = (u_{ij})$ and $\mathbf{L} = (l_{ij})$ of unlabelled and labelled mRNA, as well as the total mRNA count matrix $\mathbf{X} = (x_{ij}) = \mathbf{U} + \mathbf{L}$, where the indices $i = 1, \dots, n$ represent cells and the indices $j = 1, \dots, d$ represent genes. If $u = u_{ij}$ and $l = l_{ij}$ are a pair of observations for a cell i and gene j , then due to low transcript capture of the sequencing protocol, these observations are random samples drawn from Binomial distributions, $u \sim \text{Binomial}(\tilde{u}, \beta_i)$ and $l \sim \text{Binomial}(\tilde{l}, \beta_i)$, where $\tilde{u}, \tilde{l}$ are the mRNA counts we would observe in a sequencing experiment with 100% capture efficiency and β_i is the capture efficiency of cell i . Below, we outline the method for estimating the capture efficiency distribution directly from the data and provide formulas for the corrected (downsampled) moments of u and l .

Estimation of cell-specific capture efficiencies from scEU-seq data

The estimation of β_i is based on the assumption that cells in the same cell cycle phase have on average the same cell size. We use the total mRNA abundance $s_i = \sum_{j=1}^d x_{ij}$ of cell i of as a

proxy for its size, while the average total mRNA count $s_\theta = \frac{1}{N_\theta} \sum_{\substack{i \in \theta \\ j=1, \dots, d}} x_{ij}$ of cells in cell cycle phase θ represents the characteristic cell size in θ (N_θ is the number of cells in θ). Based on the above, we can estimate:

$$\hat{\beta}_i = \bar{\beta} \frac{s_i}{s_\theta} \quad (2.32)$$

where $\bar{\beta}$ is an estimate of the average capture efficiency of the sequencing experiment.

An additional factor that we take into account in the above estimation is that pulse-treated and chase-treated cells can overall exhibit different total mRNA count profiles (as we will show in the next section). For that reason, we consider different average capture efficiencies $\bar{\beta}_{\text{pulse}}, \bar{\beta}_{\text{chase}}$ for pulse-treated and chase-treated cells, and the normalising factors s_θ for each cell cycle phase are evaluated independently for pulse and chase cells. The estimated capture efficiency $\hat{\beta}_i$ of cell i is then given by:

$$\hat{\beta}_i = \bar{\beta}_c \frac{s_i}{s_{c,\theta}} \quad (2.33)$$

where $\bar{\beta}_c = \begin{cases} 0.1 & \text{if } c = c_{\text{pulse}} \\ 0.2 & \text{if } c = c_{\text{chase}} \end{cases}$ is the average capture efficiency of pulse/chase experiments and $s_{c,\theta} = \frac{1}{N_{c,\theta}} \sum_{\substack{i \in c \cap \theta \\ j=1, \dots, d}} x_{ij}$ is the average total count of cells in condition $c \in \{c_{\text{pulse}}, c_{\text{chase}}\}$ cell cycle phase θ . The ratio of the values we have assumed for $\bar{\beta}_c$ (capture efficiency of chase being twice the capture efficiency of pulse) is inferred from the data, while the absolute scale of 10% capture efficiency is an estimate that is reasonable for scRNA-seq protocols [145], but it is not typically directly measured.

Downsampling the mean, variance and covariance

We can now use the estimated capture efficiencies to incorporate the effect of Binomial downsampling of data in the moments. For a cell with capture efficiency β , the downsampled mean mRNA level is given by:

$$\mathbb{E}(x) = \mathbb{E}(\beta) \mathbb{E}(\tilde{x}) \quad (2.34)$$

where $\tilde{x}$ is the mRNA levels in the absence of technical noise. The downsampled variance is given by:

$$\text{Var}(x) = \mathbb{E}(\beta(1 - \beta)) \mathbb{E}(\tilde{x}) + \text{Var}(\beta) \mathbb{E}(\tilde{x}^2) + \mathbb{E}(\beta)^2 \text{Var}(\tilde{x}) \quad (2.35)$$

The formulas (2.34),(2.35) hold for both unlabelled and labelled mRNA.

Finally, the downsampled covariance of unlabelled and labelled mRNA is given by:

$$\text{Cov}(u, l) = \text{Var}(\beta) \mathbb{E}(\tilde{u} \tilde{l}) + \mathbb{E}(\beta)^2 \text{Cov}(\tilde{u}, \tilde{l}) \quad (2.36)$$

To prove the above, we use that an observed mRNA count $x_{ij} = x$ is a random variable $x \sim \text{Binomial}(\tilde{x}, \beta)$ where $\beta = \beta_i$ is the capture efficiency of cell i and $\tilde{x}$ is the observation in the absence of technical errors. The downsampled mean is given by:

$$\mathbb{E}(x) = \mathbb{E}(\mathbb{E}(x \mid \tilde{x}, \beta)) = \mathbb{E}(\beta \tilde{x}) = \mathbb{E}(\beta) \mathbb{E}(\tilde{x}) \quad (2.37)$$

and the downsampled variance is given by:

$$\text{Var}(x) = \mathbb{E}(\text{Var}(x \mid \tilde{x}, \beta)) + \text{Var}(\mathbb{E}(x \mid \tilde{x}, \beta)) \quad (2.38a)$$

$$= \mathbb{E}(\beta(1 - \beta) \tilde{x}) + \mathbb{E}(\beta^2 \tilde{x}^2) - \mathbb{E}(\beta \tilde{x})^2 \quad (2.38b)$$

$$= \mathbb{E}(\beta(1 - \beta)) \mathbb{E}(\tilde{x}) + \mathbb{E}(\beta^2) \mathbb{E}(\tilde{x}^2) - \mathbb{E}(\beta)^2 \mathbb{E}(\tilde{x})^2 \quad (2.38c)$$

$$= \mathbb{E}(\beta(1 - \beta)) \mathbb{E}(\tilde{x}) + \mathbb{E}(\beta^2) \mathbb{E}(\tilde{x}^2) - \mathbb{E}(\beta)^2 \mathbb{E}(\tilde{x})^2 - \mathbb{E}(\beta)^2 \mathbb{E}(\tilde{x}^2) + \mathbb{E}(\beta)^2 \mathbb{E}(\tilde{x}^2) \quad (2.38d)$$

$$= \mathbb{E}(\beta(1 - \beta)) \mathbb{E}(\tilde{x}) + \text{Var}(\beta) \mathbb{E}(\tilde{x}^2) + \mathbb{E}(\beta)^2 \text{Var}(\tilde{x}) \quad (2.38e)$$

Finally, the downsampled covariance of unlabelled and labelled mRNA is given by:

$$\text{Cov}(u, l) = \mathbb{E}(\text{Cov}(u, l \mid \tilde{u}, \tilde{l}, \beta)) + \text{Cov}(\mathbb{E}(u \mid \tilde{u}, \beta), \mathbb{E}(l \mid \tilde{l}, \beta)) \quad (2.39a)$$

$$= 0 + \text{Cov}(\beta \tilde{u}, \beta \tilde{l}) \quad (2.39b)$$

$$= \text{Var}(\beta) \mathbb{E}(\tilde{u} \tilde{l}) + \mathbb{E}(\beta)^2 \text{Cov}(\tilde{u}, \tilde{l}) \quad (2.39c)$$

Downsampling the 3rd central moment

As we will show in the next section, we are also considering the 3rd order moments to evaluate the predictive capacity of our model. Following similar principles as above, we derive the binomial downsampling of the 3rd order central moment. We will only consider the 3rd moment of the total mRNA count $x = u + l$ and not the mixed 3rd order moments. Using the law of total cumulance, we get:

$$\mu_3(x) = \mathbb{E}(\mu_3(x \mid \beta, \tilde{x})) + \mu_3(\mathbb{E}(x \mid \beta, \tilde{x})) + 3 \text{Cov}(\mathbb{E}(x \mid \beta, \tilde{x}), \text{Var}(x \mid \beta, \tilde{x})) \quad (2.40a)$$

$$= \underbrace{\mathbb{E}(\beta(1 - \beta)(1 - 2\beta)\tilde{x})}_{\text{gene expression}} + \underbrace{\mu_3(\beta\tilde{x})}_{\text{downsampling}} + \underbrace{3 \text{Cov}(\beta\tilde{x}, \beta(1 - \beta)\tilde{x})}_{\text{mixed effects}} \quad (2.40b)$$

where

$$\mu_3(\beta\tilde{x}) = \mathbb{E}(\beta^3)\mathbb{E}(\tilde{x}^3) - 3 \mathbb{E}(\beta)\mathbb{E}(\beta^2)\mathbb{E}(\tilde{x})\mathbb{E}(\tilde{x}^2) + 2 \mathbb{E}(\beta)^3\mathbb{E}(\tilde{x})^3 \quad (2.41)$$

and

$$\text{Cov}(\beta\tilde{x}, \beta(1 - \beta)\tilde{x}) = (\text{Var}(\beta) - \mathbb{E}(\beta^3) - \mathbb{E}(\beta)\mathbb{E}(\beta^2)) \mathbb{E}(\tilde{x})^2 + (\mathbb{E}(\beta^2) - \mathbb{E}(\beta^3)) \text{Var}(\tilde{x}) \quad (2.42)$$

2.2.4 Time-dependent summary statistics of gene expression

We have demonstrated how our model generates moments, dependent on cell age τ and labelling time t (paragraph 2.2.2), as well as how we correct them to account for technical noise (paragraph 2.2.3). Below, we describe how we further use these moments to construct a collection of summary statistics and extract interpretable information about bursty transcription dynamics.

Cell cycle-dependent mean and noise statistics

First, we construct summary statistics that describe cell cycle-dependence of mean and noise in transcript levels. Given the means of unlabelled and labelled mRNA $\mathbb{E}(u \mid \tau, t), \mathbb{E}(l \mid \tau, t)$, we compute the mean of total transcript levels $x = u + l$:

$$\mathbb{E}(x \mid \tau) = \mathbb{E}(u \mid \tau, t) + \mathbb{E}(l \mid \tau, t) \quad (2.43)$$

Note that this mean does not depend on labelling condition t . Similarly, we compute the variance of total transcripts 2nd order moments of unlabelled and labelled mRNA:

$$\text{Var}(x \mid \tau) = \text{Var}(u \mid \tau, t) + \text{Var}(l \mid \tau, t) + 2 \text{Cov}(u, l \mid \tau, t) \quad (2.44)$$

as well as the Fano factor:

$$\text{FF}(x \mid \tau) = \frac{\text{Var}(x \mid \tau)}{\mathbb{E}(x \mid \tau)} \quad (2.45)$$

These summary statistics describe temporal changes of transcription dynamics along the cell cycle that are independent of variations due to labelling.

Labelled vs total mRNA ratios capture the timescales of RNA kinetics

Next, we generate summary statistics that aim to extract information about the timescales of bursty transcription and degradation from the various pulse and chase labelling conditions. We first construct a statistic that describes how the proportion of labelled mRNA changes with labelling time:

$$\text{Ratio}(t) = \frac{\mathbb{E}(l \mid t)}{\mathbb{E}(u + l \mid t)} \quad (2.46)$$

To integrate out the cell cycle dependence, we use empirical age distributions from the data that depend on the labelling condition (paragraph 2.2.1):

$$\mathbb{E}(u \mid t) = \mathbb{E}_\tau(\mathbb{E}(u \mid \tau, t)) \quad , \quad \mathbb{E}(l \mid t) = \mathbb{E}_\tau(\mathbb{E}(l \mid \tau, t)) \quad (2.47)$$

Labelled-unlabelled mRNA correlations capture sources of transcriptional noise

Further, we construct summary statistics that describe the covariation of labelled and unlabelled mRNA. The law of total covariance gives:

$$\underbrace{\text{Cov}(u, l)}_{\text{total noise}} = \underbrace{\mathbb{E}_\tau(\text{Cov}(u, l \mid \tau))}_{\text{upstream bursty promoter noise}} + \underbrace{\text{Cov}_\tau(\mathbb{E}(u \mid \tau), \mathbb{E}(l \mid \tau))}_{\text{cell cycle variation}} \quad (2.48)$$

which shows that the unlabelled-labelled RNA modalities and the cell cycle reporter provided by the scEU-seq protocol can be used to decompose the overall noise into two contributions:

upstream noise induced by the bursty gene promoter and noise due to dependence on the cell cycle. To exploit this information, our summary statistics include two correlation coefficients - corresponding to the two noise components - that contribute to the total correlation of unlabelled-labelled mRNA:

$$\underbrace{\rho_{u,l}(t)}_{\text{total correlation}} = \underbrace{\frac{\mathbb{E}_\tau(\text{Cov}(u, l) \mid \tau, t)}{\sqrt{\text{Var}(u \mid t) \text{Var}(l \mid t)}}}_{\text{correlation coefficient 1}} + \underbrace{\frac{\text{Cov}_\tau(\mathbb{E}(u \mid \tau, t), \mathbb{E}(l \mid \tau, t))}{\sqrt{\text{Var}(u \mid t) \text{Var}(l \mid t)}}}_{\text{correlation coefficient 2}} \quad (2.49)$$

Similarly to the ratios, we compute labelling-specific variances and covariances by integrating over the labelling-specific age distributions. Note that the total variances in the denominators of (2.49) are given by the law of total variance:

$$\text{Var}(x \mid t) = \mathbb{E}_\tau(\text{Var}(x \mid \tau, t)) + \text{Var}_\tau(\mathbb{E}(x \mid \tau, t)) \quad (2.50)$$

2.2.5 Model-based noise decomposition

Now, we describe how we can use the moments generated from the model to perform transcriptional noise decomposition. Let x be the mRNA count of a gene and u, l the corresponding unlabelled and labelled modalities, so that $x = u + l$. Also, let τ denote the cell age and t the labelling time (based on experimental condition). Then, we compute the total mean transcript levels as:

$$\mathbb{E}(x) = \mathbb{E}_{\tau,t}(\mathbb{E}(x \mid \tau, t)) \quad (2.51)$$

The total variance of x can be decomposed as follows:

$$\text{Var}(x) = \underbrace{\mathbb{E}_{\tau,t}(\text{Var}(x \mid \tau, t))}_{\text{intrinsic - reaction kinetics noise}} + \underbrace{\text{Var}_{\tau,t}(\mathbb{E}(x \mid \tau, t))}_{\text{extrinsic - temporal dependence noise}} \quad (2.52)$$

The intrinsic noise component can be further decomposed as follows:

$$\underbrace{\mathbb{E}_{\tau,t}(\text{Var}(x \mid \tau, t))}_{\text{intrinsic noise}} = \underbrace{\mathbb{E}_{\tau,t}(\text{Var}(u \mid \tau, t)) + \mathbb{E}_{\tau,t}(\text{Var}(l \mid \tau, t))}_{\text{transcription and decay noise}} + \underbrace{2 \mathbb{E}_{\tau,t}(\text{Cov}(u, l \mid \tau, t))}_{\text{bursty promoter noise}} \quad (2.53)$$

We also note here that the transcription and decay term also includes noise due to random partitioning at cell division.

The extrinsic noise component is explained by variations due to cell cycle-dependence and labelling time:

$$\underbrace{\text{Var}_{\tau,t}(\mathbb{E}(x \mid \tau, t))}_{\text{extrinsic noise}} = \text{Var}_{\tau,t}(\mathbb{E}(\mathbb{E}(x \mid \tau, t) \mid \tau)) + \mathbb{E}_{\tau,t}(\text{Var}(\mathbb{E}(x \mid \tau, t) \mid \tau)) \quad (2.54a)$$

$$= \underbrace{\text{Var}_{\tau}(\mathbb{E}(x \mid \tau))}_{\text{cell cycle variation}} + \underbrace{\mathbb{E}_{\tau,t}(\text{Var}(\mathbb{E}(x \mid \tau, t) \mid \tau))}_{\text{labelling time-dependence noise}} \quad (2.54b)$$

Computing the labelling time-dependence component from the model yields a negligible, practically zero contribution and therefore we can neglect it from the formulas.

We finally obtain the total variance decomposition:

$$\text{Var}(x) = \underbrace{\mathbb{E}_{\tau,t}(\text{Var}(u \mid \tau, t) + \text{Var}(l \mid \tau, t))}_{\text{transcription and decay noise}} + \underbrace{2 \mathbb{E}_{\tau,t}(\text{Cov}(u, l \mid \tau, t))}_{\text{bursty promoter noise}} + \underbrace{\text{Var}_{\tau}(\mathbb{E}(x \mid \tau))}_{\text{cell cycle variation}} \quad (2.55)$$

Dividing the whole expression in (2.55) by the squared total mean $\mathbb{E}(x)^2$, we can obtain the formula for decomposing the total noise $\text{CV}_{\text{total}}^2(x)$.

Noise decomposition in the presence of technical noise

We now want to include the effect of technical noise due to cell-specific downsampling in the above decomposition (2.55). For each individual term in (2.55), we obtain a correction that includes the effects of technical noise. To correct the transcription and decay noise term, we use the variance downsampling formula (2.35) to get (for both $x \in \{u, l\}$):

$$\begin{aligned} \text{Var}(x \mid \tau, t) &= \mathbb{E}(\beta(1 - \beta) \mid \tau, t) \mathbb{E}(\tilde{x} \mid \tau, t) + \mathbb{E}(\beta^2 \mid \tau, t) \text{Var}(\tilde{x} \mid \tau, t) \\ &\quad + \text{Var}(\beta \mid \tau, t) \mathbb{E}(\tilde{x} \mid \tau, t)^2 \end{aligned} \quad (2.56)$$

and hence

$$\begin{aligned} \mathbb{E}_{\tau,t}(\text{Var}(x \mid \tau, t)) &= \mathbb{E}_{\tau,t}(\mathbb{E}(\beta(1 - \beta) \mid \tau, t)) \mathbb{E}_{\tau,t}(\mathbb{E}(\tilde{x} \mid \tau, t)) + \mathbb{E}_{\tau,t}(\mathbb{E}(\beta^2 \mid \tau, t)) \mathbb{E}_{\tau,t}(\text{Var}(\tilde{x} \mid \tau, t)) \\ &\quad + \mathbb{E}_{\tau,t}(\text{Var}(\beta \mid \tau, t)) \mathbb{E}_{\tau,t}(\mathbb{E}(\tilde{x} \mid \tau, t)^2) \end{aligned} \quad (2.57)$$

Similarly, to correct the bursty promoter noise term, we use the covariance downsampling formula (2.36) to get:

$$\text{Cov}(u, l \mid \tau, t) = \mathbb{E}(\beta^2 \mid \tau, t) \text{Cov}(\tilde{u}, \tilde{l} \mid \tau, t) + \text{Var}(\beta \mid \tau, t) \mathbb{E}(\tilde{u} \mid \tau, t) \mathbb{E}(\tilde{l} \mid \tau, t) \quad (2.58)$$

and hence

$$\begin{aligned} \mathbb{E}_{\tau, t}(\text{Cov}(u, l \mid \tau, t)) &= \mathbb{E}_{\tau, t}(\mathbb{E}(\beta^2 \mid \tau, t)) \mathbb{E}_{\tau, t}(\text{Cov}(\tilde{u}, \tilde{l} \mid \tau, t)) \\ &\quad + \mathbb{E}_{\tau, t}(\text{Var}(\beta \mid \tau, t)) \mathbb{E}_{\tau, t}(\mathbb{E}(\tilde{u} \mid \tau, t) \mathbb{E}(\tilde{l} \mid \tau, t)) \end{aligned} \quad (2.59)$$

Finally, in the case of the cell cycle variation we use the downsampling of the mean (2.34):

$$\begin{aligned} \text{Var}_{\tau}(\mathbb{E}(x \mid \tau)) &= \text{Var}_{\tau}(\mathbb{E}(\beta \mid \tau)) \text{Var}_{\tau}(\mathbb{E}(\tilde{x} \mid \tau)) + \mathbb{E}_{\tau}(\mathbb{E}(\beta \mid \tau))^2 \text{Var}_{\tau}(\mathbb{E}(\tilde{x} \mid \tau)) \\ &\quad + \text{Var}_{\tau}(\mathbb{E}(\beta \mid \tau)) \mathbb{E}_{\tau}(\mathbb{E}(\tilde{x} \mid \tau))^2 \end{aligned} \quad (2.60)$$

Putting it all together, we get the total variance decomposition formula in the presence of technical noise:

$$\begin{aligned} \text{Var}(x) &= \underbrace{\mathbb{E}(\beta(1 - \beta)) \mathbb{E}(\tilde{x}) + \mathbb{E}(\beta^2) \mathbb{E}(\text{Var}(\tilde{u} \mid \tau, t) + \text{Var}(\tilde{l} \mid \tau, t))}_{\text{transcription and decay noise}} \\ &\quad + \underbrace{\mathbb{E}(\text{Var}(\beta \mid \tau, t)) \mathbb{E}(\mathbb{E}(\tilde{u} \mid \tau, t)^2 + \mathbb{E}(\tilde{l} \mid \tau, t)^2)}_{\text{transcription and decay noise}} \\ &+ \underbrace{2 \mathbb{E}(\beta^2) \mathbb{E}(\text{Cov}(\tilde{u}, \tilde{l} \mid \tau, t)) + 2 \mathbb{E}(\text{Var}(\beta \mid \tau, t)) \mathbb{E}(\mathbb{E}(\tilde{u} \mid \tau, t) \mathbb{E}(\tilde{l} \mid \tau, t))}_{\text{bursty promoter noise}} \\ &+ \underbrace{\mathbb{E}(\mathbb{E}(\beta \mid \tau)^2) \text{Var}(\mathbb{E}(\tilde{x} \mid \tau)) + \text{Var}(\mathbb{E}(\beta \mid \tau)) \mathbb{E}(\tilde{x})^2}_{\text{cell cycle variation}} \end{aligned} \quad (2.61)$$

where we have dropped the indices τ, t from the total expectations and covariances to simplify notation.

2.3 Bayesian inference methodology

In the previous section, we described a data-driven stochastic modelling framework for analysing cell cycle-dependent transcription dynamics. In this section, we present a Bayesian inference framework, that we use to weigh our model against data. This will allow us to extract mechanistic insights into transcription regulation from the scEU-seq dataset. As we explained in the previous section, we consider different modelling hypotheses that relate the model parameters with different modes of cell size- and cell cycle-dependent regulation. Therefore, our inference framework includes both a parameter estimation step and a model selection scheme for investigating the full posterior distribution over models and parameters. The Julia code [200] produced to develop our Bayesian inference framework is available in the Github repository https://github.com/pthomaslab/abc_inference_transcription.

2.3.1 Approximate Bayesian Computation framework

As our model relies on moments and does not provide with a full probability distribution of transcript levels, we cannot utilise an analytical likelihood to perform exact Bayesian inference. Instead, we employ approximate Bayesian computation (ABC), a popular method for likelihood-free parameter inference in dynamical systems [107, 201]. In particular, we use ABC rejection sampling [106] that can be described in short as follows: To estimate the true parameters θ_0 , a parameter set θ is sampled from a prior distribution $\pi(\theta)$, followed by a numerical simulation of a model $f(\tilde{y} \mid \theta)$ to generate synthetic data $\tilde{y} = \tilde{y}(\theta)$. The data $\tilde{y}$ is then compared against the real data y to accept or reject θ based on whether their distance satisfies an acceptance threshold, *i.e.* $d(y, \tilde{y}(\theta)) < \varepsilon$, for some given metric d . Repeating this process a large number of times allows to approximate the true posterior distribution $\pi(\theta_0 \mid y)$ by the distribution of accepted particles $\pi(\theta \mid d(y, \tilde{y}(\theta)) < \varepsilon)$. In addition, a common approach is to compare summary statistics of the data $s(y)$ with generated summary statistics $s(\theta)$ from the model, to obtain the approximate posterior distribution. In the following paragraph, we outline the details of the ABC scheme that we employ for inference.

ABC rejection sampling for parameter inference

For each gene j in the data, we consider its associated collection of summary statistics $s_{\text{obs}}^{(j)}$ (as described in 2.2.1). The scheme for estimating gene-specific parameters is the following:

For each model index $m = 1, \dots, 5$ and for $i = 1, \dots, M$,

- Sample a parameter set $\theta^{(i)} = \theta^{(i)}(m)$ from the prior $\pi(\theta)$.
- Generate synthetic summary statistics $s^{(i)} = s^{(i)}(\theta^{(i)})$ using the model.
- Compute distances $d(s^{(i)}, s_{\text{obs}}^{(j)})$ for all $j = 1, \dots, N$.
- Accept parameter set $\theta^{(i)}$ for all $j = 1, \dots, N$ that satisfy $d(s^{(i)}, s_{\text{obs}}^{(j)}) \leq \varepsilon$, where ε is a predetermined acceptance threshold.

For each $j = 1, \dots, N$, the collection of accepted parameter sets $\{\Theta^{(j)}(m) : m = 1, \dots, 5\}$ comprises an approximate (joint) posterior distribution over parameters and models for gene j . After model selection (as we explain in more detail below), we choose the best model $m^* = m^*(j)$ for each gene j and we use the particle $\theta^{(i)}(m^*(j))$ that achieved the minimum distance $d(s^{(i)}, s_{\text{obs}}^{(j)})^*$ between model and data as the gene-specific point estimate to inform the model for downstream analyses. Figure 2.6 shows an intuitive illustration of the above algorithm.

To determine the prior distribution $\pi(\theta)$, we sample parameters from the following log-uniform distributions: $\log_{10}(k_{\text{on}}) \sim \mathcal{U}(-3, 3)$, $\log_{10}(k_{\text{off}}) \sim \mathcal{U}(-3, 3)$, $\log_{10}(\alpha) \sim \mathcal{U}(-3, 3)$, $\log_{10}(\gamma) \sim \mathcal{U}(-3, 2)$ and $\log_{10}(\lambda) \sim \mathcal{U}(-0.7, 0)$. We sample $M = 10^6$ parameter sets for each model m so we run in total $5 \cdot 10^6$ simulations. This is the computationally expensive step of our method, but is independent of the specific data for individual genes hence making our method scalable. For each sampled parameter set $\theta^{(i)} = \theta^{(i)}(m)$, we generate summary statistics $s^{(i)}(\theta^{(i)})$ from the model m (as described in 2.2.4) which we then compare to all $N = 3419$ gene-specific sets of summary statistics. The distance between a synthetic and an observed collection of summary statistics, $s = \{s_1, \dots, s_n\}$ and $s_{\text{obs}} = \{s_{\text{obs},1}, \dots, s_{\text{obs},n}\}$, is determined by the following objective function:

$$d(s, s_{\text{obs}}) = \frac{1}{n} \sum_{k=1}^n \frac{(s_k - s_{\text{obs},k})^2}{\hat{\text{se}}(s_{\text{obs}})^2 + \sigma^2 s_{\text{obs}}^2} \quad (2.62)$$

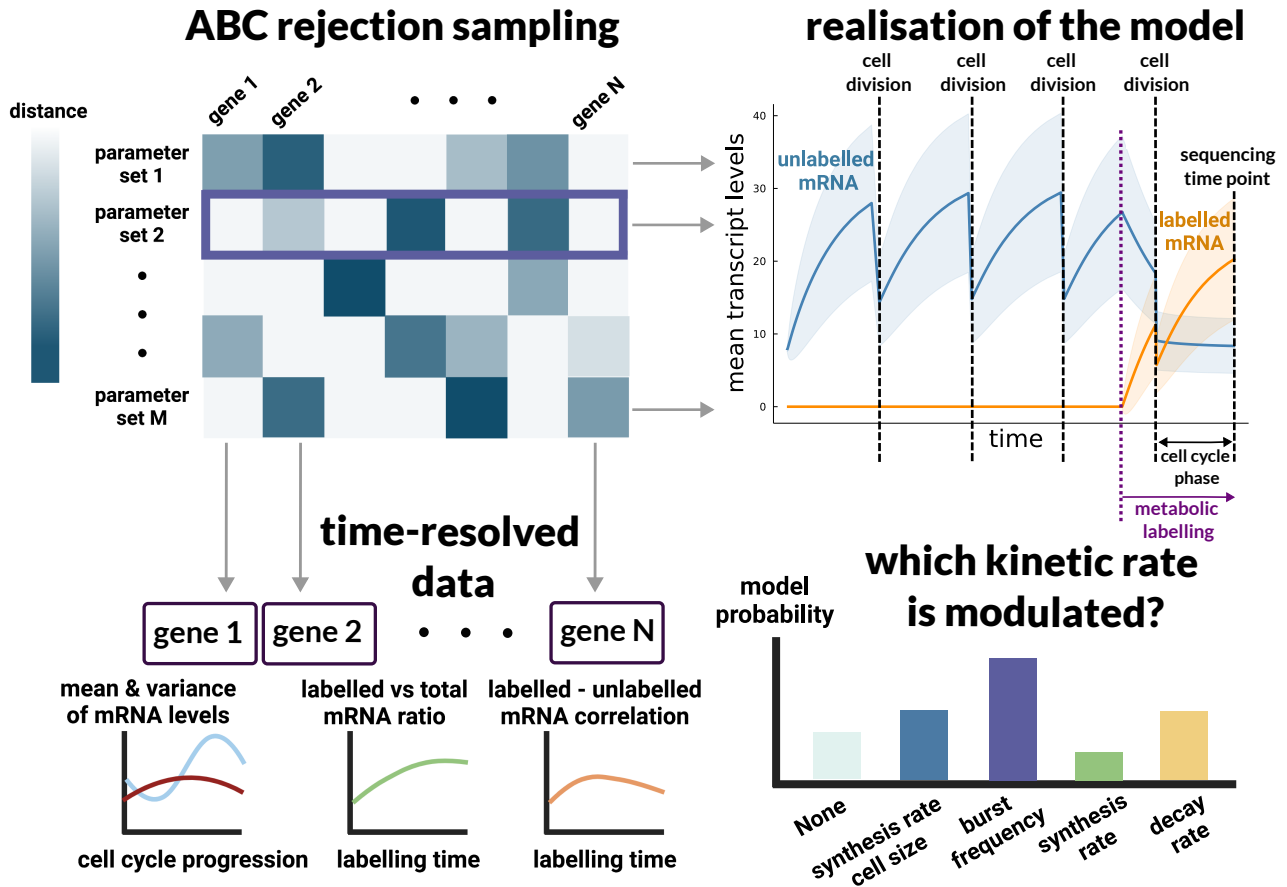


Figure 2.6: **Approximate Bayesian inference framework.** Overview of the ABC-based inference scheme: At each step, a sampled prior parameter set yields a model realisation. The parameter set is accepted/rejected by comparing the model-generated summary statistics with observed summary statistics, for all genes in the data. The relative frequency of accepted parameter sets per model provides an estimate of the posterior model probability.

where $n = 53$ is the total number of summary statistics. Our choice for normalisation of the squared distances is based on the assumption that the effect of measurement noise on a summary statistic y is of the form

$$y_{\text{noisy}} = y + \xi_1 + y \xi_2 \quad , \quad \xi_1 \sim \mathcal{N}(0, \hat{\text{se}}(y)^2) \quad , \quad \xi_2 \sim \mathcal{N}(0, \sigma^2) \quad (2.63)$$

where $\hat{\text{se}}(s_{\text{obs}})$ is the standard error of the summary statistic y due to sampling noise (see 2.2.1) and $\sigma = 0.1$ models other technical variations [202].

ABC-based model selection

As we are testing multiple modelling hypotheses against all genes in the data, we wish to assign probabilities to these hypotheses and decide which model is most likely to describe each gene. To do that, we use an ABC-based model selection approach [203], which involves two decision steps. The first one is to select whether a gene is constant or non-constant (Figure 2.4). Given this decision, the second step is to select a specific constant model (scaling, non-scaling) or a specific non-constant model (varying burst frequency, varying burst size or varying decay rate). Let H_1, H_2 denote the constant and non-constant model hypotheses and let $\mathcal{D}$ denote the data. As we have in total two constant models and three non-constant models, we set the prior probabilities of the two hypotheses to be $\frac{\pi(H_1)}{\pi(H_2)} = \frac{2}{3}$. The posterior probability of model H_i is approximated by:

$$p(H_i | \mathcal{D}) = \frac{N_{i,\varepsilon}}{N_\varepsilon} \quad (2.64)$$

where $N_{i,\varepsilon}$ is the number of particles accepted for model H_i and N_ε is the total number of particles accepted for all models, given the acceptance threshold ε . In this case, $N_{1,\varepsilon}$ includes the number of particles accepted for both constant models, while $N_{2,\varepsilon}$ includes the number of particles accepted for all three non-constant models. Model comparison is then based on the (approximate) Bayes factor

$$B_{1,2} = \frac{p(H_1 | \mathcal{D}) \pi(H_2)}{p(H_2 | \mathcal{D}) \pi(H_1)} \quad (2.65)$$

We then need a criterion for thresholding this Bayes factor to decide whether we have sufficient evidence in favour of H_1 over H_2 , to a certain level of confidence. To do that we use bootstrapping of the accepted particles for each model, and estimate confidence intervals $(L_{H_i|\mathcal{D}}, U_{H_i|\mathcal{D}})$

for each posterior probability, so that

$$\mathbb{P}\left(L_{H_i|\mathcal{D}} \leq p(H_i | \mathcal{D}) \leq U_{H_i|\mathcal{D}}\right) = 1 - \alpha \quad (2.66)$$

with $\alpha = 0.05$. For model H_1 to be selected versus model H_2 it is required that

$$p(H_1 | \mathcal{D}) > \frac{\pi(H_1)}{\pi(H_2)} p(H_2 | \mathcal{D}) \quad (2.67)$$

or equivalently, since $p(H_2 | \mathcal{D}) = 1 - p(H_1 | \mathcal{D})$, we need

$$p(H_1 | \mathcal{D}) > \frac{\frac{\pi(H_1)}{\pi(H_2)}}{1 + \frac{\pi(H_1)}{\pi(H_2)}} \quad (2.68)$$

To strengthen this criterion, we require

$$L_{H_1|\mathcal{D}} > \frac{\frac{\pi(H_1)}{\pi(H_2)}}{1 + \frac{\pi(H_1)}{\pi(H_2)}} \quad (2.69)$$

Equivalently, model M_2 is selected if

$$U_{H_1|\mathcal{D}} \leq \frac{\frac{\pi(H_1)}{\pi(H_2)}}{1 + \frac{\pi(H_1)}{\pi(H_2)}} \quad (2.70)$$

After selecting between the constant and non-constant hypotheses, we proceed to the second step of model selection, where we follow a similar approach. If the constant hypothesis is selected, we now decide between the scaling H_1 and non-scaling H_2 hypotheses, with probabilities given by (2.64), where now $N_{1,\varepsilon}, N_{2,\varepsilon}$ are the number of accepted particles of the scaling and the non-scaling model respectively and $N_\varepsilon = N_{1,\varepsilon} + N_{2,\varepsilon}$. If the non-constant hypothesis is selected, we decide between the varying burst frequency H_1 , varying burst size H_2 and varying decay rate H_3 hypotheses. In that case, the choice criterion is formed as follows. The model probabilities $p(H_i | \mathcal{D})$ are given by (2.64), where now $N_{i,\varepsilon}$ is the number of accepted particles of each non-constant model and $N_\varepsilon = N_{1,\varepsilon} + N_{2,\varepsilon} + N_{3,\varepsilon}$. If $(L_{H_i|\mathcal{D}}, U_{H_i|\mathcal{D}})$ are the corresponding confidence intervals from bootstrapping, we select the non-constant model $i \in \{1, 2, 3\}$ if

$$L_{H_i|\mathcal{D}} \geq U_{H_j|\mathcal{D}} \quad , \quad \text{for all } j \neq i \quad (2.71)$$

2.3.2 Testing the model performance on synthetic data

In addition to applying inference on the scEU-seq dataset, we generated synthetic data to test the robustness of our inference and model selection schemes. Here we describe the procedure for synthetic data generation, while the results are presented in the next section. Using the Gillespie algorithm [64], we performed stochastic simulations of the gene expression model (2.5) with reaction propensities (2.13). For each single cell and each gene, we generated unlabelled and labelled transcript counts by sampling from the respective stochastic trajectories of transcript levels at steady state. Using ground truth parameters sampled from the posterior distributions we inferred from the real data, we simulated 15 genes from each of the 5 model classes (75 genes in total). For each gene, we simulated a total of 2463 cells, across the 11 pulse-chase labelling conditions and the 5 cell cycle stages. The distribution of cells across labelling conditions and cell cycle stages corresponded to the one observed in the real scEU-seq data. Additionally, we added technical noise to the simulated count data using Binomial downsampling and cell-specific capture efficiencies that were sampled from Beta distributions. These Beta distributions were chosen by fitting the empirical capture efficiency distributions of pulse and chase cells that we previously estimated from the real data (paragraph 2.2.3).

We perform inference on the simulated data, followed by model selection, using the framework presented in 2.3.1. We also performed inference on 3 additional scenarios, as described below, and we refer to the initial scenario as the *full model* for comparison purposes:

- Small number of cells: We used half of the samples in the synthetic data ($N = 1232$ cells) to compute summary statistics.
- Small number of prior samples: We used $2 * 10^5$ prior parameter sets from each model to fit the data (1/5 of the parameters used for the full model).
- Narrow priors: The prior parameters used for inference were sampled from the following log-uniform distributions: $\log_{10}(k_{\text{on}}) \sim \mathcal{U}(-2.5, 2.5)$, $\log_{10}(k_{\text{off}}) \sim \mathcal{U}(-2.5, 2.5)$, $\log_{10}(\alpha) \sim \mathcal{U}(-2.5, 2.5)$, $\log_{10}(\gamma) \sim \mathcal{U}(-2.75, 1.75)$ and $\log_{10}(\lambda) \sim \mathcal{U}(-0.6, -0.1)$.

The results of these robustness tests will be shown in the next section.

2.4 Applications: Revealing global properties of transcriptional regulation

2.4.1 Stochastic modelling captures distinct patterns of bursty transcription dynamics along the cell cycle

We have developed a stochastic model to describe the dependence of bursty transcription dynamics on the cell cycle as well as the effects of metabolic RNA labelling, according to the scEU-seq protocol. We described in detail how our model enables testing multiple hypotheses on the cell cycle-dependence of kinetic rate parameters and identify genes characterised by distinct underlying kinetics. This allows us to utilise an approximate Bayesian computation framework to infer the mode of regulation that best describes each gene in the data, as we will demonstrate in the next paragraph (Figure 2.6).

We find that our five models yield diverse summary statistics and can identify distinct dynamic features of transcription. Example model realisations (Figure 2.7) suggest that noise statistics have a crucial role in model identifiability. For instance, even though the varying burst frequency and burst size models can give rise to overall similar summary statistics, the cell cycle-dependent Fano factor can be used to distinguish between the two models. We also find that the varying decay rate model yields very different patterns compared to the other two non-constant models, suggesting that our model can distinguish regulatory mechanisms of transcription from mechanisms regulating mRNA degradation. The constant scaling and non-scaling models can be distinguished by the cell cycle-dependent mean and noise statistics. Swapping linear for exponential growth did not affect our results as exponential and linear growth are nearly indistinguishable over one cell size doubling (Figure 2.2). Additionally, we observe that correlation coefficients have an essential role in distinguishing constant versus cell cycle-modulated kinetics. In particular, the correlation of cell cycle-dependent means, that measures noise due to cell cycle variation, is higher in non-constant compared to constant models. On the other hand, the mean (cell cycle-averaged) correlation, that measures upstream noise due to the bursty gene promoter, is higher in constant compared to non-constant models.

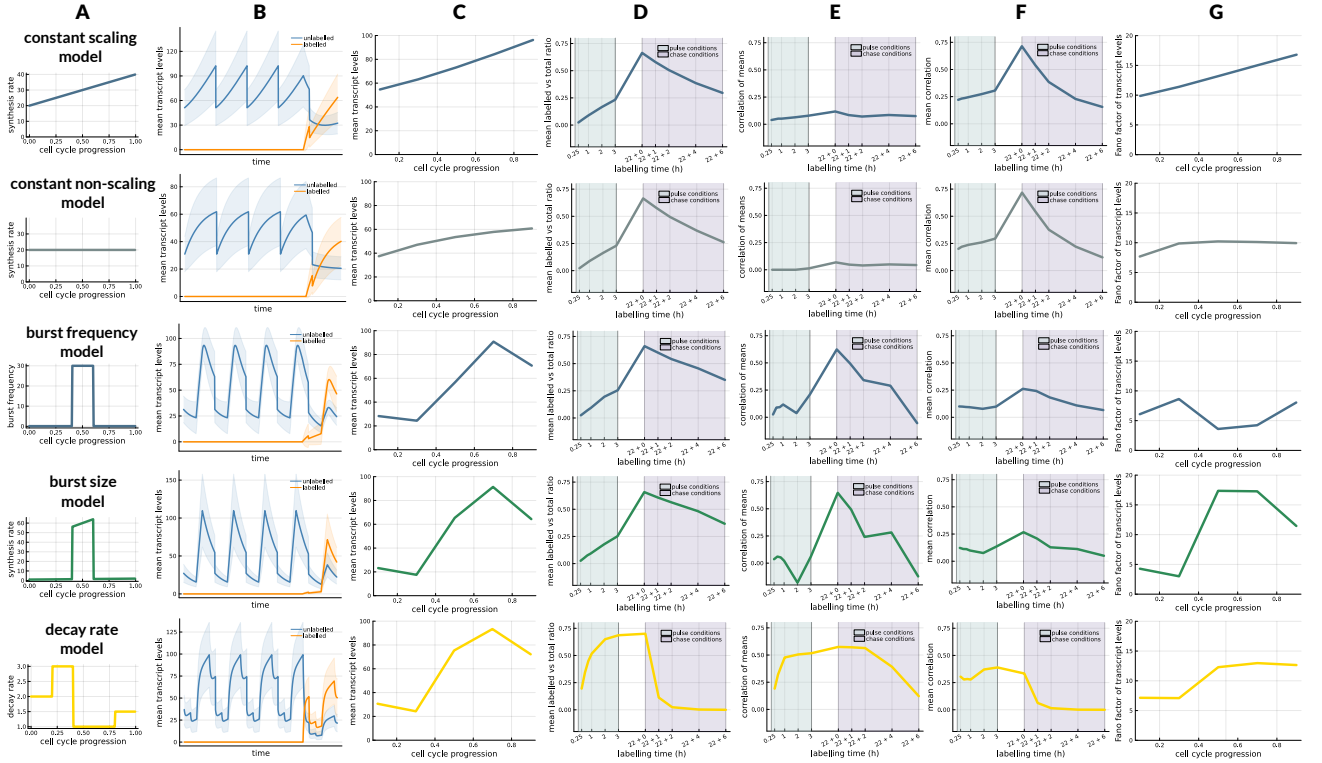


Figure 2.7: Comparison of summary statistics arising from the different models. (A) Labels of the different models and associated regulation of kinetics (see also Figure 2.4). (B) Model realisations for each parameter setup: Mean unlabelled and labelled transcript levels across a single-cell lineage over time, where the shaded region corresponds to the standard deviation of transcript levels. The 1st realisation (row 1) assumes that all kinetic rates are constant along the cell cycle, while the synthesis rate scales linearly with cell size. The 2nd realisation (row 2) assumes identical parameters with the 1st realisation, with the only difference being that the synthesis rate does not depend on cell size. The next 3 realisations (rows 3-5) assume time-varying burst frequency, synthesis rate or decay rate and scaling of synthesis rate with cell size, while the rest of the kinetic rates are constant. The chosen parameter sets in rows 3-5 are such that we obtain almost identical patterns of cell cycle-dependent mean expression (column (C)). (C-G) Summary statistics generated by each model using the cell cycle-dependent and labelling time-dependent moments of unlabelled and labelled mRNA (see paragraph 2.2.4 and Tables 2.1 and 2.2). Columns (C-E) show statistics that can be computed from mean measurements and are defined in rows (1),(3) and (5) of Table 2.1. Columns (F-G) show statistics that can be computed from noise measurements and are defined in rows (2) and (4) of Table 2.1.

2.4.2 Approximate Bayesian Computation reveals global patterns of transcriptional regulation

Motivated by the ability of the model to produce distinct summary statistics between different underlying modes of kinetics regulation, we develop an inference framework to infer these modes of regulation along with gene-specific kinetic parameters from scEU-seq data. A main challenge here is that inference requires optimising over a large parameter space across all five modelling hypotheses for each gene in the data separately. Therefore, using a maximum likelihood, optimisation-based or Markov Chain Monte Carlo approach would be computationally inefficient. For that reason, we develop a simulation-based inference framework based on ABC rejection sampling, which allows to test multiple modelling hypotheses and millions of parameter sets against the entire set of genes in a highly parallel manner (Figure 2.6). Given a parameter set sampled from a prior distribution and a realisation of the model, we generate a collection of summary statistics and simultaneously compare them to summary statistics of all genes in the data (see 2.3.1). Additionally, using an approximate Bayesian framework allows to perform model selection across several models with different levels of complexity (number of parameters) [203], *i.e.* to assign probabilities to the various hypotheses, and decide which kinetic model is most likely to describe each gene. A crucial challenge for inference is that scEU-seq data are characterised by low mRNA capture efficiency [155], meaning that summary statistics of the data and the model are not directly comparable. We address this issue by modelling cell-specific capture efficiencies and downsampling the moments to enable reliable inference [94, 145, 204], as we describe in detail in the following paragraphs.

In Figure 2.8 we show example posterior fits over the different summary statistics for a constant and a non-constant model. We observe that the mean expression of the gene *RNH1* shows a monotonous, approximately linear increase, while its Fano Factor stays relatively constant over the cell cycle (Figures 2.8 (i-ii)). Even though the constant model shows a linear increase in both mean expression and Fano factor, the posterior model fit agrees well with the data. The ratio of mean labelled vs total expression of *RNH1* increases with pulse labelling time and decreases along chase conditions as the chase duration increases (Figure 2.8 (iii)). The model behaves similarly and we see that the data lies well within the posterior model fit. The mean (cell cycle-averaged) correlation of labelled-unlabelled expression follows a similar pattern with the ratios and even though the data exhibits a more noisy behaviour, its error bars lie within

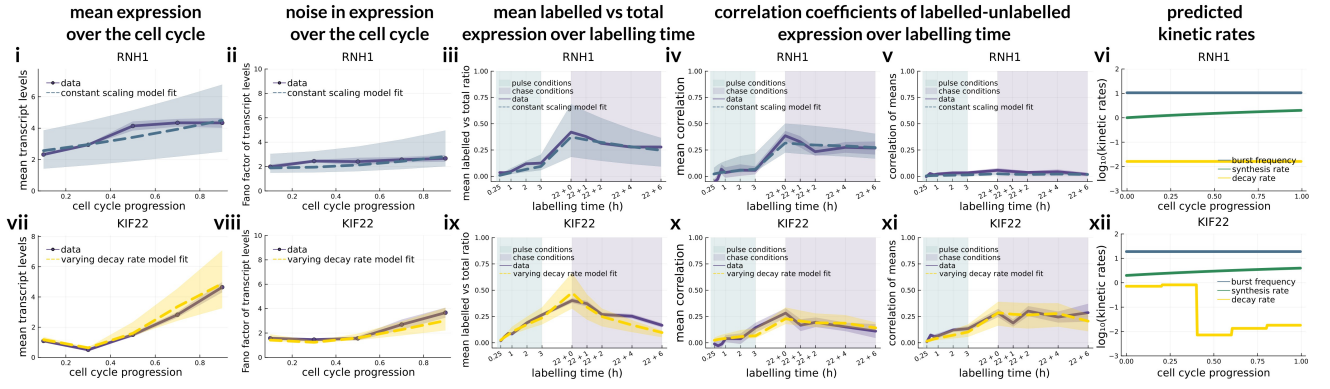


Figure 2.8: **Example posterior model fits across summary statistics.** We show fits for all summary statistics used for comparing models and data for an example constant gene (top row) and an example non-constant gene (bottom row), as well as point estimates of each gene’s kinetic rate parameters.

the posterior model fit (Figure 2.8 (iv)). The correlation of cell cycle-dependent means is almost zero across labelling conditions for both model and data (Figure 2.8 (v)).

The gene *KIF22* shows a large variation in mean and Fano factor of expression over the cell cycle (Figures 2.8 (vii-viii)) as we see a strong increase towards the late cell cycle stages. The non-constant decay rate model explains these patterns well as the error bars of the data lie within the posterior fit. The ratio of mean labelled vs total expression and the mean correlation of labelled-unlabelled expression of *KIF22* show similar behaviour with the respective statistics of *RNH1*, and even though they appear more noisy than the model, we obtain a good overall posterior fit (Figures 2.8 (ix-x)). The correlation of means is significantly higher compared to the *RNH1* case, and even though it is also very noisy, it is captured well by the non-constant posterior fit (Figure 2.8 (xi)). These model fits show that *RNH1* is explained by constant kinetics along the cell cycle and its synthesis rate scales linearly with cell size (Figure 2.8 (vi)). On the other hand, the gene *KIF22* is explained by non-constant kinetics as its decay rate varies with cell cycle progression (Figure 2.8 (xii)). Overall, our inference framework provides us with approximate posterior distributions of gene-specific kinetic parameters (Figure 2.9). From this example we see that synthesis and decay rate are better identified compared to burst size and burst frequency. Interestingly, we observe that burst size and frequency are negatively correlated as well as that decay rate estimates are independent from the rest of the parameters.

After performing parameter inference over all 3419 genes that we selected for the analysis (paragraph 2.2.1), we find that the majority (87%) are explained by constant models, and about 7% of genes are assigned to non-constant models (Figure 2.10 (A i)). Within the subset

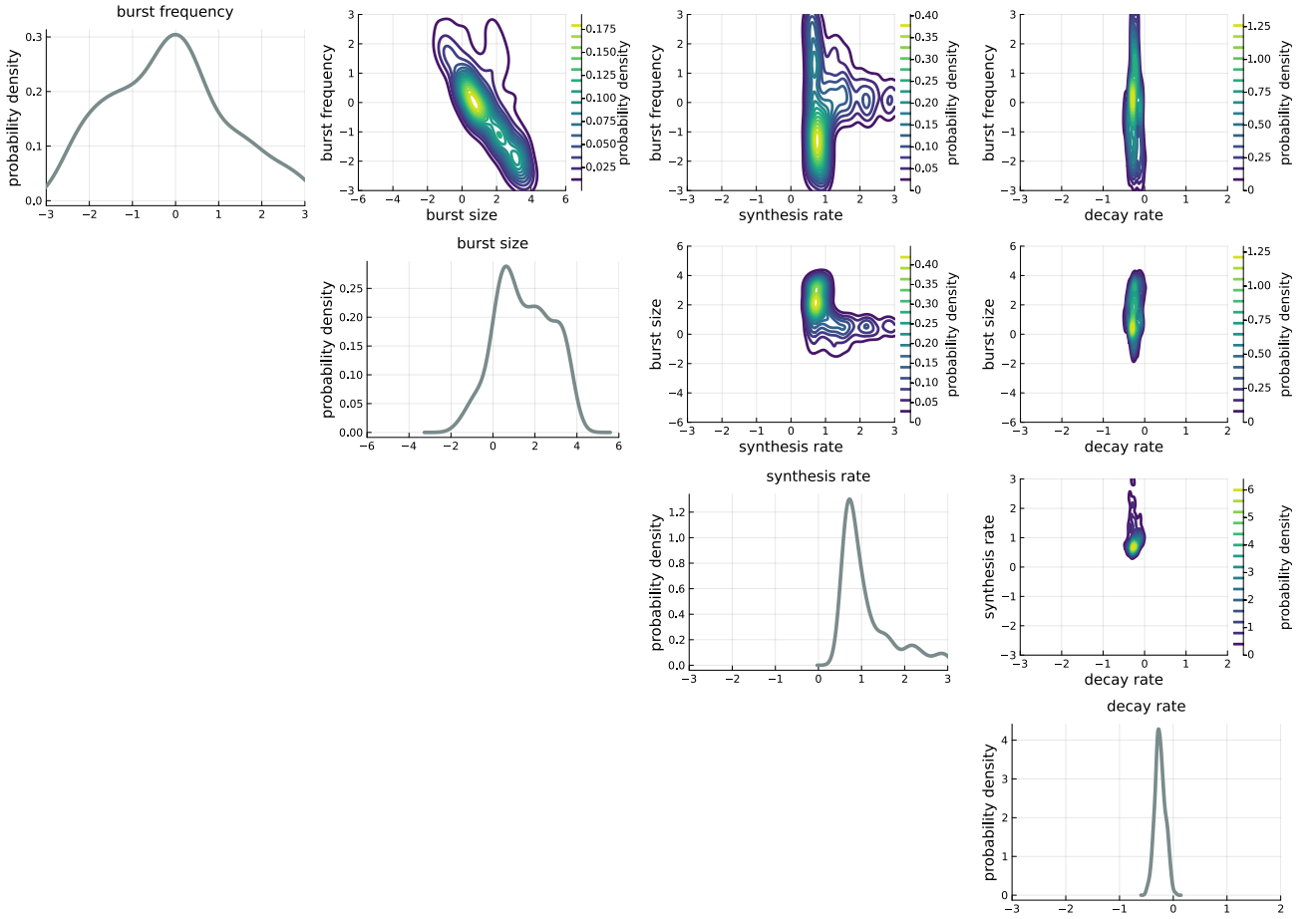


Figure 2.9: **Approximate posterior distributions of parameters for an example gene *SGK1*.** The diagonal plots show kernel density estimates of the posterior distributions of burst frequency k_{on} , burst size α/k_{off} , synthesis rate α and decay rate γ in $\log_{10}$ scale. The off-diagonal plots show kernel density estimates of the joint posterior distributions of pairs of parameters. All parameters are measured in h^{-1} , apart from burst size which is unitless.

