Stochastic foundations for single-cell RNA sequencing

Thesis by
Gennady Gorin

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy in Chemical Engineering

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2023
Defended May 19, 2023

Gennady Gorin
ORCID: 0000-0001-6097-2029

ACKNOWLEDGEMENTS

Foremost thanks to Lior Pachter, my advisor, who took a chance by accepting me, provided unwavering support throughout the program, and cast the strongest vote of confidence by recruiting capable junior students to collaborate on and continue the research programme outlined in this thesis. Thanks to my thesis committee, including Mikhail Shapiro, Rustem Ismagilov, Shasha Chong, and Zhen-Gang Wang, for their guidance, which inspired deep and careful thought and improved the work in important ways.

I am grateful to the Pachter laboratory, especially my direct collaborators on projects outlined here: Maria Carilli, Tara Chari, Meichen Fang, Catherine Felce, and Kayla Jackson; collaborators formerly in Pachter lab: Valentine Svensson and Shawn Yoshida; people who were otherwise close or immediately helpful throughout the Ph.D.: Sina Booeshaghi, Taleen Dilanyan, Anne Yeokyoung Kil, Lambda Moses, Delaney Sullivan, Laura Luebbert, and Kristján Eldjárn Hjörleifsson. This unique and exciting subfield is certainly in good hands.

Tremendous thanks to John J. Vastola for a multi-year collaboration responsible for the most interesting, impactful, personally instructive, and mathematically stimulating content in this thesis. Every seemingly simple theoretical result presented here is the distillation of years of persistent detail work. Many of these research directions would simply not have yielded results without his input; many others would be impossible without the mathematical, scientific, and communication skills in which this collaboration has trained me. Thanks to Yongin Choi, without whose contributions and support the machine learning projects — whose foundations have been laid, but whose scope is a new and wide open frontier — would never have happened. My gratitude to both of them, of course, for fully sharing in the graduate student experience with me.

Thanks to the National Institutes of Health, whose funding made this research possible.

Thanks to the researchers at Celsius Therapeutics, especially Nico Stransky, Tommy Boucher, and Mukund Varma, for their help and support throughout my internship, which enabled me to answer some fundamental questions and fully conceptualize their relevance to the applications of this work.

Several pre-Caltech mentors stand out as lasting influences. Special thanks to

Drs. Ido Golding and Heng Xu for closely supervising and teaching me, guiding my first research project to its completion and publication, and providing careful and detailed perspectives on Ph.D. projects (especially the first tentative step into the real complexities of biophysical modeling I undertook in “Interpretable and Tractable...”), and ultimately giving me most, perhaps all, of the tools I needed for this research programme.

Thanks to the Asian Studies department at Rice University, especially Drs. Richard J. Smith, Nanxiu Qian, Lisa Balabanlilar, and Susan Huang. Their impacts were innumerable; any writing skills I have can be safely attributed to them.

This impact of the humanities reminds that research is not only the abstract manipulation of data and mathematical symbols, but a lived experience. A dissertation is a snapshot of this lived experience, constrained to a structured set of insights and questions about science. It would be remiss of me not to mention the artistic influences who have directly shaped this lived experience and changed my perspectives, assumptions, priorities, and worldviews, even if their influence is subtle and indirect, in various media: Edward Tufte; the Museum of Jurassic Technology and Meow Wolf; the Documerica Project, Bernd and Hilla Becher; Steely Dan, Fat Tony, Moondog, Cab Calloway, and U.G.K.; ZA/UM and Atlus; the Van Beuren and Fleischer studios; Tati, Lynch, Kusturica, and Marx brothers; Proust, Ligotti, Eco, Borges, Hodgson, Aickman, Chiang, and Bolaño; Roueché, Saviano, and Thompson.

I am grateful to various large and small groups I have been part of: my chemical engineering Ph.D. cohort (especially Bobby Grayson); the book club; the cryptic crossword community; IMPLiCIT and TACIT at Caltech. The experience of the Ph.D. would have been pale and bleak without them.

Thanks to my family, especially my mother, an inspiration and support.

Finally, thanks to Riley Smith, who has been there for me the whole time (and Griswold, who has not).

ABSTRACT

Single-cell RNA sequencing, which quantifies cell transcriptomes, has seen widespread adoption, accompanied by proliferation of analysis methods. However, there has been relatively little systematic investigation of its best practices and their underlying assumptions, leading to challenges and discrepancies in interpretation. I present a set of generic, principled strategies for modeling the biological and technical components of sequencing experiments and use case studies to motivate their application to sequencing data.

PUBLISHED CONTENT AND CONTRIBUTIONS

The materials in the current thesis are largely drawn from the manuscripts listed below. Where relevant, the headings of sections credit the contributors for specific ideas, with the abbreviations as follows: L.P.: Lior Pachter, V.S.: Valentine Svensson, J.J.V.: John J. Vastola, M.F.: Meichen Fang, M.C.: Maria Carilli, T.C.: Tara Chari, C.F.: Catherine Felce, Y.C.: Yongin Choi, S.Y.: Shawn Yoshida, K.J.: Kayla Jackson. Although not indicated in the list below, G.G. was co-first author on manuscripts 1, 9, and 11 with M.C. and J.J.V.

[1] Maria T. Carilli, Gennady Gorin, Yongin Choi, Tara Chari, and Lior Pachter. Mechanistic modeling with a variational autoencoder for multimodal single-cell RNA sequencing data. Preprint, bioRxiv: 2023.01.13.523995, January 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.01.13.523995.
G.G. conceptualized the research, and participated in the study design, theoretical derivations, and writing of the manuscript.

[2] Gennady Gorin and Lior Pachter. Intrinsic and extrinsic noise are distinguishable in a synthesis – export – degradation model of mRNA production. Preprint, bioRxiv: 2020.09.25.312868, September 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.09.25.312868.
G.G. derived and implemented the analytical results, validated them against simulations, and participated in the research conceptualization and writing of the manuscript.

[3] Gennady Gorin and Lior Pachter. Special function methods for bursty models of transcription. *Physical Review E*, 102(2):022409, August 2020. ISSN 2470-0045, 2470-0053. doi: 10.1103/PhysRevE.102.022409. URL https://link.aps.org/doi/10.1103/PhysRevE.102.022409.
G.G. conceptualized, designed, and implemented the special function approximations, constructed the performance benchmarks, and participated in the writing of the manuscript.

[4] Gennady Gorin and Lior Pachter. Modeling bursty transcription and splicing with the chemical master equation. *Biophysical Journal*, 121(6):1056–1069, February 2022. doi: 10.1016/j.bpj.2022.02.004. URL https:

//www.cell.com/biophysj/fulltext/S0006-3495(22)00104-7.
G.G. designed, and implemented the analytical solutions, validated them against simulations, and performed the sequencing data analysis. G.G. participated in the research conceptualization and writing of the manuscript.

[5] Gennady Gorin and Lior Pachter. Distinguishing biophysical stochasticity from technical noise in single-cell RNA sequencing using *Monod*. Preprint, bioRxiv: 2022.06.11.495771, April 2023. URL https://www.biorxiv.org/content/10.1101/2022.06.11.495771v2.
G.G. designed and implemented the theory, algorithms, and sequencing data analyses. G.G. participated in the research conceptualization and writing of the manuscript.

[6] Gennady Gorin and Lior Pachter. Length biases in single-cell RNA sequencing of pre-mRNA. *Biophysical Reports*, 3(1):100097, March 2023. ISSN 26670747. doi: 10.1016/j.bpr.2022.100097. URL https://linkinghub.elsevier.com/retrieve/pii/S2667074722000544.
G.G. performed the model derivations and sequencing data analyses. G.G. participated in the research conceptualization and writing of the manuscript.

[7] Gennady Gorin and Lior Pachter. The telegraph process is not a subordinator. Preprint, bioRxiv: 2023.01.17.524309, January 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.01.17.524309.
G.G. conceptualized, designed, and implemented the critique. G.G. participated in the writing of the manuscript.

[8] Gennady Gorin, Valentine Svensson, and Lior Pachter. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biology*, 21: 39, February 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-1945-3. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1945-3.
G.G. designed and implemented the method and performed the analysis. G.G. participated in the interpretation of the results and the writing of the manuscript.

[9] Gennady Gorin, Maria Carilli, Tara Chari, and Lior Pachter. Spectral neural approximations for models of transcriptional dynamics. Preprint, bioRxiv: 2022.06.16.496448, June 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.06.16.496448.
G.G. conceptualized the approximator. G.G. participated in the study design, derivations, implementation, and writing of the manuscript.

[10] Gennady Gorin, Meichen Fang, Tara Chari, and Lior Pachter. RNA velocity unraveled. *PLOS Computational Biology*, 18(9):e1010492, September 2022. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010492.

G.G. conceptualized and designed the mathematical methodology for the formalization of RNA velocity. G.G. participated in the conceptualization, design, implementation of the critique and benchmarks, and the writing of the manuscript.

[11] Gennady Gorin, John J. Vastola, Meichen Fang, and Lior Pachter. Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments. *Nature Communications*, 13(1):7620, December 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34857-7. URL https://www.nature.com/articles/s41467-022-34857-7.
G.G. designed and implemented the gamma Ornstein–Uhlenbeck simulation. G.G. participated in the research conceptualization, derivations, sequencing data analyses, and writing of the manuscript.

[12] Gennady Gorin, Shawn Yoshida, and Lior Pachter. Transient and delay chemical master equations. Preprint, bioRxiv: 2022.10.17.512599, October 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.10.17.512599.
G.G. participated in the conceptualization of the study, the simulation design, and the writing of the manuscript. G.G. designed and implemented the mathematical framework and the sequencing data analysis.

[13] Gennady Gorin, John J. Vastola, and Lior Pachter. Studying stochastic systems biology of the cell with single-cell genomics data. Preprint, bioRxiv, May 2023. URL https://www.biorxiv.org/content/10.1101/2023.05.17.541250v1.
G.G. performed all simulated and real data analyses. G.G. participated in the conceptualization, design, and derivations reported in the manuscript.

[14] John J. Vastola, Gennady Gorin, Lior Pachter, and William R. Holmes. Analytic solution of chemical master equations involving gene switching. I: Representation theory and diagrammatic approach to exact solution. Preprint, arXiv: 2103.10992, March 2021. URL http://arxiv.org/abs/2103.10992.
G.G. participated in the design and implementation of the numerical methods and benchmarks.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Number *Page*

LIST OF TABLES

Chapter 1

INTRODUCTION AND OUTLINE

Truth is in a well.

Democritus
via Diogenes Laërtius
via Robert Drew Hicks

The past decade has seen enormous investment in the development and application of single-cell RNA sequencing, driven by inexpensive sequencing and advances in microfluidics. This widespread adoption of the technology has been matched by a profusion of analysis methods. Yet, in my view, the experimental advances have far outstripped the theory and interpretation: typical analyses use data science approaches, which are somewhat *ad hoc* and motivated by computational convenience. Although this approach is not an impediment *in principle*, in practice it has led to a crisis of best practices: different analyses produce different results, with no straightforward way to decide on the "best" strategy. I argue that these tensions stem from a reliance on data science at the expense of physical modeling. Whatever the analyses do, they should be coherent with known biophysics; conversely, if they violate physical constraints, they can catastrophically fail. Although this principle of this strategy is deceptively simple, its adoption has been surprisingly limited, in spite of the arsenal of plausible and tractable models previously developed for fluorescence transcriptomics.

The thesis attempts to unify these fields, and develop sequencing analyses that encode physical models. This project requires fairly extensive mathematical machinery, as well as a sound intuition for the physics of gene expression and sequencing. In Chapter 2, I motivate the need for mechanistic models and delineate their scope. In Chapter 3, I introduce fundamental mathematical tools. In Chapter 4, I use these tools to define a set of tractable models that combine biological and technical phenomena. In Chapter 5, I discuss strategies and challenges surrounding the practical implementation of these models. In Chapter 6, I review a common workflow, analyze its weaknesses, and use its pitfalls as a case study to motivate a more principled alternative. In Chapters 7–9, I treat the questions of model identification, inference,

and interpretation using a combination of real and simulated data. In Chapters 10 and 11, I consider the models' compatibility with gene co-expression and further experimental modalities. Finally, in Chapter 12, I summarize promising avenues for further research. Throughout the thesis, I occasionally refer to Appendix A, which contains certain useful derivations too tedious or detailed for the body of the text, and Appendix B, which discusses some of the caveats of modeling sequencing data. Very occasionally, I provide endnotes, which explicate certain qualitative insights that are only obliquely or implicitly referenced in the underlying articles.

Although broad, this thesis is not meant to be exhaustive. It is not and cannot be a review of single-cell RNA sequencing analysis methods. I have attempted to dedicate sufficient space to certain key touchpoints, but the field changes by the week, and a full survey cannot stay relevant. Worse: a review risks meeting methods on their own terms, accepting their premises, and equivocating. I have found it more fruitful to question narrow foundational assumptions; when these assumptions fail, analyses that rely on them become suspect. The thesis does not and cannot review the sprawling fields of biophysics or quantitative cell biology. It does not strive to serve as a first-principles treatment of stochastic transcriptional biophysics; I treat many deep results as a *fait accompli*, and elide the usual theoretical niceties, leaving some pedagogical gaps.

It is not even a comprehensive account of my Ph.D. work. As I have generally attempted to be thorough, and fully treat the minutiae of derivations for my own and readers' benefit, the body of the work summarized here covers, at last count, over six hundred pages across a dozen reports. Although the theoretical investigations have culminated in a common mathematical framework, which admits the individual projects as special cases, the technical details of implementation cannot be so summarized. Therefore, the thesis is not self-contained, except insofar as I unify the idiosyncrasies of my evolving notation and outline the occasionally non-obvious connections between the projects' goals. These projects, in turn, represent only a sampling of scientific questions; many others, as deserving, are omitted or given merely passing mention. But the Ph.D. is finite, and I am satisfied that the questions this thesis raises will eventually be considered and answered, either by my colleagues or by other researchers at the emerging interface of physics and bioinformatics.

Chapter 2

TECHNOLOGIES, DESIDERATA, AND AXIOMS

> And should I then presume?
> And how should I begin?
>
> ---
>
> *The Love Song of J. Alfred Prufrock*
> T.S. ELIOT

To begin, we need to understand what questions the current analyses attempt to treat, and how they answer them using the data at hand. Given a methodology, we can analyze or reverse-engineer its logic, "problematize" the assumptions by explicitly acknowledging them [14, 163], then investigate whether we could improve, validate, or falsify these assumptions to enhance the workflow.

2.1 The two perspectives on the negative binomial distribution

This section adapts portions of [115] by G.G., J.J.V., and L.P., [113] by G.G.*, J.J.V.*, M.F., and L.P., and [112] by G.G., M.F., T.C., and L.P. This perspective on the relationship between sequence census and mechanistic methods was conceptualized by G.G., J.J.V., and L.P.

This high-level course of action is, of course, far too generic, and requires an illustration. To that end, we introduce single-cell RNA sequencing (scRNA-seq) and present a case study to motivate mechanistic modeling. This motivation is one of many, but we find this one to be particularly compelling. Nevertheless, it does rely on some background knowledge of statistics and single-cell RNA sequencing analyses, and the reader from outside the field can skip to Section 2.2 for a qualitative summary if the case study proves too technical.

2.1.1 The negative binomial distribution as an effective data summary

When we perform a single-cell sequencing experiment, we obtain a collection of reads, which represent a selection of the RNA content in living cells [332]. These reads are barcoded; by judiciously using the barcodes and the sequence information, the reads can be converted to a collection of molecule counts, integer numbers x_{cg} for each cell c and gene g [196, 197]. To accomplish the goals of the downstream analysis — the typical systems biology tasks of identifying of cell types, aggregating

them into trajectories, discovering gene modules that consistently differ between cell types or throughout a differentiation trajectory, and visualizing low-dimensional summaries reflecting some component of the data structure — we manipulate the data in a way that reveals the "signal," while eliminating or controlling for the "noise" of the sequencing procedure.

There is a dizzying variety of approaches to this problem. For example, we could build a graph that encodes distances between the observed cell states $\mathbf{x}_c$, then use community detection algorithms to find cliques that coarsely represent cell types, shortest traversal paths that can correspond to trajectories, and neighborhood-preserving embeddings that summarize the graph structure in a low-dimensional visualization. Yet a naïve application of graph algorithms — essentially, a purely non-parametric, "data scientific" approach that attempts to summarize the count matrix — can be misled by the variability and noise in the data: a graph is a discrete, deterministic structure and does not "know" which data points are reliable, or how heterogeneity is to be treated. In addition, this approach restricts statistical interpretability: we can certainly claim that two cliques correspond to distinct cell types, but to justify this claim, we need to construct some measure of statistical confidence. In its simplest Platonic form, this amounts to computing an effect size (how distinct are these cell types?) and a p-value (would we plausibly see such a difference even if the cells were from a single cell type?). We are forced, then, to wrestle with uncomfortable questions like "what, precisely, do we mean by 'cell type?"' and "what is the correct noise model for the p-value computation?"

These uncomfortable questions lead us to adopt the methods of statistics. For example, we can axiomatize a cell type as an internally homogeneous population with a particular average expression, then use the central limit theorem [154] to compare the averages of discovered subpopulations and draw statistical conclusions [187]. This approach — which is ostensibly parametric, but does not make particularly strong assumptions about the RNA count distribution — is sensible, but its application may create further challenges. Single-cell RNA sequencing data are discrete and sparse; even if the central limit theorem holds in the limit, it may perform very poorly for realistic dataset sizes [206]. In addition, merely comparing averages prevents the discovery of biologically interesting cases where the RNA distributions change while keeping the mean constant [205]. Other strategies, such as "binarizing" the data — considering only the presence or absence of molecules, rather than the precise count value [36, 230] — are mathematically distinct, but conceptually

similar, as they also discard the vast majority of the data, potentially sacrificing some signal in the process.

If the central limit theorem and analogous approaches are insufficient, we can move on to parametric statistics, and improve the statistical power at the expense of possibly introducing model misspecification. Many options are available, but the discrete, positive nature of count data are a particularly natural fit for distributions on the natural numbers $\mathbb{N}_0$. For example, we attempt to represent an observation as a draw from a Poisson distribution:

$$P_{\text{Poiss}}(x_{\mathsf{cg}}; \mu_{\mathsf{cg}}) = \frac{1}{x_{\mathsf{cg}}!} \mu_{\mathsf{cg}}^{x_{\mathsf{cg}}} e^{-\mu_{\mathsf{cg}}}. \tag{2.1}$$

The Poisson distribution is straightforward to evaluate as long as we know μ_{cg}. Of course, the problem of identifying this mean parameter is grossly underspecified: if every cell c can have an different mean, and we place no restrictions on its variation, we are unable to learn anything meaningful. We can constrain the problem further, in the most extreme case by proposing that

$$\mu_{\mathsf{cg}} = \mu_{\mathsf{g}}, \tag{2.2}$$

which implies that all of the cells are independent and identically distributed draws from a common distribution. This model is insufficient to summarize real datasets: the Poisson distribution has a variance equal to the mean ($\mu = \sigma^2$), whereas gene count data typically have a variance higher than the mean ($\sigma^2 > \mu$). This "overdispersion" is ubiquitous and does not seem to be explainable by, e.g., the presence of multiple cell subpopulations, because the subpopulations are, in turn, also overdispersed (as in Fig. 1a of [160]).

The next simplest model is the negative binomial:

$$P_{\text{NB}}(x_{\mathsf{cg}}; \nu_{\mathsf{cg}}, \mu_{\mathsf{cg}}) = \frac{\Gamma(\nu_{\mathsf{cg}} + x_{\mathsf{cg}})}{x_{\mathsf{cg}}! \Gamma(\nu_{\mathsf{cg}})} \left(\frac{\nu_{\mathsf{cg}}}{\nu_{\mathsf{cg}} + \mu_{\mathsf{cg}}}\right)^{\nu_{\mathsf{cg}}} \left(\frac{\mu_{\mathsf{cg}}}{\nu_{\mathsf{cg}} + \mu_{\mathsf{cg}}}\right)^{x_{\mathsf{cg}}}. \tag{2.3}$$

This distribution gives us the correct support and distribution shape: the variance is

$$\sigma^2 = \mu + \frac{\mu^2}{\nu} > \mu, \tag{2.4}$$

which is strictly higher than the Poisson distribution with the same mean, and can, in fact, be made arbitrarily high by tuning ν.

Equation 2.3 is still overparametrized, and needs to be constrained to actually summarize data. There does not seem to be an obvious way to do so from first

principles, but analyses of pre-barcode sequencing technologies [9] have proposed that the μ term of Equation 2.4, which coincides with the Poisson variance, should be attributed to purely technical Poisson "shot noise," whereas the residual overdispersion should be attributed to a combination of technical and biological effects, which produce a gamma distribution of molecule concentrations. This approach commonly parametrizes μ as the product of a large, sample-dependent, technical "library size" parameter and a small, compositional, sample and g-dependent "fractional abundance" parameter. The g-dependent "dispersion" parameter ν is typically heavily constrained by μ or set to a constant [9, 238].

Throughout the adoption of single-cell and single-molecule barcoding technologies, this approach has persisted with only minor modifications: there are compositional abundance and "library size" parameters, now unique to a particular cell c rather than the entire sample; there is a "dispersion" parameter, which varies less arbitrarily, if at all; if we assume the compositional gene expression is Gamma-distributed, we can summarize the dataset by using a Poisson model for the technical effects. This set of assumptions produces a tractable negative binomial distribution for the RNA copy number. The specifics of the procedure vary — for example, some studies augment the basic framework with more or less complex noise terms and linking functions, and this summary is nowhere near comprehensive — but despite these numerous variations on the theme, the basic points show up time and again [52, 116, 123, 186, 191, 247, 308].

It is useful to keep in mind that the parametric approach is only one of many. The specter of the negative binomial distribution haunts the non-parametric methods nevertheless. Very few analyses are run on raw data; typically, a workflow normalizes the data with respect to the total per-cell molecule count to account for "library size" variability; afterward, some flavor of log-transformation is applied to abundance matrix to bring the gene expression values, which vary over many orders of magnitude, to a common scale [134]. These transformations are optimal for a stabilizing high-μ, uniform-ν negative binomial distributions [4, 34] under the assumptions outlined above, although they may fail elsewhere [32].

2.1.2 The negative binomial distribution as the consequence of a biophysical model

Although the foregoing description is tremendously oversimplified, it is conceptually in line with the picture presented in reviews [4, 134, 187]. Yet, by presenting it, we

have engaged in some sleight of hand: what, precisely, does the negative binomial model *mean*? What biophysical and chemical phenomena do its gamma and Poisson components represent? What biological assumptions do we make when we suppose that, e.g., ν is constant across all cells, whereas μ can vary? And can we justify these assumptions based on data external to the sequencing experiment?

To answer these questions, we can turn to the field of fluorescence transcriptomics, which uses fluorescent probes that light up when they bind to RNA, allowing us to count individual molecules [102, 233]. Yet we do not observe the gamma distributions implied by the sequencing analyses: fluorescence data are overwhelmingly overdispersed [16, 89, 102, 214, 232, 244], and negative binomial-like distributions effectively fit the observed counts [72, 91, 121, 233]. This immediately implies a problem with our interpretation of Equation 2.4: if the Poisson component μ were a purely technical consequence of the sequencing technology, we should not observe it using fluorescence imaging. Yet we do, which suggests that it is a fundamental component of the *biology*.

Indeed, the fluorescence transcriptomics field typically explains the overdispersion by appealing to the bursting behavior observed in live-cell measurements: transcription is discontinuous and intermittent; the production of RNA is relatively rare; however, when it *does* take place, it produces many molecules at once [65, 92, 161, 170, 210, 293]. The effort to fully characterize these behaviors has led to mechanistic models such as

$$\varnothing \xrightarrow{k} B \times X \xrightarrow{\gamma} \varnothing, \tag{2.5}$$

i.e., at each transcriptional event generates a random number B of molecules X; after some delay, the molecules are degraded. These two reactions happen at rates k and γ (Section A.8.1). Although this model is highly abstracted, it can represent a variety of mechanisms, such as the switching between active and inactive transcriptional states due to activator binding [219, 233]. By setting up and solving a stochastic formulation of Equation 2.5, we obtain precisely the same negative binomial distribution as in Equation 2.3. Of course, transcriptional bursting is only a part of the whole picture, and even non-bursty genes may be overdispersed due to cell-to-cell differences in transcription rates:

$$\varnothing \xrightarrow{k \sim K} X \xrightarrow{\gamma} \varnothing, \tag{2.6}$$

where K is a random variable. If K is gamma, the stationary distribution of X is negative binomial yet again, although its parameters have a different interpretation.

Such models have been used to describe the variability in transcription rates observed across cell sizes [150, 214, 272, 283]. There is no reason why these phenomena should be mutually exclusive, and it appears fair to suppose that both bursting and cell-to-cell variability have some role in the control of expression.

2.2 Motivations for mechanistic models

This section adapts portions of [115] by G.G., J.J.V., and L.P., [113] by G.G.*, J.J.V.*, M.F., and L.P., and [112] by G.G., M.F., T.C., and L.P. This review of motivations was conceptualized by G.G. and L.P.

The essential take-away from Section 2.1 is that sequencing and fluorescence transcriptomics analyses are concerned with the same problem: the summary and interpretation of noisy RNA copy number datasets. To treat this problem, they even use similar tools, such as the negative binomial distribution. However, these superficial similarities hide profound conceptual differences: the *meaning* attributed to these tools is different in the two subfields; single-molecule stochasticity is front and center in fluorescence transcriptomics, but sidelined and treated as purely technical in sequencing transcriptomics.

This observation is somewhat troubling: single-molecule stochasticity is ubiquitous [244], and its omission makes sequencing analyses incoherent with known biology. That said, in spite of these discrepancies, it is not accurate to claim that the scRNA-seq field has entirely neglected the results from fluorescence transcriptomics. Several articles explicitly point to transcriptional variation as a source of biological variability, and either directly use the solutions to mechanistic models [8, 69, 124] or augment them with a model of technical noise [37, 116, 159, 278, 279, 308]. However, this approach is comparatively rare, and has not yet gained traction as part of typical pipelines.

Here, it is reasonable to ask: *why* do the theoretical foundations matter? So far, all we have demonstrated is that both subfields use similar tools, e.g., the negative binomial distribution. Even if their bases and interpretations are subtly different, the end result is much the same. What is the actual impact of adopting one or another worldview?

It turns out that these latent problems come to a head when we attempt to treat broader questions and types of data. For example, typical single-cell analyses use the *mature* transcriptome, i.e., only the counts corresponding to exonic regions. Yet it is also possible to align to intronic regions to obtain two data matrices: the usual mature RNA matrix, as well as a *nascent* RNA matrix, containing all counts associated

with intronic regions; as introns are typically removed during RNA processing, the nascent molecules represent an earlier stage in the RNA life-cycle. The usual descriptive analyses do not have a prescription for simultaneously treating these data types: single-cell analyses omit the nascent RNA; single-nucleus analyses add the two matrices; the "best" approach is controversial, and there does not appear to be a straightforward basis for choosing between the two (Section 8.3).

Yet, in the mechanistic worldview, the solution is almost trivial. There is a causal relationship between the two modalities: nascent RNA are eventually converted to mature RNA. If are confident in the premise that transcription is bursty, we can immediately write down a reasonable model that unifies the two data types:

$$\varnothing \xrightarrow{k} B \times \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M \xrightarrow{\gamma} \varnothing. \tag{2.7}$$

Of course, this model is simplistic — the binary assignment may be overly reductive (Section B.1). Further, we have omitted ambiguities; for example, purely exonic reads may arise from either nascent or mature molecules (Section B.2). Nevertheless, we have successfully encoded the transcriptional biophysics and the causal relationship between the two species, and created a theoretical substrate for representing more sophisticated phenomena, such as technical variability. Indeed, a principled approach to "data integration" is only one of the benefits of adopting the mechanistic worldview, and there is a multi-faceted variety of arguments for its broader adoption.

The biological motivation. By investigating data through the lens of biophysical parameters, we can learn something about the mechanisms that give rise to the data, going beyond data summary to characterize the underlying biological processes. For example, finding that a gene's burst size has changed is more interpretable and actionable than finding that a negative binomial distribution's scale parameter has changed, even if these discoveries are mathematically identical: the former proposes a specific transcriptional mechanism. Just as valuably, this perspective allows us to *falsify* models: if the observed distributions cannot be reproduced by a mathematical model, our conceptualization of the underlying physics is somehow incomplete and must be adjusted.

The physical motivation. The discovery, design, and falsification of biophysical laws deserves special mention: it is part of a broad, interdisciplinary effort to ground the study of biology in physical foundations. Its origins date back to the

mid-twentieth century [24, 25, 145], and recent work in this direction [201, 221] can be bolstered by the integration of genome-wide data.

The statistical motivation, pt. I. As discussed above, to make confident summaries and predictions, accounting for uncertainty is mandatory. Although certain alternatives, such as the central limit theorem and binarization, can help, discrete models produce more statistical power in the sparse, low-copy number limit relevant to scRNA-seq data.

The statistical motivation, pt. II. The statistical advantages of parametric, mechanistic models range beyond loss function book-keeping. By instantiating models and performing a thorough mathematical analysis, we can discover which features are readily identifiable, which are more challenging to infer, and which are entirely impossible to characterize given a particular type of data. For example, the models in Equations 2.5 and 2.6 produce identical distributions at steady state, so attempting to distinguish them purely based on counts of X is futile.

The experimental motivation. We can use the results of statistical investigations to design readouts or control experiments that answer questions of interest. For instance, the aforementioned negative binomial models *can* be distinguished with two-species data (in the vein of Equation 2.7, and as discussed in Section 7.1). In addition, the explicit modeling of technical artifacts can provide a quantitative understanding of the differences between experimental workflows (Chapter 8).

The synthesis motivation. As alluded to elsewhere, if we wish to compare sequencing data to other modalities, such as fluorescence transcriptomics, we need to, on one hand, encode the premise that the underlying biology is identical, and, on the other, attribute any differences to specific technical artifacts (Section 8.2). This is easiest done through biophysical modeling.

The control motivation. Even if we choose not to invest all of our efforts into the analysis of mechanistic models, an understanding of common axioms lets us generate realistic simulated data to benchmark sequencing workflows. In addition, the mathematical framework allows us to systematically investigate implicit limitations and contradictions of common data analysis procedures (Sections 6.1 and 8.4).

The financial motivation. Experiments are expensive; computational data analysis is less so, but still requires non-negligible investment; theory is cheap. It is financially responsible to understand the limitations of experiments and analyses — i.e., which questions can we confidently answer based on a particular dataset? — before collecting any data, instead of discovering these limitations *post hoc*. In addition, a thorough, physically grounded investigation of production pipelines can help identify otherwise obscure technical artifacts and prevent target-oriented industry investigations from pursuing dead ends.

The ethical motivation. The collection of sequencing data is necessarily invasive: it requires the isolation and destruction of living cells. In a scientific context, this entails raising and euthanizing animal test subjects. In a therapeutic context, this entails collecting samples from severely ill or deceased patients. Both of these scenarios involve complicated ethical questions, but it appears most justifiable to strive to minimize invasive procedures by making the most of fewer and smaller datasets.

The synthetic biology motivation. The characterization of transcriptional kinetics has an additional, longer-term perspective: the design of synthetic gene circuits. To design a system, it is essential to understand the physics of its constituent parts; for transcriptional systems, an understanding of single-molecule stochasticity is mandatory.

2.3 Technologies and axioms

We are primarily interested in fitting readily available data from the commercial 10x Genomics platform [332]. We largely focus on the single-cell v3 version of the technology, which offers high-throughput short-read sequencing; however, we occasionally consider the older, lower-throughput v2 technology and the single-nucleus variant of v3 (Sections 8.3 and 9.3). We provide a conceptual overview of the 10x workflow in Sections 4.4.2 and 4.4.3. Nevertheless, we anticipate that the theoretical framework outlined here is applicable to other modalities that can be collected through sequencing, and we outline the prospects in Chapter 11.

We particularly focus on nascent and mature molecule counts. As discussed in Section B.1, this terminology is somewhat non-standard, and intended to emphasize that the modeling approach represents a generic two-stage RNA life-cycle. We adopt the bioinformatic conventions of [168, 197] to identify RNA with intronic content

as "nascent" and RNA without intronic content as "mature." This identification is necessarily imperfect, but helpful, as bioinformatic pipelines for distinguishing and counting these molecular species are readily available [168, 197, 264]. In one case study, we use data from a nanopore-based technology that provides considerably more resolution and a way forward for more detailed models (Section 10.2). However, this technology has not yet seen widespread adoption, so we operate with the most readily available data types at the time. We omit the treatment of ambiguity, largely for computational purposes, but we express our reservations in Section B.2. We speculate that, in many cases, ambiguity can be elided because many nominally ambiguous purely exonic reads lie in the 3′ untranslated region, which suggests they arise from fully processed, poly(A)-capped molecules [131, 217].

It remains to define a set of modeling principles and axioms. Of course, a wide variety of options are available to represent the underlying biology: we can track individual RNA bases; we can treat each molecule as continuous and track its length during production and degradation; we can treat each molecule as an interchangeable discrete entity with no internal structure; we can even take a wider view and consider molecule concentrations instead of counts. Due to the considerable success of discrete stochastic models of biology, as well as the concerning points raised in Section 2.1, we adopt the third axiom: molecules have no internal structure, and are instantaneously produced and degraded. Under certain assumptions, this can be viewed as a simplified representation of a model that *does* represent the internal structure, focusing on a single molecular region (in the vein of 5′ and 3′ probes in [319]). However, the choice is mostly motivated by theoretical and computational facility. We adopt the same framework for the experimental components of the systems we study, taking advantage of the barcodes to identify individual molecules. Again, various other options exist — such as treating reads, or even accounting for the uncertainty in sequencing individual bases — but this level of detail seems somewhat excessive at this preliminary stage.

We almost exclusively consider Markov models, whose future behavior only depends on the current state, rather than any past states. The framework we set up turns out to easily generalize to non-Markov models (Section 4.3.2), and we briefly consider and compare them against some Markov candidates (Section 7.3). However, we find that the considerably simpler Markov models largely suffice, and attempt to avoid introducing additional complexity when it does not appear to be required by the data.

In spite of their popularity (as outlined in Section 2.1), we do not generally consider "cell size" or "library size" models that couple the distributions of different genes. The sole exception is the preliminary investigation in Section 10.3, which proposes one possible basis for such variation. This choice necessarily limits our ability to describe systematic variation in molecular copy numbers, as well as co-variation between genes. However, the current theoretical understanding of these phenomena is somewhat limited, and we hesitate to make any specific modeling assumptions about them. This aspect is also somewhat beside the point. The models we present here are the base case, where we assume these sources of variability can be neglected; if desired, this assumption can be relaxed, and the models can be augmented accordingly by conditioning on some distribution. In other words, if there is some coupling variable Θ, we marginalize over Θ to compute distributions $P(x)$:

$$P(x) = \int_{\Theta} P(x;\Theta)P(\Theta)d\Theta, \tag{2.8}$$

which *still* requires computing $P(x;\Theta)$ at some stage. This is the component of the problem we consider here. We anticipate that a detailed understanding of these phenomena will require considerable further work.

Chapter 3

MATHEMATICAL TOOLS AND PRELIMINARIES

3.1 Common mathematical objects, distributions, and identities

3.1.1 Mathematical objects and key notation

We generally operate with the following hierarchy of variable spaces:

$$\mathbb{N}_0 \subset \mathbb{Z} \subset \mathbb{R} \subset \mathbb{C}, \tag{3.1}$$

where $\mathbb{N}_0$ denotes the non-negative natural numbers $0, 1, 2, ...$; $\mathbb{Z}$ denotes the integers; $\mathbb{R}$ denotes the real numbers; and $\mathbb{C}$ denotes the complex numbers. We occasionally use $\mathbb{R}_{\geq 0}$ to denote the non-negative real numbers and $\mathbb{R}_+$ to denote the positive real numbers. We typically denote variables on $\mathbb{N}_0$ and $\mathbb{Z}$ by x, on $\mathbb{R}$ by y, and on $\mathbb{C}$ by g, u, or h, giving the domains explicitly where necessary. In the context of stochastic processes, the real-valued variables represent the "spatial" degrees of freedom, i.e., the value of the process at a given instant. z is a generic variable. The variable $t \in \mathbb{R}$ denotes the process time.

Vector quantities are typically set in boldface, e.g., $\mathbf{x} = [x_1, \ldots, x_n]^\mathsf{T} \in \mathbb{N}_0^n$. Matrices are typically represented by uppercase letters, e.g.,

$$A = \begin{bmatrix} \alpha_{01} \\ \alpha_{11} \end{bmatrix} \in \mathbb{R}_{\geq 0}^{2\times 1}. \tag{3.2}$$

We use calligraphic fonts for generic mathematical objects; for example, $\mathcal{F}$ is used for neural functions, and $\mathcal{H}$ represents an operator. $\mathcal{L}$ always represents a likelihood, and $\mathcal{D}$ always represents some collection of data. Molecular species are also set in calligraphic fonts. Thus, for example, mature mRNA species are defined as $\mathcal{X}_M$, their microstates are written as x_M, and their actual observed amounts, or the associated random variable, are written as X_M; the underlying mean may be reported as μ_M and the sample mean as $\overline{X}_M$.

In a statistical context, Θ represents generic parameters or parameter vectors. Notation to the effect of $\hat{\Theta}$ represents an estimate of Θ; such estimates are typically, but not always, data-derived (for a counterexample, see Section 5.3). π or $\boldsymbol{\pi}$ represents compositional quantities, such as cell type fractions. κ indexes over cell types or subpopulations.

$\mathbb{I}$ represents the identity function, which returns unity if its argument is true and zero otherwise. δ represents the relevant (discrete or continuous) flavor of degenerate function.

i, j, and k are generic indexing subscripts. k is occasionally used to denote transcriptional burst frequencies, interchangeably with α. $\mathsf{N_c}$ denotes the total number of cells, indexed by c. $\mathsf{N_g}$ denotes the total number of genes, indexed by g. N denotes the total number of approximating terms, indexed by n. $\mathsf{N_k}$ denotes the total number of simulations, indexed by k.

Graph edges and reaction rates are always defined in the source–sink notation, i.e., a rate c_{ij} always represents a species $\mathcal{X}_i$ giving rise to species $\mathcal{X}_j$.

P represents a probability density or mass function used to define a master equation. p represents probability distributions auxiliary to the master equation, e.g., the burst size distribution. f represents generic probability densities.

3.1.2 Stochastic process framework

In the most general case, we study processes that evolve in time over a domain $N \times \mathbb{N}_0^n \times \mathbb{R}_{\geq 0}^m$. The instantaneous state of such a process is given by a collection of variables $(s, \mathbf{x}, \mathbf{y}, t)$. Thus, at time t, s gives the component of the state on a size-N finite lattice, $\mathbf{x}$ on an n-dimensional discrete infinite lattice, and $\mathbf{y}$ on an m-dimensional continuous space.

The evolution of the state over time may or may not be perfectly predictable at a time t given a set of prescribed initial conditions $\{(s^0, \mathbf{x}^0, \mathbf{y}^0, t^0)\}$, and a set of physical laws governing the state transitions. Out of physical realism, we typically impose the condition that all of $\{t^0\} \leq t$. If the physical laws are deterministic, we can study the system's evolution using dynamical systems approaches. However, if they are non-deterministic, we need to invoke the machinery of stochastic processes, and treat the system probabilistically, such that

$$1 = \int_{\mathbf{y}} \sum_{\mathbf{x}} \sum_{s} P\left(s, \mathbf{x}, \mathbf{y}, t; \{(s^0, \mathbf{x}^0, \mathbf{y}^0, t^0)\}\right) d\mathbf{y}, \tag{3.3}$$

where P is a probability distribution that generates realizations through the random variables $\{S, X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$.

A particularly tractable subset of stochastic processes has the *Markov* property, where

$$P\left(s, \mathbf{x}, \mathbf{y}, t; \{(s^0, \mathbf{x}^0, \mathbf{y}^0, t^0)\}\right) = P\left(s, \mathbf{x}, \mathbf{y}, t; s^0, \mathbf{x}^0, \mathbf{y}^0, t^0\right), \tag{3.4}$$

where t^0 is the largest value in the collection $\{t^0\}$, and $s^0, \mathbf{x}^0, \mathbf{y}^0$ is the associated state. If the Markov property holds, we need only specify the initial condition at a single time t^0 to obtain the system's statistical behavior for all $t > t^0$; we use $t^0 = 0$ with no loss of generality. It is occasionally helpful to go one step further and define a probabilistic rather than deterministic initial condition P^0:

$$P\left(s, \mathbf{x}, \mathbf{y}, t; P^0, 0\right) := \int_{\mathbf{y}^0} \sum_{\mathbf{x}^0} \sum_{s^0} P\left(s, \mathbf{x}, \mathbf{y}, t; s^0, \mathbf{x}^0, \mathbf{y}^0, 0\right) P^0(s^0, \mathbf{x}^0, \mathbf{y}^0, 0) d\mathbf{y}^0. \tag{3.5}$$

In the current context, it turns out to be mathematically simpler to use a length-N probability vector, such that

$$(\mathbf{P})_s(\cdot) := P(s, \cdot). \tag{3.6}$$

3.1.3 Generating functions

This section summarizes the mathematical machinery formalized in [115] by G.G., J.J.V., and L.P. G.G. developed this approach as a generalization of the framework constructed by J.J.V. in [113] by G.G.*, J.J.V.*, M.F., and L.P.

The analysis of stochastic processes typically proceeds through *generating functions* [95]. In the general case, the generating function (GF) is a length-N vector $\mathbf{G}$, such that

$$\begin{aligned} G_s(\mathbf{g}, \mathbf{h}) &= \int_0^\infty \cdots \int_0^\infty \sum_{x_1=0}^\infty \cdots \sum_{x_n=0}^\infty \left(\prod_{i=1}^n g_i^{x_i}\right) \left(\prod_{i=1}^m e^{h_i y_i}\right) P(s, \mathbf{x}, \mathbf{y}) dy_m \ldots dy_1 \\ \mathbf{G}(\mathbf{g}, \mathbf{h}) &:= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\mathsf{T}\mathbf{y}} \mathbf{P}\, d\mathbf{y}, \end{aligned} \tag{3.7}$$

where the second line is an abbreviated shorthand for the first. We elide the dependence on time and initial conditions for notational simplicity. The arguments $\mathbf{g} \in \mathbb{C}^n$ and $\mathbf{h} \in \mathbb{C}^m$ are spectral variables. It is frequently easier to treat the shifted coordinate $\mathbf{u} := \mathbf{g} - 1$. Strictly speaking, the mathematical object $\mathbf{G}$ is the combination of a probability-generating function (PGF) in the discrete dimensions and a moment-generating function (MGF) in the continuous dimensions. The generic discrete-only PGF is defined and condensed as follows:

$$G(\mathbf{g}) = \sum_{x_1=0}^\infty \cdots \sum_{x_n=0}^\infty \left(\prod_{i=1}^n g_i^{x_i}\right) P(\mathbf{x}) := \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} P(\mathbf{x}), \tag{3.8}$$

where P is a probability mass function (PMF).

The generic continuous-only MGF is defined and condensed as follows:

$$M(\mathbf{h}) = \int_0^\infty \cdots \int_0^\infty \left(\prod_{i=1}^{m} e^{h_i y_i} \right) P(\mathbf{y}) d\mathbf{y} := \int_{\mathbf{y}} e^{\mathbf{h}^\top \mathbf{y}} P(\mathbf{y}) d\mathbf{y}, \tag{3.9}$$

where P is a probability density function (PDF). It is straightforward to see that the MGF can be obtained by evaluating the PGF at arguments $g_i = e^{h_i}$. The converse does not hold, because the PGF is only defined for random variables on $\mathbb{N}_0$.

When it exists, the generating function allows us to reconstruct properties of the original distribution. First, evaluating a component of the PGF at $g_i = 1$ (or the MGF at $h_i = 0$) marginalizes over dimension i. Second, evaluating the derivatives of the PGF at $g_i = 1$ produces the factorial moments, such that

$$\left. \frac{\partial^k G}{\partial g_i^k} \right|_{g_i=1} = \mathbb{E}[X_i(X_i - 1) \dots (X_i - k + 1)], \tag{3.10}$$

where X_i denotes the random variable with values reported in x_i; similarly, the cross-moments can be obtained by taking mixed derivatives. Analogously, evaluating the derivatives of the MGF at $h_i = 0$ produces the raw moments:

$$\left. \frac{\partial^k M}{\partial h_i^k} \right|_{h_i=0} = \mathbb{E}[Y_i^k]. \tag{3.11}$$

Third, evaluating the derivatives of the PGF at $g_i = 0$ recovers the probability mass function:

$$\frac{1}{x_i!} \left. \frac{\partial^{x_i} G}{\partial g_i^{x_i}} \right|_{g_i=0} = P(x_i), \tag{3.12}$$

where we have assumed that all other dimensions have been marginalized out; joint distributions can be obtained by taking partial derivatives with respect to multiple dimensions.

As generating functions are spectral transforms of the original probability distributions, they inherit many other generic properties of the Fourier transform. These properties are summarized in standard texts [154, 155], and we report them as necessary for derivations.

3.1.4 Special functions

To introduce specific functional forms of stochastic processes and their solutions, it is helpful to be aware of *special functions* commonly encountered in the field. We reproduce their definitions from the standard text by Abramowitz and Stegun [2] without delving into the derivations or functional analysis properties.

The factorial $x!$ over $x \in \mathbb{N}_0$ is defined as follows:

$$x! = \prod_{k=1}^{x} k, \tag{3.13}$$

such that $0! = 1$.

A generalization of the factorial, the gamma function, is defined over $z \in \mathbb{C}$:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt, \tag{3.14}$$

such that $\Gamma(x+1) = x!$ for $x \in \mathbb{N}_0$.

The Pochhammer symbol, or rising factorial, is defined over $z \in \mathbb{C}$ and $n \in \mathbb{N}_0$:

$$(z)_n = \prod_{k=0}^{n-1} (z+k) = \frac{\Gamma(z+n)}{\Gamma(n)}. \tag{3.15}$$

The binomial coefficient is defined over $z, x \in \mathbb{C}$:

$$\binom{z}{x} = \frac{\Gamma(z+1)}{\Gamma(x+1)\Gamma(z-x+1)} = \frac{z!}{x!(z-x)!}, \tag{3.16}$$

where the second identity holds for $z, x \in \mathbb{N}_0$ such that $z \geq x$.

The upper incomplete gamma function is defined over $z, x \in \mathbb{C}$:

$$\Gamma(z, x) = \int_x^\infty t^{z-1} e^{-t} dt, \tag{3.17}$$

such that $\Gamma(z, 0) = \Gamma(z)$. Usefully, at integer arguments,

$$\Gamma(n+1, x) = n! e^{-x} \sum_{k=0}^{\infty} \frac{x^k}{k!}. \tag{3.18}$$

Kummer's confluent hypergeometric function is defined over $a, b, z \in \mathbb{C}$:

$$M(a, b, z) := {}_1F_1(a; b; z) = \sum_{n=0}^{\infty} \frac{(a)_n z^n}{(b)_n n!}. \tag{3.19}$$

When $-a \in \mathbb{N}$, this function can be expressed as a polynomial with a finite number of terms.

The hypergeometric function, or Gauss's hypergeometric function, is defined over $a, b, c, z \in \mathbb{C}$:

$$ {}_2F_1(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!}. \tag{3.20} $$

Usefully, at negative integer arguments, the summation terminates:

$$ {}_2F_1(-m, b; c; z) = \sum_{n=0}^{m} \binom{m}{n} (-1)^n \frac{(b)_n}{(c)_n} \frac{z^n}{n!}. \tag{3.21} $$

The beta function is defined over $z, x \in \mathbb{C}$:

$$ B(z, x) = \int_0^1 t^{z-1} (1-t)^{x-1} dt = \frac{\Gamma(z)\Gamma(x)}{\Gamma(z+x)}. \tag{3.22} $$

It is essential to notice that these functions can be defined by recurrence relations; for example, $\Gamma(n+1) = n\Gamma(n)$. This deceptively simple form suggests a fundamental computational challenge: we would typically like the evaluation to be independent of the particular value of n, requiring methods more sophisticated than applying the definition.

Finally, the principal branch of the Lambert W function, which does not take a combinatorial form, is implicitly defined over $x \in \mathbb{R}_+$:

$$ \text{if } ye^y = x, \text{ then } W(x) = y. \tag{3.23} $$

3.1.5 Continuous distributions

The current section defines the conventions for common continuous probability distributions. Here, we report the probability density functions and other salient properties. The definitions in this and following sections are largely reproduced from the Johnson texts [154, 155].

The normal, or Gaussian, distribution over $y \in \mathbb{R}$ is parametrized by its mean $\mu \in \mathbb{R}$ and standard deviation $\sigma \in \mathbb{R}_+$:

$$ f(y) = \left[2\pi\sigma^2\right]^{-1/2} \exp\left(-\frac{1}{\sigma^2}[y-\mu]^2\right). \tag{3.24} $$

The normal distribution is ubiquitous in statistical inference, where it arises as a consequence of the central limit theorem [154], and the study of continuous-valued

stochastic processes, where it is used to construct Gaussian processes [265]. In addition, a wide variety of data exploration, analysis, and summary techniques, such as principal component analysis, are tailored to normally-distributed variation [259]. The cumulative distribution function of the standard normal distribution with $\mu = 0$ and $\sigma = 1$ is defined to be $\Phi(y)$.

The lognormal distribution over $y \in \mathbb{R}_+$ is parametrized by the mean $\mu_l \in \mathbb{R}$ and standard deviation $\sigma_l \in \mathbb{R}_+$ of the underlying exponentiated normal distribution [51]. With some abuse of terminology, we refer to these parameters as the log-mean and the log-standard deviation.

$$f(y; \mu_l, \sigma_l) = \frac{1}{y\sigma_l\sqrt{2\pi}} \exp\left(-\frac{(\log y - \mu_l)^2}{2\sigma_l^2}\right). \tag{3.25}$$

This distribution has the following mean μ and standard deviation σ:

$$\begin{aligned} \mu &= \exp\left(\mu_l + \frac{1}{2}\sigma_l^2\right) \\ \sigma &= \mu\sqrt{\exp\sigma_l^2 - 1}. \end{aligned} \tag{3.26}$$

Usefully, we can set the parameters to produce a specified set of moments:

$$\begin{aligned} \mu_l &= \log\frac{\mu^2}{\sqrt{\sigma^2 + \mu^2}} \\ \sigma_l &= \sqrt{\log\frac{\sigma^2 + \mu^2}{\mu^2}}. \end{aligned} \tag{3.27}$$

The quantile function of the lognormal distribution have the following form:

$$F^{-1}(p; \mu_l, \sigma_l) = \exp\left(\mu_l + \sigma_l\Phi^{-1}(p)\right). \tag{3.28}$$

The bivariate lognormal distribution over $y_1, y_2 \in \mathbb{R}_+$ is parametrized by the means $\mu_{1l}, \mu_{2l} \in \mathbb{R}$, standard deviations $\sigma_{1l}, \sigma_{2l} \in \mathbb{R}_+$, and correlation $\rho_l \in [-1, 1]$ of the underlying bivariate normal distribution [51]:

$$\begin{aligned} A &:= \frac{\log y_1 - \mu_{1l}}{\sigma_{1l}} \\ B &:= \frac{\log y_2 - \mu_{2l}}{\sigma_{2l}} \\ f(y_1, y_2; \mu_{1l}, \mu_{2l}, \sigma_{1l}, \sigma_{2l}, \rho_l) &= \frac{1}{2\pi\sigma_{1l}\sigma_{2l}y_1y_2\sqrt{1-\rho_l^2}} \exp\left(-\frac{A^2 + B^2 - 2\rho_l AB}{2(1-\rho_l^2)}\right). \end{aligned} \tag{3.29}$$

The marginal distributions are lognormal with the appropriate parameters. Usefully, we can set the ρ_l to produce a desired correlation ρ:

$$\rho_l = \frac{1}{\sigma_{1l}\sigma_{2l}} \log\left(\rho\sigma_1\sigma_2 e^{\mu_{1l}+\mu_{2l}+\frac{1}{2}(\sigma_{1l}+\sigma_{2l})} + 1\right). \tag{3.30}$$

In addition, the conditional distribution over y_2, given a particular y_1, is lognormal with parameters

$$\begin{aligned} \mu_{2l|y_1} &= \mu_{2l} + \rho_l \frac{\sigma_{2l}}{\sigma_{1l}}(\log y_1 - \mu_{2l}) \\ \sigma_{2l|y_1} &= \sigma_{2l}\sqrt{1-\rho_l^2}. \end{aligned} \tag{3.31}$$

The exponential distribution over $y \in \mathbb{R}_+$ is parametrized by its scale $\theta \in \mathbb{R}_+$ or its rate $\eta = \theta^{-1}$:

$$\begin{aligned} f(y;\theta) &= \frac{1}{\theta} e^{-y/\theta} \\ f(y;\eta) &= \eta e^{-\eta y}. \end{aligned} \tag{3.32}$$

The former parametrization is less common, but more convenient for representing its MGF:

$$M(z;\theta) = \frac{1}{1-\theta z}. \tag{3.33}$$

The mean of the exponential distribution is θ. This distribution is ubiquitous in the study of Markovian stochastic processes, because exponentially-distributed waiting times are *memoryless*.

The gamma distribution over $y \in \mathbb{R}_+$ is parametrized by its shape $\nu \in \mathbb{R}_+$ and its scale θ or rate $\eta = \theta^{-1}$:

$$\begin{aligned} f(y;\nu,\theta) &= \frac{1}{\Gamma(\nu)\theta^\nu} y^{\nu-1} e^{-y/\theta} \\ f(y;\nu,\eta) &= \frac{\eta^\nu}{\Gamma(\nu)} y^{\nu-1} e^{-\eta y}. \end{aligned} \tag{3.34}$$

The former parametrization is convenient for representing its MGF:

$$M(z;\nu,\theta) = \left(\frac{1}{1-\theta z}\right)^\nu \tag{3.35}$$

The mean of the gamma distribution is $\nu\theta$. The exponential distribution is a special case of the gamma distribution ($\nu = 1$). The Erlang distribution is another special case ($\nu \in \mathbb{N}$).

The continuous uniform distribution over $y \in [a, b]$ is parametrized by its bounds:

$$f(y; a, b) = \frac{1}{b - a}. \tag{3.36}$$

The inverse Gaussian distribution over $y \in \mathbb{R}_+$ is parametrized by parameters $a, b \in \mathbb{R}_+$ [248]:

$$f(y; a, b) = \frac{a}{\sqrt{2\pi y^3}} e^{ab} \exp\left(-\frac{1}{2}[a^2 y^{-1} + b^2 y]\right). \tag{3.37}$$

The Dirac delta, or continuous degenerate, distribution over $y \in \mathbb{R}$ is defined as follows:

$$\int_{-\infty}^{\infty} f(t)\delta(t)dt = f(0) \tag{3.38}$$

for any function f. Therefore, the delta function's probability density is a point mass at zero. Translating this function and integrating $\delta(t - a)$ returns $f(a)$.

3.1.6 Discrete distributions

Here, we report the probability mass functions of common discrete distributions.

The Poisson distribution over $x \in \mathbb{N}_0$ is parametrized by its mean μ:

$$P_{\text{Poiss}}(x; \mu) = \frac{1}{x!}\mu^x e^{-\mu}. \tag{3.39}$$

Many common discrete distributions arise as Poisson-D mixtures, where D is a mixing distribution that controls the mean. Conceptually,

$$P(x; D) = \int_0^{\infty} \frac{1}{x!}\mu^x e^{-\mu} f_D(\mu)d\mu. \tag{3.40}$$

Usefully, to obtain the PGF of the Poisson mixture at spectral argument g, we can simply evaluate the MGF of the mixing distribution at $g - 1$. Standard texts report further relationships between the underlying and mixed distributions, which we do not reproduce here [157, 215].

The geometric distribution on $x \in \mathbb{N}_0$ is a Poisson-exponential mixture. It is parametrized by the scale $\theta \in \mathbb{R}_+$ of the underlying exponential distribution:

$$P(x; \theta) = \left(\frac{\theta}{1 + \theta}\right)^x \left(\frac{1}{1 + \theta}\right). \tag{3.41}$$

This parametrization is convenient for representing its PGF. The mean of the geometric distribution is θ.

The negative binomial distribution on $x \in \mathbb{N}_0$ is a Poisson-gamma mixture. It can be parametrized by the form of the underlying distribution or by the resulting shape and mean ($\mu = \nu\theta$):

$$\begin{aligned} P_{\mathrm{NB}}(x; \nu, \theta) &= \frac{\Gamma(\nu + x)}{x!\Gamma(\nu)} \left(\frac{1}{1+\theta}\right)^{\nu} \left(\frac{\theta}{1+\theta}\right)^{x} \\ P_{\mathrm{NB}}(x; \nu, \mu) &= \frac{\Gamma(\nu + x)}{x!\Gamma(\nu)} \left(\frac{\nu}{\nu+\mu}\right)^{\nu} \left(\frac{\mu}{\nu+\mu}\right)^{x}. \end{aligned} \tag{3.42}$$

The discrete degenerate distribution supported solely on $x = j$ and zero elsewhere can be represented by a Kronecker delta:

$$P(x; j) = \delta_{xj}. \tag{3.43}$$

3.2 Model selection criteria

The likelihood of parameters Θ under a proposed distribution P and a data distribution $\mathcal{D}$ is simply the total probability of the data:

$$\begin{aligned} \mathcal{L}(\Theta; \mathcal{D}) &= P(\mathcal{D}; \Theta) \\ &= \prod_{\mathsf{c}} P(\mathcal{D}_{\mathsf{c}}; \Theta_{\mathsf{c}}) \\ &= \prod_{\mathsf{c}} P(\mathcal{D}_{\mathsf{c}}; \Theta), \end{aligned} \tag{3.44}$$

where we obtain the second line by assuming observations are independent and the third line by assuming they are independent and identically distributed (i.i.d.) [208]. Much of statistical inference consists of investigating and characterizing the behavior of $\mathcal{L}$ as a function of Θ, and many of the associated challenges stem from $\mathcal{L}$ not being available in closed form.

The likelihood ratio (LR) compares the strength of evidence for various parameters or models Θ_A and Θ_B, which are treated as point estimates [208]:

$$\mathrm{LR} = \frac{\mathcal{L}(\Theta_A; \mathcal{D})}{\mathcal{L}(\Theta_B; \mathcal{D})} = \frac{P(\mathcal{D}; \Theta_A)}{P(\mathcal{D}; \Theta_B)}. \tag{3.45}$$

The Bayes factor (BF) is used to compare models M_A and M_B, and takes into account the uncertainty in their associated parameters Θ_A and Θ_B [38]:

$$\mathrm{BF} = \frac{P(\mathcal{D}; M_A)}{P(\mathcal{D}; M_B)} = \frac{\int_{\Theta_A} P(\mathcal{D}; M_A, \Theta_A) f(\Theta_A; M_A) d\Theta_A}{\int_{\Theta_B} P(\mathcal{D}; M_B, \Theta_B) f(\Theta_B; M_B) d\Theta_B}. \tag{3.46}$$

In this case, $P(\mathcal{D};\cdot)$ is the data likelihood and f is the prior. If the prior or the likelihood is Dirac-like, i.e., the parameters are deterministic, the Bayes factor is equivalent to the likelihood ratio.

The Akaike information criterion (AIC) is a penalized likelihood used to compare point estimates of models k at their optimal parameter estimates $\hat{\Theta}_k$ [38]:

$$\mathrm{AIC}_k = -2\log \mathcal{L}_k(\hat{\Theta}_k) + 2\varsigma_k, \tag{3.47}$$

where ς_k is the number of estimated model parameters. Usefully, the AIC can be used to compute posterior probabilities for a set of models:

$$\begin{aligned} \mathrm{AIC}_{\min} &= \min_k \mathrm{AIC}_k \\ \Delta_k &= \mathrm{AIC}_k - \mathrm{AIC}_{\min} \\ w_\varpi &= \frac{e^{-\frac{1}{2}\Delta_\varpi}}{\sum_k e^{-\frac{1}{2}\Delta_k}}, \end{aligned} \tag{3.48}$$

where w_ϖ is the *Akaike weight* of the model ϖ [38].

3.3 Distance measures

The Kullback-Leibler divergence (KLD) between a discrete data distribution $\mathcal{D}$, i.e., a normalized histogram over microstates $\mathbf{x}$, and a proposed distribution P is defined as follows:

$$D(\mathcal{D} \parallel P) = \sum_{\mathbf{x}} \mathcal{D}(\mathbf{x}) \log \frac{\mathcal{D}(\mathbf{x})}{P(\mathbf{x})}. \tag{3.49}$$

Evidently, only the observed microstates, with $\mathcal{D}(\mathbf{x}) > 0$, contribute to this quantity. The KLD generalizes to continuous distributions, with an integral replacing the summation. If the KLD is high, the distributions are dissimilar. In a statistical context, minimizing the KLD is equivalent to maximizing the likelihood of data under the proposed distribution.

The Jaccard distance d_J is defined as follows:

$$d_J = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}, \tag{3.50}$$

where A and B are sets and $|\cdot|$ represents the set size. If the Jaccard distance is high, the sets have little overlap. Many other distances are discussed in further detail in [73].

Chapter 4

STOCHASTIC MODELS AND SOLUTIONS

> ...as one judge said to the other,
> 'Be just and if you can't be just be arbitrary.'
>
> ---
>
> *Naked Lunch*
> WILLIAM S. BURROUGHS

4.1 Motivations for model classes

This section adapts a portion of [115] by G.G., J.J.V., and L.P. This motivating discussion was written by G.G.

We begin by defining the biological variables of interest. We seek to develop a theoretical framework that can accommodate a wide variety of biological phenomena. This is a modeling challenge that involves a series of trade-offs. On one hand, we would like to represent a broad range of phenomena; on the other, if the scope is *too* broad, the mathematical form becomes intractable. We restrict our analysis to a fairly general class of systems which afford a reasonably compact representation and can be solved by quadrature.