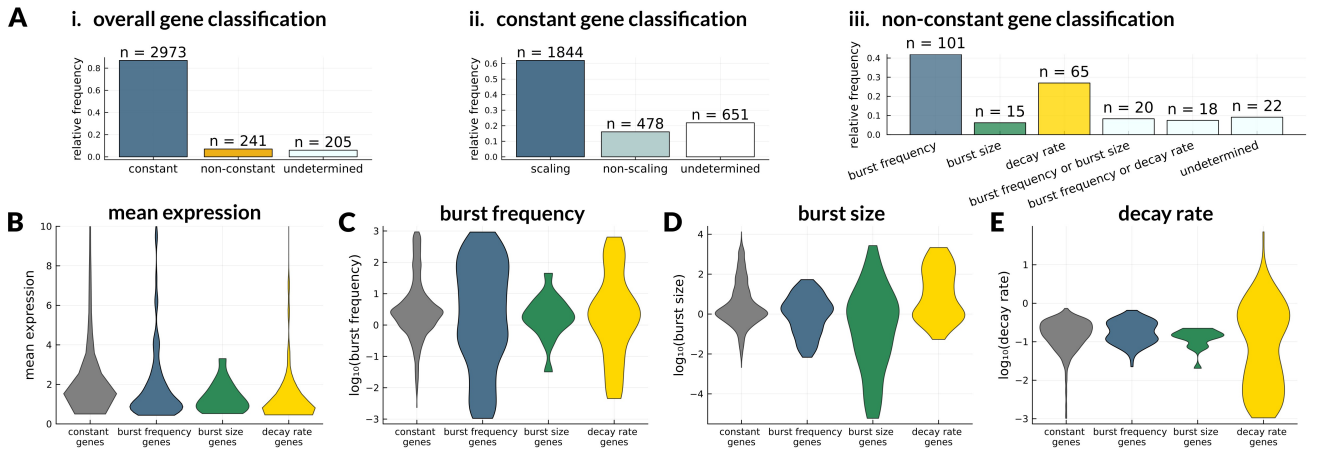


Figure 2.10: Gene classification with respect to models of kinetics regulation (A) Overall statistics of gene classification obtained from the ABC model selection. (B-D) Comparison of mean expression and kinetic rate magnitudes across different classes of genes. For each model class (constant and non-constant), proportions of specific sub-classes are shown. For constant, varying burst frequency, varying burst size and varying decay rate genes, we show distributions of (B) mean expression, (C) burst frequency, (D) burst size and (E) decay rate.

of genes that are assigned to constant models (*constant* genes), we find that the majority (62%) are assigned to the size-scaling synthesis rate model (*scaling* genes) and a smaller fraction of genes (16%) are assigned to the non-scaling synthesis rate model (*non-scaling* genes), while for the rest of constant genes we get an uncertain model choice outcome (Figure 2.10 (A ii)). Among the subset of genes assigned to non-constant models (*non-constant* genes), we find two major groups assigned to the burst frequency model (41%) and the decay rate model (27%), respectively, while a small fraction of genes are assigned to the burst size model (Figure 2.10 (A iii)). For the rest of non-constant genes, we find two noteworthy groups of genes, one that is described well by both the burst frequency and the burst size model and one that is described well by both the burst frequency and the decay rate model. For the remaining genes (6%) we either get an uncertain model selection outcome or all the models performed poorly.

Furthermore, our inference reveals general patterns of mRNA kinetics across the different groups of genes. We observe that genes in all groups span a wide range of mean expression levels, while we observe that constant genes have a slightly higher median expression level compared to non-constant genes (Figure 2.10 (B)). We find that burst frequency is highly variable across all constant and non-constant genes, while burst frequency genes exhibit the strongest variation (Figure 2.10 (C)). Additionally, cell cycle-regulated burst size shows the largest variation, reaching very low levels compared to genes with constant burst size (Figure 2.10 (D)). We similarly

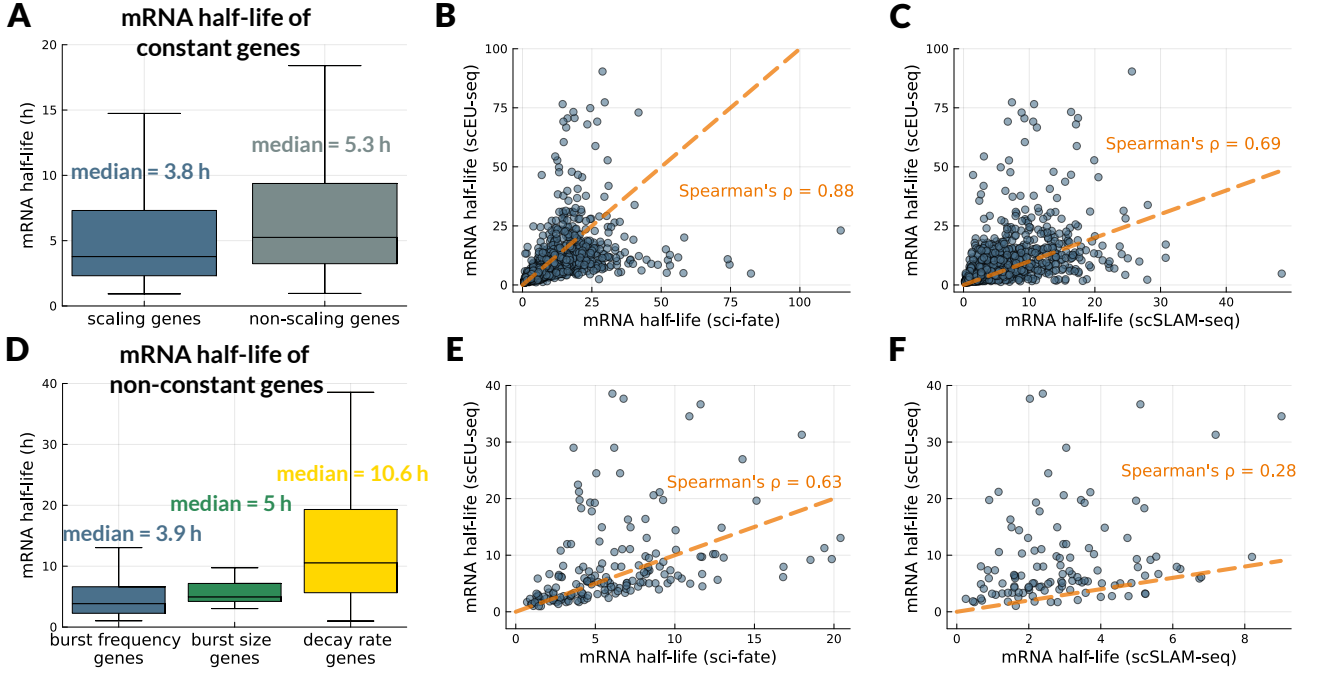


Figure 2.11: mRNA half-life distributions and comparison with previously reported estimates. (A) Estimated mRNA half-lives of constant scaling and non-scaling genes (in hours), as defined by (2.10) using the point estimates of decay rate. (B-C) Comparison of half-lives of constant genes with $n = 2209$ common genes in A549 cells as measured in the sci-fate study [190] and with $n = 1565$ common genes in NIH-3T3 cells as measured in the scSLAM-seq study [187]. These comparisons exclude half-lives that are larger than 100 hours (less than 2% of constant genes). The orange dashed lines are identity lines. (D) Estimated mRNA half-lives of non-constant genes. To compute the half-lives of varying decay rate genes, we use the average decay rate over the cell cycle. (E-F) Comparison of half-lives of non-constant genes with $n = 177$ common genes in A549 cells as measured in the sci-fate study and with $n = 119$ common genes in NIH-3T3 cells as measured in the scSLAM-seq study. These comparisons exclude half-lives that are larger than 40 hours (1% of non-constant genes). The orange dashed lines are identity lines.

see large variation in decay rate among decay rate genes, while for the rest of the genes the decay rate is constrained in similar levels (Figure 2.10 (E)).

We also find that cell cycle-dependent regulation of decay rate impacts mRNA half-life. The majority of genes have median half-lives of approximately 4-5 hours, while the median mRNA half-life of genes with cell cycle-dependent decay rate is longer than 10 hours (Figure 2.11 (A) and (D)). The predicted half-lives are thus shorter than the cell cycle duration (see 2.2.2). After comparing our estimates of mRNA half-life with two other metabolic labelling studies (sci-fate [190] and scSLAM-seq [187]), we find that the estimated half-lives of constant genes agree well with the sci-fate predictions (Spearman's $\rho \approx 0.9$, Figure 2.11 (B)), however a lower correlation is observed in the case of non-constant genes (Figure 2.11 (E)). We find good agreement between

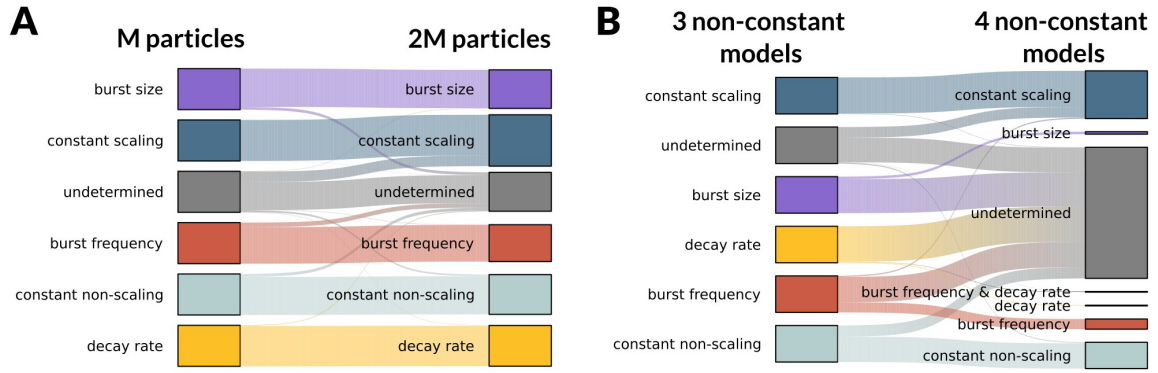


Figure 2.12: **Impact of additional particles and more complex models on model selection.** (A) 2-fold increase in the number of particles - simulations ($M = 5 * 10^6$) does not influence model selection and gene classification. (B) Including an additional model of regulation (varying burst frequency & decay rate) increases uncertainty in selecting between non-constant models.

our constant gene half-lives and the scSLAM-seq predictions (Spearman's $\rho \approx 0.7$, Figure 2.11 (C)), while there is a poor agreement in the case of non-constant genes (Figure 2.11 (F)). These discrepancies in the case of non-constant genes indicate that modelling cell cycle-dependence improves the estimation of mRNA half-life. The overall discrepancies could in addition be attributed to differences between the scEU-seq, sci-fate and scSLAM-seq metabolic labelling protocols [129] or to the fact that these estimations are based on different cell lines.

2.4.3 The inference method robustly identifies modes of regulation

Next, we validated the correctness and robustness of our inference and model selection framework. We found that doubling the number of simulations does not change our overall conclusions on model classification (Figure 2.12 (A)). We also found that when including more complex non-constant models (where more than one parameter varies with the cell cycle), our inference still robustly selects constant versus non-constant models, but there is uncertainty in distinguishing between non-constant models (Figure 2.12 (B)). This suggests that richer data are necessary to identify more complex mechanisms. Moreover, we show that our model accurately quantifies 3rd order central moments of gene expression, reflecting the predictive capacity of our model (Figure 2.13 (A-B)). We also show that these predictions capture both biological and technical noise contributions to the 3rd moments (Figure 2.13 (C-D)). These findings are indicative of sufficiency of our summary statistics, suggesting that adding higher-order moments in our inference would not improve our inference accuracy further. Additionally, we tested the statistical

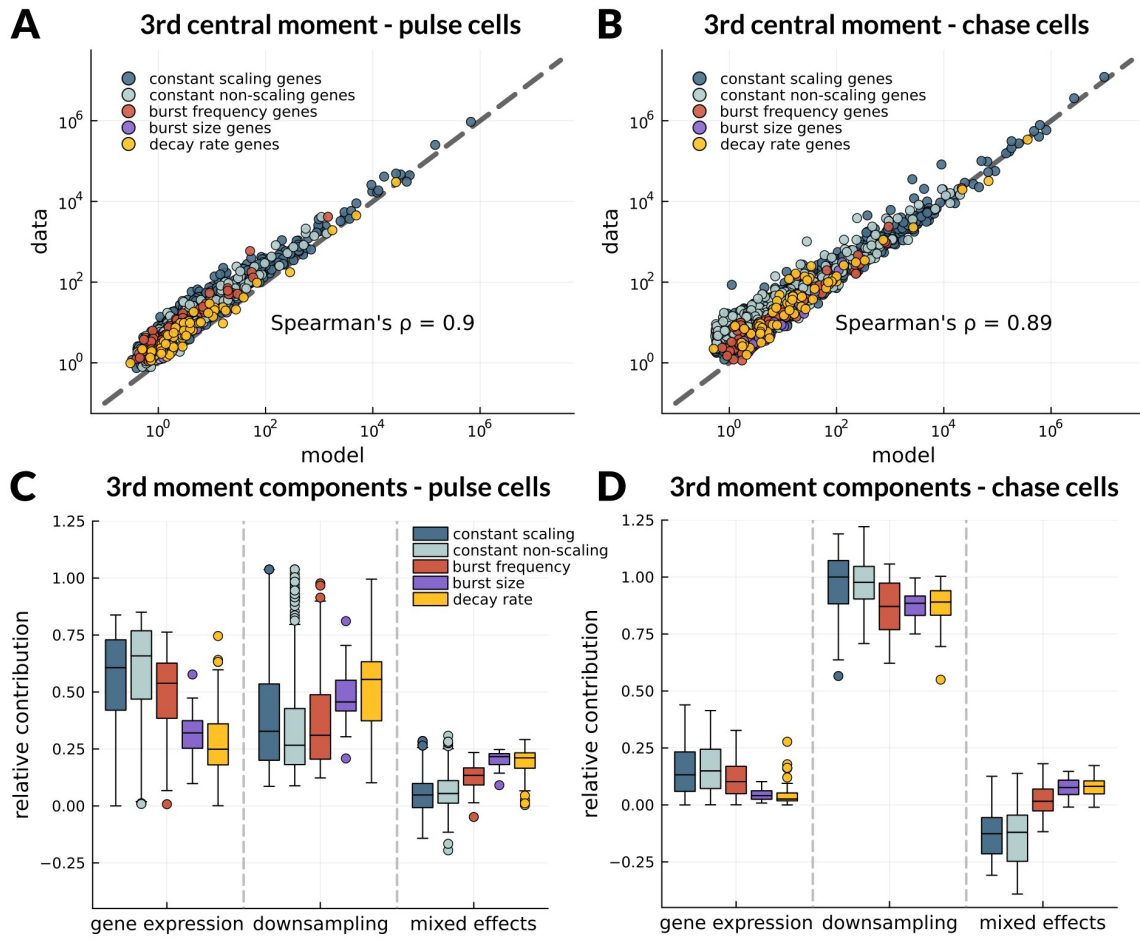


Figure 2.13: Predictive capacity of the model and sufficiency of summary statistics. (A-B) Comparison of 3rd order central moments of expression across pulse-treated and chase-treated cells as observed in the data and as predicted by the model using gene-specific MAP estimates. (C-D) Relative contributions of biological and technical components to the 3rd central moments for pulse-treated and chase-treated cells, computed using equation (2.40). The "mixed-effects" component of the 3rd moment is a covariance that can take negative values.

independence of our summary statistics and overall found zero or negligible pairwise correlations (Figure 2.14). We only found a few exceptions of small positive pairwise correlations, that might be due to sample size bias.

We next performed inference on synthetic mRNA count data and found that the model accurately recovers models and their parameters (Figure 2.15). The data was generated using Gillespie stochastic simulations, as we describe in 2.3.2. After comparing model estimates against ground truth parameter values across the 5 gene groups, we found Spearman correlations of 0.7 in the cases of burst frequency and size (Figure 2.15 (A i-ii)) and correlations higher than 0.9 in the cases of synthesis and decay rate (Figure 2.15 (A iii-iv)). Additionally, we tested the robustness of our inference method in model selection, in 4 different scenarios (as

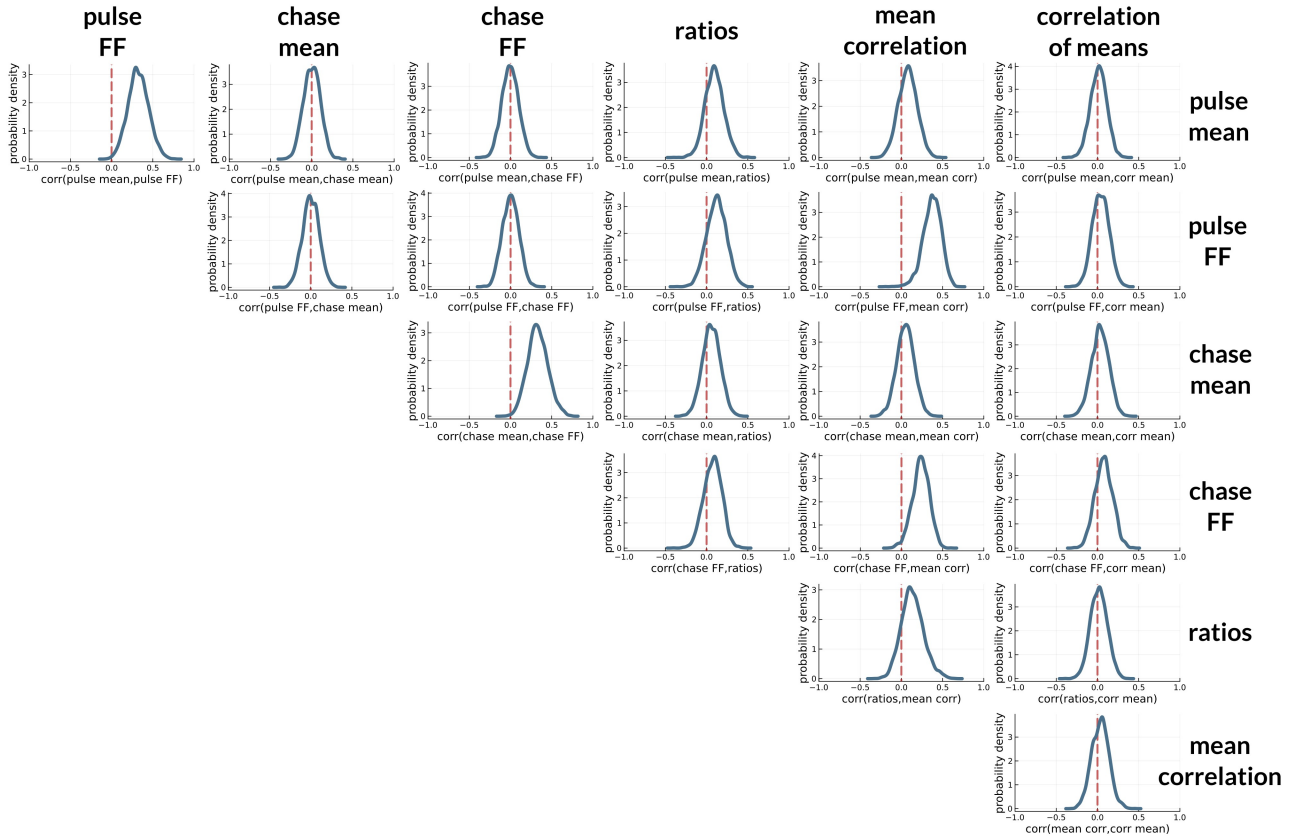


Figure 2.14: **Independence of summary statistics.** For each gene, we bootstrapped the data 100 times, we computed the summary statistics across all bootstraps and computed correlations between all different pairs of summary statistics. We then plot the distributions of these pairwise correlations across genes.

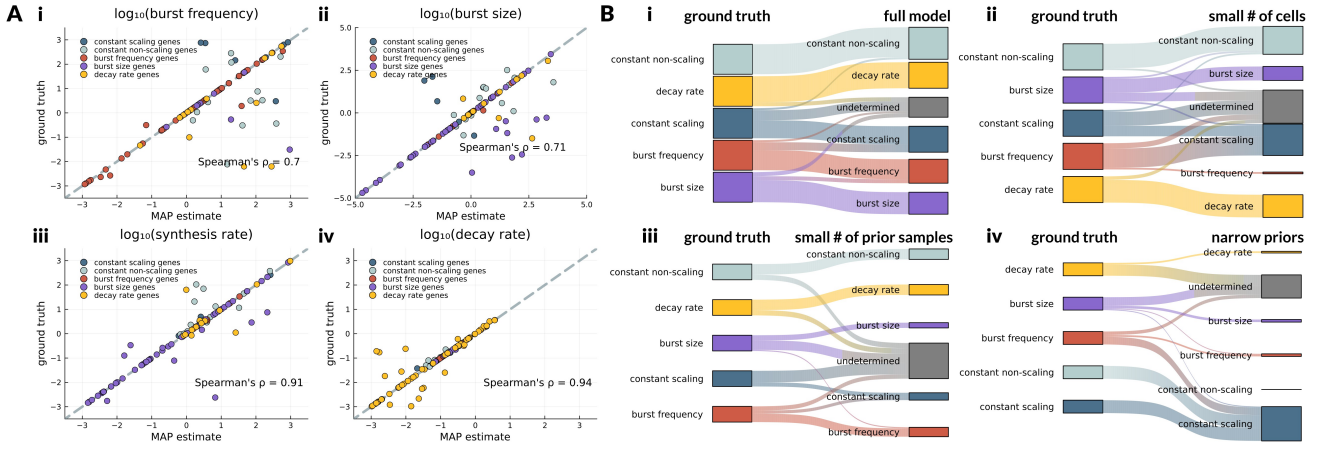


Figure 2.15: **Inference on synthetic data and model robustness.** (A) Comparison of MAP estimates vs ground truth values of (i) burst frequency, (ii) burst size, (iii) synthesis rate and (iv) decay rate across all 5 model classes. We plot all parameters corresponding to genes that were classified correctly to the ground truth model class. (B) Our method's robustness in model selection was tested in 4 different scenarios: (i) Default modelling set up (full model), (ii) small number of samples (cells), (iii) limited prior sampling, and (iv) narrow priors.

explained in 2.3.2). In the full modelling set up, the model recovers the ground truth classification with high accuracy (Figure 2.15 (B i)). In the limits of small sample size (small number of cells) and limited parameter sampling (small number of prior samples), the majority of ground truth model classes are recovered successfully, while for a fraction of genes there is significant uncertainty in model selection (Figure 2.15 (B ii-iii)). The burst frequency genes tend to be classified as constant scaling, which is reasonable as the two models are a special case of each other and can yield similar dynamics. In the case of narrow priors we observe higher uncertainty in model selection, and many genes are classified as constant scaling, indicating that simpler models are favoured. This result suggests that using a sufficiently wide prior support is essential for accurate model selection.

2.4.4 Modelling capture and labelling efficiency allows to uncouple technical and biological noise

The technical biases associated with sequencing protocols make the task of modelling scRNA-seq data challenging. Building a stochastic model that accurately explains the observed variability in the data, requires that the various biological and technical contributions to the total noise are carefully accounted for. We here demonstrate that the relationship between unlabelled and labelled mRNA levels of a single gene can reveal how different noise components

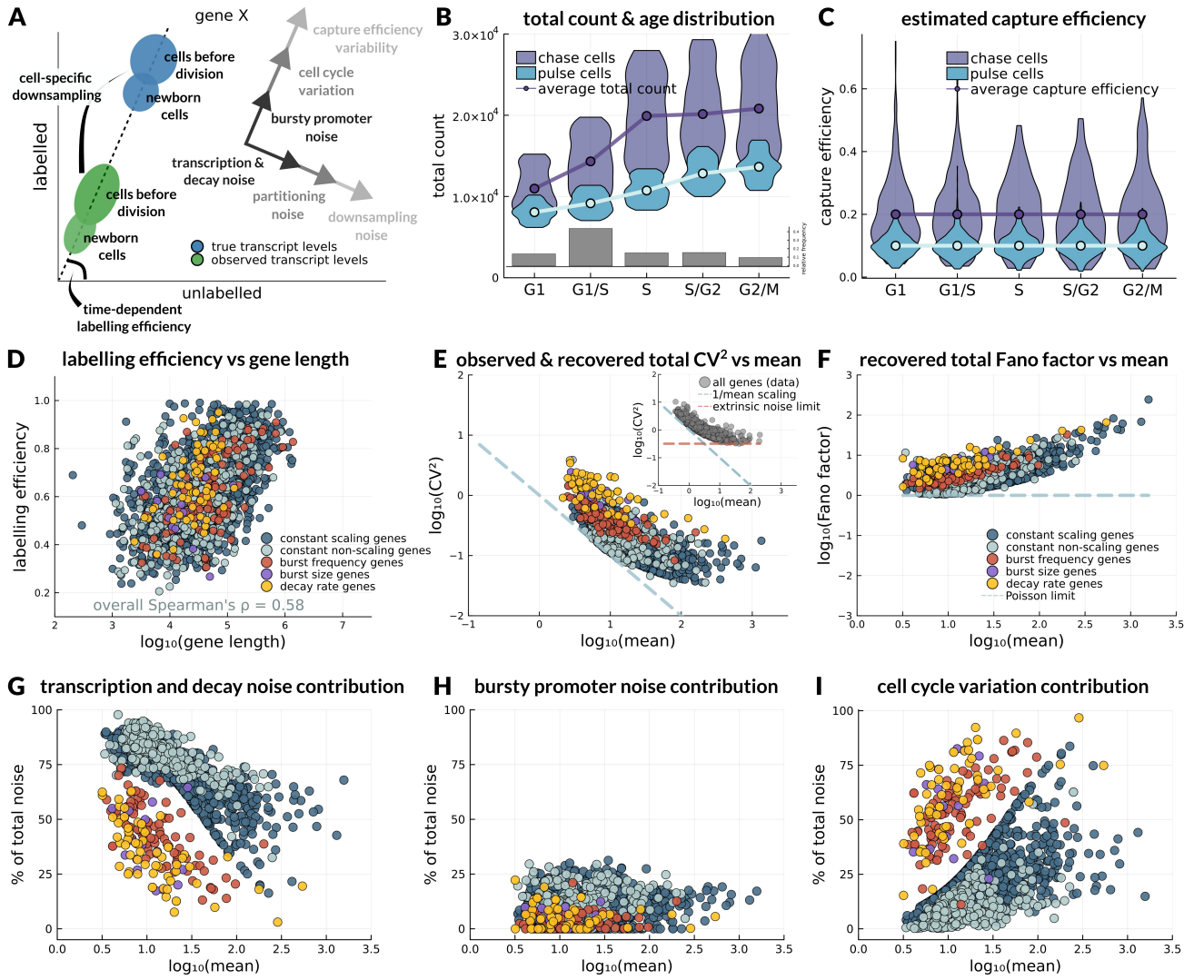


Figure 2.16: Biological and technical effects on transcriptional noise and biological noise decomposition. (A) Cartoon scatter plot of unlabelled-labelled mRNA, where the observed transcript levels (green ellipses) is a noisy version of the true transcript levels (blue ellipses) due to cell-specific downsampling. The vectors describe different noise components that contribute to the total transcript variability by correlating or decorrelating the unlabelled-labelled mRNA levels, illustrated by lengthening or widening the ellipses, respectively. Bursty promoter noise, cell cycle variation (biological) and capture efficiency variability (technical) contribute to correlating unlabelled and labelled transcript levels, while transcription & mRNA degradation noise, partitioning noise due to cell division (biological), and cell-specific downsampling (technical) contribute to decorrelating transcript levels. (B - C) Uncoupling biological and technical effects on transcript variability: (B) Total mRNA count distributions per cell cycle phase and age distribution contribute to biological variability. The average total count is shown to approximately double over the course of the cell cycle. Calculation of the empirical age distribution is described in 2.2.1. (C) Estimated capture efficiency distributions per cell cycle phase. The distributions are shown to be independent of cell cycle progression. (D) Dependence of inferred gene-specific labelling efficiency on gene length. (E-F) Observed and recovered noise-mean relationships: (E) shows the total CV^2 as observed in the data (inset scatter plot in grey) and as recovered by the different models. (F) shows the recovered total Fano factor as given by the model after removing technical noise. (G-I) Model-based biological noise decomposition: proportional contributions of the different biological noise sources to the total noise (see 2.2.5).

contribute to the total transcript variability. In the absence of technical noise, the total variance in transcript levels $x = u + l$ of a gene (u : unlabelled, l : labelled) in a time-resolved single-cell snapshot can be partitioned as follows (see details in 2.2.5):

$$\text{Var}(x) = \underbrace{\mathbb{E}(\text{Var}(u \mid \tau, t) + \text{Var}(l \mid \tau, t))}_{\text{transcription and decay noise}} + \underbrace{2 \mathbb{E}(\text{Cov}(u, l \mid \tau, t))}_{\text{bursty promoter noise}} + \underbrace{\text{Var}(\mathbb{E}(x \mid \tau))}_{\text{cell cycle variation}}$$

where τ denotes the cell cycle phase and t the pulse-chase labelling time (condition). Averaging the cell cycle- and labelling time-dependent transcript variances gives a measure of the noise due to mRNA synthesis and degradation kinetics (1st term). This term includes variation due to both active degradation and dilution, where the latter stems from stochastic partitioning at cell division. The upstream noise due to bursty gene promoter can be measured by averaging the cell cycle- and labelling time-dependent covariances of unlabelled and labelled transcripts (2nd term). The extrinsic noise due to cell cycle variation can be measured by the variance of the cell cycle-dependent mean transcript levels (3rd term). The bursty promoter noise and the cell cycle variation contribute to correlating the unlabelled-labelled transcript levels, while transcription and degradation kinetics noise reduces these correlations (Figure 2.16 (A)). This decomposition is challenging to quantify directly from the data as each component would be corrupted by multiplicative technical noise effects distorting their relative proportions (paragraph 2.2.5).

We therefore devise an approach to account for technical noise in our modelling framework. This allows us to construct reliable summary statistics that are directly comparable with the noisy data. Quantifying what part of transcript variation comes from technical noise requires access to the capture efficiency of the sequencing experiment, however this information is not provided by the scEU-seq protocol. In principle, actual and observed mRNA counts can be connected via a binomial downsampling model that depends on a cell-specific capture efficiency, but in our case, the dependence on cell size and cell cycle needs to be considered (Figure 2.16 (B)). We estimate the capture efficiencies directly from the data. The idea is that the total mRNA count of a cell is proportional to cell size and it increases along the cell cycle so it provides an estimate of the relative capture efficiency of the cell within its cell cycle phase (paragraph 2.2.3). Also, this estimate differs between pulse and chase conditions. We show that capture efficiency distributions per cell cycle phase do not scale with cell cycle progression (Figure 2.16 (C)). Our approach hence decouples cell size and age variability from technical noise effects on total mRNA count. Moreover, it normalises the total count and sample size differences between

pulse and chase conditions. Having obtained the capture efficiency estimates, we can correct the model outputs by computing downsampled versions of moments (paragraph 2.2.3).

An additional technical noise source at the single-gene level is the labelling efficiency of the pulse-chase experiments, as RNA labelling might not impact genes uniformly. To address this, we model labelling efficiency as a gene-specific parameter that is fitted simultaneously with the kinetic rates. We find that the inferred labelling efficiency shows significant variation across genes as well as that it correlates with gene length, demonstrating the importance of estimating it as a gene-specific parameter (Figure 2.16 (D)). It is also evident that this trend is independent of the selected gene-specific models, thus validating that it does not influence model selection.

2.4.5 Bayesian inference reconstructs the transcriptional noise landscape

Our approach of uncoupling technical and biological noise makes it possible to estimate the true mean expression levels and the true levels of transcriptional noise. The observed total noise (CV^2) for all 3419 genes decreases with mean expression but stays above the Poisson limit (Figure 2.16 (E) inset), and deviates from the inverse mean scaling for highly expressed genes. This deviation at high expression levels reflects extrinsic noise attributed to both biological and technical effects. Informing the model with the point estimates from the inferred kinetics posteriors, we recover and visualise the true noise-mean relationship (Figure 2.16 (E)). The dependence of noise on the mean recovered by the model follows a similar pattern to the data, but the noise is an order of magnitude lower and deviates less from the inverse mean scaling law as these deviations are purely due to biological noise. We observe distinct noise patterns among groups of genes explained by different models, with cell cycle-dependent genes commonly displaying higher noise levels than constant genes. Furthermore, using the Fano factor to quantify the noise-mean relationship, we observe that the increase in the mean is associated with increasing deviation from the Poisson limit due to extrinsic noise (Figure 2.16 (F)).

Next, we use the noise decomposition formula to quantify how the different biological noise sources contribute to the total transcript variability (Figure 2.16 (G-I)). We find that transcription and decay kinetics and cell cycle variation are the most prominent sources of noise.

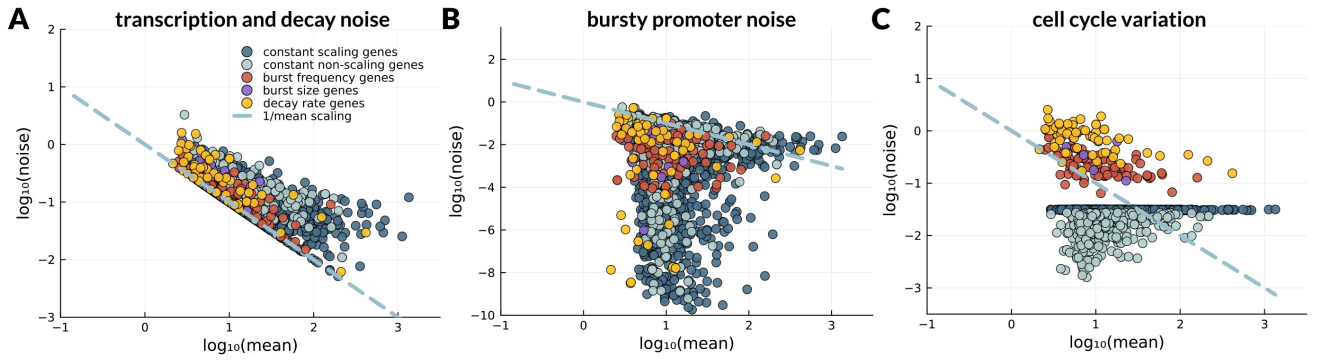


Figure 2.17: **Absolute levels of biological noise sources.** Absolute levels of noise dependence on the mean across the different gene classes, for each noise component: (A) transcription and decay noise, (B) bursty promoter noise and (C) cell cycle variation.

Interestingly, for most constant genes the largest proportion of noise is explained by mRNA synthesis and decay kinetics (Figure 2.16 (G)) that contribute to more than 50% of the total noise. On the other hand, cell cycle variation is the dominant source of noise for most non-constant genes (Figure 2.16 (I)). The bursty promoter noise is less prominent across genes, however not negligible as in a considerable number of genes it explains more than 20% of the noise. Constant and non-constant genes can be distinguished by their absolute levels of cell cycle variation (Figure 2.17).

To gain mechanistic insights into the determinants of transcriptional noise, we illustrate how the different noise sources depend on the kinetic rates. We find that the decay rate is a key determinant of all noise components. Transcription and decay noise and bursty promoter noise increase with decay rate (Figure 2.18 (A-B)). The cell cycle variation of constant scaling genes is constant and hence does not depend on the decay rate, however the cell cycle variation of non-scaling genes decreases with decay rate (Figure 2.18 (C)). These observations agree with model predictions reported in [175] and [180]. We find that transcription and decay noise is relatively independent of burst size (Figure 2.19 (A)). Additionally, bursty promoter noise increases with burst size for large decay rates (Figure 2.19 (B)), while it decreases with burst frequency for small decay rates (Figure 2.19 (C)). These findings highlight that quantifying absolute kinetic rates in combination with model-based noise decomposition reveals the sources of transcriptional variation.

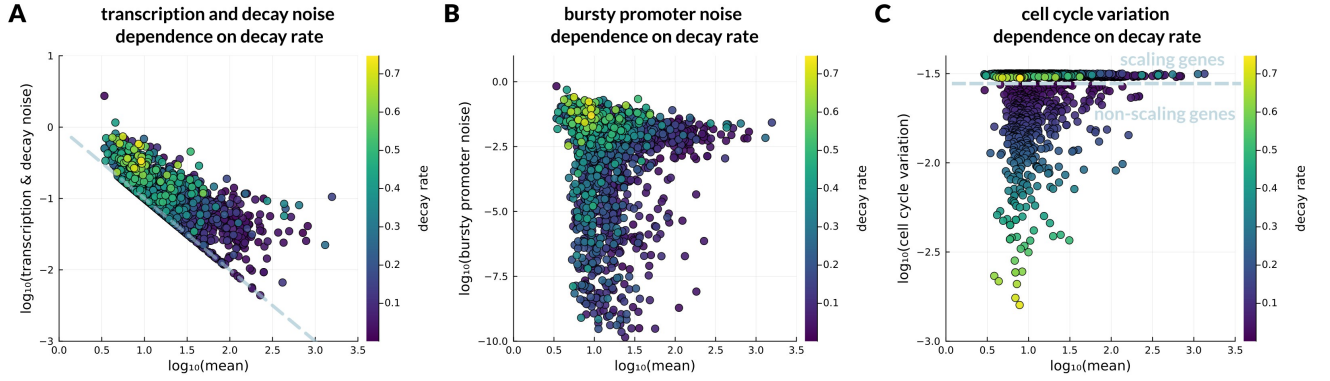


Figure 2.18: **Dependence of noise decomposition on decay rate.** CV^2 against mean expression coloured by decay rate, plotted for all three noise components (2.4.4). We only plot decay rates of constant (scaling and non-scaling) genes. Decay rate is measured in h^{-1} .

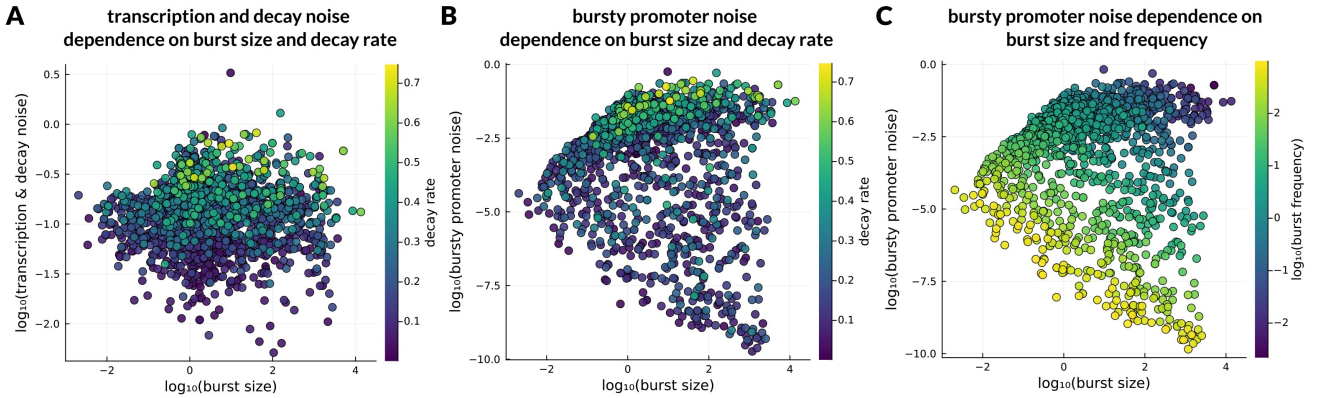


Figure 2.19: **Dependence of noise decomposition on burst size and frequency.** CV^2 of expression against burst size coloured by decay rate or burst frequency, plotted for the two intrinsic noise components. We only plot kinetic rates associated to constant (scaling and non-scaling) genes. Burst size and frequency are shown in $\log_{10}$ scale and decay rate is shown in linear scale. (A) Transcription and decay noise increases with decay rate, while there is no direct dependence on burst size. (B) Bursty promoter noise increases with burst size at the limit of high decay rate. There is no clear dependence observed in the low decay rate regime. (C) Bursty promoter noise decreases with burst size when burst frequency is fixed.

2.4.6 The model identifies scaling properties of gene expression with cell size

We next focus on genes with constant kinetics and we find that the model reveals key differences between size-dependent and size-independent gene expression. As an example, *RAB7A*, a gene that is involved in intracellular vesicle trafficking and growth-factor-mediated signalling, is explained by a constant scaling model and *MGST1*, a gene responsible for protecting endoplasmic membranes from oxidative stress, is explained by a non-scaling model (Figure 2.20 (A)). *RAB7A* shows an increase in the mean and Fano factor of transcript levels along the cell cycle (Figure 2.20 (A i-ii)), which is explained by a size-scaling mRNA synthesis rate (Figure 2.20 (A iii)), while the mean transcript levels and Fano factor of *MGST1* stay relatively constant during cell cycle progression (Figure 2.20 (A iv-v)), which is explained by a constant synthesis rate (Figure 2.20 (A vi)).

We investigate global size-scaling properties of gene expression. We find that the mean expression of scaling genes increases monotonically during the cell cycle, while the mean expression of non-scaling genes reaches a steady state towards the end of cell cycle progression. This can be observed in clustered heatmaps of both observed and recovered (after technical noise removal) mean expression of all constant genes (Figure 2.20 (B)). To validate this observation, we quantify changes in mean expression of constant genes along the cell cycle. We find that both the fold change of expression between early and late cell cycle and the mean rate of change of expression along cell cycle progression are higher for scaling compared to non-scaling genes (Figures 2.20 (C-D)). Additionally, we quantify variation due to cell cycle-dependence, including variation observed in the data and variation recovered from the model after technical noise correction (see 2.2.5). We find that in both cases, cell cycle variation of size-scaling expression is higher compared to non-scaling (Figure 2.20 (E)). This suggests that the combined signatures of mean expression and noise levels identify the scaling with cell size.

We analysed which gene functions were associated with non-scaling gene expression using the enrichment analysis provided by the DAVID bioinformatics database [205]. We find that the majority of non-scaling genes are involved in cell surface receptor signalling. Other enriched sets include protein metabolic processes, regulation of cell proliferation, tissue and system development and proteolysis. Finally, we asked how scaling of transcription with cell size relates

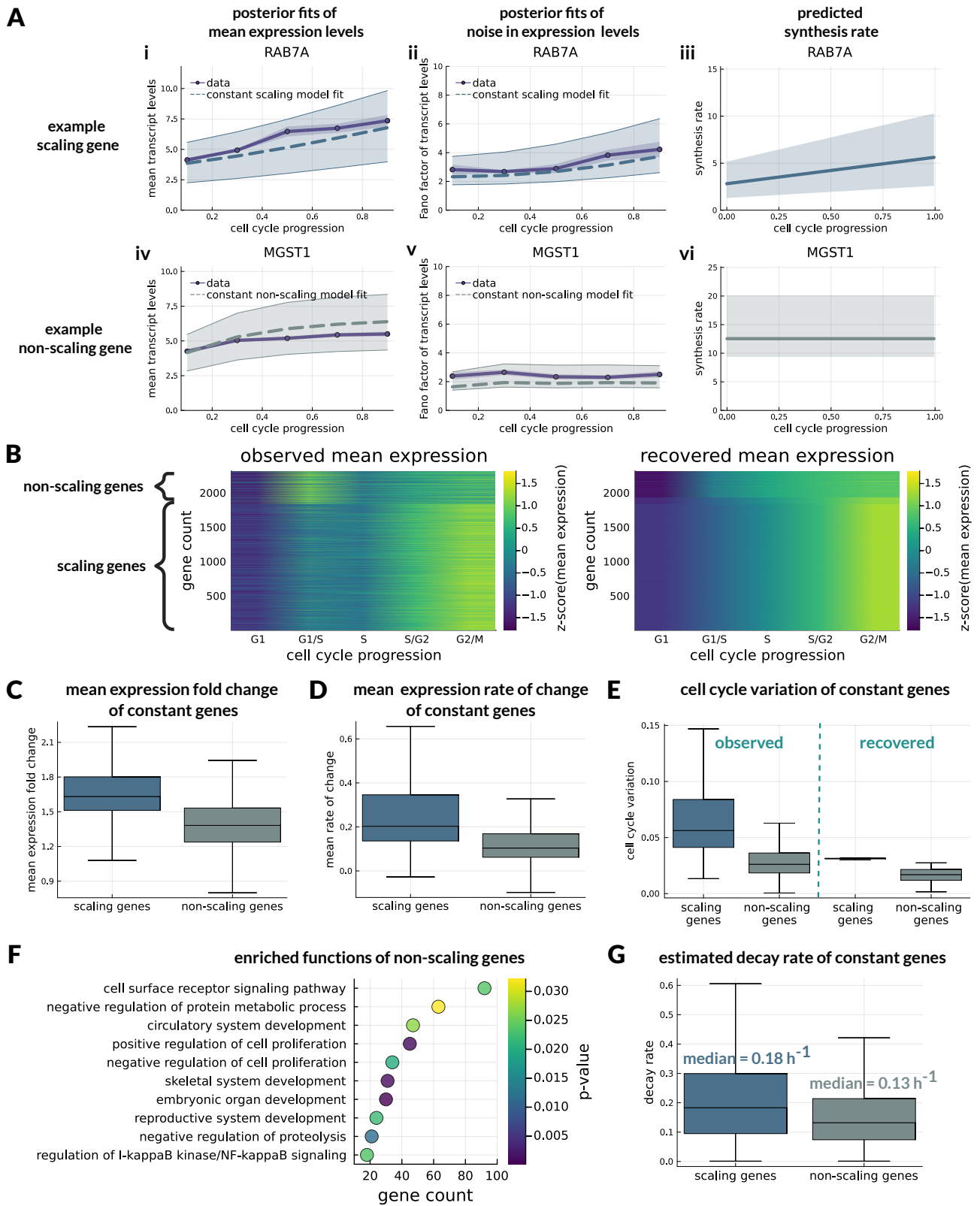


Figure 2.20: **Constant genes and size-scaling properties of gene expression.** (A) Posterior model fits for a scaling and a non-scaling gene. Panels (i)-(ii) and (iv)-(v) show cell cycle-dependent mean and Fano factor of transcript levels, and panels (iii) and (vi) show posterior estimates of scaling and non-scaling synthesis rate. (B) Clustered heatmaps of normalised mean expression along the cell cycle for constant genes, as observed in the data (left) and as recovered by the model (right), after removing technical noise effects. (C-E) Differences between scaling and non-scaling genes in cell cycle-dependent expression: (C) shows differences in the observed fold change of mean expression between G1 and G2/M phase, (D) shows differences in mean rates of change of expression along consecutive cell cycle phases and (E) shows differences in variability due to cell cycle-dependence, both observed in the data and recovered by the model. (F) Enriched GO terms of non-scaling genes. (G) Differences in decay rate distributions of scaling and non-scaling genes.

to mRNA degradation. We find that scaling genes have overall higher decay rates compared to non-scaling genes (Figure 2.20 (G)). As a consequence, scaling genes are characterised by shorter mRNA half-life, with their median mRNA half-life being shorter by more than one hour compared to non-scaling genes (Figure 2.11 (A)).

2.4.7 The model predicts patterns of cell cycle-dependent kinetics regulation

Next, we focus on genes with non-constant kinetics and we find that the model identifies three groups of genes, each characterised by a distinct mode of kinetics regulation during the cell cycle. For these groups of genes, either burst frequency, burst size or degradation rate are primarily modulated over the cell cycle, while the rest of kinetic rates stay constant. Posterior model fits for *CDK4*, *PCNA* and *CDK1*, which are genes with key functional roles in cell cycle control, display good agreement with the data (Figure 2.21 (A)). The variations in mean and Fano factor of *CDK4*, *PCNA* and *CDK1* transcript levels along the cell cycle (Figure 2.21 (A i-ii, iv-v, vii-vii)) are explained by a particular modulation of their kinetic rates (Figure 2.21 (A iii, vi, ix)). Interestingly, in all three cases the predicted kinetics show a peak in a particular cell cycle phase. The increase in mean and noise of expression of these genes in the late part of the cell cycle is achieved by a sharp increase in burst frequency or burst size about S phase, or by a sharp decrease in decay rate after G1 phase, respectively.

To obtain comprehensive insights into the global regulation of kinetics during the cell cycle, we perform hierarchical clustering of each set of varying kinetic rates based on cosine similarity (Figure 2.21 (B)). Strikingly, all emerging clusters are characterised by a unique peak in a particular cell cycle phase. Combining all clusters together, we observe a wave-like pattern in the overall kinetics regulation throughout the cell cycle. We find that varying decay rate shows peaks in the early stages of the cell cycle, while varying burst size and frequency show peaks in S, G2 and M phase. This suggests distinct regulatory strategies of transcription and degradation during the cell cycle. This non-trivial clustering of genes based on their kinetics would not be possible to extract from mean expression dynamics alone, as the different gene groups show similar expression profiles, with an increase towards the late cell cycle stages (Figure 2.22).

We also find, using gene enrichment analysis, that each mode of kinetics regulation is associ-

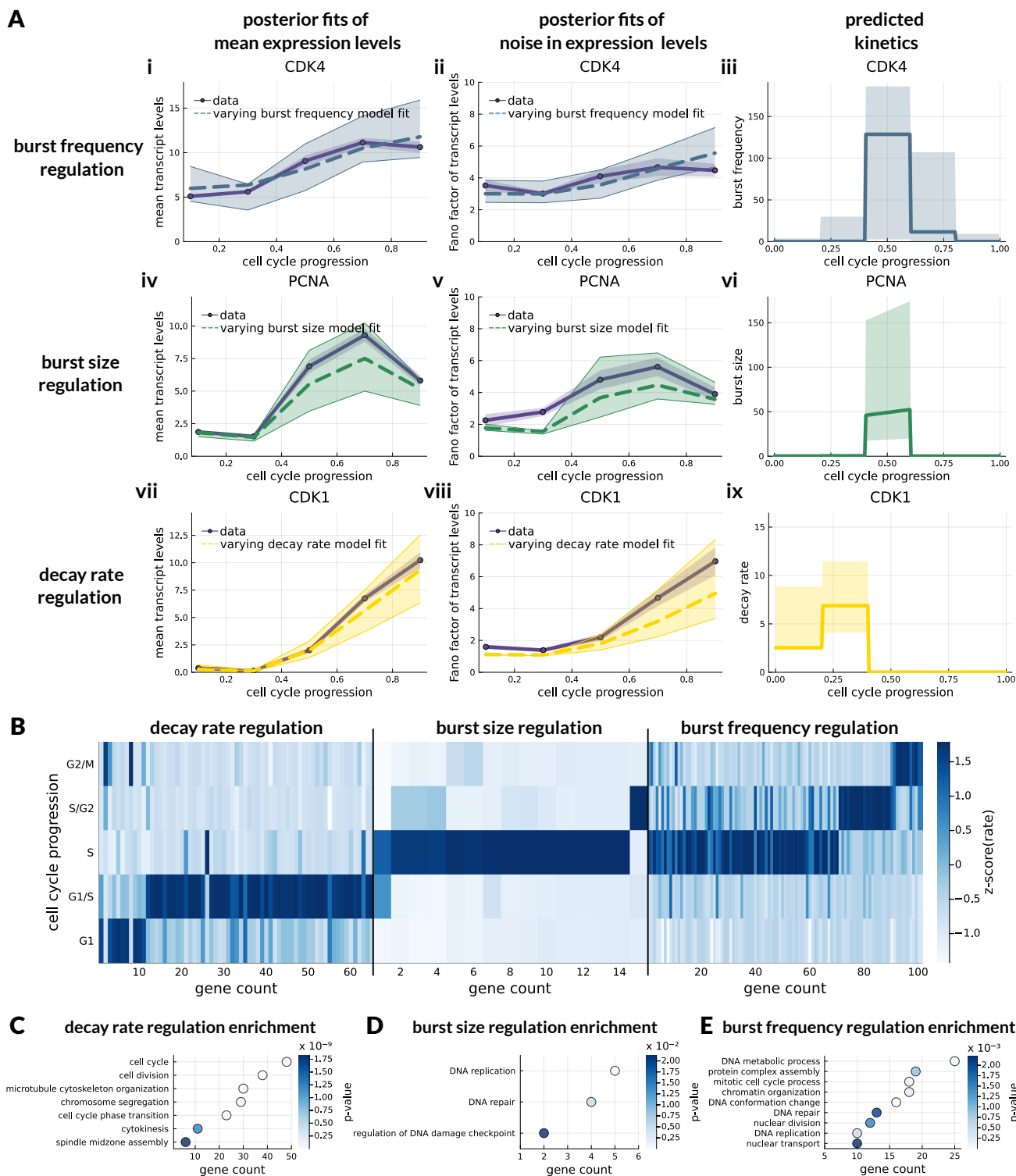


Figure 2.21: Non-constant genes and cell cycle-dependent regulation of kinetic rates. (A) Posterior model fits for three example non-constant genes. Panels (i)-(ii), (iv)-(v) and (vii)-(viii) show cell cycle-dependent mean and Fano factor of transcript levels, and panels (iii), (vi) and (ix) show posterior estimates of cell cycle-modulated kinetic rates. (B) Clustered heatmap of cell cycle-modulated kinetic rates: the different sections of the heatmap correspond to decay rate genes, burst size genes and burst frequency genes. (C-E) Enriched GO terms associated with non-constant genes: (C) shows GO terms linked with decay rate regulation, (D) shows GO terms linked with burst size regulation and (E) shows GO terms linked with burst frequency regulation.

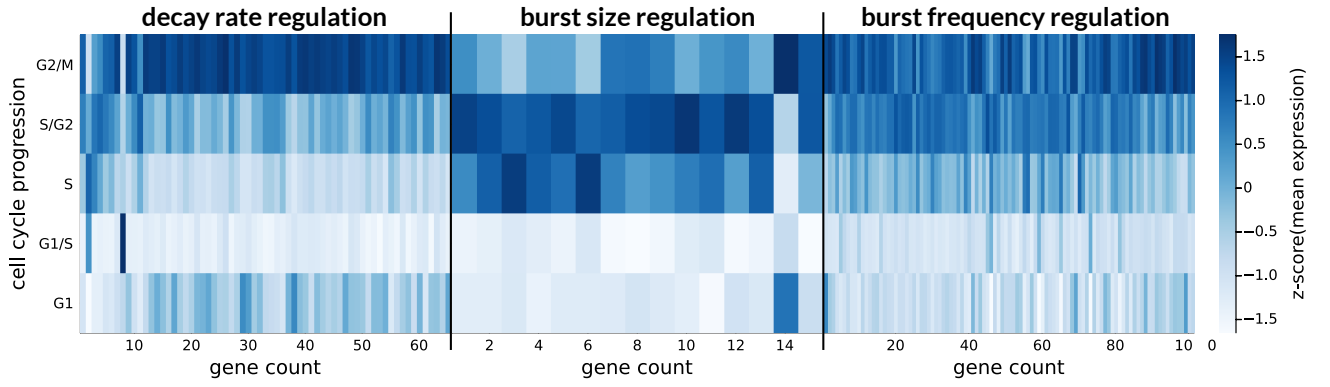


Figure 2.22: **Mean expression profiles of cell cycle-dependent genes.** Heatmaps of mean expression z-scores of decay rate genes, burst size genes and burst frequency genes. The ordering of the genes corresponds exactly to the ordering of kinetic rates in the heatmap of Figure 2.21 (B).

ated with a particular set of functions related to cell cycle control and cell division (Figures 2.21 (C-E)). Genes with varying decay rate are found to be involved in the regulation of G2/M phase transition and cell division, including chromosome segregation, spindle assembly and cytokinesis (Figure 2.21 (C)). Genes with varying burst size are primarily involved in S phase regulation and DNA replication, including DNA repair and damage checkpoint control (Figure 2.21 (D)). Varying burst frequency genes are associated with a variety of cell cycle control functions, including DNA replication, DNA repair and nuclear division (Figure 2.21 (E)). Interestingly, burst frequency regulation is also associated with other cellular functions such as protein complex assembly, chromatin organisation and DNA conformation changes, suggesting a potential link of these processes with the cell cycle. Overall, these results indicate that biological functions responsible for cell cycle regulation could be achieved by the coordinated activation of kinetic rates in particular cell cycle phases.

2.5 Discussion

In this chapter, we developed a comprehensive modelling framework to reveal global regulatory mechanisms of mRNA transcription and degradation during the cell cycle from time-resolved transcriptomics data. Our stochastic model integrates cell cycle progression and pulse-chase labelling to reveal different modes of dynamic regulation in mRNA control. Modelling capture and labelling efficiency allowed us to deconvolve technical and biological noise of the data and reveal the true sources of cellular variation that stem from distinct mechanisms of transcription regulation.