In brief, we care about models with interacting *microscopic*, *mesoscopic*, and *macroscopic* degrees of freedom. These "model scales" are defined with respect to their treatment of stochasticity. Microscopic models account for the flow of probability between discrete states, and are formalized by chemical master equations (CMEs). Mesoscopic models approximate the discrete states by a continuum, and are formalized by equivalent stochastic differential equations (SDEs) or Fokker-Planck equations (FPEs). Macroscopic models omit stochasticity altogether, and are formalized by ordinary differential equations (ODEs). As discussed in [236, 294], certain regimes of microscopic models can be effectively approximated by meso- and macroscopic dynamics. These approximations rely on strong assumptions regarding the "important" sources of stochasticity in the system, and can often be derived through perturbative approximations [297].

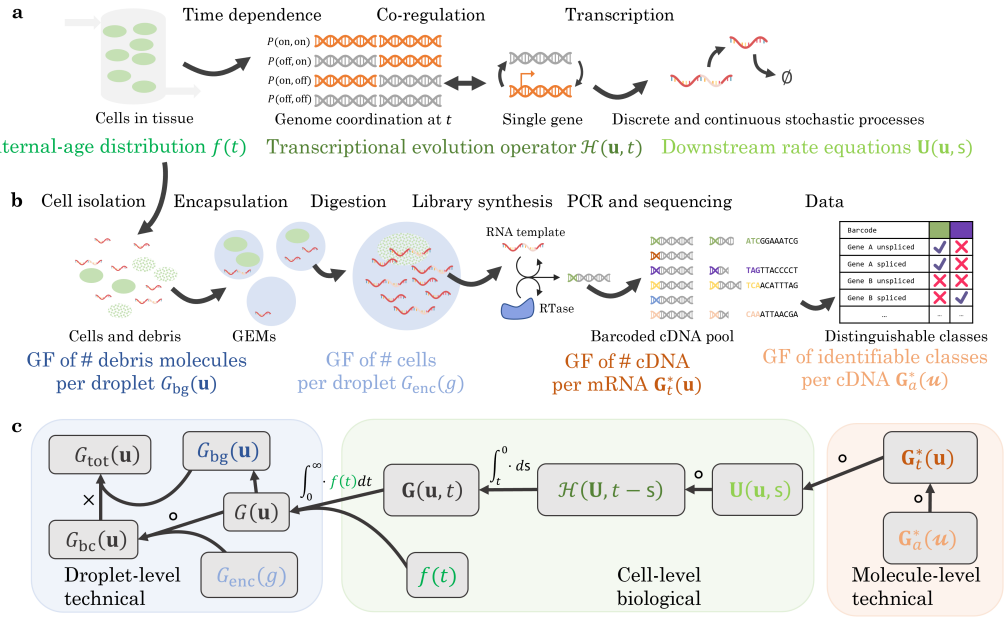

Figure 4.1: The biophysical and chemical phenomena of interest, as well as the relationships between their generating functions.
a. The biological phenomena of interest: cell influx and efflux into a tissue observed by sequencing; the time-dependent transcriptional regulation of one or more genes; downstream continuous and discrete processes.
b. The technical phenomena of interest: the encapsulation of cells and cell debris; cDNA library construction; the loss of information in transcript identification (GF: generating function).
c. The structure of the full generating function of the system in **a** and **b**: to obtain the solution, we variously compose, integrate, and multiply the generating functions of the constituent processes.

4.2 Models of RNA processing and transcriptional noise

This section summarizes the mathematical machinery formalized in [115] by G.G., J.J.V., and L.P. G.G. developed this approach as a generalization of the framework constructed by G.G. and J.J.V. in [113] by G.G.*, J.J.V.*, M.F., and L.P., as well as by J.J.V. in [299], among other publications. The description was written by G.G. and J.J.V.

Our treatment of stochastic systems considers gene state interconversion, as well as the production and processing of macromolecules such as RNA and proteins, which could treated as discrete or continuous variables depending on their concentration. We allow zero- and first-order reactions, including state-dependent bursting, interconversion, degradation, and catalysis. However, we disallow higher-order reactions, including feedback regulation. In addition, we allow various macro- and mesoscopic layers of regulation, such as state- and time-dependent variation in

transcription rates.

By setting up and writing out the relevant master equations, it turns out to be most natural by far to formalize the systems in terms of N categorical degrees of freedom, corresponding to gene states, n discrete ones, corresponding to low-copy number molecular species, and m continuous ones, corresponding to transcription rates or high-concentration species. By omitting regulation, we can split the systems into distinct "upstream" and "downstream" components. As a consequence of the *Poisson representation*, which establishes isomorphisms between discrete and continuous stochastic processes [94], the precise meaning of the "downstream" components ceases to matter: discrete and continuous degrees of freedom can be treated using the same mathematical tools. Formally, the discrete components are Poisson mixtures of the continuous processes (as in Equation 3.40).

This conceptualization happens to be particularly useful under a particular combination of assumptions (mass action kinetics, no regulation) and goals (computing dataset likelihoods). However, others alternatives are available. For example, the discrete degrees of freedom have been studied in the language of queuing theory [166, 257]. The analysis of continuous stochastic processes owes a great deal to mathematical finance [265]. The distinctions between model scales may even be translated into the language of quantum physics [5, 211, 299]: categorical states are mutually exclusive and follow fermion-like statistics; discrete states are unconstrained and follow boson-like statistics; continuous states are fundamentally classical. This conceptualization is in its nascence, but may well lead to useful and widespread mathematical tools in the future. Under the assumptions and goals we adopt, we have found that the Poisson representation approach we adopt, which exploits the properties of partial differential equations, provides the best balance of computational and analytical tractability in the multi-modal context.

4.2.1 Master equation definitions

The categorical variable, denoted by $s \in \{1, ..., N\}$, represents the instantaneous state of a multi-state gene. By assuming that the state interconversions are Markovian and independent of all other components of the system, we can define H_{ij}, the rates of transitioning from state i to state j:

$$\mathcal{S}_i \xrightarrow{H_{ij}} \mathcal{S}_j. \tag{4.1}$$

These rates can be summarized in the state transition matrix $H \in \mathbb{R}_{\geq 0}^{N \times N}$, such that $H_{ii} = -\sum_{j \neq i} H_{ij}$ and $\sum_j H_{ij} = 0$ to enforce the conservation of probability. This

set of transitions can be represented by a finite master equation, which tracks the probabilities of each state s at a time t:

$$\begin{aligned} \frac{\partial P(s,t)}{\partial t} &= \sum_{i=1}^{N} H_{is} P(i,t), \text{ or more compactly} \\ \frac{\partial \mathbf{P}(t)}{\partial t} &= H^{\mathsf{T}} \mathbf{P}. \end{aligned} \tag{4.2}$$

As this system is expressed in terms of a differential equation for an arbitrary time t, the relation holds for time-dependent H. For simplicity, we assume that H is deterministic and independent of other variables. For a review of CMEs, we recommend [95, 295, 297, 315, 336].

The nonnegative discrete variables, denoted by $\mathbf{x} \in \mathbb{N}_0^n$, represent molecular copy numbers. We assume that n molecular species participate in four classes of transitions, and can summarize their effect by considering their reaction schema and effect on x_i, the number of molecules of species i:

$$\begin{aligned} \mathcal{X}_i &\xrightarrow{c_{ij}} \mathcal{X}_j \\ \mathcal{X}_i &\xrightarrow{c_{i0}} \varnothing \\ \mathcal{X}_i &\xrightarrow{q_{ij}} \mathcal{X}_i + \mathcal{X}_j \\ \varnothing &\xrightarrow{\alpha_\omega^d} B_{i_1}\mathcal{X}_{i_1} + \cdots + B_{i_{\ell_\omega}}\mathcal{X}_{i_{\ell_\omega}}. \end{aligned} \tag{4.3}$$

First, species i can be converted to species j with rate $c_{ij}x_i$. Second, species i can spontaneously degrade with rate $c_{i0}x_i$. These classes of monomolecular transitions, which either maintain or reduce the total number of molecules in the system, can be summarized in the matrix $C^{dd} \in \mathbb{R}^{n\times n}$, such that $C^{dd}_{ij} = c_{ij}$ and $C^{dd}_{ii} = -c_{i0} - \sum_{j\neq i} c_{ij}$; $(C^{dd})^{\mathsf{T}}$ is the matrix governing the associated reaction rate equations [146]. Third, species i participate in autocatalysis at the rate q_{ii}, or catalysis of species j at the rate q_{ij}. These reactions can be summarized by the matrix $Q^d \in \mathbb{R}^{n\times n}_{\geq 0}$, such that $Q^d_{ij} = q_{ij}$.

Finally, molecules can be produced through a variety of reaction channels, indexed by ω. In the general case, a transcriptional event — a *burst* of production — simultaneously creates molecules of ℓ_ω discrete species $\{i_1, ..., i_{\ell_\omega}\}$. We assume bursts are described by a Poisson arrival process, with burst frequency α^d_ω and the nontrivial ℓ_ω-variate joint distribution $p^d_\omega(\mathbf{z})$ of non-negative burst sizes $\{B_{i_1}, \cdots, B_{\ell_\omega}\}$. In other words, $p^d_\omega(\mathbf{z})$ is well-defined over all non-negative $\mathbf{z}$, but its value is identically zero whenever any component $z_i > 0$, for all $i \notin \{i_1, ..., i_{\ell_\omega}\}$. The burst frequency and distribution may vary with gene state s.

This formulation includes the trivial case of Poisson point process production of species i, for which $\ell_\omega = 1$ and $p^d_\omega(\mathbf{z}) = \delta_{ij}$, the degenerate distribution located at unity for species i and zero for all other species.

This mass action model, which tracks molecule counts, can be represented by an equivalent discrete chemical master equation, which tracks the probability of each microstate $\mathbf{x}$:

$$\begin{aligned}\frac{\partial P(\mathbf{x},t)}{\partial t} &= \sum_{i=1}^{n} c_{i0}\left[(x_i+1)P(x_i+1,t) - x_i P(\mathbf{x},t)\right] \\ &+ \sum_{i,j=1}^{n} c_{ij}\left[(x_i+1)P(x_i+1,x_j-1,t) - x_i P(\mathbf{x},t)\right] \\ &+ \sum_{i=1}^{n} q^d_{ii}\left[(x_i-1)P(x_i-1,t) - x_i P(\mathbf{x},t)\right] \\ &+ \sum_{i,j=1}^{n} q^d_{ij}\left[x_i P(x_j-1,t) - x_i P(\mathbf{x},t)\right] \\ &+ \sum_{\omega} \alpha^d_\omega \left[\sum_{\mathbf{z}} p^d_\omega(\mathbf{z}) P(\mathbf{x}-\mathbf{z},t) - P(\mathbf{x},t)\right].\end{aligned} \tag{4.4}$$

For simplicity of notation, species that do not occur in a reaction are elided from the master equation probability terms.

As in Equation 4.2, this master equation holds even if the rates are time-dependent. For tractability, we assume only α^d_ω and p^d_ω can vary over time. Since the form of these functions is arbitrary deterministic, the dynamics of these variables represent unspecified macroscopic processes.

The nonnegative continuous variables, denoted by $\mathbf{y} \in \mathbb{R}^m_{\geq 0}$, represent mesoscopic concentrations or coarsely-modeled noise sources. We assume that these variables are governed by Ornstein–Uhlenbeck-type stochastic differential equations:

$$d\mathbf{y}_t = (C^{cc})^{\mathsf{T}}\mathbf{y}_t dt + Q^c(\mathbf{y}_t)d\mathbf{W}_t + \sum_{\omega} d\mathbf{L}_\omega(t), \tag{4.5}$$

where $\mathbf{y}_t$ is a realization of the process, $\mathbf{W}_t$ is an w-dimensional Brownian motion, and $\mathbf{L}_\omega$ is a subordinator. For a review of SDEs, we recommend [23, 57, 95, 236, 295, 297, 336].

The matrix $C^{cc} \in \mathbb{R}^{m\times m}$ sets the mean-reversion terms. In other words, a nonzero entry C^{cc}_{ij} implies that the level of species i is proportional to influx into species j. The operator $Q^c(\mathbf{y}_t) : \mathbb{R}^m_{\geq 0} \to \mathbb{R}^{m\times \mathsf{w}}_{\geq 0}$ sets the level of noise. For simplicity, we

assume the noise term takes the form of an uncoupled square-root diffusion, such that $\mathsf{w} = m$ and $Q^c(\mathbf{y}_t) = \mathrm{diag}(\boldsymbol{\sigma} \odot \sqrt{\mathbf{y}_t})$. The symbol $\odot$ denotes the elementwise, or Hadamard, product of two vectors, the square root should be interpreted as elementwise, and all elements of the constant volatility vector $\boldsymbol{\sigma}$ are non-negative. Although this choice of Q^c is somewhat restrictive, it produces a particularly simple diffusion tensor Σ:

$$\Sigma(\mathbf{y}) := \frac{1}{2} Q^c(\mathbf{y}) Q^c(\mathbf{y})^\mathsf{T} = \frac{1}{2} \mathrm{diag}(\boldsymbol{\sigma}^2 \odot \mathbf{y}), \tag{4.6}$$

where the square $\boldsymbol{\sigma}^2$ should be interpreted as elementwise.

We assume that each $\mathbf{L}_\omega$ only includes drift or compound Poisson terms. The drift terms have the form $\alpha_i^c \delta_{ij} t$. To slightly lighten the notation, we can aggregate all drift terms under $\omega = 1, \cdots, m$ as $\{\alpha_1^c dt, \cdots, \alpha_m^c dt\}$; some of these entries may be zero. The compound Poisson terms have the form $\sum_{k=0}^{\mathsf{N}_\omega(t)} (\mathbf{B}_\omega)_k$ [43], such that $\mathsf{N}_\omega(t)$ is a Poisson random variable with mean $\alpha_\omega^c t$ and $\{(\mathbf{B}_\omega)_k\}$ is a set of independent and identically distributed realizations of the random variable $\mathbf{B}_\omega$. This random variable has a nontrivial ℓ_ω-variate joint density $p_\omega^c(\mathbf{z})$ on $\mathbb{R}^m_{\geq 0}$, with the remaining $m - \ell_\omega$ dimensions concentrated at zero. We note that this formulation entails a slight abuse of notation, as ω is used to index over discrete burst processes as well as continuous drift and jump components.

This formulation can be reframed as a Fokker-Planck equation [236], which tracks the probability density of each microstate $\mathbf{y}$:

$$\begin{aligned} \frac{\partial P(\mathbf{y}, t)}{\partial t} = &- \sum_{i,j=1}^{m} C_{ij}^{cc} \frac{\partial}{\partial y_j} \left[y_i P(\mathbf{y}, t) \right] + \frac{1}{2} \sum_{i=1}^{m} \sigma_i^2 \frac{\partial^2}{\partial y_i^2} \left[y_i P(\mathbf{y}, t) \right] \\ &- \sum_{i=1}^{m} \alpha_i^c \frac{\partial P(\mathbf{y}, t)}{\partial y_i} + \sum_{\omega > m} \alpha_\omega^c \left[\int_{\mathbf{z}} p_\omega^c(\mathbf{z}) P(\mathbf{y} - \mathbf{z}, t) d\mathbf{z} - P(\mathbf{y}, t) \right]. \end{aligned} \tag{4.7}$$

As above, we assume that only the components of $\mathbf{L}_\omega$ can vary in time.

In addition to these discrete- and continuous-only terms, we need to account for these components' interactions. For example, we may want to represent the production of a discrete species controlled by a continuous variable, e.g., a time-varying transcription rate:

$$\varnothing \xrightarrow{y_i C_{ij}^{cd}} \mathcal{X}_j. \tag{4.8}$$

This reaction has the rate $y_i C_{ij}^{cd}$. This class of reactions can be summarized in the matrix $C^{cd} \in \mathbb{R}^{m \times n}_{\geq 0}$. In other words, this class of reactions contributes the following

terms to the overall master equation:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}^{cd}\left[y_i P(x_j - 1, \mathbf{y}, t) - y_i P(\mathbf{x}, \mathbf{y}, t)\right]. \tag{4.9}$$

Finally, we may want to represent the production of a continuous species from a discrete one, e.g., the rapid translation of high-abundance protein from low-abundance RNA [31]. This class of reactions simply adds a term proportional to $(C^{dc})^{\mathsf{T}}\mathbf{x}\ dt$ to the expression for $d\mathbf{y}_t$. The matrix $C^{dc} \in \mathbb{R}_{\geq 0}^{n\times m}$ contains the relevant rates, such that C_{ij}^{dc} is the rate of producing the continuous species j from discrete species i. Therefore, we append a set of drift-like terms to the Fokker-Planck equation:

$$-\sum_{i=1}^{n}\sum_{j=1}^{m} C_{ij}^{dc} x_i \frac{\partial P(\mathbf{x}, \mathbf{y}, t)}{\partial y_j}. \tag{4.10}$$

To construct the full master equation, we need to define a system of N coupled equations. To do so, we essentially add Equations 4.2, 4.4, 4.7, 4.9, and 4.10, replacing all instances of P with $\mathbf{P}(s, \mathbf{x}, \mathbf{y}, t)$. However, to account for differences in transcription between gene states, we allow the ω-associated terms to vary with s. The full master equation is reported in Equation A.1.

4.2.2 Approaches to solution

We have defined a class of master equations. To evaluate likelihoods and statistically characterize data, we need to calculate the probabilities of observations under a particular model and set of parameters. There are essentially four approaches to this problem.

4.2.2.1 Simulation

In principle, we can approximate solutions by simulation, as discussed in Section 5.5. If $m = 0$, we can use the usual form of Gillespie's stochastic simulation algorithm [98] (as in Section 5.5.1); if $m > 0$, we can use somewhat more sophisticated schema (Section 5.5.2). If we perform $\mathsf{N_k}$ simulations of a fully discrete system, indexed by k, the probability $P(s, \mathbf{x})$ can be approximated by

$$\frac{1}{\mathsf{N_k}}\sum_{\mathsf{k}=1}^{\mathsf{N_k}} \mathbb{I}\left((S, X_1, \ldots, X_n)_{\mathsf{k}} = (s, x_1, \ldots, x_n)\right), \tag{4.11}$$

where $\mathbb{I}$ is the indicator function and $(S, X_1, \ldots, X_n)_\mathsf{k}$ are the process values at time t in the kth realization. Although this method can be used when no other alternatives exist [110], it converges with the usual Monte Carlo rate of $\mathsf{N_k}^{-1/2}$ [193], which is generally unacceptably slow, and impractical even for modest n. In addition, the Gillespie approach is fundamentally "coupled": if we are interested in a subset of downstream distributions, we need to simulate the entire system, including the upstream reactions.

4.2.2.2 Matrix algorithms

Instead of considering trajectories, we can exploit the fact that the discrete terms of the master equation (Equation A.1) can be represented by matrix multiplication:

$$\begin{aligned} \frac{d\mathbf{P}}{dt} &= A\mathbf{P} \\ \mathbf{P} &= e^{At}\mathbf{P}^0, \end{aligned} \tag{4.12}$$

where A is infinite-dimensional and $\mathbf{P}$ now contains entries for all s and $\mathbf{x}$. By truncating A to a finite-dimensional matrix $\tilde{A}$, we can obtain a reasonably accurate approximation to $\mathbf{P}$:

$$\tilde{\mathbf{P}} = e^{\tilde{A}t}\tilde{\mathbf{P}}^0. \tag{4.13}$$

This is the finite state projection (FSP) algorithm [203]. If only the stationary distribution is of interest, the determination of $\tilde{\mathbf{P}}$ is equivalent to the determination of the nullspace of $\tilde{A}$ [118].

This approach is generic, and works for any combination of reactions. However, the matrix $\tilde{A}$ tends to be fairly large, and making the procedure tractable often involves a considerable degree of computational design [304]; in addition, the matrix operations have cubic time complexity, which is somewhat restrictive. More fundamentally, the FSP approach retains the "coupling" feature of the stochastic simulation algorithm: even if we only care about a certain marginal, we may need to explicitly represent all species and reactions. Finally, the probability distributions are somewhat challenging to integrate with other stochastic phenomena: for example, FSP cannot directly represent technical noise that occurs in the sequencing process, as in Section 4.4, and requires dedicated manipulation of $\tilde{\mathbf{P}}$.

4.2.2.3 Exact analysis

In some very narrow cases, the CME can be exactly solved by sheer ingenuity, e.g., by using an *ansatz* for the probability mass function [299]. For example, if we are

interested in the steady state of the simple birth–death process

$$\varnothing \xrightarrow{\alpha} X \xrightarrow{\gamma} \varnothing, \tag{4.14}$$

we can write down its master equation

$$\frac{dP(x,t)}{dt} = \alpha[P(x-1,t) - P(x,t)] + \gamma[(x+1)P(x+1,t) - xP(x,t)], \tag{4.15}$$

and notice that the steady-state probability flux equation

$$0 = \alpha[P(x-1,t) - P(x,t)] + \gamma[(x+1)P(x+1,t) - P(x,t)] \tag{4.16}$$

can be satisfied by substituting P with the Poisson distribution. For more complicated processes, we can obtain a partial differential equation equivalent to the master equation (Appendix A) and exactly solve it, either by using an *ansatz* [31], a perturbative expansion [301], or somewhat brute-force calculation and judicious use of special functions [125, 144, 299]. This approach is, however, typically only practical for some combination of $N = 2$, $n = 1$, or $t \to \infty$. It is also challenging to apply systematically: for example, even if $n = 1$ is tractable, $n > 1$ is typically not.

4.2.3 Semi-analytical spectral solution

We would like a more generic strategy. It turns out that the most straightforward way to evaluate the CME is to *almost* solve it, obtain a numerically tractable ODE, then use standard numerical packages to solve this ODE [299].

The master equation is fairly cumbersome and challenging to analyze directly. Therefore, analysis has to proceed by spectral methods (Section 3.1.3), which recast the probabilistic master equation into a deterministic partial differential equation (PDE) with respect to categorical variable s, discrete spectral variables $\mathbf{g}$, and continuous spectral variables $\mathbf{h}$. By computing the generating function of both sides of Equation A.1 (Appendix A), we find that the master equation is equivalent to a much more compact PDE system:

$$\frac{\partial \mathbf{G}}{\partial t} = H^{\mathsf{T}}\mathbf{G} + \mathbf{G} \odot \mathcal{A}(\mathbf{u}) + J\left[C\mathbf{u} + \operatorname{diag} \mathbf{u}\, D\mathbf{u}\right]. \tag{4.17}$$

This formulation relies on defining the unified variables encoded in a vector $\mathbf{u}$:

$$\mathbf{u} := \begin{bmatrix} \mathbf{g} - 1 \\ \mathbf{h} \end{bmatrix} \text{ and Jacobian } J_{si} = \frac{\partial G_s}{\partial u_i}, \tag{4.18}$$

as well as unified matrices:

$$C := \begin{bmatrix} C^{dd} + Q^d & C^{dc} \\ C^{cd} & C^{cc} \end{bmatrix} \text{ and } D := \begin{bmatrix} Q^d & C^{dc} \\ 0 & \frac{1}{2}\text{diag } \boldsymbol{\sigma}^2 \end{bmatrix}. \tag{4.19}$$

By way of analogy, we sometimes use Q^c to indicate the diffusion tensor $\frac{1}{2}\text{diag } \boldsymbol{\sigma}^2$. Each entry of the length-N vector function $\mathcal{A}$ consists of the burst and drift terms:

$$\begin{aligned} (\mathcal{A})_s &= (\boldsymbol{\alpha}^d)_s^\mathsf{T}(\mathbf{F}_s(\mathbf{u}+1) - 1) + (\boldsymbol{\alpha}^c)_s^\mathsf{T}(\mathbf{M}_s(\mathbf{u}) - 1) \\ &:= \boldsymbol{\alpha}_s^\mathsf{T}(\mathcal{M}_s(\mathbf{u}) - 1). \end{aligned} \tag{4.20}$$

The vector $\boldsymbol{\alpha}_s^d$ contains the frequencies of all discrete burst processes for state s. The first m entries of $\boldsymbol{\alpha}_s^c$ contain the continuous species' drifts in state s. The remaining entries contain the corresponding rates of continuous burst processes. $\boldsymbol{\alpha}_s$ aggregates these quantities. The vector function $\mathbf{F}_s$ contains the joint PGFs of the discrete burst processes, and only depends on the first n variables. The vector function $\mathbf{M}_s$ contains the drift terms, as well as the joint MGFs of the continuous burst processes, and only depends on the last m variables. The parameters of the $\mathcal{M}_s$ operator may vary in time.

To obtain the generating function at t, we apply the method of characteristics. First, we calculate the characteristics parametrized by the scalar variable s:

$$\begin{aligned} T(\mathsf{s}) &= t - \mathsf{s} \\ \frac{d\mathbf{U}(\mathsf{s})}{d\mathsf{s}} &= C\mathbf{U}(\mathsf{s}) + \text{diag } \mathbf{U}(\mathsf{s})\ D\mathbf{U}(\mathsf{s}) \text{ such that } \mathbf{U}(\mathsf{s}=0) = \mathbf{u}. \end{aligned} \tag{4.21}$$

This is the "downstream" ODE, which governs abundances in isolation from production and regulation.

Therefore, $\mathbf{G}$ is governed by the following system of ordinary differential equations:

$$\frac{d\mathbf{G}(\mathbf{U}(\mathsf{s}), T(\mathsf{s}))}{d\mathsf{s}} = -H(T(\mathsf{s}))^\mathsf{T}\mathbf{G} - \mathbf{G} \odot \mathcal{A}(\mathbf{U}(\mathsf{s}), T(\mathsf{s})) := \mathcal{H}(\mathbf{U}, T)\ \mathbf{G}. \tag{4.22}$$

To obtain $\mathbf{G}$ at t, we integrate this system from $\mathsf{s} = t$ to $\mathsf{s} = 0$. We use $\mathbf{G}^0(\mathbf{U}(t))$ as the initial condition, where $\mathbf{G}^0$ is the generating function of the initial distribution. This is the "upstream" ODE, which governs the full generating function.

In the general case, evaluating this system requires two applications of quadrature: first, solving the $n + m$-dimensional downstream system to obtain the values of characteristics $\mathbf{U}$ at a set of grid points over $[0, t]$, and then solving the N-dimensional upstream system to obtain the value of the generating function.

4.2.3.1 Implications

The unified treatment of continuous and discrete variables warrants dedicated mention. As discussed above, it represents an application of the Poisson representation; we can readily interconvert between equivalent continuous and discrete processes. Although the resulting problems are equally challenging, we can occasionally use standard results from the study of continuous processes in finance to solve seemingly unrelated biological problems without performing any new calculations. We use three case studies to illustrate the capabilities of this approach in Section A.8.3.

Some special cases afford simpler solutions. If $D \neq 0$, the downstream ODE takes a Riccati-like form and generally resists exact analysis [200]. However, if $D = 0$, the system takes the tractable linear form

$$\begin{aligned} \frac{d\mathbf{U}(\mathsf{s})}{d\mathsf{s}} &= C\mathbf{U}(\mathsf{s}) := V^{-1}\Lambda V\mathbf{U}(\mathsf{s}), \text{ with the solution} \\ \mathbf{U}(\mathsf{s}) &= e^{C\mathsf{s}}\mathbf{u} = Ve^{\Lambda\mathsf{s}}V^{-1}\mathbf{u}, \end{aligned} \tag{4.23}$$

where the columns of V contain the eigenvectors of C. This identity holds only when all eigenvalues of C are distinct. When they are not, $\mathbf{U}$ can be obtained analogously using generalized eigenvectors, which are a combination of polynomial and exponential functions [280]. Practically, this case only requires one application of quadrature.

If, in addition, $N = 1$, the upstream ODE reduces to a single integral:

$$\phi(t) = \int_t^0 \frac{d\phi(\mathbf{U}(\mathsf{s}), T(\mathsf{s}))}{d\mathsf{s}} d\mathsf{s} = \phi^0(\mathbf{U}(t)) + \int_0^t \mathcal{A}(\mathbf{U}(\mathsf{s}), T(\mathsf{s}))d\mathsf{s}, \tag{4.24}$$

where $\phi := \log G$, $\phi^0 = \log G^0$, and the generating function G is no longer boldfaced because only a single gene state exists.

Finally, if $\mathcal{A}$ is a linear operator $a_1u_1 + \cdots + a_{n+m}u_{n+m}$, the system is in the drift-only regime; no bursting occurs. In this case, the system reduces to

$$\phi(t) = \phi^0(\mathbf{U}(t)) + \sum_{i=1}^{n+m} \int_0^t a_i(t - \mathsf{s})U_i(\mathsf{s})d\mathsf{s}, \tag{4.25}$$

where U_i are the components of $\mathbf{U}$. As each U_i is, in turn, a weighted sum of u_i, the second term of the log-generating function is given by a sum of fairly simple convolutions that scale as $\int_0^t a_i(t - \mathsf{s})e^{-\lambda_j\mathsf{s}}d\mathsf{s}$. This system corresponds to the *constitutive* transcription process.

Finally, in the simplest case, if all eigenvalues of C are negative, the transient part of Equation 4.25 vanishes as $t \to \infty$ and the stationary log-generating function is a linear combination of u_i. This implies that the discrete distributions of the constitutive transcription process converge to multivariate independent Poisson [146].

4.3 Challenges of broader model classes

4.3.1 Regulation

This section summarizes some investigations undertaken during the writing of [115] by G.G., J.J.V., and L.P. This derivation was performed by G.G.

Thus far, we have omitted regulation. We can begin with fairly simple schema of the following form:

$$\mathcal{S}_i \xrightarrow{x_k R_{k,ij}} \mathcal{S}_j, \tag{4.26}$$

i.e., state transitions catalyzed by species $\mathcal{X}_k$. R_k is a stochastic regulation matrix analogous to H. This class of reactions leads to the following partial differential equation system, quite similar to Equation 4.17:

$$\frac{\partial \mathbf{G}}{\partial t} = H^{\mathsf{T}}\mathbf{G} + \mathbf{G} \odot \mathcal{A}(\mathbf{u}) + J\left[C\mathbf{u} + \operatorname{diag} \mathbf{u}\, D\mathbf{u}\right] + \sum_{k=1}^{n} (u_k + 1) R_k^{\mathsf{T}} \frac{\partial \mathbf{G}}{\partial u_k}. \tag{4.27}$$

As discussed in Section A.7, the coupling of "upstream" and "downstream" degrees of freedom through the regulation matrices R_k renders this problem intractable. Although this class of systems has been studied previously [141, 143, 301], and considerable ingenuity has been applied to obtain exact solutions, it is as of yet unclear whether generic strategies for solving regulation problems exist.

4.3.2 Non-Markov processes

This section is based on unpublished revisions to [114] by G.G., S.Y., and L.P. This theoretical discussion was derived and written by G.G.

In the discrete context, the vector function $\mathbf{U}$ is not arbitrary: it "correctly" propagates the initial molecule distribution into the future. In other words, if the initial condition of Equation 4.22 is degenerate, with $\mathbf{G}^0(\mathbf{u}) = \delta_{ij}(u_j + 1)$, there exists a single molecule of species $\mathcal{X}_i$ at $t = 0$. If, in addition, no production occurs and $\mathcal{H} = 0$, the generating function is trivial, and yields

$$G(\mathbf{u}, t) = U_i(\mathbf{u}, t) + 1, \tag{4.28}$$

i.e., $\mathbf{U}$ is simply the shifted generating function of the system distribution, conditional on having a single molecule at $t = 0$. In the special case of $D = 0$, the entries of $\mathbf{U}$

are generalized survival functions:

$$\begin{aligned} G(\mathbf{u}, t) &= \sum_j g_j P(x_j = 1, t | x_i = 1, 0) \\ &= \sum_j u_j P(x_j = 1, t | x_i = 1, 0) + 1 \\ &= U_i(\mathbf{u}, t) + 1. \end{aligned} \tag{4.29}$$

In other words, each U_i is a weighted sum of u_j; the weights are precisely the time-dependent conditional distributions $P(x_j = 1, t | x_i = 1, 0)$.

It turns out that we the formulation is modular: we can use *any* $\mathbf{U}$ that represents such a conditional distribution to encode non-Markovian downstream dynamics. To do so, we define an integral operator C with the following non-Markovian cases:

$$\begin{aligned} &\mathbf{U}(\mathbf{u}, \mathsf{s}) = C(\mathbf{U}), \text{ such that} \\ &C(\mathbf{U})_i = u_i F_i'(\mathsf{s}) \text{ for degradation and} \\ &\qquad = u_i F_i'(\mathsf{s}) + \int_0^{\mathsf{s}} U_j(\mathbf{u}, t^*) f_i(\mathsf{s} - t^*) dt^* \text{ for conversion to } \mathcal{X}_j. \end{aligned} \tag{4.30}$$

In this notation, F_i' is the survival function of $\mathcal{X}_i$ and f_i is the waiting time probability density function [154]. The variable t^* indicates the time at which the reaction fires. In the Markovian case, the degradation characteristic is identical, but multiple conversion routes may compete[1], yielding

$$C(\mathbf{U})_i = u_i F_i'(\mathsf{s}) + \sum_j \int_0^{\mathsf{s}} U_j(\mathbf{u}, t^*) f_{i,j}(\mathsf{s} - t^*) dt^*, \tag{4.31}$$

where j indexes over the products of isomerization. When the reaction network comprises a directed acyclic graph, Equation 4.30 can be applied to compute characteristics directly[2].

Usefully, when the species $\mathcal{X}_i$ remains in the system for a deterministic duration τ before being converted to $\mathcal{X}_j$, we find that its characteristic is given by the remarkably simple equation

$$U_i(\mathbf{u}, \mathsf{s}) = u_i \mathbb{I}(\mathsf{s} < \tau) + U_j(\mathbf{u}, \mathsf{s} - \tau). \tag{4.32}$$

Although this approach produces the correct solutions, the simplest way to prove it is far from clear. In addition, conceptualizing $\mathbf{U}$ as a collection of survival functions is useful when $D = 0$ but misleading when $D \neq 0$; however, the catalytic case does not afford a simple solution strategy, and we do not consider it further.

4.4 Models of the experimental process

This section summarizes the mathematical machinery formalized in [115] by G.G., J.J.V., and L.P. G.G. developed this approach as a generalization of the framework constructed by G.G. in [112] by G.G., M.F., T.C., and L.P., and by G.G. in [107] by G.G. and L.P.

4.4.1 Snapshot sampling

To rigorously fit transient data, we need to posit just *how* a snapshot of cells may capture multiple cell states, such that some states are the progenitors of others. The solution is not yet clear, and multiple reasonable explanations exist; for example, we may suppose that the differentiation process "lags" in certain cells (in the vein of the models of variability proposed in [270] for development and in [220, 245] for the cell cycle). In other words, all cells are captured at a time t since the beginning of a process, but H and $\mathcal{A}$ have different time dependence for different cells. Although such an explanatory model can be instantiated, it may be too challenging to fit. Further, it does not appear to be compatible with processes that operate continuously; the choice of t becomes somewhat challenging to motivate.

We propose that the simplest model for observations relies on minimal synchronization between the biology and the experimental process. To mathematically formalize it, we take inspiration from the theory of reactor modeling in chemical engineering [88, 237]. A cell enters a medium; this entrance triggers a chemical signal that begins a transient process. The dynamics of this transient process are only dependent on time since receiving the signal, and identical between cells. After a delay, the cells exit the medium. In this framework, sequencing is the uniform random sampling of cells present within this medium. Although this formulation is admittedly simplistic — it excludes the cell cycle and stochastic driving — it allows us to take the first steps with a systematic study of using snapshot data to fit transient stochastic processes. This toy model is numerically tractable, which is useful for its simulation and characterization, and possesses a stationary state invariant with the time at which the experiment is performed, which is useful for biological admissibility and realism.

Therefore, to marginalize over t, we need to augment the model with an additional property: the relationship between time along a transient process and the probability of capturing a cell. In the parlance of reactor engineering, this relationship is given by the internal-age distribution f. The simulations of transient processes in [29, 168] implicitly adopt this model and assume a particular functional form of f. We might

suppose cells enter the observation window at $t = 0$ and leave it at $t = T$, with a Dirac residence time distribution $\delta(t - T)$ and uniform sampling throughout this window. The resulting age distribution is uniform, with $f = T^{-1}$, and formally corresponds to the ideal plug flow reactor (PFR) architecture [88]. As $T \to \infty$, we obtain the $t \to \infty$ ergodic limit, if such a limit exists. On the other hand, if $f \to \delta(t - T)$, we recover the instantaneous distribution at time T; this limit formally corresponds to the batch reactor (BR).

To obtain the generating function for the cells inside a tissue, we represent the tissue as a reactor, specify its influx and efflux properties, and solve for the internal-age distribution f. This internal-age distribution yields the occupation measure of the process times, as discussed in [112], and induces the following reactor-wide generating function:

$$\begin{aligned} G &= \int_t G(t) f(t) dt, \text{ where} \\ G(t) &= \sum_s G_s(t). \end{aligned} \tag{4.33}$$

We have marginalized over the instantaneous gene state s because this variable is typically not observable.

4.4.2 Droplet encapsulation noise

The generating function G describes the biological variability due to molecular processes, transcriptional driving, and the capture of cells from a reaction medium. However, single-cell RNA sequencing does not quantify cells — it quantifies *barcodes*. Cells are randomly encapsulated into droplets with barcoded beads; to avoid the formation of "doublets," with two cells per droplet, the microfluidic protocols typically have a fairly low encapsulation rate. If we assume that a droplet may have either zero or one cells, we obtain the following generating function for the distribution of RNA on a per-barcode level:

$$G_{\text{enc}} = p_1 G + p_0 = pG + (1 - p) = G_{\text{bc}}(G), \tag{4.34}$$

where G_{bc} is the PGF of the Bernoulli distribution, with $p_1 = p$ the probability of capturing a single cell and $p_0 = 1 - p$ that of capturing none. Analogously, if we assume that doublets can occur, and the encapsulation of cells is i.i.d., we find

$$\begin{aligned} G_{\text{enc}} &= p_2 G^2 + p_1 G + p_0 = p^2 G^2 + 2p(1 - p)G + (1 - p)^2 \\ &= [pG + (1 - p)]^2 = G_{\text{bc}}(G), \end{aligned} \tag{4.35}$$

where G_{bc} is now the PGF of the binomial distribution. It is straightforward to extend this to the unconstrained case, with per-cell encapsulation *rate* λ, and obtain the analogous expression

$$\begin{aligned} G_{\text{enc}} &= p_0 + p_1 G + p_2 G^2 + p_3 G^3 + \ldots \\ &= e^{\lambda(G-1)} = G_{\text{bc}}(G), \end{aligned} \tag{4.36}$$

where G_{bc} is the PGF of the Poisson distribution.

However, even empty droplets typically contain some "background" molecules. Removing the empty droplets by filtering for cells with relatively high expression, as well as correcting for the background, is a standard part of sequencing workflows [87, 187, 256, 322, 323]. To model the joint distribution of biological and background RNA, we need to instantiate a mechanistic hypothesis about its source. The simplest hypothesis consists of two parts. First, we impose the *pseudobulk* interpretation of background: we assume that a fraction of the cells loaded in the library construction step are lysed, and produce a pool of loose molecules. Next, we assume that these molecules are free to be encapsulated into the droplets in an i.i.d. fashion. This implies the Poisson functional form for the distribution of debris entering each droplet:

$$G_{\text{bg}} = \exp\left(c \sum_i \mu_i u_i\right), \tag{4.37}$$

where c is some shared constant that reflects the pool size and the rate of diffusion, whereas $\mu_i = \left.\frac{\partial G}{\partial u_i}\right|_{u_i=0}$ is the expectation of species i over the entire cell population. This simplest model assumes that all cells are equally likely to lyse and release their contents; if this assumption is violated, μ_i needs to be obtained by computing an expectation with respect to a measure biased toward the less stable cells. Finally, the full per-droplet distribution of molecules is

$$G_{\text{tot}} = G_{\text{bc}}(G) G_{\text{bg}}(G), \tag{4.38}$$

i.e., each droplet contains contributions from the encapsulated cells, as well as the background. With some abuse of notation, we note that the first argument denotes composition, whereas the second denotes functional dependence.

4.4.3 Library construction and sequencing noise

We cannot observe the biological molecule content of each droplet: we are restricted to analyzing counts of complementary DNA (cDNA). In a typical dual-index 3′ microfluidic workflow (e.g., the commercialized 10x chemistry [332]), these cDNA are

quantified by the following sequence of reactions. First, a synthetic primer captures a poly(A) stretch in RNA, which may be an endogenous molecule or a synthetic tag [268]. The primer contains a poly(dT) oligonucleotide, a sequencing primer, a cell barcode, and a unique molecular identifier (UMI). Next, reverse transcriptase (RTase) attaches to the RNA-primer complex and synthesizes the complementary strand. When the first strand is complete, a template-switching oligonucleotide (TSO) attaches to the end, allowing RT to synthesize the second strand of cDNA. After library construction, the droplet emulsion is broken, producing a pool of long cDNA; polymerase chain reaction (PCR) is used to amplify this pool. The long cDNA molecules are enzymatically fragmented, and another sequencing primer is attached at the end of the molecule that formerly contained the TSO. Finally, another round of PCR amplifies the pool and appends sample indices and Illumina adaptors to both sides of the molecule. The pool of cDNA is loaded onto a sequencing machine and sequenced from both sides, producing two reads. One read contains the barcode and UMI bases, whereas the other contains partial information about the $3'$ end of the molecule, beginning at the fragmentation site. This sequence of reactions represents the ideal-case scenario, and the products may well include artifacts due to off-target reactions [1].

To understand the effect of technical variability on the per-barcode distributions, we need to summarize this workflow in a mechanistic model. First, we assume that the library preparation reactions occur in an i.i.d. fashion relative to each RNA molecule in the droplet, allowing us to construct a separate description of technical noise for each discrete molecular species indexed by i. At this stage, we omit the modeling of continuous species. As we quantify the number of UMIs, we can considerably simplify the description by splitting the workflow into the initial cDNA synthesis and all downstream steps. For the cDNA synthesis, we may choose one of two models:

$$\begin{aligned} &\mathcal{X}_i \to \mathcal{X}_i + \mathcal{T}_i \text{ or} \\ &\mathcal{X}_i \to \mathcal{T}_i. \end{aligned} \tag{4.39}$$

In the first model, the formation of a UMI-tagged cDNA $\mathcal{T}_i$ is non-sequestering, and the template RNA $\mathcal{X}_i$ can participate in further cDNA synthesis. In other words, a single RNA molecule can produce more than one cDNA with distinct UMIs. In the second model, the cDNA synthesis is sequestering, and each RNA can template at most one cDNA with a particular UMI. For the downstream steps, if we assume the PCR and sequencing steps produce results that are reasonably faithful to their

templates, we are essentially restricted to a single model:

$$\mathcal{T}_i \to \varnothing. \tag{4.40}$$

In other words, the sequence of steps after the formation of cDNA $\mathcal{T}_i$ may lose some UMIs, but it cannot create them. Aggregating these steps, we find the shifted per-molecule generating function for technical noise:

$$\begin{aligned} G^*_{ti} = G_{ti} - 1 = e^{\lambda_i(g_i-1)} - 1 = e^{\lambda_i u_i} - 1 \text{ in the non-sequestering case and} \\ = p_i g_i + (1-p_i) - 1 = p_i u_i \text{ in the sequestering case,} \end{aligned} \tag{4.41}$$

where $\lambda_i = \lambda_{i,c} p_{i,p}$ and $p_i = p_{i,c} p_{i,p}$. $\lambda_{i,c}$ is the overall Poisson rate of the catalytic production of cDNA $\mathcal{T}_i$ with distinct UMIs, $p_{i,c}$ is the probability of producing a single cDNA $\mathcal{T}_i$ in a non-catalytic fashion, and $p_{i,p}$ is the probability of retaining a molecule of $\mathcal{T}_i$ through the PCR steps. It is straightforward to use a Taylor expansion to observe that the limit $\lambda_{i,c} \ll 1$ yields the Bernoulli form: if non-sequestering sequencing is relatively slow or inefficient, the probability of obtaining multiple cDNA from a single RNA is low, and the mathematically simpler Bernoulli noise form approximately holds.

Using the properties of PGFs, we find that the overall generating function is given by a simple composition, plugging in G_{ti} for g_i:

$$G_{\text{tot,t}} = G_{\text{tot}}(\mathbf{G}^*_t), \tag{4.42}$$

where we use the $G_{\text{tot}}(\mathbf{u})$ parametrization, and each entry of $\mathbf{G}^*_t$ contains the shifted generating function G^*_{ti} for a particular species i.

Finally, the reads associated with each cDNA $\mathcal{T}$ are not always uniquely identifiable: for example, the sequence content is typically sufficient to identify the gene, but if a read only covers an exonic portion of the gene, it is impossible to distinguish whether or not the original molecule has been spliced [80]. To correctly represent this ambiguity, we need to transform the arguments of the generating function from a length-n vector to a length ν-vector, such that ν is the total number of mutually distinguishable classes of molecules. The simplest form of this transformation is a linear categorical partition:

$$\mathbf{g} = \mathscr{P}^a \boldsymbol{\gamma}, \tag{4.43}$$

where $\mathscr{P}^a$ is an $n \times \nu$ ambiguity matrix with $\mathscr{P}^a_{i,\dot{\imath}}$ giving the probability of molecule i being identifiable in the equivalence class $\dot{\imath}$. We assume that each molecule can be

assigned to at least one class, implying $\sum_i \mathscr{P}^a_{i,i} = 1$. In principle, only the constraint $\sum_i \mathscr{P}^a_{i,i} \leq 1$ is mandatory, but the loss of molecules can be equivalently reframed as a technical noise component in $\mathbf{G}^*_t$.

We discuss the general case of this model component in Section B.2. In summary, the entries of $\mathscr{P}^a$ are challenging to identify, but it may be possible to exploit genomic information, polymer physics, and orthogonal long-read sequencing data to construct it from first principles. This formulation admits several special cases. For example, if we cannot distinguish any distinct species at all and can only quantify the total RNA content, $n = 1$ and $\mathscr{P}^a_{i,i} = 1$ for each i. Then we yield

$$\begin{aligned} (\mathbf{g})_i &= \mathscr{g} \text{ for all } i \text{ and} \\ G(\mathscr{g}) &= G\left(\begin{bmatrix} \mathscr{g} \\ \vdots \\ \mathscr{g} \end{bmatrix}\right). \end{aligned} \tag{4.44}$$

On the other hand, if all species are perfectly identifiable, we yield $n = n$ and $\mathscr{P}^a = I_n$, the n-dimensional identity matrix. If, say, we have $n = 2$ but $n = 3$, as in the case of nascent, mature, and ambiguous molecules described in [80, 168], we yield

$$G(\boldsymbol{\mathscr{g}}) = G\left(\begin{bmatrix} \mathscr{P}^a_{1,1}\mathscr{g}_1 + \mathscr{P}^a_{1,3}\mathscr{g}_3 \\ \mathscr{P}^a_{2,2}\mathscr{g}_2 + \mathscr{P}^a_{2,3}\mathscr{g}_3 \end{bmatrix}\right), \tag{4.45}$$

where $\mathscr{g}_1$ and $\mathscr{g}_2$ correspond to two unambiguously identifiable species, whereas $\mathscr{g}_3$ corresponds to ambiguous cDNA which may have come from either source. In the general case, we find

$$\begin{aligned} \mathbf{u} &= \mathscr{P}^a \boldsymbol{\mathscr{g}} - 1 \\ &= \mathscr{P}^a(\boldsymbol{u} + 1) - 1 \\ &= \mathscr{P}^a \boldsymbol{u} \\ &= \mathbf{G}_a(\boldsymbol{u}) - 1 := \mathbf{G}^*_a(\boldsymbol{u}), \end{aligned} \tag{4.46}$$

where each entry of the vector $\mathbf{G}_a$ contains the generating function of the relevant categorical distribution that governs how species i is parsed as one of the n identifiable species:

$$(\mathbf{G}_a(\boldsymbol{u}))_i = \sum_i \mathscr{P}^a_{i,i}\mathscr{g}_i. \tag{4.47}$$

Therefore, the overall GF takes the following form:

$$G_{\text{tot,ta}} = G_{\text{tot,t}}(\mathbf{G}^*_a(\boldsymbol{u})). \tag{4.48}$$

4.5 A unified framework for scRNA-seq stochasticity

We can summarize this entire theoretical machinery for Markovian processes as follows:

$$\begin{aligned} G_{\text{tot,ta}}(\boldsymbol{u}) &= G_{\text{bc}}\left(G(\boldsymbol{u})\right) G_{\text{bg}}(G), \text{ where} \\ G(\boldsymbol{u}) &= \int_t f(t) \sum_s G_s(\boldsymbol{u}, t) dt, \\ \frac{d\mathbf{G}}{d\mathsf{s}} &= -H^{\mathsf{T}}(t-\mathsf{s})\mathbf{G} - \mathbf{G} \odot \mathcal{A}(\mathbf{U}(\boldsymbol{u}, \mathsf{s}), t-\mathsf{s}), \text{ and} \\ \frac{d\mathbf{U}}{d\mathsf{s}} &= C\mathbf{U} + \text{diag}\ \mathbf{U}\ D\mathbf{U} \text{ with the initial condition} \\ \mathbf{U}(\mathsf{s}=0) &= \mathbf{G}_t^*(\mathscr{P}^a \boldsymbol{u}). \end{aligned} \tag{4.49}$$

In the non-Markovian case, we use the methods in Section 4.3.2 to compute $\mathbf{U}$.

4.6 Commonly encountered processes

This section is a brief summary of the supplement of [106] by G.G. and L.P. G.G. performed the derivations.

Although this framework is quite generic, we typically focus on the stationary distributions of a small number of two-stage memoryless processes. These processes are hypotheses which attempt to represent the joint nascent and mature distributions in single-cell RNA sequencing datasets. These models have $N = 1$, $n = 2$, and $m = 0$, and may optionally be endowed with technical noise. Here, we report their kinetics and distributions.

4.6.1 Constitutive model

The constitutive transcription model is the simplest nontrivial two-stage representation of RNA generation and processing. It includes the following kinetics:

$$\varnothing \xrightarrow{k} \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M \xrightarrow{\gamma} \varnothing, \tag{4.50}$$

where k is the transcription rate, β is the splicing rate, and γ is the degradation rate. This yields the operators and characteristics

$$\begin{aligned} \mathbf{u} &= \begin{bmatrix} u_N \\ u_M \end{bmatrix} \qquad C^{dd} = \begin{bmatrix} -\beta & \beta \\ 0 & -\gamma \end{bmatrix} \\ \mathbf{U} &= \begin{bmatrix} U_N \\ U_M \end{bmatrix} = \begin{bmatrix} u_N e^{-\beta \mathsf{s}} + \frac{u_M \beta}{\beta - \gamma}(e^{-\gamma \mathsf{s}} - e^{-\beta \mathsf{s}}) \\ u_M e^{-\gamma \mathsf{s}} \end{bmatrix} \\ \mathcal{A}(\mathbf{u}) &= k u_N, \end{aligned} \tag{4.51}$$

with all other operators set to zero. If $\gamma = \beta$, the system degenerates and yields $U_N(\mathbf{u}, \mathsf{s}) = e^{-\gamma \mathsf{s}}(u_N + \gamma u_M \mathsf{s})$. This formulation induces the following stationary generating function:

$$\begin{aligned} \log G(\mathbf{u}) &= \int_0^\infty \mathcal{A}(\mathbf{U}(\mathsf{s}))d\mathsf{s} = k \int_0^\infty \left[u_N e^{-\beta \mathsf{s}} + \frac{u_M \beta}{\beta - \gamma}(e^{-\gamma \mathsf{s}} - e^{-\beta \mathsf{s}}) \right] d\mathsf{s}. \\ &= u_N \frac{k}{\beta} + u_M \frac{k}{\gamma}. \end{aligned} \tag{4.52}$$

The joint distribution of this model is bivariate Poisson:

$$P(x_N, x_M) = P_{\text{Poiss}}(x_N; \mu_N) \times P_{\text{Poiss}}(x_M; \mu_M), \tag{4.53}$$

where $\mu_N = k/\beta$ and $\mu_M = k/\gamma$. At steady state, we can set k to unity with no loss of generality.

4.6.2 Bursty model

The two-stage *bursty* transcription model includes the following kinetics [261]:

$$\varnothing \xrightarrow{k} B \times \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M \xrightarrow{\gamma} \varnothing, \tag{4.54}$$

with stochastic burst sizes B drawn from a geometric distribution with scale b:

$$\begin{aligned} \mathbf{u} &= \begin{bmatrix} u_N \\ u_M \end{bmatrix} \qquad C^{dd} = \begin{bmatrix} -\beta & 0 \\ \beta & -\gamma \end{bmatrix} \\ \mathbf{U} &= \begin{bmatrix} U_N \\ U_M \end{bmatrix} = \begin{bmatrix} u_N e^{-\beta \mathsf{s}} + \frac{u_M \beta}{\beta - \gamma}(e^{-\gamma \mathsf{s}} - e^{-\beta \mathsf{s}}) \\ u_M e^{-\gamma \mathsf{s}} \end{bmatrix} \\ \mathcal{A}(\mathbf{u}) &= k \left[\frac{1}{1 - b u_N} - 1 \right]. \end{aligned} \tag{4.55}$$

This formulation induces the following stationary generating function:

$$\log G(\mathbf{u}) = \int_0^\infty \mathcal{A}(\mathbf{U}(\mathsf{s}))d\mathsf{s} = k \int_0^\infty \left[\frac{1}{1 - b U_N(\mathbf{u}, \mathsf{s})} - 1 \right] d\mathsf{s}. \tag{4.56}$$

This integral is not available in closed form, but it is easy to show that the nascent marginal has negative binomial distribution with shape k/β and scale b. At steady state, we can set k to unity with no loss of generality.

4.6.3 Extrinsic noise model

The *extrinsic* transcription model accounts for non-Poisson statistics by proposing that the transcription rate varies between cells[3]. It includes the following kinetics:

$$\varnothing \xrightarrow{k \sim K} \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M \xrightarrow{\gamma} \varnothing, \tag{4.57}$$

where k is the transcription rate drawn from a gamma distribution with shape ν and scale θ, β is the splicing rate, and γ is the degradation rate. This yields the operators and characteristics

$$\begin{aligned}
\mathbf{u} &= \begin{bmatrix} u_N \\ u_M \end{bmatrix} \qquad C^{dd} = \begin{bmatrix} -\beta & \beta \\ 0 & -\gamma \end{bmatrix} \\
\mathbf{U} &= \begin{bmatrix} U_N \\ U_M \end{bmatrix} = \begin{bmatrix} u_N e^{-\beta \mathsf{s}} + \frac{u_M \beta}{\beta - \gamma}(e^{-\gamma \mathsf{s}} - e^{-\beta \mathsf{s}}) \\ u_M e^{-\gamma \mathsf{s}} \end{bmatrix} \\
\mathcal{A}(\mathbf{u}) &= k u_N,
\end{aligned} \tag{4.58}$$

with all others set to zero. This formulation induces the constitutive stationary generating function conditional on a particular value of k:

$$\begin{aligned}
\log G(\mathbf{u}|k) &= u_N \frac{k}{\beta} + u_M \frac{k}{\gamma}; \text{ marginalizing, we find} \\
G(\mathbf{u}) &= \int_K G(\mathbf{u}|k) f_K(k) dk \\
&= \int_0^\infty e^{k\left(u_N \frac{1}{\beta} + u_M \frac{1}{\gamma}\right)} f_K(k) dk \\
&= M_K\left(u_N \frac{1}{\beta} + u_M \frac{1}{\gamma}\right) \\
&= \left(\frac{1}{1 - u_N \frac{\theta}{\beta} - u_M \frac{\theta}{\gamma}}\right)^\nu,
\end{aligned} \tag{4.59}$$

where the fourth line follows from recognizing the third line is the moment-generating function of the mixing distribution f_K, evaluated at a particular argument.

The joint distribution of this model is bivariate negative binomial [83]:

$$P(x_N, x_M) = \frac{\Gamma(\nu + x_N + x_M)}{\Gamma(\nu)\, x_N! x_M!} \frac{\nu^\nu \mu_N^{x_N} \mu_M^{x_M}}{(\nu + \mu_N + \mu_M)^{\nu + x_N + x_M}}, \tag{4.60}$$

where $\mu_K = \nu\theta$, $\mu_N = \mu_K/\beta$, and $\mu_M = \mu_K/\gamma$. In addition, each marginal follows the negative binomial distribution with shape ν and the corresponding mean. At steady state, we can set θ to unity with no loss of generality.

4.6.4 Technical noise models

To compute the distributions under the sequestering Bernoulli model of library construction, we make the substitutions

$$\begin{aligned}
u_N &\leftarrow p_N u_N \\
u_M &\leftarrow p_N u_M.
\end{aligned} \tag{4.61}$$

Model	μ_N	μ_M	F'_N	F'_M	Cov_{NM}
Constitutive	$\frac{1}{\beta}$	$\frac{1}{\gamma}$	0	0	0
Bursty	$\frac{b}{\beta}$	$\frac{b}{\gamma}$	b	$\frac{b\beta}{\beta+\gamma}$	$\frac{b^2}{\beta+\gamma}$
Extrinsic	$\frac{\nu}{\beta}$	$\frac{\nu}{\gamma}$	$\frac{1}{\beta}$	$\frac{1}{\gamma}$	$\frac{\nu}{\beta\gamma}$

Table 4.1: Lower moments of the three common models without technical noise.

Model	μ_N	μ_M	F'_N	F'_M	Cov_{NM}
Constitutive	$\frac{\lambda_N}{\beta}$	$\frac{\lambda_M}{\gamma}$	λ_N	λ_M	0
Bursty	$\frac{b\lambda_N}{\beta}$	$\frac{b\lambda_M}{\gamma}$	$\lambda_N(1+b)$	$\lambda_M\left(1+\frac{b\beta}{\beta+\gamma}\right)$	$\frac{b^2\lambda_N\lambda_M}{\beta+\gamma}$
Extrinsic	$\frac{\nu\lambda_N}{\beta}$	$\frac{\nu\lambda_M}{\gamma}$	$\lambda_N\left(1+\frac{1}{\beta}\right)$	$\lambda_M\left(1+\frac{1}{\gamma}\right)$	$\frac{\nu\lambda_N\lambda_M}{\beta\gamma}$

Table 4.2: Lower moments of the three models under Poisson noise.

To compute the distributions under the non-sequestering Poisson model of library construction, we make the substitutions

$$\begin{aligned} u_N &\leftarrow e^{\lambda_N u_N} - 1 \\ u_M &\leftarrow e^{\lambda_M u_M} - 1. \end{aligned} \tag{4.62}$$

4.6.5 Moment identities

By differentiating the generating functions, it is straightforward to obtain the lower moments of the resulting distributions (Section 3.1.3). In Tables 4.1 and 4.2, we report the lower moments for the underlying biological distributions and the noise-corrupted distributions. The variances are reported in terms of the shifted Fano factor $F'_i := \sigma_i^2/\mu_i - 1$, which produces the most compact representations.

Chapter 5

COMPUTATIONAL CONSIDERATIONS

5.1 Key challenges

The generating function procedure described in Chapter 4 offers some useful advantages: under some fairly restrictive, but physically realistic assumptions, it lets us evaluate joint and marginal distributions by computing integrals and inverse Fourier transforms. These distributions are approximate. The integrals are typically intractable and require numerical quadrature, which is computationally intensive and inexact. However, even when analytical solutions are available, some error is introduced by truncation: the probabilities are evaluated on a $N \times \mathfrak{s}_1 \times \cdots \times \mathfrak{s}_n$ grid of total size $\mathfrak{s} = N \prod_i \mathfrak{s}_i$, where each $\mathfrak{s}_i$ is a non-negative integer that represents the species-specific grid dimension. The grid is restricted to have a total probability that sums to unity; therefore, errors arise when a significant portion of the probability mass lies outside the evaluation bound.

To obtain accurate estimates, we need a sufficiently large grid. Here, a more formidable challenge arises: if we seek the generating function for a fairly large number of species, increasing the grid grows $\mathfrak{s}$ exponentially in n. This problem grows more acute, and lowers the efficiency, when the number of cells $\mathsf{N_c} \ll \mathfrak{s}$. This limit is easy to achieve: for example, if we have $N = 2$, $n = 3$, and a very modest $\mathfrak{s}_i = 10$ for all three species, we yield $\mathfrak{s} = 2,000$, which is comparable to a moderate to high-abundance cell type in a single-cell dataset. In other words, if we have $\mathsf{N_c} = 1,000$ cells, we generate 2,000 probabilities, then throw away more than half, because the computation of probabilities is coupled through the Fourier transform.

This problem is by no means restricted to generating function methods. Finite state projection also requires evaluating a grid of probabilities and discarding a large fraction of the computed values: as the observed states are coupled to non-observed ones, their probabilities are mathematically related. We are perhaps justified in saying that this problem is intrinsic to discrete distributions more broadly. Even the

simplest distributions on $\mathbb{N}_0$ implicitly require recursion:

$$\begin{aligned} P(x;\mu) &= \frac{1}{x!}\mu^x e^{-\mu} = \frac{\mu}{x}P(x-1;\mu) \text{ for Poisson and} \\ P(x;\theta) &= \left(\frac{\theta}{1+\theta}\right)^x \left(\frac{1}{1+\theta}\right) = \left(\frac{\theta}{1+\theta}\right) P(x-1;\theta) \text{ for geometric.} \end{aligned} \tag{5.1}$$

On one hand, it is somewhat obvious that we need to multiply x factors to compute $x!$ or p^x. On the other, it reflects a profoundly important statistical property: the evaluation time for the likelihood of a dataset under the Poisson model is a function of *the highest value*, $\mathfrak{s}$, rather than the dataset size $\mathsf{N_c}$, because we can "recycle" probabilities for any observed $x < \mathfrak{s}$. Even if $\mathsf{N_c}$ is small, a large $\mathfrak{s}$ will create problems in evaluation. From this perspective, we may reasonably say that the foundation of discrete probability was laid by Bernoulli, Poisson, and de Moivre [266], but made practical by Stirling [78] and consequent work on the approximation of special functions [188]. For the Poisson distribution,

$$\log P(x;\mu) = -\log\Gamma(x+1) + x\log\mu - \mu. \tag{5.2}$$

If $\log\Gamma$ can be evaluated reasonably efficiently throughout its domain, this approach breaks the tyranny of recursion and removes the dependence on $\mathfrak{s}$.

In sum, despite the generating function methods' advantages, they have fundamental limitations: integrals are typically intractable, and we are forced to evaluate them on a grid of spectral coordinates. In this section, we outline some strategies for implementing or bypassing these challenges.

5.2 Special function approximations

This section summarizes the content of [104] by G.G. and L.P. The special function approximations were conceptualized, designed, and implemented by G.G.

As discussed in Section 4.6.2, the bivariate PMF of the bursty model is not available in closed form, because the integral in Equation 4.56 is not analytically tractable. Therefore, if we would like to evaluate this PMF, the vast majority of the computational burden involves numerically approximating this integral by quadrature.

We can eliminate quadrature altogether by approximating $\mathcal{A}(\mathbf{U}(\mathfrak{s})) = \frac{1}{1-bU_N} - 1$ by a series and analytically integrating the terms over $(0,\infty)$. We have set k to unity with no loss of generality at steady state. This function affords the following

expansions:

$$\begin{aligned}\frac{1}{1-bU_N} - 1 &= \sum_{n=1}^{\infty} \frac{1}{2^{n+1}}(1+bU_N)^n - \frac{1}{2} \text{ whenever } |1+bU_N| < 2 \text{ and} \\ &= -\sum_{n=0}^{\infty} \frac{1}{(bU_N)^n} \text{ whenever } |bU_N > 1|.\end{aligned} \tag{5.3}$$

The first line reports the Taylor expansion about -1, whereas the second reports the Laurent expansion. This choice of expansion produces an overlapping region of convergence; if we had selected, for instance, the simpler Taylor expansion $\sum_{n=0}^{\infty}(bU_N)^n$, integrals up to s such that $|bU_N(\mathbf{u}, s)| = 1$ would diverge. We choose $u = \frac{1}{2b}(1+\sqrt{3})$ as the $|U_N|$ threshold for switching from the inner Taylor approximation to the outer Laurent approximation, as it maximizes the distance from the bounds of the region of convergence for non-positive complex $\mathbf{u}$ (Figure 5.1a).

All positive and negative integer powers of U_N have closed-form antiderivatives. We can exploit this property as follows. First, we compute all threshold values of s such that $|U_N(\mathbf{u}, s)| = u$ using numerical root-finding. Next, we partition the domain $s \in [0, \infty)$ into disjoint sets of intervals $\{S^T\}$ and $\{S^L\}$, such that $|U_N(\mathbf{u}, s)| < u \; \forall \; s \in S^T$ and $|U_N(\mathbf{u}, s)| \geq u \; \forall \; s \in S^L$. The form of U_N guarantees that there is are at most two domains in each set, and at least one Taylor domain S^T.

Next, we calculate the antiderivatives of the powers of U_N. For $\beta = \gamma$, we find

$$\begin{aligned}T_n(s) &:= \int U_N(\mathbf{u}, s)^n ds = -\frac{e^{n u_N/u_M} u_M^n}{\gamma n^{1+n}} \Gamma\left(1+n, \frac{n}{u_M}(u_N + \gamma u_M s)\right) \\ L_n(s) &= \int U_N(\mathbf{u}, s)^{-n} ds = -\frac{e^{n u_N/u_M}(-1)^n}{\gamma u_M^n n^{1-n}} \Gamma\left(1-n, -\frac{n}{u_M}(u_N + \gamma u_M s)\right).\end{aligned} \tag{5.4}$$

Per Equation 3.18, T_n can be computed by an sum of elementary functions. Similarly, L_n can be computed by the sum of a single exponential integral and several elementary functions.

For the non-degenerate case $\beta \neq \gamma$, we find

$$\begin{aligned} z &:= \left(1 - \frac{(\beta-\gamma)u_N}{\beta u_M}\right) e^{(\gamma-\beta)\mathsf{s}} \\ T_{\mathsf{n}}(\mathsf{s}) &= -\frac{(\beta-\gamma)e^{\gamma\mathsf{s}}}{\beta\gamma\mathsf{n}u_M} U_N^{1+\mathsf{n}} {}_2F_1\left(1, 1+\frac{\beta\mathsf{n}}{\beta-\gamma}; 1+\frac{\gamma\mathsf{n}}{\beta-\gamma}; z\right) \\ &= -\frac{(\beta-\gamma)e^{\gamma\mathsf{s}}}{\beta\gamma\mathsf{n}u_M(1-z)^{1+\mathsf{n}}} U_N^{1+\mathsf{n}} {}_2F_1\left(-\mathsf{n}, \frac{\gamma\mathsf{n}}{\beta-\gamma}; 1+\frac{\gamma\mathsf{n}}{\beta-\gamma}; z\right) \\ L_{\mathsf{n}}(\mathsf{s}) &= -\frac{\mathsf{n}(\beta-\gamma)e^{\gamma\mathsf{s}}}{\beta\gamma u_M(1-z)^{1-\mathsf{n}}} U_N^{1-\mathsf{n}} {}_2F_1\left(\mathsf{n}, -\frac{\gamma\mathsf{n}}{\beta-\gamma}; 1-\frac{\gamma\mathsf{n}}{\beta-\gamma}; z\right). \end{aligned} \tag{5.5}$$

Per Equation 3.21, T_n can be computed by a sum of elementary functions. Such a decomposition is not available for L_n. For completeness, we note that the antiderivatives in Equations 5.4 and 5.5 are not well-defined when $u_M = 0$, which is the trivial negative binomial case that does not require approximation.

Next, we truncate the summation in Equation 5.3 to upper limits N_T and N_L and compute the weights of each U_N^{n} term:

$$\begin{aligned} w_{T,\mathsf{n}} &= b^{\mathsf{n}} \sum_{k=\mathsf{n}}^{\mathsf{N}_T} \frac{1}{2^{k+1}} \binom{k}{\mathsf{n}} \\ w_{L,\mathsf{n}} &= b^{-\mathsf{n}}. \end{aligned} \tag{5.6}$$

Next, we obtain the approximations for the intervals:

$$\begin{aligned} \int_{S_T} \left[\frac{1}{1-bU_N} - 1\right] d\mathsf{s} &\approx \sum_{\mathsf{n}=1}^{\mathsf{N}_T} w_{T,\mathsf{n}} \left[T_{\mathsf{n}}(\sup S_T) - T_{\mathsf{n}}(\inf S_T)\right] \\ \int_{S_L} \left[\frac{1}{1-bU_N} - 1\right] d\mathsf{s} &\approx \sum_{\mathsf{n}=0}^{\mathsf{N}_L} w_{L,\mathsf{n}} \left[L_{\mathsf{n}}(\sup S_L) - L_{\mathsf{n}}(\inf S_L)\right]. \end{aligned} \tag{5.7}$$

As shown in Figure 5.1b, even low-order approximations can accurately recapitulate distribution shapes. To quantify the performance as a function of approximation order, we used a variant of the Kolmogorov-Smirnov distance. As the generating function is not guaranteed to produce a true PMF, probabilities can be negative and this distance can exceed 1. We find that the error is largely controlled by the Taylor approximation order (Figure 5.1c), whereas the runtime is largely controlled by the Laurent approximation order (Figure 5.1d). By decreasing the Laurent order and increasing the Taylor order, we can improve the time performance while keeping the error fairly low.

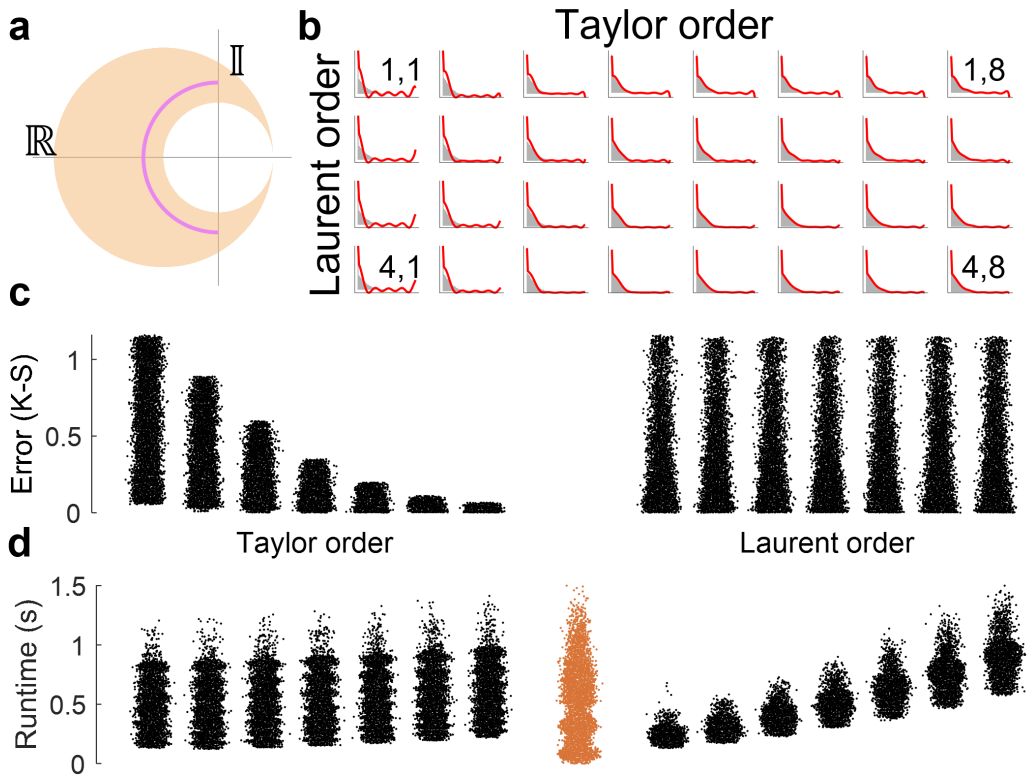

Figure 5.1: The special function approximation procedure for the two-species bursty model.

a. Taylor and Laurent approximation criterion (orange: approximations' common region of convergence; purple: threshold value of $|U_N|$).

b. Comparison of marginal mature copy number distributions for a range of approximation orders (#, # tuple and plot location: Laurent and Taylor approximation order; gray: histogram from 10^5 stochastic simulations; red line: distribution calculated from approximation; $b = 19$, $\beta = \gamma = 0.4$).

c. Kolmogorov-Smirnov error between quadrature- and expansion-based joint distributions for 2,500 $\beta = \gamma$ parameter sets on a uniform grid with $\log_{10} b \in [0.1, 2]$ and $\log_{10} \gamma \in [-1, 1]$, calculated for combinations of Taylor and Laurent orders up to 7 (black point: single parameter set; uniform jitter added).

d. Runtimes to compute approximations in **c** (black point: single parameter set computed using expansions; orange point: single parameter set computed using numerical quadrature; uniform jitter added).

As it stands, this approach is unsuited for large-scale computation. The method essentially substitutes integration with the calculation of special functions, which is helpful if these special functions can be easily approximated. We developed a bespoke algorithm for the exponential integral to implement the degenerate case $\beta = \gamma$ case, and did not implement the $\beta \neq \gamma$ case. It appears unlikely that useful approximations are forthcoming for the considerably more complicated hypergeometric function. In addition, in spite of runtime improvements, the procedure inherits the usual reliance on dense grid sampling (Section 5.1).

Nevertheless, the mathematical approach has some useful lessons for the development of solvers. With extensive prior understanding of the behaviors of functions and distributions, it is possible to develop approximators that take advantage of their properties, optimizing for features of interest while discarding others. These approximators do not generalize, and require considerable up-front work, but can outperform more naïve approaches for certain purposes.

5.3 Neural approximations

This section summarizes the content of [111] by G.G.*, M.C.*, T.C., and L.P. The MMNB and KWR approximations were conceptualized and designed by G.G. The nnNB approximation was conceptualized by G.G. and designed by M.C. The DR approximation was conceptualized and designed by M.C. All approximations were implemented by M.C. The quadrature methods were designed and implemented by G.G.

We apply these lessons to develop a solver that bypasses the grid evaluation procedure, taking inspiration from the discussion of the Stirling approximation in Section 5.1 to develop an approximator for the bursty system (Section 4.6.2). First, we recall that the nascent RNA distribution $P(x_N)$ is negative binomial. To compute the probability of a given microstate, we can apply the definition of conditional probability to find $P(x_M, x_N) = P(x_N)P(x_M|x_N)$.

This conditional is, of course, not available in closed form. Nevertheless, we have a qualitative understanding of its properties: $P(x_M|x_N)$ is unimodal, overdispersed relative to Poisson, and supported on $\mathbb{N}_0$. Thus, we may be able to construct an approximation $\hat{P}(x_M; \hat{\Theta}) \approx P(x_M|x_N)$. To ensure that $\hat{P}$ is in a reasonable class of approximators, we require it to have the qualitative properties of $P(x_M|x_N)$. To ensure that it improves the computational tractability of the problem, we require that it be a closed-form parametric distribution that can be computed in a non-recursive way for any x_M. The challenge is, then, to construct such a $\hat{P}$ and to produce a

mapping from $\Theta = \{x_N, b, \beta, \gamma\}$ to $\hat{\Theta}$.

The most obvious candidate for an approximator is $\hat{P} = P_{\mathrm{NB}}$, which has the correct distributional support and shape. It remains to specify the map $\mathcal{F} : \Theta \to \hat{\Theta}$. The conditional moments of the system are intractable. However, we know that those of the bivariate lognormal distribution *are* tractable; this five-parameter law (Equation 3.29) can be specified by defining the expectations, variances, and correlation of the two dimensions (Equations 3.27 and 3.30). If we proceed in this direction, we find that the conditional distribution is given by Equation 3.31, which is not defined at $y_1 = 0$ but produces a finite value elsewhere. Therefore, by noting that the lognormal distribution is unimodal, right-skewed, and has a strictly positive support, we can, in principle, obtain a moment-matched negative binomial (MMNB) approximation for the conditional distribution:

$$\hat{P}(x_M) = P_{\mathrm{NB}}(x_M; \hat{\nu}, \hat{\mu}). \tag{5.8}$$

First, we compute the parameters for the approximating bivariate lognormal distribution by applying Equations 3.27 and 3.30 to the moments in Table 4.1. Given this law, we compute the parameters ($\hat{\mu}_l$ and $\hat{\sigma}_l$) and moments ($\hat{\mu}$ and $\hat{\sigma}$) of the conditional lognormal distribution at $y_1 = x_M + 1$ using Equations 3.31 and 3.26. We shift the argument to ensure the conditional is well-defined for $x_M = 0$. Next, we calculate the shape parameter $\hat{\nu}$ using the following identity:

$$\hat{\nu} = \frac{\hat{\mu}^2}{\hat{\sigma}^2 - \hat{\mu}}. \tag{5.9}$$

To ensure the approximating conditional distribution is well-defined, we use Equation 5.8 only when $\hat{\nu}$ only when $\hat{\sigma}^2 > \hat{\mu}$; otherwise, we fall back to

$$\hat{P}(x_M) = P_{\mathrm{Poiss}}(x_M; \hat{\mu}). \tag{5.10}$$

This procedure comprises the function $\mathcal{F}$ used to convert x_N and biophysical parameters to the parameters of the approximating distribution.

With this coarse approximation, we can produce PMFs that roughly recapitulate the qualitative properties of the true PMF. The probabilities so obtained are unsuited to the computation of data likelihoods. Taking a broader view, we are part of the way to a usable approximation. It seems reasonable to suppose that we can get further by making minor corrections to the MMNB procedure.

Neural networks are good function approximators, and have previously been used to summarize the dynamics of physical systems [47, 179]. We take inspiration

from this approach to propose two improvements. First, we can use the procedure described above, correcting the conditional lognormal moments $\hat{\mu}$ and $\hat{\sigma}^2$:

$$\begin{aligned} \hat{\mu}^* &= \hat{\mu} c_\mu \text{ s.t. } c_\mu = s_1 \left(C_1 - C_1^{-1} \right) + C_1^{-1} \\ (\hat{\sigma}^*)^2 &= \hat{\sigma}^2 c_{\sigma^2} \text{ s.t. } c_{\sigma^2} = s_2 \left(C_2 - C_2^{-1} \right) + C_2^{-1}, \end{aligned} \tag{5.11}$$

where $s_1, s_2 \in (0, 1)$ are variables output by a neural network function $\mathcal{F}$ of Θ, whereas C_1 and C_2 are global (Θ-independent) scaling factors learned by the network. The corrected parameters $\hat{\mu}^*$ and $\hat{\sigma}^*$ so computed can be transformed into $\hat{\nu}^*$ using Equation 5.9, then substituted into Equations 5.8 and 5.10 to obtain probability estimates. By generating high-quality conditional distributions from a standard quadrature procedure, and updating $\mathcal{F}$, C_1 and C_2, we can produce a finer approximation to the true PMF. To train the network, we optimize the KLD between conditional distributions. These distributions are defined for all $x_M \in \mathbb{N}_0$, so we truncate them at $\mathfrak{s}_M = \mu_M + 4\sigma_M$ and normalize to yield strictly positive divergences. This is the neural network negative binomial (nnNB) procedure. The resulting approximations are fairly close to the true distributions, and can be improved further by tuning the neural network.

The nnNB procedure has the desired qualitative and statistical properties: it produces overdispersed bivariate distributions that bypass the Fourier grid evaluation. With a pre-trained network, to obtain the approximate likelihood $\hat{P}(x_N, x_M)$ for a parameter set $\{b, \beta, \gamma\}$, we need to calculate $\mathcal{F}$ only once. Nevertheless, the true conditionals are not negative binomial, and we can achieve better quantitative agreement by generalizing the approximator while retaining its key computational features.

To develop a better approximator, we note that finite Poisson and negative binomial mixtures can also be easily computed in a non-recursive fashion. In other words, we can propose the following functional form:

$$\begin{aligned} \hat{P}(x_M) &= \sum_{\mathsf{n}=1}^{\mathsf{N}} w_{\mathsf{n}} \hat{P}_{\text{ker}}(x_M; \hat{\nu}_{\mathsf{n}}, \hat{\mu}_{\mathsf{n}}), \text{ where} \\ \hat{P}_{\text{ker}}(x_M; \hat{\nu}, \hat{\mu}) &= P_{\text{NB}}(x_M; \hat{\nu}, \hat{\mu}) \text{ if } \hat{\nu} > 0 \text{ and} \\ &= P_{\text{Poiss}}(x_M; \hat{\mu}) \text{ otherwise.} \end{aligned} \tag{5.12}$$

This approach essentially approximates the true distribution by a weighted sum of basis functions, or kernels of the appropriate functional form. It remains to specify or learn the weights and distributional parameters of these kernels. The Nessie framework [271], seeking to fit fairly complicated univariate distributions, learns all

of these parameters simultaneously. However, as we have some qualitative insights into the distribution shape, we can simplify the procedure by judiciously placing the kernels. To improve performance in the high-probability regions, we place the approximators at the Chebyshev nodes of the lognormal quantile function F^{-1} (Equation 3.28):

$$\begin{aligned} p_{\mathsf{n}} &= \frac{1}{2}\left[\cos\left(\pi\frac{2\mathsf{n}-1}{2\mathsf{N}}\right)+1\right] \\ \hat{\mu}_{\mathsf{n}} &= F^{-1}(p_{\mathsf{n}};\hat{\mu}_l,\hat{\sigma}_l), \end{aligned} \tag{5.13}$$

where the conditional lognormal $\hat{\mu}_l$ and $\hat{\sigma}_l$ are obtained as in the MMNB procedure. Usefully, $\Phi^{-1}(p_{\mathsf{n}})$ need only be computed once. Knowing that $\hat{P}$ should be unimodal, we control the standard deviation of each kernel $\hat{\sigma}^*_{\mathsf{n}}$ by the spacing between adjacent kernels:

$$\hat{\sigma}^*_{\mathsf{n}} = c_\sigma(\hat{\mu}_{\mathsf{n}+1} - \hat{\mu}_{\mathsf{n}}) \text{ s.t. } c_\sigma = C_1 + sC_2, \tag{5.14}$$

where $s \in (0,1)$ is output by a neural function, $C_1 = 1$, and $C_2 = 5$. We somewhat arbitrarily set $\hat{\sigma}^*_{\mathsf{N}}$ to $\sqrt{\hat{\mu}_{\mathsf{N}}}$, meaning the Nth kernel is Poisson. Therefore, we can effectively approximate distributions by training a neural network function $\mathcal{F}$ of Θ, which outputs the weights $w_1,\ldots,w_{\mathsf{N}},c_\sigma$, precisely as in the nnNB case. This is the kernel weight regression (KWR) procedure.

In sum, by judiciously constructing kernel functions, then combining them using weights from a pre-trained neural network function, we can approximate conditional distributions for an intractable PMF (Figure 5.2a). To evaluate likelihoods, we can combine these conditional distributions with marginal distributions (Figure 5.2b). We used an adaptive quadrature generating function method as our training data and ground truth (QV20, evaluated using $\mathfrak{s}_i = \mu_i + 20\sigma_i$ and truncated to $\mu_i + 4\sigma_i$ for benchmarking). The accuracy of the KWR approximator was comparable to that of practical generating function methods (QV10, QV4, and FQ, or order-60 Gaussian quadrature), with runtimes per $\mathfrak{s}$ comparable to FQ. We additionally trained a direct regression (DR) method, which uses a neural function to map from $x_N, x_M, b, \beta, \gamma$ to $\hat{P}(x_N, x_M)$, in the spirit of [310]; at comparable neural network sizes, this approach yielded fairly poor performance. The differences are evident by inspection of reconstructed distributions: KWR and nnNB produce fair matches to ground truth, MMNB recapitulates the rough distribution shape, and DR suffers from extreme distortions (Figure 5.2d). Comparisons using a non-$\mathfrak{s}$-normalized metric confirm these results: KWR produces results far better than a random-w_{n} control and generally better than the other approximation strategies.

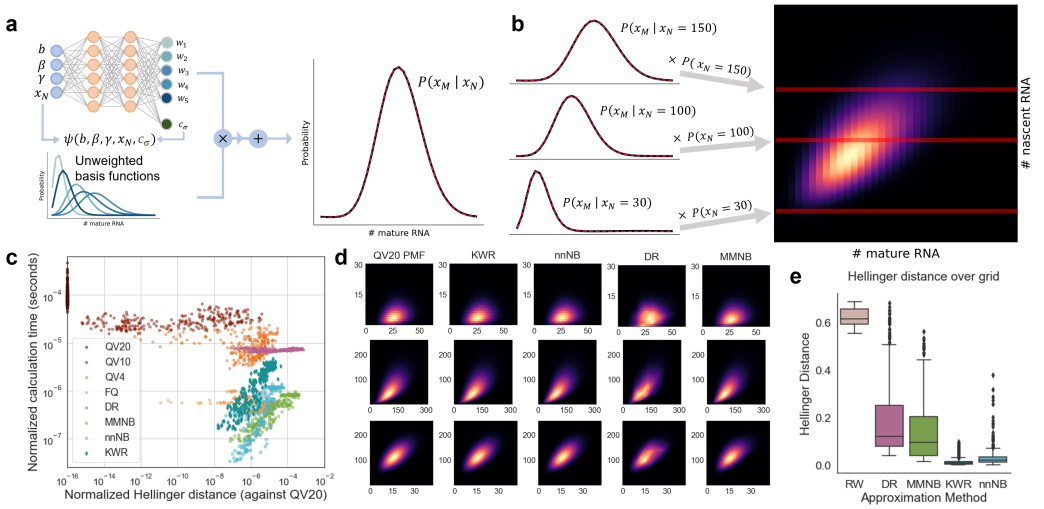

Figure 5.2: The neural network approximation procedure for the two-species bursty model.
a. Univariate conditional distributions are approximated by summing a set of kernel functions with neural network-learned weights (red dashed line: approximation; black line: ground truth distribution).
b. Bivariate distributions are reconstructed by multiplying conditional mature RNA probabilities by marginal nascent RNA probabilities (red dashed lines: approximations; black lines: ground truth distributions; heatmap: bivariate probability mass function, lighter is higher probability).
c. Runtime and accuracy of predictions, both normalized by grid size, for 256 test parameter sets, comparing three generating function-based methods (QV10, QV4, and FQ), direct regression (DR), the moment-matched negative binomial (MMNB) approximation, and kernel weight regression (KWR) to ground truth (QV20).
d. Typical distributions obtained by generating function inversion (QV20, leftmost column, ground truth) and various approximation methods.
e. Non-normalized grid reconstruction accuracy for 768 test parameter sets.

This approach appears to be quite promising and generalizable: given some prior knowledge distribution shapes, we can design a coarse approximation that exhibits the correct qualitative properties, then augment it with a neural correction. It is unlikely that this approach can generalize to arbitrary distributional dimensionality n: we are still constrained by generating high-quality training data, and the high-n cases require simulation. Nevertheless, some relevant classes of models appear to be immediately tractable. For example, although the case of $n = 1$ and $N = 2$ can produce bimodal distributions, the conditionals with respect to s are typically unimodal, suggesting the approximation

$$\hat{P}(x) = \hat{P}(s = 0, x)P(s = 0) + \hat{P}(s = 1, x)P(s = 1). \tag{5.15}$$

As it stands, the design advantages of the neural network procedures are essential for variational inference, where each cell c may have a different set of biological parameters and evaluating $\mathsf{N_c}$ Fourier transforms is impractical (Section 10.3). Even when the cells' copy number distributions are assumed to be identical, the procedure provides more stable likelihood functions. Due to the truncation implicit in defining data-based $\tilde{s}_i$ for generating function evaluation, parameter sets far from the optimum, with a large fraction of probability mass outside this bound, will produce artificially inflated likelihoods.

Nevertheless, we do not yet use the KWR solver to fit parameters when FQ is practical, for three reasons. First, likelihood inflation is a minor problem for parameter inference: method of moments estimates typically place us quite near the likelihood optimum. Second, FQ empirically shows a runtime advantage when the grid sizes are fairly small. Third, the neural approach does not allow us to easily integrate technical noise phenomena; to represent, e.g., the loss of molecules in the sequencing process, we need to retrain the network. For the Bernoulli noise model, we can partially exploit the statistical properties of distributions to forego designing a new solver. For example, the retention probabilities of nascent and mature molecules p_N and p_M are identical, we can directly use the neural solvers with b rescaled by p_N to evaluate distributions. We can similarly adapt the nnNB procedure when $p_M < p_N$, which amounts to computing the nascent marginal and conditional approximator for burst size bp_N, then weighing the conditional negative binomial scale parameter by p_M/p_N. However, this approach does not generalize to other noise behaviors, and considerable further work is necessary to implement them.

5.4 *Monod*

This section summarizes the supplementary content of [107] and [106] by G.G. and L.P. The method was conceptualized, designed, and implemented by G.G.

If we operate with bivariate data and assume that observations are independent and arise from a common distribution, it is by far easiest to numerically integrate generating functions. To this end, we developed *Monod*, a Python package for the inference of biophysical parameters.

At its heart, *Monod* is a wrapper around a *SciPy* L-BFGS-B optimizer [302]. To find the optimal parameters for a particular gene, we minimize the divergence between proposed distributions and observations. We calculate bivariate theoretical distributions by applying the inverse Fourier transform method described in [31,

261]. Specifically, we define a grid of g_N and g_M, such that

$$\begin{aligned}(g_N)_j &= e^{-\frac{2i\pi i}{\mathfrak{s}_N}} \text{ for } j = 0, \ldots, \mathfrak{s}_N - 1 \text{ and} \\ (g_M)_k &= e^{-\frac{2i\pi i}{\mathfrak{s}_M}} \text{ for } j = 0, \ldots, \lfloor \mathfrak{s}_M \rfloor + 1,\end{aligned} \tag{5.16}$$

where i denotes the unit imaginary number. Next, we define the matrix G_{jk}, such that

$$G_{jk} = G\left((g_N)_j - 1, (g_M)_k - 1\right), \tag{5.17}$$

where the function $G(u_N, u_M)$ is the distribution's generating function. To approximate the PMF on a $\mathfrak{s}_N \times \mathfrak{s}_M$ grid, we compute the inverse real fast Fourier transform of the matrix G:

$$P = \text{IRFFT}(G). \tag{5.18}$$

We use this form of the transform and truncate $\mathfrak{s}_M$ for evaluation because a probability mass function is a real-valued signal with a Hermitian Fourier transform [31].

To find the closed-form generating functions, we directly evaluate them. To find the generating functions only available in integral form, we use order-60 Gaussian quadrature [302] up to the upper bound

$$t = 10(1 + \beta^{-1} + \gamma^{-1}); \tag{5.19}$$

this scaling ensures all reactions have equilibrated.

Once we have a proposed distribution P, we optimize the Kullback-Leibler divergence (Equation 3.49) to obtain a point estimate of the biological parameters. To simplify this procedure, we begin at the method of moments estimate, if available. For numerical stability, we use the $\log_{10}$ versions of the parameters.

The usual bursty transcription model (Section 4.6.2) has three biological parameters per gene, which can be fit using the procedure outlined above. However, we would also like to learn the technical noise parameters, such as λ_N and λ_M for the Poisson sequencing model (Section 4.6.4). The evaluation of the generating function under this model amounts to applying Equation 4.62.

A naïve approach to this problem is typically futile: if we separately fit b, β, γ, λ_N, λ_M for each gene, the parameters turn out to be poorly distinguishable. However, we can use a physical argument to simplify the problem. λ_N and λ_M are *chemical*,

not biological parameters; we can reasonably suppose that they only depend on the extant molecules' properties, rather than obscure biological factors. Therefore, we make the simplest nontrivial assumption that mature molecules are all alike (i.e., have the same $\lambda_{M,\mathsf{g}} = \lambda_M$), whereas the sequencing of nascent molecules is largely controlled by the overall gene length L_g (i.e., $\lambda_{N,\mathsf{g}} = L_\mathsf{g} C_N$). Although this model is fairly crude, it turns out to produce apparently reasonable biological parameter trends (Section 8.2) and provides a foundation for constructing and testing more sophisticated hypotheses.

It remains to simultaneously obtain estimates of parameters

$$\begin{aligned} \Theta &= \{\Theta_1, \ldots, \Theta_{\mathsf{N_g}}, \Theta_t\}, \text{ such that} \\ \Theta_\mathsf{g} &= \{b_\mathsf{g}, \beta_\mathsf{g}, \gamma_\mathsf{g}\} \text{ and} \\ \Theta_t &= \{C_N, \lambda_M\}, \end{aligned} \tag{5.20}$$

such that the dataset likelihood is maximized. Formally, this is an optimization problem in $3\mathsf{N_g} + 2$ dimensions. Even if we assume the cell population is comprised of independent samples from a common distribution, the problem is not only intractable, but underspecified, and we need to make two assumptions to make inference practical.

Although the bursty model provides us with a way to evaluate the likelihood of the data $\mathcal{D}_\mathsf{g}$ for a particular gene, we formally need to optimize the KLD for a $2\mathsf{N_g}$-dimensional distribution that encodes potential patterns of co-regulation. In other words, if we independently fit each gene's data, we are intrinsically unable to reproduce correlations between between genes, because those correlations are not part of the model. In Section 10.1, we take the first tentative theoretical steps toward filling this lacuna. However, in the *Monod* implementation, we sacrifice the gene–gene relationships in favor of tractability.

Even under this implicit assumption of biological gene–gene independence, we retain coupling through the two technical noise parameters C_N and λ_M, which control all genes' likelihoods, and need to be fit simultaneously with the biological parameters (Equation 5.20). Therefore, we adopt the following schema, which is reminiscent of coordinate descent. We iterate over values of Θ_t on a grid. For a given set of Θ_t, we independently obtain gene-specific estimates $\hat{\Theta}_\mathsf{g}$, i.e., solve $\mathsf{N_g}$ relatively simple three-dimensional optimization problems. We store the resulting

parameter optima $\hat{\Theta}_g(\Theta_t)$. Next, we assign

$$\hat{\Theta}_t = \text{argmin}_{\Theta_t} \sum_{\mathsf{g}} D(\mathcal{D}_{\mathsf{g}} \parallel P(\hat{\Theta}_g(\Theta_t))), \tag{5.21}$$

where $P(\hat{\Theta}_g(\Theta_t))$ is the model probability distribution under parameters $\hat{\Theta}_g(\Theta_t)$. In other words, whichever Θ_t produces the lowest overall divergence is our optimum.

Although this approach is somewhat *ad hoc*, it provides certain advantages. For example, we can perform goodness-of-fit testing to focus on genes that fairly well recapitulate the data distributions. If we had simultaneously optimized all entries of $\hat{\Theta}$, we would need to re-run the optimization to account for the removal of some of the data. *Monod* uses three criteria for goodness-of-fit testing. First, we remove all genes whose parameters are near the search bounds, which usually represent failure to converge, excessive data sparsity, or model misspecification. Although these parameters may recapitulate distributions fairly well, they cannot be easily interpreted. Second, we remove all genes which simultaneously exceed pre-set chi-squared and Hellinger distance bounds, i.e., have unlikely and high-magnitude deviations from the proposed distribution. Typically, some 10% of the genes are discarded under the bursty model, upward of 50% under the extrinsic model, and nearly all under the constitutive model.

In addition, *Monod* allows the computation of uncertainties in $\hat{\Theta}_{\mathsf{g}}$. To bypass the problem of degeneracy with respect to the technical noise parameters, we compute these uncertainties conditional on the value of $\hat{\Theta}_t$. The uncertainties so derived are necessarily underestimates. Specifically, we use an approach based on the Fisher information matrix (FIM). Given a set of inferred parameters $\hat{\Theta}_{\mathsf{g}}$, the Fisher information matrix $\mathcal{I}$ is given by the Hessian of the Kullback-Leibler divergence:

$$\mathcal{I}_{ij} = \frac{\partial^2}{\partial \hat{\Theta}_{\mathsf{g},i}\, \partial \hat{\Theta}_{\mathsf{g},j}} D(\mathcal{D}_{\mathsf{g}} \parallel P(\hat{\Theta}_g(\hat{\Theta}_t))), \tag{5.22}$$

where i and j index over the inferred parameters $\log_{10} b$, $\log_{10} \beta$, and $\log_{10} \gamma$ [312]. The standard deviation of parameter i can be obtained from the diagonal entries of the inverse of the FIM:

$$\sigma_i = \sqrt{\left(\mathsf{N_c}^{-1} \mathcal{I}^{-1}\right)_{ii}}. \tag{5.23}$$

With these standard deviations, we use the z-score to estimate 99% confidence intervals as $2.576\sigma_i$ [208].

Given a set of fits, we can compare biological and technical noise parameters for cell populations. This procedure is somewhat limited by the precision of the inferred technical noise parameters, which are degenerate with respect to the burst size. However, we can make some progress by operating with matched samples. For example, if two cell populations are cell types collected in a single experiment, it appears reasonable to suppose they should have the same Θ_t, and any differences are purely biological, on the level of Θ_g. If we, in addition, strongly believe that the biological differences should be restricted to a handful of genes, rather than genome-wide, we can further "correct" inferred parameter values by subtracting any inter-dataset biases. We take this approach in Chapter 9. On the other hand, if they represent the same tissue processed using two different technologies, it seems reasonable to propose the Θ_g are identical, whereas Θ_t are different. We take this approach in Sections 8.3 and 8.4.

Although these procedures allow us to analyze data and draw interesting conclusions regarding the chemical and biophysical bases of transcriptome differences, the fits are only meaningful if the data meet the fairly restrictive assumptions of the model. For example, *Monod* does not include the encapsulation phenomena described in Section 4.4.2 or cell type heterogeneity. To account for the former, we remove all low-copy number barcodes, typically by the combination of the *bustools* filter [197] and knee plot thresholds. To account for the latter, we use two approaches. In the simplest one, we adopt the "marker gene" hypothesis, which supposes that intra-sample cell type differences can largely be attributed to a small number of genes, fit the datasets, and use goodness-of-fit testing to remove poorly fit genes *post hoc*. In the slightly more sophisticated one, we use pre-existing annotations to fit cell types separately. However, all of these methods represent uncomfortable compromises, and a truly comprehensive methodology should simultaneously and probabilistically account for the biological and technical effects.

To ensure the fits are informative, we further restrict our analysis to a relatively small set of genes with at least modest expression of both nascent and mature transcripts. In addition, we remove genes with excessive expression, as their likelihoods are numerically challenging to compute. This procedure typically retains several thousand genes. When we are interested in making comparisons between datasets, we restrict analysis to the genes that pass this filter in as many datasets as possible.

5.5 Simulations

In addition to numerically evaluating distributions, we frequently need to simulate from stochastic systems. In the context of data analysis, simulation provides us with a fully characterized ground truth for benchmarking various data transformations. However, this approach may be overly simplistic, as the simulation may not accurately represent all features of the underlying data-generating process. In the context of model development, simulations allow us to ensure that analytical or numerical solutions are correct.

The former use case is typically fairly straightforward. If we would like to generate realizations from a Markovian system with a particular set of physical phenomena, we either use a version of Gillespie's stochastic simulation algorithm (SSA) [98, 99] or sample directly from a numerically tractable PMF. If we seek to include other phenomena, we augment the simulator appropriately. For example, if we are interested in generating an observation from a system with cell type heterogeneity, we define cell type-specific parameters, randomly select a cell type, then generate a microstate, or molecular copy number, under the relevant parameters. Although the simulation of such Markovian systems is by no means trivial, it is well-understood, and we do not cover it in any further detail here. Instead, we report two algorithms for the exact stochastic simulation of non-Markovian systems, used to validate their numerical solutions.

5.5.1 Review of the SSA

This section extends a portion of [114] by G.G., S.Y., and L.P. The outline was written by S.Y. and G.G.

The Markovian SSA for a system with time-independent parameters takes the following form:

1. Initialize the system at time $t = t_0$ and state $\mathbf{x} = \mathbf{x}^0$.
2. Compute the instantaneous reaction rates of the μth reaction, $\phi_\mu(\mathbf{x})$ and the net state efflux rate, $\phi(\mathbf{x}) = \sum_\mu \phi_\mu(\mathbf{x})$.
3. Generate u_1 and u_2, random variables uniformly distributed over $(0, 1)$.
4. Transform u_1 to obtain the exponentially distributed residence time $\Delta t = -\frac{1}{\phi(\mathbf{x})} \log u_1$.
5. Use u_2 to compute the reaction index μ, such that $\sum_{k=1}^{\mu-1} \frac{\phi_k(\mathbf{x})}{\phi(\mathbf{x})} < u_2 \leq \sum_{k=1}^{\mu} \frac{\phi_k(\mathbf{x})}{\phi(\mathbf{x})}$.
6. Advance system to time $t \leftarrow t + \Delta t$ and state $\mathbf{x} \leftarrow \mathbf{x} + \Delta\mathbf{x}_\mu$.

7. Return to step 2 or terminate simulation.

The instantaneous propensity function for zero-order reactions, such as $\varnothing \xrightarrow{c} \mathcal{X}_i$, is $\phi_\mu(\mathbf{x}) = c$, a constant. The propensity function for first-order reactions, such as $\mathcal{X}_i \xrightarrow{c} \varnothing$, is $\phi_\mu(\mathbf{x}) = cx_i$. Higher-order propensity functions are reported in Table 2.1 of [299].

The update vector $\Delta\mathbf{x}_\mu$ consists of the entries of the stoichiometry matrix corresponding to reaction μ. This quantity can be random; for example, to simulate the bursty system in Section 4.6.2, we would generate a realization of the geometric distribution with mean b whenever a transcriptional event occurs, and add it to x_i. In addition, a stochastic burst size can be easily made time-dependent. However, time dependence in reaction rates, as well as non-Markovian dynamics, require slightly more elaborate adjustments.

5.5.2 SDE–CME systems

This section summarizes a portion of the supplement of [113] by G.G.*, J.J.V.*, M.F., and L.P. The method was conceptualized, designed, and implemented by G.G.

Hybrid continuous–discrete stochastic systems (with $n, m > 0$ and $C^{cd} \neq 0$) are popular for representing extrinsic variability in reaction rates, such as time-varying environments. However, in general, the simulation of these systems requires approximate schema. The computation of propensities in Section 5.5.1 belies the fact that the more fundamental variable is the flux $\phi(\mathbf{x}, t^*)$:

$$\begin{aligned} -\log u_1 &= \sum_\mu \int_t^{t+\Delta t} \phi_\mu(\mathbf{x}, t^*) dt^* \\ &= \sum_\mu \int_t^{t+\Delta t} \phi_\mu(\mathbf{x}) dt^* \\ &= \sum_\mu \phi_\mu(\mathbf{x})\Delta t = \Delta t \phi(\mathbf{x}), \end{aligned} \tag{5.24}$$

where the first equality holds generally [227], whereas the second holds only if all ϕ_μ are time-independent. To compute Δt if ϕ_μ is time-dependent, we typically need to use a numerical solver to find the root Δt where the first line of Equation 5.24 holds. Finally, the reaction index μ is drawn from the appropriate categorical

distribution, such that

$$\begin{aligned}\sum_{\nu=1}^{\mu-1}\frac{\phi_\mu(\mathbf{x})}{\phi(\mathbf{x})} < u_2 \leq \sum_{\nu=1}^{\mu}\frac{\phi_\mu(\mathbf{x})}{\phi(\mathbf{x})},\ \text{with}\\ \phi_\mu(\mathbf{x}) := \int_t^{t+\Delta t}\phi_\mu(\mathbf{x})dt^*\\ \phi(\mathbf{x}) := \sum_\mu \phi_\mu(\mathbf{x}).\end{aligned} \tag{5.25}$$

This task is particularly challenging when rates are time-dependent in a stochastic fashion. For example, Brownian motion is fractal, and does not afford exact roots. Therefore, a system with a Brownian motion component must be solved by sampling the process on a grid, then using interpolation to approximate fluxes [254, 305]. This may lead to errors if the grid is insufficiently fine, or excessive evaluation times if it is too fine.

Curiously, certain non-Brownian extrinsic noise sources do afford exact simulation routines. Specifically, if the stochastic driver $y_t = y(t)$ is a jump Ornstein–Uhlenbeck process, it has a non-fractal, fully specified structure [43, 57]:

$$\begin{aligned}dy_t = -\kappa y_t + dL_t \text{ such that}\\ L_t = \sum_{k=0}^{\mathsf{N}(t)} B_k,\end{aligned} \tag{5.26}$$

such that $\mathsf{N}(t)$ is a Poisson random variable with mean at. $\{B_k\}$ is a set of independent and identically distributed realizations of the positive-valued random variable B. This process has the exact solution [241]

$$y_t = \sum_{k=0}^{\mathsf{N}(t)} B_k e^{-\kappa(t-t_k)}, \tag{5.27}$$

where t_k are the arrival times of $\mathsf{N}(t)$. To account for the initial condition, we set $t_0 = 0$ and $B_0 = y_0$. This class of processes has been exhaustively studied by Barndorff-Nielsen and colleagues in the context of mathematical finance [21, 22].

To generate a single realization of the arrival process on $[0, T]$, we draw a Poisson random variable $\mathsf{N}(T)$ with mean aT. To generate the jump times, we draw $\mathsf{N}(T)$ uniform random variables on $[0, T]$ and sort them. To generate the jump sizes, we draw $\mathsf{N}(T)$ random variables from the jump distribution. Equation 5.27 immediately yields the time-dependent trajectory y_t.

If y_t drives a transcription process, such that the production of some species x_i occurs at the rate y_t, the instantaneous propensity of this reaction is simply $\phi_y(\mathbf{x}, t) = y_t$. Usefully, we can integrate it:

$$\begin{aligned}\Phi_y(t) &:= \int_0^t y(t^*)dt^* = \Phi_y(t_k) + \int_{t_k}^t y(t^*)dt^* \\ &= \Phi_y(t_k) + \frac{y(t_k)}{\kappa}\left(1 - e^{-\kappa(t-t_k)}\right),\end{aligned} \tag{5.28}$$

where $t_k < t$ but $t_{k+1} \geq t$, if one exists. This identity holds because no arrivals take place between t_k and t_{k+1}: in this region, the dynamics of y_t are a simple deterministic exponential decay. Therefore, we can immediately compute $\Phi_y(t_k)$ for all k.

Therefore, to simulate the system, we need to solve the following equation:

$$\begin{aligned}-\log u_1 &= \sum_{\mu \neq y} \int_t^{t+\Delta t} \phi_\mu(\mathbf{x}, t^*)dt^* + \int_t^{t+\Delta t} \phi_y(\mathbf{x}, t^*)dt^* \\ &= \Delta t \sum_{\mu \neq y} \phi_\mu(\mathbf{x}) + \Phi_y(t + \Delta t) - \Phi_y(t),\end{aligned} \tag{5.29}$$

where the second line holds because we have assumed all but one of the reactions have time-independent propensities.

First, suppose that $t > t_k$ for all k. In this case, we need to solve the following root-finding problem:

$$-\log u_1 = \Delta t \sum_{\mu \neq y} \phi_\mu(\mathbf{x}) + \frac{y(t)}{\kappa}\left(1 - e^{-\kappa \Delta t}\right), \tag{5.30}$$

which has an analytical solution in terms of the Lambert W function (Equation 3.23):

$$\begin{aligned}&C_1 = \sum_{\mu \neq y} \phi_\mu(\mathbf{x}), \qquad C_2 = \frac{y(t)}{\kappa}, \qquad C_3 = C_2 - \log u_1 \\ &\Delta t = \frac{1}{\kappa} W\left(\frac{\kappa C_2}{C_1} e^{\kappa C_3 / C_1}\right) - \frac{C_3}{C_1} \text{ whenever } C_1 > 0, \text{ and} \\ &\quad = -\frac{1}{\kappa}\log\left(\frac{C_3}{C_2}\right) \text{ otherwise.}\end{aligned} \tag{5.31}$$

In the nontrivial case, we use the pre-computed values of Φ_y to bound Δt. Specifically, we find the highest k such that

$$\begin{aligned} \Delta t \sum_{\mu \neq y} \phi_\mu(\mathbf{x}) + \Phi_y(t_{k+1}) - \Phi_y(t) > -\log u_1 \text{ but} \\ \Delta t \sum_{\mu \neq y} \phi_\mu(\mathbf{x}) + \Phi_y(t_k) - \Phi_y(t) < -\log u_1. \end{aligned} \tag{5.32}$$

We immediately find that $\Delta t = t_k - t + \Delta t^*$. Δt^* is the waiting time between t_k and the reaction firing time, and is computed analogously to Equation 5.31:

$$\begin{aligned} C_1 &= \sum_{\mu \neq y} \phi_\mu(\mathbf{x}) \\ C_2 &= \frac{y(t_k)}{\kappa} \\ C_3 &= C_2 - \log u_1 - (t_k - t) \sum_{\mu \neq y} \phi_\mu(\mathbf{x}) - \Phi_y(t_k) + \Phi_y(t) \\ \Delta t^* &= \frac{1}{\kappa} W\left(\frac{\kappa C_2}{C_1} e^{\kappa C_3 / C_1}\right) - \frac{C_3}{C_1} \text{ whenever } C_1 > 0, \text{ and} \\ &= -\frac{1}{\kappa} \log\left(\frac{C_3}{C_2}\right) \text{ otherwise.} \end{aligned} \tag{5.33}$$

In other words, the deterministic behavior of the trajectories between jump arrivals events allows us to pre-compute and constrain the reaction times. Once we know the region where the reaction time lies, computing it is as simple as subtracting the total flux up to the jump arrival time (the correction to C_3 in Equation 5.33) and solving a root-finding problem using a special function. The reaction index is then selected according to Equation 5.25, with

$$\phi_y(\mathbf{x}) = \Phi_y(t + \Delta t) - \Phi_y(t). \tag{5.34}$$

Attention must be paid to certain edge cases, as well as the numerical stability of the W calculation. However, overall, this approach provides a generalizable, exact strategy for simulating discrete processes driven by a jump Ornstein–Uhlenbeck process, and can be applied to a broad variety of systems that are not tractable by analytical approaches.

5.5.3 Delay master equations

This section summarizes a portion of [114] by G.G., S.Y., and L.P. The method was conceptualized, designed, and implemented for the case of deterministically delayed degradation by S.Y. and extended to the case of arbitrary delayed degradation and interconversion by G.G. The outline was written by S.Y. and G.G.

To adapt the procedure in Section 5.5.1 to the case of non-Markovian degradation

and interconversion, we make the following adjustments.

The first modification treats removal events of delayed species, i.e, transcripts that undergo reactions with non-exponential waiting times. Two empty queues are initialized: one for times and one for reaction indices. Then, if the reaction index generated in step 5 is the creation of a delayed species, the queues are populated with the time and reaction index of the removal of that species. This time is simply $t + \Delta t$, where Δt is drawn from any distribution on $\mathbb{R}_+$. If the system is initialized with delayed species, the queues of times and reaction indices must be pre-defined accordingly. For simplicity, we always assume that existing delayed species were created at $t = 0$.

The second modification alters the calculation of flux, specifically accounting for the contributions of species that don't yet exist, but will after some delay. The total flux, $\phi(\mathbf{x})$, is computed at each queued reaction event, producing a monotonically increasing, piecewise linear function of Δt. Then, the residence time corresponding to the random flux generated by $-\log u_1$ is found analytically. The computation of the residence time is essentially equivalent to the direct method outlined in [39], and amounts to linear interpolation between the arrival times of queued reactions.

The third modification ensures that all reactions happen in the correct order. After step 5, the reaction time and event are stored, and before advancing the system in step 6, all queued reactions that are to happen before the stored reaction event are sequentially applied and stored. The resulting ordered list is then converted into system times and states.

Chapter 6

SNAPSHOT INFERENCE

> Technology does everything possible so that we lose sight of the chain of cause and effect.
>
> *Turning Back the Clock: Hot Wars and Media Populism*
> UMBERTO ECO

This chapter is essentially complete, but I will need to ensure that the literature review is up to date.

6.1 Critical analysis of RNA velocity

This section summarizes the content of [112] by G.G., M.F., T.C., and L.P. The critique was conceptualized by G.G. and L.P. and implemented by G.G., M.F., and T.C.

The method of *RNA velocity* [168] aims to infer directed differentiation trajectories from snapshot single-cell transcriptomic data. Although we cannot observe the transcription rate, we can count molecules of spliced and unspliced mRNA. The unspliced mRNA content is a leading indicator of spliced mRNA, meaning that it is a predictor of the spliced mRNA content in the cell's near future. This causal relationship can be usefully exploited to identify directions of differentiation pathways without prior information about cell type relationships: "depletion" of nascent RNA suggests the gene is downregulated, whereas "accumulation" suggests it is upregulated. This qualitative premise has profound implications for the analysis of scRNA-seq data. The experimentally observed transcriptome is a snapshot of a biological process. By carefully combining snapshot data with a causal model, it is for the first time possible to reconstruct the dynamics and direction of this process without prior knowledge or dedicated experiments.

The bioinformatics field has recognized this potential, widely adopting the method and generating numerous variations on the theme. The roots of the theoretical approach date to 2011 [326], but the two most popular implementations for scRNA-seq were released in 2017–2018: *velocyto* by La Manno et al. [168], which introduced the method, and *scVelo* by Bergen et al. [29], which extended it to fit a more sophisticated dynamical model. Aside from these packages, a dizzying variety of auxiliary

methods and extensions have been developed [13, 41, 63, 77, 120, 130, 135, 171, 175, 183, 192, 229, 246, 249, 281, 311, 314, 330, 331] to incorporate additional modalities, build more complex dynamical and statistical models, and construct low-dimensional visualizations. This profusion of computational extensions has been accompanied by a much smaller volume of analytical work, including discussions of potential extensions and pitfalls [30, 50, 275, 291], as well as theoretical studies based on optimal transport [173, 329] and stochastic differential equations [178]. However, at their core, these auxiliary methods are built on top of the theory and code base from *velocyto* or *scVelo*.

Despite the popularity of RNA velocity [264, 311] and increasingly sophisticated attempts to combine it with more traditional methods for trajectory inference [171, 330], there has been little comprehensive investigation of the modeling assumptions that underlie the seemingly simple user-facing workflow (Figure 6.1a-b). The few dedicated critiques to date have largely focused on limitations of the inference and embedding steps [30, 192, 333], without questioning the foundational assumptions. This is an impediment to applying, interpreting, and refining the methods, as problems arise even in the simplest cases. Consider, for example, the result displayed in Figure 6.1b, where the outputs of the two most popular RNA velocity programs applied to exemplar human embryonic forebrain data [168] are qualitatively different. The inferred directions in the example should recapitulate a known differentiation trajectory from radial glia to mature neurons. However, *scVelo*, which "generalizes" *velocyto*, fails to identify, and even reverses the trajectory, suggesting totally different causal relationships between cell types. This type of problematic result has been reported elsewhere [29, 30, 120, 171, 175, 231], and typically used to motivate the development of new implementations. Nevertheless, the methods have produced plausible trajectories in biological studies [26, 61, 117, 126, 147, 182, 263, 317, 324], suggesting that they can identify *something* nontrivial about the underlying signal.

Motivated by such discrepancies, we systematically investigated the method's theoretical foundations, assumptions, and implementations, using a combination of simulated and biological datasets. The RNA velocity procedure combines numerous steps of data processing, reviewed in Section 6.1.1. Some are justified under a particular, very restrictive, physical model of transcription, whereas others are purely *ad hoc*. We conclude that the variable performance is, in large part, an intrinsic consequence of incompatibility between these two worldviews. An *ad hoc*

data transformation assumes some system dynamics or distributional form, which are generally incompatible with the physical model. Occasionally, the assumptions approximately hold, yielding results consistent with known biology. However, it is *a priori* impossible to predict whether they hold. The presence of somewhat arbitrary tunable hyperparameters at each step of the analysis provides an opportunity for confirmation bias to overrule the data, exacerbating the reliability problems.

6.1.1 Brief review of RNA velocity

To characterize the challenges, we briefly outline the steps of a typical velocity workflow. First, raw reads are converted to unspliced (nascent) and spliced (mature) RNA count matrices, based on the presence or absence of intronic content. After the usual filtering and normalization steps, the mature count matrix is projected to a lower-dimensional space with principal component analysis (PCA). This projection is used to construct a nearest-neighbor graph over cells, and "impute" the data matrices by replacing each cell's normalized counts with the average of its neighbors' values. In other words, the preliminary data processing effects a transformation $x_i \to y_i$, where x_i is discrete and y_i is continuous. Then, the following model is instantiated:

$$\varnothing \xrightarrow{\alpha(t)} y_N \xrightarrow{\beta} y_M \xrightarrow{\gamma} \varnothing. \tag{6.1}$$

The functional form of $\alpha(t)$ is not precisely specified. *velocyto* assumes that $\alpha(t)$ has fairly generic dynamics, but evolves slowly enough relative to β and γ to produce identifiable near-equilibrium high-expression and low-expression states. *scVelo* relaxes the assumption of equilibrium, but restricts the dynamics of $\alpha(t)$ to a much simpler function with a single step increase and decrease. In addition, the precise meaning of y_i differs by source: *velocyto* treats it as μ_i, such that x_i is drawn from a Poisson distribution with mean μ_i, whereas *scVelo* treats it as a true y_t, a continuous quantity that happens to be corrupted by isotropic noise.

Regardless of the interpretation, the following identity holds:

$$\frac{dy_M(t)}{dt} = \beta y_N(t) - \gamma y_M(t). \tag{6.2}$$

This time derivative is the "RNA velocity," the rate of change of the mature RNA abundance. By fitting a model — either extracting the y_N, y_M values at equilibria and fitting a line (*velocyto*), or using all of the data and fitting a curve (*scVelo*) — it is possible to identify γ/β and compute instantaneous velocity values for each cell.

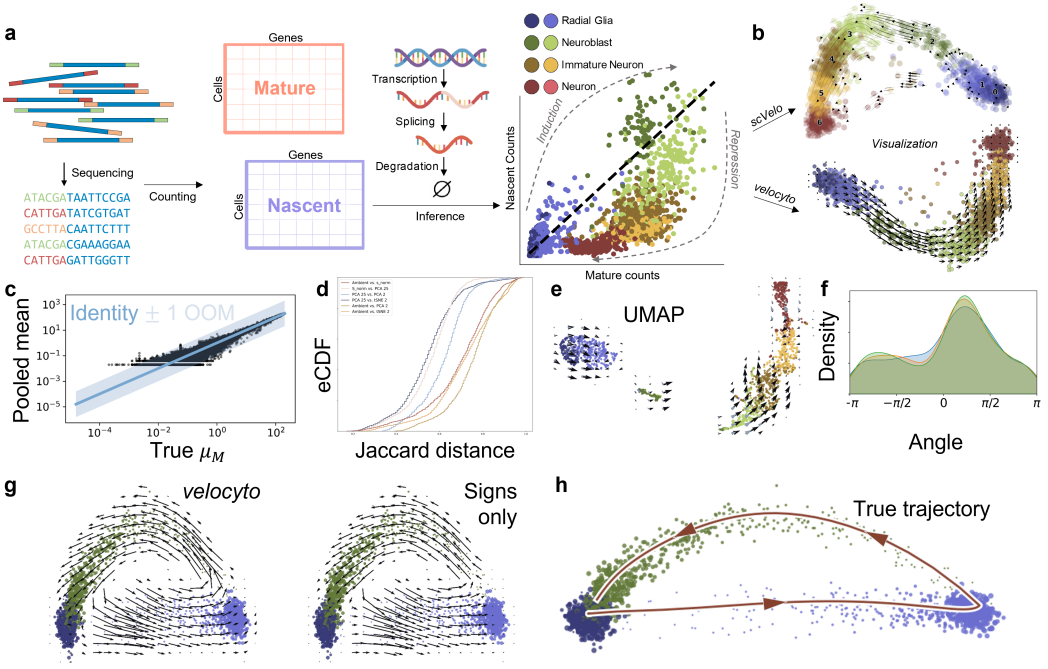

Figure 6.1: The RNA velocity workflow and its limitations.
a. A summary of the user-facing components of a typical RNA velocity workflow. Initial processing of sequencing reads produces nascent and mature counts for every cell, across all genes. Inference procedures fit a model of transcription and predict cell-level velocities, ascribing accumulation or depletion of RNA to induction or repression of the transcriptional driver (visualizations adapted from [129], forebrain data from [168]).
b. At the final stage of the workflows, cell and embedding velocities are displayed in the top two principal component dimensions, but different software implementations may disagree.
c. Smoothing and imputation introduce distortions into the simulated data, and do not recapitulate the simulated ground truth process average μ_N.
d. Normalization and dimensionality reduction distort local cell neighborhood identities (eCDF: empirical cumulative distribution function; Jaccard distance: Equation 3.50, lower is better).
e. The nonlinear UMAP embedding distorts the global cell type structure, separating cell types along a continuous trajectory.
f. Nonlinear transformations and modulation of neighborhood sizes introduce distortions in the arrow directions with respect to the simplest PCA projection (histograms: distribution of cell-specific angle deviations under different pooling neighborhood sizes).
g. Nonlinear embeddings of cell-specific velocities into PCA space, computed from simulated data, do not appear to substantially change if only the velocity signs are used.
h. If a parametric fit to the dataset is available, it can be summarized by projecting the inferred time-dependent process average into a low-dimensional space.

Next, the velocity values are summarized in a low-dimensional representation. An embedding is constructed from the PCA projection, and each cell's neighbors in that embedding are identified. The low-dimensional "direction" of the mature transcriptome is computed by calculating the degree of alignment between the RNA velocity of each cell and the directions to its embedding neighbors by passing these quantities through a kernel function. Finally, a low-dimensional, cell-specific velocity vector is produced by averaging the directions to the neighbors.

6.1.2 The velocity dynamical model is unphysical

The procedure ultimately relies on the transformation from x_i to y_i producing a quantity that follows Equation 6.2. This premise is merely asserted, never proven or quantitatively tested by the method developers, and fails on five levels.

The transformation is not theoretically founded. There is no particular reason to believe that averaging over neighbors should eliminate biological or technical stochasticity. In addition, the *space* used to identify the nearest neighbors is constructed from the data, and incorporates its noise sources. The premise of obtaining a better estimate by aggregating noisy data is superficially plausible; for example, such "local averages" are ubiquitous in time series analysis, including transcriptomics [102]. However, a data-based projection is not an externally determined experiment time, and the imputation of data is fundamentally circular in a way that a moving average is not. These, and other, pitfalls of imputation have been characterized elsewhere [10], and we describe further theoretical issues in Section B.3.

Regardless of its theoretical basis, the transformation does not actually accomplish its goals. Even in the best-case scenario, in discrete simulated data generated from a model that matches the *velocyto* assumptions, with no normalization needed, the procedure only recapitulates the true expectation μ_M on average, and is unreliable for any specific cell (Figure 6.1c). The performance is particularly poor for cells that are strongly out of equilibrium, i.e., those of the most interest for the procedure.

The alternative continuous model is inappropriate and unphysical. Although continuous approximations are reasonable in the high-concentration regime, typical scRNA-seq experiments have very low copy numbers across the genome; for the vast majority of genes, only a small fraction of cells have nonzero RNA counts. This regime contradicts the assumptions of the approximations. Making matters worse, the additive noise term used for this model in *scVelo* does not even match the multiplicative, abundance-dependent noise term that emerges from typical approx-

imations [100].

Even if we adopt the discrete model with instantaneous Poisson noise, we contradict numerous sources that suggest transcriptional activity varies with time even in stationary cell populations [16, 65, 161, 172, 205, 210, 233, 239, 244], and is effectively described by a telegraph model that stochastically switches between active and inactive states [218, 219]. Although it is possible that certain genes key to transient differentiation and development processes exhibit time-varying constitutive transcription, using this assumption to fit thousands of genes is questionable. Therefore, some variant of the bursty model appears more physically founded.

Finally, these concerns, which collectively motivate using a statistically and physically appropriate $P(\mathbf{x}, t)$, lead us to a more fundamental question: *which* t? In other words, we *a priori* know that scRNA-seq datasets are snapshots, and contain cells "earlier" and "later" in the differentiation process. Previous reports reasonably assume that t varies between cells, but do not propose a mechanism to explain how simultaneously collected cells can reside at different times along a process.