A central result of our analysis is that patterns of cell cycle- and cell size-dependent regulation of transcription dynamics are distinguished through dynamic correlations of metabolic labels. In particular, we showed that correlations between unlabelled and labelled mRNAs quantify transcriptional bursting, while transcript noise statistics quantify mRNA synthesis and degradation kinetics (Figure 2.7). Hence, these summary statistics distinguish different regulatory mechanisms. Existing models of metabolic labelling protocols are limited to averages across cells as they employ deterministic approaches [121, 155, 178, 187–191] or they are restricted to labelled transcripts [95, 194] and hence cannot make use of dynamic correlations.

Our approach, which is based on the ABC rejection method, effectively samples from the full posterior distribution across parameters and models by comparing millions of candidate simulations to any number of genes in parallel. This is equivalent to recycling all particles for each gene. The method results in an amortised inference using data-independent simulations that is more efficient than data-specific approaches such as sequential Monte Carlo [107]. Amortised simulation-based inference incurs an up-front computational cost, but is applicable to new data points efficiently [103]. Our method is thus ideally suited for the analysis of genome-wide data [102].

A distinctive feature of our approach is that it simultaneously corrects for technical variation through cell-specific capture efficiency and gene-specific metabolic labelling efficiency. Previous inference methods utilising metabolic labelling data relied on specific distribution assumptions [95, 155] or on RNA spike-ins [95] for modelling capture efficiency, while other studies modelled labelling efficiency using either a global rescaling or mixture models [121, 178, 187–189, 193, 206]. We verified that capture efficiency was cell cycle-independent, thus validating that we

can distinguish the effects of cell cycle and size. We also showed that labelling efficiency varied dramatically and correlated well with gene length but did not bias our model selection, thus demonstrating that one needs to calibrate labelling efficiency for parameter estimation. The observed correlation between labelling efficiency and gene length may provide insights into the mechanistic basis of metabolic labelling.

By removing technical noise effects, we were able to dissect sources of biological noise contributing to the total transcript variability. By quantifying absolute kinetic rates, we revealed how these sources of transcriptional variation depend on decay rate. We found that mRNA synthesis and degradation kinetics and cell cycle-dependent variation are the primary genome-wide sources of noise, while only a small fraction of the total noise stems from transcriptional bursting. The low levels of bursting noise were explained by overall low burst sizes and relatively high burst frequencies. These observations indicate that transcript abundance is characterised by minimal intrinsic stochasticity, as demonstrated in [167].

We found that most genes scale transcription with cell size. Genes that did not scale were either involved in growth-independent processes, such as cell surface receptor signalling, or in other processes such as proliferation and development. Cell surface and DNA binding proteins do not scale with cell volume and have been previously identified as potential non-scaling genes [207–209]. Size-scaling genes typically rise more rapidly during the cell cycle, have higher degradation rates, and have higher cell cycle variation than non-scaling genes. This suggests that increased synthesis rate with cell size requires higher degradation for cells to achieve homeostasis [181, 210].

Our model predicts waves of regulation across the cell cycle-dependent transcriptome, with kinetic rates peaking at specific cell cycle phases. We observed no step-wise transcriptional changes, suggesting that gene dosage effects mattered less than specific cell cycle regulation of kinetic rates. Genes with kinetics modulating decay rate peaked first in G1, then burst size peaked in S and S/G2, then burst frequency peaked in S phase and the later stages of the cell cycle. A possible explanation of our findings is that global transcription and degradation regulators act in different cell cycle phases and that burst size and frequency modulation is mediated by different transcription factors. Our predictions agree qualitatively with a similar sharp decrease in burst frequency following S phase observed using smFISH [199], and the increased synthesis rates and decreased decay rates following G1 reported in the original scEU-seq

study [155]. A mechanistic understanding of this phenomenon would require a gene regulatory network-based approach that is beyond the scope of our study.

Although we expect our inference method to generalise to more complex models and parametric assumptions, it also has several limitations. One limitation is that ABC rejection sampling can lead to weaker accuracy in parameter estimation and broader posterior distributions compared to optimisation-based schemes (see e.g. Figure 2.9). A hybrid method that uses ABC rejection sampling as a first step, followed by an SMC step where the particles for each gene are optimised, would offer a better tradeoff between parallelisation and accuracy [102, 107]. An alternative approach would be to use a synthetic likelihood or integrate ABC with a machine learning method that allows for surrogate modelling [108]. Another limitation is that we assumed deterministic cell cycles across cells, thus a potential improvement would be to account for cell cycle length variability in our moment estimates.

An important challenge for our study was that gene filtering was required as a preprocessing step due to the high sparsity of the data - a common issue in single-cell analyses [211, 212]. Even though this gene selection posed a limitation to applying our inference at full genome-wide scale, it did not limit the ability of our method to extract global patterns. Future advances in single-cell sequencing technologies with enhanced transcript detection, deeper sequencing and increased cell capture, will be essential for improving data quality and overcoming these limitations [46, 73].

Other studies use modelling to better understand transcription dynamics through mRNA splicing [115–117, 121, 127] including cell cycle-dependent effects [119] and during development [116, 193]. Therefore, splicing data could be used instead of cell cycle reporters in our inference method. Similarly, splicing kinetics could extend our analysis to understand transcription variation along developmental trajectories [115, 116, 193] or identify hidden sources of cellular variation [184].

Our study illustrates how mechanistic modelling and amortised inference can efficiently integrate data across labelled, unlabelled transcripts and cell cycle reporters and also across different pulse and chase protocols. We have demonstrated that mechanistic models can be used to overcome the challenge of technical noise across modalities commonly suffered by multi-view statistical methods [213, 214]. Data integration across conditions, time and modalities is an important challenge for single-cell multi-omics, including epigenome and proteome [215–217],

and mechanistic models like ours could provide a versatile tool for integrating these data [97].

In summary, our study reveals global mechanisms of transcription regulation through stochastic modelling of time-resolved single-cell data. Our inference methodology is highly scalable and translates seemingly to complex models of gene regulation. It is thus well suited to uncover complex mechanisms of gene expression control from single-cell transcriptomics. We expect that our framework could contribute to uncover regulatory mechanisms that drive cell fate decisions or their dysregulation in disease.

Chapter 3

Identifying modes of gene regulation in time-resolved single-cell transcriptomics

3.1 Introduction

In Chapter 2, we developed a modelling framework to quantify transcription dynamics from time-resolved single-cell RNA sequencing data. A limitation of this framework is that the dynamics of each gene is modelled independently, neglecting pairwise gene interactions. However, genes operate within regulatory networks, and their interactions significantly influence gene expression dynamics [6, 156, 218]. The precise coordination of gene-gene interactions enables cells to perform diverse functions and adapt to environmental stress [12, 25]. Consequently, gene regulatory networks (GRNs) are essential determinants of cellular decision making [14, 17, 18]. Detecting pairwise gene relationships and reconstructing GRNs is therefore of fundamental importance in biology [157, 158, 218–220]. Investigating network-level gene regulation requires computational methods that describe regulatory relationships between genes, extending beyond gene-specific modelling approaches [221, 222].

A wide range of experimental technologies have been developed for discovering genome-wide regulatory interactions [223]. As an example, methods such as ChIP-seq [224] and CUT&Tag [225] have enabled the analysis of protein-DNA interactions by identifying genome-wide binding

sites for any protein of interest. Other methods rely on genetic perturbations, to track expression changes caused by *e.g.* gene knockouts or mutations and detect putative gene relationships [222, 226, 227]. Recently, large-scale perturbation experiments have become possible at single-cell resolution, with technologies that combine pooled CRISPR screens with single-cell RNA sequencing [228–231]. However, all these methods require targeting and analysing a particular set of proteins or transcription factors, and thus do not enable a fully unbiased discovery of regulatory interactions.

It has been argued that cell-to-cell variability in gene expression can be viewed as a natural perturbation, providing a means to systematically identify gene regulatory relationships without requiring large-scale perturbation experiments [6, 25, 27, 87, 91, 159, 232–236]. Single-cell genomics technologies have created new opportunities for exploiting cellular heterogeneity to infer gene regulatory networks from data [222, 237, 238]. Numerous computational methods utilise statistical dependencies between genes across single cells to predict putative regulatory relationships [79–82, 235, 239–245]. These dependencies, typically captured by correlations or information-theoretic measures, are used to reconstruct co-expression or causal regulatory networks. A few studies have employed stochastic modelling approaches to reconstruct GRNs from single-cell transcriptomics [91, 245–247]. Yet, little emphasis is given in obtaining a mechanistic understanding of observed gene-gene covariation in single-cell data.

Detecting regulatory relationships from gene-gene correlations in scRNA-seq data is particularly challenging due to several inherent limitations. First, gene-gene interaction can cause time-shifted covariation that is hard to capture using simultaneous transcriptomic correlations [91, 232]. Transcriptional covariation can also arise from various biological sources of stochasticity, such as variation in global cellular states (*e.g.* cell cycle stage and cell size) or other upstream noise sources. Furthermore, technical noise, including cell-to-cell variability in capture efficiency [141] and transcript detection biases due to gene length [248, 249], can obscure true biological correlations or create spurious ones [142, 245]. The difficulty to distinguish these sources of covariation in noise-corrupted snapshot data poses a significant constraint in identifying gene regulatory interactions. Many existing gene network inference approaches lack a mechanistic framework to delineate sources of covariation, thus potentially failing to explain which correlations stem from true gene interaction [91].

Even when biologically meaningful correlations can be detected, it remains challenging to dis-

cern whether these correlations arise from direct gene interaction or gene-gene co-regulation, *e.g.* through a common upstream factor [250]. In principle, this problem can be addressed using a model selection approach [251], where multiple mechanistic models are tested against the data to infer possible modes of regulation underlying the observed correlations. Given the difficulty to obtain analytical likelihoods for complex models of gene regulation, one viable option would be to use approximate Bayesian computation (ABC) for comparing a large number of simulated model outputs with the data. As demonstrated in Chapter 2, ABC enables Bayesian model selection [203, 252], and provides an amortised method [103] to test candidate models against any number of unseen datapoints. Yet, this approach would suffer from high computational costs and the difficulty to select informative summary statistics, a common challenge for ABC methods [253]. An alternative approach would be to combine a simulation-based method with machine learning [103] to discover informative data features distinguishing different regulatory models [108]. This approach would effectively integrate generative mechanistic models with a discriminative method for model selection [254].

Here, we develop a method that combines mechanistic modelling with deep learning to identify modes of gene regulation in time-resolved scRNA-seq data. We extend the gene-specific modelling approach of Chapter 2 by building stochastic models of gene regulatory interaction driven by direct regulation and co-regulation. Our framework accounts for distinct sources of biological and technical covariation, enabling the detection of correlations arising from either gene interaction or extrinsic noise sources. The models incorporate RNA labelling and cell cycle phase information of scEU-seq data [155] to define temporal statistics of gene co-expression. Leveraging inferred kinetic rates of individual genes (Chapter 2), we generate synthetic data to train neural networks for classifying different regulatory models. This results in a generative-discriminative modelling method capable of predicting whether any two genes have a potential regulatory relationship and whether they directly interact or are co-regulated by a shared upstream factor. We demonstrate that distinct modes of gene regulation can be identified from non-linear combinations of dynamical correlations across RNA labelling timepoints. Our findings reveal that a large proportion of observed gene-gene correlations stem from co-regulation, including upstream noise sources. Additionally, we validate that gene pairs predicted to interact correspond to known regulatory relationships. Our study provides a mechanistic approach to interpreting transcriptional covariation and highlights possibilities of inferring gene regulatory interactions from time-resolved single-cell transcriptomics.

3.2 Stochastic modelling of gene regulatory interactions

We start by presenting stochastic reaction network descriptions for different modes of gene regulation. Importantly, our framework considers regulatory relationships in a broad sense that is not restricted to transcription factor-mediated interactions. Our models capture biological and technical noise sources, enabling the analysis of gene-gene relationships arising from both gene interaction and other sources of transcriptional covariation. We develop reaction kinetics models capturing the intrinsic dynamics of gene pairs, the extrinsic dependence of reaction rates on the cell cycle, as well as the experimental effects of metabolic labelling, effectively integrating information from time-resolved scRNA-seq data. Additionally, we formulate moment equations that govern the dynamics of 1st and 2nd order moments of the networks' molecular states. Next, we present a theory for quantifying pairwise gene relationships through their transcriptional covariation, that accounts for different biological and technical noise sources.

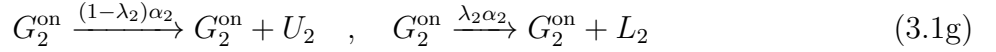
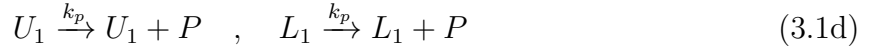
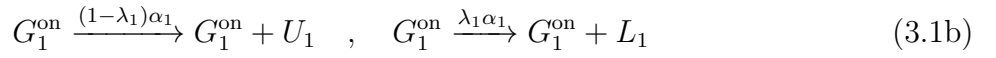
3.2.1 Stochastic reaction kinetics of interacting genes

Gene-gene interactions can be direct, for instance when a transcription factor upregulates or downregulates the expression of a downstream gene, or indirect, occurring through intermediate factors [218]. Here we consider 3 different basic models of gene interaction that can give rise to transcriptional covariation: *positive regulation* (or upregulation), *negative regulation* (or downregulation) and *co-regulation*.

Models of gene (co-)regulation

We here outline the reaction schemes encoding the different modes of gene regulation. The kinetic rates driving the reactions below include the promoter activation rate (or burst frequency) k_{on} , the promoter repression rate k_{off} , the transcription rate α , the mRNA degradation rate γ , the translation rate k_p and the protein degradation rate δ . As we want our models to capture the effects of the metabolic labelling experiments, we use λ to denote the gene-specific metabolic labelling efficiency, that controls the fraction of the mRNA molecule that is chemically labelled by the experiment. We also assume that both the unlabelled and labelled parts of an mRNA molecule contribute to protein synthesis.

Positive regulation model. The following reaction scheme encodes the transcription kinetics of a gene G_1 encoding *e.g.* a regulator or transcription factor (TF), that upregulates a target gene G_2 . To model the direct interaction of the TF with the target gene we postulate that the target gene's burst frequency explicitly depends on the protein levels of the TF, *i.e.* $k_{\text{on},2} = k_{\text{on},2}(P)$. In particular, we assume that the burst frequency of the target gene is proportional to the protein levels $k_{\text{on},2}(P) = k_{\text{on},2} \cdot P$.



Negative regulation model. Similarly to the previous model, the following reaction scheme describes a gene G_1 encoding a regulator or TF, that downregulates a target gene G_2 . The interaction is modelled by assuming that the repression rate $k_{\text{off},2}$ of the target gene is proportional

to the protein levels of the TF, *i.e.* $k_{\text{off},2}(P) = k_{\text{off},2} \cdot P$.

$$G_1^{\text{off}} \xrightleftharpoons[k_{\text{off},1}]{k_{\text{on},1}} G_1^{\text{on}} \quad (3.2a)$$

$$G_1^{\text{on}} \xrightarrow{(1-\lambda_1)\alpha_1} G_1^{\text{on}} + U_1 \quad , \quad G_1^{\text{on}} \xrightarrow{\lambda_1\alpha_1} G_1^{\text{on}} + L_1 \quad (3.2b)$$

$$U_1 \xrightarrow{\gamma_1} \emptyset \quad , \quad L_1 \xrightarrow{\gamma_1} \emptyset \quad (3.2c)$$

$$U_1 \xrightarrow{k_p} U_1 + P \quad , \quad L_1 \xrightarrow{k_p} L_1 + P \quad (3.2d)$$

$$P \xrightarrow{\delta} \emptyset \quad (3.2e)$$

$$G_2^{\text{off}} \xrightleftharpoons[k_{\text{off},2}P]{k_{\text{on},2}} G_2^{\text{on}} \quad (3.2f)$$

$$G_2^{\text{on}} \xrightarrow{(1-\lambda_2)\alpha_2} G_2^{\text{on}} + U_2 \quad , \quad G_2^{\text{on}} \xrightarrow{\lambda_2\alpha_2} G_2^{\text{on}} + L_2 \quad (3.2g)$$

$$U_2 \xrightarrow{\gamma_2} \emptyset \quad , \quad L_2 \xrightarrow{\gamma_2} \emptyset \quad (3.2h)$$

Co-regulation model. We also consider a more general model where two genes do not interact directly, but are co-regulated by a common upstream factor. The following reaction scheme encodes the upregulation of two target genes G_1 and G_2 by the factor G_0 . It is important to note here that G_0 is either a common regulator, or more broadly represents an upstream source of cellular noise that influences the activity of both target genes. In the reaction scheme, G_0 encodes a protein P that upregulates both genes G_1 and G_2 by acting on their promoter activation rates, *i.e.* $k_{\text{on},1}(P) = k_{\text{on},1} \cdot P$.

$$G_0^{\text{off}} \xrightleftharpoons[k_{\text{off},0}]{k_{\text{on},0}} G_0^{\text{on}} \quad (3.3a)$$

$$G_0^{\text{on}} \xrightarrow{k_p} G_0^{\text{on}} + P \quad (3.3b)$$

$$P \xrightarrow{\delta} \emptyset \quad (3.3c)$$

$$G_1^{\text{off}} \xrightleftharpoons[k_{\text{off},1}]{k_{\text{on},1}P} G_1^{\text{on}} \quad (3.3d)$$

$$G_1^{\text{on}} \xrightarrow{(1-\lambda_1)\alpha_1} G_1^{\text{on}} + U_1 \quad , \quad G_1^{\text{on}} \xrightarrow{\lambda_1\alpha_1} G_1^{\text{on}} + L_1 \quad (3.3e)$$

$$U_1 \xrightarrow{\gamma_1} \emptyset \quad , \quad L_1 \xrightarrow{\gamma_1} \emptyset \quad (3.3f)$$

$$G_2^{\text{off}} \xrightleftharpoons[k_{\text{off},2}]{k_{\text{on},2}P} G_2^{\text{on}} \quad (3.3g)$$

$$G_2^{\text{on}} \xrightarrow{(1-\lambda_2)\alpha_2} G_2^{\text{on}} + U_2 \quad , \quad G_2^{\text{on}} \xrightarrow{\lambda_2\alpha_2} G_2^{\text{on}} + L_2 \quad (3.3h)$$

$$U_2 \xrightarrow{\gamma_2} \emptyset \quad , \quad L_2 \xrightarrow{\gamma_2} \emptyset \quad (3.3i)$$

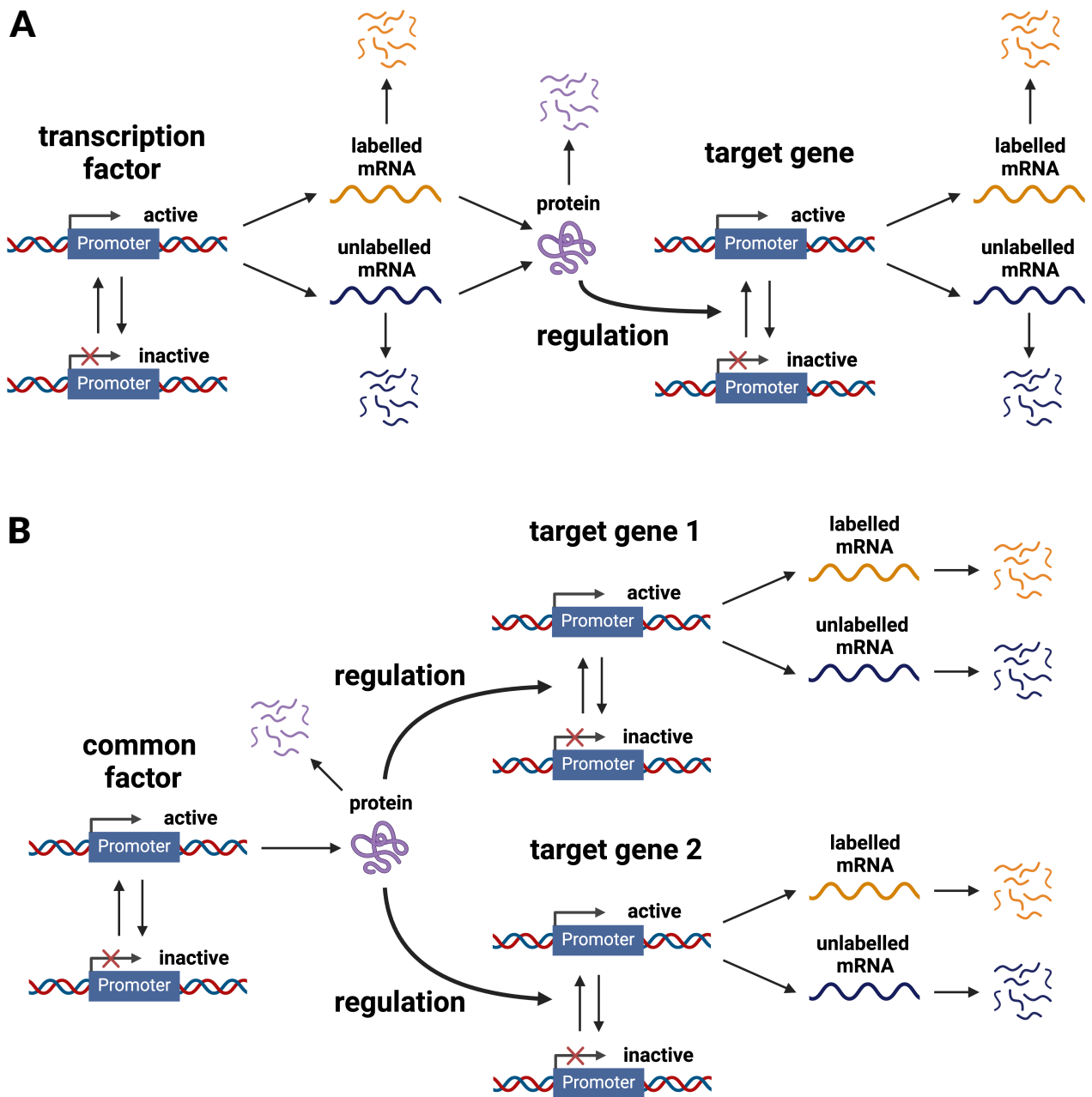


Figure 3.1: **Models of intracellular reaction kinetics of regulatory gene pairs.** (A) Models of direct regulation: A regulator gene (*e.g.* transcription factor) encodes a protein that upregulates or downregulates the expression of a downstream target gene by acting on its promoter. (B) Model of co-regulation: A common regulatory factor simultaneously upregulates the expression of two target genes through protein feedback at their promoters. The co-regulation model also broadly represents a shared upstream noise source influencing both genes.

Independent gene model. For completeness, in our analysis we also consider gene pairs that are independent, in the sense that they have no regulatory interaction. In this case, we model each gene's dynamics separately using the gene-specific model (2.5) of Chapter 2 and we will refer to this case as the *independent model*. As we will show in the next paragraph, characterising two genes as independent does not imply they are uncorrelated, as they can be influenced by extrinsic biological and technical noise sources.

Cell cycle-dependent and cell size-dependent kinetic rates

To capture extrinsic noise effects induced by the cell cycle and cell size, we allow for explicit dependence of kinetic rates on cell cycle progression, as in Chapter 2. The details are explained in 2.2.2. We here note that the scaling of both mRNA and protein expression with cell size is mediated by a cell size-dependent mRNA synthesis rate α , while the protein synthesis rate k_p stays constant.

Moment equations with time-dependent kinetic rates

We proceed with modelling the stochastic dynamics of the above gene networks using the chemical master equation and its associated moment equations. As all three networks involve the same type of unimolecular and bimolecular reactions, we will only formulate equations for the 3-gene network case (3.3), and the formulation for the simpler 2-gene network cases (3.1),(3.2) follows similarly. The network (3.3) consists of $R = 16$ reactions and $n = 8$ species $(G_0, P, G_1, U_1, L_1, G_2, U_2, L_2)$, where G_i denotes the active state of the i -th gene's promoter, and the inactive states are neglected as they have complementary stoichiometry to the active states. The time-dependent joint probability distribution $P(\mathbf{x}, t)$ of the species' abundances $\mathbf{x} = (x_1, \dots, x_n)^T$ obeys the CME:

$$\frac{d}{dt}P(\mathbf{x}, t) = \sum_{r=1}^R \left(w_r(\mathbf{x} - \mathbf{S}_r, t) P(\mathbf{x} - \mathbf{S}_r, t) - w_r(\mathbf{x}, t) P(\mathbf{x}, t) \right) \quad (3.4)$$

where $\mathbf{S}_r = (\mathbf{S}_{1r}, \dots, \mathbf{S}_{nr})^T$ denotes the r -th column of the stoichiometric matrix of the reaction system, that is:

$$\mathbf{S} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix} \quad (3.5)$$

The propensities $w_r(\mathbf{x}, t)$ of the network are:

$$\begin{aligned} w_1(\mathbf{x}, t) &= k_{\text{on},0}(t)(1 - x_1), w_2(\mathbf{x}, t) = k_{\text{off},0}(t)x_1, w_3(\mathbf{x}, t) = s(t)x_1, w_4(\mathbf{x}, t) = \delta(t)x_2 \\ w_5(\mathbf{x}, t) &= k_{\text{on},1}(t)x_2(1 - x_3), w_6(\mathbf{x}, t) = k_{\text{off},1}(t)x_3, w_7(\mathbf{x}, t) = (1 - \lambda_1)\alpha_1(t)x_3, \\ w_8(\mathbf{x}, t) &= \lambda_1\alpha_1(t)x_3, w_9(\mathbf{x}, t) = \gamma_1(t)x_4, w_{10}(\mathbf{x}, t) = \gamma_1(t)x_5 \\ w_{11}(\mathbf{x}, t) &= k_{\text{on},2}(t)x_2(1 - x_6), w_{12}(\mathbf{x}, t) = k_{\text{off},2}(t)x_6, w_{13}(\mathbf{x}, t) = (1 - \lambda_2)\alpha_2(t)x_6, \\ w_{14}(\mathbf{x}, t) &= \lambda_2\alpha_2(t)x_6, w_{15}(\mathbf{x}, t) = \gamma_1(t)x_7, w_{16}(\mathbf{x}, t) = \gamma_2(t)x_8 \end{aligned} \quad (3.6)$$

We now proceed to derive moment equations of the CME. We note that the propensities $w_5(\mathbf{x}, t), w_{11}(\mathbf{x}, t)$, that correspond to the interaction of the regulatory protein with the target genes, are non-linear. As a result, certain moment equations of 1st and 2nd order depend on higher-order moments and hence the system of moment equations is not closed [67]. In order to derive a closed system of equations we rely on the system size expansion (SSE) of the CME [69, 161] and in particular the leading order of the SSE, that is the linear noise approximation (LNA) [255]. Under the LNA, the 1st order moment equations correspond to the deterministic rate equations for the macroscopic concentrations of the species, while the 2nd order moment equations describe Gaussian fluctuations about the concentrations [99], and are given by:

$$\frac{d}{dt}\mu_i(t) = \sum_{r=1}^R \mathbf{S}_{ir} w_r(\boldsymbol{\mu}, t) \quad (3.7)$$

$$\frac{d}{dt}\Sigma_{ij}(t) = \sum_{\substack{r=1, \dots, R \\ k=1, \dots, n}} \left(\mathbf{S}_{ir} \frac{\partial w_r(\boldsymbol{\mu}, t)}{\partial \mu_k} \Sigma_{kj}(t) + \mathbf{S}_{jr} \frac{\partial w_r(\boldsymbol{\mu}, t)}{\partial \mu_k} \Sigma_{ki}(t) \right) + \sum_{r=1}^R \mathbf{S}_{ir} \mathbf{S}_{jr} w_r(\boldsymbol{\mu}, t) \quad (3.8)$$

where $R = 16, n = 8, \mu_i = \mathbb{E}(x_i)$ and $\Sigma_{ij} = \mathbb{E}((x_i - \mu_i)(x_j - \mu_j))$. In the case of non-linear propensities, $w_r(\cdot)$ denotes the approximated propensity after applying the LNA. Note that this allows us to use the linearity of the expanded propensities to write $\mathbb{E}(w_r(\mathbf{x}, t)) = w_r(\mathbb{E}(\mathbf{x}), t) = w_r(\boldsymbol{\mu}, t)$. Also, since the kinetic rates are cell cycle-dependent, all propensities are periodic functions of time and the moment equations are constrained by periodic boundary conditions, corresponding to cell division events.

Coupling to cell division dynamics

As our models assume cell cycle-dependent kinetic rates, which are periodic functions of time, we define boundary conditions for the moment equations. Similarly to Chapter 2, we use the noise partitioning formulas for the moments that we presented in paragraph 2.2.2. In addition, here we assume that the stochastic partitioning of protein molecules at cell division follows the same probabilistic rules as for the mRNA molecules.

3.2.2 Decomposition of gene-gene transcriptional covariation

In this section, we present a noise decomposition framework to analyse biological and technical contributions to the overall transcriptional covariation of gene pairs. Let $y_1 = u_1 + l_1$, $y_2 = u_2 + l_2$ be the observed mRNA levels of two genes G_1, G_2 , that includes the unlabelled (u) and labelled (l) modalities, as given by the scEU-seq protocol [155]. Also, from the data, we can obtain cell cycle-dependent and labelling time-dependent measurements of y_1 and y_2 , as shown in paragraph 2.2.1 of Chapter 2.

Gene covariation decomposition in the absence of technical noise

The total transcript covariation between the genes G_1, G_2 is given by

$$\text{Cov}(y_1, y_2) = \text{Cov}(u_1, u_2) + \text{Cov}(u_1, l_2) + \text{Cov}(l_1, u_2) + \text{Cov}(l_1, l_2) \quad (3.9)$$

In the absence of technical noise effects, each one of the terms in equation (3.9) can be decomposed as:

$$\text{Cov}(x_1, x_2) = \underbrace{\mathbb{E}_\tau(\text{Cov}(x_1, x_2 \mid \tau))}_{\text{gene interaction}} + \underbrace{\text{Cov}_\tau(\mathbb{E}(x_1 \mid \tau), \mathbb{E}(x_2 \mid \tau))}_{\text{cell cycle-induced covariation}} \quad (3.10)$$

where x denotes either unlabelled (u) or labelled (l) mRNA. The first term explains intrinsic covariance arising from gene interaction, while the second term explains extrinsic covariance arising from cell cycle dependence. This suggests that, under the assumptions of deterministic cell cycle length and negligible technical noise, conditioning on cell cycle stage can be informative of gene interaction.

These covariances can also be normalised to obtain gene-gene correlations across the four individual unlabelled-labelled modalities:

$$\rho(x_1, x_2) = \frac{\text{Cov}(x_1, x_2)}{\sqrt{\text{Var}(x_1) \text{Var}(x_2)}} \quad (3.11)$$

where $x \in \{u, l\}$ and $\text{Var}(x_1), \text{Var}(x_2)$ are gene-specific variances. Alternatively, we can compute the total gene-gene correlation:

$$\rho(y_1, y_2) = \frac{\text{Cov}(y_1, y_2)}{\sqrt{\text{Var}(y_1) \text{Var}(y_2)}} = \sum_{\substack{x_i \in \{u_i, l_i\} \\ i=1,2}} \frac{\text{Cov}(x_1, x_2)}{\sqrt{\text{Var}(y_1) \text{Var}(y_2)}} \quad (3.12)$$

where $\text{Var}(y_1), \text{Var}(y_2)$ are the total transcript variances of the two genes. This gene-specific total transcript variance is decomposed as

$$\text{Var}(y) = \text{Var}(u) + \text{Var}(l) + 2 \text{Cov}(u, l) \quad (3.13)$$

and for each modality $x \in \{u, l\}$, we have that:

$$\text{Var}(x) = \underbrace{\mathbb{E}_\tau(\text{Var}(x \mid \tau))}_{\text{intrinsic variation}} + \underbrace{\text{Var}_\tau(\mathbb{E}(x \mid \tau))}_{\text{cell cycle variation}} \quad (3.14)$$

Also, the $u - l$ covariance is decomposed as:

$$\text{Cov}(u, l) = \mathbb{E}_\tau(\text{Cov}(u, l \mid \tau)) + \text{Cov}_\tau(\mathbb{E}(u \mid \tau), \mathbb{E}(l \mid \tau)) \quad (3.15)$$

All the formulas presented above can be used to decompose the total covariance or the to-

tal correlation coefficient of genes G_1 and G_2 , into individual components, in the absence of technical noise.

Gene covariation decomposition in the presence of technical noise

We now want to extend the above formulas to include technical effects due to transcript downsampling of sequencing experiments. To do that, the different components of the total covariance need to be expressed as a function of the true mRNA levels $\tilde{x}_1, \tilde{x}_2$ of genes G_1, G_2 and the cell-specific transcript capture efficiency β . As explained in paragraph 2.2.3 of Chapter 2, we assume downsampling to be a Binomial process, meaning that for each gene we observe a Binomial sample of transcripts x out of the true total transcripts $\tilde{x}$ with probability β , *i.e.* $x \sim \text{Binomial}(\tilde{x}, \beta)$. We show that the total transcript covariance of genes G_1, G_2 , including technical effects, is decomposed as:

$$\text{Cov}(y_1, y_2) = \sum_{\substack{x_i \in \{u_i, l_i\} \\ i=1,2}} \text{Cov}(x_1, x_2) \quad (3.16)$$

where

$$\begin{aligned} \text{Cov}(x_1, x_2) = \mathbb{E}_\tau \left(\underbrace{\mathbb{E}(\beta | \tau)^2 \text{Cov}(\tilde{x}_1, \tilde{x}_2 | \tau)}_{\text{gene interaction}} + \underbrace{\text{Var}(\beta | \tau) \mathbb{E}(\tilde{x}_1 | \tau) \mathbb{E}(\tilde{x}_2 | \tau)}_{\text{capture efficiency variation}} + \underbrace{\text{Var}(\beta | \tau) \text{Cov}(\tilde{x}_1, \tilde{x}_2 | \tau)}_{\text{mixed effects}} \right) \\ + \underbrace{\text{Cov}_\tau \left(\mathbb{E}(\beta | \tau) \mathbb{E}(\tilde{x}_1 | \tau), \mathbb{E}(\beta | \tau) \mathbb{E}(\tilde{x}_2 | \tau) \right)}_{\text{cell cycle-induced covariation}} \end{aligned} \quad (3.17)$$

In the above equation, we have two main components, where the first summarises cell cycle-independent contributions to the total covariance: Covariation arising from gene interaction, covariation due to cell-to-cell variability in capture efficiency and covariation attributed to both capture efficiency and gene interaction. The second component explains covariation due to cell cycle dependence.

Interestingly, in the special case where two genes are *independent*, in the sense that there is no

gene interaction, equation (3.17) reduces to:

$$\text{Cov}(x_1, x_2) = \underbrace{\mathbb{E}_\tau \left(\text{Var}(\beta \mid \tau) \mathbb{E}(\tilde{x}_1 \mid \tau) \mathbb{E}(\tilde{x}_2 \mid \tau) \right)}_{\text{capture efficiency variation}} + \underbrace{\text{Cov}_\tau \left(\mathbb{E}(\beta \mid \tau) \mathbb{E}(\tilde{x}_1 \mid \tau), \mathbb{E}(\beta \mid \tau) \mathbb{E}(\tilde{x}_2 \mid \tau) \right)}_{\text{cell cycle-induced covariation}} \quad (3.18)$$

This shows that even though we account for extrinsic cell cycle-dependent effects, we have additional covariation arising purely from technical noise. This poses a significant challenge in distinguishing whether the observed covariance arises from technical variability or potential gene interaction.

To prove the above, we express the cell cycle-independent covariance in equation (3.16) as follows:

$$\mathbb{E}_\tau(\text{Cov}(x_1, x_2 \mid \tau)) = \mathbb{E}_\tau \left(\underbrace{\mathbb{E}(\text{Cov}(x_1, x_2 \mid \tilde{x}_1, \tilde{x}_2, \beta, \tau))}_{=0} + \text{Cov}(\mathbb{E}(x_1 \mid \tilde{x}_1, \beta, \tau), \mathbb{E}(x_2 \mid \tilde{x}_2, \beta, \tau)) \right) \quad (3.19a)$$

$$= \mathbb{E}_\tau \left(\text{Cov}(\tilde{x}_1 \beta, \tilde{x}_2 \beta \mid \tau) \right) \quad (3.19b)$$

$$= \mathbb{E}_\tau \left(\mathbb{E}(\beta^2 \mid \tau) \mathbb{E}(\tilde{x}_1 \tilde{x}_2 \mid \tau) - \mathbb{E}(\beta \mid \tau)^2 \mathbb{E}(\tilde{x}_1 \mid \tau) \mathbb{E}(\tilde{x}_2 \mid \tau) \right) \quad (3.19c)$$

$$= \mathbb{E}_\tau \left(\mathbb{E}(\beta \mid \tau)^2 \text{Cov}(\tilde{x}_1, \tilde{x}_2 \mid \tau) + \text{Var}(\beta \mid \tau) \mathbb{E}(\tilde{x}_1 \tilde{x}_2 \mid \tau) \right) \quad (3.19d)$$

$$= \mathbb{E}_\tau \left(\mathbb{E}(\beta \mid \tau)^2 \text{Cov}(\tilde{x}_1, \tilde{x}_2 \mid \tau) + \text{Var}(\beta \mid \tau) \mathbb{E}(\tilde{x}_1 \mid \tau) \mathbb{E}(\tilde{x}_2 \mid \tau) + \text{Var}(\beta \mid \tau) \text{Cov}(\tilde{x}_1, \tilde{x}_2 \mid \tau) \right) \quad (3.19e)$$

Similarly, the cell cycle-dependent covariance of equation (3.16) is:

$$\text{Cov}_\tau(\mathbb{E}(x_1 \mid \tau), \mathbb{E}(x_2 \mid \tau)) = \text{Cov}_\tau \left(\mathbb{E}(\beta \mid \tau) \mathbb{E}(\tilde{x}_1 \mid \tau), \mathbb{E}(\beta \mid \tau) \mathbb{E}(\tilde{x}_2 \mid \tau) \right) \quad (3.20)$$

This concludes our gene-gene covariation decomposition. In our downstream analyses, we will apply this decomposition to quantify covariances (3.17) between unlabelled and labelled mRNA of the two genes, across labelling conditions.

3.2.3 Time-dependent summary statistics of gene co-expression

Our models allow us to define time-dependent summary statistics describing co-expression dynamics of gene pairs to characterise the different gene regulatory models presented in 3.2.1. These statistics capture information about the mean and noise of gene expression, including both gene-specific and pair-wise gene dynamics.

To generate summary statistics from the models, we numerically solve the moment ODEs (3.7) following the same procedure as for the gene-specific transcription models of Chapter 2. As described in paragraph 2.2.2, we simulate the models across consecutive cell divisions until convergence to steady state, capturing both cell cycle-dependent dynamics and the time-dependent effects of pulse-chase labelling. These model realisations yield cell cycle- and labelling time-dependent outputs, that we utilise to compute summary statistics. Prior to computing statistics, we update the model outputs to account for binomial downsampling effects. The downsampling of moments follows the probabilistic rules that we presented in paragraph 2.2.3 of Chapter 2 and are summarised in Table 3.1. The formula for downsampling the covariance $\text{Cov}(x, y)$ applies to both the gene-specific covariance between modalities, *i.e.* $\text{Cov}(u, l)$, as well as the cross-gene covariances, *i.e.* $\text{Cov}(u_1, u_2), \text{Cov}(u_1, l_2)$, etc.

Table 3.1: Effect of binomial downsampling on 1st and 2nd order moments.

	Exact moment	Downsampled moment
(1)	$\mathbb{E}(x)$	$\mathbb{E}(\beta) \mathbb{E}(\tilde{x})$
(2)	$\text{Var}(x)$	$\mathbb{E}(\beta(1 - \beta)) \mathbb{E}(x) + \text{Var}(\beta) \mathbb{E}(x^2) + \mathbb{E}(\beta)^2 \text{Var}(x)$
(3)	$\text{Cov}(x, y)$	$\mathbb{E}(\beta)^2 \text{Cov}(x, y) + \text{Var}(\beta) \mathbb{E}(x)\mathbb{E}(y) + \text{Var}(\beta) \text{Cov}(x, y)$

We therefore generate labelling time-dependent statistics that integrate the dependence on the cell cycle and also capture technical noise effects. These statistics include gene-specific mean

unlabelled and labelled transcripts, $\mathbb{E}(u_1 | t), \mathbb{E}(l_1 | t), \mathbb{E}(u_2 | t), \mathbb{E}(l_2 | t)$, gene-specific variance-covariance of unlabelled and labelled transcripts, $\text{Var}(u_1 | t), \text{Var}(l_1 | t), \text{Cov}(u_1, l_1 | t), \text{Var}(u_2 | t), \text{Var}(l_2 | t), \text{Cov}(u_2, l_2 | t)$, and gene-gene cross-correlations between unlabelled and labelled transcripts, $\rho(u_1, u_2 | t), \rho(u_1, l_2 | t), \rho(l_1, u_2 | t), \rho(l_1, l_2 | t)$. To compute gene-gene correlations, we normalise the covariance of equation (3.17), as we show in (3.11).

3.3 Synergising machine learning and mechanistic modelling

In the previous section, we defined a stochastic framework to model transcription dynamics of gene pairs driven by various modes of regulatory interaction. By conditioning on cell cycle stage information, we distinguished between intrinsic and extrinsic components of transcriptional co-variation and demonstrated how these components are affected by technical noise. Additionally, we defined summary statistics that characterise gene-pair transcription dynamics. Our goal is to develop a method that leverages this information to classify gene pairs with respect to four hypotheses: *independence*, *positive regulation*, *negative regulation* and *co-regulation*. To achieve this, we employ a machine learning approach to address this model selection problem as a classification task. In the rest of the chapter, we will show how combining mechanistic modelling with a supervised learning approach allows to classify modes of gene regulation in an interpretable manner. In this section, we outline the classification task, and describe our methodology for integrating our stochastic models with a deep learning model for classification.

3.3.1 Generative-discriminative modelling for classification

The main idea of our approach is to use the mechanistic models to generate synthetic data, which can be then utilised as a training dataset for discriminative classification across the different models [254]. The aim of the classification is to discover distinctive features associated with modes of regulation. An important challenge is that many sets of gene co-expression statistics could arise from several candidate generative mechanisms. As we will show in the next section, the different models of gene regulation can give rise to similar statistics. Nonetheless,

they also exhibit key differences that can be used as signatures to identify them. Discovering such signatures necessitates an exhaustive search over many different model outputs.

We here outline the details of the classification problem. Assume we are given a set of observations $\mathbf{x} = (x_{ij})$ with $i = 1, \dots, n$ samples, which are outputs from different generative models of gene regulation $g(\mathbf{x} \mid c, \theta)$, where c is the particular model class and θ are the model parameters. Each model output $x_{i\cdot}$ is a vector of $j = 1, \dots, d$ features, comprising summary statistics of gene co-expression. The observations are assigned to ground-truth labels $\mathbf{y} = (y_i)$, that encode the underlying generative model class c . The classification task is to learn a (potentially non-linear) mapping $\mathbf{y} = f(\mathbf{x}; \theta)$ that assigns synthetic data $\mathbf{x}$ (gene-pair summary statistics) to modes of regulation $\mathbf{y}$, or more generally to learn the probability distribution over model classes $p(y = c \mid \mathbf{x}; \theta)$, given synthetic data $\mathbf{x}$.

Our approach is to break the classification problem into two decision steps. The first step is to classify a gene pair into the independent class or the regulation class, where in the latter it is hypothesised that there is a gene regulatory relationship that cannot be explained by the independent models. Therefore, for the first decision step, we solve a simpler binary classification task. In the second step, given that the first decision outcome is regulation, the gene pair is classified into a particular regulatory class: positive regulation, negative regulation or co-regulation. Therefore, for the second decision step, we solve a multi-class classification task.

In principle, these classification tasks can be approached using standard machine learning algorithms, such as deterministic classifiers (e.g. support vector machines [256]) or probabilistic classifiers (e.g. random forests [257]). Alternatively, deep learning models such as feed-forward and convolutional neural networks [258], are popular choices due to their capacity of approximating arbitrary non-linear functions [259, 260]. For our purposes, we are utilising feed-forward neural networks, aiming to learn complex non-linear relationships between different summary statistics, and map these relationships to potential generative models. In the rest of the chapter, we will use the symbols $\mathcal{C}_1, \mathcal{C}_2$ to refer to the first and the second classifier, respectively, that are trained for the two classification tasks.

3.3.2 Synthetic data generation

To train our classifiers, we generate a synthetic dataset $\{\mathbf{x}, \mathbf{y}\}$, where $\mathbf{x} = \{x_{ij} : i = 1, \dots, n, j = 1, \dots, d\}$ comprises summary statistics generated from the 4 stochastic models of gene interaction, and $\mathbf{y} = \{y_i : i = 1, \dots, N\}$ are the ground-truth class labels of the generative models, with $y_i \in \{0 : \text{independent}, 1 : \text{positive regulation}, 2 : \text{negative regulation}, 3 : \text{co-regulation}\}$. We generate in total $n = 4 \cdot 10^4$ training samples from each model, and each sample involves $d = 88$ features - summary statistics, that include gene-specific mean expression and gene-gene correlations across labelling timepoints, as described in paragraph 3.2.3.

To generate data, we use prior parameter sets that we sample from the gene-specific posterior distributions, obtained from our inference in Chapter 2. In particular, we focus on the set of constant scaling genes (Figure 2.10 (A ii)), from which we randomly sample 10^4 gene pairs. For each gene pair, we sample 2 parameter sets from the individual gene posterior distributions, giving us 4 pairs of parameter sets. Moreover, we sample additional latent parameters describing protein kinetics, required to simulate the 3 regulation models. For each simulated gene pair, we sample protein synthesis and degradation rates from log-normal densities that we fitted on empirical distributions reported in [131]. Additionally, for the co-regulation model (3.3), we sampled additional latent parameters for the gene promoter of the common regulator, from log-uniform distributions: $\log_{10}(k_{\text{on}}) \sim \mathcal{U}(-3, 3)$, $\log_{10}(k_{\text{off}}) \sim \mathcal{U}(-3, 3)$. These procedure gave us in total $n = 4 \cdot 10^4$ prior parameter sets that we used to produce training examples from each generative model.

A critical step in the synthetic data generation is to ensure that, for each sampled parameter set, all regulatory models produce identical gene-specific mean expression dynamics. This is essential for eliminating biases associated with mean expression differences that could arise from the different modes of regulation: for instance, the positive regulation model could give rise to higher expression levels of the target gene, compared to the negative regulation model. By standardising mean expression levels, we ensure that the classifier uses gene-specific mean expression solely as background information, and relies on gene correlations to differentiate between regulatory models. To achieve this, we adjust the parameters of the target genes in each regulation model, to match the effective transcription of the independent gene model. As an example, if $\alpha_2^{\text{eff}} = \frac{k_{\text{on},2}}{k_{\text{on},2} + k_{\text{off},2}} \alpha_2$ is the effective transcription rate under the independent model, then the corresponding effective transcription rate under the positive regulation model

is given by

$$\tilde{\alpha}_2^{\text{eff}} = \frac{p_{ss} k_{\text{on},2} + k_{\text{off},2}}{p_{ss} k_{\text{on},2}} \alpha_2^{\text{eff}} \quad (3.21)$$

where p_{ss} is the steady-state mean protein level, that is estimated as:

$$p_{ss} = \frac{k_{\text{on},1}}{k_{\text{on},1} + k_{\text{off},1}} \frac{\alpha_1}{\gamma_1 + \gamma_{\text{dil}}} \frac{k_p}{\delta + \gamma_{\text{dil}}} \quad (3.22)$$

where $\gamma_{\text{dil}} = \frac{\ln(2)}{T_d}$ is the dilution rate due to cell growth [180], and $T_d = 20$ (hours) is the cell cycle duration. Recall that we are only simulating constant size-scaling genes, and thus the protein steady state does not depend on cell cycle variation of kinetic rates.

We recall that the classifier $\mathcal{C}_1$ is trained to solve a binary classification task between *independence* and *regulation*, while the classifier $\mathcal{C}_2$ is trained to solve a multi-class classification task between *positive regulation*, *negative regulation* and *co-regulation*. To train the classifier $\mathcal{C}_1$, we use $3 \cdot 10^4$ samples from the independent model and 10^4 samples from each of the 3 regulation models, to create a balanced training between independence and regulation, with $n_1 = 6 \cdot 10^4$ training examples in total. The rest of the samples are used as our test dataset to evaluate the out-of-sample performance of our classifier. For the classifier $\mathcal{C}_2$, we use $3 \cdot 10^4$ samples from each of the 3 regulation models to get $n_2 = 9 \cdot 10^4$ training examples in total, while the rest of the samples are used as test data.

3.3.3 Neural network classifiers

We now present the deep learning model that we utilise for classification. The type of classifier we use for both classification tasks is a feed-forward neural network, or multi-layer perceptron (MLP). Each classifier receives sets of summary statistics of gene co-expression as input, and is trained to map these statistics to underlying modes of regulation (Figure 3.2). The architecture

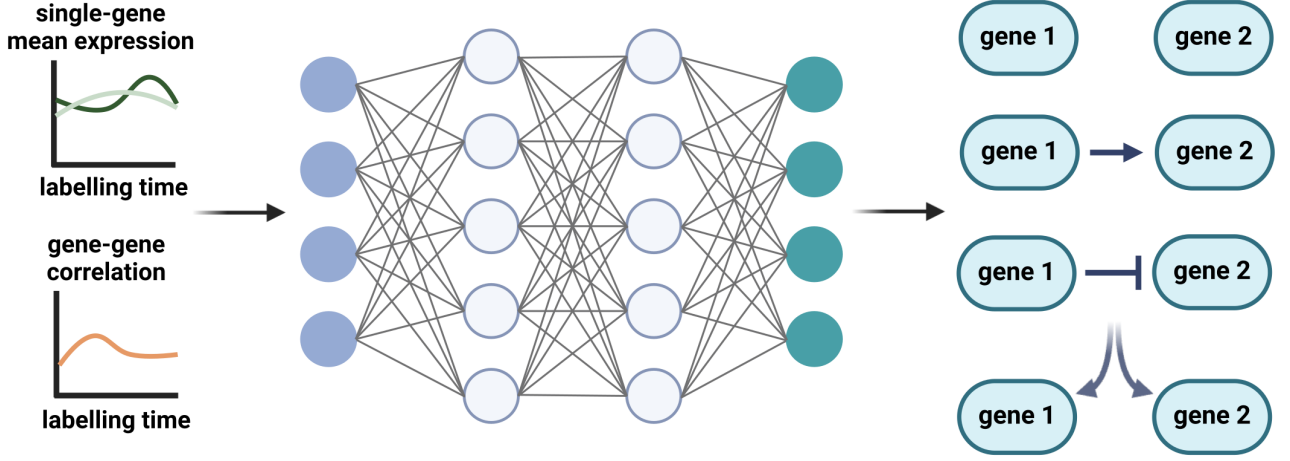


Figure 3.2: **A neural classifier of gene regulation models.** Cartoon illustration of our model classification task: A feed-forward neural network is trained with summary statistics generated from the four models of gene interaction (3.2.1) to learn the underlying model class: *independence, positive regulation, negative regulation or co-regulation*.

of the neural network is defined as:

$$f(\mathbf{x}; \mathbf{w}, \mathbf{b}) = \sigma \left(\mathbf{b}^{(L)} + \mathbf{w}^{(L)} \phi_{L-1}(\mathbf{b}^{(L-1)} + \mathbf{w}^{(L-1)} \phi_{L-2}(\cdots \phi_1(\mathbf{b}^{(1)} + \mathbf{w}^{(1)} \mathbf{x}) \cdots) \right) \quad (3.23a)$$

$$y_k = \sigma \left(b_k^{(L)} + \sum_{j=1}^{N_{L-1}} w_{jk}^{(L)} h_j^{(L-1)} \right), \quad k = 1, \dots, n_c \quad (3.23b)$$

$$h_k^{(l)} = \phi_l \left(b_k^{(l)} + \sum_{j=1}^{N_{l-1}} w_{jk}^{(l)} h_j^{(l-1)} \right), \quad k = 1, \dots, N_l, \quad l = 2, \dots, L-1 \quad (3.23c)$$

$$h_k^{(1)} = \phi_1 \left(b_k^{(1)} + \sum_{j=1}^{N_0} w_{jk}^{(1)} x_j \right), \quad k = 1, \dots, N_1 \quad (3.23d)$$

where the definitions of all the above symbols are summarised in Table 3.2. The softmax function $\sigma(u_i) = \frac{e^{u_i}}{\sum_{k=1}^{n_c} e^{u_k}} \in (0, 1)$ is applied on the output $\mathbf{h}^{(L)} = \{h_k^{(L)} : k = 1, \dots, N_L = n_c\}$ of the L -th layer to transform it into a probability vector $\mathbf{y} = \{y_k : k = 1, \dots, n_c\}$ across the n_c classes.

Neural network training

The neural networks are trained using synthetic data that are generated as described in paragraph 3.3.2. Here, we specify the details of the training procedure and our choices for the neural networks' hyperparameters.

First, to evaluate the distance between the neural network outputs and the ground-truth labels

Table 3.2: **Definition of feed-forward neural network.**

	Symbol	Definition
(1)	$\mathbf{x} = \{x_k : k = 1, \dots, d\}$	input data point with d features
(2)	$\mathbf{y} = \{y_k : k = 1, \dots, n_c\}$	output of neural network: probability distribution across n_c classes
(3)	$h_k^{(l)}$	k -th neuron (computational unit) in the l -th hidden layer
(4)	N_l	number of neurons in the l -th layer ($N_0 = d, N_L = n_c$)
(5)	$\mathbf{w}^{(l)} = \{w_{jk}^{(l)}\}$	weights on the edges connecting the $(l-1)$ -th with the l -th layer
(6)	$\mathbf{b}^{(l)} = \{b_k^{(l)}\}$	biases on the edges connecting the $(l-1)$ -th with the l -th layer
(7)	ϕ_l	non-linear activation function acting point-wise on the l -th layer
(8)	$\sigma : \mathbb{R}^{n_c} \rightarrow (0, 1)^{n_c}$	softmax function acting on the L -th (output) layer

during training, we use the cross-entropy loss function $\mathcal{L}(\mathbf{y}, \tilde{\mathbf{y}}) = -\sum_{j=1}^{n_c} \tilde{y}_j \log(y_j)$, where y_j is the probability of class $j \in 1, \dots, n_c$ as predicted by the neural network and $\tilde{y}_j = e_j$ is a one-hot vector encoding of the ground-truth class j , *e.g.* $\tilde{y}_2 = e_2 = (0, 1, 0)$ is the one-hot encoding of the negative regulation class in the second classification task.

Our choice of hyperparameters for the classifiers are as follows. The classifier $\mathcal{C}_1$ has $L = 5$ hidden layers with decreasing widths, proportional to the input size: $N_1 = 704, N_2 = 528, N_3 = 352, N_4 = 176, N_5 = 88$, resulting in $L + 1 = 6$ layers in total, including the input layer. The

classifier $\mathcal{C}_2$ has $L = 4$ hidden layers: $N_1 = 880, N_2 = 528, N_3 = 264, N_4 = 44$, resulting in $L + 1 = 5$ layers. These architecture choices were made after multiple training sessions using various combinations of network depth and layer width. For both classifiers, we used the ReLu function $h(u) = \max(0, u)$ [258] as a point-wise non-linear activation on each layer. The neural network weights were initialised using the Glorot uniform initialisation [261] and optimised using the AdaMax optimiser (a variant of the Adam algorithm based on the infinity norm [262]), with learning rate $\eta = 0.001$. For the final training, we trained both classifiers over a maximum of 150 epochs, using batch sizes of $b = 1024$ samples per epoch. We imposed an early-stopping criterion so that the training is terminated if the test loss does not improve after 30 epochs.

Evaluating the neural network performance

To evaluate the performance of classifiers $\mathcal{C}_1$ and $\mathcal{C}_2$, we utilise standard metrics for classification algorithms. During training, we compute the mean accuracy across train and test examples, by comparing the true class labels with the predicted class, that is the $\operatorname{argmax}_{j=1, \dots, n_c}(y_j)$ of the output class probabilities y_j . We evaluate the classifier's performance on the test set after training, using the confusion matrix CM, with elements

$$\text{CM}_{ij} = \frac{\#\{\text{predicted class} = j\} \cap \{\text{true class} = i\}}{\#\{\text{true class} = i\}} \quad (3.24)$$

i.e. the relative frequency of predicting class j , given true class i . Its diagonal elements CM_{ii} correspond to the *recall*, or true positive rate of class i . We also quantify the *precision* of predicting class i to be

$$\text{precision}_i = \frac{\#\{\text{predicted class} = i\} \cap \{\text{true class} = i\}}{\sum_{j=1, \dots, n_c} \#\{\text{predicted class} = i\} \cap \{\text{true class} = j\}} \quad (3.25)$$

The precision, or alternatively the *false discovery rate*, $\text{FDR} = 1 - \text{precision}$, are common metrics for controlling the false positives in classification problems and hypothesis testing [263].

To account for uncertainty in the classifier's outputs, we compute these metrics as functions of a varying *decision threshold* $p_{\min}$, that is the minimum predicted class probability required for a decision outcome. Given a test sample, we get an uncertain outcome if all classifier outputs y_j , $j = 1, \dots, n_c$, are below $p_{\min}$. The total uncertainty rate is the relative frequency of

uncertain outcomes across all test samples. For each decision threshold $p_{\min}$, we compute the confusion matrix, recall and precision across all test samples for which we get a certain outcome (*i.e.* the threshold $p_{\min}$ is satisfied). As we will show in the next section, we set $p_{\min} = 0.6$ for classifier $\mathcal{C}_1$ and $p_{\min} = 0.5$ for classifier $\mathcal{C}_2$.

Applying the neural network models to real data

After training on synthetic data, we apply our classification model to real scEU-seq data. We randomly sample 10^5 gene pairs from the constant scaling class, inferred in Chapter 2, and compute gene-specific mean unlabelled and labelled transcript levels and gene-gene correlations between unlabelled and labelled transcripts, across cells in each metabolic labelling condition. To test for both potential directions of regulation, we consider both gene orders for each pair, resulting in a test dataset of $2 \cdot 10^5$ samples. Additionally, we compute the summary statistics of each gene pair for 100 bootstrap samples of the data, to account for sampling noise of estimators.

We impose a strict condition for gene-pair classification, that accounts for both sampling noise in the data and uncertainty in the classifiers' output. Specifically, for a given gene pair, we count the number of bootstrap samples $n_{\text{dec}}(i)$ for which the classifier's outputs satisfy the decision threshold $p_{\min}$, for each class i . We classify the gene pair in class i , if $n_{\text{dec}}(i) \geq n_{\min}$, where $n_{\min} = 80$ for classifier $\mathcal{C}_1$ and $n_{\min} = 60$ for classifier $\mathcal{C}_2$. We choose a lower bootstrap threshold for the second classification, as classifier $\mathcal{C}_2$ is only presented with gene pairs that are already assigned to the regulation class by classifier $\mathcal{C}_1$.

3.4 Applications: Identifying modes of gene regulatory interaction

3.4.1 Independent gene models produce gene-gene correlations comparable to data

As an initial exploratory step, we asked whether observed gene-gene correlations in scEU-seq data [155] can be explained by individual gene expression models. To test this, we compute Pearson correlation coefficients for 750 randomly sampled gene pairs in the data and compare them with model estimates, using the inferred gene-specific kinetic parameters of Chapter 2. The model-based correlations are derived using the formulas presented in 3.2.2, which include the total correlation coefficient (3.12) and the contributions from unlabelled and labelled transcript levels, to the total correlation (equations (3.16),(3.17)). These formulas account for technical noise due to cell-to-cell capture efficiency variation, making the model estimates directly comparable to the observed data correlations. The sampled gene pairs include genes that from the same class (*e.g.* constant scaling, varying burst frequency etc., Figure 2.10) as well as from different classes.

Interestingly, we find strong overall agreement in the total correlation coefficient between model predictions and data across gene pairs (Spearman’s correlation $\rho = 0.82$, Figure 3.3 (A)). The model similarly predicts the correlations between unlabelled and labelled mRNA modalities accurately (Spearman’s correlations from $\rho = 0.67$ to $\rho = 0.74$, Figures 3.3 (B-E)). This predictive accuracy is observed regardless of the model classes to which the gene pairs belong. Additionally, we observe that correlations between cell cycle-regulated genes are notably higher compared to the majority of gene pairs, and are equally well predicted by the models (Figure 3.3 (A)). These findings suggest that, for many gene pairs, pairwise correlations can be largely explained by cell-to-cell variation in capture efficiency and cell cycle-induced covariation, the only two sources of covariance captured by the independent gene models (equation (3.18)).

Nonetheless, we observe discrepancies between observed and predicted correlations for several gene pairs, that are particularly large for certain components of the correlation (Figures 3.3 (C) and (E)). These discrepancies might suggest the presence of additional sources of covariation stemming from either gene interaction or other sources of cellular noise, that are not captured by

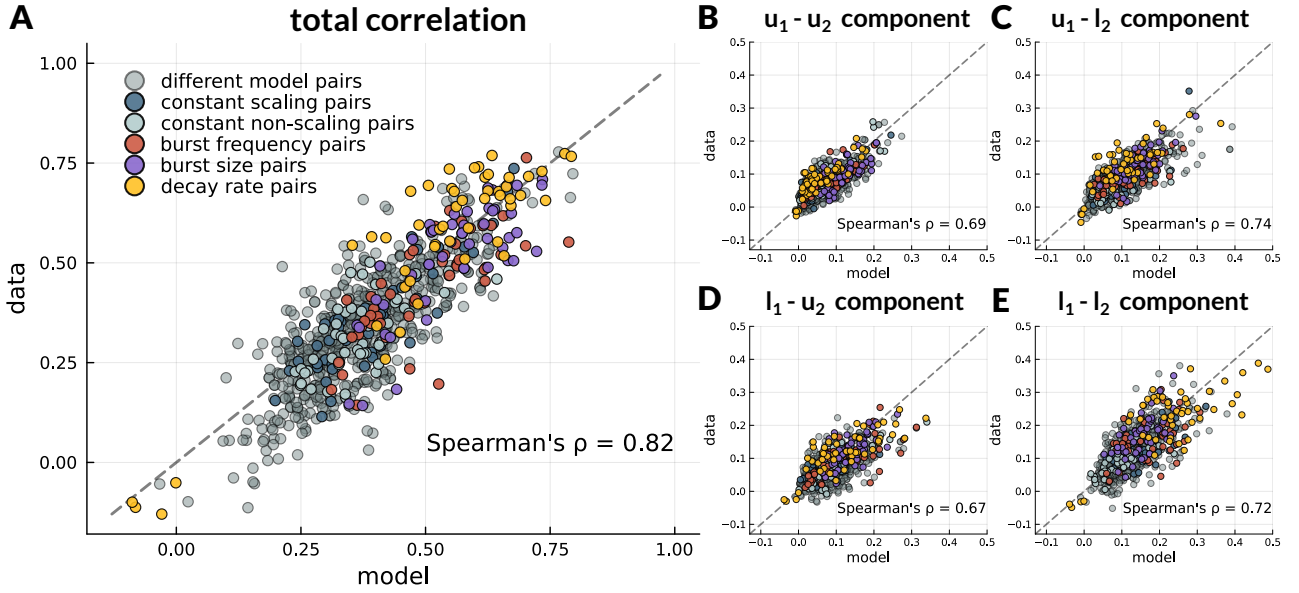


Figure 3.3: **Gene-gene correlations of independent gene model pairs.** Comparison of model-based correlations and observed correlations across gene pairs explained by different kinetic regulation models (grey points) and gene pairs explained by the same model (coloured points). All correlations are computed across cells treated with the *Chase 0* labelling condition (see paragraph 2.2.1). (A) Shows total correlation coefficients (equation (3.12)) and (B-E) show distinct components of the correlation coefficient corresponding to different pairs of unlabelled-labelled mRNA measurements (equation (3.17)).

the independent gene models. To further explore the disagreement between model and data, we scored the gene pairs based on the deviation of the predicted from the observed total correlation. Using bootstrap sampling, we compute confidence intervals for each correlation coefficient to account for sampling noise in the data, allowing us to detect statistically significant deviations. We find that approximately 70% of the predicted correlations lie within the confidence intervals of the data. This indicates that 30% of gene-gene correlations may arise from additional sources of covariation, including regulatory interactions, not explained by the independent models. Among the correlations that the model fails to predict, we detect gene pairs with known regulatory links reported in BioGRID, a genetic interaction database [264]. As an example, we find that *PPIB*, a gene involved in protein folding, cell proliferation and immune response, interacts with *CDK2*, a cell cycle regulator gene responsible for the G1/S phase transition. We also find that *CDK2* interacts with *MCM2*, a gene involved in DNA replication initiation, which in turn interacts with *BRCA1*, a tumour suppressor gene with a role in maintaining genome stability.

3.4.2 Mechanistic modelling and machine learning distinguish regulation from independent models

These observations motivate us to explore gene-gene correlations in a systematic way. Our primary goal is to address two key questions: First, given a set of gene co-expression statistics, can we discern whether a gene pair is independent or has a potential regulatory relationship? Second, if a regulatory relationship exists, can we distinguish whether there is direct gene interaction or if the two genes are co-regulated by a shared upstream factor? To tackle these questions, we develop a framework that integrates mechanistic models of gene regulation with machine learning, aiming to uncover connections between gene-gene correlations and underlying modes of regulatory interaction.

Our mechanistic models describe gene-pair transcription dynamics, driven by positive regulation, negative regulation and co-regulation (paragraph 3.2.1 and Figure 3.1), with cell cycle- and cell size-dependent kinetic rates. The models are numerically solved in steady state using moment equations, that additionally capture the temporal dependence on metabolic labelling and the effects of stochastic partitioning at cell division (paragraph 3.2.1). The model outputs, consisting of cell cycle- and labelling time-dependent moments of gene co-expression, are updated to correct for technical noise, and utilised to compute time-dependent summary statistics (paragraph 3.2.3 and Table 3.1). This framework allows us to generate synthetic data (as described in 3.3.2), comprising summary statistics of gene co-expression, that is leveraged as a training dataset for discriminative classification (Figure 3.2). Therefore, we first train and test the performance of the classifier on the synthetic data, before we proceed with applying our framework to scEU-seq data.

After training the binary classifier $\mathcal{C}_1$ (as described in 3.3.3), we find that gene-gene correlations can effectively distinguish models of gene (co-)regulation from independent gene models in synthetic data. We first verify that gene-specific mean expression alone is insufficient for this task, as training using only mean statistics results in no improvement in the mean cross entropy loss (Figure 3.4 (A i)) and mean accuracy, which remains as the random chance level (Figure 3.4 (A ii)). The classifier assigns nearly all gene pairs in the test data to the regulation model, resulting in a false positive rate of ~ 1 for the regulation class (Figure 3.4 (A iii)). In contrast, training with correlation statistics alone improves the performance, reducing the mean cross

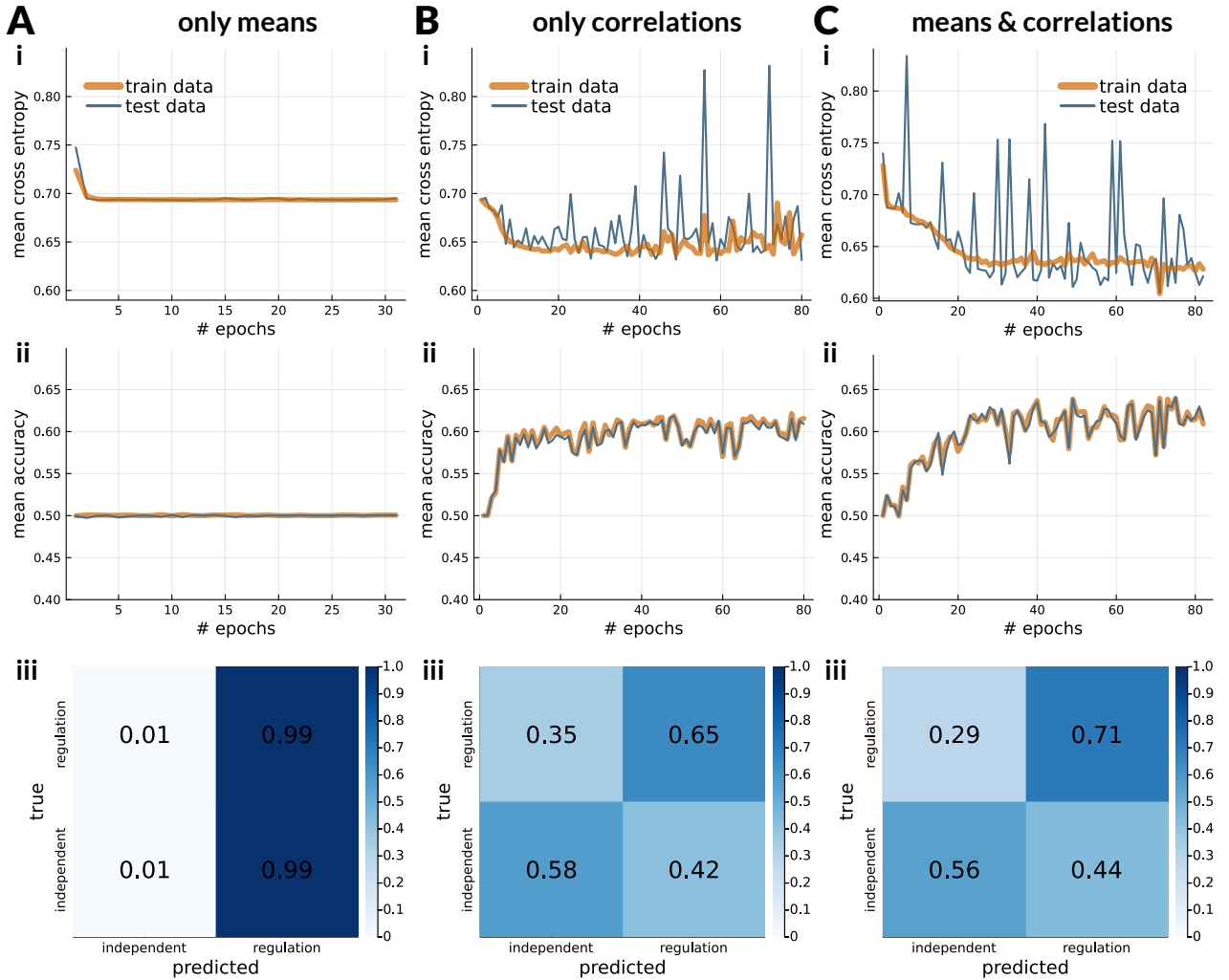


Figure 3.4: Training the binary classifier. Comparison of the binary classifier's performance after training using (A) only mean statistics, (B) only gene-gene correlation statistics and (C) both mean and correlation statistics. For all three cases, we plot (i) the mean cross entropy loss across epochs for train and test data, (ii) the mean accuracy across epochs for train and test data and (iii) the confusion matrix, that quantifies the performance of the classifier in test data (as described in 3.3.3). For this comparison, the decision threshold in all three cases is set at the minimum level, that is 50%.

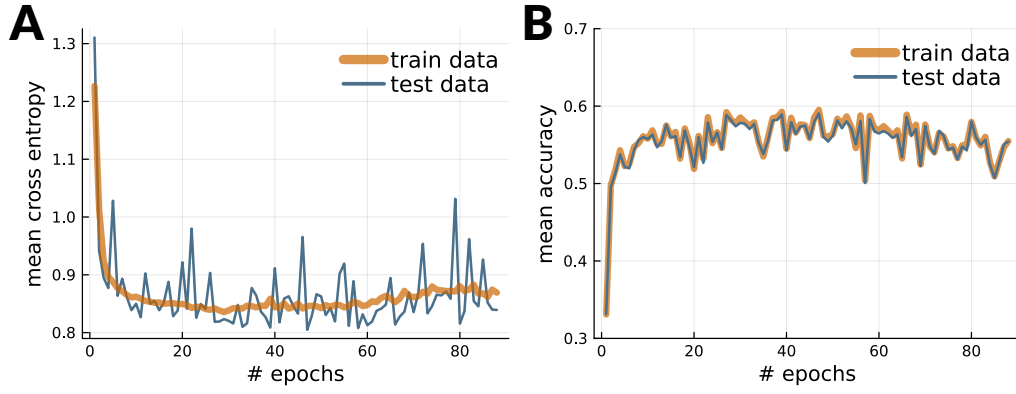


Figure 3.5: **Training the multiclass classifier.** (A) Mean cross entropy loss across epochs for train and test data. (B) Mean accuracy across epochs for train and test data.

entropy loss (Figure 3.4 (B i)) and achieving a mean accuracy of about 60% (Figure 3.4 (B ii)). In this case, the classifier correctly predicts 58% of the independent gene pairs and 64% of the regulation pairs of the test data (Figure 3.4 (B iii)). When both mean and correlation statistics are used for training, the classifier’s overall performance is further enhanced, as we observe improved training curves (Figure 3.4 (C i)) and increased mean accuracy of nearly 65% (Figure 3.4 (C ii)). In this case, the classifier successfully recovers 56% of the independent pairs and 71% of the regulation pairs (Figure 3.4 (C iii)). These findings confirm that mean expression statistics do not bias the classifier’s decisions, as we discuss in paragraph 3.3.2. Importantly, the combination of mean expression with gene-gene correlations provides the most informative set of statistics for identifying the regulation class.

Building on the insights gained from training classifier $\mathcal{C}_1$, we proceed to train the multiclass classifier $\mathcal{C}_2$, using synthetic data generated from the three gene (co-)regulation models (as described in 3.3.2). The mean cross entropy loss converges to approximately 0.8 (Figure 3.5 (A)), resulting in mean accuracy of nearly 60%, that represents a 1.8-fold increase over the random chance level (Figure 3.5 (B)). To evaluate the overall classification model’s performance, we quantify the individual performances of classifiers $\mathcal{C}_1$ and $\mathcal{C}_2$ separately, as well as the performance of $\mathcal{C}_2$ conditional on $\mathcal{C}_1$ predicting regulation. This evaluation depends on the chosen decision threshold (as explained in 3.3.3). We find that the uncertainty rate in classifier $\mathcal{C}_1$ predictions for the test set becomes higher than 60% when the decision threshold is 0.6 and exceeds 75% for thresholds above 0.7 (Figure 3.6 (A i)). For decision thresholds above 0.6, the classifier consistently recovers more than 75% of both independent and regulation pairs, with over 75% precision for both cases (Figure 3.6 (A i-ii)). Selecting a 0.6 decision threshold gives

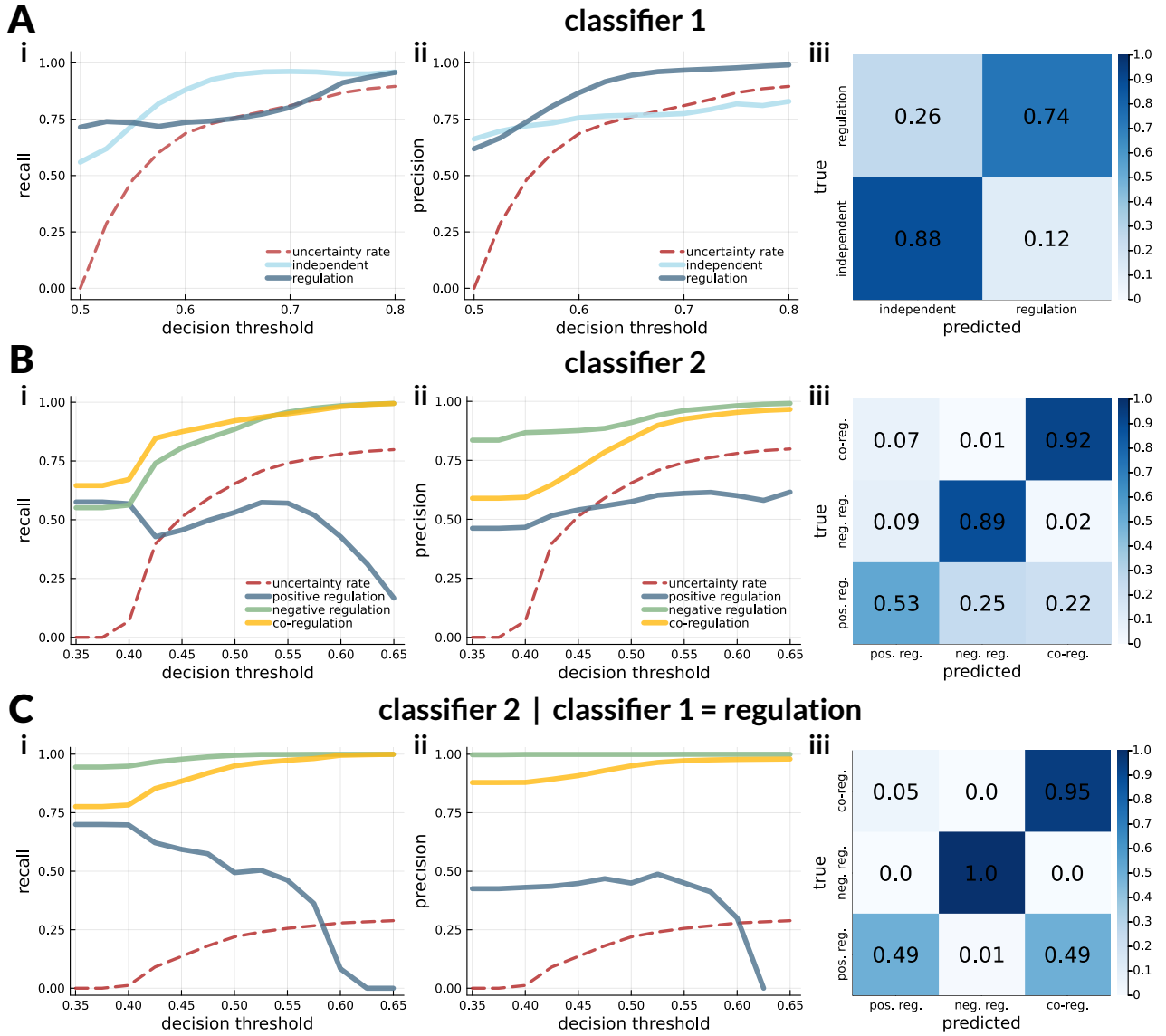


Figure 3.6: **Classification performance on synthetic test data.** Performance metrics are shown for (A) the classifier $\mathcal{C}_1$, (B) the classifier $\mathcal{C}_2$ and (C) the classifier $\mathcal{C}_2$, given that $\mathcal{C}_1$ predicts regulation. For each case, we plot (i) the recall of each class across decision thresholds, (ii) the precision of each class across decision thresholds and (iii) the confusion matrix for a fixed decision threshold, that is 0.6 in (A) and 0.5 in (B) and (C). The recall and precision plots also show uncertainty rates across thresholds. For all plots, the recall, precision and confusion matrix are evaluated among the test set examples for which the decision threshold is satisfied (as explained in 3.3.3).

a good trade-off between accuracy and uncertainty, resulting in 88% recall for the independent class and 74% recall for the regulation class (Figure 3.6 (A iii)).

The multiclass classifier $\mathcal{C}_2$ demonstrates overall good performance in identifying the three (co-)regulation models. For decision thresholds above 0.5, although the uncertainty rate exceeds 60%, both recall and precision exceed 90% for negative regulation and co-regulation pairs (Figure 3.6 (B i-ii)). The classifier correctly recovers around 55% of positive regulation pairs for thresholds between 0.5 and 0.6, with nearly 60% precision. Selecting the 0.5 decision threshold results in low false negative rates, that are below 10% for negative and co-regulation and close to 25% for positive regulation (Figure 3.6 (B iii)). Conditioning on test samples where classifier $\mathcal{C}_1$ predicts regulation, significantly reduces uncertainty, that is now below 25%, and increases recall and precision rates (Figure 3.6 (C i-ii)). In this case, classifier $\mathcal{C}_2$ recovers nearly all negative and co-regulation pairs. Relatively lower performance is observed for positive regulation, where 50% of the samples are correctly identified, whereas the remaining 50% are misclassified as co-regulation (Figure 3.6 (C iii)).

3.4.3 Dynamical gene-gene correlations identify regulatory modes of gene interaction

Analysis of the classification results reveals that dynamical gene-gene correlations of metabolic labels offer rich information for distinguishing modes of gene regulation. To better understand the role of correlations in the classification, we examine how various combinations of summary statistics influence classification outcomes. Leveraging the one-to-one correspondence in summary statistics across different regulatory models (since they were generated from the same parameters) we pinpoint different patterns of these statistics between models that can be crucial for classification decisions. Our observations overall suggest that the classifiers often require non-linear combinations of means and correlations to effectively distinguish between models.

We first observe that the classifier $\mathcal{C}_1$ successfully identifies regulation models in cases where summary statistics occupy regions of the mean-correlation space that are not captured by the independent model (Figures 3.7,3.8,3.9). For example, when the classifier is presented with co-regulation data, it correctly predicts regulation in instances where correlations extend beyond the region defined by the independent model, for given mean expression (Figure 3.7).

Notably, we also have correct classification in cases where correlations lie within the independent model region at particular labelling timepoints, yet they exceed this region at other timepoints (highlighted example, Figure 3.7). This is particularly evident in the joint distribution of the correlations $\rho(u_1, l_2)$ and $\rho(l_1, l_2)$. Conversely, the classifier fails to identify regulation in cases where both models produce highly overlapping distributions across all combinations of summary statistics.

In another example, we observe a strong overlap between the distributions of positive regulation and independent model statistics (Figure 3.8). Although the two models exhibit subtle differences in correlations for given mean expression, the joint distributions of correlations (*e.g.* $\rho(u_1, u_2)$ with $\rho(l_1, u_2)$) can be informative for correctly predicting regulation (highlighted example, Figure 3.8). However, the classifier can fail to predict regulation in cases where both models yield identical summary statistics. Similar overlap is observed between the statistics of the independent and negative regulation models (Figure 3.9). In this case, when classifier $\mathcal{C}_1$ is presented with negative regulation data, it can correctly identify regulation when correlation trajectories exceed the independent model regions (highlighted example, Figure 3.9).

Furthermore, we find that dynamical correlations are crucial for distinguishing modes of regulatory interaction. Classifier $\mathcal{C}_2$ correctly identifies co-regulation samples when correlations generated by the co-regulation model extend beyond the regions captured by the positive regulation (Figure 3.10) or negative regulation models (Figure 3.11). Importantly, accurate predictions can be made even when only specific pairs of statistics exceed the regions of positive (or negative) regulation, for at least some of the labelling timepoints (highlighted examples, Figures 3.10 and 3.11). We observe that while positive and negative regulation models produce highly overlapping distributions, identifying negative regulation can depend on particular combinations of correlations that lie outside the positive regulation regions (Figure 3.12). Similar to classifier $\mathcal{C}_1$, classifier $\mathcal{C}_2$ tends to misclassify (Figure 3.10) or return uncertain outcomes (Figures 3.11 and 3.12) when all pairs of statistics are nearly identical. Overall, our observations suggest that joint distributions of gene-gene correlations across labelling timepoints are critical for accurately identifying distinct modes of regulation.

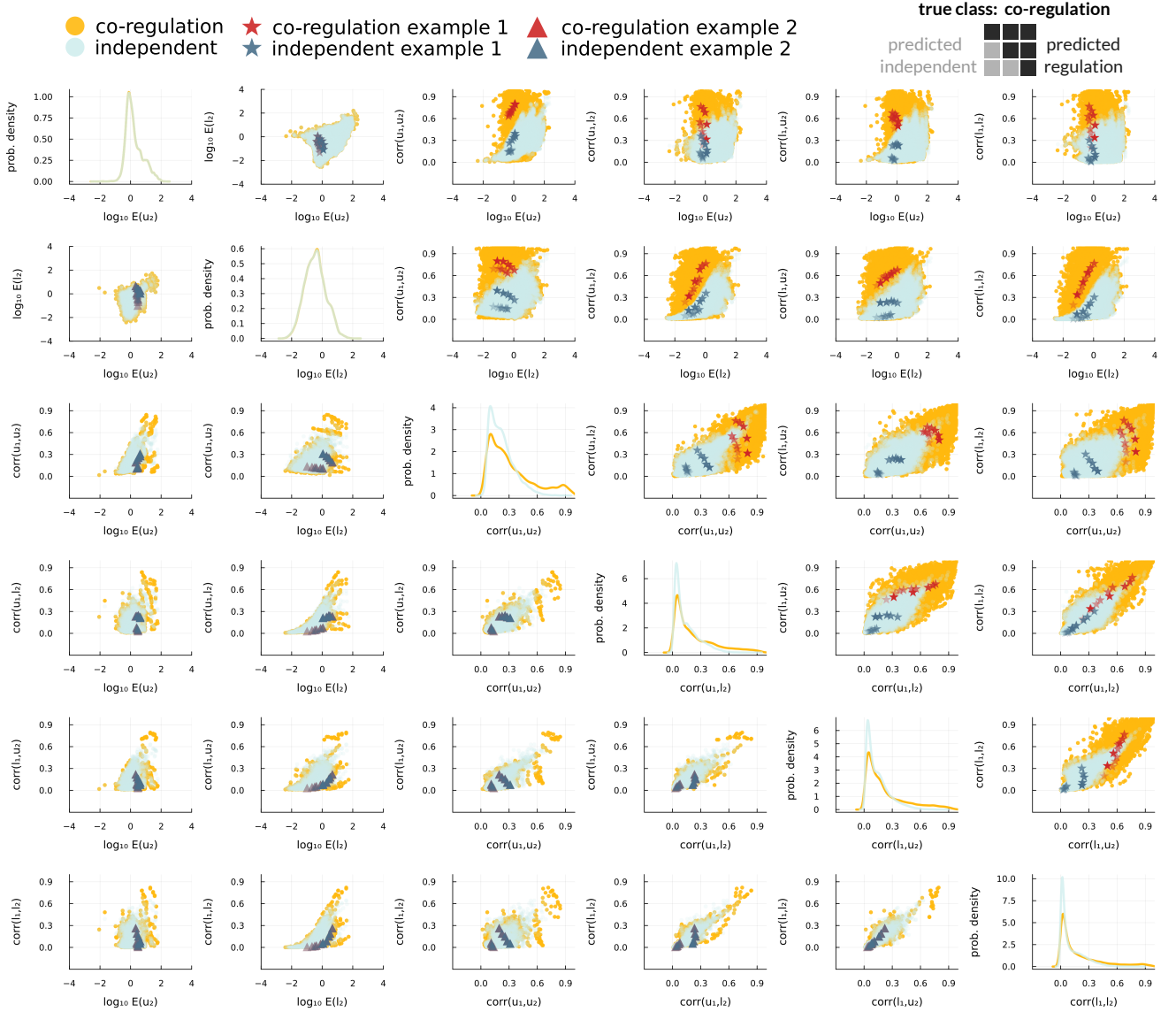


Figure 3.7: **Independent vs co-regulation statistics.** Pairplot of summary statistics of the independent model (light blue) and the co-regulation model (orange), showing empirical densities of the different statistics (diagonal) and scatter plots of statistics pairs (off-diagonal). Above and including the diagonal, we show all test samples for which classifier $\mathcal{C}_1$ correctly predicts regulation, when given co-regulation data. Below the diagonal, we show all test samples for which the classifier incorrectly predicts independence, when given co-regulation data. The independent data correspond one-to-one with the regulation data, as both were generated from the same parameter sets. Red stars mark features of a correctly classified regulation example, and blue stars mark features of the corresponding independent example. Triangles mark features of a misclassified example from both models. The increasing opacity of stars and triangles indicates progression through RNA labelling time.

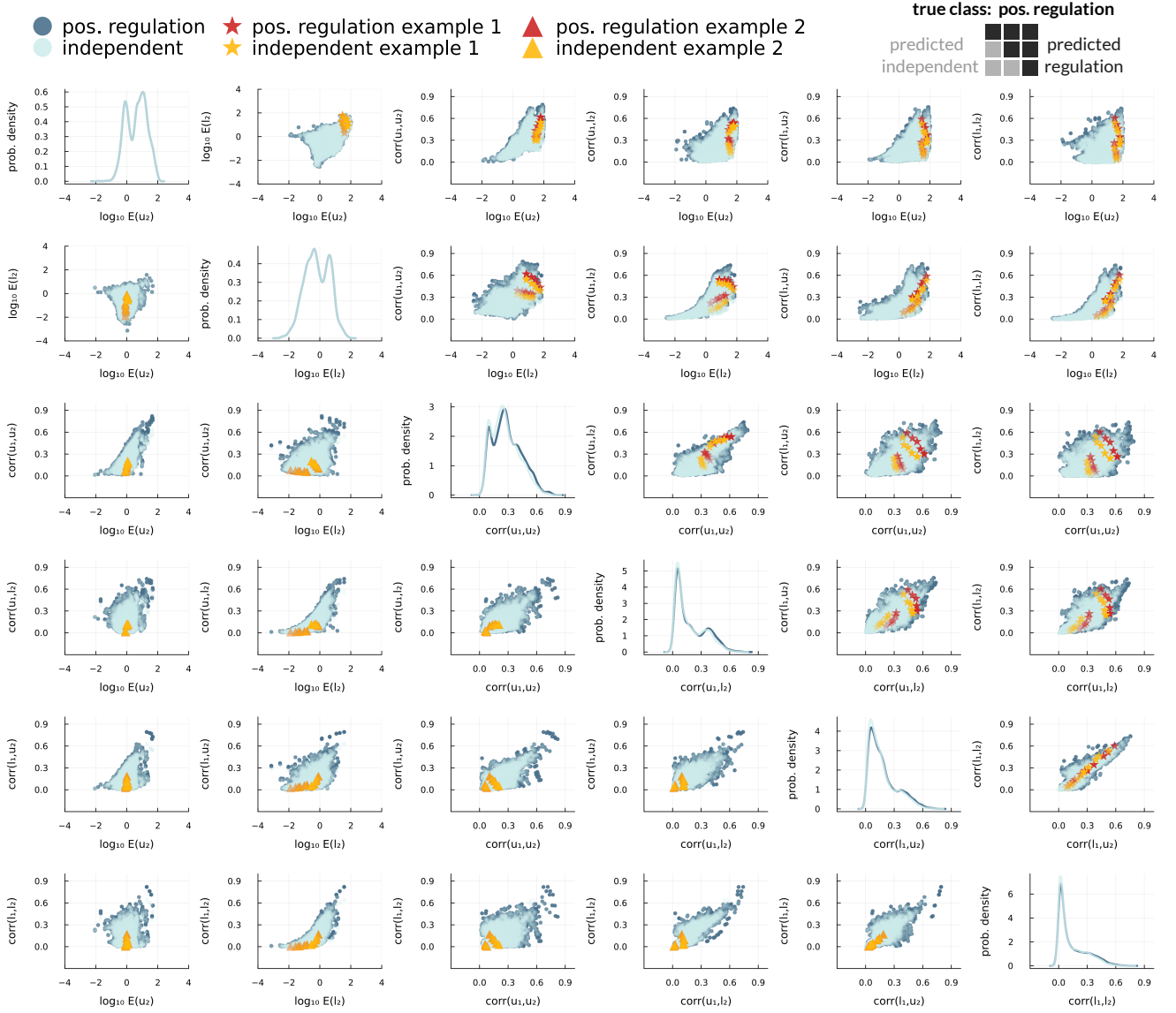


Figure 3.8: Independent vs positive regulation statistics. Pairplot of summary statistics of the independent model (light blue) and the positive regulation model (dark blue), showing empirical densities of the different statistics (diagonal) and scatter plots of statistics pairs (off-diagonal). Above and including the diagonal, we show all test samples for which classifier $\mathcal{C}_1$ correctly predicts regulation, when given positive regulation data. Below the diagonal, we show all test samples for which the classifier incorrectly predicts independence, when given positive regulation data. The independent data correspond one-to-one with the regulation data, as both were generated from the same parameter sets. Red stars mark features of a correctly classified regulation example, and orange stars mark features of the corresponding independent example. Triangles mark features of a misclassified example from both models. The increasing opacity of stars and triangles indicates progression through RNA labelling time.

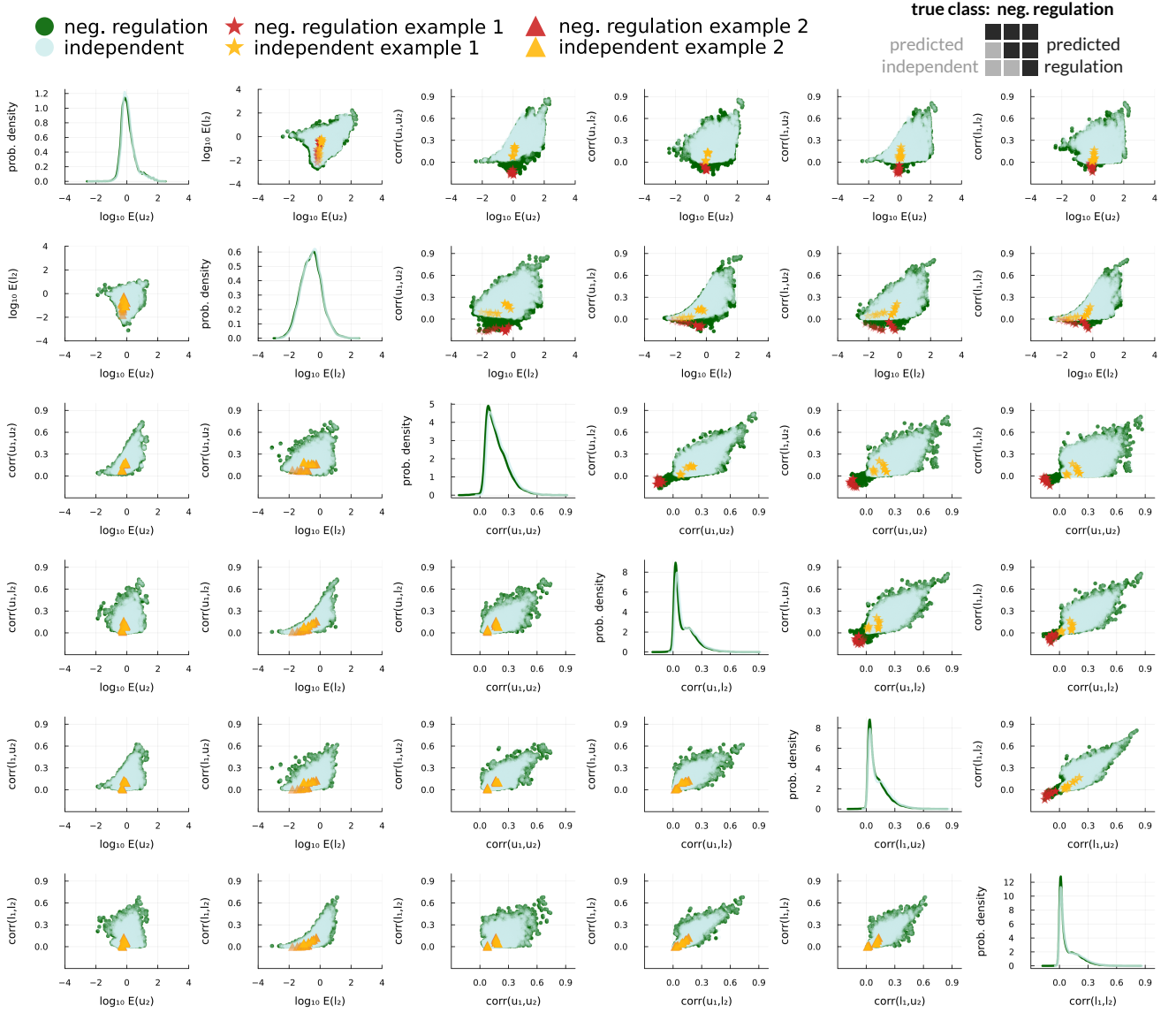


Figure 3.9: Independent vs negative regulation statistics. Pairplot of summary statistics of the independent model (light blue) and the negative regulation model (green), showing empirical densities of the different statistics (diagonal) and scatter plots of statistics pairs (off-diagonal). Above and including the diagonal, we show all test samples for which classifier $\mathcal{C}_1$ correctly predicts regulation, when given negative regulation data. Below the diagonal, we show all test samples for which the classifier incorrectly predicts independence, when given negative regulation data. The independent data correspond one-to-one with the regulation data, as both were generated from the same parameter sets. Red stars mark features of a correctly classified regulation example, and orange stars mark features of the corresponding independent example. Triangles mark features of a misclassified example from both models. The increasing opacity of stars and triangles indicates progression through RNA labelling time.

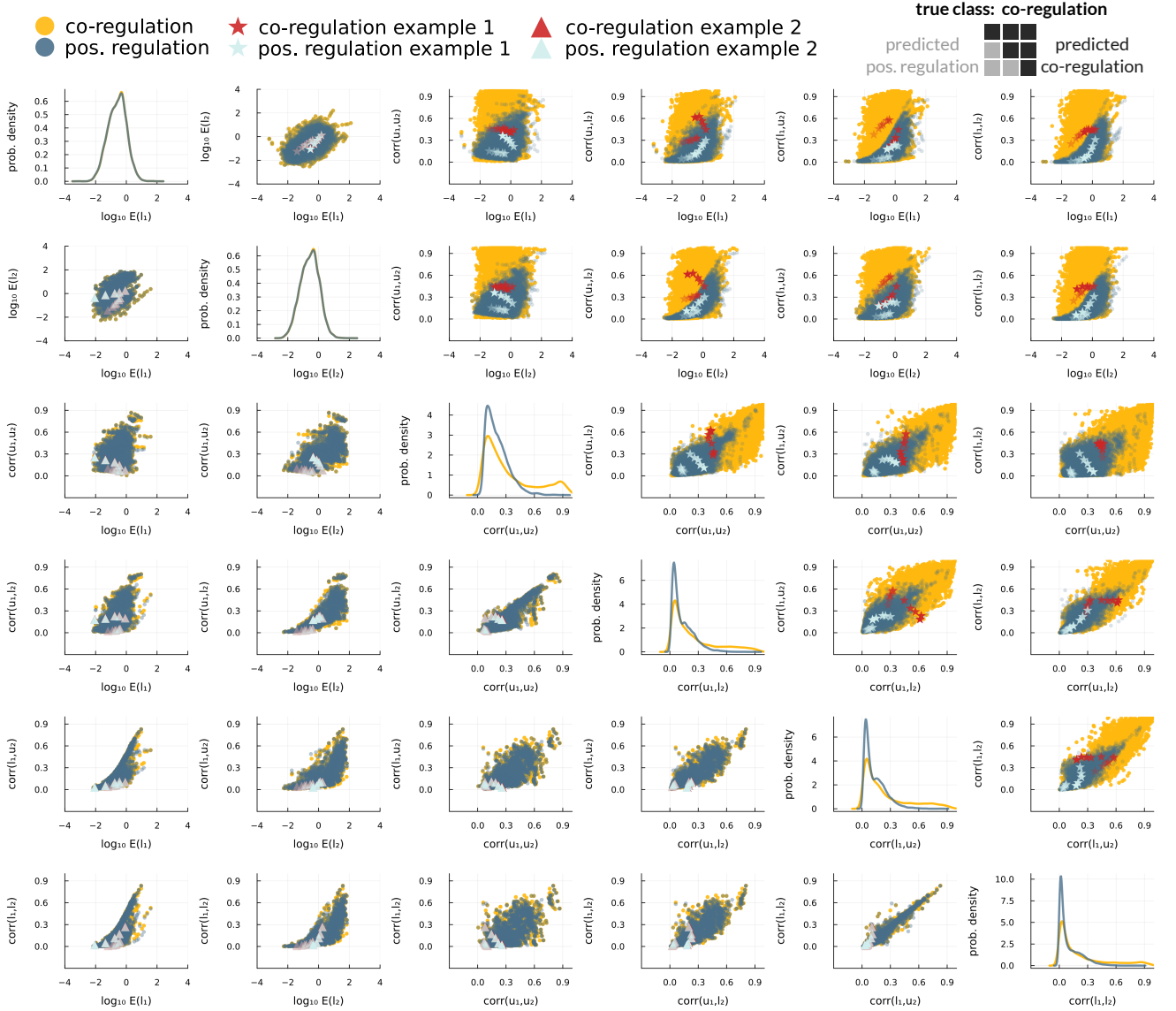


Figure 3.10: **Positive vs co-regulation statistics.** Pairplot of summary statistics of the positive regulation (dark blue) and the co-regulation model (orange), showing empirical densities of the different statistics (diagonal) and scatter plots of statistics pairs (off-diagonal). Above and including the diagonal, we show all test samples for which classifier $\mathcal{C}_2$ correctly predicts co-regulation, when given co-regulation data. Below the diagonal, we show all test samples for which the classifier incorrectly predicts positive regulation, when given co-regulation data. The positive regulation data correspond one-to-one with the co-regulation data, as both were generated from the same parameter sets. Red stars mark features of a correctly classified co-regulation example, and light blue stars mark features of the corresponding positive regulation example. Triangles mark features of a misclassified example from both models. The increasing opacity of stars and triangles indicates progression through RNA labelling time.

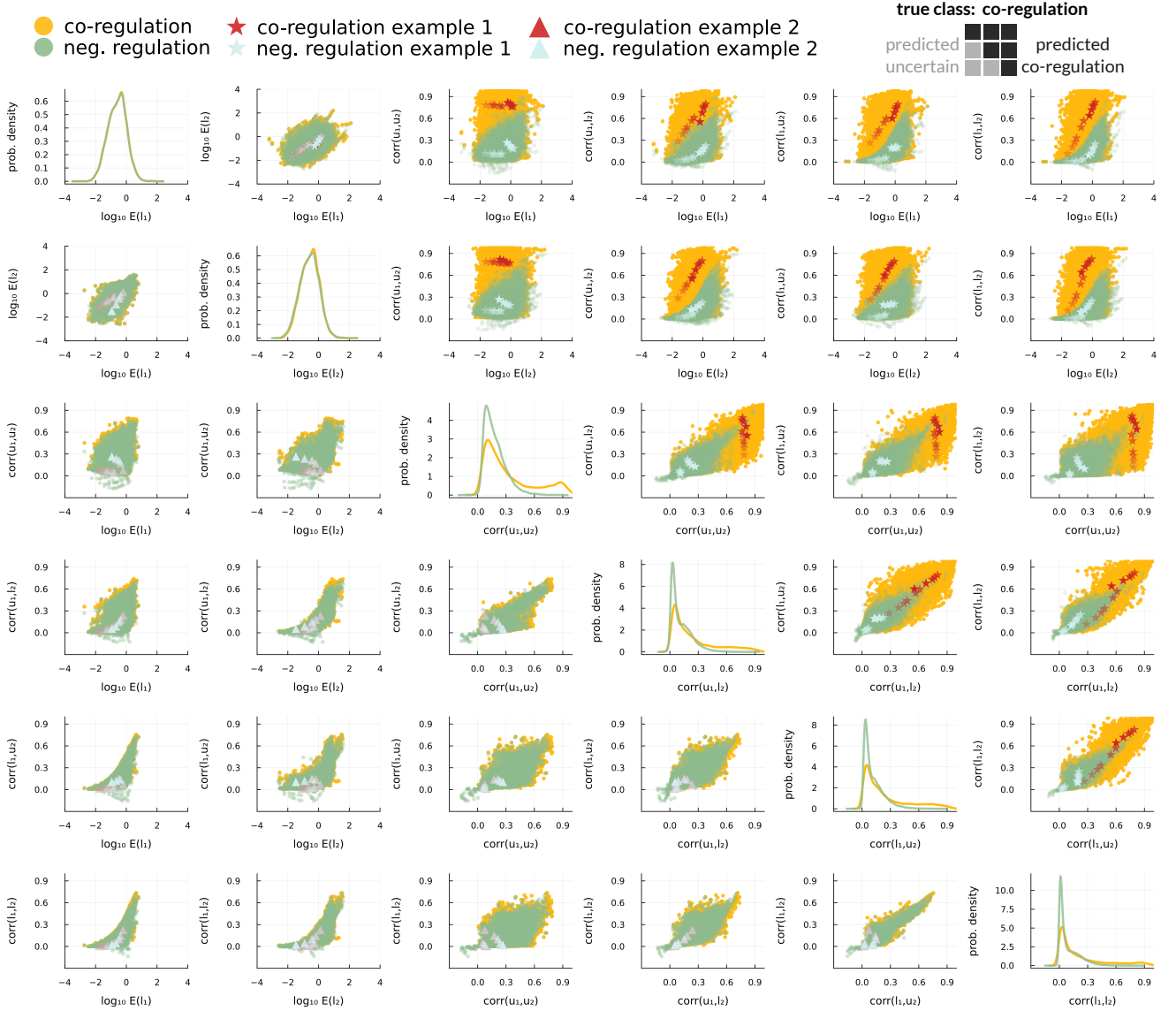


Figure 3.11: Negative vs co-regulation statistics. Pairplot of summary statistics of the negative regulation (green) and the co-regulation model (orange), showing empirical densities of the different statistics (diagonal) and scatter plots of statistics pairs (off-diagonal). Above and including the diagonal, we show all test samples for which classifier $\mathcal{C}_2$ correctly predicts co-regulation, when given co-regulation data. Below the diagonal, we show all test samples for which the classifier returns an uncertain outcome, when given co-regulation data. The negative regulation data correspond one-to-one with the co-regulation data, as both were generated from the same parameter sets. Red stars mark features of a correctly classified co-regulation example, and light blue stars mark features of the corresponding negative regulation example. Triangles mark features of an uncertain example from both models. The increasing opacity of stars and triangles indicates progression through RNA labelling time.

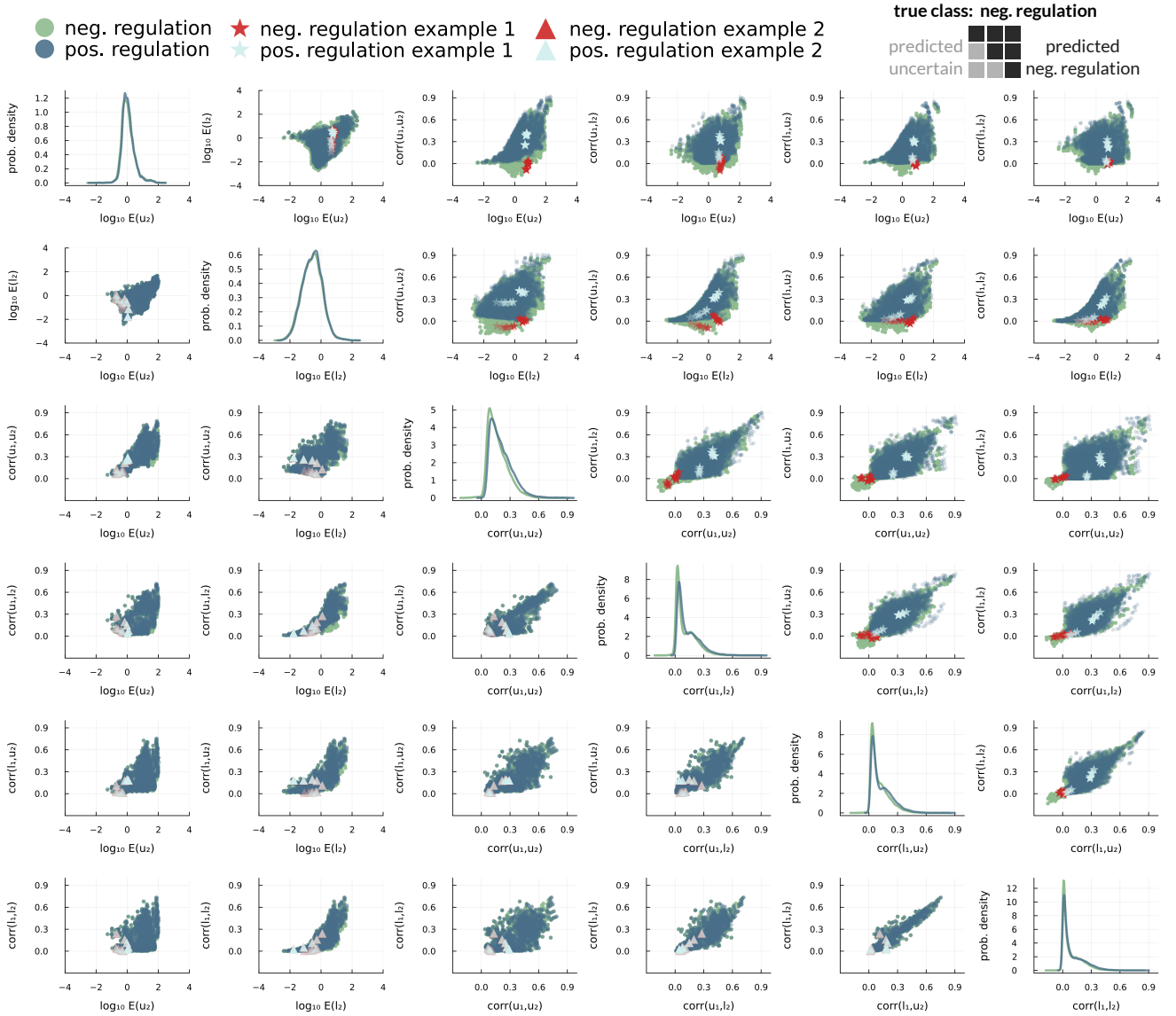


Figure 3.12: **Positive vs negative regulation statistics.** Pairplot of summary statistics of the positive regulation (dark blue) and the negative regulation model (green), showing empirical densities of the different statistics (diagonal) and scatter plots of statistics pairs (off-diagonal). Above and including the diagonal, we show all test samples for which classifier $\mathcal{C}_2$ correctly predicts negative regulation, when given negative regulation data. Below the diagonal, we show all test samples for which the classifier returns an uncertain outcome, when given negative regulation data. The negative regulation data correspond one-to-one with the co-regulation data, as both were generated from the same parameter sets. Red stars mark features of a correctly classified negative regulation example, and light blue stars mark features of the corresponding positive regulation example. Triangles mark features of an uncertain example from both models. The increasing opacity of stars and triangles indicates progression through RNA labelling time.