The range of these omissions and problems fundamentally speaks to an uneasy compromise between the descriptive and mechanistic worldviews, described in more detail in Section 2.1. Although RNA velocity uses the language of stochastic biophysics, its underlying assumptions, obscured by the user-friendly software and informal, equivocal theory, have a complex and often contradictory relationship with well-attested physical phenomena.

6.1.3 The embedding procedure is unreliable

Even outside the context of RNA velocity, linear as well as nonlinear embeddings distort local and global data relationships or suggest new ones not present in the underlying data (cf. Section 8.4 and [49, 58]). Nonlinear embeddings utilize sensitive hyperparameters that can be tuned, but do not provide well-defined criteria for an "optimal" choice [58, 162]. Tuning algorithm parameters can slightly improve some distortion metrics, though often at the expense of others [162]. In Figure 6.1d, we demonstrate the neighborhood preservation behavior of transformations used to construct low-dimensional embeddings. By the time the data have been summarized a two-dimensional embedding (gold and yellow lines), some 70–80% of the neighborhood relationships have been lost on average, largely in the initial normalization step (red line). In addition to these local distortions, which put into question the computation of directions to embedding neighbors, global distortions can occur. In

Figure 6.1e, we illustrate this point with a Uniform Manifold Approximation and Projection (UMAP) projection of the forebrain dataset, which introduces discontinuities between cell types (cf. Figure 6.1a-b). In other words, if the projection itself erases cell type relationships, RNA velocity cannot recover them.

The procedure for embedding RNA velocity in a two-dimensional space introduces further challenges, which are challenging to deconvolve. Beyond the mismatch in neighborhoods, the directions produced by the kernel-based procedure in the PCA space do not align with directions obtained by simply projecting the cell-specific velocity vectors (distribution of angle deviations shown in Figure 6.1f). Most strikingly, the embedding procedure appears to eliminate nearly all of the quantitative information obtained by the inference and velocity computation procedures: as shown using simulated data in Figure 6.1g, the results obtained using the standard *velocyto* kernel are nearly identical to those obtained using a custom kernel that only uses the signs of the direction and velocity vectors. Finally, the "Markov chains" over cells generated by the procedure are *ad hoc* and not motivated by any particular model of physiology; as discussed in full detail in Section B.4, they implicitly contradict the mechanism used to perform inference.

6.1.4 Conclusions

The standard RNA velocity framework presupposes that the evolution of every gene's transcriptional activity throughout a differentiation transient process can be described by a continuous model. It proceeds to normalize and smooth the data until the rough edges of single-molecule noise are filed off, and fits a continuous model of transcription and turnover assuming Gaussian residuals.

In the process, the stochastic dynamics that predominate in the low-copy number regime, and that characterize nearly all of mammalian transcription, are lost and cannot be recovered. Although parameters can be fit, they are distorted to an unknown extent, due to a combination of data transformation, suboptimal inference, and model misspecification. In *scVelo*, parameters are estimated under a highly restrictive model, yet applied to make broad claims about complex topologies. In *velocyto*, only the sign of the velocity is physically interpretable; if we discard everything else, we still obtain fairly consistent results, suggesting that the method fails to fully utilize valuable quantitative information. Finally, the embedding process, which produces human-interpretable visualizations, is not based on biophysics, and is not guaranteed to be stable or robust.

Nevertheless, sometimes RNA velocity works, and produces results consistent with biological intuition and orthogonal data. From the review above, the performance appears to rely on a combination of factors. First, the "signal" needs to be strong enough that the flaws in the dynamical model can be sufficiently "smoothed out" by the data processing. Second, the cell embedding needs to be faithful enough to recapitulate the features of interest. Third, the velocity embedding procedure needs to produce approximately correct results. These conditions are by no means guaranteed, and fair performance in any particular case may be attributable to hyperparameter tuning and confirmation bias. Therefore, the workflow does not yet appear to be sufficiently reliable to be used for biological discovery.

6.2 Self-consistent snapshot inference

This section unifies portions of [112] by G.G., M.F., T.C., and L.P., as well as [115] by G.G., J.J.V., and L.P. G.G., M.F., and L.P. conceptualized the theoretical alternatives to RNA velocity. G.G. conceptualized, designed, and implemented the case study shown here.

Is there no balm in Gilead? Given the foundational issues we have raised, how can the RNA velocity framework be reformulated to provide meaningful, biophysically interpretable insights? Fortunately, the natural match between stochastic models and UMI-aided molecule counting offers hope for quantitative and interpretable trajectory inference. We propose that discrete Markov modeling can directly and naturally address the fundamental issues. In particular, transient and stationary physiological models can be defined and solved via the approach in Chapter 4, which describes the time evolution of a discrete stochastic process. Since the "noise" is the data of interest, smoothing is not required. Rather, technical and extrinsic noise sources can be treated as stochastic processes in their own right, and explicit modeling of them can improve the understanding of batch and heterogeneity effects. Finally, within this framework, parameters can be inferred using standard and well-developed statistical machinery. Once these parameters are available — and only then — we may optionally summarize the findings in terms of typical low-dimensional visualizations, as with our visualization of the projection of the true μ_M in Figure 6.1h.

The inference of transient dynamics from snapshot data is a formidable problem due to a combination of theoretical and practical factors. Most fundamentally, it is not precisely clear what a snapshot *is*: how does a single measurement simultaneously capture the early and late states in a differentiation process? To develop an explanatory model, we take inspiration from the existing work on cyclostationary processes [66, 67], cell cycle ensemble measurement modeling [28, 220, 282], Markov chain occupation measure theory [167, 225, 320], and chemical reactor engineering [88, 237]. In the typical stochastic modeling context, we fit count data using stationary distributions $P(\mathbf{x})$, obtained as the limit $\lim_{t\to\infty} P(\mathbf{x}, t)$ of a transient distribution. By the ergodic theorem [95], this distribution, when it exists, coincides with the occupation measure $\lim_{T\to\infty} \frac{1}{T}\int_0^T P(\mathbf{x}, t)dt$, i.e., observations drawn from a single trajectory over a sufficiently long time horizon, rather than from multiple trajectories at once. Conveniently, the ergodic limit has time symmetry with respect to measurement: the distribution does not depend on the timing of the experiment. In the transient case, we cannot take these limits. However, we *can* retain time symmetry by proposing that the experiment samples cells at almost

surely finite times t since the beginning of the process. Therefore, we conceptualize data as coming from a set of independent cells, such that each cell's time t_c is sampled from $f(t)$, and counts are drawn from some distribution $P(\mathbf{x}, t_c)$[4]. To fit a set of data, we need to specify and motivate the distribution f.

We illustrate some of the challenges and implications of this framework using the model system shown at the bottom of Figure 6.2a. The underlying transient structure involves transitions through three cell types, each characterized by a particular transcriptional burst size. This model is more realistic but less tractable than the constitutive model implied by RNA velocity. The transient transcription process produces nascent and mature RNA trajectories for each cell; however, we only obtain a single data point per trajectory. Formally, to infer the parameters, we simultaneously need to find Θ, the biological parameters, as well as t_c, all of the cell times. The full data likelihood for a single gene takes the following form:

$$\begin{aligned} \mathcal{L}(\Theta, \{t_c\}; \mathcal{D}) &= \prod_c P(\mathcal{D}_c, t_c; \Theta) \\ \mathcal{L}(\Theta; \mathcal{D}) &= \prod_c \int_0^\infty P(\mathcal{D}_c, t_c; \Theta) f(t_c) dt_c, \end{aligned} \tag{6.3}$$

where we obtain the second line by marginalizing over $\{t_c\}$. If multiple genes are present, but their transcriptional events are not synchronized, P can be decomposed into the product of gene-specific probabilities (Section 10.1). The integral is intractable. Several approaches are available. First, we can reframe the problem as a combinatorial optimization. In this case, we can define an ordering of cell times σ, approximate the continuous distribution f by a uniform-weight discrete distribution placed at N_c quantiles t_c^*, and perform the following combinatorial optimization:

$$\hat{\Theta} = \text{argmax}_\sigma \, \text{argmax}_\Theta \sum_c \log P(\mathcal{D}_c, t_c^*; \Theta). \tag{6.4}$$

As N_c grows, the quantile approximation to f improves. However, the combinatorial optimization becomes rather challenging, as its complexity grows exponentially in N_c. Therefore, a more practical approach may involve the expectation–maximization (EM) algorithm. Such an implementation would iterate between updating the parameter estimate $\hat{\Theta}$ and cell-specific time distributions f_c, defined over a discrete grid [64, 70].

Nevertheless, even this approach requires some careful theoretical work and simulated benchmarking. Even if we have perfect information about the cell times, it is far from clear that we can accurately reconstruct the transcriptional dynamics from

snapshot data (center of Figure 6.2a). This question is essential, as it controls our ability to compute estimates $\hat{\Theta}$ in a given EM step.

In addition, we wish to know whether we can identify the *mechanism* of the snapshot collection. We can imagine cells entering and exiting the observed tissue in multiple ways, which correspond to different choices of $f(t)$. Some natural choices are uniform, which implies the cells stay in the tissue for a deterministic time [168]; decreasing over time, so cells can exit immediately; or uniform, then decreasing, so cells must stay in the tissue for some duration but are free to leave afterward. These choices can be modeled by Dirac, exponential, and Pareto residence distributions. In the parlance of chemical reactor engineering, these configurations are known as the plug flow reactor (PFR), the continuously-stirred tank reactor (CSTR), and the laminar flow reactor (LFR), respectively. Their $f(t)$, which are the reactor internal-age distributions, are well-known in the chemical engineering literature [88, 237], and shown at the top of Figure 6.2a. It is not *a priori* obvious the configurations are mutually distinguishable from count data. If they are not, the choice of $f(t)$ is immaterial for inference.

We generated snapshot data from the PFR model and fit it under all three models. To efficiently evaluate snapshot distributions, we designed an algorithm which essentially "recycles" t_c for trapezoidal quadrature. As shown in Figure 6.2b, despite only having access to a single observation per time point, all models yield results visually close to the true marginals. However, despite these superficial similarities, quantitative model identification is possible. To quantify identifiability, we use the Akaike weight w_ϖ (Equation 3.48), which transforms log-likelihood differences into model probabilities [38]. For example, if all Akaike weights are near $1/3$, the models are indistinguishable; if the correct model's weight is near 1, we can confidently identify the model from the data. For the simulated dataset shown, the true PFR model achieves an Akaike weight of $w_\varpi \approx 79\%$, whereas the CSTR and LFR both achieve $\approx 10\%$. Decreasing the dataset size substantially degrades the identifiability. Even at higher sizes, spread is considerable; for example, a 150-cell dataset gives approximately even odds ($w_\varpi > 1/2$) *on average*, but individual realizations vary from confidently correct ($w_\varpi \approx 1$) to confidently wrong ($w_\varpi \approx 0$).

To understand the robustness of model identifiability, we generated synthetic datasets at random parameter values, constrained to have fairly low expression. We observed poor identifiability, with even or better odds for the correct model in only 20% of the cases (Figure 6.2d). This performance appears to be attributable to quantitative

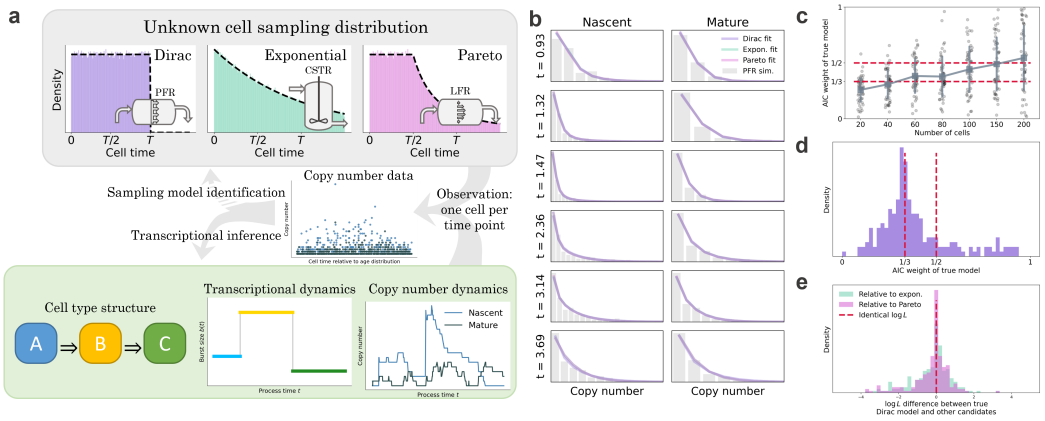

Figure 6.2: The inference of biophysical parameters and reactor configurations from snapshot data.

b. In spite of the considerable differences between the reactor architectures, they produce nearly identical molecular count marginals (histogram: data simulated from the PFR model, 200 cells; colored lines: analytical distributions at the maximum likelihood transcriptional parameter fits for each of the three reactor models. Analytical distributions nearly overlap).

a. A minimal model that accounts for the observation of transient differentiation processes in scRNA-seq: cells enter a "reactor" and receive a signal to begin transitioning from cell type A through B and to C. The change in cell type is accompanied by a step change in the burst size, which leads to variation in the nascent and mature RNA copy numbers over time. Given information about the cell type abundances and the cells' time along the process, we may fit a dynamic process to snapshot data and attempt to identify the underlying reactor type, which determines the probability of observing a cell at a particular time since the beginning of the process.

c. The true reactor model may be identified from molecule count data, but statistical performance is typically poor (points: Akaike weight values for $n = 50$ independent rounds of simulation and inference under a single set of parameters; blue markers and vertical lines: mean and standard deviation at each number of cells; blue line connects markers to summarize the trends; red lines: the Akaike weight values $1/3$, which contains no information for model selection, and $1/2$, which gives even odds for the correct model; two-species data generated from the PFR model; uniform horizontal jitter added).

d. The reactor models are poorly identifiable across a range of parameters, and rarely produce Akaike weights above $1/2$ (histogram: Akaike weight values for $n = 200$ independent rounds of parameter generation, simulation, and inference under the true PFR model; red line: the Akaike weight values $1/3$ and $1/2$; two-species data for 200 cells generated from the PFR model).

e. The challenges in reaction identification arise because all three models produce similar likelihoods (histograms: likelihood differences between candidate models and the true PFR model for $n = 200$ independent rounds of parameter generation, simulation, and inference; red line: no likelihood difference; two-species data for 200 cells generated from the PFR model).

similarities between all three models' likelihoods. As shown in Figure 6.2e, given data of this quality, we cannot even narrow the scope down to two models, as neither of the candidate models performs conspicuously worse than the true PFR configuration. Therefore, it is possible to fit snapshot data approximately equally well using a variety of models; candidates for $f(t)$ are identifiable *in principle*, but challenging to distinguish from any particular dataset. This simulated analysis implies that the details of the reactor configuration may not matter much, providing a basis for omitting this model identification problem for real data.

Chapter 7

MODEL IDENTIFICATION AND SELECTION

A considerable fraction of the variability in single-cell datasets arises from cell-to-cell and time-dependent variation in the transcription rates. These sources of variation control distribution shapes. We seek to apply the models developed in Chapter 4 to obtain insights into this variability and explain distributional differences through mechanisms. By carefully analyzing candidate models, we can characterize the prospects for model selection: for example, if different transcriptional models produce nearly identical distributions, selection is impossible and the choice of model is somewhat arbitrary.

Therefore, the probabilistic analysis of transcriptomic data entails two key challenges: the identification of parameters consistent with the data under a particular model (statistical inference) and the discrimination between distinct hypotheses (model selection). To understand how well we can distinguish different parameter regimes and models, we analyze the models' distributions and compare them using simulated and biological data.

7.1 The role of multimodal data in inference

This section adapts a portion of [115] by G.G., J.J.V., and L.P. The descriptions and mathematical foundations adapt a portion of [113] by G.G.*, J.J.V.*, M.F., and L.P. G.G. and J.J.V. performed the model development and mathematical analysis. G.G. conceptualized, designed, and implemented the case study shown here.
Portions of this case study recapitulate the methods and conclusions of [103] by G.G. and L.P. G.G. and L.P. conceptualized this study. G.G. designed and implemented the analysis.

More interestingly, such analysis can guide the design of experiments: models may be indistinguishable based on some kinds of data, but not others. This perspective has guided the interest in characterizing noise behaviors [204, 205]: distributions provide strictly more information than averages, and allow us to distinguish between regulatory behaviors. Similarly, multivariate distributions provide more information than marginal distributions. Obtaining *different* data (multiple molecular modalities) is qualitatively more useful than obtaining more data (a larger number of cells) or better data (observations less corrupted by noise).

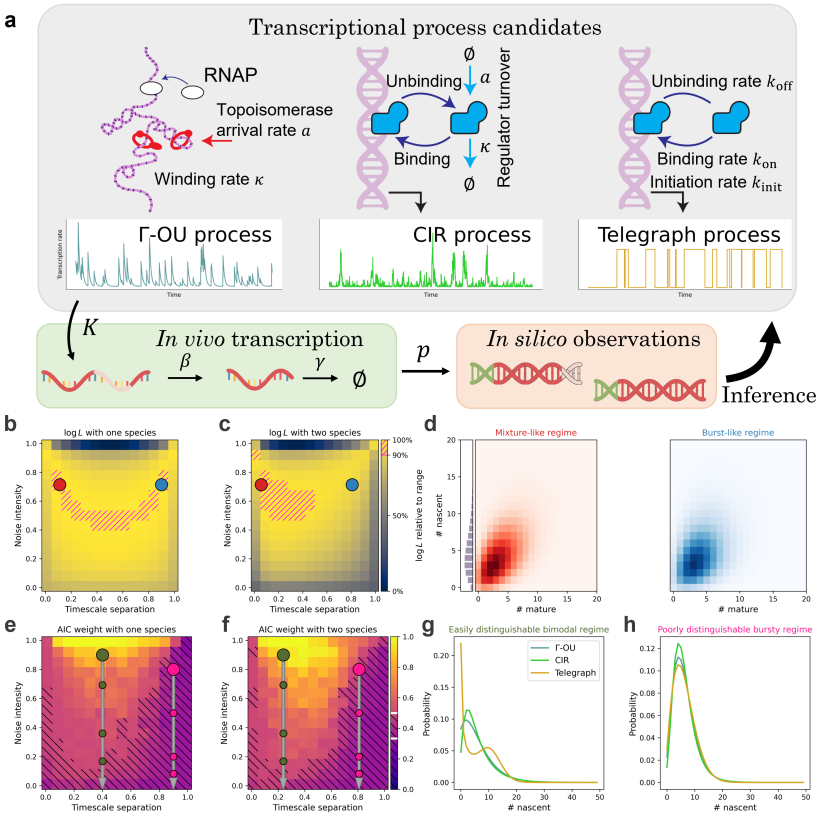

Figure 7.1: The stochastic analysis of biological and technical phenomena facilitates the identification and inference of transcriptional models.

a. A minimal model that accounts for intrinsic (single-molecule), extrinsic (cell-to-cell), and technical (experimental) variability: one of three time-varying transcriptional processes K generates molecules, which are spliced with rate β, degraded with rate γ, and observed with probability p. Given a set of observations, we can use statistics to narrow down the range of consistent models.

b. Overdispersed regimes are not mutually identifiable given a single modality (likelihood computed using nascent RNA data for 200 simulated cells; Γ-OU ground truth; red point: true parameter set in the mixture-like regime; color: log-likelihood of data, yellow is higher, 90th percentile marked with magenta hatching; blue: an illustrative parameter set in a burst-like parameter regime with a similar nascent marginal but drastically different joint structure).

c. The mixture-like and burst-like regimes become mutually identifiable with multimodal data (likelihood computed using bivariate RNA data for 200 simulated cells; all other conventions as in **b**).

d. Nascent marginal and joint distributions at the points indicated in **b** and **c**. Nascent distributions nearly overlap.

e.-f. Given a location in parameter space, models are easier to distinguish using multiple modalities. However, the performance varies widely based on the location in parameter space and the specific candidate models, and decreases with drop-out (Γ-OU Akaike weights under Γ-OU ground truth, average of $n = 50$ replicates using 200 simulated cells; color: Akaike weight of correct model, yellow is higher, regions with weight < 0.5 marked with black hatching; large circles: illustrative parameter sets; smaller circles: distributions obtained by applying $p = 50\%$, 75%, and 85% dropout to illustrative parameter sets while keeping the averages constant).

g. The telegraph model has a well-distinguishable bimodal limit when the process autocorrelation is slower than RNA dynamics, which improves its identifiability (lines: the three candidate models' nascent marginal distributions at the olive point in **e** and **f**).

h. In the bursty limit, the three models look qualitatively similar, limiting identifiability (lines: the three candidate models' nascent marginal distributions at the pink point in **e** and **f**).

We illustrate this key point by considering three transcriptional drivers coupled to a two-stage RNA process ($n = 2$). The transcription rate is time-varying, with rate K. Each biological molecule has a probability p of being observed in the final dataset, which amounts to evaluating the generating functions at pu_N and pu_M. The processes and their physical interpretations are shown in Figure 7.1a.

First, we consider two models that provide an explanatory framework for the extrinsic noise model described in Section 4.6.3. As biological distributions are overdispersed, it appears reasonable to propose that transcription rates K vary between cells, and the gamma model for K produces plausible negative binomial count distributions. However, this approach has some gaps in its physical motivation: what is the biological meaning of this distribution's parameters? And is it really reasonable to assume that the rates are "frozen," and remain as they are for all time in a given cell?

To account for possible time variation, we introduce a class of transcriptional models that balance interpretability and tractability, and generalize the mixture model. Although various biological details underlying transcription may be complicated, we assume they can be captured by an effective transcription rate K_t, which is stochastic and varies with time. This transcription rate randomly fluctuates about its mean value, with the precise nature of its fluctuations dependent upon the fine biophysical details of transcription. To guarantee that they can recapitulate the mixture model, we consider stochastic differential equation K_t with a gamma steady-state distribution.

These models' functional forms have previously been used in biology [89, 140, 228, 242, 325], but a key point has been underexplored: the resulting RNA distributions are *not* generally Poisson-gamma, and directly depend on the details of the transcriptional process dynamics. In other words, the trajectory shapes matter a great deal, and to accurately represent this form of "extrinsic" stochasticity, we need to specify the precise functional form for K_t. Conversely, given a set of candidate models, we may be able to distinguish them on the basis of RNA count distributions. As the stochastic differential equation driver models have identical count expectations and variances, we need to bring to bear the theoretical framework laid out in Chapter 4 to compute full probability mass functions.

The first candidate model is the gamma Ornstein–Uhlenbeck (Γ-OU) process ($N = 1$,

$m = 1$) [241]:

$$dy_t = -\kappa y_t dt + dL_t, \tag{7.1}$$

where the mean-reversion term $-\kappa y_t$ represents DNA winding, which makes RNA polymerase binding less favorable and causes the transcription rate to decrease, whereas the compound Poisson process jump term L_t represents the arrivals of topoisomerases, which increase propensity for transcription by exponentially distributed jumps. In the parlance of Chapter 4,

$$\mathbf{u} = \begin{bmatrix} u_N \\ u_M \\ u_K \end{bmatrix}, \; C^{cc} = -\kappa, \; C^{cd} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \text{ and } \mathcal{A}(\mathbf{u}) = a \left[\frac{1}{1 - \theta u_K} - 1 \right], \tag{7.2}$$

where a is the arrival frequency and θ is the average jump size.

The second candidate model is the Cox–Ingersoll–Ross (CIR) process ($N = 1$, $m = 1$) [62]:

$$dy_t = (a\theta - \kappa y_t)dt + \sqrt{2\kappa\theta y_t} dW_t, \tag{7.3}$$

where W_t denotes the Brownian motion, the drift term a is the production rate of a regulator, κ is its degradation rate, and θ — essentially, a "process gain" — relates the concentration of the regulator to the activity of the promoter. To derive this identity, we assume that the regulator is present at high concentration and rapidly equilibrates with the bound species. This system is characterized by the parameters and operators

$$\mathbf{u} = \begin{bmatrix} u_N \\ u_M \\ u_K \end{bmatrix}, \; C^{cc} = -\kappa, \; C^{cd} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \; Q^c = \kappa\theta, \text{ and } \mathcal{A}(\mathbf{u}) = a\theta u_K. \tag{7.4}$$

The stationary distribution of the Γ-OU and CIR processes is gamma, with shape a/κ and scale θ, i.e., mean $a\theta/\kappa$ and variance $a\theta^2/\kappa$. In addition, their autocorrelation function is $e^{-\kappa t}$.

Finally, the third candidate model is the more conventional telegraph process ($N = 2$, $m = 0$) [219], which describes the Markovian activation and deactivation of a gene. This system is characterized by

$$\mathbf{u} = \begin{bmatrix} u_N \\ u_M \end{bmatrix}, \; H = \begin{bmatrix} -k_{\text{on}} & k_{\text{on}} \\ k_{\text{off}} & -k_{\text{off}} \end{bmatrix}, \text{ and } \mathcal{A}(\mathbf{u}) = \begin{bmatrix} 0 \\ \alpha u_N \end{bmatrix}. \tag{7.5}$$

The stationary distribution of this process is Bernoulli scaled by k_{init}, with mean $\frac{k_{\text{on}}k_{\text{init}}}{k_{\text{on}}+k_{\text{off}}}$ and variance $\frac{k_{\text{on}}k_{\text{off}}k_{\text{init}}^2}{(k_{\text{on}}+k_{\text{off}})^2}$. Its autocorrelation function is $e^{-(k_{\text{on}}+k_{\text{off}})t}$ [95].

Even if the true averages of the transcriptional strength and molecular abundances are fixed, the systems can exhibit a wide variety of distribution shapes and statistical behaviors. This variety can be summarized by a two-dimensional parameter space, ranging over $(0, 1)$. The "timescale separation" governs the relative timescales of the transcriptional and molecular processes; if it is high, the transcriptional process is faster than RNA turnover. The "noise intensity" governs the variability in the transcriptional process: if it is high, the process exhibits substantial variability that translates to overdispersion in the RNA distributions. The bottom edge of this parameter space produces Poisson distributions of RNA, the top left corner produces Poisson mixtures of the law of K, and the top right corner yields bursty dynamics that do not typically have simple analytical solutions [113].

For the two-species SDE driver models, the reduced parameters take the following form:

$$\text{timescale separation } := \mathsf{x} = \frac{\kappa}{\kappa+\beta+\gamma} \text{ and noise intensity } := \mathsf{y} = \frac{\theta}{a+\theta}. \quad (7.6)$$

Equation 7.6 is defined with reference to the process parameters of the Γ-OU and CIR drivers [113]. It remains to define κ, θ, and a in terms of k_{on}, k_{off}, and k_{init} for the telegraph process. The correct identification is:

$$\begin{aligned} \kappa &= k_{\text{on}} + k_{\text{off}} \text{ is the autocorrelation timescale,} \\ a &= \frac{k_{\text{on}}\kappa}{k_{\text{off}}} \text{ is the process scaling, and} \\ \theta &= \frac{k_{\text{off}}k_{\text{init}}}{\kappa} \text{ is the gain.} \end{aligned} \quad (7.7)$$

These definitions are not arbitrary, as they endow the system with lower moments that match the SDE formulation: autocorrelation function $e^{-\kappa t}$, mean $a\theta/\kappa$, and variance $\theta\mu_K$. In addition, the system has the correct geometric burst limit ($k_{\text{init}}, k_{\text{off}} \to \infty$) with burst size $\theta/\kappa \to k_{\text{init}}/k_{\text{off}}$ and burst frequency $a \to k_{\text{on}}$ [233]; this limit matches the Γ-OU one. Therefore, given x, y, μ_K, β, and γ, we can construct the parameters of the underlying transcriptional driver.

Although different x, y regimes reflect very different transcriptional kinetics, they can produce indistinguishable distributions. Figure 7.1b demonstrates the likelihood landscape of a dataset generated from the Γ-OU transcriptional model, evaluated

using the nascent marginal and $p = 1$ (no technical noise). The mixture-like true parameters are indicated by a red point and the top decile of likelihoods is indicated by hatching. The Γ-OU model has a gamma stationary distribution, which produces approximately Poisson-gamma, or negative binomial, RNA marginals in this regime. However, the bursty regime, indicated by a blue point, also yields a negative binomial-like marginal (reported in Equation 7.8), preventing us from identifying the kinetics. On the other hand, if we evaluate likelihoods using the entire two-species dataset, we obtain the landscape in Figure 7.1c: the symmetry is broken, and the parameters can be localized to the mixture-like regime. The source of this improved performance is evident from examining the distributions, shown in Figure 7.1d. The nascent marginals are essentially identical; no amount of purely nascent count data can distinguish between them. However, the bivariate distributions show subtle differences, such as higher nascent/mature correlations in the true regime, which can be used for inference.

In addition, the timescale separation and noise intensity determine the model distinguishability. Figure 7.1e demonstrates the average Akaike weight landscape of datasets generated from the Γ-OU model, computed using the nascent distribution at the same coordinate. We indicate the region $w_{\varpi} > 1/2$ by hatching. As the Akaike weight may be interpreted as a posterior model probability [38], this threshold gives even odds for choosing the correct model, on average. The intermediate regime, indicated by a large olive green point, tends to yield fairly high Akaike weights, translating to good model identifiability. On the other hand, the burst-like regime, indicated by a large pink point, provides considerably less ability to distinguish the models. As expected, the situation improves somewhat when using bivariate data (Figure 7.1f): the Akaike weights increase throughout the parameter space, and the bursty regime data move closer to even odds for model selection. To illustrate the source of the identifiability challenges, we plot the nascent marginals of the models at the two points. In the intermediate regime, the Γ-OU and CIR models yield moderately different distributions, whereas the telegraph model is immediately distinguishable by its bimodality (Figure 7.1g). In contrast, in the bursty regime, the distributions are all unimodal and less identifiable (Figure 7.1h); the Γ-OU and telegraph marginals are particularly similar, as they converge to the same negative binomial limit.

Interestingly, this formulation fully characterizes the effect of certain forms of technical noise. If the transcriptional and observed molecular averages are fixed, but

the experiment fails to capture some molecules, the distributions are identical to those obtained by deflating the transcriptional noise intensity. In other words, even though technical noise affects the molecules, its theoretical effects are indistinguishable from decreasing the variability of the transcriptional process. As the noise levels increase, the RNA distributions are pushed toward the indistinguishable Poisson limit at the bottom edge of the reduced parameter space. We quantify how rapidly the information degrades by plotting smaller circles in Figure 7.1e-f to indicate the effect of 50%, 75%, and 85% dropout, in that order from top to bottom. This result is an extremely general and fundamental consequence of the form of the solution (Section A.8.4).

In sum, in certain overdispersed regimes, candidate drivers *are* mutually distinguishable, and the identification of transcriptional models is qualitatively and quantitatively facilitated by the collection of multimodal data. In addition, by exploiting the mathematical structure of the ODEs defining the transcriptional processes, we find that the impact of the simplest form of drop-out noise can be conceptualized as the reduction of the transcription rate scale, rendering these parameters non-identifiable.

7.2 The identification of transcriptional driving processes

This section adapts a portion of [113] by G.G.*, J.J.V.*, M.F., and L.P. The analysis was conceptualized by J.J.V. and G.G., designed by J.J.V., M.F., and G.G., and implemented by G.G. and M.F.

Even if the Γ-OU and CIR models can be distinguished and fit to data in principle, can they be distinguished and fit *in practice*? Real transcriptomic data feature additional noise due to technical errors, and possibly confounding influences due to phenomena like cell growth and division [283]. One can also face serious model misspecification problems, where one finds that even though one model fits better than others, none of them fit particularly well.

To show that these models may be observed and distinguished in real datasets, we analyzed single-cell transcriptomic data with tens of thousands of genes from the glutamatergic neurons of four mice [321]. Because neurons generally do not grow or divide, their gene expression dynamics should not be confounded by the effects of cell growth and division. To guard against spurious conclusions related to both technical noise and model misspecification, we used a multi-step filtering procedure based on a neuron subtypes from single mouse dataset to choose genes to examine.

As we are primarily interested in demonstrating whether the novel solutions for

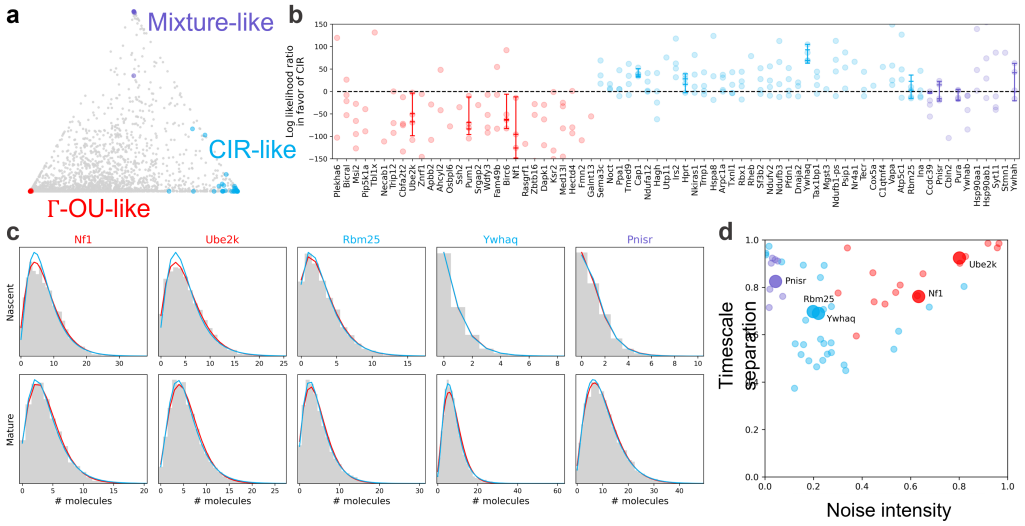

Figure 7.2: Genes from comparable single-cell RNA sequencing datasets can be consistently assigned to a particular biophysical model of transcription.
a. By fitting models in the limiting regimes and calculating model Akaike weights, visualized on a ternary diagram, we can obtain coarse gene model assignments (colors: regimes predicted by the partial fit; red: Γ-OU-like genes; blue: CIR-like genes; violet: mixture-like genes; gray: genes not consistently assigned to a limiting regime).
b. Likelihood ratios for selected genes are consistent across biological replicates, and favor categories consistent with predictions (colors: regimes predicted by the partial fit; points: likelihood ratios; horizontal line markers: Bayes factors; vertical lines: Bayes factor ranges; Bayes factor values beyond the plot bounds have been omitted. $n = 4$ biologically independent animals, with 5,343, 6,604, 5,892, and 4,497 cells per animal).
c. The differences between model best fits are reflected in raw count data (title colors: predicted regimes; lines: model fits at maximum likelihood parameter estimates; line colors: models; histograms: count data).
d. Non-distinguishable genes tend to lie in the slow-reversion and high-gain parameter regime; distinguishable genes vary more, but tend to have relatively high gain (colors: predicted regimes, large dots: genes illustrated in panel **c**. Genes with absolute log-likelihood ratios above 150 have been excluded).

Γ-OU and CIR models can be supported by data, we fit the two models' (distinct) burst-like limits, where $\mathsf{x}, \mathsf{y} \to 1$ and (identical) mixture-like limits, where $\mathsf{x} \to 0$ and $\mathsf{y} \to 1$, using *Monod*, assuming no technical noise, to five glutamatergic neuron subtypes from a single mouse. The burst-like limit of the Γ-OU model is given in Section 4.6.2, the mixture-like limit is given in Section 4.6.3, while the burst-like limit of the CIR model has the following generating function:

$$\log G(\mathbf{u}) = \frac{1}{2} \int_0^\infty \left[1 - \sqrt{1 - 4bU_N(\mathbf{u}, \mathsf{s})} \right] d\mathsf{s}, \tag{7.8}$$

where $b = \theta/\kappa$ and U_N takes the usual form in Equation 4.55 (Section A.8.2). This somewhat degenerate[5] limit describes driving by a process with infinitely many jumps in each finite time interval. Although this driver has been encountered before in the mathematical finance literature [292], the solution does not appear to have been previously reported [20, 22].

We computed the Akaike weights of the three limits for all genes (results for one subtype shown in Figure 7.2a). Finally, we selected genes that most consistently agreed with the distributions in these limits (colored points in Figure 7.2a), and extracted the genes with the best fits to the optimal models.

We fit the Γ-OU and CIR models to the 80 genes that passed the filtering step to glutamatergic neuron data from four mice, using gradient descent to find the maximum likelihood parameter set, and computed the likelihood ratios for the models (Equation 3.45), discarding poorly fit genes. The likelihood ratios for the remaining 73 genes are depicted in Figure 7.2b (points). To ensure that the likelihood ratios we obtained were not distorted by the omission of uncertainty in estimates, or potentially suboptimal fits, we further fit twelve of the genes using a Bayesian procedure, displaying the distribution of Bayes factors (Equation 3.46) in the same axes (horizontal markers).

The predictions from the coarse filter were largely concordant with the results from the full model, suggesting that it is effective for selecting genes of interest from transcriptome-wide data. The model assignments were typically consistent among datasets. Although orthogonal targeted experiments are necessary to identify whether the proposed models effectively recapitulate the live-cell transcriptional dynamics, the reproducibility of the findings suggests directions and candidate genes for such investigations. Finally, the Bayes factors were largely quantitatively consistent with the likelihood ratios, suggesting that the approximations made in the gradient descent procedure do not substantially degrade the quality of the statis-

tical results. However, we did observe several discrepancies between likelihood ratios and Bayes factors, confirming that the more computationally facile gradient descent procedure does not perfectly recapitulate the full Bayesian fit (cf. results for *Ccdc39* and *Birc6*), possibly due to substantial omitted uncertainty in some genes' parameters.

Five example fits are depicted in Figure 7.2c, with the corresponding gene names color-coded according to the best-fit model (red: Γ-OU, blue: CIR, purple: mixture). Model distinctions mostly appear to be due to differences in probability near distribution peaks. Interestingly, only either the nascent marginal or mature marginal exhibits obvious visual differences between model fits in some of the genes depicted here, further motivating the use of multimodal data.

The location of each best-fit parameter set in the qualitative regimes space is shown in Figure 7.2d. Most Γ-OU fits exist in the top right corner, suggesting we are effectively fitting a standard geometric burst model in these cases. Nonetheless, there are a number of genes for which the parameter sets reside somewhere in the center, indicating that the full complexity of the Γ-OU or CIR models is necessary to describe the corresponding data.

Despite the models' simplicity, the results suggest that single-cell RNA sequencing data may be sufficiently rich to enable Bayesian model discrimination between superficially similar regulatory schema. Certain genes demonstrate reproducible differences between the two considered SDE drivers, which may imply differences in the underlying regulatory motifs. Interpreting the specific biochemical meaning of the findings is challenging without accounting for features which have been omitted in the discussion thus far, such as technical noise and additional complexities in downstream processing of RNA. Nevertheless, fine details of transcription — including DNA mechanics and gene regulation — appear to have signatures in single-cell data, and a model-based, hypothesis-driven paradigm can help identify them. Further, these fine details can be probed using a range of tools, some more approximate and suited to genome-wide exploratory analysis, others more statistically rigorous and suited to detailed study of gene targets.

7.3 RNA processing

This section adapts a portion of [114] by G.G., S.Y., and L.P. The theoretical results and data analyses were conceptualized, designed, and implemented by G.G.

Although we have largely considered Markovian processes, this model class may

not be sufficient to describe biology. On one hand, there is considerable evidence that bacterial and mammalian transcription is typically Markovian and bursty [65, 92, 170, 239, 244], although more sophisticated extensions have been proposed and applied [125, 170, 334]. Similarly, the results of genome-wide inhibition experiments are largely consistent with Markovian RNA degradation, with exponentially decaying RNA levels over time [255].

On the other hand, the kinetics of splicing and export are rather less well-characterized. For computational and mathematical tractability, we typically assume that splicing is a single-step Markovian process, whereas export is rapid enough to neglect. These assumptions appear sufficient to fit data generated by single-cell RNA sequencing, but they have not been studied in detail. Indeed, they are counterintuitive: nuclear retention *is* important [16], and previous studies have used fairly sophisticated models to describe it [86], although others have achieved reasonable results under a Markovian hypothesis [16, 127, 206]. In addition, we have elided a considerable amount of complexity by identifying nascent RNA with intron-containing molecules and mature RNA with all others (Section B.1). Indeed, we may reasonably expect that splicing often [60] occurs co-transcriptionally [74], such that intron-containing molecules are in the process of elongation. In this case, a Markovian model is *a priori* incorrect, because elongation needs to complete before the molecule can be released and degraded [59].

These assumptions are testable using single-cell and single-nucleus RNA sequencing (snRNA-seq) datasets and the suite of models in Chapter 4. Conversely, testing them allows us to understand the qualitative differences, if any, between these data types. The splicing dynamics are relevant for all datasets, but the transport dynamics are motivated by the narrower goal of interpreting single-nucleus datasets. In single-cell datasets, we can, in principle, suppose that nuclear transport dynamics are rapid and the degradation is approximately Markovian. However, single-nucleus technologies isolate individual nuclei (Figure 7.3a), so we need to explicitly model the transport term to construct a stochastic model. First, we seek to understand whether single-nucleus data *require* models that are substantially different from single-cell data, whether the Markovian efflux hypothesis is sufficient to describe the removal of mature molecules from nuclei. Second, we seek to characterize whether the Markovian splicing hypothesis holds. Finally, we seek to understand the extent to which single-cell and single-nucleus data allow for model identification.

We consider the standard bursty model discussed in Section 4.6.2, as well as the

two models shown in Figure 7.3b, which replace one or the other of the downstream Markovian processes with a deterministic one. By applying the methods in Section 4.3.2, we find that the process with delayed efflux has the log-generating function

$$\log G(\mathbf{u}) = \frac{k/\beta}{1 - bu_M} \log\left(\frac{bU(\tau) - 1}{bu_N - 1}\right) + k\tau \frac{bu_M}{1 - bu_M} - \frac{k}{\beta} \log\left(1 - bU(\tau)\right), \text{ where}$$
$$U(t) := u_M + (u_N - u_M)e^{-\beta t}. \tag{7.9}$$

This fairly complex expression produces the expected negative binomial nascent marginal.

On the other hand, the process with delayed splicing has the log-generating function

$$\log G(\mathbf{u}) = \frac{k\tau bu_N}{1 - bu_N} - \frac{k}{\gamma} \log(1 - bu_M). \tag{7.10}$$

This generating function is separable with respect to u_N and u_M, which implies that the nascent and mature distributions are statistically independent. The mature count distribution is negative binomial, whereas the nascent one is geometric-Poisson or Pólya–Aeppli. This result generalizes the single-species case reported in [151].

We fit the three models using *Monod*, omitting technical noise. The models were separately fit to GABAergic and glutamatergic cell types from two mouse brain samples [321], one generated using single-cell sequencing and one generated using single-nucleus sequencing, as well as single-cell and single-nucleus data from pericentral, periportal, and interzonal hepatocytes from a single human liver sample [11]. Upon fitting the MLEs, we computed the models' likelihood ratios relative to the Markovian model.

We did not observe a systematic bias toward the delayed efflux model in either the whole-cell (Figure 7.3c) or the nuclear data (Figure 7.3d). The log-likelihood ratios were symmetric and near zero in all considered cases, suggesting the data were insufficient to strongly favor either model in any of the datasets. This result suggests that the Markovian model is broadly reasonable for nuclear transport.

In contrast, the delayed splicing model was considerably less favored, with log-likelihood ratios tending to be negative (Figure 7.3e-f). The strength of evidence against the models was lower in the single-nucleus data. This loss of statistical identifiability concords with intuition: the bursty Markovian and delayed-efflux models afford identical negative binomial nascent RNA marginals, requiring a large amount of mature RNA to distinguish the models, which the nuclear sequencing

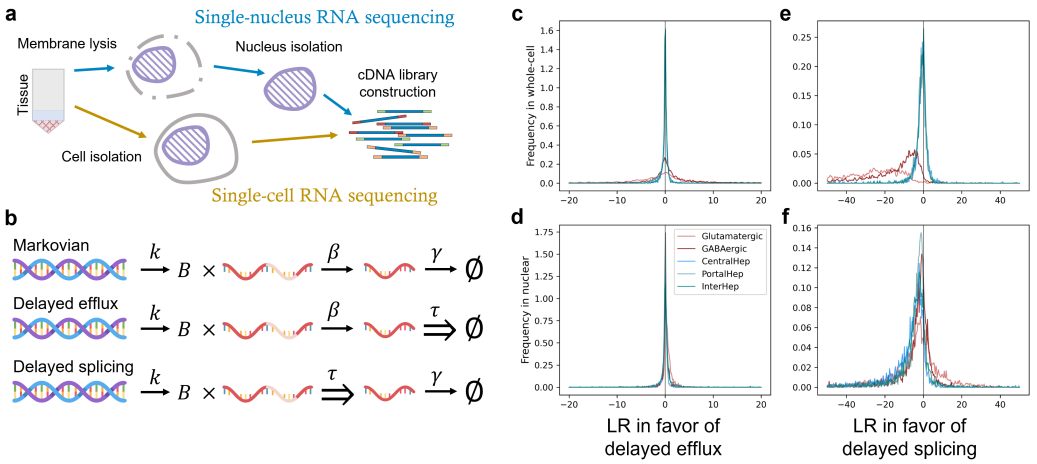

Figure 7.3: The comparison of stochastic model predictions facilitates the identification of RNA processing mechanisms compatible with data.
a. An outline of the experimental differences between single-cell and single-nucleus sequencing technologies.
b. The reaction schema of the considered models: DNA generates nascent RNA with transcriptional burst frequency k and burst size b, the nascent RNA are converted to mature RNA; the mature RNA are removed from the system, either by nuclear transport or by cytoplasmic degradation.
c. In whole-cell data, likelihood ratios do not systematically favor either the Markovian or the deterministically delayed efflux model (colors: cell types; red: Allen data; blue-green: Andrews data; lines: kernel density estimates).
d. In nuclear data, likelihood ratios do not systematically favor either the Markovian or the deterministically delayed efflux model (conventions as in **c**).
e. In whole-cell data, likelihood ratios typically favor the Markovian model over the deterministically delayed splicing model (conventions as in **c**).
f. In nuclear data, likelihood ratios typically favor the Markovian model over the deterministically delayed splicing model (conventions as in **c**).

protocol lacks. The delayed splicing model has a different nascent marginal, and appeared to be more distinguishable from the negative binomial.

Overall, the hypothesis of Markovian, one-step splicing is useful despite its simplicity and apparent conflict with the understanding of transcriptional elongation. The simplest converse assumption — that splicing has a deterministic waiting time — produces substantially worse fits across a variety of data. On the other hand, the nucleus transport dynamics are consistent with either model, and we can say very little about them because the single-nucleus data have low mature RNA content. Nevertheless, the Markovian model appears to suffice.

Chapter 8

SEQUENCING MODEL SPECIFICATION

In spite of our assumptions throughout Chapter 7, real datasets do have various technical noise sources, which need to be accounted for. Unfortunately, even the relatively simple treatment of this topic in Section 4.4 is not tractable on a genome-wide level: a fully satisfactory explanation of the multifaceted variability in single-cell datasets remains out of reach. Nevertheless, we can fruitfully attempt to treat the phenomena one at a time, assuming all others have been satisfactorily accounted for, and use biophysical hypotheses to characterize the technical noise behaviors. This incremental approach allows us to develop a more grounded alternative to typical *ad hoc* approaches to "denoising" sequencing data.

8.1 Empty droplets

This section adapts a portion of [115] by G.G., J.J.V., and L.P. G.G. conceptualized, designed, and implemented the analysis.

One of the first steps in scRNA-seq data analysis is cell quality control, which excludes cell barcodes that appear to originate from empty droplets from further analysis [187]. For computational tractability, this procedure typically uses "hard" assignment, such that barcodes associated with a total molecule count above some threshold are treated as cells, whereas barcodes below the threshold are treated as empty droplets. Threshold selection is necessary because even "empty" droplets contain ambient RNA. This ambient RNA appears to originate from cells lysed in the preparation process, and contaminates empty and cell-containing droplets alike [187]. The observation of ambient counts has led to the development of statistical methods for removing this source of noise, either by estimating and subtracting it [323] or incorporating it into a stochastic model [87, 256, 322]. Conceptually, Equation 4.38 reflects the latter approach: each droplet contains one or more cells, each with biological generating function G, and background, with a generating function G_{bg} that depends on G. To accurately model the background counts, we need to propose and justify a specific functional form for G_{bg}. Thus, under the assumption that empty and cell-containing droplets are similarly susceptible to contamination, the former provide a reasonable estimate of ambient distributions in the latter [323].

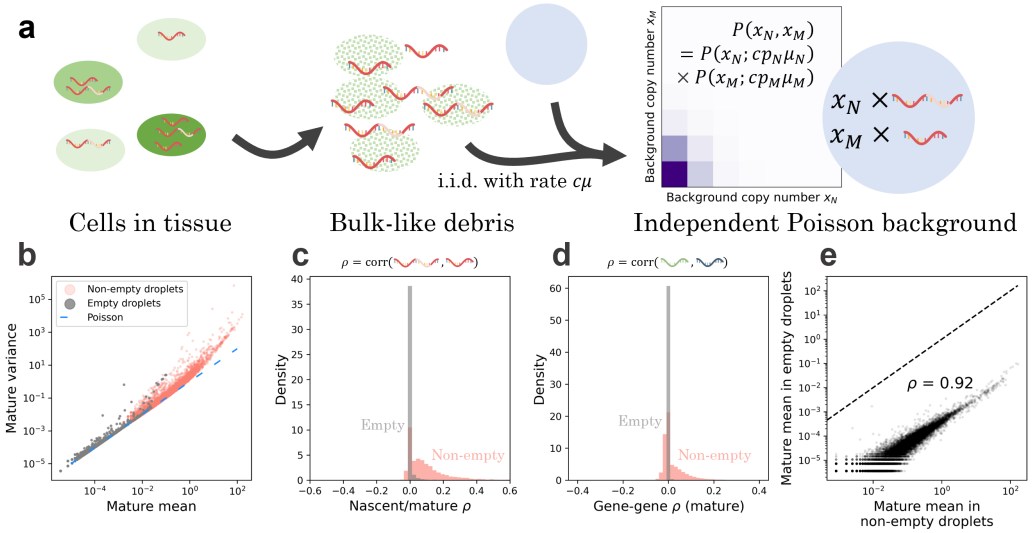

Figure 8.1: The pseudo-bulk model of background noise is quantitatively consistent with counts from a human blood cell dataset.

a. The simplest explanatory model for background noise invokes the lysis of cells (green), which creates a pool of RNA that reflects the overall transcriptome composition but retains none of the cell-level information. If the loose RNA molecules diffuse into droplets (blue) according to a memoryless and independent arrival process, the resulting background distribution (purple: higher probability mass; white: lower probability mass) observed in empty droplets should be a series of mutually independent Poisson distributions, with the mean controlled by the composition in non-empty droplets.

b. The mature transcriptome in empty droplets has a mean-variance relationship near identity (gray points, $n = 12,298$), consistent with Poisson statistics (blue line); the non-empty droplets demonstrate considerable overdispersion (red points, $n = 17,393$).

c. The mature and nascent transcripts in empty droplets have sample correlation coefficients ρ near zero, consistent with distributional independence (gray histogram, $n = 9,362$); the non-empty droplets demonstrate nontrivial statistical relationships (red histogram, $n = 14,365$).

d. The mature transcripts of different genes in empty droplets have sample correlation coefficients ρ near zero, consistent with distributional independence (gray histogram, $n = 75,614,253$); the non-empty droplets demonstrate nontrivial statistical relationships (red histogram, $n = 151,249,528$).

e. When both are nonzero, the mature count mean in empty droplets is highly correlated with the mean in the non-empty droplets, consistent with the pseudo-bulk interpretation (black points, $n = 12,107$; dashed line: identity).

The simplest model holds G_{bg} to be equivalent to a "pseudobulk" experiment, with molecules randomly sampled from the lysed cell population. If each cell is equally likely to contribute to the pool of free RNA, and diffusion occurs by a simple independent arrival process, we find that the distribution of background should be Poisson, with the mean for each species proportional to its mean in original cell population, as in, e.g., [87]. This functional form immediately induces a set of testable predictions: not only are the distributions Poisson, but they are *independent* Poisson, with no meaningful statistical structure remaining between transcripts of a single gene, as well as between different genes, as illustrated in Fig. 8.1a.

To characterize the accuracy of these predictions, we inspected datasets pseudoaligned with *kallisto* | *bustools* [197], and compared the data for barcodes passing *bustools* quality control to data for barcodes which were filtered out. As a shorthand, we call the former "non-empty" and the latter "empty" droplets, keeping in mind that this identification is approximate. We illustrate the results for a human blood dataset generated by 10x Genomics. As shown in Figure 8.1b, data from non-empty droplets are substantially overdispersed relative to Poisson, whereas data from empty droplets are largely consistent with the Poisson identity mean–variance relationship. However, a small number of relatively high-expression genes demonstrate overdispersion. In addition, intra-gene (Figure 8.1c) and inter-gene (Figure 8.1d) correlations are typically nontrivial in non-empty droplets, but consistently near zero for empty droplets, supporting distributional independence of the background counts. Finally, the mean expression in empty droplets is highly correlated with mean expression in non-empty droplets, albeit lowered by approximately four orders of magnitude (Figure 8.1e), supporting the assumption that the original cells are lysed in a uniform fashion.

To characterize the deviations from the pseudo-bulk model, we identified the genes that demonstrated overdispersion in empty droplets. A considerable fraction of these genes were associated with mitochondria or blood cells. For example, of the 21 annotated genes overdispersed in the empty droplets of a 10x Genomics mouse neuron dataset, nine were mitochondrial (*mt-Nd1*, *mt-Nd2*, *mt-Co1*, *mt-Co2*, *mt-Atp6*, *mt-Co3*, *mt-Nd3*, *mt-Nd4*, and *mt-Cytb*), three coded for hemoglobin subunits (*Hba-a1*, *Hba-a2*, and *Hbb-bs*), and two coded for blood cell-specific proteins (*Bsg*, *Vwf*) [189, 209]. On the other hand, of the 10 annotated genes overdispersed in the empty droplets of dataset generated from cultured mouse embryonic stem cells [72], six (*mt-Nd1*, *mt-Co2*, *mt-Atp6*, *mt-Co3*, *mt-Nd4*, *mt-Cytb*) were mitochondrial

and none were blood cell-specific [209].

Since overdispersion implies that contamination involves non-independent arrivals of these molecules, the results suggest that the cell-free debris contain, among other structures, entire mitochondria or erythrocytes, when they are present in the source tissue. These membrane-bound structures may diffuse into droplets, then lyse and release all of their contents at once. In other words, empty droplets do not merely have disproportionally high mitochondrial content, as has been noted previously [87, 133, 139]; they have *nontrivially distributed* mitochondrial content, which can suggest the mechanism of its incorporation and improve interpretation where simple thresholds may be misleading [139]. We speculate that cases where the model fails can be leveraged to discover more complicated forms of contamination, such as molecular aggregates [322].

In addition, we examined the total UMI counts in empty droplets, which should be Poisson (Fano = 1) if each individual gene's distribution is Poisson. For the human blood dataset demonstrated in Figure 8.1, the empty droplets had fairly significant overdispersion (Fano $\approx$ 43), which decreased, but did not disappear (Fano $\approx$ 7.6), once the 53 significantly overdispersed genes were excluded. This result suggests that, although the pseudo-bulk model is approximately valid, some residual variance, possibly due to variability in per-droplet capture rates, is present and needs to be modeled to fully describe the stochasticity in single-cell datasets.

8.2 Length biases

This section summarizes the content of [107] by G.G. and L.P. The initial interest in length bias was due to L.P. and V.S.; the model was conceptualized by L.P. and G.G.; the model was designed and implemented by G.G.

In a wide variety of 10x single-cell RNA sequencing datasets, average spliced mRNA counts do not seem to show a length dependence (Fig. 8.2a, gray lines), which is consistent with previous studies of UMI-based protocols [222]. On the other hand, unspliced mRNA counts strongly correlate with gene length [119] (Fig. 8.2a, red lines). This observation prompted us to investigate whether the discrepancy has biological origins, and raised questions about the consequences of ignoring this bias.

This bias may be explained by several models (Figure 8.2b). The first has identical, gene-specific observation probabilities p for nascent and mature species. In this model, the inferred burst size is bp, as these two parameters are not mutually

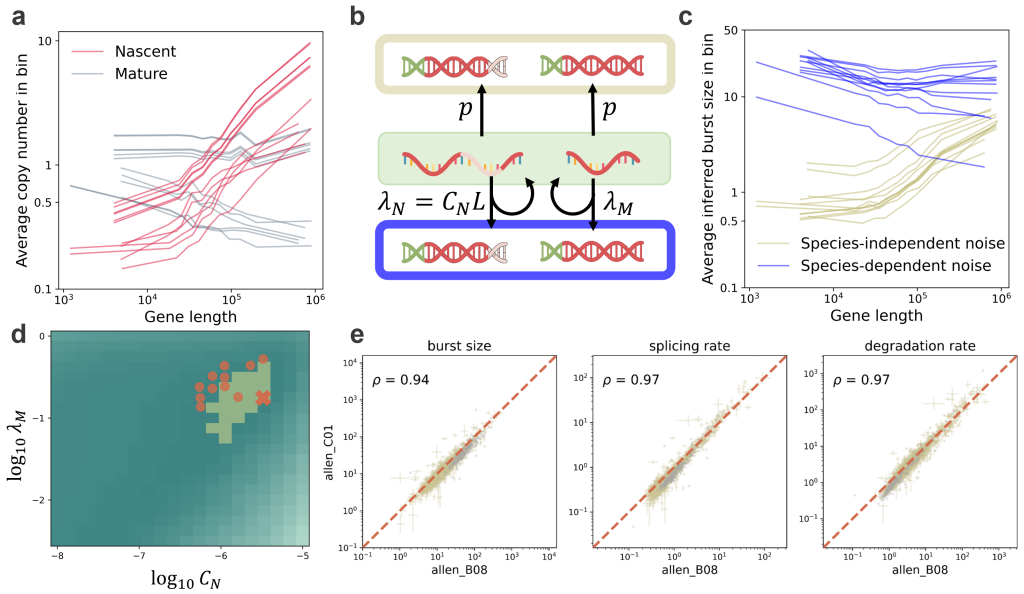

Figure 8.2: Trends in inferred transcriptional parameters allow us to distinguish between models of technical noise, and explain a pervasive length bias in molecule counts by length-dependent sequencing rates.

a. A variety of single-cell datasets produce consistent and counterintuitive length-dependent trends in nascent RNA observations (lines: average per-species gene expression, binned by gene length; red: nascent RNA observations; gray: mature RNA statistics; data for 2,500 genes shown for each dataset).

b. Two explanatory models for the trend in **a**: the species-independent bias model for length dependence in averages, which proposes nascent and mature RNA are sampled with equal probabilities, and the species-dependent bias model, which proposes nascent RNA sampling rate scales with length (top, gold: kinetics of species-independent model; bottom, blue: kinetics of species-dependent model; center, green: the source RNA molecules used to template cDNA).

c. Fits to the species-independent model show a strong positive gene length dependence for inferred burst sizes, whereas fits to the species-dependent model show a modest negative gene length dependence, which is more coherent with orthogonal data (lines: average per-gene burst size inferred by *Monod*, binned by gene length; gold: results for species-independent model; blue: results for species-dependent model; only genes that passed goodness-of-fit testing shown)

d. The likelihood over sampling parameters can be optimized to infer the parameters, which are consistent among datasets (dark teal: lower, light teal: higher total Kullback-Leibler divergence between fit and blood cell data; highlighted yellow region: 5% quantile region for the displayed landscape; orange cross: optimal sampling parameter fit for the displayed landscape; orange points: optimal sampling parameter fits for other analyzed v3 datasets).

e. Biological replicates show largely concordant inferred parameter values (orange dashed line: identity; gold: lower bounds on 99% confidence intervals; gray: fits rejected by statistical testing; splicing and degradation rates are reported in units of burst frequency).

identifiable. The second has with a gene length L-dependent technical noise term for the nascent species, coarsely representing a higher rate of priming for long molecules with abundant intronic poly(A) tracts [119, 168, 207], and a shared genome-wide term for the mature species, representing priming at the poly(A) tail (Section 4.4.3). In this model, the inferred burst size is b. Both models produce fair fits to the data.

However, the trends in the parameters inferred by *Monod* (Section 5.4) under the two models are strikingly different: the species-independent bias model predicts that longer genes have higher bp (Figure 8.2c, gold lines). Ascribing this trend to the b term — longer genes have higher burst sizes — contradicts burst size trends from fluorescence microscopy [172]. Ascribing it to the p term — longer genes have higher sampling probabilities — is physically unrealistic, because mature RNA molecules are depleted of the internal poly(A) tracts necessary for priming [217].

On the other hand, the species-dependent model predicts a modest negative relationship between length and burst size, which is more coherent with orthogonal data (Figure 8.2c, blue lines). We observe similar trends for the turnover parameters β and γ: striking length dependence under the species-independent model, which vanishes when using length as a scaling factor for the sampling rate[6]. We find that the sampling parameter optima are similar for a wide variety of datasets obtained from comparable experiments (Figure 8.2d). In addition, biological replicates [321] produce similar parameter values (Figure 8.2e).

This technical noise model is a relatively simplistic low-order approximation — all genes have the same mature molecule capture rate λ_M and length scaling C_N. Nevertheless, it foregrounds a key modeling principle: in the absence of prior information, biological parameters need to be fit on a gene-by-gene basis, but technical noise should be constructed using a common genome-wide model that varies in a mechanistic, rather than arbitrary way. In sum, the mathematics enable us to define and fit systems, but to understand whether the fits are sensible, we need to contextualize and compare them with previous results and physical intuition.

8.3 Technology differences

This section summarizes part of the content of [106] and [107] G.G. and L.P. The analysis was conceptualized and designed by G.G. and L.P, and implemented by G.G.

The explicit parametrization of biophysical models allows us to explain and account

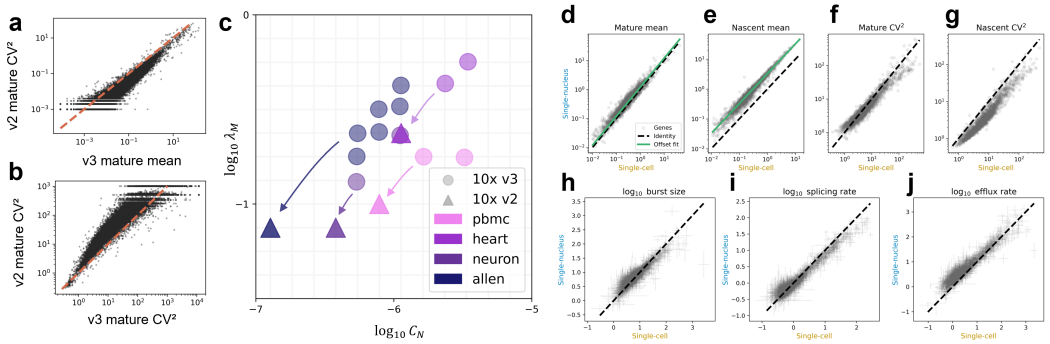

Figure 8.3: The technical noise model fits can be interpreted to analyze experimental effects.

a. 10x v2 and v3 scRNA-seq replicates generated from a single sample demonstrate discordant RNA count distributions: the v2 datasets have lower mean values (orange dashed line: identity; black: genes).

b. The v2 datasets have higher CV^2 values (conventions as in **a**).

c. The v2 datasets' distributional differences can be tentatively explained by a combination of identical biological parameters and lower technical noise parameters (C_N: coefficient for length-dependent unspliced capture rate; λ_M: spliced capture rate; colors: dataset categories; intersections of grid lines indicate the sampling parameter sets evaluated in the inference process).

d. Counterintuitively, representative paired mouse brain single-cell and single-nucleus datasets exhibit similar mature RNA levels (gray points: genes; dashed black line: line of identity; green line: the approximate average offset observed for single-nucleus data).

e. The single-nucleus dataset consistently has considerably higher nascent RNA counts, which suggests the presence of a technical effect between the two technologies (conventions as in **d**).

f. The single-nucleus dataset demonstrates slightly lower noise levels for mature count data (gray points: genes; dashed black line: line of identity).

g. The single-nucleus dataset demonstrates considerably lower noise levels for nascent count data (conventions as in **f**).

h.-i. By fitting mechanistic models to both datasets, we can identify technical noise parameters that produce consistent burst and splicing parameters between the technologies (points: maximum likelihood estimates for burst sizes and splicing rates; error bars: conditional 99% confidence intervals for inferred parameters; dashed black line: line of identity).

j. At the discovered technical noise parameters, the mature RNA efflux or turnover is considerably higher for the single-nucleus dataset, consistent with this parameter's interpretation as the rapid export from the nucleus (conventions as in **h**-**i**).

for differences between technologies. By themselves, technical noise parameters demonstrate limited identifiability. However, we can investigate the technical effects more systematically by treating replicates generated by different sequencing technologies and adopting stronger priors. We found that count data generated by the higher-efficiency v3 chemistry consistently yielded higher mean levels and lower noise (CV^2) levels than those generated by the older v2 chemistry (Fig. 8.3a-b). Intuitively, these differences should be appropriately attributed to technical effects, as the source tissues were similar or identical.

Imposing the belief that the underlying biological parameters should be the identical between all technical replicates, and treating the results for large v3 samples as a putative ground truth, we identified the set of sampling parameters for the v2 datasets that produced the best agreement to these biological parameter values. The resulting inferred sampling parameter optima are shown in Figure 8.3c: as expected, v2 datasets have lower sampling parameter values. These values are somewhat challenging to identify without enforcing the consistency criterion between transcriptional parameters: the v2 KLD landscapes are more susceptible to noise than the v3 KLD landscapes, preventing *de novo* inference. Although the current comparison is mostly relative, the framework provides a quantitative explanatory mechanism for the technical effect of sequencing chemistry.

Similarly, we can apply this approach to the analysis of single-nucleus data. The interest in single-nucleus sequencing, as well as the recognition of systematic differences in the findings from the two technologies [11, 19, 71, 286], has motivated the analysis of these differences [46] and the development of more or less *ad hoc* data integration methods [11, 169]. At least some discrepancies appears to stem from a fundamental methodological difference: single-cell analyses typically only use exonic reads, whereas single-nucleus combine intronic and exonic reads [71, 75].

We propose that scRNA-seq and snRNA-seq data may be more analyzed in a more principled way through a mechanistic lens. For single-nucleus data, we use the results in Section 7.3 to justify the bursty model (Section 4.6.2). For single-cell data, we adopt the same model, making the usual assumption that export is sufficiently rapid enough relative to degradation. Under this pair of models, the nascent RNA dynamics — i.e., transcription and splicing — should be identical for the two technologies, as the nascent RNA are confined to the nucleus.

This axiom provides a foundation for the joint analysis of the technologies. For example, Figure 8.3d-e compares the average counts for 2,000 genes in scRNA-

seq and snRNA-seq datasets generated from a single mouse brain tissue sample by 10x Genomics. Surprisingly, in spite of the depletion of cytoplasmic RNA, the mature count averages were visually similar, whereas the nuclear count averages were approximately half an order of magnitude higher in the single-nucleus dataset. Quantitatively, 83% of the mature and over 99% of the nascent averages were higher in the snRNA-seq sample. To explain this difference, we adopt the usual "marker gene" paradigm, i.e., that closely related cell types typically differ in the expression of a small number of genes [187], whereas the other genes have similar distributions. Under this assumption, we are immediately led to conclude that the difference is purely technical, and cannot be attributed to enrichment of certain cell types in one or the other technology. In other words, due to the details of the nuclear sequencing protocol, the procedure retains considerably more RNA of both types. This assumption appears to be supported by Figure 8.3f-g: both species exhibited an overall decrease in the noise levels (66% of the mature and 98% of the nascent CV^2 values), which is consistent with decreased molecule loss. The difference in mature RNA amounts should, then, be explained by the combination of two competing effects: the depletion of cytoplasmic RNA, as well as more effective capture of remaining molecules, in the single-nucleus protocol.

To quantify the efflux rates, we fit the datasets using *Monod* and inferred the technical noise parameters for the single-cell dataset. Next, we identified the set of single-nucleus technical noise parameters that provided the best match to the burst size and splicing rate parameters (Figure 8.3h-i); the discovered set of technical noise parameters had higher (more effective) sampling rates. The inferred efflux rates at this set were considerably higher for the single-nucleus dataset, both visually (Figure 8.3j) and statistically: the t-test $\{t, p\}$ values were $\{-2.7, 7.3 \times 10^{-3}\}$ for the burst size, $\{1.6, 0.11\}$ for the splicing rate, and $\{-11, 2.1 \times 10^{-27}\}$ for the efflux rate.

The procedure we have outlined has significant limitations: for example, we have neglected nuclear efflux in the single-cell data and cell type heterogeneity, both of which are physiologically important [17, 321] likely contributors to deviations in Figure 8.3h-i. In addition, single-nucleus sequencing may harbor as of yet poorly-understood technical noise phenomena particular to the technology. Nevertheless, the model formulation provides a foundation for the incorporation of more sophisticated nuclear retention phenomena [85] jointly with technical noise. In addition, the strategy provides a principled solution to the dilemma of incorporating intronic reads: all the available data should be used, with species differences encoded in a

multivariate mechanistic model. If its assumptions are explicitly formulated, the model can be fit, or extended to account for violations, based on experimental data.

8.4 Limitations of normalization procedures

This section summarizes part of the content of [106] by G.G. and L.P. The control was conceptualized and designed by L.P.; the derivation was performed by G.G. The method was implemented by G.G.

The modeling framework provides an appealing and self-consistent alternative to typical methods for the treatment of technical variability. Due to the scale of scRNA-seq data, standard analyses heavily use data transformation and dimensionality reduction to produce a version of the data more amenable to statistics [187]. For example, a typical analysis of cell type heterogeneity may apply size normalization (e.g, proportional fitting or PF, which treats RNA counts as compositional quantities [34]), log-transformation, principal component analysis (PCA), and Uniform Manifold Approximation and Projection (UMAP) [187, 195]. Each of these steps has a specific purpose; for the four steps above, the purposes are, in turn, to remove variability due to technical heterogeneity, to obtain easily tractable normal-like log-abundance distributions, to select the latent data dimensions that contain the most variability, and to visualize the cell type structure [187]. These transformations rely on implicit assumptions about the structure of the data; these assumptions may be mutually contradictory, and their violation may produce results that range from suboptimal to catastrophically incorrect.

These limitations and failure modes have previously been investigated. Size normalization privileges relative, rather than absolute RNA species abundance; occasionally, this approach produces inconsistent results across the genome [123] and retains apparently technical variation [34, 56]. Log-transformation is optimal for homogeneous, high-expression, approximately negative binomial data [4, 34, 187], and relies on an arbitrary genome-wide "pseudocount" hyperparameter that can distort the distributions [4, 34, 123, 290]. PCA is optimal for multivariate normal data, and can be misled by the large zero fractions observed in single-cell data [290]. Finally, UMAP appears to be optimal for data with uniform, low-noise coverage of a latent manifold, with risk of distortions due to violated assumptions and stochastic initialization [49, 58] (Section 6.1.3). A comprehensive treatment of the distortions induced or ameliorated by each step appears, however, to be out of reach, as the transformations' results are heavily data-dependent and elude theoretical analysis.

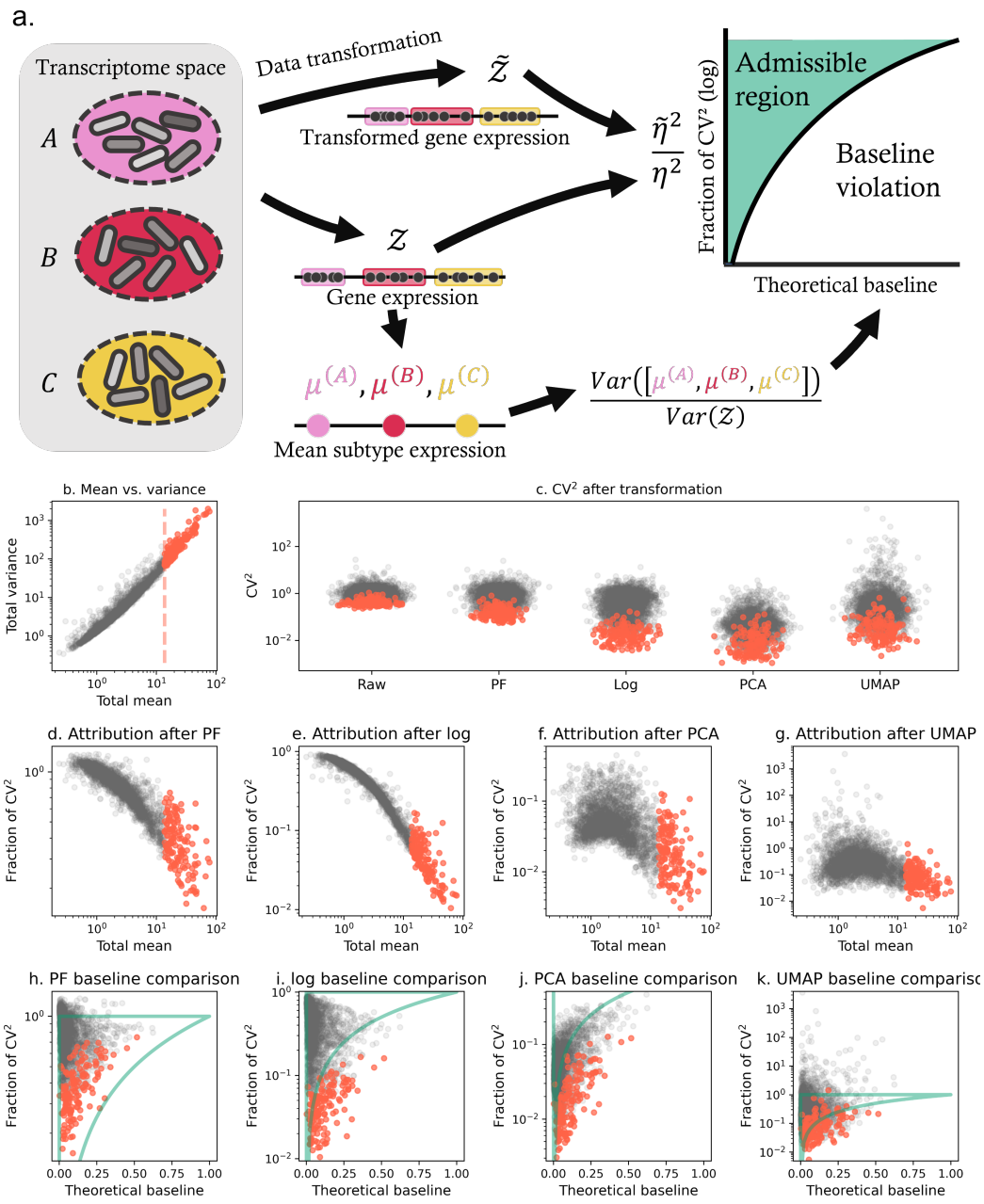

Figure 8.4: Normalization and dimensionality reduction distort and underestimate biological variation, especially in high-expression genes.

a. A proposed baseline for the analysis of residual variation after data transformation: the fraction of biological variability can be bounded by a theoretical baseline, which is computed from the variation in average subpopulation expression. If this baseline is violated, the data transformation has discarded some biophysically meaningful variation.

b. High-expression genes have high variance (gray points: genes below the 95th percentile by mature RNA expression; red points: genes above the 95th percentile by mean mature RNA expression, red line: percentile threshold).

c. Proportional fitting size normalization (PF), log-transformation (log), and principal component analysis (PCA) globally deflate the squared coefficient of variation (CV^2), whereas Uniform Manifold Approximation and Projection (UMAP) globally inflates it (gray and red points: as in **b**).

d.-g. All four of the steps substantially deflate high-expression genes' CV^2 relative to raw data, implicitly attributing their variability to nuisance technical effects (gray and red points: as in **b**).

h.-k. The deflation of variability results in the violation of the theoretical lower bound computed from cell subpopulation differences, particularly for high-expression genes (gray and red points: as in **b**; curved teal line: identity baseline, below which biological variability is removed; horizontal teal line: threshold, above which variability is inflated relative to raw data).

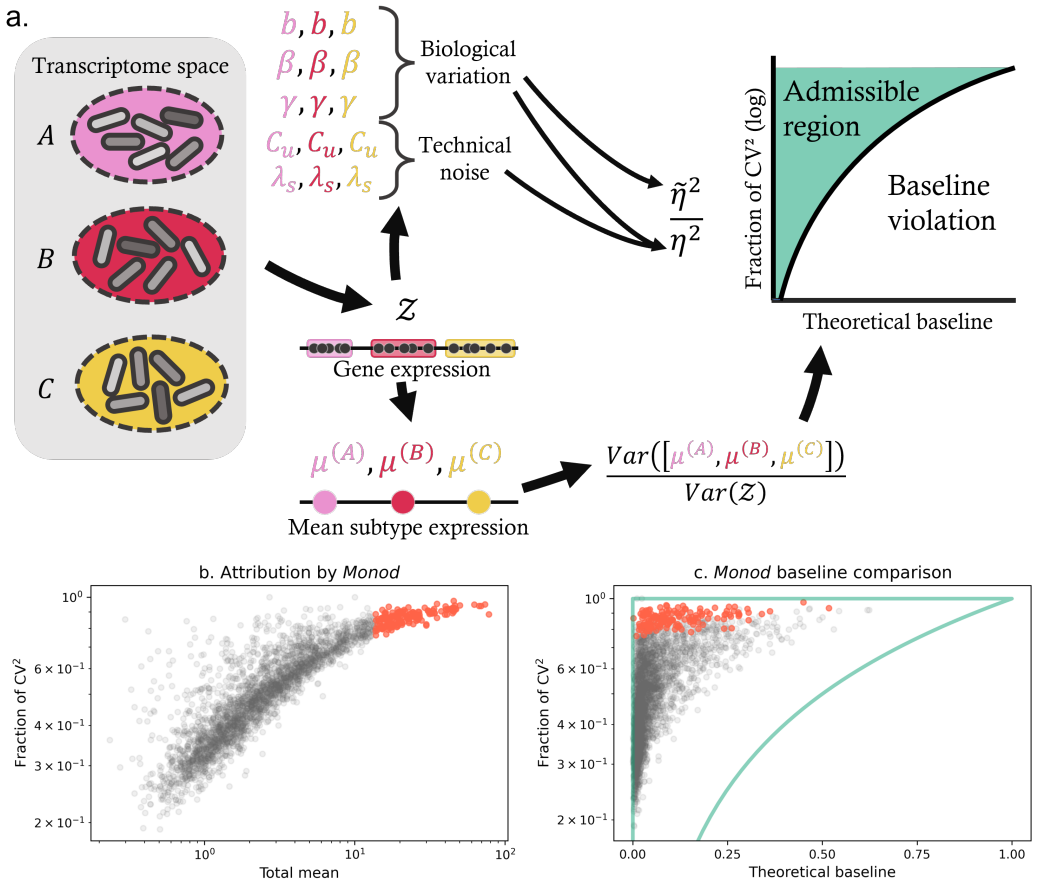

Figure 8.5: The *Monod* mechanistic analysis of biological and technical variability produces coherent results.

a. The baseline introduced in Figure 8.4a may be compared to point estimates of the biological variability fractions, which follow immediately from a fit to a parametric model of transcription and sequencing.

b. The *Monod* fits explicitly attribute the variability in high-expression genes to biological phenomena (gray and red points: as in Figure 8.4b).

c. The *Monod* results lie entirely within the admissible region (gray and red points: as in **b**; curved teal line: identity baseline, below which inferred biological variability is lower than inter-cell population variability; horizontal teal line: threshold, above which inferred biological variability exceeds that of raw data).

In Figure 8.4a, we propose a procedure for the quantitative benchmarking of data transformations relative to an internal baseline. Each step transforms the data distribution, purportedly retaining relevant biological variability — such as cell type differences — and removing incidental or technical variability, quantified by the squared coefficient of variation (CV^2). Therefore, by removing some fraction of variability, a data transformation implies this component is immaterial to analysis, whereas the residual fraction of variation — the CV^2 ratio for the distribution after and prior to transformation, denoted by $\tilde{\eta}^2/\eta^2$ — is attributed to biology. For dimensionality reduction techniques, this procedure involves some subtleties, and requires the existence of an inverse transformation that can map the lower-dimensional representation back to the high-dimensional space.

This residual fraction should not vary arbitrarily; under mild assumptions, we can bound the biological fraction of CV^2 from below by the variability in cell subpopulation averages. Specifically, we can write down the following identities for biological lower moments:

$$\begin{aligned} \tilde{\mu} &= \mathbb{E}_{\boldsymbol{\pi}}[\tilde{\mu}_\kappa] = \sum_\kappa \pi_\kappa \tilde{\mu}_\kappa \\ \tilde{\sigma}^2 &= \sum_\kappa \pi_\kappa \tilde{\sigma}_\kappa^2 + \sum_\kappa \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2, \end{aligned} \tag{8.1}$$

where $\boldsymbol{\pi}$ consists of the cell subpopulation proportions. If we assume sequencing samples evenly from the subpopulation, and rescales the mean and variance by scalars ξ_κ and Ξ_κ, we obtain the observed moments

$$\begin{aligned} \mu &= \sum_\kappa \pi_\kappa \mu_\kappa = \sum_\kappa \pi_\kappa \xi_\kappa \tilde{\mu}_\kappa \\ \sigma^2 &= \sum_\kappa \pi_\kappa \sigma_\kappa^2 + \sum_\kappa \pi_\kappa (\mu_\kappa - \mu)^2 \\ &= \sum_\kappa \pi_\kappa \Xi_\kappa \tilde{\sigma}_\kappa^2 + \sum_\kappa \pi_\kappa (\xi_\kappa \tilde{\mu}_\kappa - \mu)^2. \end{aligned} \tag{8.2}$$

We further assume that $\xi_\kappa = \xi$ for all κ. In other words, we suppose that, for a particular gene and on average, all cell types are chemically and statistically identical with respect to the sequencing process. We find that the lower moments of the observed distributions can be rewritten in terms of the lower moments of the

biological distributions:

$$\begin{aligned}\mu &= \xi \sum_{\kappa} \pi_\kappa \tilde{\mu}_\kappa = \xi \tilde{\mu} \\ \sigma^2 &= \sum_{\kappa} \pi_\kappa \Xi_\kappa \tilde{\sigma}_\kappa^2 + \sum_{\kappa} \pi_\kappa (\xi \tilde{\mu}_\kappa - \xi \tilde{\mu})^2 \\ &= \sum_{\kappa} \pi_\kappa \Xi_\kappa \tilde{\sigma}_\kappa^2 + \xi^2 \sum_{\kappa} \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2.\end{aligned} \tag{8.3}$$

Using the definition of the squared coefficient of variation ($\mathrm{CV}^2 = \sigma^2/\mu^2 := \eta^2$), we find that the fraction of CV^2 due to biology can be bounded from below:

$$\begin{aligned}\frac{\tilde{\eta}^2}{\eta^2} = \frac{\mu^2}{\tilde{\mu}^2}\frac{\tilde{\sigma}^2}{\sigma^2} = \frac{\xi^2\tilde{\mu}^2}{\tilde{\mu}^2}\frac{\tilde{\sigma}^2}{\sigma^2} = \xi^2\frac{\tilde{\sigma}^2}{\sigma^2} &= \frac{\xi^2 \sum_\kappa \pi_\kappa \tilde{\sigma}_\kappa^2 + \xi^2 \sum_\kappa \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2}{\sum_\kappa \pi_\kappa \Xi_\kappa \tilde{\sigma}_\kappa^2 + \xi^2 \sum_\kappa \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2} \\ &\geq \frac{\xi^2 \sum_\kappa \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2}{\sum_\kappa \pi_\kappa \Xi_\kappa \tilde{\sigma}_\kappa^2 + \xi^2 \sum_\kappa \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2},\end{aligned} \tag{8.4}$$

i.e., the fraction of biological variability is at least as high as the fraction of variability attributable to the inter-population mean differences. The bound affords the consistent estimator

$$f_{\text{baseline}} = \frac{1}{S^2} \sum_{\kappa} \frac{(\mathsf{N_c})_\kappa}{\mathsf{N_c}} \left((\overline{X}_M)_\kappa - \overline{X}_M \right)^2, \tag{8.5}$$

where S^2 is the sample variance over all $\mathsf{N_c}$ cells, $(\overline{X}_M)_\kappa$ is the average expression in cell subpopulation κ, and $(\mathsf{N_c})_\kappa$ is the number of cells in that subpopulation.

To compare the results of the transformation procedures to this baseline, we analyzed a mouse glutamatergic neuron dataset [321], using pre-annotated subtypes to produce a lower bound. We considered several thousand genes, emphasizing the top 5% by dataset-wide average; these high-variability genes are typically of most interest in single-cell analyses (Figure 8.4b). The iterative application of transformations up to PCA typically deflated the gene-specific CV^2 values, particularly for the high-expression genes and in the log-transformation step. However, the application of UMAP inflated CV^2 throughout. We found that the high-expression genes' variability was typically deflated relative to the raw data, suggesting that the data transformations attribute overdispersion to nuisance technical effects (Figure 8.4d-g).

Log-transformation, PCA, and UMAP violated the baseline computed from inter-subtype variation, particularly for the high-expression genes. In addition, a considerable fraction of genes demonstrated variability exceeding that of the original

data after PF and UMAP. After the final step, more than a third of the genes in the dataset had, at some point in the analysis, gone below the lower bound. This result suggests that ubiquitous transformations efface meaningful biological signal. UMAP attempts to recover it by inflating cell type differences; however, since this inflation is genome-wide, it does not restore the quantitative information lost in previous steps, and may generate false findings.

We propose that a mechanistic approach provides a more reliable avenue for the analysis of sequencing data. In this worldview, all assumptions about the noise behaviors are explicit rather than implicit; count data are not to be denoised, but fit to a first-principles model that includes biological and technical noise terms. Once a satisfactory parametric fit is available, the fractions of biological and technical variability follow immediately (Section 4.6.5, using the bursty model in Section 4.6.2). This approach is outlined schematically in Figure 8.5a: given annotations, we can separately fit cell subtypes, obtain their biophysical parameters, and aggregate them to obtain the fraction of biological variability. The details of the calculation amount to applying the following definitions:

$$\begin{aligned} \tilde{\eta}^2 &= \frac{\mathbb{E}_{\boldsymbol{\pi}}[\tilde{\sigma}_\kappa^2] + \mathrm{Var}_{\boldsymbol{\pi}}(\tilde{\mu}_\kappa)}{\mathbb{E}_{\boldsymbol{\pi}}[\tilde{\mu}_\kappa]^2}, \\ \eta^2 &= \frac{\mathbb{E}_{\boldsymbol{\pi}}[\sigma_\kappa^2] + \mathrm{Var}_{\boldsymbol{\pi}}(\mu_\kappa)}{\mathbb{E}_{\boldsymbol{\pi}}[\mu_\kappa]^2}, \end{aligned} \tag{8.6}$$

inserting the plug-in estimates of the subtype-specific means and variances with (Table 4.2) and without (Table 4.1) technical noise. The fit, implemented in *Monod*, attributes overdispersion in high-expression genes to biological variability (Figure 8.5b), in striking contrast to the non-parametric transformations. As a consequence, the inferred fraction of biological variability coheres with the baseline (Figure 8.5c).

Interestingly, this agreement is not merely a consequence of independently fitting cell subtypes and aggregating the variance. We used *Monod* to fit the entire glutamatergic dataset, introducing some error due to the neglect of subtype heterogeneity. Quantitatively, this approach simply compares the coefficients of variation implied by Tables 4.1 and 4.2, without using $\boldsymbol{\pi}$. We obtain similar results, with a single violation of the bound. This control suggests that the mechanistic procedure largely explains biological variability by transcriptional bursting, rather than subtype differences.

Chapter 9

DETERMINATION OF BIOLOGICAL DIFFERENCES

The theoretical (Section 8.2) and numerical (Section 5.4) tools we have introduced provide a relatively simple framework for the determination of biophysical and technical parameters consistent with data under a particular set of hypotheses about the biophysics of transcription and chemistry of sequencing. Although the parameters are often challenging to identify precisely, we have shown technical differences can be satisfactorily accounted for by assuming that that biological differences between paired samples are minimal (Section 8.3).

A probabilistic understanding of technical noise is mandatory, and omitting it leads to the serious problems outlined in Sections 8.4 and B.3. Yet it is, in many important ways, secondary: we perform experiments to learn something about the biology of living cells, rather than nuisance technical effects. To that end, the promise of the mechanistic worldview consists of providing a biophysical, rather than phenomenological, alternative to analyses of biological heterogeneity. Once we have accounted for technical effects, we can ascribe differences in expression to specific mechanisms of regulation. These insights are necessarily incomplete: we may be able to find that certain biophysical parameters have changed, but cannot determine *how*. Nevertheless, this approach is promising in light of orthogonal work studying the relationship between regulatory mechanisms and affected parameters [210]. In addition, the biophysical approach provides an avenue to probe subtle differences in copy number distributions that would not be identifiable using standard statistical methods [72].

9.1 The role of multimodal data in differential expression

This section summarizes principles originally introduced in [107] and elaborated upon in [106] by G.G. and L.P. The approach was conceptualized by G.G.

The collection of multimodal data, such as nascent and mature RNA counts, provides qualitatively different and more actionable information than the quantification of a single modality, and deserves particular attention. In Section 7.1, we found that bivariate data improve the identifiability of models and distinct parameter regimes. In the context of differential expression testing, the availability of multimodal data allows us to distinguish regulation trends that would otherwise be ambiguous.

To see why, we can compare the single-species and two-species cases. The single-species model of bursty transcription

$$\varnothing \xrightarrow{k} B \times \mathcal{X} \xrightarrow{\gamma} \varnothing \tag{9.1}$$

yields a negative binomial distribution with shape k/γ and scale b. Therefore, we can identify these two parameters and no more based on steady-state measurements, regardless of the amount of data. If we observe a difference in b between conditions, we can confidently attribute it to a change in the burst size. However, if we observe a difference in k/γ, we cannot uniquely attribute it to the turnover rate γ or the burst frequency k. We can *assume* that turnover is less likely to be modulated than the transcription rate [201], but this assumption is impossible to justify based on the data alone.

If we fit the two-species model in Section 4.6.2, we obtain estimates of k/β, k/γ, and b. Superficially, these estimates suffer from the same problem: we cannot uniquely identify k, β, and γ. However, this scenario is fundamentally different: it is implausible that β and γ are simultaneously modulated, because these processes occur in distinct compartments. This assumption is quite a bit milder than excluding the modulation of turnover altogether. Therefore, if we observe *synchronized* changes in k/β and k/γ, with the same sign and similar magnitudes, we may treat them as evidence for the modulation of the burst frequency.

Indeed, differences in inferred normalized splicing and degradation rates demonstrate striking and pervasive correlations (third panel of Figure 9.1a), suggesting that they should appropriately be attributed to burst frequency modulation. Certain genes lie off the diagonal, suggesting that turnover modulation has a role to play in certain narrow cases. Therefore, if the approximate equality $\Delta \log_{10} \frac{\beta}{k} \approx \Delta \log_{10} \frac{\gamma}{k}$ holds, we generally assume that $\Delta \log_{10} k$ has a similar magnitude, but the opposite sign. We average the two to estimate the burst frequency modulation:

$$\Delta \log_{10} k \approx -\frac{1}{2} \left(\Delta \log_{10} \frac{\beta}{k} + \Delta \log_{10} \frac{\gamma}{k} \right) \tag{9.2}$$

9.2 Mechanistic differential expression

This section summarizes a portion of [107] and [106] by G.G. and L.P. The analysis was conceptualized by G.G. and L.P. and designed and implemented by G.G.

With this physical and statistical machinery in hand, we strive to generalize the identification of expression differences and ascribe them to specific mechanisms.

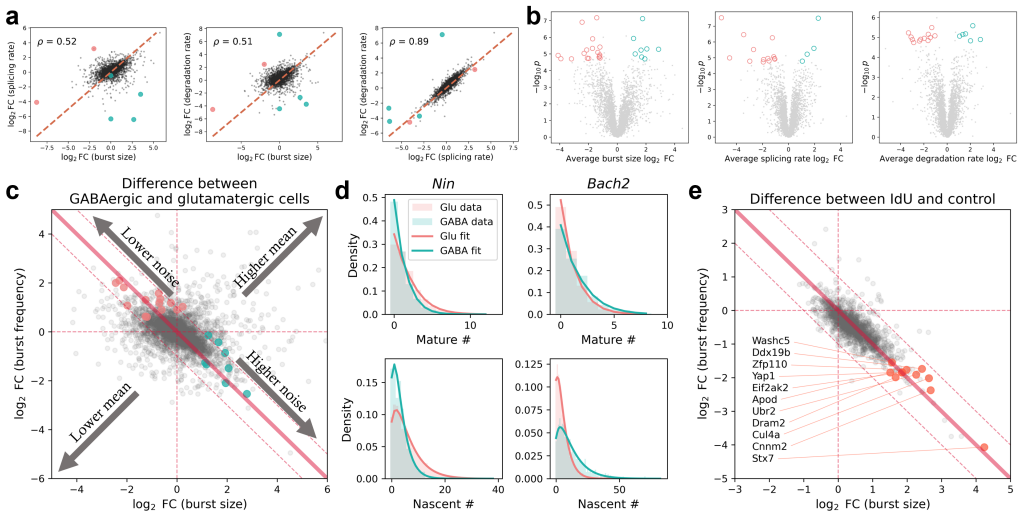

Figure 9.1: The *Monod* mechanistic framework generalizes differential expression testing to the identification of genes with distributional differences, without requiring substantial changes in average expression.

a. Mouse neuron cell types show strong co-variation in normalized splicing and degradation rate differences, suggesting potential burst frequency modulation (orange dashed line: identity; black: genes retained after statistical testing; red: known glutamatergic markers; light teal: known GABAergic markers).

b. Differential expression analysis identifies genes that exhibit consistent inter-cell type parameter modulation in neuron populations (gray: parameters for genes not identified as differentially expressed by the *t*-test and a fold change (FC) criterion; light red: parameters identified as higher in the glutamatergic cell type; light teal: parameters identified as higher in the GABAergic cell type).

c. The differences between mouse glutamatergic and GABAergic cell types, computed from four independent replicates, include genes with substantial noise enhancement but little to no change in average expression, which may reflect biophysically important compensation mechanisms (light red points: genes with significantly higher noise in glutamatergic cells; light teal points: genes with significantly higher noise in GABAergic cells; gray points: all other genes; solid diagonal line: parameter combinations where burst size and frequency differences compensate to maintain a constant average expression; dashed diagonal lines: ± 1 $\log_2$ expression fold change region about the constant-average expression line; vertical and horizontal lines: parameter combinations where burst size and frequency, respectively, do not change).

d. Differences in inferred noise behaviors reflect differences in distribution shapes (light red: glutamatergic cell type; light teal: GABAergic cell type; histograms: raw counts; lines: *Monod* fits; top row: mature RNA marginal; bottom row: nascent RNA marginal).

e. Perturbation by IdU, which triggers DNA damage and repair, rarely changes expression levels, but induces genome-wide noise enhancement [40] detectable by *Monod* (lines and gray points: as in **c**; red points and labels: well-fit, moderate-expression genes identified as highly noise-enhanced).

In typical transcriptomics workflows, the determination of differences between cell types or conditions often reduces to the determination of differentially expressed (DE) genes, which exhibit statistically significant differences in their average copy numbers. However, the identification of DE genes requires careful accounting for technical covariates [187]. In addition, the data may exhibit *compensating* mechanistic effects that change the distribution while keeping the averages constant, which would not be identifiable by standard statistical methods [69, 72, 205].

We propose that differential expression testing should be generalized to the identification of modulated parameters. We use the notation "DE-Θ" to denote criteria using Θ — which may be a data moment or an inferred biophysical parameter — as a test statistic. The mechanistic DE approach is reminiscent of negative binomial regression methods previously proposed for scRNA-seq data [9, 123, 186], but has key distinctions: the model is not assumed to be closed-form or univariate, and the differences are explained in terms of specific regulatory mechanisms rather than descriptive parameters.

If we do not have any replicates — for example, if we seek to compare the differences between cell types in a single tissue — we can essentially use outlier calling procedures on the distributions of parameter differences (the marginals of panels of Figure 9.1a) to propose DE-Θ genes. This approach complements, rather than substitutes typical statistical procedures: in the demonstrated example, many well-known marker genes for the GABAergic and glutamatergic cell types (red: glutamatergic; light teal: GABAergic) exhibit such low expression outside their characteristic cell types that their parameters cannot be accurately identified, and the differences are uninterpretable.

If we *do* have have independent biological replicates, we can use standard statistical procedures, such as the t-test. We illustrate DE-b, DE-β/k, and DE-γ/k genes with consistent parameter differences between GABAergic and glutamatergic cell types, computed from four mouse neuron datasets, in Figure 9.1b. Per Section 9.1, these differences can be particularly naturally summarized in terms of burst size and frequency differences (Figure 9.1c), in the spirit of [65, 172].

Most interestingly, we found several genes that consistently exhibited transcriptional parameter modulation but exhibited approximately constant mean mature RNA expression between cell types, and would not be identifiable by standard statistical procedures (colored points in Figure 9.1c). The discovered genes are indicated according to the cell type differences' effect on noise: genes highlighted in red exhibit

more overdispersion in the glutamatergic population, whereas genes highlighted in light teal exhibit more overdispersion in the GABAergic population. These genes exhibit only minor differences in average expression, and fall fairly close to the line of expression identity (solid diagonal line in Figure 9.1c), where an increase in burst size is precisely compensated by a decrease in the burst frequency.

The differences in parameters are reflected in the data distributions and the model fits. For example, *Nin* and *Bach2* visually exhibit higher noise in the glutamatergic and GABAergic populations, respectively (Figure 9.1d). The mature count averages are, on the other hand, fairly close (*Nin* Glu: 1.7, GABA: 0.98; *Bach2* Glu: 0.87, GABA: 1.4).

Many of the genes identified by this procedure were associated with neuronal structure and development. *Socs2*, *Igf1r*, *Itga4*, and *Dpysl3* are involved in differentiation and neurite outgrowth [152, 164, 250, 284]. *Bach2* and *Cxxc4* induce feedback in neuronal development, apparently to maintain differentiated status in neurons [93, 258]. *Mid2* and *Nin* are associated with neural development regulation through microtubule organization [18, 273]. *Egln1* is linked to neuronal apoptosis [174]. *Fam174a* is involved in lipid metabolism and membrane structure [142]. *Rnf152* and *Rgmb* are broadly implicated in neural development [55, 212, 243]. *Scg3* appears to have a functional role in secretory granule biogenesis [177].

Some identified genes have less clear mechanistic connections to brain structure and function. *Ankrd40* is uncharacterized and is not known to have neural functions [82], but the similar gene product *Ankrd6* has an obscure neurodevelopmental role [288]. *Stx4a* is localized on synaptic membranes [6]. *Slc39a11* is a zinc transporter, involved in brain function [68]. *Mblac2* codes for an obscure protein that may have enzymatic activity [190]. *Ccdc136* appears to have a DNA regulatory role [194], but may be involved in neural speech pathology [3]. The role of *Crtc3* in the rodent brain appears to be restricted to stress response [240]. *Il34* is a microglial marker; microglia have immune and regulatory functions in the brain [15].

Although these distinctions are statistically identifiable, the import and basis of cell type differences in distribution rather than average expression is as of yet obscure. These differences appear to be associated with compensatory mechanisms and motivate further study of the role of noise in biophysical systems.

9.3 Genome-wide noise modulation

This section summarizes a portion of [106] by G.G. and L.P. The analysis was conceptualized, designed, and implemented by G.G.

Such compensatory mechanisms, where substantial distributional changes are associated with only minor average expression changes, have long been explored using theoretical tools [205]. These theoretical studies have come to fruition: recent studies have found that the introduction of the modified nucleotide 5-iodo-2′-deoxyuridine (IdU) to a culture medium enhances transcriptional noise, but keeps average expression constant, hinting at a genome-wide mechanism for compensation [40, 72].

The mechanistic approach also enables the summary of such far-reaching perturbations, which move beyond the usual marker gene paradigm. Although the model required to fully recapitulate the dynamics of DNA damage repair involved in this process is sophisticated, we found that we could characterize the effects of IdU using a simple bursty model. We fit the nascent and mature data from control and IdU datasets using *Monod*. As in Section 8.3, technical noise parameters were not readily identifiable from the 10x v2 sequencing data. We assumed the parameters were in a region we previously discovered for this technology (Figure 8.3c), and analyzed biophysical parameters under that assumption.

We found that the IdU-perturbed cell culture exhibited striking noise amplification, with very limited differences in mean expression (Figure 9.1e). This result strongly contrasts, e.g., Figure 9.1c, which shows fairly symmetric noise amplification and reduction between cell types. The asymmetry in the findings are consistent with the authors' conclusions and orthogonal validation, which likewise found that burst size increases and burst frequency decreases in the IdU condition [40].

We selected a set of well-fit genes that exhibited particularly high modulation and had average expression greater than 1 in at least one of the conditions for further analysis, identifying *Stx7*, *Washc5*, *Apod*, *Eif2ak2*, *Ubr2*, *Cnnm2*, *Dram2*, *Zfp110*, *Cul4a*, *Ddx19b*, and *Yap1* (red points in Figure 9.1c). Interestingly, two of these genes are directly related to the DNA damage activity of IdU: *Dram2* is involved in the autophagic response to DNA damage repair, whereas *Cul4a* is involved in the turnover of DNA repair proteins. Several other genes more generally mediate the cellular stress response: *Zfp110*, *Eif2ak2*, and *Yap1* regulate apoptosis, whereas *Ddx19b* may be active in stress granules. The role of the remaining genes is obscure: *Stx7* and *Washc5* are related to vesicular function, *Apod* is involved in lipid

metabolism, *Ubr2* controls ubiquitination, and *Cnnm2* appears to be involved in ion transport [209].

We were able to partially compare our results for *Sox2*, *Nanog*, and *Mtpap*, whose transcriptional parameters were computed from fluorescence data in [40, 72]. We did not observe *Sox2* expression in either dataset. *Nanog* was rejected by our goodness-of-fit procedure. This is, in principle, consistent with the results in Table S2 of [72], which report gene on fractions near 30-55%; this regime violates the assumptions of the bursty model (gene on fraction tending to zero). The inferred signs for *Mtpap* parameter modulation agreed with Figure S4 of [40], although we obtained rather different magnitudes ($\log_2$ fold changes of ≈ -0.3 by smFISH vs. ≈ -1.5 by *Monod* for burst frequency; ≈ 2 by smFISH vs. ≈ 1.3 by *Monod* for burst size). Therefore, although the genome-wide trends broadly recapitulate the mechanistic explanations provided by the authors, and some of the high-noise genes appear to be implicated in DNA repair and stress, the quantitative comparison of fluorescence and sequencing data requires further analytical work.

In sum, the mechanistic DE framework offers several avenues for the identification of biophysical mechanisms. Substantial differences in expression can be quantitatively explained by the effects of transcriptional burst size and frequency modulation. In addition, differences in *distributions* can be explained by simultaneous, and counteracting, modulation of both parameters, revealing complexity that would be lost by a simple consideration of the averages and suggesting directions for further experiments.

Chapter 10

MODELING MULTI-GENE SYSTEMS

10.1 Key goals and context

This section summarizes a portion of [115] by G.G., J.J.V. and L.P. The analysis was conceptualized, designed, and implemented by G.G. The models proposed originate from [112] by G.G., M.F., T.C., and L.P., [105] by G.G. and L.P., [44] by M.C.*, G.G.*, Y.C., T.C., and L.P., [113] by G.G.*, J.J.V.*, M.F., and L.P., and unpublished research undertaken by C.F. and G.G.

In Chapters 8 and 9, we have shown that fitting fairly simple, two-species models can provide some insight into the biophysics of transcription and the chemistry of single-cell sequencing. However, throughout the process, we have essentially focused on statistically homogeneous populations, treating cells as independent and identically distributed draws from a common distribution and treating the genes as independent. As discussed in Section 5.4, this approach omits all gene–gene relationships by design, and breaks with standard analyses, which use these relationships to characterize dataset structure [187].

For a multitude of reasons, we cannot build a comprehensive model using the tools in Chapter 4. To do so, we would need to explicitly represent regulation, which is excluded from these models (as discussed in Section 4.3.1). In addition, regulation typically proceeds through protein signaling cascades; as we do not have protein data, a regulation model would be woefully underdetermined for a sequencing assay. The construction, parametrization, and inference of such models falls under the purview of the whole-cell modeling and systems biology fields [260, 269, 303]. We anticipate that a full synthesis of systems biology, bioinformatics, and stochastic biophysics, if not altogether futile, will require some decades of interdisciplinary work.

Nevertheless, certain kinds of co-regulation *can* be represented in this framework. To leverage the master equation models outlined in Chapter 4 to describe correlations between genes, we need to specify how upstream interactions lead to co-expression. As the simplest illustrative model system, we can consider the co-regulation of two genes, indexed by j, with $U_j = u_j e^{-\gamma_j \mathsf{s}}$. We outline several relatively simple classes of candidate models which induce expression coupling.

In the simplest case, $\mathcal{H}(\mathbf{u}, t) = \sum_j \mathcal{H}_j(u_j, t)$. In other words, the genes' dynamics are fully separable, and produce solutions in the form $G(\mathbf{u}, t) = \prod_j G_j(u_j, t)$. This formulation produces independent distributions at each t, but the *trajectories* may possess nontrivial statistical relationships. For example, if both genes start at $x_1 = x_2 = 0$, their trajectories will be correlated over a finite timespan $[0, T]$, with the correlation decaying as $T \to \infty$. This is the model implicit in the RNA velocity framework (Chapter 6). Therefore, this model class ascribes gene–gene relationships to transient phenomena, but cannot produce nontrivial stationary correlations.

In the next simplest case, co-regulation is the consequence of parameter differences in subpopulations. For example, the full cell population may consist of cell types indexed by κ. If we suppose each cell type has the abundance π_κ and transcriptional parameters Θ_κ, we yield

$$G(\mathbf{u}, t) = \sum_\kappa \pi_\kappa G(\mathbf{u}, t; \Theta_\kappa) = \sum_\kappa \pi_\kappa \prod_j G_j(u_j, t; \Theta_{j,\kappa}); \tag{10.1}$$

i.e., the generating function decomposes into a product of independent generating functions *conditional on* a particular cell type, but not globally. In other words, even if transcriptional processes are independent, cell type structure can produce nontrivial relationships between genes. However, this model cannot produce correlations *within* a cell type.

Equation 10.1 is immediately recognizable as the special discrete case of a more general mixture model:

$$G(\mathbf{u}, t) = \int_\Theta G(\mathbf{u}, t; \Theta) df_\Theta = \int_\Theta \prod_j G_j(u_j, t; \Theta_j) df_\Theta. \tag{10.2}$$

In other words, the cell types do not have to be point masses: the variation of parameters throughout the cell population may well be continuous, with higher-f_Θ regions corresponding to "cell states." This is the model implicit in the *scVI* variational autoencoder framework [185].

Alternatively, we can propose a model of co-regulation by the categorical variables. For example, two neighboring genes may prefer to have the same or opposite accessibility, depending on the polymeric properties of DNA. Assuming, for the purposes of illustration, that the system is symmetric, we yield the following $N = 4$

form:

$$s \in \{\text{both off, gene 1 on, gene 2 on, both on}\}$$

$$H = \begin{bmatrix} -2k_{\text{on}} & k_{\text{on}} & k_{\text{on}} & 0 \\ \varepsilon^{-1}k_{\text{off}} & -\varepsilon^{-1}(k_{\text{on}} + k_{\text{off}}) & 0 & \varepsilon^{-1}k_{\text{on}} \\ \varepsilon^{-1}k_{\text{off}} & 0 & -\varepsilon^{-1}(k_{\text{on}} + k_{\text{off}}) & \varepsilon^{-1}k_{\text{on}} \\ 0 & k_{\text{off}} & k_{\text{off}} & -2k_{\text{off}} \end{bmatrix} \tag{10.3}$$

$$\mathcal{A} = \begin{bmatrix} 0 \\ k_{\text{init}}u_1 \\ k_{\text{init}}u_2 \\ k_{\text{init}}(u_1 + u_2) \end{bmatrix}.$$

This form encodes the co-regulation of two genes. If $\varepsilon \ll 1$, the intermediate states are unstable and the genes tend to be either both on or both off. If $\varepsilon \gg 1$, the intermediate states are particularly stable, and only one of the genes tends to be on at a time. If $\varepsilon = 1$, we recover the independent case.

We can define a similar model for co-regulation by a continuous variable y_1, as an extension of Chapter 7.2 or the paired activation motif discussed in [205]. For example, there may be a latent regulator, such as the concentration of an activator, that controls multiple loci: if it is high, both have a high transcription rate; otherwise, both are inactive. This amounts to appending the following reactions to the master equation:

$$C_{1j}^{cd} y_1 [P(x_j - 1) - P(x_j)], \tag{10.4}$$

where the C^{cd} matrix encodes the relationship between the concentration and the transcription rate. Therefore, the genes become mutually correlated through the trajectory of y_1, although the extent of correlation depends on the dynamics.

If the categorical or continuous driving process is bursty, we can approximate it by a co-bursting module. For example, in the limit of $\varepsilon \to 0$, the dynamics of the system in Equation 10.3 converge to the $N = 2$ formulation

$$H = \begin{bmatrix} -k_{\text{on}}^* & k_{\text{on}}^* \\ k_{\text{off}}^* & -k_{\text{off}}^* \end{bmatrix} \text{ and } \mathcal{A} = \begin{bmatrix} 0 \\ k_{\text{init}}(u_1 + u_2) \end{bmatrix}, \text{ where}$$
$$k_{\text{on}}^* = \frac{2k_{\text{on}}^2}{k_{\text{on}} + k_{\text{off}}} \text{ and } k_{\text{off}}^* = \frac{2k_{\text{off}}^2}{k_{\text{on}} + k_{\text{off}}}. \tag{10.5}$$

If, in addition, $k_{\text{off}}^*, k_{\text{init}} \to \infty$, we obtain the $N = 1$ module characterized by

$$\mathcal{A} = k_{\text{on}}^* \left[\frac{1}{1 - b(u_1 + u_2)} - 1 \right], \tag{10.6}$$

where $b := k_{\text{init}}/k^*_{\text{off}}$. This is the bursty limit of Equation 10.3, which possesses the more general form

$$\mathcal{A} = \alpha \left[\frac{1}{1 - \sum_i b_j u_j} - 1 \right], \tag{10.7}$$

where each transcription event produces B_j molecules of X_j, with B_j drawn from a geometric distribution with mean b_j. Due to the structure of the burst distribution, the different gene products' burst sizes are correlated.

Interestingly, that mechanism also possesses a slow mixture limit. If $\varepsilon \to \infty$ while $k_{\text{on}}, k_{\text{off}} \to 0$, we obtain a special case of Equation 10.1, with $\pi_\kappa = 1/2$ and mutually exclusive expression in the "cell types," or long-lived gene states.

Even when we restrict our analysis to simple feed-forward regulation, this outline of motifs is nowhere near exhaustive. Nevertheless, the "mixture" and "bursty" limits are particularly natural starting points, as their distributions are straightforward to construct. In other words, we speculate that the careful analysis of co-expression models can distinguish relationships due to "slow" variation between cell types and "fast" variation due to coupled transcriptional events.

10.2 Biophysical constraints on "fast" transcript–transcript covariation

This section summarizes a portion of [105] by G.G. and L.P. The data analysis was conceptualized, designed, and implemented by G.G.

We cannot directly fit the "fast" variation models, as they are severely underspecified: we do not know *which* genes are co-regulated in this fashion. However, we can treat certain closely related problems and use the functional form of Equation 10.7 to *constrain* gene–gene relationships. In other words, although we cannot possibly fit all genes, if a particular pair of genes *were* expressed in simultaneous bursts, their expression must meet certain constraints. It turns out that the correlation between the two species has the correlation coefficient

$$\rho = \frac{2\sqrt{\gamma_1/\gamma_2}}{1 + \gamma_1/\gamma_2} \sqrt{\frac{1}{(1 + b_1^{-1})(1 + b_2^{-1})}}. \tag{10.8}$$

As the marginals are negative binomial, we could easily estimate b_i and γ_i from the marginals, without having to fit the full distribution. Once we have these estimates, we could predict the correlation between the genes. Interestingly, this equation is invariant under Bernoulli technical noise, with $p_j b_j$ taking the place of b_j. Nevertheless, as the set of assumptions is fairly severe, Equation 10.8 should be

treated as the upper bound on correlation coefficients between genes for this model class, in the spirit of [137, 138]. If real genes routinely violate this upper bound, the co-bursting model is insufficient to describe the gene–gene relationships in the dataset, and other model components need to be invoked.

This class of models can also describe *intra*-gene correlations. Specifically, a single "parent" transcript X_0 can give rise to multiple downstream transcripts X_j:

$$\begin{aligned} \varnothing &\xrightarrow{\alpha} B \times X_0 \\ X_0 &\xrightarrow{\beta_j} X_j \\ X_j &\xrightarrow{\gamma_j} \varnothing, \end{aligned} \tag{10.9}$$

where the first line represents the bursty transcription of X_0, the second encodes multiple splicing routes, and the third represents the degradation or isomerization to secondary transcript forms, which we do not consider. If all β_j are high relative to all γ_j, the distribution of each X_j is negative binomial with shape α/γ_j and scale $b\beta_j/\beta$, where $\beta := \sum_j \beta_j$ is the total efflux rate from X_0. In addition, the correlation between any two transcripts is given by Equation 10.8, with $b\beta_j/\beta$ taking the place of b_j.

To understand whether real systems actually follow this bound on transcript–transcript correlations, we obtained data from the recent FLT-seq (full-length transcript sequencing by sampling) protocol [287], which uses nanopore technology to obtain long reads amenable to the identification of transient transcripts. As this experimental technique has molecular and cellular barcodes, the data are interpretable as discrete transcript counts sampled from a distribution. To minimize transient effects, such as cell cycling and differentiation, we selected a dataset generated from cultured mouse stem cells. To limit biological heterogeneity due to discrete cell subpopulations (as in Equation 10.1), we filtered for cell barcodes corresponding to the activated cell subset (136 barcodes) according to the authors' annotations. In all downstream analyses, we treated this filtered dataset as biologically homogeneous up to endogenous stochasticity.

The FLT-seq protocol produces full-length reads, which can be used to discover new isoforms, but does not reveal causal relationships between those isoforms. Nevertheless, we can use the tools of discrete mathematics to partially infer these relationships. Splicing removes introns, but cannot insert them. We can use this relationship to constrain the splicing graph: if transcript X_j can be obtained by removing part of the sequence in transcript X_i, there must be a path from X_i to X_j.

On the other hand, if $\mathcal{X}_i$ contains the sequence I_i but omits the sequence I_j, whereas $\mathcal{X}_j$ contains the sequence I_j but omits the sequence I_i, the transcripts are *mutually exclusive* and must be generated from the parent transcript by distinct pathways.

For each gene, we enumerate the transcripts observed in the data and split them into elementary intervals, contiguous stretches that are either present or absent in each transcript (denoted by the colors in Figure 10.1a). These elementary intervals constrain the relationships between transcripts, and we can use their presence or absence in each transcript to construct an accessibility graph. The internal structure of this graph is underspecified, but immaterial: the negative binomial model implied by the operator in Equation 10.7 describes the *roots*, mutually exclusive transcripts that must be generated directly from the parent transcript (indicated in orange in Figure 10.1b). We fit the distributions of these roots, discarding any data that are underdispersed, overly sparse, or fail to converge to a fit. The satisfactory fits for the sample gene *Rpl13* are shown in Figure 10.1c.

The negative binomial fit yields burst sizes b_i and non-dimensionalized efflux rates γ_i. We substitute these quantities into Equation 10.8, compute hypothetical correlations ρ_{theo}, and compare them to sample correlations ρ_{samp} in Figure 10.1d. These results represent the 4,885 nontrivial correlation matrix entries between 1,978 transcripts from 500 genes. 302 transcripts were rejected due to underdispersion, 542 due to sparsity, and 100 due to poor fits. The theoretical constraint (sample correlation equal to or lower than predicted correlation) was met in 4,606 cases (94.3%).

The results suggest that the model is not sufficient to recapitulate the full dynamics, but *does* provide an effective, and nontrivial, theoretical constraint. We hypothesize that the "consistent" regime ($\rho_{\text{samp}} \in (0, \rho_{\text{theo}})$, 3,856 entries) represents the degradation of correlations due to technical noise in the sequencing process and stochastic intermediates. The "inconsistent" regime ($\rho_{\text{samp}} \in (\rho_{\text{theo}}, 1)$, 279 entries) may stem from model misidentification, and could be explained by coupling between splicing events. Some of these apparently inconsistent correlations may also be due to the small sample sizes, as the bootstrap 95% confidence intervals only rarely lie outside the theoretical bound (29 entries). Finally, the "negative" regime ($\rho_{\text{samp}} < 0$, 750 entries) technically meets the constraint, but cannot actually be reproduced by the model. This does not appear to be an artifact of sample sizes. Instead, we speculate that enrichment in negative correlations is the signature of a more complicated regulatory schema which preferentially synthesizes some isoforms to the exclusion of others, rather than choosing the splicing pathway randomly.

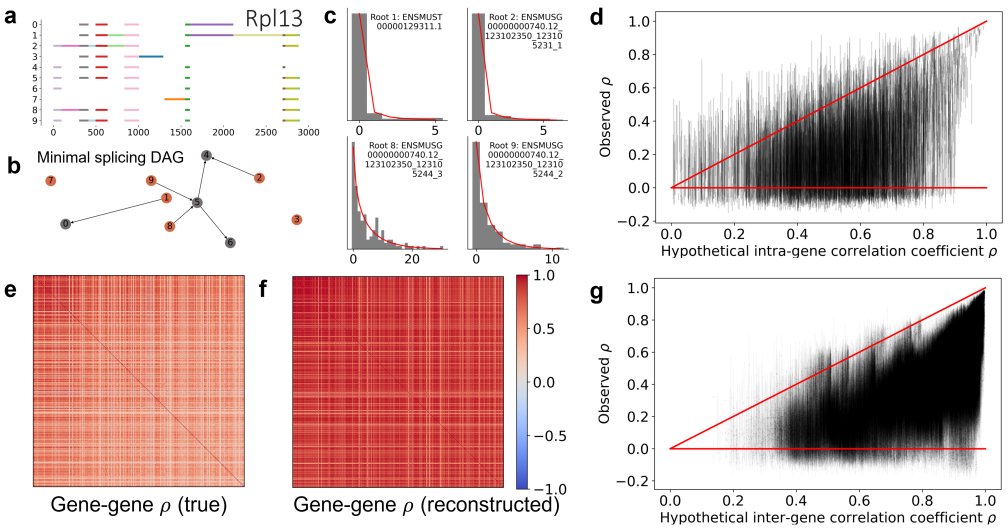

Figure 10.1: The synchronized-burst model can be leveraged to constrain transcript-transcript correlations.

a. By inspecting exon co-expression structures in long-read sequencing data, we can split genes into elementary intervals.

b. Although sequencing data are not sufficient to identify the relationships between various transcripts, they can provide information about "roots" of the splicing graph (highlighted in orange), which must be produced from the parent transcript by mutually exclusive pathways.

c. The root transcript copy number distributions are well-described by negative binomial laws (gray histograms: raw marginal count data; red lines: fits).

d. The co-bursting model is not sufficient to accurately predict transcript-transcript correlations, but does serve as a nontrivial upper bound: few sample correlations exceed the model-based predictions obtained from Equation 10.8 (points: transcript-transcript correlation matrix entries for mutually exclusive "root" transcripts of a single gene; error bars: bootstrap 95% confidence intervals; red line: theory/experiment identity line).

e. The highest-expressed transcripts across the top 500 genes show distinctive, and generally positive, correlation patterns.

f. We can use an analogous model to predict and reconstruct the gene–gene correlation matrix based solely on marginal data.

g. As before, the model is not sufficient to accurately predict gene–gene correlations, but provides an effective and nontrivial upper bound (points: gene–gene correlation matrix entries; error bars: bootstrap 95% confidence intervals; red line: theory/experiment identity line).

Analogously, we can exploit the inter-gene model encoded in Equation 10.7 to predict the gene-gene correlation matrix (Figure 10.1e) based solely on the marginals, supposing *all* pairs of 500 highest-expressed genes fire simultaneously as a limiting case. For each gene, we consider the highest-abundance root transcript that can be fit by a negative binomial distribution, and identify its marginal burst size and efflux rate. Substituting these parameter estimates into Equation 10.8, we obtain theoretical correlations ρ_{theo} and reconstruct the correlation matrix (Figure 10.1f). Finally, we compare the intra-gene sample correlations ρ_{samp} to the theoretical values in Figure 10.1g. These results represent the 119,805 nontrivial correlation matrix entries based on the 490 genes with well-fit roots. The theoretical constraint (sample correlation equal to or lower than predicted correlation) was met in 119,503 cases (99.7%), with only five confidence intervals above the bound.

Yet again, the model provides a nontrivial bound. We hypothesize that the "consistent" regime ($\rho_{\text{samp}} \in (0, \rho_{\text{theo}})$, 117,542 entries) represents the degradation of correlations due to stochastic effects outside the model, much as before. The correlations in the "inconsistent" regime ($\rho_{\text{samp}} \in (\rho_{\text{theo}}, 1)$, 302 entries) lie very close to the identity line, so we hypothesize they are mostly explained by small sample sizes. Finally, the "negative" regime ($\rho_{\text{samp}} < 0$, 1,961 entries) is rare, and we expect these observations also emerge from small sample sizes.

This model is extremely simple: we have largely omitted the realistic description of technical noise, the modeling of transient intermediates, and the accurate inference of parameters. Nevertheless, for nearly every pair of transcripts we observe, the distribution shapes are consistent with the nontrivial bound obtained by assuming the co-bursting model holds. This model cannot recapitulate the precise quantitative details; such an effort would require considerably more involved modeling and statistics. Nevertheless, it does suggest that the conception of "fast" gene–gene variation has some predictive value, and provides a foundation for developing more sophisticated models. In addition, the analytical procedure provides a framework for testing the consistency of models prior to performing a computationally intensive full fit.

10.3 Multimodal variational autoencoder models for "slow" covariation

This section summarizes the content of [44] by M.C.*, G.G.*, Y.C., T.C., and L.P. The *biVI* approach was conceptualized by G.G., designed by G.G., M.C., Y.C., and T.C., and implemented by M.C., Y.C., and T.C. The statistical derivations were performed by G.G. and M.C.

Alternatively, we may explain co-variation in gene expression by cell type differences within a sample. This approach may be as simple as fitting a mixture model to Equation 10.1, but one promising alternative direction has used neural networks to approximate the more general mixture in Equation 10.2. For example, the popular tool *scVI* is a variational autoencoder (VAE) that uses neural networks to encode scRNA-seq counts to a low-dimensional representation. This representation is decoded by another neural network to a set of cell- and gene- specific parameters for conditional likelihood distributions of observed counts. These Poisson or negative binomial[7] distributions are chosen *post hoc* to be consistent with the discrete, over-dispersed nature of scRNA-seq counts, but can be derived from biophysical models (Sections 2.1 and 4.6).

Extensions of *scVI* to bimodal data have been attempted for protein [96] and chromatin measurements [12] by jointly encoding data modalities to a single latent space, then employing two decoding networks to produce parameters for *independent* conditional likelihoods specific to each datatype. Nascent and mature transcripts [168, 197] could be similarly treated (Figure 10.2a). However, using independent conditional likelihoods for bimodal measurements derived from the same gene ignores the inherent causality between observations and has no biophysical basis: the generative model is merely part of a neural "black box" used to summarize data.