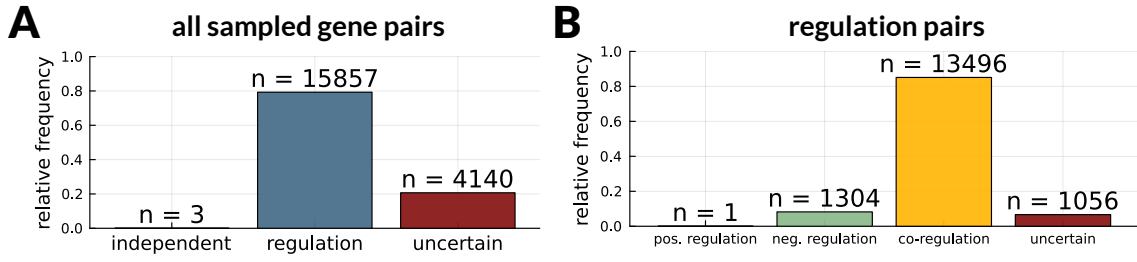


Figure 3.13: **Classification of sampled gene pairs across modes of regulation.** Frequencies of the classified gene pairs across models of gene regulation. (A) Predictions of classifier $\mathcal{C}_1$ for all gene pairs. (B) Predictions of classifier $\mathcal{C}_2$ for gene pairs that were assigned to the regulation class by classifier $\mathcal{C}_1$.

3.4.4 The model predicts modes of regulation from time-resolved scRNA-seq data

Applying the trained classifiers to scEU-seq data allows us to associate gene pairs with underlying models of gene regulation. We generate a test set comprising gene-specific mean unlabelled and labelled expression and pairwise gene correlations between unlabelled and labelled modalities, across metabolic labelling conditions, for $2 \cdot 10^5$ sampled gene pairs (as described in paragraph 3.3.3). Strikingly, our classifiers assign the vast majority of gene pairs to the regulation class, specifically to the co-regulation model. Approximately 80% of all pairs are classified as regulation, with only 3 pairs classified as independent, whereas around 20% of pairs yield an uncertain outcome (Figure 3.13 (A)). Among the regulation pairs, over 85% are classified as co-regulation, less than 10% as negative regulation and only 1 pair as positive regulation, with an uncertain outcome for about 6% of gene pairs (Figure 3.13 (B)). The overall low percentage of uncertain outcomes, despite the strict decision thresholds imposed for real data classification (paragraph 3.3.3), demonstrates high confidence in our predictions.

Next, we find that our model successfully identifies known gene regulatory relationships. After querying the BioGRID database [264], we verify the known direct interaction between *RPS25* and *RPL19* (genes encoding ribosomal proteins) which our model classifies as positive regulation. Similarly, our model recovers known interactions of gene pairs assigned to the negative regulation class. Within a subset of 200 pairs that are classified with high certainty (consistent predictions across all 100 bootstrapped samples), we confirm gene-gene interactions that are either direct or indirect, mediated by an intermediate gene. Example gene pairs with direct interaction are *AK2* - *NEK7* (genes involved in kinase and transferase activity), *VAPB*-

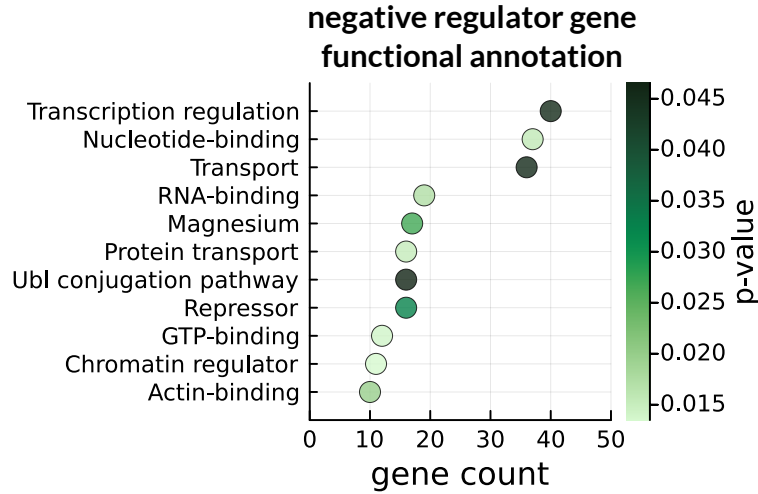


Figure 3.14: **Functional annotation of identified negative regulator genes.** Enriched biological processes and molecular functions associated with negative regulator genes, along with their frequencies and p-values.

NDUFA12 (genes encoding intracellular membrane proteins) and *ARF1-SLC38A3* (genes involved in transmembrane transporter activity). Example gene pairs with indirect interaction are *CIC* - *PSMD14* (a transcriptional repressor and a gene involved in ATP-dependent protein degradation, respectively) and *MRPL22* - *OCIAD1* (a mitochondrial ribosomal protein and a membrane protein involved in stem cell homeostasis, respectively). Additionally, we explore the functional annotation of the regulators within the 200 negative regulation pairs, using the DAVID bioinformatics database [205, 265]. We find enrichment in functions that are highly relevant to negative gene regulation, such as transcription regulation, nucleotide-binding, RNA-binding, repression and chromatin regulation (Figure 3.14).

To gain further insights into the classification results, we examine the distributions of the observed gene-pair summary statistics and their dependence on labelling time. Overall, we find that the distributions of correlation coefficients are comparable to the ones predicted by our models (diagonal panels, Figure 3.15). Additionally, we observe the presence of negative correlations, which are exclusively predicted by our negative regulation model. Notably, the joint distributions of all statistics pairs exhibit strong dependence on labelling conditions, with generally higher correlations observed during chase compared to pulse timepoints (off-diagonal panels, Figure 3.15). Moreover, for certain labelling timepoints, correlation coefficients can occupy regions that are not captured by the independent, positive and negative regulation models. This indicates the presence of additional correlations that can only be explained by the co-regulation model.

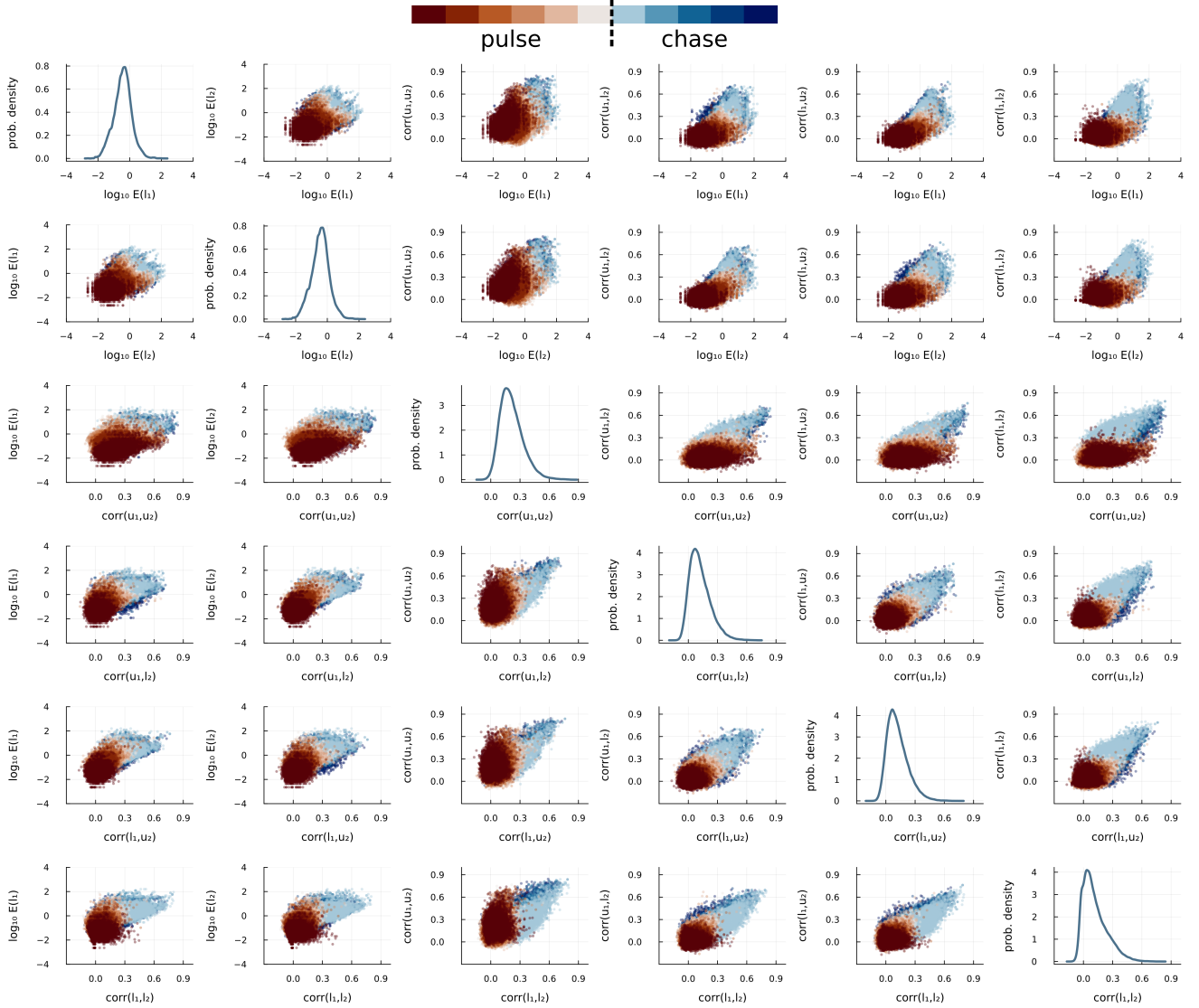


Figure 3.15: **Time-resolved gene co-expression statistics in scEU-seq data.** Pairplot of gene co-expression summary statistics for 10^5 sampled gene pairs in scEU-seq data, showing empirical densities of the different statistics (diagonal) and scatter plots of statistics pairs (off-diagonal). The data shown include labelled mean expression of the individual genes and gene-gene correlations between unlabelled and labelled transcripts, across labelling time. The red-to-blue colour gradient indicates the different RNA labelling timepoints.

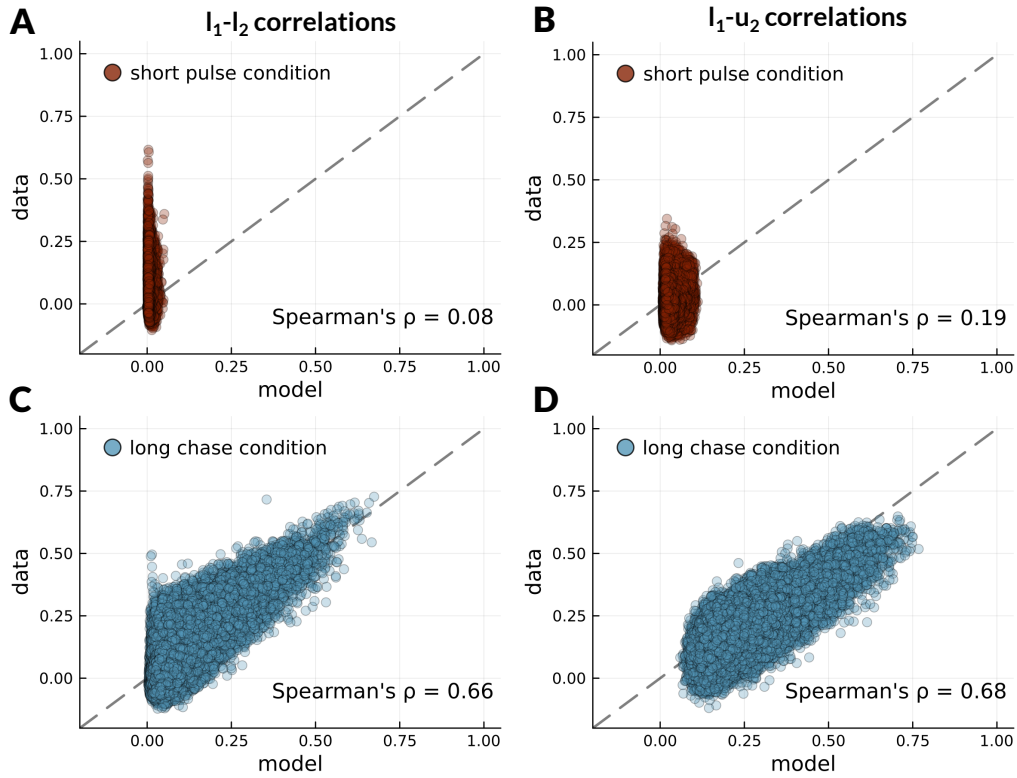


Figure 3.16: **Gene-gene correlations at particular labelling timepoints.** Comparison of independent model correlation estimates against observed correlations for 10⁵ sampled pairs of constant scaling genes. (A-B) show correlations in a short pulse labelling condition, and (C-D) show correlations in a long chase condition (see paragraph 2.2.1). (A) and (C) show $\rho(l_1, l_2)$ coefficients, while (B) and (D) show $\rho(l_1, u_2)$ coefficients.

We further confirm that independent models fail to capture correlations in particular labelling timepoints. Specifically, the data reveals high correlation coefficients at early pulse timepoints, that are not captured by the independent models (Figure 3.16 (A-B)). In contrast, correlation coefficients are more accurately predicted at chase timepoints, which involve long pulse labelling (Figure 3.16 (C-D)). Similar predictions can be expected from the positive regulation model, since it produces statistics very similar to those of the independent model (as shown in Figure 3.8). These observations help explain why independent models and positive regulation models are penalised in gene-pair classification and highlight the importance of temporal correlations in identifying modes of regulation.

3.5 Discussion

In this chapter, we developed a modelling framework to systematically investigate pairwise gene relationships through correlations in time-resolved single-cell transcriptomics. The framework integrates different models of gene regulation along with distinct biological and technical noise sources, to enable a mechanistic understanding of gene-gene transcriptional covariation. Our method effectively combines generative modelling with machine learning to leverage temporal statistics of gene co-expression and identify modes of gene regulatory interaction.

Our method addresses a model selection problem through a deep learning-enhanced simulation-based approach [103, 203], which has multiple advantages. First, it overcomes the challenge of selecting informative summary statistics, by leveraging neural networks to learn the most informative data features. Second, it reduces computational costs associated with purely generative approaches (*e.g.* ABC), due to the generalisation power of neural networks [266], that can uncover links between generative models and their outputs with a reasonable amount of training data. Additionally, our approach bypasses the challenging step of inferring the individual parameters of the regulatory models, as it harnesses prior knowledge inferred from the gene-specific transcription models (Chapter 2). These priors capture rich information about observed genome-wide expression distributions, allowing us to generate synthetic data that recapitulate these distributions. This makes our method amortised, in the sense that it can be applied, in a scalable manner, to unseen genome-wide co-expression data.

Our study highlights that dynamical gene-gene correlations across labelling timepoints are critical for identifying modes of gene regulation. Specifically, we demonstrated that different regulatory models can be distinguished from non-linear combinations of temporal correlation coefficients between unlabelled and labelled mRNA. These combinations are exploited by the classifiers to detect differences in the joint distributions of mean and correlation statistics produced by the different models. Notably, co-regulation could be distinguished by its significantly wider correlation distributions compared to other models. Nonetheless, temporal statistics also proved crucial in instances where the models produced highly overlapping distributions.

Our findings reveal that most observed gene-gene correlations are best explained by regulation models. This suggests that the overall transcriptional covariation revealed by the data cannot be explained by the independent gene models. However, we also demonstrated that independent

models can reasonably predict correlations at specific labelling timepoints, suggesting that a fraction of the data correlations stems from cell cycle covariation and technical variability. Additionally, we found that the vast majority of gene pairs are explained by the co-regulation model. This may suggest that observed correlations involve significant contributions from extrinsic noise, thereby favouring the selection of co-regulation over direct interaction models. A potential extension of our framework would be to explicitly model global extrinsic noise influencing all genes. This could help distinguish between global upstream noise effects and local co-regulatory interaction.

Our method also identified a group of gene pairs associated with negative regulation. One explanation of this classification is the presence of negative correlations in the data, that could only be explained by our negative regulation model. We validated that several of these gene pairs correspond to known regulatory links, and found that the predicted regulators are enriched in functions such as repression and nucleotide binding. These findings indicate that our method can successfully recover genetic interactions.

Although neural networks proved to be a powerful tool for our classification problem, they remain black-box models, making it challenging to explain their predictions [267, 268]. Nonetheless, we were able to interpret our classifiers' predictions by analysing how key differences in the summary statistics between models influence classification outcomes. This was possible due to the one-to-one correspondence between the outputs of the different regulatory models, ensured by our synthetic data generation process (paragraph 3.3.2). Further analysis on how classification decisions relate to the underlying gene expression parameters could provide deeper mechanistic insights. Additionally, our framework could benefit from explainable AI techniques [268, 269], such as feature attribution [270, 271], to quantify the importance of different summary statistics in the neural network predictions.

An important challenge for gene regulatory inference methods is the control of false discovery rates (FDR), a common issue in multiple hypothesis testing and multi-class classification [263, 272]. In our study, we controlled for false positives by evaluating the classifiers' performance on synthetic test data across various decision thresholds and empirically selecting a threshold that balances uncertainty and precision. We showed that our model is robust to sample size uncertainty, despite the strict decision thresholds imposed for real data classification. However, employing a more rigorous statistical approach to control FDR, would allow us to better assess

the correlations' significance and improve the overall precision [245, 273].

It is now a well established belief that time-shifted gene-gene correlations can reveal regulatory activity, by capturing stochastic expression fluctuations propagated from a regulator to a target gene [27, 91, 232, 233, 236]. However, detecting temporal covariation caused by gene interaction from simultaneous transcript correlations is particularly hard, especially in the presence of global extrinsic noise. In this study we have demonstrated that dynamical correlations over labelling time offer richer signatures for identifying regulation, compared to conventional single-cell snapshot data. Additionally, we showed that accounting for biological and technical sources of stochasticity can help distinguish extrinsic noise from true regulation. A potential extension of our framework would be to model cross-correlations between pre-mature (unspliced) and mature (spliced) mRNA, which can serve as an alternative to time-shifted correlations [91, 251, 274].

In summary, we have presented a comprehensive framework for modelling pairwise gene correlations in time-resolved single-cell transcriptomics. Our method provides a mechanistic foundation for interpreting transcriptional covariation and synergises generative modelling with machine learning to identify complex mechanisms of gene regulation from single-cell data. We anticipate that further extensions of this framework will provide a valuable tool for the systematic discovery of gene regulatory interactions.

Chapter 4

Approximation of stochastic reaction networks under timescale separation

4.1 Introduction

Inside living cells, molecules are synthesised and interact as part of intricate biochemical reaction networks, such as gene regulatory networks, protein-protein interaction networks and metabolic networks [275, 276]. These networks evolve in time under the effect of random fluctuations, inherent to the stochastic nature of biochemical interactions [4, 21]. Theoretical analyses of these systems typically rely on stochastic models to describe the noisy dynamics of the involved biochemical species. The standard tools of choice for stochastic modelling are the chemical master equation (CME) [63] and the Gillespie stochastic simulation algorithm (SSA) [65], which are used to analyse the probability distributions of molecular species' abundances, across cells and over time.

For most biochemical reaction networks, the CME is analytically intractable and stochastic simulations can be computationally intensive, posing challenges in developing accurate stochastic descriptions. To overcome these challenges, a variety of approximation methods have been developed [86], including methods for estimating probability distributions or moments of the CME [67, 69, 70, 277–281], approximate stochastic simulation algorithms [66, 282–284], and stochastic model reduction approaches [285–292]. These methods enable efficient computation of statistics to characterise the stochasticity of biochemical systems. As an example, in Chap-

ters 2 and 3 we used moment equations as an approximation of the CME, to model stochastic gene (co-)expression dynamics.

A commonly used method for approximating the exact discrete description of the CME is the system size expansion, which is a perturbative expansion with respect to the inverse system size, originally developed by van Kampen [161]. The expansion enables deriving equations for the mean macroscopic dynamics as well as the variance of the fluctuations of the CME about the mean dynamics. Specifically, truncating the system size expansion leads to the linear noise approximation (LNA) of the CME. The LNA is a linear Fokker-Planck equation (FPE) governing the time-dependent probability density of the molecular states' fluctuations about the deterministic macroscopic concentrations. Since the solution of this FPE is a multivariate Gaussian [293], the LNA serves as a practical tool for estimating the mean and noise statistics of the exact system [86]. For instance, in Chapter 3, we utilised the LNA to obtain a closed system of moment equations for the different models of gene regulation.

Another popular tool for approximating the CME is the chemical Langevin equation (CLE) [277], a stochastic differential equation (SDE) that governs the time-evolution of the biochemical network states as a continuous stochastic process. Its associated Fokker-Planck equation (the chemical FPE or CFPE) was originally derived by Kramers and Moyal through a Taylor expansion of the CME, truncated at second-order derivatives [294, 295]. The CFPE approximates the probability mass function described by the CME with a continuous probability density function, while the CLE governs the associated approximating diffusion process [293]. Notably, the CLE becomes an exact approximation of the CME in the thermodynamic limit of large molecule numbers [296]. It has also been demonstrated that the CLE is accurate even in low molecule number regimes, generally providing a better approximation of the CME statistics compared to the LNA [297]. Although the CFPE is non-linear and typically hard to solve analytically for most biochemical networks, numerical simulations of the CLE, e.g. through the Euler-Maruyama scheme [298], can be computationally more efficient compared to discrete Gillespie stochastic simulations [86]. Overall, the CLE provides a versatile tool, balancing the precision of exact, yet complex, discrete stochastic descriptions with the computation efficiency of approximate, deterministic descriptions.

Although the LNA and the CLE are useful tools for modelling biochemical reaction systems, their simulation can still be computationally expensive, particularly when the systems involve

a large number of biochemical species. Addressing this limitation necessitates methods that provide low-dimensional representations of biochemical systems. These methods enable the simulation of only a subset of the full system, reducing computational costs while capturing essential system dynamics. Many biochemical networks evolve under highly heterogeneous timescales [299]. As an example, during gene expression, mRNA molecules are typically produced and degraded within a short timescale, often up to a few hours [300], whereas protein turnover typically occurs over a longer timescale including several mRNA lifetimes [131]. Timescale separation is also prevalent in metabolic pathways, where short-lived enzymes mediate the conversion of intermediate metabolites into end products, which evolve over long timescales [301]. A classic example of this mechanism is Michaelis-Menten enzyme catalysis [302, 303]. Enzyme-catalysed reactions with multiple timescales are also found in the process of phosphorylation-dephosphorylation, a fundamental mechanism in signalling pathways [304].

The property of biochemical systems to evolve under well-separated timescales has motivated the development of approximation methods that leverage this property to reduce model complexity [54, 162, 163, 282–284, 286, 287, 289, 305–314]. In timescale separation conditions, reaction networks are commonly partitioned into *fast* and *slow* sub-systems. This partitioning is based on the frequency of occurrence of the networks’ reactions or on the characteristic timescales of change of the biochemical species, distinguishing those that evolve fast from those that evolve more slowly. The key assumption of approximation methods is that the fast sub-system rapidly reaches a quasi-steady state, when the system is observed with respect to the slow timescale. This assumption leads to a reduced system description where the dynamics of the fast sub-system is effectively averaged-out and only the dynamics of the slow sub-system is described explicitly. Originally, timescale separation methods were developed for deterministic models of reaction networks, to reduce the dimensionality of the ordinary differential equations (ODEs) governing the time-dependent macroscopic concentrations of biochemical species [315]. Popular deterministic reduction methods include the *quasi-steady state approximation* (QSSA) [316], which assumes a separation of the network into fast and slow species, and the *quasi-equilibrium approximation* (QEA) [301], which assumes a separation into fast and slow reactions. Applying these methods results in a coarse-grained network with non-elementary reaction rates that contain information about the eliminated fast sub-system dynamics.

Building on deterministic approximation methods, numerous approaches have been developed

that extend timescale separation techniques to stochastic systems [163, 317]. A common approach is applying the deterministic QSSA to the full system and utilising the obtained non-elementary reaction rates to derive a heuristic CME for the reduced system [162, 282]. However, the validity of this heuristic stochastic QSSA (sQSSA) has been questioned, as it can often fail to match the predictions of the deterministic QSSA in certain parameter regimes [318, 319]. Variants of the QEA method have also been developed for stochastic systems that can be partitioned into fast and slow reactions [283, 284, 307, 314]. Nonetheless, the separation into fast and slow reactions is a stronger assumption, and in many cases can be seen as a special case of the timescale separation of biochemical species [288].

Other studies have developed alternative methods combining timescale separation assumptions with approximations of the CME. For instance, a stochastic reduction method based on the linear noise approximation has been proposed as a rigorous alternative to the heuristic stochastic QSSA [288, 289]. Additionally, model reduction methods have been developed particularly for the CLE, that leverage adiabatic elimination techniques [285] or tools from the multiscale theory of SDEs [163, 311, 312, 320]. Yet, these methods rely on timescale separation of reactions, while no alternative to the QSSA is available for approximating the CLE description of stochastic reaction networks.

Here, we develop a novel approximation method for reducing the exact chemical Langevin equation of a biochemical reaction network under timescale separation. We derive the *slow-scale CLE* (*ssCLE*), an equation of reduced dimensionality governing the slow sub-system dynamics, that is uncoupled from the fast sub-system dynamics. Our method solely depends on the minimal assumption of separating the characteristic timescales of biochemical species, making it a rigorous alternative to the existing stochastic quasi-steady state approximations. Notably, the drift and diffusion terms of our slow-scale CLE capture contributions from both the intrinsic dynamics of the slow sub-system and the extrinsic dynamics of the fast sub-system, providing a faithful representation of the exact stochastic dynamics of the full biochemical system. We demonstrate the accuracy of our method in the classic example of Michaelis-Menten enzyme kinetics. We show that the reduced description accurately predicts the steady-state probability distributions and statistics of the exact CME, providing corrections to the estimates of the heuristic sQSSA. Our approximation method can be applied to a broad range of timescale-separated biochemical networks, including gene regulatory networks.

4.2 A reduced diffusion approximation of stochastic reaction networks

In this section, we develop the full theory for approximating the CLE description of stochastic reaction networks under timescale separation. We first formulate the exact discrete stochastic description of a reaction network using the CME, and its diffusion approximation, using the CFPE and the CLE. We then rescale the CFPE with respect to the slow timescale and utilise a projection operator method to map the exact CFPE to a reduced equation for the slow sub-system. As the resulting equation is computationally intractable, we utilise a conditional linear noise approximation approach to evaluate the contributions of the fast sub-system. This allows us to decouple the fast variables from the description and derive a tractable CFPE (and its associated CLE) for the slow sub-system.

4.2.1 Exact and approximate descriptions of stochastic reaction networks

A stochastic reaction network with d distinct biochemical species Z_i interacting via R reactions is encoded in the scheme:

$$\sum_{i=1}^d \nu_{ir}^- Z_i \xrightarrow{k_r} \sum_{i=1}^d \nu_{ir}^+ Z_i \quad , \quad r = 1, \dots, R \quad (4.1)$$

where k_r is the macroscopic rate coefficient of the r -th reaction and $\nu_{ir}^\pm$ are the stoichiometric coefficients of the i -th species in the r -th reaction. These coefficients comprise the stoichiometric matrix $\mathbf{S} \in \mathbb{Z}^{d \times R}$ of the system, with elements $S_{ir} = \nu_{ir}^+ - \nu_{ir}^-$, describing the change in the i -th species' abundance when the r -th reaction occurs. Under timescale separation conditions, the network can be partitioned into a set of d_f *fast* species Z_f , $f = 1, \dots, d_f$ and a set of $d_s = d - d_f$ *slow* species Z_s , $s = d_f + 1, \dots, d_f + d_s$, where the symbols f, s will be used throughout to denote fast and slow species' indices, respectively. We characterise species as fast or slow based on their characteristic decay (or characteristic lifetime) and we define τ_f and τ_s to be the characteristic fast and slow timescales of change respectively, so that $\gamma = \frac{\tau_s}{\tau_f} \gg 1$. Accordingly, the stoichiometric matrix can be partitioned into two sub-matrices $\mathbf{S}^f \in \mathbb{Z}^{d_f \times R}$ and $\mathbf{S}^s \in \mathbb{Z}^{d_s \times R}$, where each sub-matrix includes rows corresponding to fast and slow species,

respectively.

Chemical Master Equation description

To provide an exact discrete description of the reaction network (4.1), we encode the possible states or abundances of the fast and slow species with the vectors $\mathbf{x} = (x_1, \dots, x_{d_f}) \in \mathcal{X} \equiv \mathbb{N}^{d_f}$ and $\mathbf{y} = (y_1, \dots, y_{d_s}) \in \mathcal{Y} \equiv \mathbb{N}^{d_s}$. Then, the stochastic evolution of the fast and slow states in time is modelled by a continuous-time Markov chain $(\mathbf{X}_t, \mathbf{Y}_t)_{t \geq 0}$ with state space $\mathcal{X} \times \mathcal{Y} = \mathbb{N}^{d_f} \times \mathbb{N}^{d_s}$. The time evolution of the probability $P(\mathbf{x}, \mathbf{y}, t) = \mathbb{P}(\mathbf{X}_t = \mathbf{x}, \mathbf{Y}_t = \mathbf{y})$ to find the system in state $(\mathbf{x}, \mathbf{y})$ at time t , is given by the chemical master equation (CME):

$$\frac{\partial P(\mathbf{x}, \mathbf{y}, t)}{\partial t} = \sum_{r=1}^R \left(w_r(\mathbf{x} - \mathbf{S}_r^f, \mathbf{y} - \mathbf{S}_r^s) P(\mathbf{x} - \mathbf{S}_r^f, \mathbf{y} - \mathbf{S}_r^s, t) - w_r(\mathbf{x}, \mathbf{y}) P(\mathbf{x}, \mathbf{y}, t) \right) \quad (4.2)$$

where $w_r(\mathbf{x}, \mathbf{y})$ denotes the propensity of the r -th reaction, such that the probability that the r -th reaction occurs in an infinitesimal time-interval dt is $w_r(\mathbf{x}, \mathbf{y}) dt$. Additionally, $\mathbf{S}_r^f = (S_{fr})$, $\mathbf{S}_r^s = (S_{sr})$ denote the r -th column of the fast and slow stoichiometric sub-matrices, respectively. For notational convenience, we rewrite the CME using the step operator notation:

$$\frac{\partial P(\mathbf{x}, \mathbf{y}, t)}{\partial t} = \sum_r \left(\prod_f \mathcal{E}_{\mathbf{x}}^{-S_{fr}} \prod_s \mathcal{E}_{\mathbf{y}}^{-S_{sr}} - 1 \right) w_r(\mathbf{x}, \mathbf{y}) P(\mathbf{x}, \mathbf{y}, t) \quad (4.3)$$

where the action of the step operators on a function of fast and slow variables $g(\mathbf{x}, \mathbf{y})$ is:

$$\mathcal{E}_{\mathbf{x}}^{-S_{fr}} \mathcal{E}_{\mathbf{y}}^{-S_{sr}} g(\mathbf{x}, \mathbf{y}) = g(x_1, \dots, x_f - S_{fr}, \dots, x_{d_f}, y_1, \dots, y_s - S_{sr}, \dots, y_{d_s}) \quad (4.4)$$

We now bring equation (4.3) to an equivalent form to isolate reactions changing only fast species:

$$\frac{\partial P(\mathbf{x}, \mathbf{y}, t)}{\partial t} = (\mathcal{W}_f + \mathcal{W}_{\text{int}}) P(\mathbf{x}, \mathbf{y}, t) \quad (4.5a)$$

$$\mathcal{W}_f = \underbrace{\sum_r \left(\prod_f \mathcal{E}_{\mathbf{x}}^{-S_{fr}} - 1 \right) w_r(\mathbf{x}, \mathbf{y})}_{\text{isolated reactions changing fast species}} \quad (4.5b)$$

$$\mathcal{W}_{\text{int}} = \underbrace{\sum_r \prod_f \mathcal{E}_{\mathbf{x}}^{-S_{fr}} \left(\prod_s \mathcal{E}_{\mathbf{y}}^{-S_{sr}} - 1 \right) w_r(\mathbf{x}, \mathbf{y})}_{\text{intermediate reactions changing fast and slow species}} \quad (4.5c)$$

In this form of the CME, the transition matrix $\mathcal{W}_f$ only involves reactions changing the fast states, while the transition matrix $\mathcal{W}_{\text{int}}$ involves all *intermediate* reactions changing both fast and slow states.

Deterministic description

The reaction system (4.1) also admits a deterministic macroscopic description, given by the coupled system of ODEs:

$$\frac{d[\mathbf{x}]}{dt} = \sum_r \mathbf{S}_r^f f_r([\mathbf{x}], [\mathbf{y}]) \quad (4.6a)$$

$$\frac{d[\mathbf{y}]}{dt} = \sum_r \mathbf{S}_r^s f_r([\mathbf{x}], [\mathbf{y}]) \quad (4.6b)$$

where $[\mathbf{x}], [\mathbf{y}]$ denote macroscopic concentrations of fast and slow species. Also $f_r([\mathbf{x}], [\mathbf{y}])$ is the macroscopic rate functions associated to the r -th reaction, which is linked to the r -th microscopic propensity of the network through the large system size limit:

$$f_r([\mathbf{x}], [\mathbf{y}]) = \lim_{\Omega \rightarrow \infty} \frac{w_r(\Omega[\mathbf{x}], \Omega[\mathbf{y}])}{\Omega} \quad (4.7)$$

Diffusion approximation of the CME

Here, we are particularly interested in a continuous approximation of the CME (4.2), namely the diffusion approximation, in which the CME is approximated by a (chemical) Fokker Planck equation (CFPE) and its associated chemical Langevin equation (CLE) [277]. It is thus assumed that the molecular state $(\mathbf{x}, \mathbf{y})$ of the system is a continuous variable and so $(\mathbf{X}_t, \mathbf{Y}_t)_{t \geq 0}$ is a continuous-time Markov process with continuous state-space $\mathcal{X} \times \mathcal{Y} := \mathbb{R}^{d_f} \times \mathbb{R}^{d_s}$. The CFPE governs the time-evolution of the probability density $\rho(\mathbf{x}, \mathbf{y}, t)$ of $(\mathbf{X}_t, \mathbf{Y}_t)$ and is obtained via the Kramers-Moyal expansion of the CME [161, 293], which is a Taylor expansion of the step operators, truncated to the 2nd order:

$$\prod_f \mathcal{E}_{\mathbf{x}}^{-S_{fr}} = 1 - \sum_f S_{fr} \frac{\partial}{\partial x_f} + \frac{1}{2} \sum_{f,f'} S_{fr} S_{f'r} \frac{\partial^2}{\partial x_f \partial x_{f'}} \quad (4.8a)$$

$$\prod_s \mathcal{E}_{\mathbf{y}}^{-S_{sr}} = 1 - \sum_s S_{sr} \frac{\partial}{\partial y_s} + \frac{1}{2} \sum_{s,s'} S_{sr} S_{s'r} \frac{\partial^2}{\partial y_s \partial y_{s'}} \quad (4.8b)$$

Applying these expansions to (4.5), we obtain the CFPE:

$$\frac{\partial}{\partial t} \rho(\mathbf{x}, \mathbf{y}, t) = (\mathcal{W}_f + \mathcal{W}_{\text{int}}) \rho(\mathbf{x}, \mathbf{y}, t) \quad (4.9a)$$

$$\mathcal{W}_f(\mathbf{x}, \mathbf{y}) = - \sum_f \frac{\partial}{\partial x_f} \mathbf{A}_f(\mathbf{x}, \mathbf{y}) + \frac{1}{2} \sum_{f,f'} \frac{\partial^2}{\partial x_f \partial x_{f'}} \mathbf{B}_{ff'}(\mathbf{x}, \mathbf{y}) \quad (4.9b)$$

$$\mathcal{W}_{\text{int}}(\mathbf{x}, \mathbf{y}) = - \sum_s \frac{\partial}{\partial y_s} \mathbf{A}_s(\mathbf{x}, \mathbf{y}) + \frac{1}{2} \sum_{s,s'} \frac{\partial^2}{\partial y_s \partial y_{s'}} \mathbf{B}_{ss'}(\mathbf{x}, \mathbf{y}) + \sum_{s,f} \frac{\partial^2}{\partial y_s \partial x_f} \mathbf{B}_{sf}(\mathbf{x}, \mathbf{y}) \quad (4.9c)$$

where we define:

$$\mathbf{A}_f(\mathbf{x}, \mathbf{y}) = \sum_{r=1}^R S_{fr} w_r(\mathbf{x}, \mathbf{y}), \quad \mathbf{A}_s(\mathbf{x}, \mathbf{y}) = \sum_{r=1}^R S_{sr} w_r(\mathbf{x}, \mathbf{y}) \quad (4.10a)$$

$$\mathbf{B}_{ff'}(\mathbf{x}, \mathbf{y}) = \sum_{r=1}^R S_{fr} S_{f'r} w_r(\mathbf{x}, \mathbf{y}), \quad \mathbf{B}_{ss'}(\mathbf{x}, \mathbf{y}) = \sum_{r=1}^R S_{sr} S_{s'r} w_r(\mathbf{x}, \mathbf{y}), \quad \mathbf{B}_{sf}(\mathbf{x}, \mathbf{y}) = \sum_{r=1}^R S_{sr} S_{fr} w_r(\mathbf{x}, \mathbf{y}) \quad (4.10b)$$

The diffusion process $(\mathbf{X}_t, \mathbf{Y}_t)$ satisfies the associated d -dimensional system of chemical Langevin equations (Itô SDEs):

$$d\mathbf{X}_t = \sum_r \mathbf{S}_r^f w_r(\mathbf{X}_t, \mathbf{Y}_t) dt + \sum_r \mathbf{S}_r^f \sqrt{w_r(\mathbf{X}_t, \mathbf{Y}_t)} dV_t^{(r)} \quad (4.11a)$$

$$d\mathbf{Y}_t = \sum_r \mathbf{S}_r^s w_r(\mathbf{X}_t, \mathbf{Y}_t) dt + \sum_r \mathbf{S}_r^s \sqrt{w_r(\mathbf{X}_t, \mathbf{Y}_t)} dV_t^{(r)} \quad (4.11b)$$

where $\mathbf{V}_t = (V_t^{(1)}, \dots, V_t^{(R)})$ is an R -dimensional Wiener process. This is the exact CLE representation of the reaction network (4.1) (Figure 4.1 (A)).

The objective of the subsequent sections is to derive an approximation of the CLE description (4.11) of the reaction network (4.1), by a CLE of the form

$$d\mathbf{Y}_t = \boldsymbol{\mu}(\mathbf{Y}_t) dt + \boldsymbol{\sigma}(\mathbf{Y}_t) d\mathbf{V}_t \quad (4.12)$$

that provides a reduced description of the stochastic dynamics of the slow sub-system, decoupled from the fast sub-system dynamics (Figure 4.1 (B)).

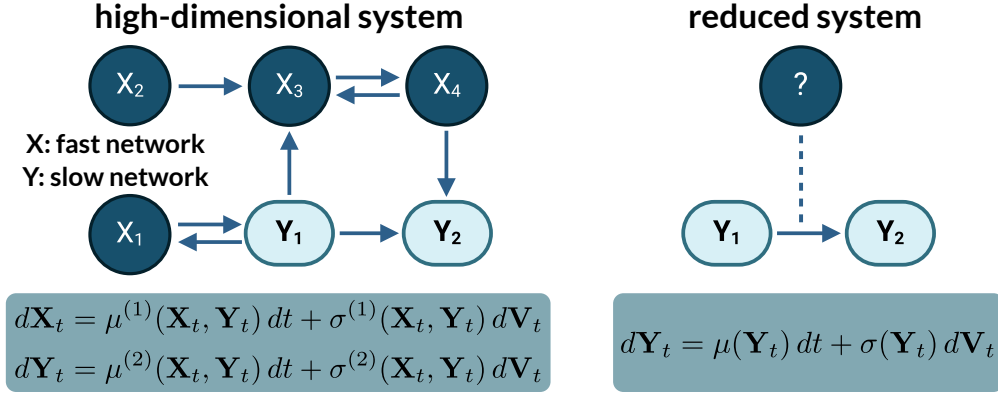


Figure 4.1: **Approximation of stochastic reaction networks with the ssCLE.** (A) Cartoon illustration of a biochemical reaction network under timescale separation. The fast and slow sub-system dynamics are governed by a coupled system of CLEs. (B) Reduced representation of the reaction network. The slow sub-system dynamics is governed by the ssCLE, which captures stochastic effects from the fast sub-system.

4.2.2 A heuristic stochastic approximation

We start by presenting a simple approach to approximating the exact CLE (4.11) of the full system, which is a heuristic stochastic alternative to the deterministic quasi-steady state approximation (QSSA) [319]. The deterministic QSSA is obtained by solving the fast sub-system ODEs (4.6a) in steady state and substituting the exact dependence on the fast species' concentrations by their steady state in the slow sub-system ODEs (4.6b):

$$\sum_r \mathbf{S}_r^f f_r([\mathbf{x}_{ss}], [\mathbf{y}]) \approx 0 \Rightarrow [\mathbf{x}_{ss}] = g([\mathbf{y}]) \quad (4.13a)$$

$$\frac{d[\mathbf{y}]}{dt} = \sum_r \mathbf{S}_r^s f_r(g([\mathbf{y}]), [\mathbf{y}]) \quad (4.13b)$$

Thus, the QSSA yields a reduced ODE (4.13b) for the slow sub-system concentrations that does not depend on fast concentrations. The macroscopic rate functions $f_r(\cdot)$, as given by the QSSA, can be used to obtain heuristic propensities for the slow molecule states by substituting $[\mathbf{y}] \rightarrow \mathbf{y}/\Omega$:

$$\tilde{w}_r(\mathbf{y}) = \Omega f_r(g(\mathbf{y}/\Omega), \mathbf{y}/\Omega) \quad (4.14)$$

Using these non-elementary propensities, one can derive a heuristic reduced CLE (hCLE) that is a stochastic version of the QSSA:

$$d\mathbf{Y}_t = \sum_r \mathbf{S}_r^s \tilde{w}_r(\mathbf{Y}_t) dt + \sum_r \mathbf{S}_r^s \sqrt{\tilde{w}_r(\mathbf{Y}_t)} dV_t^{(r)} \quad (4.15)$$

Therefore, this simple approximation method gives us an equation of the form (4.12) describing the slow sub-system dynamics that is uncoupled from the fast sub-system. However, this reduced representation does not faithfully capture the stochasticity of the full reaction network (4.1). We observe that the drift and diffusion terms of the heuristic CLE are proportional to the slow stoichiometry $\mathbf{S}^s$, meaning that they only capture the reactions that change the slow species. Thus, this equation completely neglects stochasticity induced by changes in the fast species.

In the following sections, we derive an alternative stochastic approximation that effectively captures the dynamics of both slow and fast species dynamics. The resulting equation, that we will refer to as *slow-scale CLE* (ssCLE), proposes corrections to terms of the hCLE (4.15). We will show that the drift of the ssCLE is of the form:

$$\boldsymbol{\mu}(\mathbf{y}) = \underbrace{\mathbf{S}^s \tilde{\mathbf{w}}(\mathbf{y})}_{\text{hCLE drift}} + \underbrace{\boldsymbol{\xi}(\mathbf{y})}_{\text{ssCLE correction}} \quad (4.16)$$

where the correction term $\boldsymbol{\xi}(\mathbf{y})$ involves additional effects induced by fast species dynamics.

4.2.3 Projection of the system to the slow timescale

The first step of our approximation method is to project the CFPE (4.9) to an approximate equation of the form

$$\frac{\partial}{\partial \tau} \rho(\mathbf{y}, \tau) = \mathcal{W}_s \rho(\mathbf{y}, \tau) \quad (4.17)$$

by rescaling time with respect to the slow timescale and averaging out the exact dependence on the fast dynamics.

Integrating out the fast dynamics

The averaging process relies on the assumption that, under the slow timescale, the fast dynamics relaxes to a quasi-steady state, which is formally defined as follows. Given a slow variable state $\mathbf{y} \in \mathcal{Y}$, the fast dynamics operator $\mathcal{W}_f$ defines a conditional stationary density $\pi_{\mathbf{y}}(\mathbf{x}) = \pi(\mathbf{x} \mid \mathbf{y})$

over the fast state space $\mathcal{X}$, via the detailed balance equation:

$$\mathcal{W}_f \pi(\mathbf{x} \mid \mathbf{y}) = 0 \quad (4.18)$$

This implies that the fast dynamics can be averaged with respect to $\pi_{\mathbf{y}}(\mathbf{x})$. The conditional stationary expectation of any function $g(\mathbf{x})$ of fast variables given a slow state $\mathbf{y}$ is:

$$\langle g(\mathbf{x}) \mid \mathbf{y} \rangle_{\pi} = \int_{\mathcal{X}} g(\mathbf{x}) \pi(\mathbf{x} \mid \mathbf{y}) d\mathbf{x} \quad (4.19)$$

Our goal is to eliminate the exact dependence of intermediate reactions $\mathcal{W}_{\text{int}}$ on the fast variables. To do that, we will assume that the transition matrix $\mathcal{W}_{\text{int}}$ admits a representation of the form:

$$\mathcal{W}_{\text{int}} = \langle \mathcal{W}_{\text{int}} \rangle_{\pi} + \tilde{\mathcal{W}} \quad (4.20)$$

where $\langle \mathcal{W}_{\text{int}} \rangle_{\pi}$ is the conditional stationary average over the transition rates of intermediate reactions and $\tilde{\mathcal{W}}$ describes fluctuations about this stationary average. This representation will allow us to decouple the fast from the slow dynamics.

Time rescaling

Assuming that τ_f, τ_s are the characteristic fast and slow timescales, we rescale time with respect to the slow timescale, that is $\tau = \frac{t}{\tau_s}$, and we consider the timescale separation constant $\gamma = \frac{\tau_f}{\tau_s}$. Given the new time coordinate τ , we assume the following rescaling of the transition matrices:

$$\mathcal{W}_f \rightarrow \gamma \mathcal{W}_f \quad (4.21a)$$

$$\mathcal{W}_{\text{int}} \rightarrow \langle \mathcal{W}_{\text{int}} \rangle_{\pi} + \gamma^{1/2} \tilde{\mathcal{W}} \quad (4.21b)$$

where $\mathcal{W}_f$ is scaled by γ as it involves changes in fast states exclusively, $\langle \mathcal{W}_{\text{int}} \rangle_{\pi}$ is not scaled as it only involves changes in slow states, and the fluctuations matrix $\tilde{\mathcal{W}}$ is assumed to scale with $\gamma^{1/2}$. The time-rescaled CFPE reads:

$$\frac{\partial}{\partial \tau} \rho(\mathbf{x}, \mathbf{y}, \tau) = \left(\gamma \mathcal{W}_f + \gamma^{1/2} \tilde{\mathcal{W}} + \langle \mathcal{W}_{\text{int}} \rangle_{\pi} \right) \rho(\mathbf{x}, \mathbf{y}, \tau) \quad (4.22)$$

We will now approximate this equation using an adiabatic elimination approach [285] that will yield a reduced equation describing the slow sub-system.

Projecting to the slow timescale

To carry out the adiabatic elimination, we use a *projection operator* $\mathcal{P}$, that acts on functions $g(\mathbf{x}, \mathbf{y})$ of fast and slow variables by marginalising over the fast states and rescaling by the stationary density $\pi_{\mathbf{y}}(\mathbf{x})$:

$$\mathcal{P}g(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{x} | \mathbf{y}) \int_{\mathcal{X}} g(\mathbf{x}, \mathbf{y}) d\mathbf{x} \quad (4.23)$$

We will use this projector to map the FPE (4.22) to a reduced slow-scale FPE, following the approach of [288]. We first show that $\mathcal{P}$ satisfies the following properties:

$$\mathcal{P}\tilde{\mathcal{W}}\mathcal{P} = 0 \quad (4.24a)$$

$$\mathcal{W}_f\mathcal{P} = 0 \quad (4.24b)$$

$$\mathcal{P}\mathcal{W}_f = 0 \quad (4.24c)$$

To prove (4.24a), we have that for a function $g(\mathbf{x}, \mathbf{y})$:

$$\mathcal{P}\tilde{\mathcal{W}}\mathcal{P}g(\mathbf{x}, \mathbf{y}) = \mathcal{P}\tilde{\mathcal{W}}\pi(\mathbf{x} | \mathbf{y}) \int_{\mathcal{X}} g(\mathbf{x}, \mathbf{y}) d\mathbf{x} \quad (4.25a)$$

$$= \pi(\mathbf{x} | \mathbf{y}) \int_{\mathcal{X}} \tilde{\mathcal{W}}\pi(\mathbf{x} | \mathbf{y}) \int_{\mathcal{X}} g(\mathbf{x}', \mathbf{y}) d\mathbf{x}' d\mathbf{x} \quad (4.25b)$$

$$= \pi(\mathbf{x} | \mathbf{y}) \langle \tilde{\mathcal{W}} \rangle_{\pi} \int_{\mathcal{X}} g(\mathbf{x}', \mathbf{y}) d\mathbf{x}' = 0 \quad (4.25c)$$

since $\langle \tilde{\mathcal{W}} \rangle_{\pi} = \langle \mathcal{W}_{\text{int}} - \langle \mathcal{W}_{\text{int}} \rangle_{\pi} \rangle_{\pi} = 0$. Similarly, to prove (4.24b) we write:

$$\mathcal{W}_f\mathcal{P}g(\mathbf{x}, \mathbf{y}) = \mathcal{W}_f\pi(\mathbf{x} | \mathbf{y}) \int_{\mathcal{X}} g(\mathbf{x}, \mathbf{y}) d\mathbf{x} = 0 \quad (4.26)$$

where we apply the definition of $\pi(\mathbf{x} | \mathbf{y})$ (4.18). To prove (4.24c), we write:

$$\begin{aligned} \mathcal{P}\mathcal{W}_f g(\mathbf{x}, \mathbf{y}) &= \pi(\mathbf{x} | \mathbf{y}) \int_{\mathcal{X}} \mathcal{W}_f g(\mathbf{x}, \mathbf{y}) d\mathbf{x} \\ &= \pi(\mathbf{x} | \mathbf{y}) \int_{\mathcal{X}} \sum_f \frac{\partial}{\partial x_f} \left(-\mathbf{A}_f(\mathbf{x}, \mathbf{y}) + \frac{1}{2} \sum_{f'} \frac{\partial}{\partial x_{f'}} \mathbf{B}_{ff'}(\mathbf{x}, \mathbf{y}) \right) g(\mathbf{x}, \mathbf{y}) d\mathbf{x} = 0 \end{aligned} \quad (4.27)$$

where we use that the integral involves divergence in the fast variables and hence it vanishes due to zero boundary conditions at infinity:

$$\int_{\mathcal{X}} \nabla_{\mathbf{x}} \cdot (\dots) d\mathbf{x} = 0 \quad (4.28)$$

We now define the following projections of the probability density $\rho(\tau) = \rho(\mathbf{x}, \mathbf{y}, \tau)$:

$$f(\tau) = \mathcal{P}\rho(\tau), \quad g(\tau) = \mathcal{Q}\rho(\tau) \quad (4.29)$$

with $\mathcal{Q} = 1 - \mathcal{P}$, as well as their Laplace transformations:

$$\tilde{f}(s) = \int_0^\infty e^{-s\tau} f(\tau) d\tau, \quad \tilde{g}(s) = \int_0^\infty e^{-s\tau} g(\tau) d\tau \quad (4.30)$$

Using integration by parts, we get:

$$s\tilde{f}(s) = \int_0^\infty s e^{-s\tau} f(\tau) d\tau = f(0) + \int_0^\infty s e^{-s\tau} f(\tau) d\tau \quad (4.31)$$

Using this property as well as the properties of $\mathcal{P}$ (4.24a),(4.24b),(4.24c), we get:

$$\begin{aligned} s\tilde{f}(s) - f(0) &= \int_0^\infty e^{-s\tau} \left(\gamma \mathcal{P}\mathcal{W}_f + \gamma^{1/2} \mathcal{P}\tilde{\mathcal{W}} + \mathcal{P}\langle \mathcal{W}_{\text{int}} \rangle_\pi \right) \rho(\tau) d\tau \\ &= \int_0^\infty e^{-s\tau} \left(\gamma^{1/2} \mathcal{P}\tilde{\mathcal{W}}(\mathcal{P} + \mathcal{Q}) + \mathcal{P}\langle \mathcal{W}_{\text{int}} \rangle_\pi(\mathcal{P} + \mathcal{Q}) \right) \rho(\tau) d\tau \\ &= \left(\gamma^{1/2} \mathcal{P}\tilde{\mathcal{W}} + \mathcal{P}\langle \mathcal{W}_{\text{int}} \rangle_\pi \right) \tilde{g}(s) + \mathcal{P}\langle \mathcal{W}_{\text{int}} \rangle_\pi \tilde{f}(s) \end{aligned} \quad (4.32)$$

Similarly:

$$\begin{aligned} s\tilde{g}(s) - g(0) &= \int_0^\infty e^{-s\tau} \left(\gamma \mathcal{Q}\mathcal{W}_f + \gamma^{1/2} \mathcal{Q}\tilde{\mathcal{W}} + \mathcal{Q}\langle \mathcal{W}_{\text{int}} \rangle_\pi \right) \rho(\tau) d\tau \\ &= \int_0^\infty e^{-s\tau} \left(\gamma \mathcal{W}_f \mathcal{Q} + \gamma^{1/2} \mathcal{Q}\tilde{\mathcal{W}}(\mathcal{P} + \mathcal{Q}) + \mathcal{Q}\langle \mathcal{W}_{\text{int}} \rangle_\pi(\mathcal{P} + \mathcal{Q}) \right) \rho(\tau) d\tau \\ &= \left(\gamma \mathcal{W}_f + \gamma^{1/2} \mathcal{Q}\tilde{\mathcal{W}} + \mathcal{Q}\langle \mathcal{W}_{\text{int}} \rangle_\pi \right) \tilde{g}(s) + \left(\gamma^{1/2} \mathcal{Q}\tilde{\mathcal{W}} + \mathcal{Q}\langle \mathcal{W}_{\text{int}} \rangle_\pi \right) \tilde{f}(s) \end{aligned} \quad (4.33)$$

We solve for $\tilde{g}(s)$ in (4.33) and substitute in (4.32) to get:

$$\begin{aligned} s\tilde{f}(s) - f(0) &= \left(\mathcal{P}\langle \mathcal{W}_{\text{int}} \rangle_\pi + (\gamma^{1/2} \mathcal{P}\tilde{\mathcal{W}} + \mathcal{P}\langle \mathcal{W}_{\text{int}} \rangle_\pi) \mathcal{H}(\gamma) (\gamma^{1/2} \mathcal{Q}\tilde{\mathcal{W}} + \mathcal{Q}\langle \mathcal{W}_{\text{int}} \rangle_\pi) \right) \tilde{f}(s) \\ &\quad + \left(\gamma^{1/2} \mathcal{P}\tilde{\mathcal{W}} + \mathcal{P}\langle \mathcal{W}_{\text{int}} \rangle_\pi \right) \mathcal{H}(\gamma) g(0) \end{aligned} \quad (4.34)$$

where we have defined $\mathcal{H}(\gamma) = \left(s - \gamma \mathcal{W}_f - \gamma^{1/2} \mathcal{Q} \mathcal{W}' - \mathcal{Q} \langle \mathcal{W}_{\text{int}} \rangle_\pi \right)^{-1}$. Rearranging terms, we get:

$$\begin{aligned} s \tilde{f}(s) - f(0) = & \left(\mathcal{P} \langle \mathcal{W}_{\text{int}} \rangle_\pi + \left(\mathcal{P} \tilde{\mathcal{W}} \gamma \mathcal{H}(\gamma) \mathcal{Q} \tilde{\mathcal{W}} + \mathcal{P} \tilde{\mathcal{W}} \gamma^{1/2} \mathcal{H}(\gamma) \mathcal{Q} \langle \mathcal{W}_{\text{int}} \rangle_\pi + \mathcal{P} \langle \mathcal{W}_{\text{int}} \rangle_\pi \gamma^{1/2} \mathcal{H}(\gamma) \mathcal{Q} \tilde{\mathcal{W}} \right. \right. \\ & \left. \left. + \mathcal{P} \langle \mathcal{W}_{\text{int}} \rangle_\pi \mathcal{H}(\gamma) \mathcal{Q} \langle \mathcal{W}_{\text{int}} \rangle_\pi \right) \right) \tilde{f}(s) + \left(\gamma^{1/2} \mathcal{P} \tilde{\mathcal{W}} + \mathcal{P} \langle \mathcal{W}_{\text{int}} \rangle_\pi \right) \mathcal{H}(\gamma) g(0) \end{aligned} \quad (4.35)$$

We now make use of the timescale separation assumption, that is $\gamma \rightarrow \infty$, to get

$$\gamma \mathcal{H}(\gamma) \rightarrow -\mathcal{W}_f^{-1}, \quad \gamma^{1/2} \mathcal{H}(\gamma) \rightarrow 0, \quad \mathcal{H}(\gamma) \rightarrow 0 \quad (4.36)$$

and reduce (4.35) to

$$s \tilde{f}(s) - f(0) = \mathcal{P} \left(\langle \mathcal{W}_{\text{int}} \rangle_\pi - \tilde{\mathcal{W}} \mathcal{W}_f^{-1} \mathcal{Q} \tilde{\mathcal{W}} \right) \tilde{f}(s) \quad (4.37)$$

Inverting the Laplace transformation, we obtain

$$\frac{\partial}{\partial \tau} f(\tau) = \mathcal{P} \left(\langle \mathcal{W}_{\text{int}} \rangle_\pi - \tilde{\mathcal{W}} \mathcal{W}_f^{-1} \mathcal{Q} \tilde{\mathcal{W}} \right) f(\tau) \quad (4.38)$$

Equivalently:

$$\begin{aligned} \frac{\partial}{\partial \tau} \mathcal{P} \rho(\mathbf{x}, \mathbf{y}, \tau) &= \mathcal{P} \left(\langle \mathcal{W}_{\text{int}} \rangle_\pi - \tilde{\mathcal{W}} \mathcal{W}_f^{-1} \tilde{\mathcal{W}} + \tilde{\mathcal{W}} \mathcal{W}_f^{-1} \mathcal{P} \tilde{\mathcal{W}} \right) \mathcal{P} \rho(\mathbf{x}, \mathbf{y}, \tau) \\ &= \mathcal{P} \left(\langle \mathcal{W}_{\text{int}} \rangle_\pi - \tilde{\mathcal{W}} \mathcal{W}_f^{-1} \tilde{\mathcal{W}} \right) \mathcal{P} \rho(\mathbf{x}, \mathbf{y}, \tau) \end{aligned} \quad (4.39)$$

We note that, for a function $\phi(\mathbf{x}, \mathbf{y})$ of fast and slow variables:

$$\mathcal{P} \langle \mathcal{W}_{\text{int}} \rangle_\pi \mathcal{P} \phi(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{x} | \mathbf{y}) \int_{\mathcal{X}} \langle \mathcal{W}_{\text{int}} \rangle_\pi \pi(\mathbf{x} | \mathbf{y}) \int_{\mathcal{X}} \phi(\mathbf{x}', \mathbf{y}) d\mathbf{x}' d\mathbf{x} \quad (4.40a)$$

$$= \pi(\mathbf{x} | \mathbf{y}) \langle \mathcal{W}_{\text{int}} \rangle_\pi \int_{\mathcal{X}} \phi(\mathbf{x}', \mathbf{y}) d\mathbf{x}' \quad (4.40b)$$

and

$$\mathcal{P}(\tilde{\mathcal{W}} \mathcal{W}_f^{-1} \tilde{\mathcal{W}}) \mathcal{P} \phi(\mathbf{x}', \mathbf{y}) = \pi(\mathbf{x} | \mathbf{y}) \int_{\mathcal{X}} \tilde{\mathcal{W}} \mathcal{W}_f^{-1} \tilde{\mathcal{W}} \pi(\mathbf{x} | \mathbf{y}) \int_{\mathcal{X}} \phi(\mathbf{x}', \mathbf{y}) d\mathbf{x}' d\mathbf{x} \quad (4.41a)$$

$$= \pi(\mathbf{x} | \mathbf{y}) \langle \tilde{\mathcal{W}} \mathcal{W}_f^{-1} \tilde{\mathcal{W}} \rangle_\pi \int_{\mathcal{X}} \phi(\mathbf{x}', \mathbf{y}) d\mathbf{x}' \quad (4.41b)$$

Finally, using that $\mathcal{P}\rho(\mathbf{x}, \mathbf{y}, \tau) = \pi(\mathbf{x} | \mathbf{y})\rho(\mathbf{y}, \tau)$ and substituting in equation (4.39), we obtain

$$\pi(\mathbf{x} | \mathbf{y}) \frac{\partial}{\partial \tau} \rho(\mathbf{y}, \tau) = \pi(\mathbf{x} | \mathbf{y}) \left(\langle \mathcal{W}_{\text{int}} \rangle_{\pi} - \langle \tilde{\mathcal{W}} \mathcal{W}_f^{-1} \tilde{\mathcal{W}} \rangle_{\pi} \right) \rho(\mathbf{y}, \tau) \quad (4.42)$$

Integrating over $\mathcal{X}$ yields the equation

$$\frac{\partial}{\partial \tau} \rho(\mathbf{y}, \tau) = \left(\langle \mathcal{W}_{\text{int}} \rangle_{\pi} - \langle \tilde{\mathcal{W}} \mathcal{W}_f^{-1} \tilde{\mathcal{W}} \rangle_{\pi} \right) \rho(\mathbf{y}, \tau) \quad (4.43)$$

that is a reduced CFPE on the slow state space $\mathcal{Y}$.

The reduced CFPE

We have derived a projection of the initial CFPE (4.9) to the reduced slow-timescale equation:

$$\frac{\partial}{\partial \tau} \rho(\mathbf{y}, \tau) = \mathcal{W}_s \rho(\mathbf{y}, \tau) \quad (4.44)$$

with

$$\mathcal{W}_s = \underbrace{\langle \mathcal{W}_{\text{int}} \rangle_{\pi}}_{\text{averaged dynamics}} + \underbrace{\langle \mathcal{W}_{\text{ext}} \rangle_{\pi}}_{\text{external fluctuations}} \quad (4.45a)$$

$$\mathcal{W}_{\text{ext}} = -\tilde{\mathcal{W}} \mathcal{W}_f^{-1} \tilde{\mathcal{W}} \quad (4.45b)$$

Interestingly, the slow-scale operator $\mathcal{W}_s$ consists of two parts: The averaged dynamics component $\langle \mathcal{W}_{\text{int}} \rangle_{\pi}$ describes the effect of all intermediate reactions, from which the fast dynamics has been averaged out. The external fluctuations component $\langle \mathcal{W}_{\text{ext}} \rangle_{\pi}$ describes fluctuations about the averaged dynamics, and hence summarises the stochastic effects of the fast sub-system that is now decoupled from the slow sub-system.

The averaged dynamics reads:

$$\langle \mathcal{W}_{\text{int}} \rangle_{\pi} = - \sum_s \frac{\partial}{\partial y_s} \langle \mathbf{A}_s(\mathbf{x}, \mathbf{y}) \rangle_{\pi} + \frac{1}{2} \sum_{s, s'} \frac{\partial^2}{\partial y_s \partial y_{s'}} \langle \mathbf{B}_{ss'}(\mathbf{x}, \mathbf{y}) \rangle_{\pi} + \sum_s \frac{\partial}{\partial y_s} \underbrace{\left\langle \sum_f \frac{\partial}{\partial x_f} \mathbf{B}_{sf}(\mathbf{x}, \mathbf{y}) \right\rangle_{\pi}}_{=0} \quad (4.46)$$

where the third term vanishes, as it is an integral of the divergence in the fast variables, due to zero boundary values of the probability density $\pi_{\mathbf{y}}(\mathbf{x})$ at infinity. Also, the intermediate

fluctuations matrix $\tilde{\mathcal{W}}$ reads:

$$\tilde{\mathcal{W}} = \mathcal{W}_{\text{int}} - \langle \mathcal{W}_{\text{int}} \rangle_{\pi} \quad (4.47a)$$

$$\begin{aligned} &= - \sum_s \frac{\partial}{\partial y_s} (\mathbf{A}_s(\mathbf{x}, \mathbf{y}) - \langle \mathbf{A}_s(\mathbf{x}, \mathbf{y}) \rangle_{\pi}) + \frac{1}{2} \sum_{s,s'} \frac{\partial^2}{\partial y_s \partial y_{s'}} (\mathbf{B}_{ss'}(\mathbf{x}, \mathbf{y}) - \langle \mathbf{B}_{ss'}(\mathbf{x}, \mathbf{y}) \rangle_{\pi}) \\ &\quad + \sum_{s,f} \frac{\partial^2}{\partial y_s \partial x_f} \mathbf{B}_{sf}(\mathbf{x}, \mathbf{y}) \end{aligned} \quad (4.47b)$$

A key challenge here is that it is not possible to obtain an exact expression for $\mathcal{W}_s$ in the general case, as calculating the stationary density (4.18) for reaction networks with second or higher order reactions is intractable. To overcome that, we approximate the non-linear propensities of the network by applying the Linear Noise Approximation (LNA) on the fast sub-system, assuming a sufficiently large system size (or large molecule copy numbers).

4.2.4 Conditional linear noise approximation of the fast sub-system

The starting point of the approximation is Van Kampen's ansatz [161] that relates molecule numbers with macroscopic concentrations given a system size Ω . Under this ansatz, given a fixed slow state $\mathbf{y} \in \mathcal{Y}$, the fast molecular state $\mathbf{x} \in \mathcal{X}$ satisfy:

$$\frac{\mathbf{x} | \mathbf{y}}{\Omega} = [\mathbf{x} | \mathbf{y}] + \Omega^{-1/2} \boldsymbol{\epsilon}^{\mathbf{x}} \quad (4.48)$$

where $[\mathbf{x} | \mathbf{y}]$ is the macroscopic concentration of $\mathbf{x} | \mathbf{y}$ and $\boldsymbol{\epsilon}^{\mathbf{x}}$ quantifies fluctuations about the concentration. The (conditional) steady state mean of the fluctuations with respect to $\pi_{\mathbf{y}}(\mathbf{x})$ is zero, *i.e.* $\langle \boldsymbol{\epsilon}^{\mathbf{x}} \rangle_{\pi} = 0$. This implies that the steady-state mean of the fast molecule numbers can be obtained directly from the fast ODEs (4.6a) solved at steady state:

$$\langle \mathbf{x} | \mathbf{y} \rangle_{\pi} = \Omega [\mathbf{x} | \mathbf{y}] \quad (4.49)$$

This steady-state mean corresponds to the steady-state concentration obtained in a deterministic quasi-state approximation [316]. Note that the states $\mathbf{x} | \mathbf{y}$ are realisations of the conditional stochastic process $\mathbf{X}_t | \mathbf{y}$, implying that the fast process is observed given a fixed slow state. Due to timescale separation, we assume, that at each time point $t \geq 0$, the fast stochastic process relaxes rapidly to a steady state $\langle \mathbf{x} | \mathbf{y} \rangle_{\pi}$. This allows us to replace the exact fast

dynamics with its steady state in the CLE description, while also quantifying the effects of fast fluctuations $\epsilon^{\mathbf{x}}$ in the reduced system.

System size expansion of the propensities

We assume that the propensities have an expansion with respect to the system size Ω of the form

$$w_r(\mathbf{x}, \mathbf{y}) = \Omega f_r\left(\frac{\mathbf{x}}{\Omega}, \frac{\mathbf{y}}{\Omega}\right) + \mathcal{O}(\Omega^0) \quad (4.50)$$

where the first term of this expansion is the deterministic macroscopic rate f_r at the large system size limit (4.7). This implies the approximation

$$\frac{w_r(\mathbf{x}, \mathbf{y})}{\Omega} \approx f_r([\mathbf{x}], [\mathbf{y}]) \quad (4.51)$$

We now rewrite the expansion (4.50) using the ansatz (4.49):

$$w_r(\mathbf{x}, \mathbf{y}) = w_r(\Omega [\mathbf{x} | \mathbf{y}] + \Omega^{1/2} \epsilon^{\mathbf{x}}, \mathbf{y}) = \Omega f_r\left([\mathbf{x} | \mathbf{y}] + \Omega^{-1/2} \epsilon^{\mathbf{x}}, \frac{\mathbf{y}}{\Omega}\right) + \mathcal{O}(\Omega^0) \quad (4.52a)$$

$$= \Omega f_r([\mathbf{x} | \mathbf{y}], [\mathbf{y}]) + \Omega^{1/2} \sum_f \frac{\partial}{\partial [x_f]} f_r([\mathbf{x} | \mathbf{y}], [\mathbf{y}]) \epsilon_f^{\mathbf{x}} + \mathcal{O}(\Omega^0) \quad (4.52b)$$

Equivalently, we rewrite the expansion with respect to propensities using (4.51):

$$w_r(\mathbf{x}, \mathbf{y}) = w_r(\Omega [\mathbf{x} | \mathbf{y}], \mathbf{y}) + \Omega^{-1/2} \sum_f \frac{\partial}{\partial [x_f]} w_r(\Omega [\mathbf{x} | \mathbf{y}], \mathbf{y}) \epsilon_f^{\mathbf{x}} + \mathcal{O}(\Omega^{-1}) \quad (4.53)$$

Note that the first term in the expansion (4.53) is an averaged propensity $a_r(\mathbf{y}) = w_r(\langle \mathbf{x} | \mathbf{y} \rangle_\pi, \mathbf{y})$ where the fast variables $\mathbf{x}$ are replaced by their steady-state mean $\langle \mathbf{x} | \mathbf{y} \rangle_\pi$, while the second term describes fluctuations that scale with $\Omega^{-1/2}$.

As a consequence, we get the following expansions:

$$\mathbf{A}_s(\mathbf{x}, \mathbf{y}) = \sum_r S_{sr} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) + \Omega^{-1/2} \sum_r S_{sr} \sum_f \frac{\partial}{\partial[x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \epsilon_f^x + \mathcal{O}(\Omega^{-1}) \quad (4.54)$$

$$\mathbf{B}_{ss'}(\mathbf{x}, \mathbf{y}) = \sum_r S_{sr} S_{s'r} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) + \Omega^{-1/2} \sum_r S_{sr} S_{s'r} \sum_f \frac{\partial}{\partial[x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \epsilon_f^x + \mathcal{O}(\Omega^{-1}) \quad (4.55)$$

$$\mathbf{B}_{sf}(\mathbf{x}, \mathbf{y}) = \sum_r S_{sr} S_{fr} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) + \Omega^{-1/2} \sum_r S_{sr} S_{fr} \sum_f \frac{\partial}{\partial[x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \epsilon_f^x + \mathcal{O}(\Omega^{-1}) \quad (4.56)$$

which implies that:

$$\mathbf{A}_s(\mathbf{x}, \mathbf{y}) - \langle \mathbf{A}_s(\mathbf{x}, \mathbf{y}) \rangle_\pi = \Omega^{-1/2} \sum_r S_{sr} \sum_f \frac{\partial}{\partial[x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \epsilon_f^x + \mathcal{O}(\Omega^{-1}) \quad (4.57)$$

$$\mathbf{B}_{ss'}(\mathbf{x}, \mathbf{y}) - \langle \mathbf{B}_{ss'}(\mathbf{x}, \mathbf{y}) \rangle_\pi = \Omega^{-1/2} \sum_r S_{sr} S_{s'r} \sum_f \frac{\partial}{\partial[x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \epsilon_f^x + \mathcal{O}(\Omega^{-1}) \quad (4.58)$$

Additionally, from the ansatz (4.49) we have that:

$$\frac{\partial}{\partial x_f} = \Omega^{-1/2} \frac{\partial}{\partial \epsilon_f^x} \quad (4.59)$$

and thus:

$$\sum_f \frac{\partial}{\partial x_f} \mathbf{B}_{sf}(\mathbf{x}, \mathbf{y}) = \Omega^{-1/2} \sum_r S_{sr} \sum_f S_{fr} \frac{\partial}{\partial \epsilon_f^x} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) + \mathcal{O}(\Omega^{-1}) \quad (4.60)$$

We will now use these expansions to obtain a system size expansion of the reduced FPE (4.44).

System size expansion of the slow-scale FPE

Gathering the above expansions we can obtain truncated expansions of all terms contributing to the slow-scale operator (4.45). The averaged dynamics operator becomes:

$$\langle \mathcal{W}_{\text{int}} \rangle_\pi = - \sum_s \frac{\partial}{\partial y_s} \sum_r S_{sr} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) + \frac{1}{2} \sum_{s,s'} \frac{\partial^2}{\partial y_s \partial y_{s'}} \sum_r S_{sr} S_{s'r} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \quad (4.61)$$

Also, we get an expansion for equation (4.47):

$$\begin{aligned}\tilde{\mathcal{W}} = & - \sum_s \frac{\partial}{\partial y_s} \sum_r S_{sr} \left(\Omega^{-1/2} \sum_f \frac{\partial}{\partial [x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \epsilon_f^{\mathbf{x}} - \Omega^{-1/2} \sum_f S_{fr} \frac{\partial}{\partial \epsilon_f^{\mathbf{x}}} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \right) \\ & + \frac{1}{2} \sum_{s,s'} \frac{\partial^2}{\partial y_s \partial y_{s'}} \Omega^{-1/2} \sum_r S_{sr} S_{s'r} \sum_f \frac{\partial}{\partial [x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \epsilon_f^{\mathbf{x}} + \mathcal{O}(\Omega^{-1})\end{aligned}\quad (4.62)$$

Finally, we obtain a truncated expansion of the external fluctuations operator (4.45b):

$$\begin{aligned}\langle \mathcal{W}_{\text{ext}} \rangle_{\pi} = & - \frac{1}{\Omega} \sum_{r,r'} \sum_{s,s'} \frac{\partial}{\partial y_s} S_{sr} \sum_{f,f'} \frac{\partial}{\partial [x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \frac{\partial}{\partial y_{s'}} S_{s'r'} \left\langle \epsilon_f^{\mathbf{x}} \mathcal{W}_f^{-1} \epsilon_{f'}^{\mathbf{x}} \right\rangle_{\pi} \frac{\partial}{\partial [x_{f'}]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \\ & + \frac{1}{\Omega} \sum_{r,r'} \sum_{s,s'} \frac{\partial}{\partial y_s} S_{sr} \sum_{f,f'} \frac{\partial}{\partial [x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \frac{\partial}{\partial y_{s'}} \left\langle \epsilon_f^{\mathbf{x}} \mathcal{W}_f^{-1} \frac{\partial}{\partial \epsilon_{f'}^{\mathbf{x}}} \right\rangle_{\pi} S_{f'r'} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \\ & + \frac{1}{\Omega} \sum_{r,r'} \sum_{s,s'} \frac{\partial}{\partial y_s} S_{sr} \sum_{f,f'} S_{fr} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \frac{\partial}{\partial y_{s'}} \left\langle \frac{\partial}{\partial \epsilon_f^{\mathbf{x}}} \mathcal{W}_f^{-1} \epsilon_{f'}^{\mathbf{x}} \right\rangle_{\pi} \frac{\partial}{\partial [x_{f'}]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \\ & - \frac{1}{\Omega} \sum_{r,r'} \sum_{s,s'} \frac{\partial}{\partial y_s} S_{sr} \sum_{f,f'} S_{f'r} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \frac{\partial}{\partial y_{s'}} \left\langle \frac{\partial}{\partial \epsilon_f^{\mathbf{x}}} \mathcal{W}_f^{-1} \frac{\partial}{\partial \epsilon_{f'}^{\mathbf{x}}} \right\rangle_{\pi} S_{f'r'} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y})\end{aligned}\quad (4.63)$$

Note that the third term in the expansion of $\tilde{\mathcal{W}}$ (4.62), when multiplied with other terms, gives rise to higher order derivatives (non-Fokker-Planck terms) and hence it is omitted from the above expansion. Also, the terms $\left\langle \frac{\partial}{\partial \epsilon_f^{\mathbf{x}}} \mathcal{W}_f^{-1} \epsilon_{f'}^{\mathbf{x}} \right\rangle_{\pi} = \left\langle \frac{\partial}{\partial \epsilon_f^{\mathbf{x}}} \mathcal{W}_f^{-1} \frac{\partial}{\partial \epsilon_{f'}^{\mathbf{x}}} \right\rangle_{\pi} = 0$ as explained in (4.28). Thus:

$$\begin{aligned}\langle \mathcal{W}_{\text{ext}} \rangle_{\pi} = & - \sum_{r,r'} \sum_{s,s'} \frac{\partial}{\partial y_s} S_{sr} \sum_{f,f'} \frac{\partial}{\partial [x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \frac{\partial}{\partial y_{s'}} S_{s'r'} \cdot \\ & \cdot \left(\frac{1}{\Omega} \left\langle \epsilon_f^{\mathbf{x}} \mathcal{W}_f^{-1} \epsilon_{f'}^{\mathbf{x}} \right\rangle_{\pi} \frac{\partial}{\partial [x_{f'}]} w_{r'}(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) - \frac{1}{\Omega} \left\langle \epsilon_f^{\mathbf{x}} \mathcal{W}_f^{-1} \frac{\partial}{\partial \epsilon_{f'}^{\mathbf{x}}} \right\rangle_{\pi} S_{f'r'} w_{r'}(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \right)\end{aligned}\quad (4.64)$$

or equivalently:

$$\begin{aligned}\langle \mathcal{W}_{\text{ext}} \rangle_{\pi} = & - \sum_{s,s'} \frac{\partial}{\partial y_s} \sum_f \sum_r S_{sr} \frac{\partial}{\partial [x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \frac{\partial}{\partial y_{s'}} \cdot \\ & \cdot \sum_{f'} \left(\frac{1}{\Omega} \left\langle \epsilon_f^{\mathbf{x}} \mathcal{W}_f^{-1} \epsilon_{f'}^{\mathbf{x}} \right\rangle_{\pi} \sum_{r'} S_{s'r'} \frac{\partial}{\partial [x_{f'}]} w_{r'}(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) - \frac{1}{\Omega} \left\langle \epsilon_f^{\mathbf{x}} \mathcal{W}_f^{-1} \frac{\partial}{\partial \epsilon_{f'}^{\mathbf{x}}} \right\rangle_{\pi} \sum_{r'} S_{f'r'} S_{s'r'} w_{r'}(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \right)\end{aligned}\quad (4.65)$$

We now note that the product rule of differentiation gives:

$$\sum_r S_{sr} \frac{\partial}{\partial [x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \frac{\partial}{\partial y_{s'}} = \frac{\partial}{\partial y_{s'}} \sum_r S_{sr} \frac{\partial}{\partial [x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) - \delta_{ss'} \sum_r S_{s'r} \frac{\partial^2}{\partial y_{s'} \partial [x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \quad (4.66)$$

which implies that:

$$\begin{aligned} \langle \mathcal{W}_{\text{ext}} \rangle_\pi &= \sum_{s,s'} \left(\frac{\partial}{\partial y_s} \delta_{ss'} \sum_f \sum_r S_{s'r} \frac{\partial^2}{\partial y_{s'} \partial [x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) - \frac{\partial^2}{\partial y_s \partial y_{s'}} \sum_f \sum_r S_{sr} \frac{\partial}{\partial [x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \right) \cdot \\ &\cdot \sum_{f'} \left(\frac{1}{\Omega} \left\langle \boldsymbol{\epsilon}_f^x \mathcal{W}_f^{-1} \boldsymbol{\epsilon}_{f'}^x \right\rangle_\pi \sum_{r'} S_{s'r'} \frac{\partial}{\partial [x_{f'}]} w_{r'}(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) - \frac{1}{\Omega} \left\langle \boldsymbol{\epsilon}_f^x \mathcal{W}_f^{-1} \frac{\partial}{\partial \boldsymbol{\epsilon}_{f'}^x} \right\rangle_\pi \sum_{r'} S_{f'r'} S_{s'r'} w_{r'}(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \right) \end{aligned} \quad (4.67)$$

We use known results from [288] to evaluate:

$$\frac{1}{\Omega} \left\langle \boldsymbol{\epsilon}_f^x \mathcal{W}_f^{-1} \boldsymbol{\epsilon}_{f'}^x \right\rangle_\pi = -\frac{1}{2} [\mathbf{J}^f(\mathbf{y})^{-1} \mathbf{D}^f(\mathbf{y}) \mathbf{J}^f(\mathbf{y})^{-T}]_{ff'} \quad (4.68)$$

$$\frac{1}{\Omega} \left\langle \boldsymbol{\epsilon}_f^x \mathcal{W}_f^{-1} \frac{\partial}{\partial \boldsymbol{\epsilon}_{f'}^x} \right\rangle_\pi = -\mathbf{J}_{ff'}^f(\mathbf{y})^{-1} \quad (4.69)$$

where the matrices $\mathbf{J}^f(\mathbf{y})$, $\mathbf{D}^f(\mathbf{y})$ are defined as:

$$\mathbf{J}_{ff'}^f(\mathbf{y}) = \sum_r S_{fr} \frac{\partial}{\partial [x_{f'}]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \quad (4.70)$$

$$\mathbf{D}_{ff'}^f(\mathbf{y}) = \sum_r S_{fr} S_{f'r} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \quad (4.71)$$

The external fluctuations operator finally becomes:

$$\begin{aligned} \langle \mathcal{W}_{\text{ext}} \rangle_\pi &= \sum_{s,s'} \left(\frac{\partial}{\partial y_s} \delta_{ss'} \sum_f \mathbf{H}_{s'f}^{sf}(\mathbf{y}) - \sum_f \frac{\partial^2}{\partial y_s \partial y_{s'}} \mathbf{J}_{sf}^{sf}(\mathbf{y}) \right) \cdot \\ &\cdot \sum_{f'} \left(-\frac{1}{2} [\mathbf{J}^f(\mathbf{y})^{-1} \mathbf{D}^f(\mathbf{y}) \mathbf{J}^f(\mathbf{y})^{-T}]_{ff'} \mathbf{J}_{f's'}^{sf}(\mathbf{y})^T + \mathbf{J}_{ff'}^f(\mathbf{y})^{-1} \mathbf{D}_{f's'}^{fs}(\mathbf{y}) \right) \end{aligned} \quad (4.72)$$

where we have introduced the matrices $\mathbf{J}^{sf}(\mathbf{y})$, $\mathbf{H}^{sf}(\mathbf{y})$, $\mathbf{D}^{fs}(\mathbf{y})$ with elements:

$$\mathbf{J}_{sf}^{sf}(\mathbf{y}) = \sum_r S_{sr} \frac{\partial}{\partial [x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \quad (4.73)$$

$$\mathbf{H}_{sf}^{sf}(\mathbf{y}) = \sum_r S_{sr} \frac{\partial^2}{\partial y_s \partial [x_f]} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \quad (4.74)$$

$$\mathbf{D}_{fs}^{fs}(\mathbf{y}) = \sum_r S_{fr} S_{sr} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \quad (4.75)$$

Combining (4.61) and (4.72) we obtain the expansion of the slow-scale FPE operator $\mathcal{W}_s = \langle \mathcal{W}_{\text{int}} \rangle_\pi + \langle \mathcal{W}_{\text{ext}} \rangle_\pi$:

$$\begin{aligned} \mathcal{W}_s = & - \sum_s \frac{\partial}{\partial y_s} \left(\sum_r S_{sr} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) + \sum_{f,f'} \mathbf{H}_{sf}^{sf}(\mathbf{y}) \mathbf{J}_{ff'}^f(\mathbf{y})^{-1} \left(\frac{1}{2} [\mathbf{D}^f(\mathbf{y}) \mathbf{J}^f(\mathbf{y})^{-\text{T}}]_{ff'} \mathbf{J}_{f's}^{sf}(\mathbf{y})^{\text{T}} - \mathbf{D}_{f's}^{fs}(\mathbf{y}) \right) \right) \\ & + \sum_{s,s'} \frac{\partial^2}{\partial y_s \partial y_{s'}} \left(\frac{1}{2} \mathbf{D}_{ss'}^s(\mathbf{y}) + \sum_{f,f'} \mathbf{J}_{sf}^{sf}(\mathbf{y}) \mathbf{J}_{ff'}^f(\mathbf{y})^{-1} \left(\frac{1}{2} [\mathbf{D}^f(\mathbf{y}) \mathbf{J}^f(\mathbf{y})^{-\text{T}}]_{ff'} \mathbf{J}_{f's'}^{sf}(\mathbf{y})^{\text{T}} - \mathbf{D}_{f's'}^{fs}(\mathbf{y}) \right) \right) \end{aligned} \quad (4.76)$$

where we have also introduced the matrix $\mathbf{D}^s(\mathbf{y})$, with elements:

$$\mathbf{D}_{ss'}^s(\mathbf{y}) = \sum_r S_{sr} S_{s'r} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) \quad (4.77)$$

We also note that $\sum_{f,f'} \mathbf{J}_{sf}^{sf}(\mathbf{y}) \mathbf{J}_{ff'}^f(\mathbf{y})^{-1} \mathbf{D}_{f's'}^{fs}(\mathbf{y}) = \sum_r S_{sr} S_{s'r} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) = \mathbf{D}_{ss'}^s(\mathbf{y})$, so we simplify the diffusion term in (4.76):

$$\begin{aligned} \mathcal{W}_s = & - \sum_s \frac{\partial}{\partial y_s} \left(\sum_r S_{sr} w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) + \sum_{f,f'} \mathbf{H}_{sf}^{sf}(\mathbf{y}) \mathbf{J}_{ff'}^f(\mathbf{y})^{-1} \left(\frac{1}{2} [\mathbf{D}^f(\mathbf{y}) \mathbf{J}^f(\mathbf{y})^{-\text{T}}]_{ff'} \mathbf{J}_{f's}^{sf}(\mathbf{y})^{\text{T}} - \mathbf{D}_{f's}^{fs}(\mathbf{y}) \right) \right) \\ & + \sum_{s,s'} \frac{\partial^2}{\partial y_s \partial y_{s'}} \left(-\frac{1}{2} \mathbf{D}_{ss'}^s(\mathbf{y}) + \sum_{f,f'} \frac{1}{2} \mathbf{J}_{sf}^{sf}(\mathbf{y}) [\mathbf{J}^f(\mathbf{y})^{-1} \mathbf{D}^f(\mathbf{y}) \mathbf{J}^f(\mathbf{y})^{-\text{T}}]_{ff'} \mathbf{J}_{f's'}^{sf}(\mathbf{y})^{\text{T}} \right) \end{aligned} \quad (4.78)$$

We have therefore obtained a final, tractable form of the slow-scale Fokker-Planck operator $\mathcal{W}_s$. We now use this operator to derive the slow-scale chemical Langevin equation.

4.2.5 The slow-scale chemical Langevin equation

In the previous section, we derived the slow-scale FPE (4.44)

$$\frac{\partial}{\partial \tau} \rho(\mathbf{y}, \tau) = \mathcal{W}_s \rho(\mathbf{y}, \tau) \quad (4.79)$$

and applied a conditional linear noise approximation on the fast sub-system (4.2.4), to obtain an expression for the operator $\mathcal{W}_s$. This is given, in matrix notation, by:

$$\mathcal{W}_s = - \sum_s \frac{\partial}{\partial y_s} \boldsymbol{\mu}_s(\mathbf{y}) + \frac{1}{2} \sum_{s,s'} \frac{\partial^2}{\partial y_s \partial y_{s'}} (\boldsymbol{\sigma} \boldsymbol{\sigma}^T)_{ss'}(\mathbf{y}) \quad (4.80a)$$

$$\boldsymbol{\mu}(\mathbf{y}) = \underbrace{\mathbf{S}^s \boldsymbol{\alpha}(\mathbf{y})}_{\text{averaged dynamics}} + \underbrace{\text{diag} \left(\mathbf{H}^{sf}(\mathbf{y}) \mathbf{J}^f(\mathbf{y})^{-1} \left(\frac{1}{2} \mathbf{D}^f(\mathbf{y}) \mathbf{J}^f(\mathbf{y})^{-T} \mathbf{J}^{sf}(\mathbf{y})^T - \mathbf{D}^{fs}(\mathbf{y}) \right) \right)}_{\text{external fluctuations}} \quad (4.80b)$$

$$(\boldsymbol{\sigma} \boldsymbol{\sigma}^T)(\mathbf{y}) = \underbrace{-\mathbf{D}^s(\mathbf{y})}_{\text{averaged dynamics}} + \underbrace{\mathbf{J}^{sf}(\mathbf{y}) \mathbf{J}^f(\mathbf{y})^{-1} \mathbf{D}^f(\mathbf{y}) \mathbf{J}^f(\mathbf{y})^{-T} \mathbf{J}^{sf}(\mathbf{y})^T}_{\text{external fluctuations}} \quad (4.80c)$$

where $\boldsymbol{\alpha}(\mathbf{y})$ denotes the vector of averaged propensities $a_r(\mathbf{y}) = w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y})$. Interestingly, we observe that both the drift and the diffusion involve external fluctuations contributions. This shows that the derived reduced description effectively captures external noise induced by the fast process in the slow sub-system.

To derive the CLE associated to the above FPE, we note that the diffusion can be factorised as:

$$(\boldsymbol{\sigma} \boldsymbol{\sigma}^T)(\mathbf{y}) = \left(\mathbf{S}^s - \mathbf{J}^{sf}(\mathbf{y}) \mathbf{J}^f(\mathbf{y})^{-1} \mathbf{S}^f \right) \text{diag}(\boldsymbol{\alpha}(\mathbf{y})) \left(\mathbf{S}^s - \mathbf{J}^{sf}(\mathbf{y}) \mathbf{J}^f(\mathbf{y})^{-1} \mathbf{S}^f \right)^T \quad (4.81)$$

and hence

$$\boldsymbol{\sigma}(\mathbf{y}) = \left(\mathbf{S}^s - \mathbf{J}^{sf}(\mathbf{y}) \mathbf{J}^f(\mathbf{y})^{-1} \mathbf{S}^f \right) \text{diag} \left(\sqrt{\boldsymbol{\alpha}(\mathbf{y})} \right) \quad (4.82)$$

The **slow-scale CLE** describing the evolution of the slow stochastic process is

$$d\mathbf{Y}_t = \boldsymbol{\mu}(\mathbf{Y}_t) dt + \boldsymbol{\sigma}(\mathbf{Y}_t) d\mathbf{V}_t \quad (4.83)$$

where $\mathbf{V}_t$ is an R -dimensional Wiener process.

We now observe that the averaged dynamics term $\mathbf{S}^s \boldsymbol{\alpha}(\mathbf{y})$ of the ssCLE drift matches the drift of the hCLE, as the averaged propensities $a_r(\mathbf{y}) = w_r(\Omega[\mathbf{x} | \mathbf{y}], \mathbf{y}) = \tilde{w}_r(\mathbf{y})$ were obtained from the conditional LNA of the fast sub-system, using the fast steady-state concentration $[\mathbf{x} | \mathbf{y}]$ given by the QSSA. We have therefore shown that the drift of the ssCLE, given by (4.80b),

proposes a correction to the drift of the heuristic CLE:

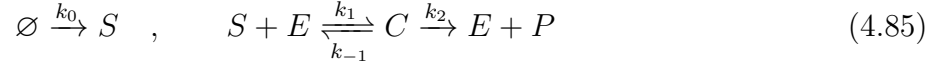
$$\boldsymbol{\mu}(\mathbf{y}) = \underbrace{\mathbf{S}^s \boldsymbol{\alpha}(\mathbf{y})}_{\text{intrinsic drift}} + \underbrace{\boldsymbol{\xi}(\mathbf{y})}_{\text{external fluctuations}} = \underbrace{\mathbf{S}^s \tilde{\mathbf{w}}(\mathbf{y})}_{\text{hCLE drift}} + \underbrace{\boldsymbol{\xi}(\mathbf{y})}_{\text{ssCLE correction}} \quad (4.84)$$

In the following section, we demonstrate through an example that this correction to the drift is crucial for the accuracy of the two stochastic reduction methods.

4.3 Application: Michaelis-Menten enzyme kinetics

In this section, we demonstrate the effectiveness of our approximation method through an example, the Michaelis-Menten enzyme catalysis mechanism, a system that has been widely studied in evaluating multiscale deterministic and stochastic approximations [289, 303, 319].

The mechanism is encoded in the following reaction network:



and describes the production of a substrate S that binds reversibly to an enzyme E to form a complex C , from which a product P (e.g. protein) is produced and the original enzyme is regenerated. In this example, it is commonly assumed that E, C have short characteristic lifetimes compared to S, P that have longer lifetimes, and hence we will consider E, C to be fast and S, P to be slow species. The fast and slow stoichiometry of the network is

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}^f \\ \mathbf{S}^s \end{pmatrix} = \begin{pmatrix} 0 & -1 & 1 & 1 \\ 0 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (4.86)$$

where the rows $i = 1, 2, 3, 4$ of $\mathbf{S}$ correspond to the stoichiometry of the species E, C, S, P , respectively. Denoting with $\phi_E, \phi_C, \phi_S, \phi_P$ the macroscopic concentrations of the species, the macroscopic description of the system is given by the rate ODEs:

$$\frac{d\phi_E(t)}{dt} = -k_1\phi_E(t)\phi_S(t) + (k_{-1} + k_2)(E_0 - \phi_E(t)) \quad (4.87a)$$

$$\frac{d\phi_C(t)}{dt} = -\frac{d\phi_E(t)}{dt} \quad (4.87b)$$

$$\frac{d\phi_S(t)}{dt} = k_0 - k_1\phi_E(t)\phi_S(t) + k_{-1}(E_0 - \phi_E(t)) \quad (4.87c)$$

$$\frac{d\phi_P(t)}{dt} = k_2(E_0 - \phi_E(t)) \quad (4.87d)$$

We note that the system involves the conservation laws $E_0 = \phi_E(t) + \phi_C(t)$ and $C_0 = \phi_S(t) + \phi_C(t) + \phi_P(t)$. Therefore, it is sufficient to use the enzyme molecule states $x = x_E$ to describe the fast sub-system, and the substrate molecule states $y = y_S$ to describe the slow sub-system.

The propensities of the network are:

$$w_0(x, y) = \Omega k_0, \quad w_1(x, y) = \frac{k_1}{\Omega} x y, \quad w_{-1}(x, y) = k_{-1}(e_0 - x), \quad w_2(x, y) = k_2(e_0 - x) \quad (4.88)$$

where Ω denotes the volume of the system and $e_0 = \Omega E_0$ is the total enzyme molecule number. In what follows, to simplify notation, we set $\Omega = 1$, and we will use the total enzyme number $E_0 = e_0$ as a proxy to vary the system size. Accordingly, in the numerical examples shown in the following section, we rescale the reaction rate constants that depend on system size: $k_0 \rightarrow E_0 k_0$ and $k_1 \rightarrow k_1/E_0$.

We now consider the diffusion approximation of the CME describing the Michaelis-Menten system. The coupled system of CLEs describing the enzyme (fast) process $(X_t)_{t \geq 0}$ and the substrate (slow) process $(Y_t)_{t \geq 0}$ is:

$$\begin{aligned} dX_t = & -(k_1 X_t Y_t - (k_{-1} + k_2)(E_0 - X_t)) dt - \sqrt{k_1 X_t Y_t} dV_t^{(2)} + \sqrt{k_{-1}(E_0 - X_t)} dV_t^{(3)} \\ & + \sqrt{k_2(E_0 - X_t)} dV_t^{(4)} \end{aligned} \quad (4.89a)$$

$$dY_t = -(k_1 X_t Y_t - k_0 - k_{-1}(E_0 - X_t)) dt + \sqrt{k_0} dV_t^{(1)} - \sqrt{k_1 X_t Y_t} dV_t^{(2)} + \sqrt{k_{-1}(E_0 - X_t)} dV_t^{(3)} \quad (4.89b)$$

where $V_t^{(r)}, r = 1, \dots, 4$ are independent standard Wiener processes corresponding to the 4 respective reactions.

Reduced Michaelis-Menten system

Under the timescale separation assumption, for any state $y \in \mathcal{Y}$ of the slow process, the fast process rapidly reaches a quasi-stationary distribution, conditional on the slow state. Using the conditional LNA of the fast sub-system, that we describe in 4.2.4, we approximate the stationary mean of the fast process by the steady-state solution of the corresponding fast ODE (4.87a). We find that the steady-state concentration of the enzyme ϕ_E , given the substrate molecular state y , is given by:

$$\phi_E \mid y = \frac{K_M E_0}{K_M + y} \quad (4.90)$$

where we denote with $K_M = \frac{k_{-1} + k_2}{k_1}$ the Michaelis-Menten constant. Therefore, we approximate the (conditional) stationary mean of the fast process by:

$$\langle x \mid y \rangle_\pi \approx \Omega \phi_E \mid y = \frac{K_M E_0}{K_M + y} \quad (4.91)$$

We can now evaluate the averaged propensities $\alpha_r(y) = w_r(\langle x \mid y \rangle_\pi, y)$, where the fast molecular state x in the original propensities $w_r(x, y)$ is approximated by its stationary mean $\langle x \mid y \rangle_\pi$:

$$a_0(y) = k_0, \quad a_1(y) = \frac{E_0 k_1 K_M y}{K_M + y}, \quad a_{-1}(y) = \frac{E_0 k_{-1} y}{K_M + y}, \quad a_2(y) = \frac{E_0 k_2 y}{K_M + y} \quad (4.92)$$

We now derive the reduced CLE (ssCLE) for the substrate process, by evaluating its drift and diffusion using the formulas (4.80b) and (4.82). The ssCLE reads

$$dY_t = \mu(Y_t) dt + \sigma(Y_t) dV_t \quad (4.93)$$

with

$$\mu(y) = \underbrace{k_0 - k_2 \Lambda(y)}_{\text{intrinsic drift}} - \underbrace{\frac{k_{-1} \Lambda(y)}{K_M + y} - \frac{k_2 K_M \Lambda(y)}{(K_M + y)^2}}_{\text{external fluctuations}} \quad (4.94a)$$

$$\sigma(y) = \text{diag} \left(\sqrt{k_0}, -\frac{k_2 K_M \sqrt{\Lambda(y)}}{(K_M + y) \sqrt{k_{-1} + k_2}}, \frac{k_2 K_M \sqrt{k_{-1} \Lambda(y)}}{(K_M + y)(k_{-1} + k_2)}, -\frac{((k_{-1} + k_2)y + K_M k_{-1}) \sqrt{k_2 \Lambda(y)}}{(K_M + y)(k_{-1} + k_2)} \right) \quad (4.94b)$$

where we have defined $\Lambda(y) = \frac{E_0 y}{K_M + y}$. We observe that the drift $\mu(y)$ has two components: the *intrinsic drift*, capturing the averaged dynamics of the slow process, and the *external fluctuations* contribution, capturing fluctuations in the dynamics induced by the fast sub-system. Accordingly, the diffusion $\sigma(y)$ also captures contributions from both intrinsic and external dynamics.

Evaluation of the approximation

To evaluate the effectiveness of our stochastic reduction method, we compare the accuracy of the ssCLE in predicting the exact stationary distribution of substrate levels to the prediction of the heuristic stochastic approximation (see 4.2.2). As the Gillespie stochastic simulation

algorithm (SSA) gives access to the exact marginal stationary solutions of the CME, we use it as ground truth to benchmark the two approximation methods against simulations. To obtain the heuristic CLE for the Michaelis-Menten system, we apply the QSSA to the full system of ODEs (4.87), and we substitute the steady-state of the enzyme concentration (4.90) in the substrate ODE (4.87c) to get:

$$\frac{d\phi_S(t)}{dt} \approx k_0 - \frac{k_2 E_0 \phi_S(t)}{K_M + \phi_S(t)} \quad (4.95)$$

We now use the QSSA to construct a CLE with heuristic propensities, that we will refer to as hCLE (heuristic CLE), governing the stochastic substrate process:

$$dY_t = \left(k_0 - \frac{k_2 E_0 Y_t}{K_M + Y_t} \right) dt + \sqrt{k_0} d\tilde{V}_t^{(1)} - \sqrt{\frac{k_2 E_0 Y_t}{K_M + Y_t}} d\tilde{V}_t^{(2)} \quad (4.96)$$

We observe that the hCLE (4.96) represents a coarse-grained reaction network of the form



that approximates the full reaction network (4.85). The reduced network describes the production of the substrate S with rate k_0 and its conversion to a downstream product P with rate

$$\tilde{k} = \frac{k_2 E_0 y}{K_M + y} \quad (4.98)$$

The rate $\tilde{k}$ summarises the effect of the (fast) enzyme process on the slow sub-system, during its binding with the substrate. However, this representation is limited and does not fully capture the stochasticity induced by the fast sub-system. The ssCLE (4.93) proposes a correction to the drift of the hCLE (4.96), that captures additional external contributions from the fast sub-system:

$$\underbrace{\mu(y)}_{\text{ssCLE drift}} = \underbrace{k_0}_{\text{substrate input}} - \underbrace{k_2 \frac{E_0 y}{K_M + y}}_{\text{binding}} - \underbrace{\frac{k_{-1} E_0 y}{(K_M + y)^2}}_{\text{unbinding}} - \underbrace{\frac{k_2 K_M E_0 y}{(K_M + y)^3}}_{\text{catalysis}} \quad (4.99)$$

Equation (4.99) shows that the hCLE only captures contributions from the first two reactions, whereas the ssCLE captures effects of the full reaction network. As an example, even though the catalysis reaction $C \xrightarrow{k_2} E + P$ does not directly affect the (slow) substrate dynamics, it

affects the (fast) enzyme dynamics, which in turn influences the slow sub-system. This indirect effect is expressed by the 4th component of the ssCLE's drift. Similarly, the diffusion terms of the ssCLE represent the noise effects induced by all four reactions of the network. These observations suggest that the ssCLE approximation is more reliable in capturing the overall stochasticity of the full reaction network.

We next demonstrate that the ssCLE is more accurate than the hCLE in a numerical example. We choose the model parameter values to be $k_0 = 8E_0$, $k_{-1} = 0.5$, $k_2 = 10$, $K_M = 10$, $k_1 = (k_{-1} + k_2)/K_M E_0 = 1.05/E_0$ and we use the total enzyme number E_0 as a proxy for system size, ranging in $\{1, 5, 10, 20\}$. The parameter values are chosen such that the timescale separation assumption is satisfied. We validate this by showing that the deterministic QSSA (4.95) exactly matches the solution of the exact substrate concentration ODE (4.87c) (Figure 4.2).

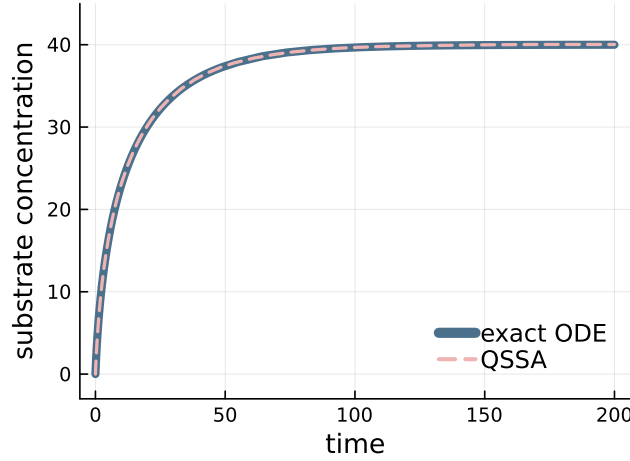


Figure 4.2: **Timescale separation in the system.** The exact substrate ODE (4.87c) and the reduced ODE obtained by the QSSA (4.95) are numerically solved to validate the timescale separation assumption for the chosen parameter regime.

To compare the predictive accuracy of the ssCLE and the hCLE, we evaluate probability densities of substrate numbers in steady state, given by the stationary solutions of their respective Fokker-Planck equations [293]:

$$p_{ss}(y) = \frac{\mathcal{C}}{(\boldsymbol{\sigma}\boldsymbol{\sigma})^T(y)} \exp\left(2 \int_0^y \frac{\mu(u)}{(\boldsymbol{\sigma}\boldsymbol{\sigma})^T(u)} du\right) \quad (4.100)$$

where $\mathcal{C}$ is a normalisation constant such that $\int_0^{y_{\max}} p_{ss}(y) dy = 1$ and $\mu(y), \boldsymbol{\sigma}\boldsymbol{\sigma}^T(y)$ are the drift and diffusion terms of the Fokker-Planck equations. We compare the evaluated stationary densities with the marginal probability distributions obtained from Gillespie simulation of the

full Michaelis-Menten network (4.85). We find that the ssCLE perfectly matches the simulated distributions across all system sizes, while the hCLE fails to capture these distributions (Figure 4.3 (A-D)). In particular, we observe that the hCLE overestimates the dispersion of the substrate distribution, while the ssCLE accurately estimates the full probability distribution. Notably, the ssCLE is accurate even at the small system size limit, in the presence of only $E_0 = 1$ enzyme molecule (Figure 4.3 (A)).

We show that both the ssCLE and hCLE predict the steady-state mean substrate concentration accurately, with an absolute relative error smaller than 2% compared to the SSA, that becomes zero as the system size increases (Figure 4.4 (A)). However, the hCLE overestimates the noise in substrate abundance. We observe a consistent deviation (approximately 8.5%) in the hCLE’s prediction of the coefficient of variation across system sizes (Figure 4.4 (B)) and an increasing deviation (higher than 18%) in its prediction of the Fano Factor (Figure 4.4 (C)), compared to stochastic simulations. On the other hand, the ssCLE correctly predicts both noise statistics in substrate concentration (Figures 4.4 (B-C)). Overall, these results validate our expectation, that our approximation method accurately captures the stochasticity of the full reaction network, making it a more reliable approximation of the chemical master equation compared to the heuristic method.

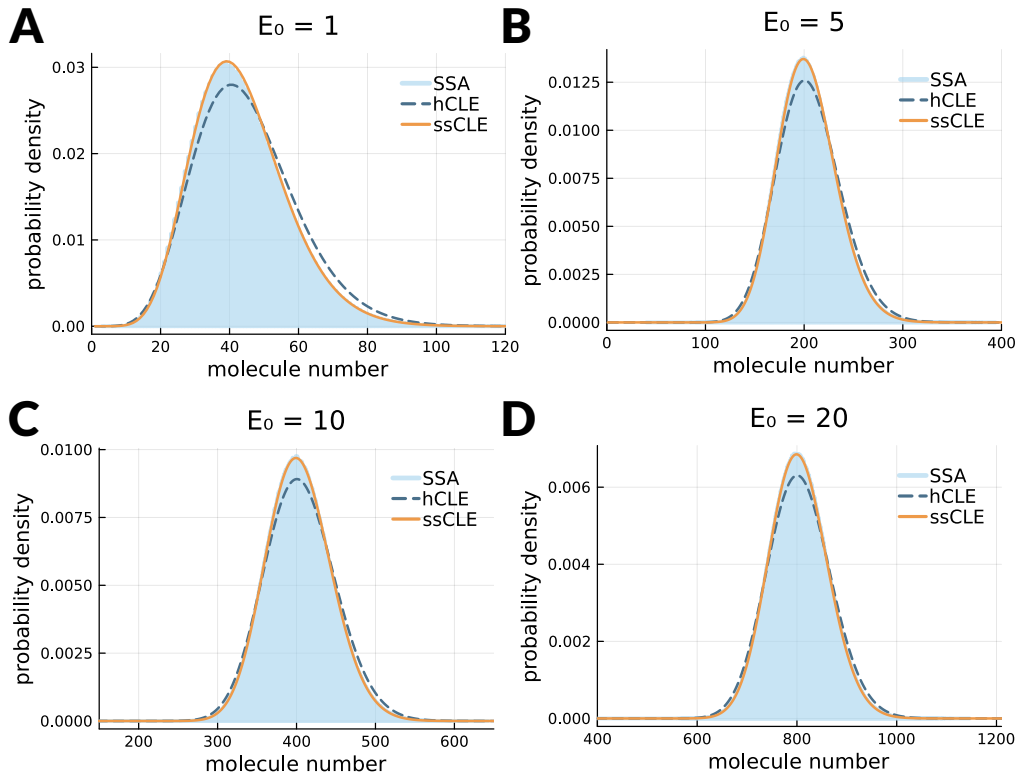


Figure 4.3: **Steady-state probability distributions of the slow species molecule numbers.** Stationary probability density of substrate numbers, as predicted by Gillespie stochastic simulations (SSA), the heuristic CLE (hCLE) and the slow-scale CLE (ssCLE), across different system sizes. The system size is represented by the total enzyme number, with values (A) $E_0 = 1$, (B) $E_0 = 5$, (C) $E_0 = 10$ and (D) $E_0 = 20$. The SSA distribution corresponds to relative frequencies of substrate states, obtained from a long stochastic trajectory of the full Michaelis-Menten network (4.85).

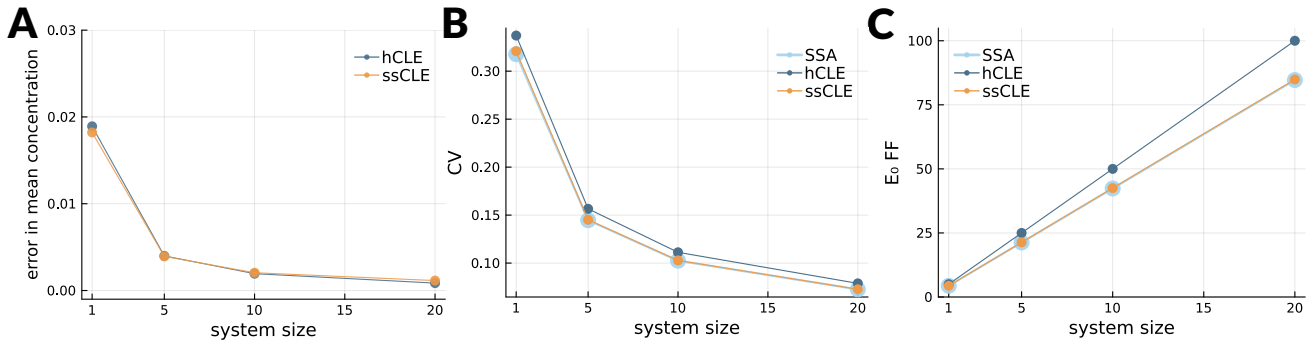


Figure 4.4: **Comparison of steady-state mean and noise statistics of the slow species abundance.** (A) Absolute relative error between the hCLE / ssCLE predictions and the Gillespie simulation (SSA) predictions of the mean substrate concentration across system sizes. (B-C) Predictions of the noise in substrate concentration across system sizes made by the SSA, the hCLE and the ssCLE, measured by the coefficient of variation (CV) and the Fano Factor (FF), respectively. In (C), the Fano factor is multiplied by the system size E_0 to get a non-dimensional measure of the concentration noise.