Nevertheless, good causal model candidates for the nascent–mature distributions are available, such as the extensively validated [65, 233, 244] bursty model of transcription (Section 4.6.2). While the joint steady-state distribution induced by the bursty model is analytically intractable [261], we have previously shown that it can be approximated by a set of basis functions with neural-network learned weights (Section 5.3). To that end, we introduce *biVI*, a strategy that adapts *scVI* to work with well-characterized stochastic models of transcription.

First, we propose a parameterization of the bursty process that could give rise to bivariate count distributions for nascent and mature transcripts, such that the univariate case matches the *scVI* assumptions. Specifically, *scVI* assumes that the conditional distribution represents contributions from a gene-specific dispersion parameter ν_g, a cell-specific "size" parameter ℓ_c, and cell- and gene-specific compositional parameter ρ_{cg}, such that the distribution is negative binomial with shape ν_g and mean $\mu_{cg} = \ell_c \rho_{cg}$.

Formalizing this descriptive model requires specifying the precise mechanistic meaning of ℓ_g. Previous reports equivocate [96], appealing to a combination of

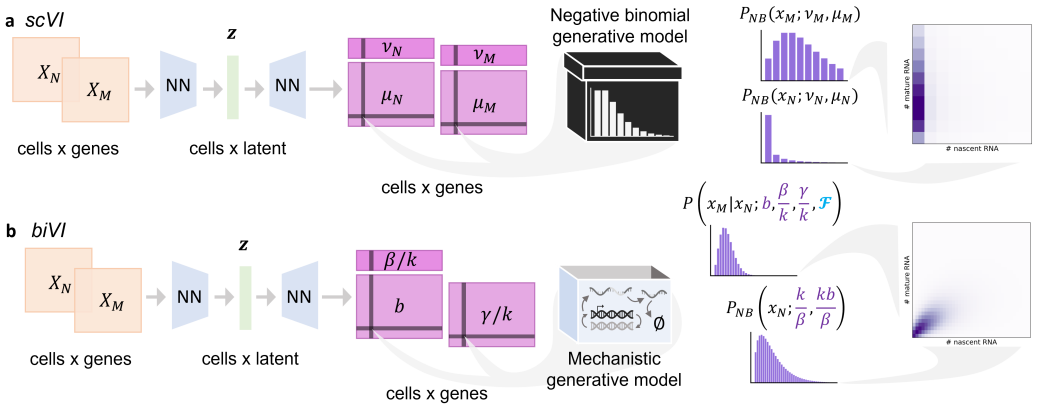

Figure 10.2: *biVI* reinterprets and extends *scVI* to infer biophysical parameters.
a. *scVI* can take in concatenated nascent (X_N) and mature (X_M) RNA count matrices, encode each cell to a low-dimensional space **z**, and learn per-cell parameters μ_N and μ_M and per-gene parameters ν_N and ν_M for independent nascent and mature count distributions. This approach is not motivated by any specific biophysical model.
b. Operating conditional on the bursty model of transcription, *biVI* can take in nascent and mature count matrices, produce a low-dimensional representation for each cell, and output per-cell parameters b and γ/k, as well as the per-gene parameters β/k, for a mechanistically motivated joint distribution of nascent and mature counts.

"cell size," cell-wide effects on the biology (in the spirit of [81, 124]), or "sequencing depth," technical variability in the sequencing process (in the spirit of [308]). In other words, the former scenario represents, e.g., systematic differences in the concentrations of relevant macromolecules, such as RNA polymerase, whereas the latter scenario represents random differences in the amount of sequencing primers between 10x beads.

For simplicity, we only treat the first case here, although either one may be used as the basis for a mechanistic formulation. If we introduce a genome-wide scaling factor C and recall the basis of the bursty model (Section A.8.1), we find that the

following interpretation of the parameters matches *scVI* for a univariate model:

$$\begin{aligned}
\mu_{\mathsf{cg}} &= \frac{k_{\mathsf{g}} b_{\mathsf{cg}}}{\gamma_{\mathsf{g}}} \\
\nu_{\mathsf{g}} &:= \frac{k_{\mathsf{g}}}{\gamma_{\mathsf{g}}} \\
b_{\mathsf{cg}} &:= b_{\mathsf{cg,RNAP}}[\mathrm{RNAP}]_{\mathsf{c}} \\
&= \frac{b_{\mathsf{cg,RNAP}}}{C} \times C[\mathrm{RNAP}]_{\mathsf{c}} \\
&:= \frac{\rho_{\mathsf{cg}} \ell_{\mathsf{c}}}{\nu_{\mathsf{g}}}.
\end{aligned} \tag{10.10}$$

In other words, the burst size consists of a cell- and gene-specific term, which describes the *scaling* of the burst size with respect with respect to the polymerase concentration $[\mathrm{RNAP}]_{\mathsf{c}}$, as well as a cell-specific term, which encodes this concentration. The dependence on c encodes the mixture model in Equation 10.2. By setting the units appropriately and making C fairly large, we can find small ρ_{cg} and large ℓ_{c} that produce acceptable fits to the data under the usual *scVI* priors and functional assumptions. The rest of the variability is encoded in ν_{g}, and implicitly assumes that the burst frequency and degradation rate do not change between cells.

Next, we construct a two-species bursty model that retains these assumptions. The simplest one, with no technical noise, takes the following form:

$$\begin{aligned}
\mu_{\mathsf{cg}}^{(N)} &= \frac{k_{\mathsf{g}} b_{\mathsf{cg}}}{\beta_{\mathsf{g}}} \\
\mu_{\mathsf{cg}}^{(M)} &= \frac{k_{\mathsf{g}} b_{\mathsf{cg}}}{\gamma_{\mathsf{cg}}} \\
\nu_{\mathsf{g}} &= \frac{k_{\mathsf{g}}}{\beta_{\mathsf{g}}},
\end{aligned} \tag{10.11}$$

and b_{cg} defined as in Equation 10.10. We have somewhat arbitrarily assumed that the fixed ν should correspond to a fixed nascent negative binomial marginal shape parameter in the bivariate case. This yields the following parameter definitions:

$$\begin{aligned}
\rho_{\mathsf{cg}}^{(N)} &= \nu_{\mathsf{g}} \frac{b_{\mathsf{cg,RNAP}}}{C} \\
\rho_{\mathsf{cg}}^{(M)} &= \nu_{\mathsf{g}} \frac{\beta_{\mathsf{g}}}{\gamma_{\mathsf{cg}} C} b_{\mathsf{cg,RNAP}} \\
\ell_{\mathsf{c}} &= C[\mathrm{RNAP}]_{\mathsf{c}}.
\end{aligned} \tag{10.12}$$

Given a particular set of ν_{g}, $\rho_{\mathsf{cg}}^{(N)}$, $\rho_{\mathsf{cg}}^{(M)}$, and ℓ_{c}, we can immediately compute the cell- and gene-specific parameters:

$$\begin{aligned} \frac{\beta_{\mathsf{g}}}{k_{\mathsf{g}}} &= \frac{1}{\nu_{\mathsf{g}}} \\ b_{\mathsf{cg}} &= \frac{\rho_{\mathsf{cg}}^{(N)} \ell_{\mathsf{c}}}{\nu_{\mathsf{g}}} \\ \frac{\gamma_{\mathsf{cg}}}{k_{\mathsf{g}}} &= \frac{\rho_{\mathsf{cg}}^{(N)}}{\nu_{\mathsf{g}} \rho_{\mathsf{cg}}^{(M)}}. \end{aligned} \tag{10.13}$$

Using the neural solver in Section 5.3, we can compute distributions and incrementally increasing the likelihood of data $\mathcal{D}_{\mathsf{cg}}$ under the model by training the network and updating the parameters; ν values are treated as deterministic, while ℓ values may treated as probabilistic or simply use the total molecule count as a plug-in estimate. Each latent vector $\mathbf{z}_{\mathsf{c}}$ is decoded to a pair of $\rho_{\mathsf{cg}}^{(N)}, \rho_{\mathsf{cg}}^{(M)}$, such that $\sum_{\mathsf{g}} \left[\rho_{\mathsf{cg}}^{(N)} + \rho_{\mathsf{cg}}^{(M)} \right] = 1$.

Per Equation 10.13, the inferred likelihood parameters have biophysical interpretations under a specific mechanistic model of transcriptional dynamics. Although we focus on the bursty model, *biVI* also implements the closed-form constitutive (Section 4.6.1) and extrinsic (Section 4.6.3) noise models [81, 124].

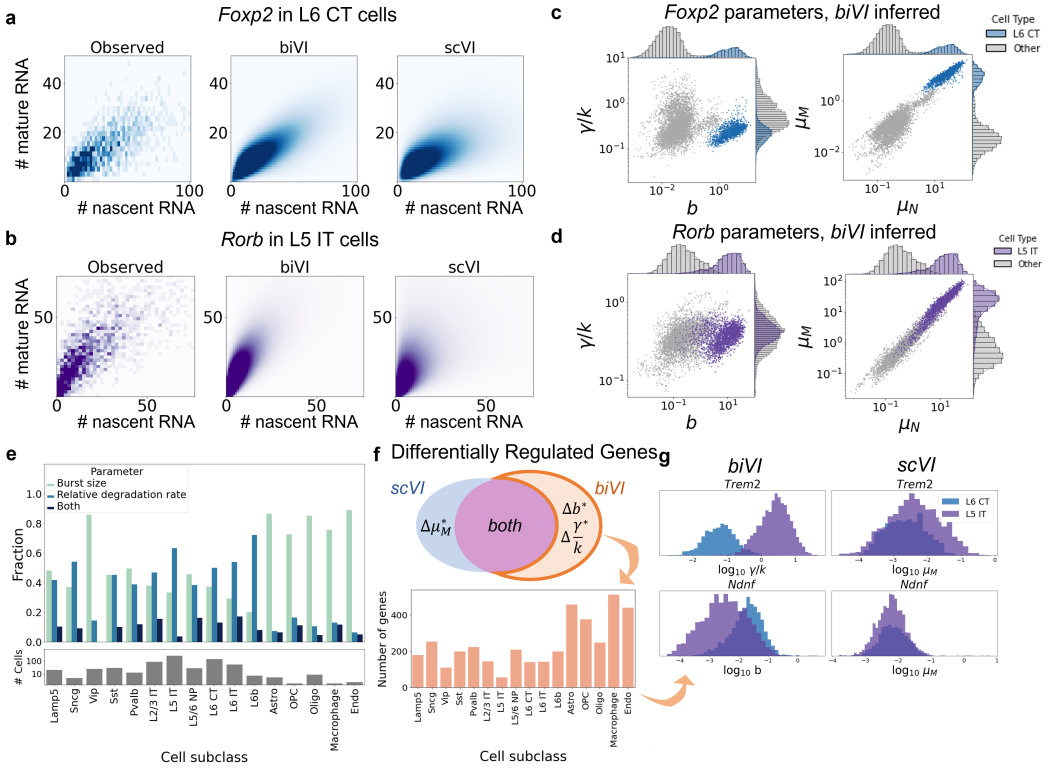

Figure 10.3: *biVI* successfully fits single-cell neuron data and suggests the biophysical basis for expression differences.

a.-b. Observed, *scVI*, and *biVI* reconstructed distributions of *Foxp2*, a marker gene for L6 CT (layer 6 corticothalamic) cells, and *Rorb*, a marker gene for L5 IT (layer 5 intratelencephalic) cells, restricted to respective cell type.

c.-d. Cell-specific parameters inferred for *Foxp2* and *Rorb* demonstrate identifiable differences in means and parameters in the marked cell types.

e. Cell subclasses show different modulation patterns, with especially pronounced distinctions in non-neuronal cells (top: fractions of genes exhibiting differences in each parameter; bottom: number of cells in each subclass).

f. *biVI* allows the identification of cells which exhibit differences in burst size or relative degradation rate, without necessarily demonstrating differences in mature mean expression. Hundreds of genes demonstrate this modulation behavior, with variation across cell subclasses.

g. Histograms of *biVI* parameters and *scVI* mature means for two genes that exhibit parameter modulation without identifiable mature mean modulation. *Trem2* (top) shows differences in the degradation rate in L5 IT cells, whereas *Ndnf* (bottom) shows differences in burst size in L6 CT cells.

Under comparable conditions, *biVI* recapitulates observed bivariate RNA distributions better than *scVI* (Figure 10.3a-b). In addition, the latent space structure effectively recapitulates cell subtypes from existing annotations of the mouse neuron dataset under consideration [321]. Beyond this empirical agreement, it allows us to *interpret* differences in the generative model parameters in terms of biophysics, in the spirit of Section 9.2. For example, in Figure 10.3c-d, we illustrate that the upregulation of markers *Foxp2* and *Rorb* can be ascribed to an increase in burst size; these differences are starkly evident in the distribution of parameters but much less so in the distribution of averages.

We extend and exploit this approach to find "marker genes" that demonstrate substantial modulation in the values of $b_{\mathrm{cg,RNAP}}$ and $\gamma_{\mathrm{cg}}/k_{\mathrm{g}}$ between cell types using a Bayesian procedure analogous to the approach in [96]. The results are visualized in Figure 10.3e. Surprisingly, even in this high-level summary, variation between cell types is quite considerable: neuronal cells appear to regulate gene expression via a mix of regulatory strategies, while non-neuronal cells seem to preferentially modulate burst size.

As in Section 9.2, many genes that demonstrated substantial parameter differences did not demonstrate strong differences in the mature RNA averages. For some cell subclasses, there were several hundred such genes (Figure 10.3f). For example, the relative degradation rate of the gene coding for the triggering receptor expressed on myeloid cells-2 (TREM2), variants of which are strongly associated with increased risk of Alzheimer's disease [296], was found to be greater in L5 IT neurons than in other subclasses (Figure 10.3g, top row). Similarly, the gene *Ndnf*, which codes for the neuron derived neurotrophic factor NDNF and promotes the growth, migration, and survival of neurons [165], demonstrated a statistically significant difference in the *biVI* inferred burst size, but not *scVI* inferred mature mean, in L6 CT neurons (Figure 10.3g, bottom row).

Such a mechanistic description provides a framework for characterizing the connection between a gene's role and a cell's regulatory strategies beyond a mere change in mean expression [204, 206]. The neural framework enables us to relax many of the assumptions of simple mechanistic models, treat non-homogeneous cell populations, and *discover* internal differences. This marriage of the neural and the mechanistic provides an actionable implementation of the themes developed in [137, 156]: the known physics are represented explicitly; the obscure unspecified networks and parameters are relegated to a neural network "black box." This network can, in

turn, be made more "transparent" by using a linear, rather than neural decoder to map from the low-dimensional latent space to the biophysical parameters, and we obtained reasonable results by implementing such an architecture [276].

We believe that the design of variational autoencoders with neural and mechanistic components presents an exciting avenue for single-cell data analysis: this approach already scales to hundreds of thousands of cells [97] and can easily be extended to more sophisticated models by working through the necessary mathematics. However, the construction of compatible likelihood functions (Section 5.3 and [271, 310]) is no trivial task, and typically requires developing bespoke routines and training approximators anew with each addition. We anticipate that the development of more realistic technical noise models is necessary, especially in light of Section 8.2.

On the other hand, the conclusions we can draw are only as good as the models. The bursty model of transcription is fairly well-attested, but other assumptions we have made may not be. The first implementation of *biVI* strives to be consistent with *scVI*; in this quest for consistency, it sacrifices the ability to describe variation in burst frequencies, which are surely important to the differences between cell types (Chapter 9 and [65]). We could, for example, conceptualize a model that allows the transcriptional parameters to vary while keeping the turnover rates constant:

$$\begin{aligned} \mu_{\mathsf{cg}}^{(N)} &= \frac{k_{\mathsf{cg}} b_{\mathsf{cg}}}{\beta_{\mathsf{g}}} \\ \mu_{\mathsf{cg}}^{(M)} &= \frac{k_{\mathsf{cg}} b_{\mathsf{cg}}}{\gamma_{\mathsf{g}}}, \end{aligned} \tag{10.14}$$

keeping $\beta_{\mathsf{g}}/\gamma_{\mathsf{g}}$ constant. In this case, we may be able to interpret the "cell size" scaling as mass constraint. In a "strong" mass constraint, we would not allow the total number of molecules to exceed some preset bound; this form of constraint gives rise to intractable distributions. Instead, we would impose a "weak" mass constraint, such that genes cannot have simultaneously have arbitrarily high averages. This implies the following form:

$$\begin{aligned} k_{\mathsf{cg}} b_{\mathsf{cg}} &= \rho_{\mathsf{cg}}^{\mu} \ell_{\mathsf{c}} \\ k_{\mathsf{cg}} &= \rho_{\mathsf{cg}}^{k} \rho_{\mathsf{cg}}^{\mu} \ell_{\mathsf{c}} \\ b_{\mathsf{cg}} &= (1 - \rho_{\mathsf{cg}}^{k}) \rho_{\mathsf{cg}}^{\mu} \ell_{\mathsf{c}}, \end{aligned} \tag{10.15}$$

such that $\sum_{\mathsf{g}} \rho_{\mathsf{cg}}^{\mu} = 1$, whereas $\rho_{\mathsf{cg}}^{b} \in (0, 1)$ but not otherwise constrained. It remains to learn or specify β_{g} and γ_{g}, which may not be mutually identifiable. Overall, the "correct" way to implement such a constraint is far from clear, and this direction remains an area of active research.

In addition to representing more realistic biological phenomena, we anticipate that the VAE framework can be relatively straightforwardly integrated with technical noise phenomena discussed in Sections 4.4.2 and 8.1. Specifically, if the biological RNA distribution is negative binomial with shape ν and scale θ, whereas the background distribution is Poisson with mean μ, the PGF of the overall distribution is given by the product of the individual PGFs:

$$G(u) = \left(\frac{1}{1-\theta u}\right)^{\nu} e^{\mu u}. \tag{10.16}$$

By directly applying Equation 3.12 and the definition of Kummer's confluent hypergeometric function M in Equation 3.19, we find that the molecule generative distribution is

$$P(x) = \frac{\Gamma(\nu+x)}{x!\Gamma(\nu)} \frac{e^{-\mu}\theta^x}{(1+\theta)^{\nu+x}} M\left(-x, 1-x-\nu, \frac{\mu\theta}{1+\theta}\right). \tag{10.17}$$

Whenever $\mu = 0$, the hypergeometric and exponential terms are unity, yielding the expected negative binomial distribution. We expect that a reasonable estimate for μ can be obtained by rescaling the dataset-wide average expression. However, the optimal way to implement Equation 10.17 is somewhat obscure. Although Kummer's function M can be written down in closed form for $x \in \mathbb{N}_0$, the expression is somewhat unwieldy; to integrate this form of variation into a VAE, it may be more fruitful to use an approximation of the function tailored to the low-μ regime. Whatever the implementation, the design of such a generative model requires careful consideration of the basis and meaning of "cell size" effects, and remains a compelling target for future investigations.

Chapter 11

MODELING FURTHER CLASSES OF MULTIOMIC DATA

There was set before me a mighty hill,
And long days I climbed
Through regions of snow.
When I had before me the summit-view,
It seemed that my labour
Had been to see gardens
Lying at impossible distances.

The Black Riders and Other Lines, XXVI
Stephen Crane

Technology hardly stands still, and recent years have seen the development of assays that quantify RNA alongside other modalities, such as chromatin accessibility, epigenetic modifications, and protein content [213]. These technological advances have been accompanied by a variety of more or less *ad hoc* methods for data integration [12, 96, 128, 180]. Although these methods produce correlated results, they are not perfect proxies for each other [198], nor should they be treated as such: their sources of technical and biological stochasticity are fundamentally different.

There is some hope that we can "integrate" these data types by appealing to the central dogma, and treating observations as realizations of a common process. This strategy extends our discussion thus far: there are conventional models for a single RNA species; to represent nascent and mature RNA, we merely append another reaction; to represent other modalities, we can extend the stochastic systems further, by appending chromatin state transitions to represent DNA states and translation reactions to represent proteins. However, although this conceptual picture is very much in line with the rest of the thesis, the details — i.e., the "correct" ways to represent these phenomena — are as of yet obscure. In this chapter, we speculate about some promising directions for modeling and data analysis.

11.1 Protein velocity and acceleration

This section summarizes the content of [109] by G.G., V.S., and L.P. The method was conceptualized by V.S. and L.P., and designed and implemented by G.G.

The RNA velocity pipeline, reviewed in Section 6.1.1, can be easily, and self-consistently, extended to protein species with abundance y_P [275]:

$$\begin{aligned} \frac{dy_M(t)}{dt} &= \beta y_N(t) - \gamma y_M(t) \\ \frac{dy_P(t)}{dt} &= \beta_P y_M(t) - \gamma_P y_P(t), \end{aligned} \tag{11.1}$$

where β_P is the translation rate and γ_P is the protein degradation rate. In other words, if we have protein observations, we can define an "RNA velocity" and a "protein velocity." The key distinction involves the interpretation. The RNA velocity allows us to extrapolate into the future, because the current nascent RNA content is a leading indicator of the mature RNA content. On the other hand, the protein velocity allows us to extrapolate into the past, because the current protein content is a lagging indicator of the mature RNA content.

Thus, in the standard implementation of RNA velocity, we extrapolate the mature RNA matrix into the future, compare the direction of the extrapolation vector to the directions of the embedding neighbors in mature RNA space, and use this comparison to build a low-dimensional projection of the "future." In the case of proteins — treating this as more of an analogy then a rigorous derivation — we extrapolate the protein matrix into the past, compare the direction to the directions of the embedding neighbors in protein space, and construct another low-dimensional projection, now reflecting the "past" of the system. We obtain two vectors per cell, whose directions may not match. If they are particularly misaligned, the system exhibits high "acceleration," which is a second-order, mostly qualitative, characterization of changes in the mature transcriptome.

This approach is theoretically consistent with RNA velocity, easily generalizes, and produces apparently reproducible trends across a variety of datasets. Although our description extends the *velocyto* assumptions, this line of argument appears to have inspired an alternative, *scVelo*-based approach to such trivariate models [313]. However, the usual pitfalls of these techniques apply (Section 6.1). The data processing and inference procedures are *ad hoc*, although the simplistic noise and dynamics assumptions may actually be more legitimate for high-abundance proteins than for low-abundance RNA. The embedding procedure is arbitrary in much the same ways as elsewhere.

More problematically, there is somewhat of a mismatch between the modalities. RNA sequencing captures genome-wide, endogenous nuclear and cytoplasmic RNA,

as well as various background noise. On the other hand, the protein quantification technologies built on top of scRNA-seq exploit synthetic, oligo-tagged antibodies bound to membrane proteins. Therefore, key features of the system, such as the cytoplasmic protein content and the precise stoichiometry of antibody binding, are obscure and apparently impossible to determine from the data. In addition to these factors, previous attempts to model this data type have found that proteins exhibit unique and fairly complicated technical artifacts [96, 322]. Much more fundamentally, the process chemistry restricts the approach to *a priori* well-characterized cell types with commercially available antibodies, i.e., blood cell immune profiling, which somewhat limits the breadth of investigations.

The strategy set out in [31, 253] and outlined in Chapter 4 may hold more promise: RNA and proteins have a causal relationship, so fitting a model of transcription and translation may tell us about the underlying biophysics. However, given the considerable difficulties of solving such models, we anticipate that a study of model behaviors and feasibility would be more appropriate at these early stages, in the spirit of [138].

11.2 Chromatin accessibility

This section summarizes unpublished research undertaken by C.F. and G.G. The form of the model is due to C.F.; the overarching motivation and connection to Glauber dynamics is due to G.G.

In addition to molecular modalities, we would like to self-consistently "integrate" DNA measurements, such as epigenetic markers or chromatin accessibility. For example, if we are interested in the latter, we need to propose a model that defines the dynamics of chromatin opening and closing, and endow it with realistic technical noise. Thus far, statistical approaches to this problem have been largely descriptive [12, 181], with limited use of mechanistic models.

In Section 10.1, we discussed potentially promising ways to model multi-gene systems. In particular, a cursory examination of Equation 10.3 reveals that co-regulation by categorical variables, with a parameter ε controlling the strength of neighbor interactions, is essentially identical to the continuous-time Glauber formulation [101, 132] of the Ising model of lattice spins [48, 234]. The Ising model is familiar from statistical thermodynamics, and encodes interacting spins on a lattice. At equilibrium, the distribution of states is Boltzmann:

$$P(\sigma) \sim e^{-\beta H(\sigma)}, \tag{11.2}$$

where σ is a particular combination of spin states, β is the inverse temperature, and $H(\sigma)$ is the energy associated with σ, encoded in the Hamiltonian. The Hamiltonian, in turn, encodes the interactions between adjacent spins, which drive them to align in a parallel or anti-parallel way, and the field strength, which drives them to align with to the field. Although the Ising model is typically studied at steady state, the Glauber formulation constructs a continuous-time Markov chain with the correct steady state, and allows us to couple the spin dynamics to downstream transcription processes. With some algebra, it is straightforward to see that k_{on} and k_{off} control the field strength, whereas ε controls the interactions. Therefore, it seems legitimate to associate the "on" state with open chromatin and the "off" state with closed chromatin, and attempt to fit joint distributions of RNA counts and DNA states.

Ising-style models are appealing and provide certain advantages. The model structure is simple, but encodes two key ideas: on one hand, neighboring genes are co-regulated [84]; on the other, they are bursty when considered individually. The statistics of the Ising model are well-understood, and it is likely that we can obtain important properties of the RNA distributions in terms of the underlying state kinetics. Although the technical noise behaviors for ATAC and RNA-seq technologies are likely quite different, we can begin to construct simple hypotheses. For example, if 00110 denotes a state vector with three occluded and two exposed sites, we should be able to obtain measurements of 00110, 00100, 00010, and 00000, where the exposed sites are erroneously reported as occluded due to stochastic loss of reads; on the other hand, we should not obtain measurements like 10110. However, the details are somewhat obscure and require further study; for example, it is likely that polymerase and transcription factor occupancy can interfere with ATAC readouts, and systematically lead to active sites being reported as occluded.

Usefully, the approach generalizes: the Ising approach has intuitive "knobs" that can be "tuned" to incorporate more sophisticated phenomena. If the model is too simplistic to fit data, we can easily relax certain assumptions, e.g., allow the field or interaction strengths to vary between sites. Most interestingly, once we have begun to operate in the Ising framework, we are not restricted to simple lattice models: we may be able to leverage chromatin structure measurements, such as Hi-C [176], to construct generic gene–gene interaction graphs and model co-regulation accordingly.

Although rare, this class of models is precedented in bioinformatics, and has been

applied to the study of methylation [148] and DNA–protein interactions [199]. Ultimately, however, this direction is very much in its nascence, and considerable further research will be necessary to understand whether Glauber-like dynamics are at all consistent with real data.

11.3 Spatial transcriptomics

This section summarizes unpublished research undertaken by K.J. and G.G.

Finally, it is worthwhile to discuss the compatibility between the mechanistic approach and the recent spate of sequencing-based spatial transcriptomics technologies [202]. Essentially, the commercial methods provide a grid, rather than suspension, of barcoded beads; a tissue is placed on a grid and its RNA is reverse-transcribed and sequenced; the barcode design allows for the reconstruction of the original spatial configuration of the beads. We can begin to construct a model by proposing that a single cell's RNA content, for a particular gene g, is drawn from a negative binomial distribution with shape $k_{\mathsf{g}}/\gamma_{\mathsf{g}}$ and scale $b_{\mathsf{g}}p$. All of these parameters vary with two-dimensional location $\mathbf{z}$, but in different fashions: intuitively, we should expect the endogenous parameters k_{g}, γ_{g}, and b_{g} to depend on the cell type, and the technical parameter p to only depend on the grid. Of course, some beads may be associated with more than one cell, in which case the molecule distribution per barcode would be a sum of negative binomials. In addition, if RNA freely diffuse and contaminate non-cell-associated regions, we may also append a Poisson technical noise term, in the spirit of Section 4.4.2. Therefore, the full generating function may take the following form:

$$G_{\text{tot,t}} = G_{\text{bc}}\left(G(\mathbf{z}, p(\mathbf{z})), \mathbf{z}\right) \times G_{\text{bg}}(\mathbf{z}, p(\mathbf{z})), \tag{11.3}$$

where $G_{\text{bg}}(\mathbf{z})$ encodes the background at the grid point $\mathbf{z}$, G_{bc} encodes the density of cells captured per bead at $\mathbf{z}$ (where we have assumed that a single bead can only capture cells from a single cell type), $p(\mathbf{z})$ is the local molecule capture probability, and G is the negative binomial PGF. This formula is fairly generic, but can be used to, e.g., generate synthetic spatial data by making assumptions about the biological and technical components of variability. For example, we can make G_{bc} degenerate (one cell per barcode), $p(\mathbf{z})$ a low-frequency Gaussian process transformed to be within $(0, 1)$, G_{bg} the usual pseudobulk Poisson distribution (Section 8.1), and G the negative binomial PGF, with parameters that are piecewise constant functions of $\mathbf{z}$.

The extension of the modeling framework to spatial transcriptomic data is an active area of research, and the optimal way to actually fit distributions is far from clear for now. However, we note that the use of the notation $\mathbf{z}$, matching Section 10.3, is not incidental: we may be able to treat the location as a predictor, then use a generic neural function to learn the parameters' dependence on $\mathbf{z}$, hopefully recapitulating the cell types. Further, the underlying assumption of cell–cell independence appears to be somewhat restrictive in this context, and it is plausible that agent-based models, in the spirit of [283, 285] are more appropriate for spatial data.

Chapter 12

DISCUSSION AND CONCLUSION

> But a poem is never actually finished.
> It just stops moving.
>
> *Sayori, Doki Doki Literature Club*
> Dan Salvato

12.1 Future challenges

The work presented here is only the first step toward a physical treatment of sequencing data. Much work remains.

Although the modeling framework is fairly generic, the connections to real data are still obscure. I have operated with "nascent" and "mature" matrices, using counts aligning to intronic and exonic counts. But this binary is questionable, and I raise many potential counterpoints in Sections B.1 and B.2. Ultimately, a comprehensive model should represent transcript elongation *and* splicing *and* imperfect capture *and* ambiguities in assignment *and* stochasticity in all of the above. In the same vein, I have alluded to the construction of more sophisticated models for RNA capture, but have not attempted this. Therefore, although the fits to real data are at least fair, many fundamental questions are still outstanding.

The "feed-forward" systems I have outlined afford fairly simple solution strategies. Explicit regulation does not. Aside from the very brief discussions in Sections 4.3.1 and A.7, I have essentially ignored this key part of biology: the mathematics are intractable, and the (usually protein-based) mechanisms cannot be constrained using (RNA) data. However, further theoretical study is certainly worthwhile.

The solutions I have outlined rely on generating functions, and can, at least in principle, be combined to represent transcription genome-wide, using the toolbox in Section 10.1 to "couple" gene modules. However, in practice, computing, inverting, and discarding the vast majority of enormous n-dimensional arrays is impractical, and new solvers are necessary. It is possible that the methods outlined in Sections 5.3 and 10.3 can be used to this end. However, my feeling is that the current approach

is still far too bespoke and reliant on brute force, and alternative strategies need to be invented.

Although I have used a handful of statistical techniques, the core of the thesis is not about statistics. Instead, it represents foundational work meant to *enable* rigorous investigations by trained statisticians. Although the results thus far may have frustrating limitations, I believe that the general outline of the mechanistic approach provides a more promising foundation for future work than the "data science" methods critiqued at length in Sections 6.1, 8.4, and B.3, and B.4. These critiques, in turn, are still incomplete, and many other methods used in single-cell sequencing data analysis — thermodynamic, landscape, and graph analogies[8], nearest-neighbor graphs, various clustering algorithms — give me pause; their compatibility with single-molecule noise is obscure. However, the comprehensive analysis of these methods is a substantial undertaking pursuing a rapidly moving target.

12.2 Concluding notes

The stochastic worldview offers us a principled way to ask questions of single-cell RNA sequencing data. Even though the conclusions are limited, and do not add up to a grand theory of biology, on having read (or written) this thesis, we are not Goethe's Faust, who "...here, poor fool! with all [his] lore... stand[s], no wiser than before" [306]. We have learned something. We know how to solve certain seemingly imposing equations, considerably reducing the mathematical ingenuity necessary to model biophysical phenomena. We have gained a healthy unease with standard practices. We have learned to think about single-cell technologies in a way that brings them closer to "full communion" with the tradition of transcriptomics.

What do I hope to actually accomplish with this thesis?

In the near term, I have raised doubts about standard analysis procedures (normalization, RNA velocity), and proposed a (flawed, limited, but precedented and principled) alternative. Where this will lead is unclear. Every week, a new velocity, graph analysis, normalization, machine learning method is released. Perhaps the critiques reported here will counteract some of the momentum and lead to more carefully weighed claims, models, and benchmarks. Per Samuel Karlin, modeling lets us "sharpen the questions" about data [289], and the critiques and hypotheses I have outlined here are intended to illustrate the power of this worldview.

In the medium term, I have attempted to draw connections across disciplines and en-

courage researchers in related fields — chemical engineering, physics, fluorescence transcriptomics, finance, machine learning — to seriously consider lending their experience in stochastic modeling to the single-cell RNA sequencing field. Time will tell whether this will lead to the transfer of expertise. A few people in the right places, asking the right questions, may make all the necessary difference.

In the long term, nature is a forest at dawn, veiled in mist; all of research is a stream flowing through this forest; this stream is in flux; some patches of its surface reflect the forest, some are murky, some are opaque, placid and isolated against the current, but for all that no less a part of it; and this work is a drop in the stream, painstakingly made discrete and unified for a brief moment, here clarifying its surroundings, here obscuring, but from here on ultimately dissipating at the will of the stream.

Bibliography

[1] 10x Genomics. Interpreting Intronic and Antisense Reads in 10x Genomics Single Cell Gene Expression Data. Technical Note CG000376, 10x Genomics, August 2021. URL https://www.10xgenomics.com/support/single-cell-gene-expression/documentation/steps/sequencing/interpreting-intronic-and-antisense-reads-in-10-x-genomics-single-cell-gene-expression-data.

[2] Milton Abramowitz and Irene Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. United States National Bureau of Standards, 9 edition, 1970.

[3] Andrew K. Adams, Shelley D. Smith, Dongnhu T. Truong, Erik G. Willcutt, Richard K. Olson, John C. DeFries, Bruce F. Pennington, and Jeffrey R. Gruen. Enrichment of putatively damaging rare variants in the DYX2 locus and the reading-related genes CCDC136 and FLNC. *Human Genetics*, 136 (11-12):1395–1405, November 2017. ISSN 0340-6717, 1432-1203. doi: 10.1007/s00439-017-1838-z. URL http://link.springer.com/10.1007/s00439-017-1838-z.

[4] Constantin Ahlmann-Eltze and Wolfgang Huber. Comparison of transformations for single-cell RNA-seq data. *Nature Methods*, April 2023. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-023-01814-1. URL https://www.nature.com/articles/s41592-023-01814-1.

[5] Jaroslav Albert. Path integral approach to generating functions for multistep post-transcription and post-translation processes and arbitrary initial conditions. *Journal of Mathematical Biology*, 79(6-7):2211–2236, December 2019. ISSN 0303-6812, 1432-1416. doi: 10.1007/s00285-019-01426-4. URL http://link.springer.com/10.1007/s00285-019-01426-4.

[6] Melissa J. Alldred, Karen E. Duff, and Stephen D. Ginsberg. Microarray analysis of CA1 pyramidal neurons in a mouse model of tauopathy reveals progressive synaptic dysfunction. *Neurobiology of Disease*, 45(2):751–762, February 2012. ISSN 09699961. doi: 10.1016/j.nbd.2011.10.022. URL https://linkinghub.elsevier.com/retrieve/pii/S0969996111003548.

[7] Lisa Amrhein. *Stochastic Modeling of Heterogeneous Low-Input Gene Expression: Linking Single-Cell Probability Distributions to Transcription Mechanisms*. PhD Dissertation, Technische Universitat Munchen, Munich, June 2021.

[8] Lisa Amrhein, Kumar Harsha, and Christiane Fuchs. A mechanistic model for the negative binomial distribution of single-cell mRNA counts. Preprint, bioRxiv: 657619, June 2019. URL http://biorxiv.org/lookup/doi/10.1101/657619.

[9] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[10] Tallulah Andrews and Martin Hemberg. False signals induced by single-cell imputation. *F1000Research*, 7:1740, 2019. URL https://f1000research.com/articles/7-1740/v2.

[11] Tallulah S. Andrews, Jawairia Atif, Jeff C. Liu, Catia T. Perciani, Xue-Zhong Ma, Cornelia Thoeni, Michal Slyper, Gökcen Eraslan, Asa Segerstolpe, Justin Manuel, Sai Chung, Erin Winter, Iulia Cirlan, Nicholas Khuu, Sandra Fischer, Orit Rozenblatt-Rosen, Aviv Regev, Ian D. McGilvray, Gary D. Bader, and Sonya A. MacParland. Single-Cell, Single-Nucleus, and Spatial RNA Sequencing of the Human Liver Identifies Cholangiocyte and Mesenchymal Heterogeneity. *Hepatology Communications*, 6(4):821–840, 2022. doi: https://doi.org/10.1002/hep4.1854. URL https://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/hep4.1854.

[12] Tal Ashuach, Daniel A. Reidenbach, Adam Gayoso, and Nir Yosef. PeakVI: A deep generative model for single-cell chromatin accessibility analysis. *Cell Reports Methods*, 2(3):100182, March 2022. ISSN 26672375. doi: 10.1016/j.crmeth.2022.100182. URL https://linkinghub.elsevier.com/retrieve/pii/S2667237522000376.

[13] Lyla Atta, Arpan Sahoo, and Jean Fan. VeloViz: RNA velocity-informed embeddings for visualizing cellular trajectories. *Bioinformatics*, 38(2):391–396, September 2021. doi: 10.1093/bioinformatics/btab653.

[14] Carol Bacchi. The Turn to Problematization: Political Implications of Contrasting Interpretive and Poststructural Adaptations. *Open Journal of Political Science*, 05(01):1–12, 2015. ISSN 2164-0505, 2164-0513. doi: 10.4236/ojps.2015.51001. URL http://www.scirp.org/journal/doi.aspx?DOI=10.4236/ojps.2015.51001.

[15] Ana Badimon, Hayley J. Strasburger, Pinar Ayata, Xinhong Chen, Aditya Nair, Ako Ikegami, Philip Hwang, Andrew T. Chan, Steven M. Graves, Joseph O. Uweru, Carola Ledderose, Munir Gunes Kutlu, Michael A. Wheeler, Anat Kahan, Masago Ishikawa, Ying-Chih Wang, Yong-Hwee E. Loh, Jean X. Jiang, D. James Surmeier, Simon C. Robson, Wolfgang G. Junger, Robert Sebra, Erin S. Calipari, Paul J. Kenny, Ukpong B. Eyo, Marco Colonna, Francisco J. Quintana, Hiroaki Wake, Viviana Gradinaru, and Anne Schaefer. Negative feedback control of neuronal activity by microglia. *Nature*, 586(7829):417–423, October 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2777-8. URL https://www.nature.com/articles/s41586-020-2777-8.

[16] Keren Bahar Halpern, Sivan Tanami, Shanie Landen, Michal Chapal, Liran Szlak, Anat Hutzler, Anna Nizhberg, and Shalev Itzkovitz. Bursty

Gene Expression in the Intact Mammalian Liver. *Molecular Cell*, 58 (1):147–156, April 2015. ISSN 10972765. doi: 10.1016/j.molcel.2015.01.027. URL https://linkinghub.elsevier.com/retrieve/pii/S1097276515000507.

[17] Keren Bahar Halpern, Inbal Caspi, Doron Lemze, Maayan Levy, Shanie Landen, Eran Elinav, Igor Ulitsky, and Shalev Itzkovitz. Nuclear Retention of mRNA in Mammalian Tissues. *Cell Reports*, 13(12):2653–2662, December 2015. ISSN 22111247. doi: 10.1016/j.celrep.2015.11.036. URL https://linkinghub.elsevier.com/retrieve/pii/S2211124715013510.

[18] Douglas H. Baird, Kenneth A. Myers, Mette Mogensen, David Moss, and Peter W. Baas. Distribution of the microtubule-related protein ninein in developing neurons. *Neuropharmacology*, 47(5):677–683, October 2004. ISSN 00283908. doi: 10.1016/j.neuropharm.2004.07.016. URL https://linkinghub.elsevier.com/retrieve/pii/S0028390804002096.

[19] Trygve E. Bakken, Rebecca D. Hodge, Jeremy A. Miller, Zizhen Yao, Thuc Nghi Nguyen, Brian Aevermann, Eliza Barkan, Darren Bertagnolli, Tamara Casper, Nick Dee, Emma Garren, Jeff Goldy, Lucas T. Graybuck, Matthew Kroll, Roger S. Lasken, Kanan Lathia, Sheana Parry, Christine Rimorin, Richard H. Scheuermann, Nicholas J. Schork, Soraya I. Shehata, Michael Tieu, John W. Phillips, Amy Bernard, Kimberly A. Smith, Hongkui Zeng, Ed S. Lein, and Bosiljka Tasic. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLOS ONE*, 13(12):e0209648, December 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0209648. URL https://dx.plos.org/10.1371/journal.pone.0209648.

[20] O. E. Barndorff-Nielsen and J. Schmiegel. A Stochastic Differential Equation Framework for the Timewise Dynamics of Turbulent Velocities. *Theory of Probability & Its Applications*, 52(3):372–388, January 2008. ISSN 0040-585X, 1095-7219. doi: 10.1137/S0040585X9798316X. URL http://epubs.siam.org/doi/10.1137/S0040585X9798316X.

[21] Ole E Barndorff-Nielsen and Neil Shephard. Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in Financial economics. *Journal of the Royal Statistical Society: Series B*, 63:167–241, 2001. doi: 10.1111/1467-9868.00282. URL https://rss.onlinelibrary.wiley.com/doi/10.1111/1467-9868.00282.

[22] Ole E. Barndorff-Nielsen and Neil Shephard. Integrated OU Processes and Non-Gaussian OU-based Stochastic Volatility Models. *Scandinavian Journal of Statistics*, 30(2):277–295, June 2003. ISSN 0303-6898, 1467-9469. doi: 10.1111/1467-9469.00331. URL http://doi.wiley.com/10.1111/1467-9469.00331.

[23] Ole E. Barndorff-Nielsen, Sidney I. Resnick, and Thomas Mikosch, editors. *Lévy Processes*. Birkhäuser Boston, Boston, MA, 2001. ISBN 978-1-4612-6657-0 978-1-4612-0197-7. doi: 10.1007/978-1-4612-0197-7. URL http://link.springer.com/10.1007/978-1-4612-0197-7.

[24] Anthony F. Bartholomay. On the linear birth and death processes of biology as Markoff chains. *The Bulletin of Mathematical Biophysics*, 20(2):97–118, June 1958. ISSN 0007-4985, 1522-9602. doi: 10.1007/BF02477571. URL http://link.springer.com/10.1007/BF02477571.

[25] Anthony F. Bartholomay. Stochastic models for chemical reactions: I. Theory of the unimolecular reaction process. *The Bulletin of Mathematical Biophysics*, 20(3):175–190, September 1958. ISSN 0007-4985, 1522-9602. doi: 10.1007/BF02478297. URL http://link.springer.com/10.1007/BF02478297.

[26] Ayse Bassez, Hanne Vos, Laurien Van Dyck, Giuseppe Floris, Ingrid Arijs, Christine Desmedt, Bram Boeckx, Marlies Vanden Bempt, Ines Nevelsteen, Kathleen Lambein, Kevin Punie, Patrick Neven, Abhishek D. Garg, Hans Wildiers, Junbin Qian, Ann Smeets, and Diether Lambrechts. A single-cell map of intratumoral changes during anti-PD1 treatment of patients with breast cancer. *Nature Medicine*, 27(5):820–832, May 2021. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-021-01323-8. URL http://www.nature.com/articles/s41591-021-01323-8.

[27] Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Control of Transcript Variability in Single Mammalian Cells. *Cell*, 163(7):1596–1610, December 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.11.018. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867415014981.

[28] Casper H. L. Beentjes, Ruben Perez-Carrasco, and Ramon Grima. Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics. *Physical Review E*, 101(3):032403, March 2020. ISSN 2470-0045, 2470-0053. doi: 10.1103/PhysRevE.101.032403. URL https://link.aps.org/doi/10.1103/PhysRevE.101.032403.

[29] Volker Bergen, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, August 2020. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-020-0591-3. URL http://www.nature.com/articles/s41587-020-0591-3.

[30] Volker Bergen, Ruslan A Soldatov, Peter V Kharchenko, and Fabian J Theis. RNA velocity—current challenges and future perspectives. *Molecular Systems Biology*, 17(8), August 2021. ISSN 1744-4292, 1744-4292. doi: 10.15252/msb.202110282. URL https://onlinelibrary.wiley.com/doi/10.15252/msb.202110282.

[31] Pavol Bokes, John R. King, Andrew T. A. Wood, and Matthew Loose. Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. *Journal of Mathematical Biology*, 64(5):829–854, April 2012. ISSN 0303-6812, 1432-1416. doi: 10.1007/s00285-011-0433-5. URL http://link.springer.com/10.1007/s00285-011-0433-5.

[32] A Sina Booeshaghi and Lior Pachter. Normalization of single-cell RNA-seq counts by log(x + 1) or log(1 + x). *Bioinformatics*, 37 (15):2223–2224, August 2021. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btab085. URL https://academic.oup.com/bioinformatics/article/37/15/2223/6155989.

[33] A. Sina Booeshaghi, Zizhen Yao, Cindy van Velthoven, Kimberly Smith, Bosiljka Tasic, Hongkui Zeng, and Lior Pachter. Isoform cell-type specificity in the mouse primary motor cortex. *Nature*, 598(7879):195–199, October 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03969-3. URL https://www.nature.com/articles/s41586-021-03969-3.

[34] A. Sina Booeshaghi, Ingileif B. Hallgrímsdóttir, Angel Gálvez-Merchán, and Lior Pachter. Depth normalization for single-cell genomics count data. Preprint, bioRxiv: 2022.05.06.490859, May 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.05.06.490859.

[35] Jorge L. Borges. *Collected Fictions*. Penguin Classics, 1999. ISBN 978-0-14-028680-9.

[36] Gerard A. Bouland, Ahmed Mahfouz, and Marcel J. T. Reinders. Consequences and opportunities arising due to sparser single-cell RNA-seq datasets. *Genome Biology*, 24(1):86, April 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-02933-w. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-02933-w.

[37] Jérémie Breda, Mihaela Zavolan, and Erik van Nimwegen. Bayesian inference of gene expression states from single-cell RNA-seq data. *Nature Biotechnology*, 39(8):1008–1016, August 2021. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-021-00875-x. URL https://www.nature.com/articles/s41587-021-00875-x.

[38] Kenneth P. Burnham and David Raymond Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York, 2nd ed edition, 2002. ISBN 978-0-387-95364-9. OCLC: ocm48557578.

[39] Xiaodong Cai. Exact stochastic simulation of coupled chemical reactions with delays. *The Journal of Chemical Physics*, 126(12):124108, March 2007. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.2710253. URL http://aip.scitation.org/doi/10.1063/1.2710253.

[40] Giuliana P Calia, Xinyue Chen, Binyamin Zuckerman, and Leor S Weinberger. Comparative analysis between single-cell RNA-seq and single-molecule RNA FISH indicates that the pyrimidine nucleobase idoxuridine (IdU) globally amplifies transcriptional noise. Preprint, bioRxiv: 2023.03.14.532632, March 2023. URL https://www.biorxiv.org/content/10.1101/2023.03.14.532632v1.full.

[41] Robrecht Cannoodt, Wouter Saelens, Louise Deconinck, and Yvan Saeys. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. *Nature Communications*, 12(1):3942, December 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-24152-2. URL http://www.nature.com/articles/s41467-021-24152-2.

[42] 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(9):4682–4692, March 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1910888117. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1910888117.

[43] Jessica Cariboni and Wim Schoutens. Jumps in intensity models: investigating the performance of Ornstein-Uhlenbeck processes in credit risk modeling. *Metrika*, 69(2-3):173–198, March 2009. ISSN 0026-1335, 1435-926X. doi: 10.1007/s00184-008-0213-4. URL http://link.springer.com/10.1007/s00184-008-0213-4.

[44] Maria T. Carilli, Gennady Gorin, Yongin Choi, Tara Chari, and Lior Pachter. Mechanistic modeling with a variational autoencoder for multimodal single-cell RNA sequencing data. Preprint, bioRxiv: 2023.01.13.523995, January 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.01.13.523995.

[45] Lewis Carroll. *Sylvie and Bruno Concluded*. Macmillan and Co., London, 1894.

[46] John T. Chamberlin, Younghee Lee, Gabor T. Marth, and Aaron R. Quinlan. Variable RNA sampling biases mediate concordance of single-cell and nucleus sequencing across cell types. Preprint, bioRxiv: 2022.08.01.502392, August 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.08.01.502392.

[47] Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45):22445–22451, November 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1906995116. URL https://pnas.org/doi/full/10.1073/pnas.1906995116.

[48] David Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, New York, 1987.

[49] Tara Chari, Joeyta Banerjee, and Lior Pachter. The Specious Art of Single-Cell Genomics. Preprint, bioRxiv: 2021.08.25.457696, September 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.08.25.457696.

[50] Mohammed Charrout, Marcel J.T. Reinders, and Ahmed Mahfouz. Untangling biological factors influencing trajectory inference from single cell data. Preprint, bioRxiv: 2020.02.11.942102, February 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.02.11.942102.

[51] Yung-Sung Cheng. Bivariate Lognormal Distribution for Characterizing Asbestos Fiber Aerosols. *Aerosol Science and Technology*, 5(3): 359–368, January 1986. ISSN 0278-6826, 1521-7388. doi: 10.1080/02786828608959100. URL http://www.tandfonline.com/doi/abs/10.1080/02786828608959100.

[52] Yongin Choi, Ruoxin Li, and Gerald Quon. siVAE: interpretable deep generative models for single-cell transcriptomes. *Genome Biology*, 24(1): 29, February 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-02850-y. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-02850-y.

[53] Sandeep Choubey. Nascent RNA kinetics: Transient and steady state behavior of models of transcription. *Physical Review E*, 97(2):022402, 2018. ISSN 2470-0045, 2470-0053. doi: 10.1103/PhysRevE.97.022402.

[54] Sandeep Choubey, Jane Kondev, and Alvaro Sanchez. Deciphering Transcriptional Dynamics In Vivo by Counting Nascent RNA Molecules. *PLOS Computational Biology*, 11(11):e1004345, 2015. ISSN 1553-7358.

[55] Siu Yu A. Chow, Kazuki Nakayama, Tatsuya Osaki, Maki Sugiyama, Maiko Yamada, Hirotaka Takeuchi, and Yoshiho Ikeuchi. Human sensory neurons modulate melanocytes through secretion of RGMB. *Cell Reports*, 40 (12):111366, September 2022. ISSN 22111247. doi: 10.1016/j.celrep.2022.111366. URL https://linkinghub.elsevier.com/retrieve/pii/S2211124722011986.

[56] Michael B. Cole, Davide Risso, Allon Wagner, David DeTomaso, John Ngai, Elizabeth Purdom, Sandrine Dudoit, and Nir Yosef. Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *Cell Systems*, 8(4):315–328.e8, April 2019. ISSN 24054712. doi: 10.1016/j.cels.2019.03.010. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471219300808.

[57] Rama Cont and Peter Tankov. *Financial Modeling with Jump Processes*. Financial Mathematics. Chapman & Hall, 2004.

[58] Shamus M. Cooley, Timothy Hamilton, J. Christian J. Ray, and Eric J. Deeds. A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-Seq data. Preprint, bioRxiv: 689851, September 2020. URL https://www.biorxiv.org/content/10.1101/689851v4.

[59] Adam M Corrigan, Edward Tunnacliffe, Danielle Cannon, and Jonathan R Chubb. A continuum model of transcriptional bursting. *eLife*, 5:e13051, February 2016. ISSN 2050-084X. doi: 10.7554/eLife.13051. URL https://elifesciences.org/articles/13051.

[60] Allison Coté, Chris Coté, Sareh Bayatpour, Heather L Drexler, Katherine A Alexander, Fei Chen, Asmamaw T Wassie, Edward S Boyden, Shelley Berger, L Stirling Churchman, and Arjun Raj. pre-mRNA spatial distributions suggest that splicing can occur post-transcriptionally. Preprint, bioRxiv: 2020.04.06.028092, June 2021. URL https://doi.org/10.1101/2020.04.06.028092.

[61] Charles P. Couturier, Shamini Ayyadhury, Phuong U. Le, Javad Nadaf, Jean Monlong, Gabriele Riva, Redouane Allache, Salma Baig, Xiaohua Yan, Mathieu Bourgey, Changseok Lee, Yu Chang David Wang, V. Wee Yong, Marie-Christine Guiot, Hamed Najafabadi, Bratislav Misic, Jack Antel, Guillaume Bourque, Jiannis Ragoussis, and Kevin Petrecca. Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nature Communications*, 11(1):3406, December 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17186-5. URL http://www.nature.com/articles/s41467-020-17186-5.

[62] John C. Cox, Jonathan E. Ingersoll, and Stephen A. Ross. A Theory of the Term Structure of Interest Rates. *Econometrica*, 53(2):385, March 1985. ISSN 00129682. doi: 10.2307/1911242. URL https://www.jstor.org/stable/1911242?origin=crossref.

[63] Haotian Cui, Hassaan Maan, and Bo Wang. DeepVelo: Deep Learning extends RNA velocity to multi-lineage systems with cell-specific kinetics. Preprint, bioRxiv: 2022.04.03.486877, April 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.04.03.486877.

[64] Bernie J Daigle, Min K Roh, Linda R Petzold, and Jarad Niemi. Accelerated maximum likelihood parameter estimation for stochastic biochemical systems. *BMC Bioinformatics*, 13(1):68, December 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-68. URL http://link.springer.com/10.1186/1471-2105-13-68.

[65] R. D. Dar, B. S. Razooky, A. Singh, T. V. Trimeloni, J. M. McCollum, C. D. Cox, M. L. Simpson, and L. S. Weinberger. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences*, 109(43):17454–17459,

October 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1213530109. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1213530109.

[66] Justine Dattani. *Exact solutions of master equations for the analysis of gene transcription models*. PhD Dissertation, Imperial College London, November 2015.

[67] Justine Dattani and Mauricio Barahona. Stochastic models of gene transcription with upstream drives: exact solution and sample path characterization. *Journal of The Royal Society Interface*, 14(126):20160833, January 2017. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2016.0833. URL https://royalsocietypublishing.org/doi/10.1098/rsif.2016.0833.

[68] Chiara A. De Benedictis, Claudia Haffke, Simone Hagmeyer, Ann Katrin Sauer, and Andreas M. Grabrucker. Expression Analysis of Zinc Transporters in Nervous Tissue Cells Reveals Neuronal and Synaptic Localization of ZIP4. *International Journal of Molecular Sciences*, 22(9):4511, April 2021. ISSN 1422-0067. doi: 10.3390/ijms22094511. URL https://www.mdpi.com/1422-0067/22/9/4511.

[69] Mihails Delmans and Martin Hemberg. Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics*, 17:110, December 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-0944-6. URL http://www.biomedcentral.com/1471-2105/17/110.

[70] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 0035-9246. URL https://www.jstor.org/stable/2984875. Publisher: [Royal Statistical Society, Wiley].

[71] Elena Denisenko, Belinda B. Guo, Matthew Jones, Rui Hou, Leanne de Kock, Timo Lassmann, Daniel Poppe, Olivier Clément, Rebecca K. Simmons, Ryan Lister, and Alistair R. R. Forrest. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biology*, 21(1):130, December 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02048-6. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02048-6.

[72] Ravi V. Desai, Xinyue Chen, Benjamin Martin, Sonali Chaturvedi, Dong Woo Hwang, Weihan Li, Chen Yu, Sheng Ding, Matt Thomson, Robert H. Singer, Robert A. Coleman, Maike M. K. Hansen, and Leor S. Weinberger. A DNA repair pathway can regulate transcriptional noise to promote cell fate transitions. *Science*, 373(6557):eabc6506, August 2021. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abc6506. URL https://www.sciencemag.org/lookup/doi/10.1126/science.abc6506.

[73] M. Deza and Elena Deza. *Encyclopedia of distances*. Springer Verlag, Dordrecht : New York, 2009. ISBN 978-3-642-00233-5 978-3-642-00234-2. OCLC: ocn310400730.

[74] Fangyuan Ding and Michael B. Elowitz. Constitutive splicing and economies of scale in gene expression. *Nature structural & molecular biology*, 26(6): 424–432, June 2019. ISSN 1545-9993. doi: 10.1038/s41594-019-0226-x. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6663491/.

[75] Jiarui Ding, Xian Adiconis, Sean K. Simmons, Monika S. Kowalczyk, Cynthia C. Hession, Nemanja D. Marjanovic, Travis K. Hughes, Marc H. Wadsworth, Tyler Burks, Lan T. Nguyen, John Y. H. Kwon, Boaz Barak, William Ge, Amanda J. Kedaigle, Shaina Carroll, Shuqiang Li, Nir Hacohen, Orit Rozenblatt-Rosen, Alex K. Shalek, Alexandra-Chloé Villani, Aviv Regev, and Joshua Z. Levin. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature Biotechnology*, 38(6):737–746, June 2020. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-020-0465-8. URL https://www.nature.com/articles/s41587-020-0465-8.

[76] Heather L. Drexler, Karine Choquet, and L. Stirling Churchman. Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Molecular Cell*, 77(5):985–998.e8, March 2020. ISSN 10972765. doi: 10.1016/j.molcel.2019.11.017. URL https://linkinghub.elsevier.com/retrieve/pii/S1097276519308652.

[77] Jin-Hong Du, Ming Gao, and Jingshu Wang. Model-based Trajectory Inference for Single-Cell RNA Sequencing Using Deep Learning with a Mixture Prior. Preprint, bioRxiv: 2020.12.26.424452, December 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.12.26.424452.

[78] Jacques Dutka. The early history of the factorial function. *Archive for History of Exact Sciences*, 43(3):225–249, 1991. ISSN 0003-9519, 1432-0657. doi: 10.1007/BF00389433. URL http://link.springer.com/10.1007/BF00389433.

[79] Marcelo R. Ebert and Michael Reissig. *Methods for Partial Differential Equations*. Springer International Publishing, Cham, 2018. ISBN 978-3-319-66455-2 978-3-319-66456-9. doi: 10.1007/978-3-319-66456-9. URL http://link.springer.com/10.1007/978-3-319-66456-9.

[80] Kristján Eldjárn Hjörleifsson, Delaney K. Sullivan, Guillaume Holley, Páll Melsted, and Lior Pachter. Accurate quantification of single-nucleus and single-cell RNA-seq transcripts. Preprint, bioRxiv: 2022.12.02.518832, December 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.12.02.518832.

[81] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic Gene Expression in a Single Cell. *Science*, 297(5584):1183–1186, 2002. doi: 10.1126/science.1070919.

[82] W. L. Ernst, Y. Zhang, J. W. Yoo, S. J. Ernst, and J. L. Noebels. Genetic Enhancement of Thalamocortical Network Activity by Elevating 1G-Mediated Low-Voltage-Activated Calcium Current Induces Pure Absence Epilepsy. *Journal of Neuroscience*, 29(6):1615–1625, February 2009. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.2081-08.2009. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.2081-08.2009.

[83] Felix Famoye. On the bivariate negative binomial regression model. *Journal of Applied Statistics*, 37(6):969–981, June 2010. ISSN 0266-4763, 1360-0532. doi: 10.1080/02664760902984618. URL https://www.tandfonline.com/doi/full/10.1080/02664760902984618.

[84] Alexander Feuerborn and Peter R. Cook. Why the activity of a gene depends on its neighbors. *Trends in Genetics*, 31(9):483–490, September 2015. ISSN 01689525. doi: 10.1016/j.tig.2015.07.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0168952515001298.

[85] Tatiana Filatova, Nikola Popovic, and Ramon Grima. Statistics of Nascent and Mature RNA Fluctuations in a Stochastic Model of Transcriptional Initiation, Elongation, Pausing, and Termination. *Bulletin of Mathematical Biology*, 83(1):3, January 2021. ISSN 0092-8240, 1522-9602. doi: 10.1007/s11538-020-00827-7. URL http://link.springer.com/10.1007/s11538-020-00827-7.

[86] Tatiana Filatova, Nikola Popović, and Ramon Grima. Modulation of nuclear and cytoplasmic mRNA fluctuations by time-dependent stimuli: Analytical distributions. *Mathematical Biosciences*, 347:108828, May 2022. ISSN 00255564. doi: 10.1016/j.mbs.2022.108828. URL https://linkinghub.elsevier.com/retrieve/pii/S0025556422000372.

[87] Stephen J. Fleming, Mark D. Chaffin, Alessandro Arduini, Amer-Denis Akkad, Eric Banks, John C. Marioni, Anthony A. Philippakis, Patrick T. Ellinor, and Mehrtash Babadi. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. Preprint, bioRxiv: 791699, October 2019. URL http://biorxiv.org/lookup/doi/10.1101/791699.

[88] H. Scott Fogler. *Elements of chemical reaction engineering*. Prentice Hall PTR international series in the physical and chemical engineering sciences. Prentice Hall PTR, Upper Saddle River, NJ, 4th ed edition, 2006. ISBN 978-0-13-047394-3. OCLC: ocm56956313.

[89] Nir Friedman, Long Cai, and X. Sunney Xie. Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. *Physical Review Letters*, 97(16):168302, October 2006. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.97.168302. URL https://link.aps.org/doi/10.1103/PhysRevLett.97.168302.

[90] Xiaoming Fu, Heta P. Patel, Stefano Coppola, Libin Xu, Zhixing Cao, Tineke L. Lenstra, and Ramon Grima. Accurate inference of stochastic gene expression from nascent transcript heterogeneity. Preprint, bioRxiv: 2021.11.09.467882, November 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.11.09.467882.