4.4 Discussion

In this chapter, we developed a novel approximation method for stochastic reaction networks, evolving under distinct fast and slow timescales. The method provides an effective and accurate diffusion approximation of the chemical master equation, the slow-scale CLE, that governs the dynamics of the slowly evolving sub-network. Our theory can be applied to obtain an approximate stochastic description of any reaction network satisfying two minimal conditions: (i) The slow sub-network involves species present in not very low abundances (in the order of at least tens of molecules) and (ii) a clear partitioning can be imposed on the characteristic timescales of the involved biochemical species. The timescale separation condition is satisfied if the transients of the fast species decay over a short timescale, while the transients of the slow species decay over a relatively longer timescale. In some cases, these decay rates are known from experimental measurements of relevant biochemical processes, such as *e.g.* the characteristic half-lives of mRNA and protein during gene expression [300, 321].

We demonstrated that our approximation method provides a faithful reduced representation of stochastic reaction networks. The reduced CLE description is decoupled from explicit dependence on fast variables, yet it captures contributions from all reactions of the original network. The drift of the ssCLE captures contributions from fluctuations in the fast variables, including effects from reactions that only change the fast sub-system. These drift terms are crucial for accurately estimating the long-term average dynamics of the full stochastic system. Similarly, the ssCLE diffusion captures stochastic effects induced by the entire set of reactions, thereby reliably approximating the noise in the original system. In contrast, we showed that the heuristic approximation, that utilises non-elementary propensities of the QSSA to form a CLE for the slow sub-network, does not provide a reliable representation of stochastic reaction networks. The heuristic CLE (hCLE) only explains noise induced by reactions changing the slow species, neglecting stochastic effects induced by the eliminated fast sub-system. Therefore, our approximation method provides a more reliable extension of the deterministic QSSA to the CLE description of stochastic systems.

Several methods have proposed alternative stochastic versions of the QSSA, such as the stochastic total QSSA (stQSSA) [322, 323] and the stochastic low-state QSSA (slQSSA) [324], by constructing alternative non-elementary reaction propensities. These approaches eliminate re-

actions of the full network, leading to low-dimensional descriptions and enabling stochastic simulations with reduced computational cost. However, the validity of these methods depends on the architecture of the reaction network [322, 324]. Importantly, we have shown that, in general, all reactions contribute to the noise of the slow sub-system. Therefore, eliminating reactions can lead to inaccuracies in capturing the overall stochasticity of the original reaction network. Furthermore, a few reduction methods have been developed specifically for the CLE description of biochemical networks, that rely on separating the timescales of both reactions and species [311, 312, 325]. Our approach generalises this condition by only assuming partitioning of the species’ timescales, making it more broadly applicable. Overall, our method provides a versatile and improved alternative to existing multiscale stochastic reduction methods.

We demonstrated the efficiency of our method in approximating the Michaelis-Menten enzyme kinetics network. We showed that the ssCLE accurately predicts the exact stationary probability distribution and noise statistics of the slow species abundance across different system sizes, as given by Gillespie simulations. We explored small system sizes of only a few tens of slow species’ molecules, to larger sizes of up to a thousand molecules. Notably, the ssCLE remains highly accurate in the small system size regime. On the other hand, the hCLE overestimates the noise levels and fails to match the probability distributions in all cases, including the large system size limit. This provides evidence that the corrections to the drift and diffusion of the hCLE, as proposed by the ssCLE, are critical for approximating probability distributions of stochastic reaction networks.

Our approach can be broadly applied to a variety of biochemical systems with heterogeneous timescales, including gene regulatory networks. An application of particular interest is genetic feedback loops, including positive and negative autoregulation [326, 327], in which the transcription of a gene is either activated or repressed by its own protein products. Autoregulatory feedback loops are widespread mechanisms that suppress gene expression noise, ensuring stability and contributing to the regulation of essential biological processes [328, 329]. As an example, positive feedback mechanisms, involving cooperative protein binding at the gene promoter [327], are known to create genetic switches where proteins are present in two stable steady states [330]. This bistability in protein levels is crucial for cellular decision making, as it can determine the cell’s response to external signals or its commitment to a specific fate during differentiation [331]. Another key example is negative autoregulation mechanisms, known to produce sus-

tained transcriptional oscillations [332–334]. Oscillations underlie biological rhythms, such as circadian clocks [335, 336], which are essential for cellular physiology [337]. Therefore, it would be interesting to test our theory in bistable or oscillatory networks, as the CLE is better tool for handling these systems compared to deterministic methods and the linear noise approximation [289, 334].

One limitation of our method is that it relies on a continuous approximation of the chemical master equation. This poses a constraint in studying reaction networks in which the slow species of interest are typically present in very low copy numbers [68]. Yet, the slow-scale CLE is more well-suited than the slow-scale LNA [289] for approximating distributions in low copy-number regimes [297]. A potential extension of our theory would be to derive a rigorous approximation of the chemical master equation under timescale separation, which remains an open problem. An additional limitation of our method is that it utilises the LNA to obtain tractable expressions for the fast variable contributions. However, the LNA may provide poor estimations of the noise in systems involving high-order reactions [297, 338]. This could compromise the accuracy of the ssCLE in biochemical networks with highly non-linear reactions, where the slow species interact with multiple fast species. Therefore, a potential extension of our framework would be to introduce higher order terms of the system size expansion for approximating the fast variables [278].

In summary, we demonstrated that the slow-scale chemical Langevin equation enables efficient and accurate estimation of the statistics of stochastic reaction networks. Our method addresses some of the constraints in modelling complex biochemical systems, by providing a tractable alternative to the chemical master equation and reducing the computational cost associated with stochastic simulations. Therefore, we anticipate the ssCLE to become a valuable tool for stochastic analyses of biological systems.

Chapter 5

Discussion

5.1 Summary of contributions

In this thesis, we have developed methodologies for data-driven mathematical modelling of gene regulation. The contribution of this thesis is two-fold. First, it develops frameworks that integrate stochastic modelling with Bayesian inference and machine learning to enable the discovery of gene regulatory mechanisms from single-cell gene expression data. Second, it introduces a novel model reduction method that enables efficient stochastic analysis of complex biochemical systems, including gene regulatory networks.

The main contribution of Chapter 2 is a modelling framework that reveals global properties of transcriptional regulation and its coupling to the cell cycle from time-resolved single-cell data. The complexity and technical limitations characterising single-cell RNA sequencing pose significant challenges for inferring mechanistic models of gene expression from data. Three key aspects of our method allowed to tackle these challenges. First, we constructed a stochastic model that exploits temporal information of cell cycle dynamics and RNA labelling to identify various mechanisms underlying gene expression control. To infer these mechanisms from data, we introduced a scalable Bayesian computation framework, that enabled efficient sampling of the full posterior distribution over models and their parameters. Modelling technical noise sources of sequencing data, enabled reliable parameter inference and model selection. This methodology resulted in accurate genome-wide quantification of absolute burst frequency, burst size and degradation rate throughout the cell cycle. Our inference revealed that most genes scale

transcription rates with cell size and identified a wave-like pattern of transcriptional regulation across the cell cycle. Finally, our stochastic model allowed us to disentangle technical and biological variation in the data and quantify true sources of transcriptional noise. Our method thus provides novel insights into the coordination of transcription with cell size and cell cycle on a genome-wide scale.

The main contribution of Chapter 3 is a modelling framework for the mechanistic investigation of gene-gene correlations in single-cell transcriptomics. Transcriptional covariation revealed by single-cell data serves as a valuable tool for exploring genome-wide regulatory relationships. While multiple noise sources are known to contribute to gene-gene covariation, limited emphasis is given in dissecting these sources to identify true gene regulation. To address this, we built stochastic models of gene co-expression driven by different modes of regulatory interaction, capturing distinct biological and technical noise contributions, and integrating cell cycle progression and metabolic labelling timepoints. We employed a deep learning approach that harnesses synthetic data generated from these models to uncover links between gene co-expression patterns and underlying modes of regulation. Our framework predicts whether observed gene co-expression statistics can be purely explained by cell cycle covariation and technical noise, or whether they stem from direct gene interaction, or co-regulation, including additional upstream noise. We demonstrated that dynamical gene-gene correlations provide rich signatures for distinguishing these modes of regulation. Our analysis revealed that, while cell cycle covariation and technical variability contribute significantly to observed gene-gene correlations, most correlations are associated with co-regulation, indicative of additional extrinsic noise present in the data. Additionally, we demonstrated that our model successfully recovers known regulatory relationships. Our method thus provides a systematic approach for interpreting gene-gene covariation in single-cell data and highlights the potential of time-resolved correlations for inferring gene regulatory interactions.

The main contribution of Chapter 4 is a method for approximating stochastic reaction networks that evolve under multiple timescales. Modelling complex biochemical networks through the chemical master equation (CME) and the Gillespie stochastic simulation algorithm (SSA) often presents analytical and computational challenges, hindering efficient stochastic analysis. A common approach to mitigate these challenges is to exploit the timescale separation property of biochemical systems to obtain low-dimensional models, that solely describe dynam-

ics occurring over long timescales and average-out the explicit dependence on short-timescale dynamics. We introduced a novel stochastic model reduction tool, the slow-scale chemical Langevin equation (ssCLE), that overcomes limitations of existing approximation methods and is more broadly applicable. Many existing stochastic model reduction approaches utilise non-elementary propensities, or heuristic propensities given by deterministic approximations, to eliminate reactions and obtain coarse-grained networks [318, 324]. However, we showed that all reactions of the original reaction network can contribute to the noise of the reduced system, and hence these coarse-grained descriptions can fail to explain the stochasticity of the full network. The ssCLE describes the exact dynamics of the sub-system of interest and the stochastic effects from all reactions changing the eliminated sub-system, providing a faithful representation of the original reaction network. Finally, we demonstrated that the ssCLE approximates accurately the Michaelis-Menten enzyme catalysis network, a system of central interest for timescale separation methods. Therefore, our approximation method provides a novel valuable tool for stochastic modelling of biochemical reaction networks.

5.2 Conclusions and future directions

Gene expression is a highly complex, stochastic process, involving multiple regulatory steps [2]. Elucidating this process necessitates comprehensive and efficient computational methods that can extract rich information from biological data. Chapters 2 and 3 demonstrate that methods synergising mechanistic models of stochastic gene expression with Bayesian inference and machine learning, can effectively harness single-cell data and reveal mechanistic principles of gene regulation. Mechanistic models can simulate regulatory processes that generate diverse gene expression patterns, enabling analysis of their quantitative properties. Complex models can be approximated and simplified, as we show in Chapter 4, to gain quantitative insights more efficiently. Bayesian inference provides the means to explore the model space that best represents the true underlying regulatory processes, given the data. Machine learning allows to uncover links between generative processes and observed data, even when statistical inference is impractical. Therefore, the methods developed in this thesis provide a powerful toolset for extracting mechanisms of gene expression regulation from single-cell data.

5.2.1 Inferring complex modes of gene regulation from time-resolved transcriptomics

The findings and methodologies presented in this thesis open up new opportunities for investigating gene regulation through mechanistic modelling and single-cell data. We showed that time-resolved single-cell transcriptomics offers rich signatures for identifying regulatory mechanisms, bypassing limitations of conventional single-cell RNA sequencing. The framework introduced in Chapter 2 can be broadly applied to study transcription dynamics using any time-resolved transcriptomics protocol [129]. Of particular interest is the TEMPOMap protocol reported in [191], that enriches single-cell transcriptomes with pulse-chase labelling timepoints, cell cycle phases and spatial transcriptomic measurements, including nuclear and cytoplasmic mRNA. Our framework can be used to infer complex models of transcription regulation that also account for nuclear mRNA export. Additionally, applying our method to data reported in [193], that integrates pulse labelling conditions with developmental timepoints, could provide valuable insights into the role of transcriptional regulation in development.

The framework developed in Chapter 3, extended independent gene models to models of 2-gene networks for exploring gene-pair regulatory relationships. An additional extension of this framework that can help distinguish direct regulatory interaction from co-regulation, would be to consider 3-gene network models. These models allow to investigate conditional correlations, so that the covariation between two genes is conditioned on the expression of the third gene. This enables testing all possible (co-)regulatory relationships within a 3-gene network. This concept is related to Bayesian networks, that provide a framework for identifying causal interactions from conditional dependencies [226, 339]. Time-resolved transcriptomics would allow to quantify dynamic conditional correlations across labelling timepoints, which can provide rich information. Additionally, incorporating RNA splicing information to explore (conditional) correlations between unspliced mRNA of regulators and spliced mRNA of potential targets, can offer additional temporal resolution [274, 340].

Motivated by the findings of Chapter 3, another important extension of our framework would be to systematically model extrinsic noise that influences gene regulatory relationships globally. Apart from variation in cellular states, global extrinsic noise could represent interactions of the 2-gene or 3-gene network with a larger regulatory network in which it is embedded. Stochastic

model reduction methods, such as the method developed in Chapter 4, can prove very useful for modelling extrinsic noise effects in these gene networks [341].

It can be argued that 2-gene and 3-gene networks remain simplistic models, considering that genes operate within large regulatory networks [156]. Yet, the complexity and sparsity of single-cell RNA sequencing data limit the statistical power required to identify complex networks [245, 342]. The modelling frameworks presented here, provide a principled approach to test this statistical power and the potential to infer complex regulatory interactions.

Furthermore, Chapters 2 and 3 demonstrated that modelling technical noise is critical for reliable inference of regulatory mechanisms from single-cell data. Our modelling frameworks account for technical variability caused by cell-specific transcript capture efficiency and gene-specific metabolic labelling efficiency. However, additional technical noise factors associated with sequencing protocols need to be considered [144, 148]. An important factor is gene length, that is known to cause transcript detection biases and generate spurious gene-gene correlations [248, 249]. In Chapter 2, we showed that gene labelling efficiency correlates with gene length, indicating that gene-gene correlations between unlabelled and labelled mRNA may be impacted by length biases. Therefore, future work involves developing a more detailed technical noise model that controls for transcript detection biases through a gene- and cell-specific capture efficiency.

5.2.2 Mechanistic modelling, Bayesian inference and machine learning for single-cell multi-omics

Advances in single-cell genomics technologies have enabled high-throughput profiling of cellular modalities beyond single-cell RNA sequencing, including epigenomics and proteomics [343–345], as well as spatial epigenomics and transcriptomics [346, 347]. Recently, simultaneous profiling of multiple modalities in the same cell has been made possible with single-cell multi-omics technologies [217, 348–351]. These technologies open new avenues for investigating complex gene regulation in single cells with computational methods. The methodologies presented in this thesis can provide valuable tools for interpretable analyses of single-cell multi-omics data. As shown in Chapter 2, mechanistic stochastic models can naturally integrate multiple modalities and experimental conditions, without requiring statistical integration [215, 352],

to quantitatively explain multiple sources of heterogeneity [353]. A mechanistic modelling approach for multi-omics has recently been proposed, integrating chromatin accessibility with transcription dynamics [354]. Additionally, mathematical models provide a natural means for studying single-cell dynamics across temporal and spatial dimensions revealed by these technologies [216, 355].

To fully exploit the power and scale of single-cell multi-omics, mechanistic modelling must be combined with efficient statistical inference and machine learning methods. As demonstrated in Chapter 2, amortised Bayesian computation offers a highly scalable approach for approximating the posterior over competing modelling hypotheses [103, 107]. Machine learning techniques offer additional versatility, by approximating intractable models [97, 280], enhancing the performance of simulation-based methods [108] and enabling efficient approximate Bayesian inference [103, 356–358]. In summary, the methodologies presented in this thesis promise to drive further advances in computational methods for modelling gene regulation with single-cell genomics.

Bibliography

- [1] Bruce Alberts. *Molecular Biology of the Cell*. 6th edition. W.W. Norton & Company Inc., 2015. DOI: [10.1201/9781315735368](https://doi.org/10.1201/9781315735368).
- [2] Christopher Buccitelli and Matthias Selbach. “mRNAs, proteins and the emerging principles of gene expression control”. *Nature Reviews Genetics* 21 (2020), pp. 630–644. DOI: [10.1038/s41576-020-0258-4](https://doi.org/10.1038/s41576-020-0258-4).
- [3] William J. Blake et al. “Noise in eukaryotic gene expression”. *Nature* 422 (2003), pp. 633–637. DOI: [10.1038/nature01546](https://doi.org/10.1038/nature01546).
- [4] Mukund Thattai and Alexander van Oudenaarden. “Intrinsic noise in gene regulatory networks”. *Proceedings of the National Academy of Sciences* 98 (2001), pp. 8614–8619. DOI: [10.1073/pnas.151588598](https://doi.org/10.1073/pnas.151588598).
- [5] Johan Paulsson. “Summing up the noise in gene networks”. *Nature* 427 (2004), pp. 415–418. DOI: [10.1038/nature02257](https://doi.org/10.1038/nature02257).
- [6] Juan M. Pedraza and Alexander van Oudenaarden. “Noise Propagation in Gene Networks”. *Science* 307 (2005), pp. 1965–1969. DOI: [10.1126/science.1109090](https://doi.org/10.1126/science.1109090).
- [7] H. McAdams Harley and Adam Arkin. “Stochastic mechanisms in gene expression”. *Proceedings of the National Academy of Sciences* 94 (1997), pp. 814–819. DOI: [10.1073/pnas.94.3.814](https://doi.org/10.1073/pnas.94.3.814).
- [8] Michael B. Elowitz et al. “Stochastic Gene Expression in a Single Cell”. *Science* 297 (2002), pp. 1183–1186. DOI: [10.1126/science.1070919](https://doi.org/10.1126/science.1070919).
- [9] Narendra Maheshri and Erin K. O’Shea. “Living with Noisy Genes: How Cells Function Reliably with Inherent Variability in Gene Expression”. *Annual Review of Biophysics and Biomolecular Structure* 36 (2007), pp. 413–434. DOI: [10.1146/annurev.biophys.36.040306.132705](https://doi.org/10.1146/annurev.biophys.36.040306.132705).

- [10] Mads Kærn et al. “Stochasticity in gene expression: from theories to phenotypes”. *Nature Reviews Genetics* 6 (2005), pp. 451–464. DOI: [10.1038/nrg1615](https://doi.org/10.1038/nrg1615).
- [11] Jonathan M. Raser and Erin K. O’Shea. “Noise in Gene Expression: Origins, Consequences, and Control”. *Science* 309 (2005), pp. 2010–2013. DOI: [10.1126/science.1105891](https://doi.org/10.1126/science.1105891).
- [12] Luis López-Maury, Samuel Marguerat, and Jürg Bähler. “Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation”. *Nature Reviews Genetics* 9 (2008), pp. 583–593. DOI: [10.1038/nrg2398](https://doi.org/10.1038/nrg2398).
- [13] Avigdor Eldar and Michael B. Elowitz. “Functional roles for noise in genetic circuits”. *Nature* 467 (2010), pp. 167–173. DOI: [10.1038/nature09326](https://doi.org/10.1038/nature09326).
- [14] P. Thomas, N. Popovic, and R. Grima. “Phenotypic switching in gene regulatory networks”. *Proceedings of the National Academy of Sciences* 111 (2014), pp. 6994–6999. DOI: [10.1073/pnas.1400049111](https://doi.org/10.1073/pnas.1400049111).
- [15] Richard Losick and Claude Desplan. “Stochasticity and Cell Fate”. *Science* 320 (2008), pp. 65–68. DOI: [10.1126/science.1147888](https://doi.org/10.1126/science.1147888).
- [16] Arjun Raj and Alexander van Oudenaarden. “Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences”. *Cell* 135 (2008), pp. 216–226. DOI: [10.1016/j.cell.2008.09.050](https://doi.org/10.1016/j.cell.2008.09.050).
- [17] Gábor Balázsi, Alexander van Oudenaarden, and James J. Collins. “Cellular Decision Making and Biological Noise: From Microbes to Mammals”. *Cell* 144 (2011), pp. 910–925. DOI: [10.1016/j.cell.2011.01.030](https://doi.org/10.1016/j.cell.2011.01.030).
- [18] Orsolya Symmons and Arjun Raj. “What’s Luck Got to Do with It: Single Cells, Multiple Fates, and Biological Nondeterminism”. *Molecular Cell* 62 (2016), pp. 788–802. DOI: <https://doi.org/10.1016/j.molcel.2016.05.023>.
- [19] Peter S. Swain, Michael B. Elowitz, and Eric D. Siggia. “Intrinsic and extrinsic contributions to stochasticity in gene expression”. *Proceedings of the National Academy of Sciences* 99 (2002), pp. 12795–12800. DOI: [10.1073/pnas.162041399](https://doi.org/10.1073/pnas.162041399).
- [20] Andreas Hilfinger and Johan Paulsson. “Separating intrinsic from extrinsic fluctuations in dynamic biological systems”. *Proceedings of the National Academy of Sciences* 108 (2011), pp. 12167–12172. DOI: [10.1073/pnas.1018832108](https://doi.org/10.1073/pnas.1018832108).

- [21] Vahid Shahrezaei and Peter S Swain. “The stochastic nature of biochemical networks”. *Current Opinion in Biotechnology* 19 (2008), pp. 369–374. DOI: [10.1016/j.copbio.2008.06.011](https://doi.org/10.1016/j.copbio.2008.06.011).
- [22] Dmitri Volfson et al. “Origins of extrinsic variability in eukaryotic gene expression”. *Nature* 439 (2005), pp. 861–864. DOI: [10.1038/nature04281](https://doi.org/10.1038/nature04281).
- [23] Ertugrul M. Ozbudak et al. “Regulation of noise in the expression of a single gene”. *Nature Genetics* 31 (2002), pp. 69–73. DOI: [10.1038/ng869](https://doi.org/10.1038/ng869).
- [24] Jonathan M. Raser and Erin K. O’Shea. “Control of Stochasticity in Eukaryotic Gene Expression”. *Science* 304 (2004), pp. 1811–1814. DOI: [10.1126/science.1098641](https://doi.org/10.1126/science.1098641).
- [25] Nitzan Rosenfeld et al. “Gene Regulation at the Single-Cell Level”. *Science* 307 (Mar. 2005), pp. 1962–1965. DOI: [10.1126/science.1106914](https://doi.org/10.1126/science.1106914).
- [26] Alejandro Colman-Lerner et al. “Regulated cell-to-cell variation in a cell-fate decision system”. *Nature* 437 (2005), pp. 699–706. DOI: [10.1038/nature03998](https://doi.org/10.1038/nature03998).
- [27] Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. “Using Gene Expression Noise to Understand Gene Regulation”. *Science* 336 (2012), pp. 183–187. DOI: [10.1126/science.1216379](https://doi.org/10.1126/science.1216379).
- [28] Berend Snijder and Lucas Pelkmans. “Origins of regulated cell-to-cell variability”. *Nature Reviews Molecular Cell Biology* 12 (2011), pp. 119–125. DOI: [10.1038/nrm3044](https://doi.org/10.1038/nrm3044).
- [29] Rutu Rinott, Ariel Jaimovich, and Nir Friedman. “Exploring transcription regulation through cell-to-cell variability”. *Proceedings of the National Academy of Sciences* 108 (2011), pp. 6329–6334. DOI: [10.1073/pnas.1013148108](https://doi.org/10.1073/pnas.1013148108).
- [30] Clive G. Bowsher and Peter S. Swain. “Identifying sources of variation and the flow of information in biochemical networks”. *Proceedings of the National Academy of Sciences* 109 (2012). DOI: [10.1073/pnas.1119407109](https://doi.org/10.1073/pnas.1119407109).
- [31] Ido Golding and Edward C. Cox. “RNA dynamics in live Escherichia coli cells”. *Proceedings of the National Academy of Sciences* 101 (2004), pp. 11310–11315. DOI: [10.1073/pnas.0404443101](https://doi.org/10.1073/pnas.0404443101).
- [32] Susan M Janicki et al. “From Silencing to Gene Expression: Real-Time Analysis in Single Cells”. *Cell* 116 (2004), pp. 683–698. DOI: [10.1016/s0092-8674\(04\)00171-0](https://doi.org/10.1016/s0092-8674(04)00171-0).

- [33] Long Cai, Nir Friedman, and X. Sunney Xie. “Stochastic protein expression in individual cells at the single molecule level”. *Nature* 440 (2006), pp. 358–362. DOI: [10.1038/nature04599](https://doi.org/10.1038/nature04599).
- [34] Diane Longo and Jeff Hasty. “Dynamics of single-cell gene expression”. *Molecular Systems Biology* 2 (2006), p. 64. DOI: [10.1038/msb4100110](https://doi.org/10.1038/msb4100110).
- [35] Benjamin B Kaufmann and Alexander van Oudenaarden. “Stochastic gene expression: from single molecules to the proteome”. *Current Opinion in Genetics & Development* 17 (2007), pp. 107–112. DOI: [10.1016/j.gde.2007.02.007](https://doi.org/10.1016/j.gde.2007.02.007).
- [36] Daniel R Larson. “What do expression dynamics tell us about the mechanism of transcription?” *Current Opinion in Genetics & Development* 21 (2011), pp. 591–599. DOI: [10.1016/j.gde.2011.07.010](https://doi.org/10.1016/j.gde.2011.07.010).
- [37] Gene-Wei Li and X. Sunney Xie. “Central dogma at the single-molecule level in living cells”. *Nature* 475 (2011), pp. 308–315. DOI: [10.1038/nature10315](https://doi.org/10.1038/nature10315).
- [38] Ineke Brouwer and Tineke L Lenstra. “Visualizing transcription: key to understanding gene expression dynamics”. *Current Opinion in Chemical Biology* 51 (2019), pp. 122–129. DOI: [10.1016/j.cbpa.2019.05.031](https://doi.org/10.1016/j.cbpa.2019.05.031).
- [39] Ido Golding et al. “Real-Time Kinetics of Gene Activity in Individual Bacteria”. *Cell* 123 (2005), pp. 1025–1036. DOI: [10.1016/j.cell.2005.09.031](https://doi.org/10.1016/j.cell.2005.09.031).
- [40] Jonathan R. Chubb et al. “Transcriptional Pulsing of a Developmental Gene”. *Current Biology* 16 (2006), pp. 1018–1025. DOI: [10.1016/j.cub.2006.03.092](https://doi.org/10.1016/j.cub.2006.03.092).
- [41] Fuchou Tang et al. “mRNA-Seq whole-transcriptome analysis of a single cell”. *Nature Methods* 6 (2009), pp. 377–382. DOI: [10.1038/nmeth.1315](https://doi.org/10.1038/nmeth.1315).
- [42] Grace X. Y. Zheng et al. “Massively parallel digital transcriptional profiling of single cells”. *Nature Communications* 8 (2017), p. 14049. DOI: [10.1038/ncomms14049](https://doi.org/10.1038/ncomms14049).
- [43] Paul Datlinger et al. “Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing”. *Nature Methods* 18 (2021), pp. 635–642. DOI: [10.1038/s41592-021-01153-z](https://doi.org/10.1038/s41592-021-01153-z).
- [44] Aleksandra A. Kolodziejczyk et al. “The Technology and Biology of Single-Cell RNA Sequencing”. *Molecular Cell* 58 (2015), pp. 610–620. DOI: [10.1016/j.molcel.2015.04.005](https://doi.org/10.1016/j.molcel.2015.04.005).

- [45] Florian Buettner et al. “Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells”. *Nature Biotechnology* 33 (2015), pp. 155–160. DOI: [10.1038/nbt.3102](https://doi.org/10.1038/nbt.3102).
- [46] Antoine-Emmanuel Saliba et al. “Single-cell RNA-seq: advances and future challenges”. *Nucleic Acids Research* 42 (2014), pp. 8845–8860. DOI: [10.1093/nar/gku555](https://doi.org/10.1093/nar/gku555).
- [47] Amos Tanay and Aviv Regev. “Scaling single-cell genomics from phenomenology to mechanism”. *Nature* 541 (2017), pp. 331–338. DOI: [10.1038/nature21350](https://doi.org/10.1038/nature21350).
- [48] Ashwinikumar Kulkarni et al. “Beyond bulk: a review of single cell transcriptomics methodologies and applications”. *Current Opinion in Biotechnology* 58 (2019), pp. 129–136. DOI: <https://doi.org/10.1016/j.copbio.2019.03.001>.
- [49] Ashraful Haque et al. “A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications”. *Genome Medicine* 9 (2017), p. 75. DOI: [10.1186/s13073-017-0467-4](https://doi.org/10.1186/s13073-017-0467-4).
- [50] Dezhi Huang et al. “Advances in single-cell RNA sequencing and its applications in cancer research”. *Journal of Hematology & Oncology* 16 (2023), p. 98. DOI: [10.1186/s13045-023-01494-6](https://doi.org/10.1186/s13045-023-01494-6).
- [51] Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. “Computational and analytical challenges in single-cell transcriptomics”. *Nature Reviews Genetics* 16 (2015), pp. 133–145. DOI: [10.1038/nrg3833](https://doi.org/10.1038/nrg3833).
- [52] Johan Paulsson. “Models of stochastic gene expression”. *Physics of Life Reviews* 2 (2005), pp. 157–175. DOI: [10.1016/j.plrev.2005.03.003](https://doi.org/10.1016/j.plrev.2005.03.003).
- [53] Thomas B. Kepler and Timothy C. Elston. “Stochasticity in Transcriptional Regulation: Origins, Consequences, and Mathematical Representations”. *Biophysical Journal* 81 (2001), pp. 3116–3136. DOI: [10.1016/s0006-3495\(01\)75949-8](https://doi.org/10.1016/s0006-3495(01)75949-8).
- [54] Vahid Shahrezaei and Peter S. Swain. “Analytical distributions for stochastic gene expression”. *Proceedings of the National Academy of Sciences* 105 (2008), pp. 17256–17261. DOI: [10.1073/pnas.0803850105](https://doi.org/10.1073/pnas.0803850105).
- [55] Vahid Shahrezaei, Julien F Ollivier, and Peter S Swain. “Colored extrinsic fluctuations and stochastic gene expression”. *Molecular Systems Biology* 4 (2008), p. 196. DOI: <https://doi.org/10.1038/msb.2008.31>.

- [56] Abhyudai Singh and Mohammad Soltani. “Quantifying Intrinsic and Extrinsic Variability in Stochastic Gene Expression Models”. *PLOS ONE* 8 (2014), pp. 1–12. DOI: [10.1371/journal.pone.0084301](https://doi.org/10.1371/journal.pone.0084301).
- [57] Lucy Ham et al. “Exactly solvable models of stochastic gene expression”. *The Journal of Chemical Physics* 152 (2020), p. 144106. DOI: [10.1063/1.5143540](https://doi.org/10.1063/1.5143540).
- [58] Zhixing Cao and Ramon Grima. “Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells”. *Proceedings of the National Academy of Sciences* 117 (2020), pp. 4682–4692. DOI: [10.1073/pnas.1910888117](https://doi.org/10.1073/pnas.1910888117).
- [59] Niraj Kumar, Abhyudai Singh, and Rahul V. Kulkarni. “Transcriptional Bursting in Gene Expression: Analytical Results for General Stochastic Models”. *PLOS Computational Biology* 11 (2015), e1004292. DOI: [10.1371/journal.pcbi.1004292](https://doi.org/10.1371/journal.pcbi.1004292).
- [60] J. Peccoud and B. Ycart. “Markovian Modeling of Gene-Product Synthesis”. *Theoretical Population Biology* 48 (1995), pp. 222–234. DOI: [10.1006/tpbi.1995.1027](https://doi.org/10.1006/tpbi.1995.1027).
- [61] J. Goutsias and G. Jenkinson. “Markovian dynamics on complex reaction networks”. *Physics Reports* 529 (2013), pp. 199–264. DOI: <https://doi.org/10.1016/j.physrep.2013.03.004>.
- [62] P H Constantino et al. “Modeling stochasticity in biochemical reaction networks”. *Journal of Physics D: Applied Physics* 49 (2016), p. 093001. DOI: [10.1088/0022-3727/49/9/093001](https://doi.org/10.1088/0022-3727/49/9/093001).
- [63] Daniel T. Gillespie. “A rigorous derivation of the chemical master equation”. *Physica A: Statistical Mechanics and its Applications* 188 (1992), pp. 404–425. DOI: [10.1016/0378-4371\(92\)90283-v](https://doi.org/10.1016/0378-4371(92)90283-v).
- [64] Daniel T. Gillespie. “Exact stochastic simulation of coupled chemical reactions”. *The Journal of Physical Chemistry* 81 (1977), pp. 2340–2361. DOI: [10.1021/j100540a008](https://doi.org/10.1021/j100540a008).
- [65] Daniel T. Gillespie. “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions”. *Journal of Computational Physics* 22 (1976), pp. 403–434. DOI: [10.1016/0021-9991\(76\)90041-3](https://doi.org/10.1016/0021-9991(76)90041-3).
- [66] Daniel T. Gillespie, Andreas Hellander, and Linda R. Petzold. “Perspective: Stochastic algorithms for chemical kinetics”. *The Journal of Chemical Physics* 138 (2013), p. 170901. DOI: [10.1063/1.4801941](https://doi.org/10.1063/1.4801941).

- [67] Chang Hyeong Lee, Kyeong-Hun Kim, and Pilwon Kim. “A moment closure method for stochastic reaction networks”. *The Journal of Chemical Physics* 130 (2009), p. 134107. DOI: [10.1063/1.3103264](https://doi.org/10.1063/1.3103264).
- [68] Daniel T. Gillespie. “The chemical Langevin equation”. *The Journal of Chemical Physics* 113 (2000), pp. 297–306. DOI: [10.1063/1.481811](https://doi.org/10.1063/1.481811).
- [69] A. F. Peralta and R. Toral. “System-size expansion of the moments of a master equation”. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28 (2018), p. 106303. DOI: [10.1063/1.5039817](https://doi.org/10.1063/1.5039817).
- [70] Brian Munsky and Mustafa Khammash. “The finite state projection algorithm for the solution of the chemical master equation”. *The Journal of Chemical Physics* 124 (2006). DOI: [10.1063/1.2145882](https://doi.org/10.1063/1.2145882).
- [71] David J. Warne, Ruth E. Baker, and Matthew J. Simpson. “Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art”. *Journal of The Royal Society Interface* 16 (2019), p. 20180943. DOI: [10.1098/rsif.2018.0943](https://doi.org/10.1098/rsif.2018.0943).
- [72] Giulia Simoni et al. “Stochastic simulation algorithms for computational systems biology: Exact, approximate, and hybrid methods”. *WIREs Systems Biology and Medicine* 11 (2019), e1459. DOI: [10.1002/wsbm.1459](https://doi.org/10.1002/wsbm.1459).
- [73] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. “Exponential scaling of single-cell RNA-seq in the past decade”. *Nature Protocols* 13 (2018), pp. 599–604. DOI: [10.1038/nprot.2017.149](https://doi.org/10.1038/nprot.2017.149).
- [74] Peter V Kharchenko. “The triumphs and limitations of computational methods for scRNA-seq”. *Nature Methods* 18 (2021), pp. 723–732. DOI: [10.1038/s41592-021-01171-x](https://doi.org/10.1038/s41592-021-01171-x).
- [75] Malte D Luecken and Fabian J Theis. “Current best practices in single-cell RNA-seq analysis: a tutorial”. *Molecular Systems Biology* 15 (2019), e8746. DOI: [10.15252/msb.20188746](https://doi.org/10.15252/msb.20188746).
- [76] Allon Wagner, Aviv Regev, and Nir Yosef. “Revealing the vectors of cellular identity with single-cell genomics”. *Nature Biotechnology* 34 (2016), pp. 1145–1160. DOI: [10.1038/nbt.3711](https://doi.org/10.1038/nbt.3711).

- [77] Cole Trapnell. “Defining cell types and states with single-cell genomics”. *Genome Research* 25 (2015), pp. 1491–1498. DOI: [10.1101/gr.190595.115](https://doi.org/10.1101/gr.190595.115).
- [78] Zhichao Miao et al. “Putative cell type discovery from single-cell gene expression data”. *Nature Methods* 17 (2020), pp. 621–628. DOI: [10.1038/s41592-020-0825-9](https://doi.org/10.1038/s41592-020-0825-9).
- [79] Vân Anh Huynh-Thu et al. “Inferring Regulatory Networks from Expression Data Using Tree-Based Methods”. *PLoS ONE* 5 (2010), e12776. DOI: [10.1371/journal.pone.0012776](https://doi.org/10.1371/journal.pone.0012776).
- [80] Andrea Ocone et al. “Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data”. *Bioinformatics* 31 (2015), pp. 89–96. DOI: [10.1093/bioinformatics/btv257](https://doi.org/10.1093/bioinformatics/btv257).
- [81] Thalia E. Chan, Michael P. H. Stumpf, and Ann C. Babbie. “Gene Regulatory Networks from Single Cell Data for Exploring Cell Fate Decisions”. *Computational Stem Cell Biology*. Springer New York, 2019, pp. 211–238. DOI: [10.1007/978-1-4939-9224-9_10](https://doi.org/10.1007/978-1-4939-9224-9_10).
- [82] Aditya Pratapa et al. “Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data”. *Nature Methods* 17 (2020), pp. 147–154. DOI: [10.1038/s41592-019-0690-6](https://doi.org/10.1038/s41592-019-0690-6).
- [83] Cole Trapnell et al. “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells”. *Nature Biotechnology* 32 (2014), pp. 381–386. DOI: [10.1038/nbt.2859](https://doi.org/10.1038/nbt.2859).
- [84] Ann C. Babbie, Thalia E. Chan, and Michael P.H. Stumpf. “Learning regulatory models for cell development from single cell transcriptomic data”. *Current Opinion in Systems Biology* 5 (2017), pp. 72–81. DOI: [10.1016/j.coisb.2017.07.013](https://doi.org/10.1016/j.coisb.2017.07.013).
- [85] Junyue Cao et al. “The single-cell transcriptional landscape of mammalian organogenesis”. *Nature* 566 (2019), pp. 496–502. DOI: [10.1038/s41586-019-0969-x](https://doi.org/10.1038/s41586-019-0969-x).
- [86] David Schnoerr, Guido Sanguinetti, and Ramon Grima. “Approximation and inference methods for stochastic biochemical kinetics—a tutorial review”. *Journal of Physics A: Mathematical and Theoretical* 50 (2017), p. 093001. DOI: [10.1088/1751-8121/aa54d9](https://doi.org/10.1088/1751-8121/aa54d9).
- [87] Brian Munsky, Brooke Trinh, and Mustafa Khammash. “Listening to the noise: random fluctuations reveal gene network parameters”. *Molecular Systems Biology* 5 (2009), p. 318. DOI: [10.1038/msb.2009.75](https://doi.org/10.1038/msb.2009.75).

- [88] Jong Kim and John C Marioni. “Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data”. *Genome Biology* 14 (2013), R7. DOI: [10.1186/gb-2013-14-1-r7](https://doi.org/10.1186/gb-2013-14-1-r7).
- [89] Bernie J. Daigle et al. “Inferring single-cell gene expression mechanisms using stochastic simulation”. *Bioinformatics* 31 (2015), pp. 1428–1435. DOI: [10.1093/bioinformatics/btv007](https://doi.org/10.1093/bioinformatics/btv007).
- [90] Ian S. Hsu and Alan M. Moses. “Stochastic models for single-cell data: Current challenges and the way forward”. *The FEBS Journal* 289 (2021), pp. 647–658. DOI: [10.1111/febs.15760](https://doi.org/10.1111/febs.15760).
- [91] Anika Gupta et al. “Inferring gene regulation from stochastic transcriptional variation across single cells at steady state”. *Proceedings of the National Academy of Sciences* 119 (2022), e2207392119. DOI: [10.1073/pnas.2207392119](https://doi.org/10.1073/pnas.2207392119).
- [92] Anton J. M. Larsson et al. “Genomic encoding of transcriptional burst kinetics”. *Nature* 565 (2019), pp. 251–254. DOI: [10.1038/s41586-018-0836-1](https://doi.org/10.1038/s41586-018-0836-1).
- [93] Samuel O Skinner et al. “Single-cell analysis of transcription kinetics across the cell cycle”. *eLife* 5 (2016), e12175. DOI: [10.7554/elife.12175](https://doi.org/10.7554/elife.12175).
- [94] Wenhao Tang et al. “Modelling capture efficiency of single-cell RNA-sequencing data improves inference of transcriptome-wide burst kinetics”. *Bioinformatics* 39 (2023), btad395. DOI: [10.1093/bioinformatics/btad395](https://doi.org/10.1093/bioinformatics/btad395).
- [95] David M. Edwards, Philip Davies, and Daniel Hebenstreit. “Synergising single-cell resolution and 4sU labelling boosts inference of transcriptional bursting”. *Genome Biology* 24 (2023), p. 138. DOI: [10.1186/s13059-023-02977-y](https://doi.org/10.1186/s13059-023-02977-y).
- [96] Songhao Luo et al. “Inferring transcriptional bursting kinetics from single-cell snapshot data using a generalized telegraph model”. *Royal Society Open Science* 10 (2023), p. 221057. DOI: [10.1098/rsos.221057](https://doi.org/10.1098/rsos.221057).
- [97] Maria Carilli et al. “Biophysical modeling with variational autoencoders for bimodal, single-cell RNA sequencing data”. *Nature Methods* (2024), pp. 1466–1469. DOI: [10.1038/s41592-024-02365-9](https://doi.org/10.1038/s41592-024-02365-9).

- [98] Fabian Fröhlich et al. “Scalable Parameter Estimation for Genome-Scale Biochemical Reaction Networks”. *PLOS Computational Biology* 13 (2017), e1005331. DOI: [10.1371/journal.pcbi.1005331](https://doi.org/10.1371/journal.pcbi.1005331).
- [99] Fabian Fröhlich et al. “Inference for Stochastic Chemical Kinetics Using Moment Equations and System Size Expansion”. *PLOS Computational Biology* 12 (2016), e1005030. DOI: [10.1371/journal.pcbi.1005030](https://doi.org/10.1371/journal.pcbi.1005030).
- [100] Peter Milner, Colin S. Gillespie, and Darren J. Wilkinson. “Moment closure based parameter inference of stochastic kinetic models”. *Statistics and Computing* 23 (2012), pp. 287–295. DOI: [10.1007/s11222-011-9310-8](https://doi.org/10.1007/s11222-011-9310-8).
- [101] Zekai Li, Mauricio Barahona, and Philipp Thomas. “Moment-based parameter inference with error guarantees for stochastic reaction networks”. *arXiv* (2024). DOI: [10.48550/arXiv.2406.17434](https://doi.org/10.48550/arXiv.2406.17434).
- [102] Simon Martina Perez, Heba Sailem, and Ruth E. Baker. “Efficient Bayesian inference for mechanistic modelling with high-throughput data”. *PLOS Computational Biology* 18 (2022), p. 1010191. DOI: [10.1371/journal.pcbi.1010191](https://doi.org/10.1371/journal.pcbi.1010191).
- [103] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. “The frontier of simulation-based inference”. *Proceedings of the National Academy of Sciences* 117 (2020), pp. 30055–30062. DOI: [10.1073/pnas.1912789117](https://doi.org/10.1073/pnas.1912789117).
- [104] Andrew Golightly and Darren J. Wilkinson. “Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo”. *Interface Focus* 1 (2011), pp. 807–820. DOI: [10.1098/rsfs.2011.0047](https://doi.org/10.1098/rsfs.2011.0047).
- [105] Christoph Zechner et al. “Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings”. *Nature Methods* 11 (2014), pp. 197–202. DOI: [10.1038/nmeth.2794](https://doi.org/10.1038/nmeth.2794).
- [106] Mikael Sunnåker et al. “Approximate Bayesian Computation”. *PLoS Computational Biology* 9 (2013), e1002803. DOI: [10.1371/journal.pcbi.1002803](https://doi.org/10.1371/journal.pcbi.1002803).
- [107] Tina Toni et al. “Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems”. *Journal of The Royal Society Interface* 6 (2008), pp. 187–202. DOI: [10.1098/rsif.2008.0172](https://doi.org/10.1098/rsif.2008.0172).

- [108] Andreas Christ Sølvesten Jørgensen et al. “Efficient Bayesian inference for stochastic agent-based models”. *PLOS Computational Biology* 18 (2022), e1009508. DOI: [10.1371/journal.pcbi.1009508](https://doi.org/10.1371/journal.pcbi.1009508).
- [109] Caleb Weinreb et al. “Fundamental limits on dynamic inference from single-cell snapshots”. *Proceedings of the National Academy of Sciences* 115 (2018). DOI: [10.1073/pnas.1714723115](https://doi.org/10.1073/pnas.1714723115).
- [110] Michael M Saint-Antoine and Abhyudai Singh. “Network inference in systems biology: recent developments, challenges, and applications”. *Current Opinion in Biotechnology* 63 (2020), pp. 89–98. DOI: [10.1016/j.copbio.2019.12.002](https://doi.org/10.1016/j.copbio.2019.12.002).
- [111] Wouter Saelens et al. “A comparison of single-cell trajectory inference methods”. *Nature Biotechnology* 37 (2019), pp. 547–554. DOI: [10.1038/s41587-019-0071-9](https://doi.org/10.1038/s41587-019-0071-9).
- [112] Louise Deconinck et al. “Recent advances in trajectory inference from single-cell omics data”. *Current Opinion in Systems Biology* 27 (2021), p. 100344. DOI: [10.1016/j.coisb.2021.05.005](https://doi.org/10.1016/j.coisb.2021.05.005).
- [113] Sophie Triteschler et al. “Concepts and limitations for learning developmental trajectories from single cell genomics”. *Development* 146 (2019). DOI: [10.1242/dev.170506](https://doi.org/10.1242/dev.170506).
- [114] Meichen Fang, Gennady Gorin, and Lior Pachter. “Trajectory inference from single-cell genomics data with a process time model”. *bioRxiv* (2024). DOI: [10.1101/2024.01.26.577510](https://doi.org/10.1101/2024.01.26.577510).
- [115] Amit Zeisel et al. “Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli”. *Molecular Systems Biology* 7 (2011), p. 529. DOI: [10.1038/msb.2011.62](https://doi.org/10.1038/msb.2011.62).
- [116] Gioele La Manno et al. “RNA velocity of single cells”. *Nature* 560 (2018), pp. 494–498. DOI: [10.1038/s41586-018-0414-6](https://doi.org/10.1038/s41586-018-0414-6).
- [117] Volker Bergen et al. “Generalizing RNA velocity to transient cell states through dynamical modeling”. *Nature Biotechnology* 38 (2020), pp. 1408–1414. DOI: [10.1038/s41587-020-0591-3](https://doi.org/10.1038/s41587-020-0591-3).
- [118] Valentine Svensson and Lior Pachter. “RNA Velocity: Molecular Kinetics from Single-Cell RNA-Seq”. *Molecular Cell* 72 (2018), pp. 7–9. DOI: <https://doi.org/10.1016/j.molcel.2018.09.026>.

- [119] Andrea Riba et al. “Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning”. *Nature Communications* 13 (2022), p. 2865. DOI: [10.1038/s41467-022-30545-8](https://doi.org/10.1038/s41467-022-30545-8).
- [120] Gennady Gorin, Valentine Svensson, and Lior Pachter. “Protein velocity and acceleration from single-cell multiomics experiments”. *Genome Biology* 21 (2020), p. 39. DOI: [10.1186/s13059-020-1945-3](https://doi.org/10.1186/s13059-020-1945-3).
- [121] Xiaojie Qiu et al. “Mapping transcriptomic vector fields of single cells”. *Cell* 185 (2022), pp. 690–711. DOI: [10.1016/j.cell.2021.12.045](https://doi.org/10.1016/j.cell.2021.12.045).
- [122] Zhanlin Chen et al. “DeepVelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations”. *Science Advances* 8 (2022), eabq3745. DOI: [10.1126/sciadv.abq3745](https://doi.org/10.1126/sciadv.abq3745).
- [123] Philipp Weiler et al. “CellRank 2: unified fate mapping in multiview single-cell data”. *Nature Methods* 21 (2024), pp. 1196–1205. DOI: [10.1038/s41592-024-02303-9](https://doi.org/10.1038/s41592-024-02303-9).
- [124] Cassandra Burdziak et al. “scKINETICS: inference of regulatory velocity with single-cell transcriptomics data”. *Bioinformatics* 39 (2023), pp. i394–i403. DOI: [10.1093/bioinformatics/btad267](https://doi.org/10.1093/bioinformatics/btad267).
- [125] Rohit Singh et al. “Causal gene regulatory analysis with RNA velocity reveals an interplay between slow and fast transcription factors”. *Cell Systems* 15 (2024), pp. 462–474. DOI: [10.1016/j.cels.2024.04.005](https://doi.org/10.1016/j.cels.2024.04.005).
- [126] Charlotte Soneson et al. “Preprocessing choices affect RNA velocity results for droplet scRNA-seq data”. *PLOS Computational Biology* 17 (2021), e1008585. DOI: [10.1371/journal.pcbi.1008585](https://doi.org/10.1371/journal.pcbi.1008585).
- [127] Gennady Gorin et al. “RNA velocity unraveled”. *PLOS Computational Biology* 18 (2022), e1010492. DOI: [10.1371/journal.pcbi.1010492](https://doi.org/10.1371/journal.pcbi.1010492).
- [128] Volker Bergen et al. “RNA velocity—current challenges and future perspectives”. *Molecular Systems Biology* 17 (2021). DOI: [10.15252/msb.202110282](https://doi.org/10.15252/msb.202110282).
- [129] Florian Erhard et al. “Time-resolved single-cell RNA-seq using metabolic RNA labelling”. *Nature Reviews Methods Primers* 2 (2022), p. 77. DOI: [10.1038/s43586-022-00157-z](https://doi.org/10.1038/s43586-022-00157-z).

- [130] Lars Dölken et al. “High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay”. *RNA* 14 (2008), pp. 1959–1972. ISSN: 1469-9001. DOI: [10.1261/rna.1136108](https://doi.org/10.1261/rna.1136108).
- [131] Björn Schwanhäusser et al. “Global quantification of mammalian gene expression control”. *Nature* 473 (2011), pp. 337–342. DOI: [10.1038/nature10098](https://doi.org/10.1038/nature10098).
- [132] Michal Rabani et al. “Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells”. *Nature Biotechnology* 29 (2011), pp. 436–442. DOI: [10.1038/nbt.1861](https://doi.org/10.1038/nbt.1861).
- [133] Veronika A Herzog et al. “Thiol-linked alkylation of RNA to assess expression dynamics”. *Nature Methods* 14 (2017), pp. 1198–1204. DOI: [10.1038/nmeth.4435](https://doi.org/10.1038/nmeth.4435).
- [134] Christopher Jürges, Lars Dölken, and Florian Erhard. “Dissecting newly transcribed and old RNA using GRAND-SLAM”. *Bioinformatics* 34 (2018), pp. i218–i226. DOI: [10.1093/bioinformatics/bty256](https://doi.org/10.1093/bioinformatics/bty256).
- [135] Angela R Wu et al. “Quantitative assessment of single-cell RNA-sequencing methods”. *Nature Methods* 11 (2013), pp. 41–46. DOI: [10.1038/nmeth.2694](https://doi.org/10.1038/nmeth.2694).
- [136] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. “Validation of noise models for single-cell transcriptomics”. *Nature Methods* 11 (2014), pp. 637–640. DOI: [10.1038/nmeth.2930](https://doi.org/10.1038/nmeth.2930).
- [137] Philip Brennecke et al. “Accounting for technical noise in single-cell RNA-seq experiments”. *Nature Methods* 10 (2013), pp. 1093–1095. DOI: [10.1038/nmeth.2645](https://doi.org/10.1038/nmeth.2645).
- [138] Peter V Kharchenko, Lev Silberstein, and David T Scadden. “Bayesian approach to single-cell differential expression analysis”. *Nature Methods* 11 (2014), pp. 740–742. DOI: [10.1038/nmeth.2967](https://doi.org/10.1038/nmeth.2967).
- [139] Aaron T. L. Lun, Karsten Bach, and John C. Marioni. “Pooling across cells to normalize single-cell RNA sequencing data with many zero counts”. *Genome Biology* 17 (2016), p. 75. DOI: [10.1186/s13059-016-0947-7](https://doi.org/10.1186/s13059-016-0947-7).
- [140] Bo Ding et al. “Normalization and noise reduction for single cell RNA-seq experiments”. *Bioinformatics* 31 (2015), pp. 2225–2227. DOI: [10.1093/bioinformatics/btv122](https://doi.org/10.1093/bioinformatics/btv122).
- [141] Catalina A Vallejos et al. “Normalizing single-cell RNA sequencing data: challenges and opportunities”. *Nature methods* 14 (2017), pp. 565–571. DOI: [10.1038/nmeth.4292](https://doi.org/10.1038/nmeth.4292).

- [142] Ruoyu Zhang, Gurinder S. Atwal, and Wei Keat Lim. “Noise regularization removes correlation artifacts in single-cell RNA-seq data preprocessing”. *Patterns* 2 (2021), p. 100211. DOI: [10.1016/j.patter.2021.100211](https://doi.org/10.1016/j.patter.2021.100211).
- [143] Abhishek Sarkar and Matthew Stephens. “Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis”. *Nature Genetics* 53 (2021), pp. 770–777. DOI: [10.1038/s41588-021-00873-4](https://doi.org/10.1038/s41588-021-00873-4).
- [144] Gennady Gorin and Lior Pachter. “Distinguishing biophysical stochasticity from technical noise in single-cell RNA sequencing using Monod”. *bioRxiv* (2022). DOI: [10.1101/2022.06.11.495771](https://doi.org/10.1101/2022.06.11.495771).
- [145] Wenhao Tang et al. “bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data”. *Bioinformatics* 36 (2019), pp. 1174–1181. DOI: [10.1093/bioinformatics/btz726](https://doi.org/10.1093/bioinformatics/btz726).
- [146] Mo Huang et al. “SAVER: gene expression recovery for single-cell RNA sequencing”. *Nature methods* 15 (2018), pp. 539–542. DOI: [10.1038/s41592-018-0033-z](https://doi.org/10.1038/s41592-018-0033-z).
- [147] Jérémie Breda, Mihaela Zavolan, and Erik van Nimwegen. “Bayesian inference of gene expression states from single-cell RNA-seq data”. *Nature Biotechnology* 39 (2021), pp. 1008–1016. DOI: [10.1038/s41587-021-00875-x](https://doi.org/10.1038/s41587-021-00875-x).
- [148] Gennady Gorin, John J. Vastola, and Lior Pachter. “Studying stochastic systems biology of the cell with single-cell genomics data”. *bioRxiv* (2023). DOI: [10.1101/2023.05.17.541250](https://doi.org/10.1101/2023.05.17.541250).
- [149] Stephanie C Hicks et al. “Missing data and technical variability in single-cell RNA-sequencing experiments”. *Biostatistics* 19 (2017), pp. 562–578. DOI: [10.1093/biostatistics/kxx053](https://doi.org/10.1093/biostatistics/kxx053).
- [150] Po-Yuan Tung et al. “Batch effects and the effective design of single-cell gene expression studies”. *Scientific Reports* 7 (2017), p. 39921. DOI: [10.1038/srep39921](https://doi.org/10.1038/srep39921).
- [151] Ilya Korsunsky et al. “Fast, sensitive and accurate integration of single-cell data with Harmony”. *Nature Methods* 16 (2019), pp. 1289–1296. DOI: [10.1038/s41592-019-0619-0](https://doi.org/10.1038/s41592-019-0619-0).
- [152] Jacob Zhurinsky et al. “A Coordinated Global Control over Cellular Transcription”. *Current Biology* 20 (2010), pp. 2010–2015. DOI: [10.1016/j.cub.2010.10.002](https://doi.org/10.1016/j.cub.2010.10.002).