[91] Xiaoming Fu, Heta P Patel, Stefano Coppola, Libin Xu, Zhixing Cao, Tineke L Lenstra, and Ramon Grima. Quantifying how post-transcriptional noise and gene copy number variation bias transcriptional parameter inference from mRNA distributions. *eLife*, 11:e82493, October 2022. ISSN 2050-084X. doi: 10.7554/eLife.82493. URL https://elifesciences.org/articles/82493.

[92] Takashi Fukaya, Bomyi Lim, and Michael Levine. Enhancer Control of Transcriptional Bursting. *Cell*, 166(2):358–368, July 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.05.025. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867416305736.

[93] Jie Gao, Yue Ma, Hua-Lin Fu, Qian Luo, Zhen Wang, Yu-Huan Xiao, Hao Yang, Da-Xiang Cui, and Wei-Lin Jin. Non-catalytic roles for TET1 protein negatively regulating neuronal differentiation through srGAP3 in neuroblastoma cells. *Protein & Cell*, 7(5):351–361, May 2016. ISSN 1674-8018. doi: 10.1007/s13238-016-0267-4. URL https://doi.org/10.1007/s13238-016-0267-4.

[94] C. W. Gardiner and S. Chaturvedi. The poisson representation. I. A new technique for chemical master equations. *Journal of Statistical Physics*, 17(6):429–468, December 1977. ISSN 0022-4715, 1572-9613. doi: 10.1007/BF01014349. URL http://link.springer.com/10.1007/BF01014349.

[95] Crispin Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. Springer, third edition, 2004.

[96] Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L. Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3):272–282, March 2021. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-020-01050-x. URL http://www.nature.com/articles/s41592-020-01050-x.

[97] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard

Mehlman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mariano Gabitto, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Vitalii Kleshchevnikov, Carlos Talavera-López, Lior Pachter, Fabian J. Theis, Aaron Streets, Michael I. Jordan, Jeffrey Regier, and Nir Yosef. A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2):163–166, February 2022. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-021-01206-w. URL https://www.nature.com/articles/s41587-021-01206-w.

[98] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, December 1976. ISSN 00219991. doi: 10.1016/0021-9991(76)90041-3. URL https://linkinghub.elsevier.com/retrieve/pii/0021999176900413.

[99] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, December 1977. ISSN 0022-3654, 1541-5740. doi: 10.1021/j100540a008. URL https://pubs.acs.org/doi/abs/10.1021/j100540a008.

[100] Daniel T. Gillespie. The chemical Langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, July 2000. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.481811. URL http://aip.scitation.org/doi/10.1063/1.481811.

[101] Roy J. Glauber. Time-Dependent Statistics of the Ising Model. *Journal of Mathematical Physics*, 4(2):294–307, February 1963. ISSN 0022-2488, 1089-7658. doi: 10.1063/1.1703954. URL http://aip.scitation.org/doi/10.1063/1.1703954.

[102] Ido Golding, Johan Paulsson, Scott M. Zawilski, and Edward C. Cox. Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell*, 123 (6):1025–1036, December 2005. ISSN 00928674. doi: 10.1016/j.cell.2005.09.031. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867405010378.

[103] Gennady Gorin and Lior Pachter. Intrinsic and extrinsic noise are distinguishable in a synthesis – export – degradation model of mRNA production. Preprint, bioRxiv: 2020.09.25.312868, September 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.09.25.312868.

[104] Gennady Gorin and Lior Pachter. Special function methods for bursty models of transcription. *Physical Review E*, 102(2):022409, August 2020. ISSN 2470-0045, 2470-0053. doi: 10.1103/PhysRevE.102.022409. URL https://link.aps.org/doi/10.1103/PhysRevE.102.022409.

[105] Gennady Gorin and Lior Pachter. Modeling bursty transcription and splicing with the chemical master equation. *Biophysical Journal*, 121(6):1056–1069, February 2022. doi: 10.1016/j.bpj.2022.02.004. URL https://www.cell.com/biophysj/fulltext/S0006-3495(22)00104-7.

[106] Gennady Gorin and Lior Pachter. Distinguishing biophysical stochasticity from technical noise in single-cell RNA sequencing using *Monod*. Preprint, bioRxiv: 2022.06.11.495771, April 2023. URL https://www.biorxiv.org/content/10.1101/2022.06.11.495771v2.

[107] Gennady Gorin and Lior Pachter. Length biases in single-cell RNA sequencing of pre-mRNA. *Biophysical Reports*, 3(1):100097, March 2023. ISSN 26670747. doi: 10.1016/j.bpr.2022.100097. URL https://linkinghub.elsevier.com/retrieve/pii/S2667074722000544.

[108] Gennady Gorin and Lior Pachter. The telegraph process is not a subordinator. Preprint, bioRxiv: 2023.01.17.524309, January 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.01.17.524309.

[109] Gennady Gorin, Valentine Svensson, and Lior Pachter. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biology*, 21:39, February 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-1945-3. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1945-3.

[110] Gennady Gorin, Mengyu Wang, Ido Golding, and Heng Xu. Stochastic simulation and statistical inference platform for visualization and estimation of transcriptional kinetics. *PLOS ONE*, 15(3):e0230736, March 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0230736. URL https://dx.plos.org/10.1371/journal.pone.0230736.

[111] Gennady Gorin, Maria Carilli, Tara Chari, and Lior Pachter. Spectral neural approximations for models of transcriptional dynamics. Preprint, bioRxiv: 2022.06.16.496448, June 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.06.16.496448.

[112] Gennady Gorin, Meichen Fang, Tara Chari, and Lior Pachter. RNA velocity unraveled. *PLOS Computational Biology*, 18(9):e1010492, September 2022. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010492.

[113] Gennady Gorin, John J. Vastola, Meichen Fang, and Lior Pachter. Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments. *Nature Communications*, 13(1):7620, December 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34857-7. URL https://www.nature.com/articles/s41467-022-34857-7.

[114] Gennady Gorin, Shawn Yoshida, and Lior Pachter. Transient and delay chemical master equations. Preprint, bioRxiv: 2022.10.17.512599, October 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.10.17.512599.

[115] Gennady Gorin, John J. Vastola, and Lior Pachter. Stochastic systems biology of the cell using single-cell genomics data. Preprint, bioRxiv: in prep., April 2023. URL Inprep.

[116] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640, June 2014. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.2930. URL http://www.nature.com/articles/nmeth.2930.

[117] Jingtao Guo, Edward J. Grow, Hana Mlcochova, Geoffrey J. Maher, Cecilia Lindskog, Xichen Nie, Yixuan Guo, Yodai Takei, Jina Yun, Long Cai, Robin Kim, Douglas T. Carrell, Anne Goriely, James M. Hotaling, and Bradley R. Cairns. The adult human testis transcriptional cell atlas. *Cell Research*, 28(12):1141–1157, December 2018. ISSN 1001-0602, 1748-7838. doi: 10.1038/s41422-018-0099-2. URL http://www.nature.com/articles/s41422-018-0099-2.

[118] Ankit Gupta, Jan Mikelson, and Mustafa Khammash. A finite state projection algorithm for the stationary solution of the chemical master equation. *The Journal of Chemical Physics*, 147(15):154101, October 2017. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.5006484. URL http://aip.scitation.org/doi/10.1063/1.5006484.

[119] Anushka Gupta, Farnaz Shamsi, Nicolas Altemose, Gabriel F. Dorlhiac, Aaron M. Cypess, Andrew P. White, Nir Yosef, Mary Elizabeth Patti, Yu-Hua Tseng, and Aaron Streets. Characterization of transcript enrichment and detection bias in single-nucleus RNA-seq for mapping of distinct human adipocyte lineages. *Genome Research*, 32(2):242–257, February 2022. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.275509.121. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.275509.121.

[120] R. Gupta, D. Cerletti, G. Gut, A. Oxenius, and M. Claassen. Cytopath: Simulation based inference of differentiation trajectories from RNA velocity fields. Preprint, bioRxiv: 2020.12.21.423801, December 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.12.21.423801.

[121] Mariana Gómez-Schiavon, Liang-Fu Chen, Anne E. West, and Nicolas E. Buchler. BayFish: Bayesian inference of transcription dynamics from population snapshots of single-molecule RNA FISH in single cells. *Genome Biology*, 18(1):164, December 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1297-9. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1297-9.

[122] Brian J Haas, Melissa Chin, Chad Nusbaum, Bruce W Birren, and Jonathan Livny. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics*, 13(1):734, December 2012. ISSN 1471-2164. doi: 10.1186/1471-2164-13-734. URL https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-734.

[123] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20:296, December 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1874-1. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1874-1.

[124] Lucy Ham, Rowan D. Brackston, and Michael P. H. Stumpf. Extrinsic Noise and Heavy-Tailed Laws in Gene Expression. *Physical Review Letters*, 124 (10):108101, March 2020. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.124.108101. URL https://link.aps.org/doi/10.1103/PhysRevLett.124.108101.

[125] Lucy Ham, David Schnoerr, Rowan D. Brackston, and Michael P. H. Stumpf. Exactly solvable models of stochastic gene expression. *The Journal of Chemical Physics*, 152(14):144106, April 2020. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.5143540. URL http://aip.scitation.org/doi/10.1063/1.5143540.

[126] Xiaoping Han, Ziming Zhou, Lijiang Fei, Huiyu Sun, Renying Wang, Yao Chen, Haide Chen, Jingjing Wang, Huanna Tang, Wenhao Ge, Yincong Zhou, Fang Ye, Mengmeng Jiang, Junqing Wu, Yanyu Xiao, Xiaoning Jia, Tingyue Zhang, Xiaojie Ma, Qi Zhang, Xueli Bai, Shujing Lai, Chengxuan Yu, Lijun Zhu, Rui Lin, Yuchi Gao, Min Wang, Yiqing Wu, Jianming Zhang, Renya Zhan, Saiyong Zhu, Hailan Hu, Changchun Wang, Ming Chen, He Huang, Tingbo Liang, Jianghua Chen, Weilin Wang, Dan Zhang, and Guoji Guo. Construction of a human cell landscape at single-cell level. *Nature*, 581(7808):303–309, May 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2157-4. URL http://www.nature.com/articles/s41586-020-2157-4.

[127] Maike M.K. Hansen, Ravi V. Desai, Michael L. Simpson, and Leor S. Weinberger. Cytoplasmic Amplification of Transcriptional Noise Generates Substantial Cell-to-Cell Variability. *Cell Systems*, 7(4):384–397.e6, October 2018. ISSN 24054712. doi: 10.1016/j.cels.2018.08.002. URL https://linkinghub.elsevier.com/retrieve/pii/S240547121830317X.

[128] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish,

Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021. ISSN 00928674. doi: 10.1016/j.cell.2021.04.048. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867421005833.

[129] Ashraful Haque, Jessica Engel, Sarah A. Teichmann, and Tapio Lönnberg. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(1):75, December 2017. ISSN 1756-994X. doi: 10.1186/s13073-017-0467-4. URL http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0467-4.

[130] Akdes Serin Harmanci, Arif O Harmanci, Xiaobo Zhou, Benjamin Deneen, Ganesh Rao, Tiemo Klisch, and Akash Patel. scRegulocity: Detection of local RNA velocity patterns in embeddings of single cell RNA-Seq data. Preprint, bioRxiv: 2021.06.01.446674, June 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.06.01.446674.

[131] Dongze He, Charlotte Soneson, and Rob Patro. Understanding and evaluating ambiguity in single-cell and single-nucleus RNA-sequencing. Preprint, bioRxiv: 2023.01.04.522742, January 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.01.04.522742.

[132] S. P. Heims. Master Equation for Ising Model. *Physical Review*, 138 (2A):A587–A590, April 1965. ISSN 0031-899X. doi: 10.1103/PhysRev.138.A587. URL https://link.aps.org/doi/10.1103/PhysRev.138.A587.

[133] Cody N. Heiser, Victoria M. Wang, Bob Chen, Jacob J. Hughey, and Ken S. Lau. Automated quality control and cell identification of droplet-based single-cell data using dropkick. *Genome Research*, 31(10):1742–1752, October 2021. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.271908.120. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.271908.120.

[134] Lukas Heumos, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, Daniel C. Strobl, Juan Henao, Fabiola Curion, Single-cell Best Practices Consortium, Hananeh Aliee, Meshal Ansari, Pau Badia-i Mompel, Maren Büttner, Emma Dann, Daniel Dimitrov, Leander Dony, Amit Frishberg, Dongze He, Soroor Hediyeh-zadeh, Leon Hetzel, Ignacio L. Ibarra, Matthew G. Jones, Mohammad Lotfollahi, Laura D. Martens, Christian L. Müller, Mor Nitzan, Johannes Ostner, Giovanni Palla, Rob Patro, Zoe Piran, Ciro Ramírez-Suástegui, Julio Saez-Rodriguez, Hirak Sarkar, Benjamin Schubert, Lisa Sikkema, Avi Srivastava, Jovan Tanevski, Isaac Virshup, Philipp Weiler, Herbert B. Schiller, and Fabian J. Theis. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, March 2023. ISSN 1471-0056, 1471-

0064. doi: 10.1038/s41576-023-00586-w. URL https://www.nature.com/articles/s41576-023-00586-w.

[135] Brian L. Hie, Kevin K. Yang, and Peter S. Kim. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Systems*, 13(4):274–285.e6, April 2022. ISSN 24054712. doi: 10.1016/j.cels.2022.01.003. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471222000382.

[136] A. Hilfinger and J. Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences*, 108(29):12167–12172, July 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1018832108. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1018832108.

[137] Andreas Hilfinger, Thomas M. Norman, Glenn Vinnicombe, and Johan Paulsson. Constraints on Fluctuations in Sparsely Characterized Biological Systems. *Physical Review Letters*, 116(5):058101, February 2016. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.116.058101. URL https://link.aps.org/doi/10.1103/PhysRevLett.116.058101.

[138] Andreas Hilfinger, Thomas M. Norman, and Johan Paulsson. Exploiting Natural Fluctuations to Identify Kinetic Mechanisms in Sparsely Characterized Systems. *Cell Systems*, 2(4):251–259, April 2016. ISSN 24054712. doi: 10.1016/j.cels.2016.04.002. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471216301107.

[139] Ariel A. Hippen, Matias M. Falco, Lukas M. Weber, Erdogan Pekcan Erkan, Kaiyang Zhang, Jennifer Anne Doherty, Anna Vähärautio, Casey S. Greene, and Stephanie C. Hicks. miQC: An adaptive probabilistic framework for quality control of single-cell RNA-sequencing data. *PLOS Computational Biology*, 17(8):e1009290, August 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1009290. URL https://dx.plos.org/10.1371/journal.pcbi.1009290.

[140] Bo Hu, David A. Kessler, Wouter-Jan Rappel, and Herbert Levine. How input fluctuations reshape the dynamics of a biological switching system. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 86(6 Pt 1):061910, December 2012. ISSN 1539-3755. doi: 10.1103/PhysRevE.86.061910. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5836738/.

[141] Lifang Huang, Zhanjiang Yuan, Peijiang Liu, and Tianshou Zhou. Feedback-induced counterintuitive correlations of gene expression noise with bursting kinetics. *Physical Review E*, 90(5):052702, November 2014. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.90.052702. URL https://link.aps.org/doi/10.1103/PhysRevE.90.052702.

[142] Virginie Imbault, Chiara Dionisi, Gilles Naeije, David Communi, and Massimo Pandolfo. Cerebrospinal Fluid Proteomics in Friedreich Ataxia Reveals Markers of Neurodegeneration and Neuroinflammation. *Frontiers in Neuroscience*, 16:885313, July 2022. ISSN 1662-4548. doi: 10.3389/fnins.2022.885313. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9326443/.

[143] Srividya Iyer-Biswas and C. Jayaprakash. Mixed Poisson distributions in exact solutions of stochastic auto-regulation models. *Physical Review E*, 90 (5):052712, November 2014. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.90.052712. URL http://arxiv.org/abs/1110.2804. arXiv: 1110.2804.

[144] Srividya Iyer-Biswas, F. Hayot, and C. Jayaprakash. Stochasticity of gene products from transcriptional pulsing. *Physical Review E*, 79(3):031911, March 2009. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.79.031911. URL https://link.aps.org/doi/10.1103/PhysRevE.79.031911.

[145] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356, June 1961. doi: 10.1016/S0022-2836(61)80072-7. URL https://www.sciencedirect.com/science/article/pii/S0022283661800727.

[146] Tobias Jahnke and Wilhelm Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54:1–26, September 2006. ISSN 0303-6812, 1432-1416. doi: 10.1007/s00285-006-0034-x. URL http://link.springer.com/10.1007/s00285-006-0034-x.

[147] Selina Jansky, Ashwini Kumar Sharma, Verena Körber, Andrés Quintero, Umut H. Toprak, Elisa M. Wecht, Moritz Gartlgruber, Alessandro Greco, Elad Chomsky, Thomas G. P. Grünewald, Kai-Oliver Henrich, Amos Tanay, Carl Herrmann, Thomas Höfer, and Frank Westermann. Single-cell transcriptomic analyses provide insights into the developmental origins of neuroblastoma. *Nature Genetics*, 53(5):683–693, May 2021. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-021-00806-1. URL http://www.nature.com/articles/s41588-021-00806-1.

[148] Garrett Jenkinson, Jordi Abante, Andrew P. Feinberg, and John Goutsias. An information-theoretic approach to the modeling and analysis of whole-genome bisulfite sequencing data. *BMC Bioinformatics*, 19(1):87, December 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2086-5. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2086-5.

[149] Chen Jia. Kinetic Foundation of the Zero-Inflated Negative Binomial Model for Single-Cell RNA Sequencing Data. *SIAM Journal on Applied Mathematics*, 80(3):1336–1355, January 2020. ISSN 0036-1399, 1095-712X. doi: 10.1137/19M1253198. URL https://epubs.siam.org/doi/10.1137/19M1253198.

[150] 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(1):105746, January 2023. ISSN 25890042. doi: 10.1016/j.isci.2022.105746. URL https://linkinghub.elsevier.com/retrieve/pii/S2589004222020193.

[151] Qingchao Jiang, Xiaoming Fu, Shifu Yan, Runlai Li, Wenli Du, Zhixing Cao, Feng Qian, and Ramon Grima. Neural network aided approximation and parameter inference of non-Markovian models of gene expression. *Nature Communications*, 12(1):2618, December 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22919-1. URL http://www.nature.com/articles/s41467-021-22919-1.

[152] Jing Jin, Priyadarshini Ravindran, Danila Di Meo, and Andreas W. Püschel. Igf1R/InsR function is required for axon extension and corpus callosum formation. *PLOS ONE*, 14(7):e0219362, July 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0219362. URL https://dx.plos.org/10.1371/journal.pone.0219362.

[153] Fritz John. *Partial Differential Equations*. Springer US, New York, NY, 1978. ISBN 978-1-4684-0059-5 978-1-4684-0061-8. URL http://public.eblib.com/choice/publicfullrecord.aspx?p=3082466. OCLC: 859156366.

[154] Norman Lloyd Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous univariate distributions, Vol. 1*. Wiley series in probability and mathematical statistics. Wiley, New York, 2nd ed edition, 1994. ISBN 978-0-471-58495-7 978-0-471-58494-0.

[155] Norman Lloyd Johnson, Adrienne W. Kemp, and Samuel Kotz. *Univariate discrete distributions*. Wiley, Hoboken, N.J, 3rd ed edition, 2005. ISBN 978-0-471-27246-5.

[156] Euan Joly-Smith, Zitong Jerry Wang, and Andreas Hilfinger. Inferring gene regulation dynamics from static snapshots of gene expression variability. *Physical Review E*, 104(4):044406, October 2021. ISSN 2470-0045, 2470-0053. doi: 10.1103/PhysRevE.104.044406. URL https://link.aps.org/doi/10.1103/PhysRevE.104.044406.

[157] Dimitris Karlis and Evdokia Xekalaki. Mixed Poisson Distributions. *International Statistical Review / Revue Internationale de Statistique*, 73(1):

35–58, 2005. ISSN 0306-7734. URL http://www.jstor.org/stable/25472639.

[158] O Kessler, Y Jiang, and L A Chasin. Order of intron removal during splicing of endogenous adenine phosphoribosyltransferase and dihydrofolate reductase pre-mRNA. *Molecular and Cellular Biology*, 13(10):6211–6222, October 1993. ISSN 0270-7306, 1098-5549. doi: 10.1128/MCB.13.10.6211. URL http://mcb.asm.org/lookup/doi/10.1128/MCB.13.10.6211.

[159] Jong Kim and John C Marioni. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biology*, 14:R7, 2013. ISSN 1465-6906. doi: 10.1186/gb-2013-14-1-r7. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-1-r7.

[160] Tae Hyun Kim, Xiang Zhou, and Mengjie Chen. Demystifying "drop-outs" in single-cell UMI data. *Genome Biology*, 21:196, December 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02096-y. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02096-y.

[161] Alena Klindziuk and Anatoly B. Kolomeisky. Understanding the molecular mechanisms of transcriptional bursting. *Physical Chemistry Chemical Physics*, page 10.1039.D1CP03665C, 2021. ISSN 1463-9076, 1463-9084. doi: 10.1039/D1CP03665C. URL http://xlink.rsc.org/?DOI=D1CP03665C.

[162] Dmitry Kobak and Philipp Berens. The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, 10(1):5416, December 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-13056-x. URL http://www.nature.com/articles/s41467-019-13056-x.

[163] Colin Koopman. Problematization. In Leonard Lawlor and John Nale, editors, *The Cambridge Foucault Lexicon*, pages 399–403. Cambridge University Press, 1 edition, April 2014. ISBN 978-1-139-02230-9 978-0-521-11921-4. doi: 10.1017/CBO9781139022309.070. URL https://www.cambridge.org/core/product/identifier/9781139022309%23c11921-68-1/type/book_part.

[164] Renata Kowara, Michel Ménard, Leslie Brown, and Balu Chakravarthy. Co-localization and interaction of DPYSL3 and GAP43 in primary cortical neurons. *Biochemical and Biophysical Research Communications*, 363(1):190–193, November 2007. ISSN 0006291X. doi: 10.1016/j.bbrc.2007.08.163. URL https://linkinghub.elsevier.com/retrieve/pii/S0006291X07018840.

[165] XL. Kuang, XM. Zhao, HF. Xu, YY. Shi, JB. Deng, and GT. Sun. Spatio-temporal expression of a novel neuron-derived neurotrophic factor

(ndnf) in mouse brains during development. *BMC Neuroscience*, 11:137, 2010. doi: 10.1186/1471-2202-11-137. URL https://bmcneurosci.biomedcentral.com/articles/10.1186/1471-2202-11-137.

[166] Niraj Kumar, Thierry Platini, and Rahul V. Kulkarni. Exact Distributions for Stochastic Gene Expression Models with Bursting and Feedback. *Physical Review Letters*, 113(26):268105, December 2014. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.113.268105. URL https://link.aps.org/doi/10.1103/PhysRevLett.113.268105.

[167] Juan Kuntz, Philipp Thomas, Guy-Bart Stan, and Mauricio Barahona. The Exit Time Finite State Projection Scheme: Bounding Exit Distributions and Occupation Measures of Continuous-Time Markov Chains. *SIAM Journal on Scientific Computing*, 41(2):A748–A769, January 2019. ISSN 1064-8275, 1095-7197. doi: 10.1137/18M1168261. URL https://epubs.siam.org/doi/10.1137/18M1168261.

[168] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, August 2018. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-018-0414-6. URL http://www.nature.com/articles/s41586-018-0414-6.

[169] Blue B. Lake, Simone Codeluppi, Yun C. Yung, Derek Gao, Jerold Chun, Peter V. Kharchenko, Sten Linnarsson, and Kun Zhang. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Scientific Reports*, 7(1):6031, July 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-04426-w. URL https://www.nature.com/articles/s41598-017-04426-w.

[170] Nicholas C. Lammers, Yang Joon Kim, Jiaxi Zhao, and Hernan G. Garcia. A matter of time: Using dynamics and theory to uncover mechanisms of transcriptional bursting. *Current Opinion in Cell Biology*, 67:147–157, December 2020. ISSN 09550674. doi: 10.1016/j.ceb.2020.08.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0955067420300971.

[171] Marius Lange, Volker Bergen, Michal Klein, Manu Setty, Bernhard Reuter, Mostafa Bakhti, Heiko Lickert, Meshal Ansari, Janine Schniering, Herbert B. Schiller, Dana Pe'er, and Fabian J. Theis. CellRank for directed single-cell fate mapping. *Nature Methods*, 19(2):159–170, February 2022. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-021-01346-6. URL https://www.nature.com/articles/s41592-021-01346-6.

[172] Anton J. M. Larsson, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R. Faridani, Björn Reinius, Asa Segerstolpe, Chloe M. Rivera, Bing Ren, and Rickard Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254, January 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-018-0836-1. URL http://www.nature.com/articles/s41586-018-0836-1.

[173] Hugo Lavenant, Stephen Zhang, Young-Heon Kim, and Geoffrey Schiebinger. Towards a mathematical theory of trajectory inference. Preprint, arXiv: 2102.09204, February 2021. URL http://arxiv.org/abs/2102.09204.

[174] Sungwoo Lee, Eijiro Nakamura, Haifeng Yang, Wenyi Wei, Michelle S. Linggi, Mini P. Sajan, Robert V. Farese, Robert S. Freeman, Bruce D. Carter, William G. Kaelin, and Susanne Schlisio. Neuronal apoptosis linked to EglN3 prolyl hydroxylase and familial pheochromocytoma genes: Developmental culling and cancer. *Cancer Cell*, 8(2):155–167, August 2005. ISSN 15356108. doi: 10.1016/j.ccr.2005.06.015. URL https://linkinghub.elsevier.com/retrieve/pii/S1535610805002242.

[175] Chen Li, Maria Virgilio, Kathleen L. Collins, and Joshua D. Welch. Single-cell multi-omic velocity infers dynamic and decoupled gene regulation. Preprint, bioRxiv: 2021.12.13.472472, December 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.12.13.472472.

[176] Fang-Zhen Li, Zhi-E Liu, Xiu-Yuan Li, Li-Mei Bu, Hong-Xia Bu, Hui Liu, and Cai-Ming Zhang. Chromatin 3D structure reconstruction with consideration of adjacency relationship among genomic loci. *BMC Bioinformatics*, 21(1):272, December 2020. ISSN 1471-2105. doi: 10.1186/s12859-020-03612-4. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03612-4.

[177] Fengrui Li, Xiaofei Tian, Yishu Zhou, Lanhui Zhu, Baojie Wang, Mei Ding, and Hao Pang. Dysregulated expression of secretogranin III is involved in neurotoxin-induced dopaminergic neuron apoptosis. *Journal of Neuroscience Research*, 90(12):2237–2246, December 2012. ISSN 03604012. doi: 10.1002/jnr.23121. URL https://onlinelibrary.wiley.com/doi/10.1002/jnr.23121.

[178] Tiejun Li. On the Mathematics of RNA Velocity I: Theoretical Analysis. *CSIAM Transactions on Applied Mathematics*, 2(1):1–55, June 2021. ISSN 2708-0560, 2708-0579. doi: 10.4208/csiam-am.SO-2020-0001. URL http://global-sci.org/intro/article_detail/csiam-am/18653.html.

[179] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier Neural Operator for Parametric Partial Differential Equations. Preprint, arXiv: 2010.08895, May 2021. URL http://arxiv.org/abs/2010.08895.

[180] Xiang Lin, Tian Tian, Zhi Wei, and Hakon Hakonarson. Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nature Communications*, 13(1):7705, December 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-35031-9. URL https://www.nature.com/articles/s41467-022-35031-9.

[181] Zhixiang Lin, Mahdi Zamanighomi, Timothy Daley, Shining Ma, and Wing Hung Wong. Model-Based Approach to the Joint Analysis of Single-Cell Data on Chromatin Accessibility and Gene Expression. *Statistical Science*, 35(1), February 2020. ISSN 0883-4237. doi: 10.1214/19-STS714. URL https://projecteuclid.org/journals/statistical-science/volume-35/issue-1/Model-Based-Approach-to-the-Joint-Analysis-of-Single-Cell/10.1214/19-STS714.full.

[182] Monika Litviňuková, Carlos Talavera-López, Henrike Maatz, Daniel Reichart, Catherine L. Worth, Eric L. Lindberg, Masatoshi Kanda, Krzysztof Polanski, Matthias Heinig, Michael Lee, Emily R. Nadelmann, Kenny Roberts, Liz Tuck, Eirini S. Fasouli, Daniel M. DeLaughter, Barbara McDonough, Hiroko Wakimoto, Joshua M. Gorham, Sara Samari, Krishnaa T. Mahbubani, Kourosh Saeb-Parsy, Giannino Patone, Joseph J. Boyle, Hongbo Zhang, Hao Zhang, Anissa Viveiros, Gavin Y. Oudit, Omer Ali Bayraktar, J. G. Seidman, Christine E. Seidman, Michela Noseda, Norbert Hubner, and Sarah A. Teichmann. Cells of the adult human heart. *Nature*, 588(7838):466–472, December 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2797-4. URL https://www.nature.com/articles/s41586-020-2797-4.

[183] Ruishan Liu, Angela Oliveira Pisco, Emelie Braun, Sten Linnarsson, and James Zou. Dynamical Systems Model of RNA Velocity Improves Inference of Single-cell Trajectory, Pseudo-time and Gene Regulation. *Journal of Molecular Biology*, 434(15):167606, August 2022. ISSN 00222836. doi: 10.1016/j.jmb.2022.167606. URL https://linkinghub.elsevier.com/retrieve/pii/S0022283622001863.

[184] Manyuan Long and Michael Deutsch. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Research*, 27(15):3219–3228, August 1999. ISSN 1362-4962, 0305-1048. doi: 10.1093/nar/27.15.3219. URL https://academic.oup.com/nar/article/27/15/3219/2549228.

[185] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-018-0229-2. URL http://www.nature.com/articles/s41592-018-0229-2.

[186] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome*

Biology, 15(12):550, December 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

[187] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, June 2019. ISSN 1744-4292, 1744-4292, 1744-4292. doi: 10.15252/msb.20188746. URL http://msb.embopress.org/lookup/doi/10.15252/msb.20188746.

[188] Yudell L Luke. *The Special Functions and Their Approximations, Vol 1*. Academic Press, London ; New York, 1969.

[189] Gudrun Lutsch, Roland Vetter, Ulrike Offhauss, Martin Wieske, Hermann-Josef Gröne, Roman Klemenz, Ingolf Schimke, Joachim Stahl, and Rainer Benndorf. Abundance and Location of the Small Heat Shock Proteins HSP25 and aB-Crystallin in Rat and Human Heart. *Circulation*, 96 (10):3466–3476, 1997. doi: 10.1161/01.CIR.96.10.3466. URL https://www.ahajournals.org/doi/abs/10.1161/01.CIR.96.10.3466.

[190] Martin Ian Paguio Malgapo. *Structure and function of the palmitoyltransferase dhhc20 and the acyl coa hydrolase mblac2*. PhD Dissertation, Cornell, Ithaca, NY, December 2019. URL https://ecommons.cornell.edu/handle/1813/70073.

[191] Aanchal Malhotra, Samarendra Das, and Shesh N. Rai. Analysis of Single-Cell RNA-Sequencing Data: A Step-by-Step Guide. *BioMedInformatics*, 2(1):43–61, December 2021. ISSN 2673-7426. doi: 10.3390/biomedinformatics2010003. URL https://www.mdpi.com/2673-7426/2/1/3.

[192] Valérie Marot-Lassauzaie, Brigitte Joanne Bouman, Fearghal Declan Donaghy, Yasmin Demerdash, Marieke Alida Gertruda Essers, and Laleh Haghverdi. Towards reliable quantification of cell state velocities. *PLOS Computational Biology*, 18(9):e1010031, September 2022. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1010031. URL https://dx.plos.org/10.1371/journal.pcbi.1010031.

[193] S. Mauch and M. Stalzer. An efficient method for computing steady state solutions with Gillespie's direct method. *The Journal of Chemical Physics*, 133(14):144108, October 2010. ISSN 0021-9606. doi: 10.1063/1.3489354. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2973983/.

[194] Maxime Mazille, Katarzyna Buczak, Peter Scheiffele, and Oriane Mauger. Stimulus-specific remodeling of the neuronal transcriptome through nuclear intron-retaining transcripts. *The EMBO Journal*, 41(21):e110192, 2022. doi: https://doi.org/10.15252/embj.2021110192. URL https://www.embopress.org/doi/abs/10.15252/embj.2021110192.

[195] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint, arXiv: 1802.03426v2, December 2018. URL http://arxiv.org/abs/1802.03426. arXiv: 1802.03426.

[196] Páll Melsted, Vasilis Ntranos, and Lior Pachter. The barcode, UMI, set format and BUStools. *Bioinformatics*, page btz279, 2019. doi: 10.1093/bioinformatics/btz279.

[197] Páll Melsted, A. Sina Booeshaghi, Lauren Liu, Fan Gao, Lambda Lu, Kyung Hoi Min, Eduardo da Veiga Beltrame, Kristján Eldjárn Hjörleifsson, Jase Gehring, and Lior Pachter. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature Biotechnology*, 39(7):813–818, July 2021. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-021-00870-2. URL http://www.nature.com/articles/s41587-021-00870-2.

[198] Eleni P. Mimitou, Caleb A. Lareau, Kelvin Y. Chen, Andre L. Zorzetto-Fernandes, Yuhan Hao, Yusuke Takeshima, Wendy Luo, Tse-Shun Huang, Bertrand Z. Yeung, Efthymia Papalexi, Pratiksha I. Thakore, Tatsuya Kibayashi, James Badger Wing, Mayu Hata, Rahul Satija, Kristopher L. Nazor, Shimon Sakaguchi, Leif S. Ludwig, Vijay G. Sankaran, Aviv Regev, and Peter Smibert. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature Biotechnology*, 39(10):1246–1258, October 2021. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-021-00927-2. URL https://www.nature.com/articles/s41587-021-00927-2.

[199] Qianxing Mo and Faming Liang. Bayesian Modeling of ChIP-chip Data Through a High-Order Ising Model. *Biometrics*, 66(4):1284–1294, 2010. ISSN 0006-341X. URL https://www.jstor.org/stable/40962526. Publisher: [Wiley, International Biometric Society].

[200] Elliott W. Montroll. On Coupled Rate Equations with Quadratic Nonlinearities. *Proceedings of the National Academy of Sciences of the United States of America*, 69(9):2532–2536, 1972. ISSN 0027-8424. URL https://www.jstor.org/stable/61810. Publisher: National Academy of Sciences.

[201] Muir Morrison, Manuel Razo-Mejia, and Rob Phillips. Reconciling kinetic and thermodynamic models of bacterial transcription. *PLOS Computational Biology*, 17(1):e1008572, January 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008572. URL https://dx.plos.org/10.1371/journal.pcbi.1008572.

[202] Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546, May 2022. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-022-01409-2. URL https://www.nature.com/articles/s41592-022-01409-2.

[203] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics*, 124(4):044104, 2006. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.2145882.

[204] Brian Munsky, Brooke Trinh, and Mustafa Khammash. Listening to the noise: random fluctuations reveal gene network parameters. *Molecular Systems Biology*, 5:318, October 2009. ISSN 1744-4292. doi: 10.1038/msb.2009.75. URL https://www.embopress.org/doi/full/10.1038/msb.2009.75.

[205] Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using Gene Expression Noise to Understand Gene Regulation. *Science*, 336(6078):183–187, April 2012. doi: 10.1126/science.1216379.

[206] Brian Munsky, Guoliang Li, Zachary R. Fox, Douglas P. Shepherd, and Gregor Neuert. Distribution shapes govern the discovery of predictive models for gene regulation. *Proceedings of the National Academy of Sciences*, 115 (29):7533–7538, 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1804060115.

[207] D. K. Nam, S. Lee, G. Zhou, X. Cao, C. Wang, T. Clark, J. Chen, J. D. Rowley, and S. M. Wang. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proceedings of the National Academy of Sciences*, 99(9):6152–6156, April 2002. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.092140899. URL http://www.pnas.org/cgi/doi/10.1073/pnas.092140899.

[208] Ilya Narsky and Frank C. Porter. *Statistical Analysis Techniques in Particle Physics: Fits, Density Estimation and Supervised Learning*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, November 2013. ISBN 978-3-527-67732-0 978-3-527-41086-6. doi: 10.1002/9783527677320. URL http://doi.wiley.com/10.1002/9783527677320.

[209] National Library of Medicine. Gene [Internet], 2004. URL https://www.ncbi.nlm.nih.gov/gene/.

[210] Damien Nicolas, Nick E. Phillips, and Felix Naef. What shapes eukaryotic transcriptional bursting? *Molecular BioSystems*, 13(7):1280–1290, 2017. ISSN 1742-206X, 1742-2051. doi: 10.1039/C7MB00154A. URL http://xlink.rsc.org/?DOI=C7MB00154A.

[211] Jun Ohkubo. Karlin-McGregor-like formula in a simple time-inhomogeneous birth–death process. *Journal of Physics A: Mathematical and Theoretical*, 47(40):405001, October 2014. ISSN 1751-8113, 1751-8121. doi: 10.1088/1751-8113/47/40/405001. URL https://iopscience.iop.org/article/10.1088/1751-8113/47/40/405001.

[212] Takumi Okamoto, Kazunori Imaizumi, and Masayuki Kaneko. The Role of Tissue-Specific Ubiquitin Ligases, RNF183, RNF186, RNF182 and RNF152, in Disease and Biological Function. *International Journal of Molecular Sciences*, 21(11):3921, May 2020. ISSN 1422-0067. doi: 10.3390/ijms21113921. URL https://www.mdpi.com/1422-0067/21/11/3921.

[213] Jonathan Packer and Cole Trapnell. Single-Cell Multi-omics: An Engine for New Quantitative Models of Gene Regulation. *Trends in Genetics*, 34(9):653–665, September 2018. ISSN 01689525. doi: 10.1016/j.tig.2018.06.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0168952518301082.

[214] Olivia Padovan-Merhar, Gautham P. Nair, Andrew G. Biaesch, Andreas Mayer, Steven Scarfone, Shawn W. Foley, Angela R. Wu, L. Stirling Churchman, Abhyudai Singh, and Arjun Raj. Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms. *Molecular Cell*, 58(2):339–352, April 2015. ISSN 10972765. doi: 10.1016/j.molcel.2015.03.005. URL https://linkinghub.elsevier.com/retrieve/pii/S1097276515001707.

[215] Harry H Panjer. Mixed Poisson Distributions. In *Encyclopedia of Actuarial Science*. John Wiley & Sons, Ltd, 2004. ISBN 978-0-470-01250-5.

[216] Nikolaos Papadopoulos, Parra R Gonzalo, and Johannes Söding. PROSSTT: probabilistic simulation of single-cell RNA-seq data for complex differentiation processes. *Bioinformatics*, 35(18):3517–3519, September 2019. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btz078. URL https://academic.oup.com/bioinformatics/article/35/18/3517/5305637.

[217] Ralph Patrick, David T. Humphreys, Vaibhao Janbandhu, Alicia Oshlack, Joshua W.K. Ho, Richard P. Harvey, and Kitty K. Lo. Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome Biology*, 21(1):167, December 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02071-7. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02071-7.

[218] Johan Paulsson. Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157–175, June 2005. ISSN 15710645. doi: 10.1016/j.plrev.2005.03.003. URL https://linkinghub.elsevier.com/retrieve/pii/S1571064505000138.

[219] Jean Peccoud and Bernard Ycard. Markovian Modeling of Gene Product Synthesis. *Theoretical Population Biology*, 48(2):222–234, 1995. doi: 10.1006/tpbi.1995.1027.

[220] 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(168): 20200360, July 2020. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2020.0360. URL https://royalsocietypublishing.org/doi/10.1098/rsif.2020.0360.

[221] Rob Phillips. Napoleon Is in Equilibrium. *Annual Review of Condensed Matter Physics*, 6(1):85–111, March 2015. ISSN 1947-5454, 1947-5462. doi: 10.1146/annurev-conmatphys-031214-014558. URL http://www.annualreviews.org/doi/10.1146/annurev-conmatphys-031214-014558.

[222] Belinda Phipson, Luke Zappia, and Alicia Oshlack. Gene length and detection bias in single cell RNA sequencing protocols. *F1000Research*, 6, April 2017. ISSN 2046-1402. doi: 10.12688/f1000research.11290.1. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5428526/.

[223] Harold Pimentel, John G. Conboy, and Lior Pachter. Keep Me Around: Intron Retention Detection and Analysis. Preprint, arXiv: 1510.00696, October 2015. URL http://arxiv.org/abs/1510.00696.

[224] Harold Pimentel, Marilyn Parra, Sherry L. Gee, Narla Mohandas, Lior Pachter, and John G. Conboy. A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Research*, 44(2):838–851, January 2016. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkv1168. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1168.

[225] J. W. Pitman. Occupation Measures for Markov Chains. *Advances in Applied Probability*, 9(1):69–86, 1977. ISSN 0001-8678. doi: 10.2307/1425817. URL https://www.jstor.org/stable/1425817. Publisher: Applied Probability Trust.

[226] Alex A Pollen, Tomasz J Nowakowski, Joe Shuga, Xiaohui Wang, Anne A Leyrat, Jan H Lui, Nianzhen Li, Lukasz Szpankowski, Brian Fowler, Peilin Chen, Naveen Ramalingam, Gang Sun, Myo Thu, Michael Norris, Ronald Lebofsky, Dominique Toppani, Darnell W Kemp, Michael Wong, Barry Clerkson, Brittnee N Jones, Shiquan Wu, Lawrence Knutsson, Beatriz Alvarado, Jing Wang, Lesley S Weaver, Andrew P May, Robert C Jones, Marc A Unger, Arnold R Kriegstein, and Jay A A West. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology*, 32(10):1053–1058, October 2014. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.2967. URL http://www.nature.com/articles/nbt.2967.

[227] A. Prados, J. J. Brey, and B. Sánchez-Rey. A Dynamical Monte Carlo Algorithm for Master Equations with Time-Dependent Transition Rates. *Journal*

of Statistical Physics, 89(3-4):709–734, November 1997. ISSN 0022-4715, 1572-9613. doi: 10.1007/BF02765541. URL http://link.springer.com/10.1007/BF02765541.

[228] Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, February 2020. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0690-6. URL http://www.nature.com/articles/s41592-019-0690-6.

[229] Qian Qin, Eli Bingham, Gioele La Manno, David M Langenau, and Luca Pinello. Pyro-Velocity: Probabilistic RNA Velocity inference from single-cell data. Preprint, bioRxiv: 2022.09.12.507691, October 2022. URL https://www.biorxiv.org/content/10.1101/2022.09.12.507691v2.

[230] Peng Qiu. Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications*, 11(1):1169, December 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-14976-9. URL http://www.nature.com/articles/s41467-020-14976-9.

[231] Xiaojie Qiu, Yan Zhang, Jorge D. Martin-Rufino, Chen Weng, Shayan Hosseinzadeh, Dian Yang, Angela N. Pogson, Marco Y. Hein, Kyung Hoi (Joseph) Min, Li Wang, Emanuelle I. Grody, Matthew J. Shurtleff, Ruoshi Yuan, Song Xu, Yian Ma, Joseph M. Replogle, Eric S. Lander, Spyros Darmanis, Ivet Bahar, Vijay G. Sankaran, Jianhua Xing, and Jonathan S. Weissman. Mapping transcriptomic vector fields of single cells. *Cell*, page S0092867421015774, February 2022. ISSN 00928674. doi: 10.1016/j.cell.2021.12.045. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867421015774.

[232] Arjun Raj and Alexander van Oudenaarden. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell*, 135(2):216–226, October 2008. ISSN 00928674. doi: 10.1016/j.cell.2008.09.050. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867408012439.

[233] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biology*, 4(10):e309, September 2006. ISSN 1545-7885. doi: 10.1371/journal.pbio.0040309. URL https://dx.plos.org/10.1371/journal.pbio.0040309.

[234] Linda E Reichl. *A Modern Course in Statistical Physics*. Wiley-VCH Verlag GmbH & Co. KGaA, 4th edition, 2016.

[235] Kirsten A. Reimer, Claudia A. Mimoso, Karen Adelman, and Karla M. Neugebauer. Co-transcriptional splicing regulates 3' end cleavage during mammalian erythropoiesis. *Molecular Cell*, 81(5):998–1012.e7, March

2021. ISSN 10972765. doi: 10.1016/j.molcel.2020.12.018. URL https://linkinghub.elsevier.com/retrieve/pii/S1097276520309370.

[236] H. Risken. *The Fokker-Planck equation: methods of solution and applications*. Number v. 18 in Springer series in synergetics. Springer-Verlag, New York, 2nd ed edition, 1996. ISBN 978-3-540-61530-9.

[237] G. W. Roberts. *Chemical reactions and chemical reactors*. John Wiley & Sons, Hoboken, NJ, 2008. ISBN 978-0-471-74220-3. OCLC: ocn176897332.

[238] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010. ISSN 1367-4811, 1367-4803. doi: 10.1093/bioinformatics/btp616. URL https://academic.oup.com/bioinformatics/article/26/1/139/182458.

[239] Joseph Rodriguez and Daniel R. Larson. Transcription in Living Cells: Molecular Mechanisms of Bursting. *Annual Review of Biochemistry*, 89 (1):189–212, June 2020. ISSN 0066-4154, 1545-4509. doi: 10.1146/annurev-biochem-011520-105250. URL https://www.annualreviews.org/doi/10.1146/annurev-biochem-011520-105250.

[240] Laura Rubió Ferrarons. *Insights into the CREB-regulated transcription coactivators (CRTCs) in neurons and astrocytes*. PhD Dissertation, Universitat Autònoma de Barcelona, October 2020. URL https://ddd.uab.cat/record/241504.

[241] Piergiacomo Sabino and Nicola Cufaro Petroni. Gamma-related Ornstein–Uhlenbeck processes and their simulation. *Journal of Statistical Computation and Simulation*, 91(6):1108–1133, April 2021. ISSN 0094-9655, 1563-5163. doi: 10.1080/00949655.2020.1842408. URL https://www.tandfonline.com/doi/full/10.1080/00949655.2020.1842408.

[242] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, May 2019. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-019-0071-9. URL http://www.nature.com/articles/s41587-019-0071-9.

[243] T. A. Samad. DRAGON: A Member of the Repulsive Guidance Molecule-Related Family of Neuronal- and Muscle-Expressed Membrane Proteins Is Regulated by DRG11 and Has Neuronal Adhesive Properties. *Journal of Neuroscience*, 24(8):2027–2036, February 2004. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.4115-03.2004. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.4115-03.2004.

[244] A. Sanchez and I. Golding. Genetic Determinants and Cellular Constraints in Noisy Gene Expression. *Science*, 342(6163):1188–1193, December 2013.

ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1242975. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1242975.

[245] Sara Sanders, Kunaal Joshi, Petra Levin, and Srividya Iyer-Biswas. Single cells tell their own story: An updated framework for understanding stochastic variations in cell cycle progression in bacteria. Preprint, bioRxiv: 2022.03.15.484524, March 2022. URL http://biorxiv.org/content/early/2022/03/16/2022.03.15.484524.abstract.

[246] Mahakaran Sandhu, Matthew A. Spence, and Colin J. Jackson. Evo-velocity: Protein language modeling accelerates the study of evolution. *Cell Systems*, 13(4):271–273, April 2022. ISSN 24054712. doi: 10.1016/j.cels.2022.03.004. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471222001314.

[247] Abhishek K Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis. Preprint, bioRxiv: 2020.04.07.030007, April 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.04.07.030007.

[248] Wim Schoutens. *Lévy Processes in Finance*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, UK, March 2003. ISBN 978-0-470-85156-2 978-0-470-87023-5. doi: 10.1002/0470870230. URL http://doi.wiley.com/10.1002/0470870230.

[249] Daniel Schwabe, Sara Formichetti, Jan Philipp Junker, Martin Falcke, and Nikolaus Rajewsky. The transcriptome dynamics of single cells during the cell cycle. *Molecular Systems Biology*, 16(11), November 2020. ISSN 1744-4292, 1744-4292. doi: 10.15252/msb.20209946. URL https://onlinelibrary.wiley.com/doi/10.15252/msb.20209946.

[250] Hannah J. Scott, Martin J. Stebbing, Claire E. Walters, Samuel McLenachan, Mark I. Ransome, Nancy R. Nichols, and Ann M. Turnley. Differential effects of SOCS2 on neuronal differentiation and morphology. *Brain Research*, 1067(1):138–145, January 2006. ISSN 00068993. doi: 10.1016/j.brainres.2005.10.032. URL https://linkinghub.elsevier.com/retrieve/pii/S0006899305014587.

[251] Adrien Senecal, Brian Munsky, Florence Proux, Nathalie Ly, Floriane E. Braye, Christophe Zimmer, Florian Mueller, and Xavier Darzacq. Transcription Factors Modulate c-Fos Transcriptional Bursts. *Cell Reports*, 8(1):75–83, July 2014. ISSN 22111247. doi: 10.1016/j.celrep.2014.05.053. URL https://linkinghub.elsevier.com/retrieve/pii/S2211124714004471.

[252] Sheel Shah, Yodai Takei, Wen Zhou, Eric Lubeck, Jina Yun, Chee-Huat Linus Eng, Noushin Koulena, Christopher Cronin, Christoph Karp, Eric J. Liaw, Mina Amin, and Long Cai. Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. *Cell*, 174(2):363–376.e16, July

2018. ISSN 00928674. doi: 10.1016/j.cell.2018.05.035. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867418306470.

[253] V. Shahrezaei and P. S. Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105 (45):17256–17261, November 2008. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0803850105. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0803850105.

[254] Vahid Shahrezaei, Julien F Ollivier, and Peter S Swain. Colored extrinsic fluctuations and stochastic gene expression. *Molecular Systems Biology*, 4(1):196, January 2008. ISSN 1744-4292, 1744-4292. doi: 10.1038/msb.2008.31. URL https://onlinelibrary.wiley.com/doi/10.1038/msb.2008.31.

[255] L. V. Sharova, A. A. Sharov, T. Nedorezov, Y. Piao, N. Shaik, and M. S.H. Ko. Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells. *DNA Research*, 16(1):45–58, January 2009. ISSN 1340-2838, 1756-1663. doi: 10.1093/dnares/dsn030. URL https://academic.oup.com/dnaresearch/article-lookup/doi/10.1093/dnares/dsn030.

[256] Caibin Sheng, Rui Lopes, Gang Li, Sven Schuierer, Annick Waldt, Rachel Cuttat, Slavica Dimitrieva, Audrey Kauffmann, Eric Durand, Giorgio G. Galli, Guglielmo Roma, and Antoine de Weck. Probabilistic machine learning ensures accurate ambient denoising in droplet-based single-cell omics. Preprint, bioRxiv: 2022.01.14.476312, January 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.01.14.476312.

[257] Changhong Shi, Yiguo Jiang, and Tianshou Zhou. Queuing Models of Gene Expression: Analytical Distributions and Beyond. *Biophysical Journal*, 119(8):1606–1616, October 2020. ISSN 0006-3495. doi: 10.1016/j.bpj.2020.09.001. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7642270/.

[258] Ki Shuk Shim, Margit Rosner, Angelika Freilinger, Gert Lubec, and Markus Hengstschläger. Bach2 is involved in neuronal differentiation of N1E-115 neuroblastoma cells. *Experimental Cell Research*, 312(12):2264–2278, July 2006. ISSN 00144827. doi: 10.1016/j.yexcr.2006.03.018. URL https://linkinghub.elsevier.com/retrieve/pii/S0014482706001194.

[259] Jonathon Shlens. A Tutorial on Principal Component Analysis. Preprint, arXiv: 1404.1100, April 2014. URL http://arxiv.org/abs/1404.1100.

[260] Daniel Silk, Paul D. W. Kirk, Chris P. Barnes, Tina Toni, and Michael P. H. Stumpf. Model Selection in Systems Biology Depends on Experimental Design. *PLoS Computational Biology*, 10(6):e1003650, June 2014. ISSN

1553-7358. doi: 10.1371/journal.pcbi.1003650. URL https://dx.plos.org/10.1371/journal.pcbi.1003650.

[261] Abhyudai Singh and Pavol Bokes. Consequences of mRNA Transport on Stochastic Variability in Protein Levels. *Biophysical Journal*, 103 (5):1087–1096, September 2012. ISSN 00063495. doi: 10.1016/j.bpj.2012.07.015. URL https://linkinghub.elsevier.com/retrieve/pii/S0006349512007904.

[262] Samuel O Skinner, Heng Xu, Sonal Nagarkar-Jaiswal, Pablo R Freire, Thomas P Zwaka, and Ido Golding. Single-cell analysis of transcription kinetics across the cell cycle. *eLife*, 5:e12175, January 2016. ISSN 2050-084X. doi: 10.7554/eLife.12175. URL https://elifesciences.org/articles/12175.

[263] Ruslan Soldatov, Marketa Kaucka, Maria Eleni Kastriti, Julian Petersen, Tatiana Chontorotzea, Lukas Englmaier, Natalia Akkuratova, Yunshi Yang, Martin Häring, Viacheslav Dyachuk, Christoph Bock, Matthias Farlik, Michael L. Piacentino, Franck Boismoreau, Markus M. Hilscher, Chika Yokota, Xiaoyan Qian, Mats Nilsson, Marianne E. Bronner, Laura Croci, Wen-Yu Hsiao, David A. Guertin, Jean-Francois Brunet, Gian Giacomo Consalez, Patrik Ernfors, Kaj Fried, Peter V. Kharchenko, and Igor Adameyko. Spatiotemporal structure of cell fate decisions in murine neural crest. *Science*, 364(6444):eaas9536, June 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aas9536. URL http://www.sciencemag.org/lookup/doi/10.1126/science.aas9536.

[264] Charlotte Soneson, Avi Srivastava, Rob Patro, and Michael B. Stadler. Preprocessing choices affect RNA velocity results for droplet scRNA-seq data. *PLOS Computational Biology*, 17(1):e1008585, January 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008585. URL https://dx.plos.org/10.1371/journal.pcbi.1008585.

[265] J. Michael Steele. *Stochastic calculus and financial applications*. Number 45 in Applications of mathematics. Springer, New York, 2001. ISBN 978-0-387-95016-7.

[266] Stephen M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press, March 1990. URL https://www.hup.harvard.edu/catalog.php?isbn=9780674403413.

[267] Adam R. Stinchcombe, Charles S. Peskin, and Daniel Tranchina. Population density approach for discrete mRNA distributions in generalized switching models for stochastic gene expression. *Physical Review E*, 85(6):061919, June 2012. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.85.061919. URL https://link.aps.org/doi/10.1103/PhysRevE.85.061919.

[268] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, September 2017. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4380. URL http://www.nature.com/articles/nmeth.4380.

[269] Michael P.H. Stumpf. Inferring better gene regulation networks from single-cell data. *Current Opinion in Systems Biology*, 27:100342, September 2021. ISSN 24523100. doi: 10.1016/j.coisb.2021.05.003. URL https://linkinghub.elsevier.com/retrieve/pii/S2452310021000275.

[270] Patrick S. Stumpf, Rosanna C.G. Smith, Michael Lenz, Andreas Schuppert, Franz-Josef Müller, Ann Babtie, Thalia E. Chan, Michael P.H. Stumpf, Colin P. Please, Sam D. Howison, Fumio Arai, and Ben D. MacArthur. Stem Cell Differentiation as a Non-Markov Stochastic Process. *Cell Systems*, 5(3):268–282.e7, September 2017. ISSN 24054712. doi: 10.1016/j.cels.2017.08.009. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471217303423.

[271] Augustinas Sukys, Kaan Öcal, and Ramon Grima. Approximating solutions of the Chemical Master equation using neural networks. *iScience*, 25(9):105010, August 2022. ISSN 2589-0042. doi: 10.1016/j.isci.2022.105010. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9474291/.

[272] Xi-Ming Sun, Anthony Bowman, Miles Priestman, Francois Bertaux, Amalia Martinez-Segura, Wenhao Tang, Chad Whilding, Dirk Dormann, Vahid Shahrezaei, and Samuel Marguerat. Size-Dependent Increase in RNA Polymerase II Initiation Rates Mediates Gene Expression Scaling with Cell Size. *Current Biology*, 30(7):1217–1230.e7, April 2020. ISSN 09609822. doi: 10.1016/j.cub.2020.01.053. URL https://linkinghub.elsevier.com/retrieve/pii/S096098222030097X.

[273] Makoto Suzuki, Yusuke Hara, Chiyo Takagi, Takamasa S. Yamamoto, and Naoto Ueno. *MID1* and *MID2* are required for *Xenopus* neural tube closure through the regulation of microtubule organization. *Development*, 138(2):385–385, January 2011. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.062976. URL https://journals.biologists.com/dev/article/138/2/385/44835/MID1-and-MID2-are-required-for-Xenopus-neural-tube.

[274] Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, February 2020. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-019-0379-5. URL https://www.nature.com/articles/s41587-019-0379-5.

[275] Valentine Svensson and Lior Pachter. RNA Velocity: Molecular Kinetics from Single-Cell RNA-Seq. *Molecular Cell*, 72(1):7–9, October

2018. ISSN 10972765. doi: 10.1016/j.molcel.2018.09.026. URL https://linkinghub.elsevier.com/retrieve/pii/S1097276518307974.

[276] Valentine Svensson, Adam Gayoso, Nir Yosef, and Lior Pachter. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421, June 2020. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btaa169. URL https://academic.oup.com/bioinformatics/article/36/11/3418/5807606.

[277] P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, October 2002. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.162041399. URL http://www.pnas.org/cgi/doi/10.1073/pnas.162041399.

[278] Wenhao Tang, François Bertaux, Philipp Thomas, Claire Stefanelli, Malika Saint, Samuel Marguerat, and Vahid Shahrezaei. bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. *Bioinformatics*, 36(4):1174–1181, February 2020. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btz726. URL https://academic.oup.com/bioinformatics/article/36/4/1174/5581401.

[279] Wenhao Tang, Andreas Christ Sølvsten Jørgensen, Samuel Marguerat, Philipp Thomas, and Vahid Shahrezaei. Modelling capture efficiency of single cell RNA-sequencing data improves inference of transcriptome-wide burst kinetics. Preprint, bioRxiv: 2023.03.06.531327, March 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.03.06.531327.

[280] Michael E. Taylor. *Partial differential equations*. Number v. 115-<116> in Applied mathematical sciences. Springer, New York, 2nd ed edition, 2011. ISBN 978-1-4419-7054-1 978-1-4419-7055-8 978-1-4419-7051-0 978-1-4419-7052-7 978-1-4419-7048-0 978-1-4419-7049-7. OCLC: ocn681675116.

[281] Martina Tedesco, Francesca Giannese, Dejan Lazarević, Valentina Giansanti, Dalia Rosano, Silvia Monzani, Irene Catalano, Elena Grassi, Eugenia R. Zanella, Oronza A. Botrugno, Leonardo Morelli, Paola Panina Bordignon, Giulio Caravagna, Andrea Bertotti, Gianvito Martino, Luca Aldrighetti, Sebastiano Pasqualato, Livio Trusolino, Davide Cittaro, and Giovanni Tonon. Chromatin Velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin. *Nature Biotechnology*, October 2021. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-021-01031-1. URL https://www.nature.com/articles/s41587-021-01031-1.

[282] Philipp Thomas. Making sense of snapshot data: ergodic principle for clonal cell populations. *Journal of The Royal Society Interface*, 14

(136):20170467, November 2017. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2017.0467. URL https://royalsocietypublishing.org/doi/10.1098/rsif.2017.0467.

[283] 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(178):20210274, May 2021. ISSN 1742-5662. doi: 10.1098/rsif.2021.0274. URL https://royalsocietypublishing.org/doi/10.1098/rsif.2021.0274.

[284] Surangrat Thongkorn, Songphon Kanlayaprasit, Pawinee Panjabud, Thanit Saeliw, Thanawin Jantheang, Kasidit Kasitipradit, Suthathip Sarobol, Depicha Jindatip, Valerie W. Hu, Tewin Tencomnao, Takako Kikkawa, Tatsuya Sato, Noriko Osumi, and Tewarit Sarachana. Sex differences in the effects of prenatal bisphenol A exposure on autism-related genes and their relationships with the hippocampus functions. *Scientific Reports*, 11(1):1241, December 2021. ISSN 2045-2322. doi: 10.1038/s41598-020-80390-2. URL http://www.nature.com/articles/s41598-020-80390-2.

[285] B. C. Thorne, A. M. Bailey, and S. M. Peirce. Combining experiments with multi-cell agent-based modeling to study biological tissue patterning. *Briefings in Bioinformatics*, 8(4):245–257, March 2007. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbm024. URL https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbm024.

[286] Nicola Thrupp, Carlo Sala Frigerio, Leen Wolfs, Nathan G. Skene, Nicola Fattorelli, Suresh Poovathingal, Yannick Fourne, Paul M. Matthews, Tom Theys, Renzo Mancuso, Bart de Strooper, and Mark Fiers. Single-Nucleus RNA-Seq Is Not Suitable for Detection of Microglial Activation Genes in Humans. *Cell Reports*, 32(13):108189, September 2020. ISSN 22111247. doi: 10.1016/j.celrep.2020.108189. URL https://linkinghub.elsevier.com/retrieve/pii/S2211124720311785.

[287] Luyi Tian, Jafar S. Jabbari, Rachel Thijssen, Quentin Gouil, Shanika L. Amarasinghe, Oliver Voogd, Hasaru Kariyawasam, Mei R. M. Du, Jakob Schuster, Changqing Wang, Shian Su, Xueyi Dong, Charity W. Law, Alexis Lucattini, Yair David Joseph Prawer, Coralina Collar-Fernández, Jin D. Chung, Timur Naim, Audrey Chan, Chi Hai Ly, Gordon S. Lynch, James G. Ryall, Casey J. A. Anttila, Hongke Peng, Mary Ann Anderson, Christoffer Flensburg, Ian Majewski, Andrew W. Roberts, David C. S. Huang, Michael B. Clark, and Matthew E. Ritchie. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biology*, 22(1):310, December 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02525-6. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02525-6.