- [153] Martin Fischer et al. “Coordinating gene expression during the cell cycle”. *Trends in Biochemical Sciences* 47 (2022), pp. 1009–1022. DOI: [10.1016/j.tibs.2022.06.007](https://doi.org/10.1016/j.tibs.2022.06.007).
- [154] David O. Morgan. *The cell cycle: principles of control*. Oxford University Press, 2006.
- [155] Nico Battich et al. “Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies”. *Science* 367 (2020), pp. 1151–1156. DOI: [10.1126/science.aax3072](https://doi.org/10.1126/science.aax3072).
- [156] Soumya Bhattacharjee et al. “Combinatorial Control of Gene Expression”. *BioMed Research International* (2013), pp. 1–11. DOI: [10.1155/2013/407263](https://doi.org/10.1155/2013/407263).
- [157] Smadar Ben-Tabou de-Leon and Eric H. Davidson. “Gene Regulation: Gene Control Network in Development”. *Annual Review of Biophysics and Biomolecular Structure* 36 (2007), pp. 191–212. DOI: [10.1146/annurev.biophys.35.040405.102002](https://doi.org/10.1146/annurev.biophys.35.040405.102002).
- [158] Scott D. Pope and Ruslan Medzhitov. “Emerging Principles of Gene Expression Programs and Their Regulation”. *Molecular Cell* 71 (2018), pp. 389–397. DOI: [10.1016/j.molcel.2018.07.017](https://doi.org/10.1016/j.molcel.2018.07.017).
- [159] Olivia Padovan-Merhar and Arjun Raj. “Using variability in gene expression as a tool for studying gene regulation”. *WIREs Systems Biology and Medicine* 5 (2013), pp. 751–759. DOI: [10.1002/wsbm.1243](https://doi.org/10.1002/wsbm.1243).
- [160] Darren J. Wilkinson. “Stochastic modelling for quantitative description of heterogeneous biological systems”. *Nature Reviews Genetics* 10 (2009), pp. 122–133. DOI: [10.1038/nrg2509](https://doi.org/10.1038/nrg2509).
- [161] N.G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, 2007. DOI: [10.1016/b978-0-444-52965-7.x5000-4](https://doi.org/10.1016/b978-0-444-52965-7.x5000-4).
- [162] Eric L. Haseltine and James B. Rawlings. “On the origins of approximations for stochastic chemical kinetics”. *The Journal of Chemical Physics* 123 (2005). DOI: [10.1063/1.2062048](https://doi.org/10.1063/1.2062048).
- [163] Hye-Won Kang and Thomas G. Kurtz. “Separation of time-scales and model reduction for stochastic reaction networks”. *The Annals of Applied Probability* 23 (2013), pp. 529–583. DOI: [10.1214/12-AAP841](https://doi.org/10.1214/12-AAP841).
- [164] Arjun Raj et al. “Stochastic mRNA Synthesis in Mammalian Cells”. *PLOS Biology* 4 (2006), pp. 1–13. DOI: [10.1371/journal.pbio.0040309](https://doi.org/10.1371/journal.pbio.0040309).

- [165] David M. Suter et al. “Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics”. *Science* 332 (2011), pp. 472–474. DOI: [10.1126/science.1198817](https://doi.org/10.1126/science.1198817).
- [166] Roy D. Dar et al. “Transcriptional burst frequency and burst size are equally modulated across the human genome”. *Proceedings of the National Academy of Sciences* 109 (2012), pp. 17454–17459. DOI: [10.1073/pnas.1213530109](https://doi.org/10.1073/pnas.1213530109).
- [167] Nico Battich, Thomas Stoeger, and Lucas Pelkmans. “Control of Transcript Variability in Single Mammalian Cells”. *Cell* 163 (2015), pp. 1596–1610. DOI: [10.1016/j.cell.2015.11.018](https://doi.org/10.1016/j.cell.2015.11.018).
- [168] Megan A. Coomer, Lucy Ham, and Michael P.H. Stumpf. “Noise distorts the epigenetic landscape and shapes cell-fate decisions”. *Cell Systems* 13 (2022), pp. 83–102. DOI: <https://doi.org/10.1016/j.cels.2021.09.002>.
- [169] Natalie A. Cookson et al. “Cell cycle-dependent variations in protein concentration”. *Nucleic Acids Research* 38 (2009), pp. 2676–2681. DOI: [10.1093/nar/gkp1069](https://doi.org/10.1093/nar/gkp1069).
- [170] Veronika Bierbaum and Stefan Klumpp. “Impact of the cell division cycle on gene circuits”. *Physical Biology* 12 (2015), e066003. DOI: [10.1088/1478-3975/12/6/066003](https://doi.org/10.1088/1478-3975/12/6/066003).
- [171] Philipp Thomas. “Making sense of snapshot data: ergodic principle for clonal cell populations”. *Journal of The Royal Society Interface* 14 (2017), p. 20170467. DOI: [10.1098/rsif.2017.0467](https://doi.org/10.1098/rsif.2017.0467).
- [172] Noreen Walker, Philippe Nghe, and Sander J. Tans. “Generation and filtering of gene expression noise by the bacterial cell cycle”. *BMC Biology* 14 (2016), p. 11. DOI: [10.1186/s12915-016-0231-z](https://doi.org/10.1186/s12915-016-0231-z).
- [173] Joseph R. Peterson et al. “Effects of DNA replication on mRNA noise”. *Proceedings of the National Academy of Sciences* 112 (2015), pp. 15886–15891. DOI: [10.1073/pnas.1516246112](https://doi.org/10.1073/pnas.1516246112).
- [174] Dann Huh and Johan Paulsson. “Non-genetic heterogeneity from stochastic partitioning at cell division”. *Nature Genetics* 43 (2010), pp. 95–100. DOI: [10.1038/ng.729](https://doi.org/10.1038/ng.729).
- [175] Philipp Thomas. “Intrinsic and extrinsic noise of gene expression in lineage trees”. *Scientific Reports* 9 (2019), p. 474. DOI: [10.1038/s41598-018-35927-x](https://doi.org/10.1038/s41598-018-35927-x).

- [176] Ruben Perez-Carrasco, Casper Beentjes, and Ramon Grima. “Effects of cell cycle variability on lineage and population measurements of messenger RNA abundance”. *Journal of The Royal Society Interface* 17 (2020), p. 20200360. DOI: [10.1098/rsif.2020.0360](https://doi.org/10.1098/rsif.2020.0360).
- [177] C. J. Zopf et al. “Cell-Cycle Dependence of Transcription Dominates Noise in Gene Expression”. *PLoS Computational Biology* 9 (2013), e1003161. DOI: [10.1371/journal.pcbi.1003161](https://doi.org/10.1371/journal.pcbi.1003161).
- [178] Haiyue Liu et al. “SLAM-Drop-seq reveals mRNA kinetic rates throughout the cell cycle”. *Molecular Systems Biology* (2023), e11427. DOI: [10.15252/msb.202211427](https://doi.org/10.15252/msb.202211427).
- [179] Xi-Ming Sun et al. “Size-Dependent Increase in RNA Polymerase II Initiation Rates Mediates Gene Expression Scaling with Cell Size”. *Current Biology* 30 (2020), pp. 1217–1230. DOI: [10.1016/j.cub.2020.01.053](https://doi.org/10.1016/j.cub.2020.01.053).
- [180] Philipp Thomas and Vahid Shahrezaei. “Coordination of gene expression noise with cell size: analytical results for agent-based models of growing cell populations”. *Journal of The Royal Society Interface* 18 (2021), p. 20210274. DOI: [10.1098/rsif.2021.0274](https://doi.org/10.1098/rsif.2021.0274).
- [181] Scott Berry et al. “Feedback from nuclear RNA on transcription promotes robust RNA concentration homeostasis in human cells”. *Cell Systems* 13 (2022), pp. 454–470. DOI: [10.1016/j.cels.2022.04.005](https://doi.org/10.1016/j.cels.2022.04.005).
- [182] Chen Jia and Ramon Grima. “Coupling gene expression dynamics to cell size dynamics and cell cycle events: Exact and approximate solutions of the extended telegraph model”. *iScience* 26 (2023), p. 105746. DOI: [10.1016/j.isci.2022.105746](https://doi.org/10.1016/j.isci.2022.105746).
- [183] Songhao Luo et al. “Genome-wide inference reveals that feedback regulations constrain promoter-dependent transcriptional burst kinetics”. *Nucleic Acids Research* 51 (2022), pp. 68–83. DOI: [10.1093/nar/gkac1204](https://doi.org/10.1093/nar/gkac1204).
- [184] Lucy Ham, Marcel Jackson, and Michael PH Stumpf. “Pathway dynamics can delineate the sources of transcriptional noise in gene expression”. *eLife* 10 (2021), e69324. DOI: [10.7554/elife.69324](https://doi.org/10.7554/elife.69324).
- [185] Gennady Gorin et al. “Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments”. *Nature Communications* 13 (2022), p. 7620. DOI: [10.1038/s41467-022-34857-7](https://doi.org/10.1038/s41467-022-34857-7).

- [186] Gennady Gorin et al. “Stochastic simulation and statistical inference platform for visualization and estimation of transcriptional kinetics”. *PLOS ONE* 15 (2020), e0230736. DOI: [10.1371/journal.pone.0230736](https://doi.org/10.1371/journal.pone.0230736).
- [187] Florian Erhard et al. “scSLAM-seq reveals core features of transcription dynamics in single cells”. *Nature* 571 (2019), pp. 419–423. DOI: [10.1038/s41586-019-1369-y](https://doi.org/10.1038/s41586-019-1369-y).
- [188] Gert-Jan Hendriks et al. “NASC-seq monitors RNA synthesis in single cells”. *Nature Communications* 10 (2019), p. 3138. DOI: [10.1038/s41467-019-11028-9](https://doi.org/10.1038/s41467-019-11028-9).
- [189] Qi Qiu et al. “Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq”. *Nature Methods* 17 (2020), pp. 991–1001. DOI: [10.1038/s41592-020-0935-4](https://doi.org/10.1038/s41592-020-0935-4).
- [190] Junyue Cao et al. “Sci-fate characterizes the dynamics of gene expression in single cells”. *Nature Biotechnology* 38 (2020), pp. 980–988. DOI: [10.1038/s41587-020-0480-9](https://doi.org/10.1038/s41587-020-0480-9).
- [191] Jingyi Ren et al. “Spatiotemporally resolved transcriptomics reveals the subcellular RNA kinetic landscape”. *Nature Methods* 20 (2023), pp. 695–705. DOI: [10.1038/s41592-023-01829-8](https://doi.org/10.1038/s41592-023-01829-8).
- [192] Shichao Lin et al. “Well-TEMP-seq as a microwell-based strategy for massively parallel profiling of single-cell temporal RNA dynamics”. *Nature Communications* 14 (2023). DOI: [10.1038/s41467-023-36902-5](https://doi.org/10.1038/s41467-023-36902-5).
- [193] Rory J. Maizels, Daniel M. Snell, and James Briscoe. “Reconstructing developmental trajectories using latent dynamical systems and time-resolved transcriptomics”. *Cell Systems* 15 (2024), 411–424.e9. DOI: [10.1016/j.cels.2024.04.004](https://doi.org/10.1016/j.cels.2024.04.004).
- [194] Qiangwei Peng, Xiaojie Qiu, and Tiejun Li. “Storm: Incorporating transient dynamics to infer the RNA velocity with metabolic labeling information”. *bioRxiv* (2023). DOI: [10.1101/2023.06.21.545990](https://doi.org/10.1101/2023.06.21.545990).
- [195] Dimitris Volterras, Vahid Shahrezaei, and Philipp Thomas. “Global transcription regulation revealed from dynamical correlations in time-resolved single-cell RNA sequencing”. *Cell Systems* 15 (2024), pp. 694–708. DOI: [10.1016/j.cels.2024.07.002](https://doi.org/10.1016/j.cels.2024.07.002).
- [196] Delaney K. Sullivan et al. “kallisto, bustools and kb-python for quantifying bulk, single-cell and single-nucleus RNA-seq”. *Nature Protocols* (2024). DOI: [10.1038/s41596-024-01057-0](https://doi.org/10.1038/s41596-024-01057-0).

- [197] Hidenori Tani and Nobuyoshi Akimitsu. “Genome-wide technology for determining RNA stability in mammalian cells: historical perspective and recent advantages based on modified nucleotide labeling”. *RNA Biology* 9 (2012), pp. 1233–1238. DOI: [10.4161/rna.22036](https://doi.org/10.4161/rna.22036).
- [198] Hermannus Kempe et al. “The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise”. *Molecular Biology of the Cell* 26 (2015), pp. 797–804. DOI: [10.1091/mbc.e14-08-1296](https://doi.org/10.1091/mbc.e14-08-1296).
- [199] Olivia Padovan-Merhar et al. “Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms”. *Molecular Cell* 58 (2015), pp. 339–352. DOI: [10.1016/j.molcel.2015.03.005](https://doi.org/10.1016/j.molcel.2015.03.005).
- [200] Jeff Bezanson et al. “Julia: A fresh approach to numerical computing”. *SIAM review* 59 (2017), pp. 65–98. DOI: [10.1137/141000671](https://doi.org/10.1137/141000671).
- [201] Mark A Beaumont, Wenyang Zhang, and David J Balding. “Approximate Bayesian Computation in Population Genetics”. *Genetics* 162 (2002), pp. 2025–2035. DOI: [10.1093/genetics/162.4.2025](https://doi.org/10.1093/genetics/162.4.2025).
- [202] Mark K Transtrum and Peng Qiu. “Optimal experiment selection for parameter estimation in biological differential equation models”. *BMC Bioinformatics* 13 (2012). DOI: [10.1186/1471-2105-13-181](https://doi.org/10.1186/1471-2105-13-181).
- [203] Juliane Liepe et al. “A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation”. *Nature Protocols* 9 (2014), pp. 439–456. DOI: [10.1038/nprot.2014.025](https://doi.org/10.1038/nprot.2014.025).
- [204] Ramon Grima and Pierre-Marie Esmenjaud. “Quantifying and correcting bias in transcriptional parameter inference from single-cell data”. *bioRxiv* (2023). DOI: [10.1101/2023.06.19.545536](https://doi.org/10.1101/2023.06.19.545536).
- [205] Brad T Sherman et al. “DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update)”. *Nucleic Acids Research* 50 (2022), pp. 216–221. DOI: [10.1093/nar/gkac194](https://doi.org/10.1093/nar/gkac194).
- [206] Christopher Jürges, Lars Dölken, and Florian Erhard. “Dissecting newly transcribed and old RNA using GRAND-SLAM”. *Bioinformatics* 34 (2018), pp. 218–226. DOI: [10.1093/bioinformatics/bty256](https://doi.org/10.1093/bioinformatics/bty256).

- [207] Samuel Marguerat and Jürg Bähler. “Coordinating genome expression with cell size”. *Trends in Genetics* 28 (2012), pp. 560–565. DOI: [10.1016/j.tig.2012.07.003](https://doi.org/10.1016/j.tig.2012.07.003).
- [208] Malika Saint et al. “Single-cell imaging and RNA sequencing reveal patterns of gene expression heterogeneity during fission yeast growth and adaptation”. *Nature Microbiology* 4 (2019), pp. 480–491. DOI: [10.1038/s41564-018-0330-4](https://doi.org/10.1038/s41564-018-0330-4).
- [209] Kora-Lee Claude et al. “Transcription coordinates histone amounts and genome content”. *Nature Communications* 12 (2021), p. 4202. DOI: [10.1038/s41467-021-24451-8](https://doi.org/10.1038/s41467-021-24451-8).
- [210] Scott Berry and Lucas Pelkmans. “Mechanisms of cellular mRNA transcript homeostasis”. *Trends in Cell Biology* 32 (2022), pp. 655–668. DOI: [10.1016/j.tcb.2022.05.003](https://doi.org/10.1016/j.tcb.2022.05.003).
- [211] Guo-Cheng Yuan et al. “Challenges and emerging directions in single-cell analysis”. *Genome Biology* 18 (2017), p. 84. DOI: [10.1186/s13059-017-1218-y](https://doi.org/10.1186/s13059-017-1218-y).
- [212] David Lähnemann et al. “Eleven grand challenges in single-cell data science”. *Genome Biology* 21 (2020), p. 31. DOI: [10.1186/s13059-020-1926-6](https://doi.org/10.1186/s13059-020-1926-6).
- [213] Theodoulos Rodosthenous, Vahid Shahrezaei, and Marina Evangelou. “Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study”. *Bioinformatics* 36 (2020), pp. 4616–4625. DOI: [10.1093/bioinformatics/btaa530](https://doi.org/10.1093/bioinformatics/btaa530).
- [214] Theodoulos Rodosthenous, Vahid Shahrezaei, and Marina Evangelou. *Multi-view Data Visualisation via Manifold Learning*. 2021. DOI: [10.48550/arxiv.2101.06763](https://doi.org/10.48550/arxiv.2101.06763).
- [215] Ricard Argelaguet et al. “Computational principles and challenges in single-cell data integration”. *Nature Biotechnology* 39 (2021), pp. 1202–1215. DOI: [10.1038/s41587-021-00895-7](https://doi.org/10.1038/s41587-021-00895-7).
- [216] Britta Velten and Oliver Stegle. “Principles and challenges of modeling temporal and spatial omics data”. *Nature Methods* (2023). DOI: [10.1038/s41592-023-01992-y](https://doi.org/10.1038/s41592-023-01992-y).
- [217] Alev Baysoy et al. “The technological landscape and applications of single-cell multi-omics”. *Nature Reviews Molecular Cell Biology* 24 (2023), pp. 695–713. DOI: [10.1038/s41580-023-00615-w](https://doi.org/10.1038/s41580-023-00615-w).

- [218] Lesley T. MacNeil and Albertha J.M. Walhout. “Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression”. *Genome Research* 21 (2011), pp. 645–657. DOI: [10.1101/gr.097378.109](https://doi.org/10.1101/gr.097378.109).
- [219] Michael Levine and Eric H. Davidson. “Gene regulatory networks for development”. *Proceedings of the National Academy of Sciences* 102 (2005), pp. 4936–4942. DOI: [10.1073/pnas.0408031102](https://doi.org/10.1073/pnas.0408031102).
- [220] Albertha J.M. Walhout. “Unraveling transcription regulatory networks by protein–DNA and protein–protein interaction mapping”. *Genome Research* 16 (2006), pp. 1445–1454. DOI: [10.1101/gr.5321506](https://doi.org/10.1101/gr.5321506).
- [221] Hidde de Jong. “Modeling and Simulation of Genetic Regulatory Systems: A Literature Review”. *Journal of Computational Biology* 9 (2002), pp. 67–103. DOI: [10.1089/10665270252833208](https://doi.org/10.1089/10665270252833208).
- [222] Michael Hecker et al. “Gene regulatory network inference: Data integration in dynamic models—A review”. *Biosystems* 96 (2009), pp. 86–103. DOI: [10.1016/j.biosystems.2008.12.004](https://doi.org/10.1016/j.biosystems.2008.12.004).
- [223] Daniele Mercatelli et al. “Gene regulatory network inference resources: A practical overview”. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1863 (2020), p. 194430. DOI: [10.1016/j.bbagr.2019.194430](https://doi.org/10.1016/j.bbagr.2019.194430).
- [224] David S. Johnson et al. “Genome-Wide Mapping of in Vivo Protein-DNA Interactions”. *Science* 316 (2007), pp. 1497–1502. DOI: [10.1126/science.1141319](https://doi.org/10.1126/science.1141319).
- [225] Hatice S. Kaya-Okur et al. “CUT&Tag for efficient epigenomic profiling of small samples and single cells”. *Nature Communications* 10 (2019). DOI: [10.1038/s41467-019-09982-5](https://doi.org/10.1038/s41467-019-09982-5).
- [226] Dana Pe’er et al. “Inferring subnetworks from perturbed expression profiles”. *Bioinformatics* 17 (2001), S215–S224. DOI: [10.1093/bioinformatics/17.suppl_1.s215](https://doi.org/10.1093/bioinformatics/17.suppl_1.s215).
- [227] Molly Gasperini et al. “A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens”. *Cell* 176 (2019), pp. 377–390. DOI: [10.1016/j.cell.2018.11.029](https://doi.org/10.1016/j.cell.2018.11.029).

- [228] Atray Dixit et al. “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens”. *Cell* 167 (2016), pp. 1853–1866. DOI: [10.1016/j.cell.2016.11.038](https://doi.org/10.1016/j.cell.2016.11.038).
- [229] Paul Datlinger et al. “Pooled CRISPR screening with single-cell transcriptome readout”. *Nature Methods* 14 (2017), pp. 297–301. DOI: [10.1038/nmeth.4177](https://doi.org/10.1038/nmeth.4177).
- [230] Daniel Schraivogel et al. “Targeted Perturb-seq enables genome-scale genetic screens in single cells”. *Nature Methods* 17 (2020), pp. 629–635. DOI: [10.1038/s41592-020-0837-5](https://doi.org/10.1038/s41592-020-0837-5).
- [231] Thomas M. Norman et al. “Exploring genetic interaction manifolds constructed from rich single-cell phenotypes”. *Science* 365 (2019), pp. 786–793. DOI: [10.1126/science.aax4438](https://doi.org/10.1126/science.aax4438).
- [232] Mary J Dunlop et al. “Regulatory activity revealed by dynamic correlations in gene expression noise”. *Nature Genetics* 40 (2008), pp. 1493–1498. DOI: [10.1038/ng.281](https://doi.org/10.1038/ng.281).
- [233] Aryeh Warmflash and Aaron R. Dinner. “Signatures of combinatorial regulation in intrinsic biological noise”. *Proceedings of the National Academy of Sciences* 105 (2008), pp. 17262–17267. DOI: [10.1073/pnas.0809314105](https://doi.org/10.1073/pnas.0809314105).
- [234] Hannah Dueck, James Eberwine, and Junhyong Kim. “Variation is function: Are single cell differences functionally important?: Testing the hypothesis that single cell variation is required for aggregate function”. *BioEssays* 38 (2015), pp. 172–180. DOI: [10.1002/bies.201500124](https://doi.org/10.1002/bies.201500124).
- [235] Thalia E. Chan, Michael P.H. Stumpf, and Ann C. Babbie. “Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures”. *Cell Systems* 5 (2017), pp. 251–267. DOI: [10.1016/j.cels.2017.08.014](https://doi.org/10.1016/j.cels.2017.08.014).
- [236] Euan Joly-Smith et al. “Exploiting fluctuations in gene expression to detect causal interactions between genes”. *eLife* (2024). DOI: [10.7554/eLife.92497.1](https://doi.org/10.7554/eLife.92497.1).
- [237] Pau Badia-i-Mompel et al. “Gene regulatory network inference in the era of single-cell multi-omics”. *Nature Reviews Genetics* (2023), pp. 739–754. DOI: [10.1038/s41576-023-00618-5](https://doi.org/10.1038/s41576-023-00618-5).

- [238] Daniel Kim et al. “Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data”. *npj Systems Biology and Applications* 9 (2023), p. 51. DOI: [10.1038/s41540-023-00312-6](https://doi.org/10.1038/s41540-023-00312-6).
- [239] Alicia T Specht and Jun Li. “LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering”. *Bioinformatics* 33 (2016), pp. 764–766. DOI: [10.1093/bioinformatics/btw729](https://doi.org/10.1093/bioinformatics/btw729).
- [240] Sara Aibar et al. “SCENIC: single-cell regulatory network inference and clustering”. *Nature Methods* 14 (2017), pp. 1083–1086. DOI: [10.1038/nmeth.4463](https://doi.org/10.1038/nmeth.4463).
- [241] David van Dijk et al. “Recovering Gene Interactions from Single-Cell Data Using Data Diffusion”. *Cell* 174 (2018), pp. 716–729. DOI: [10.1016/j.cell.2018.05.061](https://doi.org/10.1016/j.cell.2018.05.061).
- [242] Xiaojie Qiu et al. “Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe”. *Cell Systems* 10 (2020), pp. 265–274. DOI: [10.1016/j.cels.2020.02.003](https://doi.org/10.1016/j.cels.2020.02.003).
- [243] Silvia Giulia Galfrè et al. “COTAN: scRNA-seq data analysis based on gene co-expression”. *NAR Genomics and Bioinformatics* 3 (2021), lqab072. DOI: [10.1093/nargab/lqab072](https://doi.org/10.1093/nargab/lqab072).
- [244] Wei Vivian Li and Yanzeng Li. “scLink: Inferring Sparse Gene Co-expression Networks from Single-cell Expression Data”. *Genomics, Proteomics & Bioinformatics* 19 (2021), pp. 475–492. DOI: [10.1016/j.gpb.2020.11.006](https://doi.org/10.1016/j.gpb.2020.11.006).
- [245] Kai Silkwood et al. “Leveraging gene correlations in single cell transcriptomic data”. *bioRxiv* (2023). DOI: [10.1101/2023.03.14.532643](https://doi.org/10.1101/2023.03.14.532643).
- [246] Ulysse Herbach et al. “Inferring gene regulatory networks from single-cell data: a mechanistic approach”. *BMC Systems Biology* 11 (2017), p. 105. DOI: [10.1186/s12918-017-0487-0](https://doi.org/10.1186/s12918-017-0487-0).
- [247] Arnaud Bonnafeux et al. “WASABI: a dynamic iterative framework for gene regulatory network inference”. *BMC Bioinformatics* 20 (2019), p. 220. DOI: [10.1186/s12859-019-2798-1](https://doi.org/10.1186/s12859-019-2798-1).
- [248] Belinda Phipson, Luke Zappia, and Alicia Oshlack. “Gene length and detection bias in single cell RNA sequencing protocols”. *F1000 Research* 6 (2017), p. 595. DOI: [10.12688/f1000research.11290.1](https://doi.org/10.12688/f1000research.11290.1).

- [249] Gennady Gorin and Lior Pachter. “Length biases in single-cell RNA sequencing of pre-mRNA”. *Biophysical Reports* 3 (2023), p. 100097. DOI: [10.1016/j.bpr.2022.100097](https://doi.org/10.1016/j.bpr.2022.100097).
- [250] Vân Anh Huynh-Thu and Guido Sanguinetti. “Gene Regulatory Network Inference: An Introductory Survey”. *Gene Regulatory Networks*. Springer New York, 2018, pp. 1–23. DOI: [10.1007/978-1-4939-8882-2_1](https://doi.org/10.1007/978-1-4939-8882-2_1).
- [251] Michael P.H. Stumpf. “Inferring better gene regulation networks from single-cell data”. *Current Opinion in Systems Biology* 27 (2021), p. 100342. DOI: [10.1016/j.coisb.2021.05.003](https://doi.org/10.1016/j.coisb.2021.05.003).
- [252] Andreas Christ Sølvsten Jørgensen et al. “Bayesian model discovery for reverse-engineering biochemical networks from data”. *bioRxiv* (2023). DOI: [10.1101/2023.09.15.557764](https://doi.org/10.1101/2023.09.15.557764).
- [253] M. G. B. Blum et al. “A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation”. *Statistical Science* 28 (2013), pp. 189–208. DOI: [10.1214/12-sts406](https://doi.org/10.1214/12-sts406).
- [254] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL: <http://probml.github.io/book2>.
- [255] Ramon Grima. “Linear-noise approximation and the chemical master equation agree up to second-order moments for a class of chemical systems”. *Physical Review E* 92 (2015), p. 042124. DOI: [10.1103/physreve.92.042124](https://doi.org/10.1103/physreve.92.042124).
- [256] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. *Machine Learning* 20 (1995), pp. 273–297. DOI: [10.1007/bf00994018](https://doi.org/10.1007/bf00994018).
- [257] Tin Kam Ho. “Random decision forests”. *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1, pp. 278–282. DOI: [10.1109/icdar.1995.598994](https://doi.org/10.1109/icdar.1995.598994).
- [258] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: <http://www.deeplearningbook.org>.
- [259] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. “Multilayer feedforward networks are universal approximators”. *Neural Networks* 2 (1989), pp. 359–366. DOI: [10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8).
- [260] Dmitry Yarotsky. “Universal approximations of invariant maps by neural networks”. *arXiv* (2018). DOI: [10.48550/arXiv.1804.10306](https://doi.org/10.48550/arXiv.1804.10306).

- [261] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Vol. 9. 2010, pp. 249–256. URL: <http://proceedings.mlr.press/v9/glorot10a.html>.
- [262] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. *arXiv* (2017). DOI: [10.48550/arXiv.1412.6980](https://doi.org/10.48550/arXiv.1412.6980). eprint: [1412.6980](https://arxiv.org/abs/1412.6980).
- [263] Yoav Benjamini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 57 (1995), pp. 289–300. DOI: [10.1111/j.2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x).
- [264] Rose Oughtred et al. “The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions”. *Protein Science* 30 (2020), pp. 187–200. DOI: [10.1002/pro.3978](https://doi.org/10.1002/pro.3978).
- [265] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources”. *Nature Protocols* 4 (2008), pp. 44–57. DOI: [10.1038/nprot.2008.211](https://doi.org/10.1038/nprot.2008.211).
- [266] Behnam Neyshabur et al. “Exploring Generalization in Deep Learning”. *arXiv* (2017). DOI: [10.48550/arXiv.1706.08947](https://doi.org/10.48550/arXiv.1706.08947).
- [267] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. *arXiv* (2016). DOI: [10.48550/arXiv.1602.04938](https://doi.org/10.48550/arXiv.1602.04938).
- [268] W. James Murdoch et al. “Definitions, methods, and applications in interpretable machine learning”. *Proceedings of the National Academy of Sciences* 116 (2019), pp. 22071–22080. DOI: [10.1073/pnas.1900654116](https://doi.org/10.1073/pnas.1900654116).
- [269] David Gunning et al. “XAI—Explainable artificial intelligence”. *Science Robotics* 4 (2019), eaay7120. DOI: [10.1126/scirobotics.aay7120](https://doi.org/10.1126/scirobotics.aay7120).
- [270] David Baehrens et al. “How to Explain Individual Classification Decisions”. *arXiv* (2009). DOI: [10.48550/arXiv.0912.1128](https://doi.org/10.48550/arXiv.0912.1128).
- [271] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic Attribution for Deep Networks”. *arXiv* (2017). DOI: [10.48550/arXiv.1703.01365](https://doi.org/10.48550/arXiv.1703.01365).

- [272] Daniel Morgan et al. “A generalized framework for controlling FDR in gene regulatory network inference”. *Bioinformatics* 35 (2018), pp. 1026–1032. DOI: [10.1093/bioinformatics/bty764](https://doi.org/10.1093/bioinformatics/bty764).
- [273] Eric Kernfeld et al. “Transcriptome data are insufficient to control false discoveries in regulatory network inference”. *Cell Systems* 15 (2024), pp. 709–724. DOI: [10.1016/j.cels.2024.07.006](https://doi.org/10.1016/j.cels.2024.07.006).
- [274] Lingfeng Xue, Yan Wu, and Yihan Lin. “Dissecting and improving gene regulatory network inference using single-cell transcriptome data”. *Genome Research* 33 (2023), pp. 1609–1621. DOI: [10.1101/gr.277488.122](https://doi.org/10.1101/gr.277488.122).
- [275] E Alm. “Biological networks”. *Current Opinion in Structural Biology* 13 (2003), pp. 193–202. DOI: [10.1016/s0959-440x\(03\)00031-9](https://doi.org/10.1016/s0959-440x(03)00031-9).
- [276] Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, 2006. DOI: [10.1201/9781420011432](https://doi.org/10.1201/9781420011432).
- [277] Daniel T. Gillespie. “The chemical Langevin equation”. *The Journal of Chemical Physics* 113 (2000), pp. 297–306. DOI: [10.1063/1.481811](https://doi.org/10.1063/1.481811).
- [278] Philipp Thomas and Ramon Grima. “Approximate probability distributions of the master equation”. *Physical Review E* 92 (2015). DOI: [10.1103/physreve.92.012120](https://doi.org/10.1103/physreve.92.012120).
- [279] Angelique Ale, Paul Kirk, and Michael P. H. Stumpf. “A general moment expansion method for stochastic kinetic models”. *The Journal of Chemical Physics* 138 (2013), p. 174101. DOI: [10.1063/1.4802475](https://doi.org/10.1063/1.4802475).
- [280] Augustinas Sukys, Kaan Öcal, and Ramon Grima. “Approximating solutions of the Chemical Master equation using neural networks”. *iScience* 25 (2022), p. 105010. DOI: [10.1016/j.isci.2022.105010](https://doi.org/10.1016/j.isci.2022.105010).
- [281] Mohammad Soltani, Cesar Augusto Vargas-Garcia, and Abhyudai Singh. “Conditional Moment Closure Schemes for Studying Stochastic Dynamics of Genetic Circuits”. *IEEE Transactions on Biomedical Circuits and Systems* 9 (2015), pp. 518–526. DOI: [10.1109/TBCAS.2015.2453158](https://doi.org/10.1109/TBCAS.2015.2453158).
- [282] Christopher V. Rao and Adam P. Arkin. “Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm”. *The Journal of Chemical Physics* 118 (2003), pp. 4999–5010. DOI: [10.1063/1.1545446](https://doi.org/10.1063/1.1545446).

- [283] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. “The slow-scale stochastic simulation algorithm”. *The Journal of Chemical Physics* 122 (2005), p. 014116. DOI: [10.1063/1.1824902](https://doi.org/10.1063/1.1824902).
- [284] Eric L. Haseltine and James B. Rawlings. “Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics”. *The Journal of Chemical Physics* 117 (2002), pp. 6959–6969. DOI: [10.1063/1.1505860](https://doi.org/10.1063/1.1505860).
- [285] C. W. Gardiner. “Adiabatic elimination in stochastic systems. I. Formulation of methods and application to few-variable systems”. *Physical Review A* 29 (1984), pp. 2814–2822. DOI: [10.1103/physreva.29.2814](https://doi.org/10.1103/physreva.29.2814).
- [286] Tatsuo Shibata. “Reducing the master equations for noisy chemical reactions”. *The Journal of Chemical Physics* 119 (2003), pp. 6629–6634. DOI: [10.1063/1.1603738](https://doi.org/10.1063/1.1603738).
- [287] Simone Pigolotti and Angelo Vulpiani. “Coarse graining of master equations with fast and slow states”. *The Journal of Chemical Physics* 128 (2008), p. 154114. DOI: [10.1063/1.2907242](https://doi.org/10.1063/1.2907242).
- [288] Philipp Thomas, Ramon Grima, and Arthur V. Straube. “Rigorous elimination of fast stochastic variables from the linear noise approximation using projection operators”. *Phys. Rev. E* 86 (2012), p. 041110. DOI: [10.1103/PhysRevE.86.041110](https://doi.org/10.1103/PhysRevE.86.041110).
- [289] Philipp Thomas, Arthur V Straube, and Ramon Grima. “The slow-scale linear noise approximation: an accurate, reduced stochastic description of biochemical networks under timescale separation conditions”. *BMC Systems Biology* 6 (2012), p. 39. DOI: [10.1186/1752-0509-6-39](https://doi.org/10.1186/1752-0509-6-39).
- [290] Leo Bronstein and Heinz Koepl. “Marginal process framework: A model reduction tool for Markov jump processes”. *Phys. Rev. E* 97 (2018), p. 062147. DOI: [10.1103/PhysRevE.97.062147](https://doi.org/10.1103/PhysRevE.97.062147).
- [291] Zhixing Cao and Ramon Grima. “Linear mapping approximation of gene regulatory networks with stochastic dynamics”. *Nature Communications* 9 (2018), p. 3305. DOI: [10.1038/s41467-018-05822-0](https://doi.org/10.1038/s41467-018-05822-0).
- [292] Chen Jia and Ramon Grima. “Holimap: an accurate and efficient method for solving stochastic gene network dynamics”. *Nature Communications* 15 (2024), p. 6557. DOI: [10.1038/s41467-024-50716-z](https://doi.org/10.1038/s41467-024-50716-z).

- [293] C. Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer Berlin Heidelberg, 2009.
- [294] H.A. Kramers. “Brownian motion in a field of force and the diffusion model of chemical reactions”. *Physica* 7 (1940), pp. 284–304. DOI: [10.1016/s0031-8914\(40\)90098-2](https://doi.org/10.1016/s0031-8914(40)90098-2).
- [295] J. E. Moyal. “Stochastic Processes and Statistical Physics”. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 11 (1949), pp. 150–210. DOI: [10.1111/j.2517-6161.1949.tb00030.x](https://doi.org/10.1111/j.2517-6161.1949.tb00030.x).
- [296] Thomas G. Kurtz. *Limit theorems and diffusion approximations for density dependent Markov chains*. Springer Berlin Heidelberg, 1976, pp. 67–78. DOI: [10.1007/BFb0120765](https://doi.org/10.1007/BFb0120765).
- [297] Ramon Grima, Philipp Thomas, and Arthur V. Straube. “How accurate are the nonlinear chemical Fokker-Planck and chemical Langevin equations?” *The Journal of Chemical Physics* 135 (2011), p. 084103. DOI: [10.1063/1.3625958](https://doi.org/10.1063/1.3625958).
- [298] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer Berlin Heidelberg, 1992. DOI: [10.1007/978-3-662-12616-5](https://doi.org/10.1007/978-3-662-12616-5).
- [299] Juan Kuntz, Diego Oyarzún, and Guy-Bart Stan. “Model Reduction of Genetic-Metabolic Networks via Time Scale Separation”. *A Systems Theoretic Approach to Systems and Synthetic Biology I: Models and System Characterizations*. Springer Netherlands, 2014, pp. 181–210. DOI: [10.1007/978-94-017-9041-3_7](https://doi.org/10.1007/978-94-017-9041-3_7).
- [300] Caroline C. Friedel et al. “Conserved principles of mammalian transcriptional regulation revealed by RNA half-life”. *Nucleic Acids Research* 37 (2009). DOI: [10.1093/nar/gkp542](https://doi.org/10.1093/nar/gkp542).
- [301] Reinhart Heinrich and Stefan Schuster. *The Regulation of Cellular Systems*. Springer US, 1996. DOI: [10.1007/978-1-4613-1161-4](https://doi.org/10.1007/978-1-4613-1161-4).
- [302] J. D. Murray. *Mathematical Biology: I. An Introduction*. Springer New York, 2002. DOI: [10.1007/b98868](https://doi.org/10.1007/b98868).
- [303] Philipp Thomas, Arthur V. Straube, and Ramon Grima. “Stochastic theory of large-scale enzyme-reaction networks: Finite copy number corrections to rate equation models”. *The Journal of Chemical Physics* 133 (2010), p. 195101. DOI: [10.1063/1.3505552](https://doi.org/10.1063/1.3505552).

- [304] Morten Gram Pedersen et al. “The total quasi-steady-state approximation for complex enzyme reactions”. *Mathematics and Computers in Simulation* 79 (2008), pp. 1010–1019. DOI: [10.1016/j.matcom.2008.02.009](https://doi.org/10.1016/j.matcom.2008.02.009).
- [305] Tatsuo Shibata. “Fluctuating reaction rates and their application to problems of gene expression”. *Phys. Rev. E* 67 (2003), p. 061906. DOI: [10.1103/PhysRevE.67.061906](https://doi.org/10.1103/PhysRevE.67.061906).
- [306] Carlos A. Gómez-Urbe, George C. Verghese, and Abraham R. Tzafriri. “Enhanced identification and exploitation of time scales for model reduction in stochastic chemical kinetics”. *The Journal of Chemical Physics* 129 (2008), p. 244112. DOI: [10.1063/1.3050350](https://doi.org/10.1063/1.3050350).
- [307] John Goutsias. “Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems”. *The Journal of Chemical Physics* 122 (2005), p. 184102. DOI: [10.1063/1.1889434](https://doi.org/10.1063/1.1889434). URL: <https://doi.org/10.1063/1.1889434>.
- [308] Pavol Bokes et al. “Multiscale stochastic modelling of gene expression”. *Journal of Mathematical Biology* 65 (2011), pp. 493–520. DOI: [10.1007/s00285-011-0468-7](https://doi.org/10.1007/s00285-011-0468-7).
- [309] Michael W. Chevalier and Hana El-Samad. “A rigorous framework for multiscale simulation of stochastic cellular networks”. *The Journal of Chemical Physics* 131 (2009), p. 054102. DOI: [10.1063/1.3190327](https://doi.org/10.1063/1.3190327).
- [310] Simon L. Cotter et al. “A constrained approach to multiscale stochastic simulation of chemically reacting systems”. *The Journal of Chemical Physics* 135 (2011), pp. 94–102. DOI: [10.1063/1.3624333](https://doi.org/10.1063/1.3624333).
- [311] Marie-Nathalie Contou-Carrere et al. “Model reduction of multi-scale chemical Langevin equations”. *Systems & Control Letters* 60 (2011), pp. 75–86. DOI: [10.1016/j.sysconle.2010.10.011](https://doi.org/10.1016/j.sysconle.2010.10.011).
- [312] Fuke Wu et al. “Approximate method for stochastic chemical kinetics with two-time scales by chemical Langevin equations”. *The Journal of Chemical Physics* 144 (2016), p. 174112. DOI: [10.1063/1.4948407](https://doi.org/10.1063/1.4948407).
- [313] Weinan E, Di Liu, and Eric Vanden-Eijnden. “Nested stochastic simulation algorithms for chemical kinetic systems with multiple time scales”. *Journal of Computational Physics* 221 (2007), pp. 158–180. DOI: [10.1016/j.jcp.2006.06.019](https://doi.org/10.1016/j.jcp.2006.06.019).

- [314] Yang Cao, Dan Gillespie, and Linda Petzold. “Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems”. *Journal of Computational Physics* 206 (2005), pp. 395–411. DOI: [10.1016/j.jcp.2004.12.014](https://doi.org/10.1016/j.jcp.2004.12.014).
- [315] Justin Eilertsen and Santiago Schnell. “The quasi-steady-state approximations revisited: Timescales, small parameters, singularities, and normal forms in enzyme kinetics”. *Mathematical Biosciences* 325 (2020), p. 108339. DOI: [10.1016/j.mbs.2020.108339](https://doi.org/10.1016/j.mbs.2020.108339).
- [316] I. Stoleriu, F.A. Davidson, and J.L. Liu. “Quasi-steady state assumptions for non-isolated enzyme-catalysed reactions”. *Journal of Mathematical Biology* 48 (2004), pp. 82–104. DOI: [10.1007/s00285-003-0225-7](https://doi.org/10.1007/s00285-003-0225-7).
- [317] Ethan A. Mastny, Eric L. Haseltine, and James B. Rawlings. “Two classes of quasi-steady-state model reductions for stochastic kinetics”. *The Journal of Chemical Physics* 127 (2007), p. 094106. DOI: [10.1063/1.2764480](https://doi.org/10.1063/1.2764480).
- [318] Philipp Thomas, Arthur V. Straube, and Ramon Grima. “Communication: Limitations of the stochastic quasi-steady-state approximation in open biochemical reaction networks”. *The Journal of Chemical Physics* 135 (2011), p. 181103. DOI: [10.1063/1.3661156](https://doi.org/10.1063/1.3661156).
- [319] Jae Kyoung Kim, Krešimir Josić, and Matthew R. Bennett. “The relationship between stochastic and deterministic quasi-steady state approximations”. *BMC Systems Biology* 9 (2015), p. 87. DOI: [10.1186/s12918-015-0218-3](https://doi.org/10.1186/s12918-015-0218-3).
- [320] Grigorios A Pavliotis and Andrew M Stuart. *Multiscale methods: Averaging and Homogenization*. Springer New York, 2008. DOI: [10.1007/978-0-387-73829-1](https://doi.org/10.1007/978-0-387-73829-1).
- [321] Eran Eden et al. “Proteome Half-Life Dynamics in Living Human Cells”. *Science* 331 (2011), pp. 764–768. DOI: [10.1126/science.1199784](https://doi.org/10.1126/science.1199784).
- [322] Debashis Barik et al. “Stochastic Simulation of Enzyme-Catalyzed Reactions with Disparate Timescales”. *Biophysical Journal* 95 (2008), pp. 3563–3574. DOI: [10.1529/biophysj.108.129155](https://doi.org/10.1529/biophysj.108.129155).
- [323] Jae Kyoung Kim, Krešimir Josić, and Matthew R. Bennett. “The Validity of Quasi-Steady-State Approximations in Discrete Stochastic Simulations”. *Biophysical Journal* 107 (2014), pp. 783–793. DOI: [10.1016/j.bpj.2014.06.012](https://doi.org/10.1016/j.bpj.2014.06.012).

- [324] Yun Min Song, Hyukpyo Hong, and Jae Kyoung Kim. “Universally valid reduction of multiscale stochastic biochemical systems using simple non-elementary propensities”. *PLOS Computational Biology* 17 (2021), e1008952. DOI: [10.1371/journal.pcbi.1008952](https://doi.org/10.1371/journal.pcbi.1008952).
- [325] Yueheng Lan, Timothy C. Elston, and Garegin A. Papoian. “Elimination of fast variables in chemical Langevin equations”. *The Journal of Chemical Physics* 129 (2008), p. 214115. DOI: [10.1063/1.3027499](https://doi.org/10.1063/1.3027499).
- [326] Attila Becskei and Luis Serrano. “Engineering stability in gene networks by autoregulation”. *Nature* 405 (2000), pp. 590–593. DOI: [10.1038/35014651](https://doi.org/10.1038/35014651).
- [327] James Holehouse, Zhixing Cao, and Ramon Grima. “Stochastic Modeling of Autoregulatory Genetic Feedback Loops: A Review and Comparative Study”. *Biophysical Journal* 118 (2020), pp. 1517–1525. DOI: [10.1016/j.bpj.2020.02.016](https://doi.org/10.1016/j.bpj.2020.02.016).
- [328] Abhyudai Singh and Joao P. Hespanha. “Optimal Feedback Strength for Noise Suppression in Autoregulatory Gene Networks”. *Biophysical Journal* 96 (2009), pp. 4013–4023. DOI: [10.1016/j.bpj.2009.02.064](https://doi.org/10.1016/j.bpj.2009.02.064).
- [329] Chen Jia et al. “Stochastic fluctuations can reveal the feedback signs of gene regulatory networks at the single-molecule level”. *Scientific Reports* 7 (2017), p. 16037. DOI: [10.1038/s41598-017-15464-9](https://doi.org/10.1038/s41598-017-15464-9).
- [330] Joshua L Cherry and Frederick R Adler. “How to make a Biological Switch”. *Journal of Theoretical Biology* 203 (2000), pp. 117–133. DOI: [10.1006/jtbi.2000.1068](https://doi.org/10.1006/jtbi.2000.1068).
- [331] Stephen T. Crews and Joseph C. Pearson. “Transcriptional autoregulation in development”. *Current Biology* 19 (2009), R241–R246. DOI: [10.1016/j.cub.2009.01.015](https://doi.org/10.1016/j.cub.2009.01.015).
- [332] Michael B. Elowitz and Stanislas Leibler. “A synthetic oscillatory network of transcriptional regulators”. *Nature* 403 (2000), pp. 335–338. DOI: [10.1038/35002125](https://doi.org/10.1038/35002125).
- [333] Didier Gonze and Wassim Abou-Jaoudé. “The Goodwin Model: Behind the Hill Function”. *PLoS ONE* 8 (2013), e69573. DOI: [10.1371/journal.pone.0069573](https://doi.org/10.1371/journal.pone.0069573).
- [334] Philipp Thomas et al. “Signatures of nonlinearity in single cell noise-induced oscillations”. *Journal of Theoretical Biology* 335 (2013), pp. 222–234. DOI: [10.1016/j.jtbi.2013.06.021](https://doi.org/10.1016/j.jtbi.2013.06.021).

- [335] Jae Kyoung Kim. “Tick, Tock, Circadian Clocks”. *Case Studies in Systems Biology*. Springer International Publishing, 2021, pp. 79–94. DOI: [10.1007/978-3-030-67742-8_6](https://doi.org/10.1007/978-3-030-67742-8_6).
- [336] Aleksandra Eremina et al. “Environmental and molecular noise buffering by the cyanobacterial clock in individual cells”. *bioRxiv* (2024). DOI: [10.1101/2024.08.02.605997](https://doi.org/10.1101/2024.08.02.605997).
- [337] Béla Novák and John J. Tyson. “Design principles of biochemical oscillators”. *Nature Reviews Molecular Cell Biology* 9 (2008), pp. 981–991. DOI: [10.1038/nrm2530](https://doi.org/10.1038/nrm2530).
- [338] Philipp Thomas, Hannes Matuschek, and Ramon Grima. “How reliable is the linear noise approximation of gene regulatory networks?” *BMC Genomics* 14 (2013), S5. DOI: [10.1186/1471-2164-14-s4-s5](https://doi.org/10.1186/1471-2164-14-s4-s5).
- [339] Nir Friedman. “Inferring Cellular Networks Using Probabilistic Graphical Models”. *Science* 303 (2004), pp. 799–805. DOI: [10.1126/science.1094068](https://doi.org/10.1126/science.1094068).
- [340] Stephen Y Zhang and Michael P H Stumpf. “Learning cell-specific networks from dynamics and geometry of single cells”. *bioRxiv* (2023). DOI: [10.1101/2023.01.08.523176](https://doi.org/10.1101/2023.01.08.523176).
- [341] Barbara Bravi, Katy J. Rubin, and Peter Sollich. “Systematic model reduction captures the dynamics of extrinsic noise in biochemical subnetworks”. *The Journal of Chemical Physics* 153 (2020), p. 025101. DOI: [10.1063/5.0008304](https://doi.org/10.1063/5.0008304).
- [342] Cole Trapnell. “Revealing gene function with statistical inference at single-cell resolution”. *Nature Reviews Genetics* 25 (2024), pp. 623–638. DOI: [10.1038/s41576-024-00750-w](https://doi.org/10.1038/s41576-024-00750-w).
- [343] Darren A. Cusanovich et al. “Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing”. *Science* 348 (2015), pp. 910–914. DOI: [10.1126/science.aab1601](https://doi.org/10.1126/science.aab1601).
- [344] Sébastien A Smallwood et al. “Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity”. *Nature Methods* 11 (2014), pp. 817–820. DOI: [10.1038/nmeth.3035](https://doi.org/10.1038/nmeth.3035).
- [345] Hayley M. Bennett et al. “Single-cell proteomics enabled by next-generation sequencing or mass spectrometry”. *Nature Methods* 20 (2023), pp. 363–374. DOI: [10.1038/s41592-023-01791-5](https://doi.org/10.1038/s41592-023-01791-5).

- [346] Lambda Moses and Lior Pachter. “Museum of spatial transcriptomics”. *Nature Methods* 19 (2022), pp. 534–546. DOI: [10.1038/s41592-022-01409-2](https://doi.org/10.1038/s41592-022-01409-2).
- [347] Tian Lu, Cheen Euong Ang, and Xiaowei Zhuang. “Spatially resolved epigenomic profiling of single cells in complex tissues”. *Cell* 185 (2022), pp. 4448–4464. DOI: [10.1016/j.cell.2022.09.035](https://doi.org/10.1016/j.cell.2022.09.035).
- [348] Marlon Stoeckius et al. “Simultaneous epitope and transcriptome measurement in single cells”. *Nature Methods* 14 (2017), pp. 865–868. DOI: [10.1038/nmeth.4380](https://doi.org/10.1038/nmeth.4380).
- [349] Junyue Cao et al. “Joint profiling of chromatin accessibility and gene expression in thousands of single cells”. *Science* 361 (2018), pp. 1380–1385. DOI: [10.1126/science.aau0730](https://doi.org/10.1126/science.aau0730).
- [350] Jeongwoo Lee, Do Young Hyeon, and Daehee Hwang. “Single-cell multiomics: technologies and data analysis methods”. *Experimental & Molecular Medicine* 52 (2020), pp. 1428–1442. DOI: [10.1038/s12276-020-0420-2](https://doi.org/10.1038/s12276-020-0420-2).
- [351] Katy Vandereyken et al. “Methods and applications for single-cell and spatial multiomics”. *Nature Reviews Genetics* 24 (2023), pp. 494–515. DOI: [10.1038/s41576-023-00580-2](https://doi.org/10.1038/s41576-023-00580-2).
- [352] Lukas Heumos et al. “Best practices for single-cell analysis across modalities”. *Nature Reviews Genetics* 24 (2023), pp. 550–572. DOI: [10.1038/s41576-023-00586-w](https://doi.org/10.1038/s41576-023-00586-w).
- [353] Philipp Thomas. “Stochastic Modeling Approaches for Single-Cell Analyses”. *Systems Medicine*. Elsevier, 2021, pp. 45–55. DOI: [10.1016/b978-0-12-801238-3.11539-9](https://doi.org/10.1016/b978-0-12-801238-3.11539-9).
- [354] Catherine Felce, Gennady Gorin, and Lior Pachter. “A Biophysical Model for ATAC-seq Data Analysis”. *bioRxiv* (2024). DOI: [10.1101/2024.01.25.577262](https://doi.org/10.1101/2024.01.25.577262).
- [355] Jun Ding, Nadav Sharon, and Ziv Bar-Joseph. “Temporal modelling using single-cell transcriptomics”. *Nature Reviews Genetics* 23 (2022), pp. 355–368. DOI: [10.1038/s41576-021-00444-7](https://doi.org/10.1038/s41576-021-00444-7).
- [356] Diederik P Kingma and Max Welling. “Auto-Encoding Variational Bayes”. *arXiv* (2022). DOI: [10.48550/arXiv.1312.6114](https://doi.org/10.48550/arXiv.1312.6114).
- [357] Yanzhi Chen et al. “Neural Approximate Sufficient Statistics for Implicit Models”. *arXiv* (2021). DOI: [10.48550/arXiv.2010.10079](https://doi.org/10.48550/arXiv.2010.10079).

- [358] David T. Frazier and Christopher Drovandi. “Robust Approximate Bayesian Inference with Synthetic Likelihood”. *arXiv* (2020). DOI: [10.48550/arXiv.1904.04551](https://doi.org/10.48550/arXiv.1904.04551).