[288] F. Tissir, I. Bar, A.M. Goffinet, and C. Lambert De Rouvroit. Expression of the ankyrin repeat domain 6 gene (Ankrd6) during mouse brain development. *Developmental Dynamics*, 224(4):465–469, August 2002. ISSN 1058-8388, 1097-0177. doi: 10.1002/dvdy.10126. URL https://onlinelibrary.wiley.com/doi/10.1002/dvdy.10126.

[289] Nestor V. Torres and Guido Santos. The (Mathematical) Modeling Process in Biosciences. *Frontiers in Genetics*, 6:354, December 2015. ISSN 1664-8021. doi: 10.3389/fgene.2015.00354. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4686688/.

[290] F. William Townes, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20(1):295, December 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1861-6. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1861-6.

[291] Sophie Tritschler, Maren Büttner, David S. Fischer, Marius Lange, Volker Bergen, Heiko Lickert, and Fabian J. Theis. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development*, 146(12):dev170506, June 2019. ISSN 0950-1991, 1477-9129. doi: 10.1242/dev.170506. URL http://dev.biologists.org/lookup/doi/10.1242/dev.170506.

[292] S. T. Tse and Justin W. L. Wan. Low-bias simulation scheme for the Heston model by Inverse Gaussian approximation. *Quantitative Finance*, 13(6):919–937, June 2013. ISSN 1469-7688, 1469-7696. doi: 10.1080/14697688.2012.696678. URL http://www.tandfonline.com/doi/abs/10.1080/14697688.2012.696678.

[293] Edward Tunnacliffe and Jonathan R. Chubb. What Is a Transcriptional Burst? *Trends in Genetics*, 36(4):288–297, April 2020. ISSN 01689525. doi: 10.1016/j.tig.2020.01.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0168952520300056.

[294] M. Ullah and O. Wolkenhauer. Family tree of Markov models in systems biology. *IET Systems Biology*, 1(4):247–254, July 2007. ISSN 1751-8849, 1751-8857. doi: 10.1049/iet-syb:20070017. URL https://digital-library.theiet.org/content/journals/10.1049/iet-syb_20070017.

[295] Mukhtar Ullah and Olaf Wolkenhauer. *Stochastic approaches for systems biology*. Springer, New York, 2011. ISBN 978-1-4614-0477-4 978-1-4614-0478-1. OCLC: ocn733239594.

[296] T. K. Ulland and M. Colonna. Trem2 - a key player in microglial biology and alzheimer disease. *Nature Reviews Neurology*, 14:667–675, 2018. doi: 10.

1038/s41582-018-0072-1. URL https://www.nature.com/articles/s41582-018-0072-1.

[297] N. G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, third edition, 2007. ISBN 978-0-444-52965-7.

[298] Erik van Nimwegen. Inferring intrinsic and extrinsic noise from a dual fluorescent reporter. Preprint, bioRxiv: 049486, April 2016. URL http://biorxiv.org/lookup/doi/10.1101/049486.

[299] John J Vastola. *In search of a coherent theoretical framework for stochastic gene regulation*. PhD thesis, Vanderbilt, March 2021. URL https://ir.vanderbilt.edu/handle/1803/16646.

[300] John J. Vastola. Solving the chemical master equation for monomolecular reaction systems and beyond: a Doi-Peliti path integral view. *Journal of Mathematical Biology*, 83(5):48, November 2021. ISSN 0303-6812, 1432-1416. doi: 10.1007/s00285-021-01670-7. URL https://link.springer.com/10.1007/s00285-021-01670-7.

[301] Frits Veerman, Carsten Marr, and Nikola Popović. Time-dependent propagators for stochastic models of gene expression: an analytical method. *Journal of Mathematical Biology*, 77(2):261–312, August 2018. ISSN 0303-6812, 1432-1416. doi: 10.1007/s00285-017-1196-4. URL http://link.springer.com/10.1007/s00285-017-1196-4.

[302] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1.0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael

Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, March 2020. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0686-2. URL http://www.nature.com/articles/s41592-019-0686-2.

[303] Sean T. Vittadello and Michael P.H. Stumpf. Open problems in mathematical biology. *Mathematical Biosciences*, 354:108926, December 2022. ISSN 00255564. doi: 10.1016/j.mbs.2022.108926. URL https://linkinghub.elsevier.com/retrieve/pii/S0025556422001158.

[304] Huy D. Vo and Roger B. Sidje. An adaptive solution to the chemical master equation using tensors. *The Journal of Chemical Physics*, 147(4):044102, July 2017. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.4994917. URL http://aip.scitation.org/doi/10.1063/1.4994917.

[305] Margaritis Voliotis, Philipp Thomas, Ramon Grima, and Clive G. Bowsher. Stochastic Simulation of Biomolecular Networks in Dynamic Environments. *PLOS Computational Biology*, 12(6):e1004923, June 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004923. URL https://dx.plos.org/10.1371/journal.pcbi.1004923.

[306] Johann Wolfgang von Goethe. *Faust*. The World Publishing Company, Cleveland, Ohio, 1870.

[307] Yihan Wan, Dimitrios G. Anastasakis, Joseph Rodriguez, Murali Palangat, Prabhakar Gudla, George Zaki, Mayank Tandon, Gianluca Pegoraro, Carson C. Chow, Markus Hafner, and Daniel R. Larson. Dynamic imaging of nascent RNA reveals general principles of transcription dynamics and stochastic splice site selection. *Cell*, 184(11):2878–2895.e20, May 2021. ISSN 00928674. doi: 10.1016/j.cell.2021.04.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867421004918.

[308] Jingshu Wang, Mo Huang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, John Murray, Arjun Raj, Mingyao Li, and Nancy R. Zhang. Gene expression distribution deconvolution in single-cell RNA sequencing. *Proceedings of the National Academy of Sciences*, 115(28):E6437–E6446, July 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1721085115. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1721085115.

[309] Mengyu Wang, Jing Zhang, Heng Xu, and Ido Golding. Measuring transcription at a single gene copy reveals hidden drivers of bacterial individuality.

Nature Microbiology, 4:2118–2127, September 2019. ISSN 2058-5276. doi: 10.1038/s41564-019-0553-z. URL http://www.nature.com/articles/s41564-019-0553-z.

[310] Shangying Wang and Simone Bianco. AI-assisted Biology: Predict the Conditional Probability Distributions from Noisy Measurements. Preprint, bioRxiv: 2021.10.07.463577, October 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.10.07.463577.

[311] Xin Wang. Velo-Predictor: an ensemble learning pipeline for RNA velocity prediction. *BMC Bioinformatics*, page 12, 2021.

[312] Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Number 25 in Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2009. ISBN 978-0-511-65153-3.

[313] Philipp Weiler. Protein Velocity in Single Cells using Multi-Omics Modelling. Master's thesis, EPFL, TU Munich, January 2021.

[314] Guangzheng Weng, Junil Kim, and Kyoung Jae Won. VeTra: a tool for trajectory inference based on RNA velocity. *Bioinformatics*, page btab364, May 2021. doi: 10.1093/bioinformatics/btab364.

[315] Darren James Wilkinson. *Stochastic modelling for systems biology*. Chapman & Hall/CRC mathematical and computational biology. CRC Press, Taylor & Francis Group, Boca Raton, third edition edition, 2019. ISBN 978-1-138-54928-9.

[316] Barbara Wold and Richard M Myers. Sequence census methods for functional genomics. *Nature Methods*, 5(1):19–21, January 2008. ISSN 1548-7105. doi: 10.1038/nmeth1157. URL https://doi.org/10.1038/nmeth1157.

[317] Haiqing Xiong, Yingjie Luo, Yanzhu Yue, Jiejie Zhang, Shanshan Ai, Xin Li, Xuelian Wang, Yun-Long Zhang, Yusheng Wei, Hui-Hua Li, Xinli Hu, Cheng Li, and Aibin He. Single-Cell Transcriptomics Reveals Chemotaxis-Mediated Intraorgan Crosstalk During Cardiogenesis. *Circulation Research*, 125(4):398–410, August 2019. ISSN 0009-7330, 1524-4571. doi: 10.1161/CIRCRESAHA.119.315243. URL https://www.ahajournals.org/doi/10.1161/CIRCRESAHA.119.315243.

[318] Heng Xu, Leonardo A Sepúlveda, Lauren Figard, Anna Marie Sokac, and Ido Golding. Combining protein and mRNA quantification to decipher transcriptional regulation. *Nature Methods*, 12(8):739–742, August 2015. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3446. URL http://www.nature.com/articles/nmeth.3446.

[319] Heng Xu, Samuel O. Skinner, Anna Marie Sokac, and Ido Golding. Stochastic Kinetics of Nascent RNA. *Physical Review Letters*, 117(12):128101,

2016. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.117.128101. URL https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.117.128101.

[320] Yunan Yang, Levon Nurbekyan, Elisa Negrini, Robert Martin, and Mirjeta Pasha. Optimal Transport for Parameter Identification of Chaotic Dynamics via Invariant Measures. Preprint, arXiv: 2104.15138, May 2021. URL http://arxiv.org/abs/2104.15138.

[321] Zizhen Yao, Hanqing Liu, Fangming Xie, Stephan Fischer, Ricky S. Adkins, Andrew I. Aldridge, Seth A. Ament, Anna Bartlett, M. Margarita Behrens, Koen Van den Berge, Darren Bertagnolli, Hector Roux de Bézieux, Tommaso Biancalani, A. Sina Booeshaghi, Héctor Corrada Bravo, Tamara Casper, Carlo Colantuoni, Jonathan Crabtree, Heather Creasy, Kirsten Crichton, Megan Crow, Nick Dee, Elizabeth L. Dougherty, Wayne I. Doyle, Sandrine Dudoit, Rongxin Fang, Victor Felix, Olivia Fong, Michelle Giglio, Jeff Goldy, Mike Hawrylycz, Brian R. Herb, Ronna Hertzano, Xiaomeng Hou, Qiwen Hu, Jayaram Kancherla, Matthew Kroll, Kanan Lathia, Yang Eric Li, Jacinta D. Lucero, Chongyuan Luo, Anup Mahurkar, Delissa McMillen, Naeem M. Nadaf, Joseph R. Nery, Thuc Nghi Nguyen, Sheng-Yong Niu, Vasilis Ntranos, Joshua Orvis, Julia K. Osteen, Thanh Pham, Antonio Pinto-Duarte, Olivier Poirion, Sebastian Preissl, Elizabeth Purdom, Christine Rimorin, Davide Risso, Angeline C. Rivkin, Kimberly Smith, Kelly Street, Josef Sulc, Valentine Svensson, Michael Tieu, Amy Torkelson, Herman Tung, Eeshit Dhaval Vaishnav, Charles R. Vanderburg, Cindy van Velthoven, Xinxin Wang, Owen R. White, Z. Josh Huang, Peter V. Kharchenko, Lior Pachter, John Ngai, Aviv Regev, Bosiljka Tasic, Joshua D. Welch, Jesse Gillis, Evan Z. Macosko, Bing Ren, Joseph R. Ecker, Hongkui Zeng, and Eran A. Mukamel. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*, 598(7879):103–110, October 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03500-8. URL https://www.nature.com/articles/s41586-021-03500-8.

[322] Yuan Yin, Masanao Yajima, and Joshua D. Campbell. Characterization and decontamination of background noise in droplet-based single-cell protein expression data with DecontPro. Preprint, bioRxiv: 2023.01.27.525964v2, February 2023. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9979990/.

[323] Matthew D Young and Sam Behjati. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience*, 9 (12):giaa151, December 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa151. URL https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giaa151/6049831.

[324] Leqian Yu, Yulei Wei, Jialei Duan, Daniel A. Schmitz, Masahiro Sakurai, Lei Wang, Kunhua Wang, Shuhua Zhao, Gary C. Hon, and Jun Wu.

Blastocyst-like structures generated from human pluripotent stem cells. *Nature*, 591(7851):620–626, March 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03356-y. URL https://www.nature.com/articles/s41586-021-03356-y.

[325] Christoph Zechner and Heinz Koeppl. Uncoupled Analysis of Stochastic Reaction Networks in Fluctuating Environments. *PLoS Computational Biology*, 10(12):e1003942, December 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003942. URL https://dx.plos.org/10.1371/journal.pcbi.1003942.

[326] A. Zeisel, W. J. Kostler, N. Molotski, J. M. Tsai, R. Krauthgamer, J. Jacob-Hirsch, G. Rechavi, Y. Soen, S. Jung, Y. Yarden, and E. Domany. Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Molecular Systems Biology*, 7(1):529–529, September 2011. ISSN 1744-4292. doi: 10.1038/msb.2011.62. URL http://msb.embopress.org/cgi/doi/10.1038/msb.2011.62.

[327] Daniel Zenklusen, Daniel R Larson, and Robert H Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology*, 15(12):1263–1271, 2008. ISSN 1545-9993, 1545-9985. doi: 10.1038/nsmb.1514.

[328] Martin Jinye Zhang, Vasilis Ntranos, and David Tse. Determining sequencing depth in a single-cell RNA-seq experiment. *Nature Communications*, 11(1):774, February 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-14482-y. URL https://www.nature.com/articles/s41467-020-14482-y.

[329] Stephen Zhang, Anton Afanassiev, Laura Greenstreet, Tetsuya Matsumoto, and Geoffrey Schiebinger. Optimal transport analysis reveals trajectories in steady-state systems. *PLOS Computational Biology*, 17(12):e1009466, December 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1009466. URL https://dx.plos.org/10.1371/journal.pcbi.1009466.

[330] Ziqi Zhang and Xiuwei Zhang. Inference of high-resolution trajectories in single cell RNA-Seq data from RNA velocity. Preprint, bioRxiv: 2020.09.30.321125, October 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.09.30.321125.

[331] Ziqi Zhang and Xiuwei Zhang. VeloSim: Simulating single cell gene-expression and RNA velocity. Preprint, bioRxiv: 2021.01.11.426277, January 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.01.11.426277.

[332] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y.

Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8 (1):14049, April 2017. ISSN 2041-1723. doi: 10.1038/ncomms14049. URL http://www.nature.com/articles/ncomms14049.

[333] Shijie C. Zheng, Genevieve Stein-O'Brien, Leandros Boukas, Loyal A Goff, and Kasper D Hansen. Pumping the brakes on RNA velocity – understanding and interpreting RNA velocity estimates. Preprint, bioRxiv: 2022.06.19.494717, June 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.06.19.494717.

[334] Tianshou Zhou and Jiajun Zhang. Analytical Results for a Multistate Gene Model. *SIAM Journal on Applied Mathematics*, 72(3):789–818, January 2012. ISSN 0036-1399, 1095-712X. doi: 10.1137/110852887. URL http://epubs.siam.org/doi/10.1137/110852887.

[335] Liucun Zhu, Ying Zhang, Wen Zhang, Sihai Yang, Jian-Qun Chen, and Dacheng Tian. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics*, 10(1):47, December 2009. ISSN 1471-2164. doi: 10.1186/1471-2164-10-47. URL https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-10-47.

[336] Péter Érdi and Gábor Lente. *Stochastic chemical kinetics: theory and (mostly) systems biological applications*. Springer Complexity. Springer, New York, 2014. ISBN 978-1-4939-0386-3.

Appendix A

SUPPLEMENTARY GENERATING FUNCTION DERIVATIONS

Most of this appendix summarizes the mathematical machinery outlined in the supplement to [115] by G.G., J.J.V., and L.P. G.G. developed this approach as a generalization of the framework constructed by G.G. and J.J.V. in [113] by G.G.*, J.J.V.*, M.F., and L.P., as well as by J.J.V. in [299], among other publications. The description was written by G.G. and J.J.V.

Section A.8.2 was adapted by G.G. from a derivation by J.J.V. and G.G. in [113] by G.G.*, J.J.V.*, M.F., and L.P.

Section A.8.3.1 was adapted from [105] by G.G. and L.P. The derivation was performed by G.G.

A.1 The full master equation

The full master equation for each s takes the following form:

$$
\begin{aligned}
\frac{\partial P(s,\mathbf{x},\mathbf{y},t)}{\partial t} &= \sum_{i=1}^{N} H_{is}(t) P(i,\mathbf{x},\mathbf{y},t) \\
&+ \sum_{i=1}^{n} c_{i0}\left[(x_i+1)P(s,x_i+1,\mathbf{y},t) - x_i P(s,\mathbf{x},\mathbf{y},t)\right] \\
&+ \sum_{i,j=1}^{n} c_{ij}\left[(x_i+1)P(s,x_i+1,x_j-1,\mathbf{y},t) - x_i P(s,\mathbf{x},\mathbf{y},t)\right] \\
&+ \sum_{i=1}^{n} Q^d_{ii}\left[(x_i-1)P(s,x_i-1,\mathbf{y},t) - x_i P(s,\mathbf{x},\mathbf{y},t)\right] \\
&+ \sum_{i,j=1}^{n} Q^d_{ij}\left[x_i P(s,x_j-1,\mathbf{y},t) - x_i P(s,\mathbf{x},\mathbf{y},t)\right] \\
&+ \sum_{\omega} \alpha^d_{s,\omega}(t)\left[\sum_{\mathbf{z}} p^d_{s,\omega}(\mathbf{z},t) P(s,\mathbf{x}-\mathbf{z},\mathbf{y},t) - P(s,\mathbf{x},\mathbf{y},t)\right] \\
&- \sum_{i,j=1}^{m} C^{cc}_{ij} \frac{\partial}{\partial y_j}\left[y_i P(s,\mathbf{x},\mathbf{y},t)\right] \\
&+ \frac{1}{2}\sum_{i=1}^{m} \sigma_i^2 \frac{\partial^2}{\partial y_i^2}\left[y_i P(s,\mathbf{x},\mathbf{y},t)\right] \\
&- \sum_{i=1}^{m} \alpha^c_{s,i}(t) \frac{\partial P(s,\mathbf{x},\mathbf{y},t)}{\partial y_i} \\
&+ \sum_{\omega>m} \alpha^c_{s,\omega}(t)\left[\int_{\mathbf{z}} p^c_{s,\omega}(\mathbf{z},t) P(s,\mathbf{x},\mathbf{y}-\mathbf{z},t)\,d\mathbf{z} - P(s,\mathbf{x},\mathbf{y},t)\right] \\
&+ \sum_{i=1}^{m}\sum_{j=1}^{n} C^{cd}_{ij}\left[y_i P(s,x_j-1,\mathbf{y},t) - y_i P(s,\mathbf{x},\mathbf{y},t)\right] \\
&- \sum_{i=1}^{n}\sum_{j=1}^{m} C^{dc}_{ij} x_i \frac{\partial P(s,\mathbf{x},\mathbf{y},t)}{\partial y_j}.
\end{aligned}
\tag{A.1}
$$

We annotate the terms in Table A.1, eliding the arguments that do not explicitly appear in the reactions.

To convert the master equation into a partial differential equation, we need to enumerate the functional forms of master equations terms and their generating functions. We begin with the definition of the generating function at time t. In the current derivation, whenever the argument of P is not explicitly specified, it consists

Term	Interpretation
$H_{is}(t)P(i)$	Transition from categorical state i to s
$c_{i0}\left[(x_i+1)P(x_i+1)-x_iP(s,\mathbf{x},\mathbf{y},t)\right]$	Degradation of discrete species i
$c_{ij}\left[(x_i+1)P(x_i+1,x_j-1)-x_iP\right]$	Conversion of discrete species i to j
$Q^d_{ii}\left[(x_i-1)P(x_i-1)-x_iP\right]$	Autocatalysis of discrete species i
$Q^d_{ij}\left[x_iP(x_j-1)-x_iP\right]$	Catalysis of discrete species j by i
$\alpha^d_{s,\omega}(t)\left[\sum_{\mathbf{z}} p^d_{s,\omega}(\mathbf{z},t)P(\mathbf{x}-\mathbf{z})-P\right]$	Bursty production of discrete species
$-C^{cc}_{ij}\frac{\partial}{\partial y_j}\left[y_iP\right]$	Increase in continuous species j proportional to level of continuous species i
$\sigma_i^2\frac{\partial^2}{\partial y_i^2}\left[y_iP\right]$	Square-root noise in continuous species i
$-\alpha^c_{s,i}(t)\frac{\partial P}{\partial y_i}$	Drift in continuous species i
$\alpha^c_{s,\omega}(t)\left[\int_{\mathbf{z}} p^c_{s,\omega}(\mathbf{z},t)P(\mathbf{y}-\mathbf{z})d\mathbf{z}-P\right]$	Bursting in continuous species
$C^{cd}_{ij}\left[y_iP(x_j-1)-y_iP\right]$	Production of discrete species j proportional to level of continuous species i
$-C^{dc}_{ij}x_i\frac{\partial P}{\partial y_j}$	Drift in continuous species j proportional to number of discrete species i

Table A.1: Components of the full master equation.

of s, $\mathbf{x}$, and $\mathbf{y}$. Analogously, whenever the argument of G is not explicitly specified, it consists of s, $\mathbf{g}$, and $\mathbf{h}$.

$$G=\int_{\mathbf{y}}\sum_{\mathbf{x}}\mathbf{g}^{\mathbf{x}}e^{\mathbf{h}^\top\mathbf{y}}Pd\mathbf{y}. \tag{A.2}$$

Evidently, the generating function of all terms in Equation A.1 that scale as P is G.

A.2 Fully discrete master equation terms

Multiplying G through by g_i, we obtain:

$$g_iG=\int_{\mathbf{y}}\sum_{\mathbf{x}}\mathbf{g}^{\mathbf{x}}e^{\mathbf{h}^\top\mathbf{y}}g_iPd\mathbf{y}=\int_{\mathbf{y}}\sum_{\mathbf{x}}\mathbf{g}^{\mathbf{x}}e^{\mathbf{h}^\top\mathbf{y}}P(x_i-1)d\mathbf{y}. \tag{A.3}$$

This follows from rewriting $\sum_{x_i=0}^{\infty}g_i^{x_i+1}P(x_i)$ as $\sum_{x_i=-1}^{\infty}g_i^{x_i+1}P(x_i)$, noting that $P(x_i)=0$ whenever $x_i<0$, and reindexing to obtain the equivalent sum $\sum_{x_i=0}^{\infty}g_i^{x_i}P(x_i-1)$. Therefore, the generating function of all terms that scale as $P(x_i-1)$ is g_iG.

Differentiating with respect to g_i:

$$\begin{aligned}\frac{\partial G}{\partial g_i}&=\int_{\mathbf{y}}\sum_{\mathbf{x}}\mathbf{g}^{\mathbf{x}}e^{\mathbf{h}^\top\mathbf{y}}x_ig_i^{-1}Pd\mathbf{y},\ \text{i.e.,}\\ g_i\frac{\partial G}{\partial g_i}&=\int_{\mathbf{y}}\sum_{\mathbf{x}}\mathbf{g}^{\mathbf{x}}e^{\mathbf{h}^\top\mathbf{y}}x_iPd\mathbf{y}.\end{aligned} \tag{A.4}$$

Therefore, the generating function of all terms that scale as $x_i P$ is $g_i \frac{\partial G_s}{\partial g_i}$.

Alternatively, we can note that the $x_i = 0$ term of $\sum_{x_i=0}^{\infty} g_i^{x_i-1} x_i P(x_i)$ is zero, and rewrite as the equivalent expression $\sum_{x_i=0}^{\infty} g_i^{x_i}(x_i+1)P(x_i+1)$. Therefore, the generating function of all terms that scale as $(x_i+1)P(x_i+1)$ is $\frac{\partial G_s}{\partial g_i}$.

Multiplying this equation through by g_j:

$$\begin{aligned} g_j \frac{\partial G}{\partial g_i} &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} g_j (x_i+1) P(x_i+1) d\mathbf{y} \\ &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} (x_i+1) P(x_i+1, x_j-1) d\mathbf{y}. \end{aligned} \tag{A.5}$$

This follows from rewriting $\sum_{x_j=0}^{\infty} g_j^{x_j+1} P(x_i+1, j)$ as $\sum_{x_j=-1}^{\infty} g_j^{x_j+1} P(x_i+1, x_j)$, noting that $P(x_j) = 0$ whenever $x_j < 0$, and reindexing to obtain the equivalent sum $\sum_{x_j=0}^{\infty} g_j^{x_j} P(x_i+1, x_j-1)$. Therefore, the generating function of all terms that scale as $(x_i+1)P(x_i+1, x_j-1)$ is $g_j \frac{\partial G}{\partial g_i}$.

Multiplying the derivative by g_i twice:

$$\begin{aligned} g_i^2 \frac{\partial G}{\partial g_i} &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} x_i g_i P d\mathbf{y} \\ &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} (x_i-1) P(x_i-1) d\mathbf{y}. \end{aligned} \tag{A.6}$$

This follows from rewriting $\sum_{x_i=0}^{\infty} g_i^{x_i+1} x_i P(x_i)$ as $\sum_{x_i=-1}^{\infty} g_i^{x_i+1} x_i P(x_i)$, noting that $P(x_i) = 0$ whenever $x_i < 0$, and reindexing to obtain the equivalent sum $\sum_{x_i=0}^{\infty} g_i^{x_i}(x_i-1)P(x_i-1)$. Therefore, the generating function of all terms that scale as $(x_i-1)P(x_i-1)$ is $g_i^2 \frac{\partial G}{\partial g_i}$.

Multiplying the derivative by $g_i g_j$:

$$\begin{aligned} g_i g_j \frac{\partial G}{\partial g_i} &= g_j \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} x_i P d\mathbf{y} \\ &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} x_i g_j P d\mathbf{y} \\ &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} x_i P(x_j-1) d\mathbf{y}. \end{aligned} \tag{A.7}$$

This follows from rewriting $\sum_{x_i=0}^{\infty} g_j^{x_j+1} P(x_j)$ as $\sum_{x_j=-1}^{\infty} g_j^{x_j+1} P(x_j)$, noting that $P(x_j) = 0$ whenever $x_j < 0$, and reindexing to obtain the equivalent sum $\sum_{x_j=0}^{\infty} g_j^{x_j} P(x_j)$. Therefore, the generating function of all terms that scale as $x_i P(x_j-1)$ is $g_i g_j \frac{\partial G}{\partial g_i}$.

Multiplying the generating function by a probability-generating function F of a discrete burst distribution p:

$$\begin{aligned} FG &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\intercal \mathbf{y}} F P d\mathbf{y} \\ &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\intercal \mathbf{y}} \sum_{\mathbf{z}} p(\mathbf{z}) P(\mathbf{x}-\mathbf{z}) d\mathbf{y}. \end{aligned} \tag{A.8}$$

This identity may be derived from three equivalent directions. Most simply, it is a statement of the convolution theorem. Alternatively, it may be proven directly using the repeated (n-fold) application of Cauchy products. Finally, the master equation term essentially aggregates two independent random variables – the process values and the burst sizes – whose sum is equal to $\mathbf{x}$. The generating function of the sum of independent variates is the product of their generating functions. Therefore, the generating function of all terms that scale as $\sum_{\mathbf{z}} p(\mathbf{z}) P(\mathbf{x}-\mathbf{z})$ is FG.

A.3 Fully continuous master equation terms

Multiplying G through by h_i, we obtain:

$$\begin{aligned} h_i G &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\intercal \mathbf{y}} h_i P d\mathbf{y} = \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} \frac{\partial}{\partial y_i} \left[e^{\mathbf{h}^\intercal \mathbf{y}} \right] [P] \, d\mathbf{y} \\ &= -\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\intercal \mathbf{y}} \frac{\partial P}{\partial y_i} d\mathbf{y}. \end{aligned} \tag{A.9}$$

This follows from integrating by parts. The product term $e^{\mathbf{h}^\intercal \mathbf{y}} P$ does not contribute to this expression because P is a density with zero mass at any particular value of $\mathbf{y}$. Therefore, the generating function of all terms that scale as $\frac{\partial P}{\partial y_i}$ is $-h_i G$.

Differentiating with respect to h_i:

$$\frac{\partial G}{\partial h_i} = \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\intercal \mathbf{y}} y_i P d\mathbf{y}. \tag{A.10}$$

Therefore, the generating function of all terms that scale as $y_i P$ is $\frac{\partial G_s}{\partial h_i}$.

Multiplying through by h_j:

$$\begin{aligned} h_j \frac{\partial G}{\partial h_i} &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\intercal \mathbf{y}} h_j y_i P d\mathbf{y} = \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} \frac{\partial}{\partial y_j} \left[e^{\mathbf{h}^\intercal \mathbf{y}} \right] [y_i P] \, d\mathbf{y} \\ &= -\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\intercal \mathbf{y}} \frac{\partial \left[y_i P \right]}{\partial y_j} d\mathbf{y}. \end{aligned} \tag{A.11}$$

This follows from integrating by parts. The product term $e^{\mathbf{h}^\top \mathbf{y}} y_i P$ does not contribute to this expression because it is identically zero at $y_i = 0$ and P vanishes as $y_i \to \infty$. Therefore, the generating function of all terms that scale as $\frac{\partial [y_i P]}{\partial y_j}$ is $-h_j \frac{\partial G}{\partial h_i}$.

Multiplying the derivative by h_i twice:

$$\begin{aligned} h_i^2 \frac{\partial G}{\partial h_i} &= -h_i \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} \frac{\partial \left[y_i P \right]}{\partial y_i} d\mathbf{y} \\ &= -\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} \frac{\partial}{\partial y_i} \left[e^{\mathbf{h}^\top \mathbf{y}} \right] \frac{\partial \left[y_i P \right]}{\partial y_i} d\mathbf{y} \\ &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} \frac{\partial^2 \left[y_i P \right]}{\partial y_i^2} d\mathbf{y}. \end{aligned} \tag{A.12}$$

This follows from integrating by parts. Again, the product term $e^{\mathbf{h}^\top \mathbf{y}} \frac{\partial [y_i P]}{\partial y_i} = e^{\mathbf{h}^\top \mathbf{y}} y_i \frac{\partial P}{\partial y_i} + e^{\mathbf{h}^\top \mathbf{y}} P$ does not contribute to this expression because P is a density that vanishes as $y_i \to \infty$. Therefore, the generating function of all terms that scale as $\frac{\partial^2 [y_i P]}{\partial y_i^2}$ is $h_i^2 \frac{\partial G}{\partial h_i}$.

Multiplying the generating function by a moment-generating function M of continuous burst distribution p:

$$MG = \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} M P d\mathbf{y} = \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} \int_{\mathbf{z}} p(\mathbf{z}) P(\mathbf{y} - \mathbf{z}) d\mathbf{z} d\mathbf{y}, \tag{A.13}$$

which may be derived from the convolution theorem, or MGF identities, identically to Equation A.8. Therefore, the generating function of all terms that scale as $p(\mathbf{z}) P(\mathbf{y} - \mathbf{z}) d\mathbf{z}$ is MG.

A.4 Mixed master equation terms

Considering the case where a continuous process drives a discrete one, and multiplying $\frac{\partial G}{\partial h_i}$ by g_j:

$$\begin{aligned} g_j \frac{\partial G}{\partial h_i} &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} y_i g_j P d\mathbf{y} \\ &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} y_i P (x_j - 1) d\mathbf{y}. \end{aligned} \tag{A.14}$$

The derivation is identical to Equation A.3. Therefore, the generating function of all terms that scale as $y_i P (x_j - 1)$ is $g_j \frac{\partial G}{\partial h_i}$.

Considering the case where a discrete process drives a continuous one, and multiplying $g_i \frac{\partial G}{\partial g_i}$ by h_j:

$$\begin{aligned} h_j g_i \frac{\partial G}{\partial g_i} &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} x_i h_j P d\mathbf{y} \\ &= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} \frac{\partial}{\partial y_j} \left[e^{\mathbf{h}^\top \mathbf{y}} \right] [x_i P] \, d\mathbf{y} \\ &= - \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} \frac{\partial \, [x_i P]}{\partial y_j} d\mathbf{y}. \end{aligned} \tag{A.15}$$

This follows from integrating by parts; as before, the product term does not appear because P is a density. Therefore, the generating function of all terms that scale as $\frac{\partial [x_i P]}{\partial y_j} = x_i \frac{\partial P}{\partial y_j}$ is $-h_j g_i \frac{\partial G}{\partial g_i}$. This concludes the enumeration of generating function identities.

A.5 Converting the master equation to a partial differential equation

By exploiting the identities derived above and the linearity of the generating function, we can represent Equation A.1 by an equivalent deterministic partial differential equation. We begin by considering the expressions for each entry of $\mathbf{G}$ separately, eliding the gene state s.

Each entry of the second term on the right-hand side, which represents degradation of the discrete species, takes the form

$$\begin{aligned} &\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} \left[(x_i + 1) P(x_i + 1) - x_i P \right] d\mathbf{y} \\ &= \frac{\partial G}{\partial g_i} - g_i \frac{\partial G}{\partial g_i} = (1 - g_i) \frac{\partial G}{\partial g_i}, \text{ yielding} \\ &\sum_{i=1}^{n} c_{i0} (1 - g_i) \frac{\partial G}{\partial g_i}. \end{aligned} \tag{A.16}$$

Each entry of the third term, which represents interconversion of the discrete species, takes the form

$$\begin{aligned} &\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^\top \mathbf{y}} \left[(x_i + 1) P(x_i + 1, x_j - 1) - x_i P \right] d\mathbf{y} \\ &= g_j \frac{\partial G}{\partial g_i} - g_i \frac{\partial G}{\partial g_i} = (g_j - g_i) \frac{\partial G}{\partial g_i}, \text{ yielding} \\ &\sum_{i,j=1}^{n} c_{ij} (g_j - g_i) \frac{\partial G}{\partial g_i}. \end{aligned} \tag{A.17}$$

Each entry of the fourth term, which represents autocatalysis of the discrete species, takes the form

$$\begin{aligned}&\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^{\mathsf{T}}\mathbf{y}} \left[(x_i - 1)P(x_i - 1) - x_i P\right] d\mathbf{y} \\ &= g_i^2 \frac{\partial G}{\partial g_i} - g_i \frac{\partial G}{\partial g_i} = (g_i - 1) g_i \frac{\partial G}{\partial g_i}, \text{ yielding} \\ &\sum_{i=1}^{n} Q_{ii}^{d} (g_i - 1) g_i \frac{\partial G}{\partial g_i}.\end{aligned} \tag{A.18}$$

Each entry of the fifth term, which represents catalysis of the discrete species, takes the form

$$\begin{aligned}&\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^{\mathsf{T}}\mathbf{y}} \left[x_i P(x_j - 1) - x_i P\right] d\mathbf{y} \\ &= g_i g_j \frac{\partial G}{\partial g_i} - g_i \frac{\partial G}{\partial g_i} = (g_j - 1) g_i \frac{\partial G}{\partial g_i}, \text{ yielding} \\ &\sum_{i,j=1}^{n} Q_{ij}^{d} (g_j - 1) g_i \frac{\partial G}{\partial g_i}.\end{aligned} \tag{A.19}$$

Each entry of the sixth term, which represents bursty production of the discrete species, takes the form

$$\begin{aligned}&\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^{\mathsf{T}}\mathbf{y}} \left[\sum_{\mathbf{z}} p_{s,\omega}^{d}(\mathbf{z}, t) P(\mathbf{x} - \mathbf{z}) - P\right] d\mathbf{y} \\ &= FG - G = (F - 1)G, \text{ yielding} \\ &\sum_{\omega} \alpha_{s,\omega}^{d}(t)(F_{s,\omega} - 1)G.\end{aligned} \tag{A.20}$$

Each entry of the seventh term, which represents the deterministic dynamics of the continuous species, takes the form

$$\begin{aligned}&\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^{\mathsf{T}}\mathbf{y}} \frac{\partial}{\partial y_j} \left[y_i P\right] d\mathbf{y} \\ &= -h_j \frac{\partial G}{\partial h_i}, \text{ yielding} \\ &\sum_{i,j=1}^{m} C_{ij}^{cc} h_j \frac{\partial G}{\partial h_i}.\end{aligned} \tag{A.21}$$

Each entry of the eighth term, which represents the diffusion dynamics of the continuous species, takes the form

$$\begin{aligned}&\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^{\top}\mathbf{y}} \frac{\partial^2}{\partial y_i^2} [y_i P] \, d\mathbf{y} \\ &= h_i^2 \frac{\partial G}{\partial h_i}, \text{ yielding} \\ &\frac{1}{2} \sum_{i=1}^{m} \sigma_i^2 h_i^2 \frac{\partial G}{\partial h_i}.\end{aligned} \tag{A.22}$$

Each entry of the ninth term, which represents the drift of the continuous species, takes the form

$$\begin{aligned}&\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^{\top}\mathbf{y}} \frac{\partial P}{\partial y_i} d\mathbf{y} \\ &= -h_i G, \text{ yielding} \\ &\sum_{i=1}^{m} \alpha_{s,i}^{c} h_i G.\end{aligned} \tag{A.23}$$

Each entry of the tenth term, which represents the bursty production of the continuous species, takes the form

$$\begin{aligned}&\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^{\top}\mathbf{y}} \left[\int_{\mathbf{z}} p_{\omega}^{c}(\mathbf{z}) P(\mathbf{y} - \mathbf{z}, t) d\mathbf{z} - P \right] d\mathbf{y} \\ &= MG - G = (M-1)G, \text{ yielding} \\ &\sum_{\omega > m} \alpha_{s,\omega}^{c} (M_{s,\omega} - 1) G.\end{aligned} \tag{A.24}$$

Each entry of the eleventh term, which represents a continuous species driving a discrete one, takes the form

$$\begin{aligned}&\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^{\top}\mathbf{y}} \left[y_i P(x_j - 1) - y_i P \right] d\mathbf{y} \\ &= g_j \frac{\partial G}{\partial h_i} - \frac{\partial G}{\partial h_i} = (g_j - 1) \frac{\partial G}{\partial h_i}, \text{ yielding} \\ &\sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij}^{cd} (g_j - 1) \frac{\partial G}{\partial h_i}.\end{aligned} \tag{A.25}$$

Each entry of the twelfth term, which represents a discrete species driving a continuous one, takes the form

$$\begin{aligned}&\int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^{\mathsf{T}}\mathbf{y}} x_i \frac{\partial P}{\partial y_j} d\mathbf{y} \\ &= h_j g_i \frac{\partial G}{\partial g_i}, \text{ yielding} \\ &\sum_{i=1}^{n}\sum_{j=1}^{m} C_{ij}^{dc} h_j g_i \frac{\partial G}{\partial g_i}.\end{aligned} \tag{A.26}$$

Therefore, the PDE form of the master equation is

$$\begin{aligned}\frac{\partial G}{\partial t} = &\sum_{i=1}^{N} H_{is} G_i + \sum_{i=1}^{n} c_{i0}(1-g_i)\frac{\partial G}{\partial g_i} + \sum_{i,j=1}^{n} c_{ij}(g_j - g_i)\frac{\partial G}{\partial g_i} \\ &+ \sum_{i=1}^{n} Q_{ii}^{d}(g_i - 1) g_i \frac{\partial G}{\partial g_i} + \sum_{i,j=1}^{n} Q_{ij}^{d}(g_j - 1) g_i \frac{\partial G}{\partial g_i} + \sum_{\omega} \alpha_{s,\omega}^{d}(F_{s,\omega} - 1)G \\ &+ \sum_{i,j=1}^{m} C_{ij}^{cc} h_j \frac{\partial G}{\partial h_i} + \frac{1}{2}\sum_{i=1}^{m} \sigma_i^2 h_i^2 \frac{\partial G}{\partial h_i} + \sum_{i=1}^{m} \alpha_{s,i}^{c} h_i G + \sum_{\omega > m} \alpha_{s,\omega}^{c}(M_{s,\omega} - 1)G \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}^{cd}(g_j - 1)\frac{\partial G}{\partial h_i} + \sum_{i=1}^{n}\sum_{j=1}^{m} C_{ij}^{dc} h_j g_i \frac{\partial G}{\partial g_i}.\end{aligned} \tag{A.27}$$

This equation governs the dynamics of G_s, one of N coupled PDEs. We elide this subscript when it does not directly factor into the calculation. As in Equation A.1, the terms that scale with G, rather than one of its derivatives, may be time- and s-dependent.

A.6 Representing the PDE in matrix form

This formulation in Equation A.27 is somewhat more compact, but may be simplified further. First, we note that the second and third terms on the right-hand side can be represented by the matrix equation

$$(\nabla^d G)^{\mathsf{T}} C^{dd} (\mathbf{g} - 1), \tag{A.28}$$

where ∇^d is the length-n column vector gradient with respect to entries of $\mathbf{g}$.

Next, the fourth and fifth terms can be represented by

$$(\nabla^d G)^{\mathsf{T}} (\text{diag } \mathbf{g}) Q^d (\mathbf{g} - 1). \tag{A.29}$$

The sixth term can be represented by constructing the vector $\boldsymbol{\alpha}^d$ and vector function $\mathbf{F}$, indexed by ω, with elided dependence on s:

$$G(\boldsymbol{\alpha}^d)^\mathsf{T}(\mathbf{F}(\mathbf{g}) - 1). \tag{A.30}$$

The seventh term can be represented similarly to second and third:

$$(\nabla^c G)^\mathsf{T} C^{cc}\mathbf{h}, \tag{A.31}$$

where ∇^c is the gradient with respect to entries of $\mathbf{h}$.

The eighth term can be represented analogously to the fourth and fifth:

$$(\nabla^c G)^\mathsf{T}(\operatorname{diag}\mathbf{h})\left(\frac{1}{2}\operatorname{diag}\boldsymbol{\sigma}^2\right)\mathbf{h}. \tag{A.32}$$

The ninth and tenth term can be represented analogously to the sixth; we construct vector $\boldsymbol{\alpha}^c$ and vector function $\mathbf{M}$, indexed by ω, with elided dependence on s:

$$G(\boldsymbol{\alpha}^c)^\mathsf{T}(\mathbf{M}(\mathbf{h}) - 1), \tag{A.33}$$

The first m entries of $\boldsymbol{\alpha}^c$ contain the m scalar drift rates, whereas the other entries contain jump rates, i.e., $(\mathbf{M})_i := h_i + 1$ for $i \leq m$.

The eleventh term takes a form analogous to those for second, third, and seventh:

$$(\nabla^c G)^\mathsf{T} C^{cd}(\mathbf{g} - 1). \tag{A.34}$$

Finally, the twelfth is analogous to fourth and fifth:

$$(\nabla^d G)^\mathsf{T}(\operatorname{diag}\mathbf{g})\, C^{dc}\, \mathbf{h}. \tag{A.35}$$

Therefore, Equation A.27 can be condensed further for a particular s:

$$\begin{aligned}\frac{\partial G}{\partial t} = &\sum_{i=1}^{N} H_{is}G_i \\ &+ (\nabla^d G)^\mathsf{T} C^{dd}(\mathbf{g} - 1) + (\nabla^d G)^\mathsf{T}(\operatorname{diag}\mathbf{g})\, Q^d(\mathbf{g} - 1) + G(\boldsymbol{\alpha}^d)^\mathsf{T}(\mathbf{F}(\mathbf{g}) - 1) \\ &+ (\nabla^c G)^\mathsf{T} C^{cc}\mathbf{h} + (\nabla^c G)^\mathsf{T}(\operatorname{diag}\mathbf{h})\left(\frac{1}{2}\operatorname{diag}\boldsymbol{\sigma}^2\right)\mathbf{h} + G(\boldsymbol{\alpha}^c)^\mathsf{T}(\mathbf{M}(\mathbf{h}) - 1) \\ &+ (\nabla^c G)^\mathsf{T} C^{cd}(\mathbf{g} - 1) + (\nabla^d G)^\mathsf{T}(\operatorname{diag}\mathbf{g})\, C^{dc}\, \mathbf{h}.\end{aligned} \tag{A.36}$$

Collecting terms:

$$\begin{aligned}\frac{\partial G}{\partial t} &= (\nabla^d G)^\mathsf{T}\left[C^{dd}(\mathbf{g}-1) + (\operatorname{diag}\mathbf{g})\,Q^d(\mathbf{g}-1) + (\operatorname{diag}\mathbf{g})\,C^{dc}\,\mathbf{h}\right] \\ &\quad + (\nabla^c G)^\mathsf{T}\left[C^{cc}\mathbf{h} + (\operatorname{diag}\mathbf{h})\left(\frac{1}{2}\operatorname{diag}\boldsymbol{\sigma}^2\right)\mathbf{h} + C^{cd}(\mathbf{g}-1)\right] \\ &\quad + G\left[(\boldsymbol{\alpha}^d)^\mathsf{T}(\mathbf{F}(\mathbf{g})-1) + (\boldsymbol{\alpha}^c)^\mathsf{T}(\mathbf{M}(\mathbf{h})-1)\right] + \sum_{i=1}^{N} H_{is}G_i.\end{aligned} \tag{A.37}$$

In other words, the PDE separates into the usual first-order linear form, with terms corresponding to G, $\nabla^d G$, and $\nabla^c G$.

A.6.1 Unifying the discrete and continuous species

However, analyzing Equation A.37 as is obfuscates the mathematical similarities of the discrete and continuous species.

To exploit them, we first introduce the variable $\mathbf{u}$, which is a shifted version of $\mathbf{g}$ concatenated to $\mathbf{h}$:

$$\mathbf{u} := \begin{bmatrix}\mathbf{u}_d \\ \mathbf{u}_c\end{bmatrix} = \begin{bmatrix}\mathbf{g}-1 \\ \mathbf{h}\end{bmatrix}. \tag{A.38}$$

This yields the following form for the gradient-dependent terms:

$$\begin{aligned}&(\nabla^d G)^\mathsf{T}\left[C^{dd}\mathbf{u}_d + (\operatorname{diag}\mathbf{u}_d + I)Q^d\mathbf{u}_d + (\operatorname{diag}\mathbf{u}_d + I)C^{dc}\mathbf{u}_c\right] \\ &+ (\nabla^c G)^\mathsf{T}\left[C^{cc}\mathbf{u}_c + (\operatorname{diag}\mathbf{u}_c)\left(\frac{1}{2}\operatorname{diag}\boldsymbol{\sigma}^2\right)\mathbf{u}_c + C^{cd}\mathbf{u}_d\right] \\ &= (\nabla^d G)^\mathsf{T}\left[\left(C^{dd} + Q^d\right)\mathbf{u}_d + C^{dc}\mathbf{u}_c + (\operatorname{diag}\mathbf{u}_d)Q^d\mathbf{u}_d + (\operatorname{diag}\mathbf{u}_d)C^{dc}\mathbf{u}_c\right] \\ &+ (\nabla^c G)^\mathsf{T}\left[C^{cd}\mathbf{u}_d + C^{cc}\mathbf{u}_c + (\operatorname{diag}\mathbf{u}_c)\left(\frac{1}{2}\operatorname{diag}\boldsymbol{\sigma}^2\right)\mathbf{u}_c\right].\end{aligned} \tag{A.39}$$

Next, we define the full gradient of G, such that ∇G contains the derivatives with respect to all entries of $\mathbf{u}$. We define common jump rates and generating functions:

$$\begin{aligned}\boldsymbol{\alpha} &:= \begin{bmatrix}\boldsymbol{\alpha}^d \\ \boldsymbol{\alpha}^c\end{bmatrix} \\ \mathcal{M}(\mathbf{u}) &:= \begin{bmatrix}\mathbf{F}(1 + \mathbf{u}_{1,\cdots,n}) \\ \mathbf{M}(\mathbf{u}_{n+1,\cdots,n+m})\end{bmatrix}.\end{aligned} \tag{A.40}$$

This notation emphasizes that $\mathcal{M}$ is formally defined over all values of $\mathbf{u}$; however, each entry, which corresponds to a specific influx process (i.e., a burst, drift, or jump

term) possesses nontrivial dependence only on the relevant (discrete or continuous) indices.

We define the common interconversion matrix:

$$C := \begin{bmatrix} C^{dd} + Q^d & C^{dc} \\ C^{cd} & C^{cc} \end{bmatrix}, \tag{A.41}$$

as well as the common diffusion matrix:

$$D := \begin{bmatrix} Q^d & C^{dc} \\ 0 & \frac{1}{2}\text{diag } \boldsymbol{\sigma}^2 \end{bmatrix} := \begin{bmatrix} Q^d & C^{dc} \\ 0 & Q^c \end{bmatrix}. \tag{A.42}$$

This yields the following unified expression for a single state:

$$\frac{\partial G}{\partial t} = (\nabla G)^\mathsf{T} \left[C\mathbf{u} + \text{diag } \mathbf{u}\, D\mathbf{u}\right] + G\left[\boldsymbol{\alpha}^\mathsf{T}(\mathcal{M}(\mathbf{u}) - 1)\right] + \sum_{i=1}^{N} H_{is} G_i. \tag{A.43}$$

In this equation, ∇G is the gradient of G with respect to $\mathbf{u}$. It remains to generalize this expression to multiple states. Of the biological parameters, only the entries of H, $\boldsymbol{\alpha}$, and $\mathcal{M}$ depend on gene state. To specify the influx dynamics, we need to define the full bursting operator, which is a length-N vector function:

$$\mathcal{A}(\mathbf{u}) = \begin{bmatrix} \boldsymbol{\alpha}_1^\mathsf{T}(\mathcal{M}_1(\mathbf{u}) - 1) \\ \cdots \\ \boldsymbol{\alpha}_N^\mathsf{T}(\mathcal{M}_N(\mathbf{u}) - 1) \end{bmatrix}, \tag{A.44}$$

where the subscripts of $\boldsymbol{\alpha}$ and $\mathcal{M}$ now indicate the gene state. Finally, recalling that the full Jacobian has entries $J_{si} = \frac{\partial G_s}{\partial u_i}$, the full PDE system takes the following form:

$$\frac{\partial \mathbf{G}}{\partial t} = H^\mathsf{T}\mathbf{G} + \mathbf{G} \odot \mathcal{A}(\mathbf{u}) + J\left[C\mathbf{u} + \text{diag } \mathbf{u}\, D\mathbf{u}\right]. \tag{A.45}$$

A.6.2 Solving the partial differential equation

We seek to integrate this PDE to obtain the generating function at an arbitrary time t. The form is conducive to applying the method of characteristics. First, we define the characteristic variable s. By taking a total derivative with respect to s, we obtain

$$\frac{dG_s}{d\mathsf{s}} = \frac{\partial G_s}{\partial T}\frac{dT}{d\mathsf{s}} + \sum_{i=1}^{n+m} \frac{\partial G_s}{\partial U_i}\frac{dU_i}{d\mathsf{s}}. \tag{A.46}$$

Next, we rewrite the PDE to match the form of the total derivative:

$$-H^T\mathbf{G} - \mathbf{G} \odot \mathcal{A}(\mathbf{u}) = -\frac{\partial \mathbf{G}}{\partial t} + J\left[C\mathbf{u} + \text{diag}\ \mathbf{u}\ D\mathbf{u}\right]. \tag{A.47}$$

The characteristic curves emanating from $(t, \mathbf{u})$ that satisfy the PDE are given by:

$$\begin{aligned}
&\frac{dT(\mathsf{s})}{d\mathsf{s}} = -1 \text{ such that } T(\mathsf{s}=0) = t, \text{ i.e., } T(\mathsf{s}) = t - \mathsf{s} \text{ and} \\
&\frac{dU_i(\mathsf{s})}{d\mathsf{s}} = \left(C\mathbf{U}(\mathsf{s}) + \text{diag}\ \mathbf{U}(\mathsf{s})\ D\mathbf{U}(\mathsf{s})\right)_i \text{ such that } U_i(\mathsf{s}=0) = u_i, \text{ i.e.,} \\
&\frac{d\mathbf{U}(\mathsf{s})}{d\mathsf{s}} = C\mathbf{U}(\mathsf{s}) + \text{diag}\ \mathbf{U}(\mathsf{s})\ D\mathbf{U}(\mathsf{s}) \text{ such that } \mathbf{U}(\mathsf{s}=0) = \mathbf{u}.
\end{aligned} \tag{A.48}$$

This is the "downstream" ODE, which governs abundances in isolation from production and regulation.

Therefore, $\mathbf{G}$ is governed by the following system of ordinary differential equations:

$$\frac{d\mathbf{G}(\mathbf{U}(\mathsf{s}), T(\mathsf{s}))}{d\mathsf{s}} = -H(T(\mathsf{s}))^T\mathbf{G} - \mathbf{G} \odot \mathcal{A}(\mathbf{U}(\mathsf{s}), T(\mathsf{s})) := \mathcal{H}(\mathbf{U}, T)\ \mathbf{G}. \tag{A.49}$$

To obtain $\mathbf{G}$ at t, we integrate this matrix system from $\mathsf{s} = t$ to $\mathsf{s} = 0$. We use $\mathbf{G}^0(\mathbf{U}(t))$ as the initial condition, where $\mathbf{G}^0$ is the generating function of the initial distribution. This is the "upstream" ODE, which governs the full generating function.

A.7 Regulation extensions

This section summarizes some investigations undertaken during the writing of [115] by G.G., J.J.V., and L.P. This derivation was performed by G.G.

We have summarized a considerable breadth of biological phenomena in a common framework. Yet the really "interesting" ones, such as regulation, are still elusive. To see why, we can formalize the challenges of feedback, using the $m = 0$, $N > 1$, $n > 0$ case as an example.

First, we write down the master equation. We are interested in *non-sequestering* catalysis of state switching, such that reactions of the form

$$\mathcal{S}_i \xrightarrow{x_k R_{k,ij}} \mathcal{S}_j \tag{A.50}$$

are allowed, whereas reactions of the form

$$\mathcal{S}_i + R_{k,ij} \xrightarrow{x_k R_{k,ij}} \mathcal{S}_j \tag{A.51}$$

are disallowed. This is mostly a mathematical convenience: in addition to catalysis, we would like to retain the usual non-catalytic switching (encoded in a matrix H), and restricting the allowed reactions in this way avoid strange and nonphysical edge cases in the vein of

$$\mathcal{S}_i \to \mathcal{S}_j \to \mathcal{S}_i + \mathcal{X}_k, \tag{A.52}$$

i.e., the spontaneous generation of molecules[9].

The master equation terms corresponding to the switching reactions are

$$\sum_{k=1}^{n} x_k \left[\sum_{i=1}^{N} R_{k,is} P(i, \mathbf{x}, t) - \sum_{i=1}^{N} R_{k,si} P(s, \mathbf{x}, t) \right], \tag{A.53}$$

i.e., any species, indexed by k, can, in principle, catalyze any transition between states. Ostensibly, the summation excludes self-transitions. From Equation A.4, we immediately obtain that the corresponding generating function terms are

$$\sum_{k=1}^{n} g_k \left[\sum_{i=1}^{N} R_{k,is} \frac{\partial G_i}{\partial g_k} - \sum_{i=1}^{N} R_{k,si} \frac{\partial G_s}{\partial g_k} \right]. \tag{A.54}$$

Defining the diagonal elements of R_k in the usual fashion, such that $R_{k,ss} := \sum_{i \neq s} R_{k,si}$, we yield the matrix form

$$\sum_{k=1}^{n} g_k R_k^{\mathsf{T}} \frac{\partial \mathbf{G}}{\partial g_k}, \tag{A.55}$$

where the partial derivative is elementwise. In the case of $n = 1$, this reduces to the somewhat simpler case

$$\begin{aligned} &gR^{\mathsf{T}}J, \text{ yielding the full expression} \\ \frac{\partial \mathbf{G}}{\partial t} &= H^{\mathsf{T}}\mathbf{G} + \mathbf{G} \odot \mathcal{A}(u) + J\left[Cu + Du^2\right] + (u+1)R^{\mathsf{T}}J. \end{aligned} \tag{A.56}$$

The *right*-multiplication by the Jacobian matrix J makes all the difference: the system ceases to be tractable by the method of characteristics, as we cannot specify a "downstream" component.

We can illustrate this point more easily by considering the usual $N = 2$, $D = 0$ case with Poisson process transcription in the on state [301]. This yields

$$\begin{aligned} \frac{\partial \mathbf{G}}{\partial t} &= H^{\mathsf{T}}\mathbf{G} + \boldsymbol{\alpha} \odot \mathbf{G}u + JCu + (u+1)R^{\mathsf{T}}J \\ &= \begin{bmatrix} -k_{\text{on}} & k_{\text{off}} \\ k_{\text{on}} & -k_{\text{off}} \end{bmatrix} \begin{bmatrix} G_{\text{off}} \\ G_{\text{on}} \end{bmatrix} + \begin{bmatrix} 0 \\ k_{\text{init}} G_{\text{on}} \end{bmatrix} u - \gamma u \begin{bmatrix} \frac{\partial G_{\text{off}}}{\partial u} \\ \frac{\partial G_{\text{on}}}{\partial u} \end{bmatrix} \\ &+ (u+1) \begin{bmatrix} -R_{\text{on}} & R_{\text{off}} \\ R_{\text{on}} & -R_{\text{off}} \end{bmatrix} \begin{bmatrix} \frac{\partial G_{\text{off}}}{\partial u} \\ \frac{\partial G_{\text{on}}}{\partial u} \end{bmatrix}, \end{aligned} \tag{A.57}$$

where R_{on} and R_{off} are the mass action rates of transition catalysis. This matrix equation is equivalent to the system

$$\begin{aligned} \frac{\partial G_{\text{off}}}{\partial t} &= -k_{\text{on}} G_{\text{off}} + k_{\text{off}} G_{\text{on}} - \gamma u \frac{\partial G_{\text{off}}}{\partial u} \\ &\quad - R_{\text{on}}(u+1)\frac{\partial G_{\text{off}}}{\partial u} + R_{\text{off}}(u+1)\frac{\partial G_{\text{on}}}{\partial u} \\ \frac{\partial G_{\text{on}}}{\partial t} &= k_{\text{on}} G_{\text{off}} - k_{\text{off}} G_{\text{on}} - \gamma u \frac{\partial G_{\text{on}}}{\partial u} + k_{\text{init}} u G_{\text{on}} \\ &\quad + R_{\text{on}}(u+1)\frac{\partial G_{\text{off}}}{\partial u} - R_{\text{off}}(u+1)\frac{\partial G_{\text{on}}}{\partial u}, \end{aligned} \tag{A.58}$$

which combines the autoactivation and autorepression cases in Equations 2.7 and 2.8 of [301]. Parenthetically, we question the authors' justification for treating these phenomena as mutually exclusive because they "cancel out." For example, if $R_{\text{on}} = R_{\text{off}}$, but both are both extremely high, we obtain a trivial Poisson distribution of RNA, which is qualitatively different from the possibly bimodal $R_{\text{on}} = R_{\text{off}} = 0$ case. In addition, this justification ceases to hold when considering $N > 2$.

Let us treat the simplest case, with $k_{\text{on}} = k_{\text{off}} = \gamma = 0$. Then we obtain

$$\begin{aligned}
\frac{\partial G_{\text{off}}}{\partial t} &= -R_{\text{on}}(u+1)\frac{\partial G_{\text{off}}}{\partial u} + R_{\text{off}}(u+1)\frac{\partial G_{\text{on}}}{\partial u} \\
\frac{\partial G_{\text{on}}}{\partial t} &= k_{\text{init}} u G_{\text{on}} + R_{\text{on}}(u+1)\frac{\partial G_{\text{off}}}{\partial u} - R_{\text{off}}(u+1)\frac{\partial G_{\text{on}}}{\partial u}, \text{ i.e.,} \\
0 &= \frac{\partial \mathbf{G}}{\partial t} - (u+1)R^{\mathsf{T}}\frac{\partial \mathbf{G}}{\partial u} - \text{diag}\,\boldsymbol{\alpha}\,\mathbf{G}u.
\end{aligned} \tag{A.59}$$

The next steps are somewhat obscure. We can use eigendecomposition:

$$\begin{aligned}
-(u+1)R^{\mathsf{T}} &= V\Lambda V^{-1} \\
V &= \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 \end{bmatrix} = \begin{bmatrix} R_{\text{off}} & -1 \\ R_{\text{on}} & 1 \end{bmatrix} \\
\Lambda &= \begin{bmatrix} 0 & 0 \\ 0 & (R_{\text{on}} + R_{\text{off}})(u+1) \end{bmatrix}.
\end{aligned} \tag{A.60}$$

Defining $V\tilde{\mathbf{G}} := \mathbf{G}$, we obtain

$$0 = V\frac{\partial \tilde{\mathbf{G}}}{\partial t} + V\Lambda V^{-1} V\frac{\partial \tilde{\mathbf{G}}}{\partial u} - u\,\text{diag}\,\boldsymbol{\alpha} V\tilde{\mathbf{G}}, \tag{A.61}$$

and multiplying by V^{-1} from the left,

$$\begin{aligned}
0 &= \frac{\partial \tilde{\mathbf{G}}}{\partial t} + \Lambda\frac{\partial \tilde{\mathbf{G}}}{\partial u} - uV^{-1}\,\text{diag}\,\boldsymbol{\alpha} V\tilde{\mathbf{G}} \\
&:= \frac{\partial \tilde{\mathbf{G}}}{\partial t} + \Lambda\frac{\partial \tilde{\mathbf{G}}}{\partial u} - A\tilde{\mathbf{G}}.
\end{aligned} \tag{A.62}$$

This produces *two* characteristic curves, defined by

$$\begin{aligned}
\frac{\partial U_1}{\partial \mathsf{s}} &= 0 \rightarrow U_1 = u \\
\frac{\partial U_2}{\partial \mathsf{s}} &= (R_{\text{on}} + R_{\text{off}})(U_2 + 1) \rightarrow U_2 = (u+1)e^{(R_{\text{on}}+R_{\text{off}})\mathsf{s}} - 1.
\end{aligned} \tag{A.63}$$

Although this functional form is certainly precedented in the study of partial differential equations (see, e.g., Section 22.4 of [79] and Section 2.5 of [153]), it does not appear to lead to a closed-form solution, and the development of a reasonably generic procedure for treating regulation in the same framework remains out of reach for now. Although we do not treat the more general cases with nontrivial H and γ, they produce largely the same challenges.

A.8 Stochastic process identities

In this section, we outline some useful identities and demonstrate the mathematical capabilities of the current approach.

A.8.1 The telegraph process converges to the jump subordinator

This section adapts a portion of the supplement of [105] by G.G. and L.P. This derivation was performed by G.G.

In Equation 4.22, we have summarized the "upstream" degrees of freedom in terms of a state interconversion matrix H and a state-dependent transcription operator $\mathcal{A}$. The entries of the operator essentially encode the distribution of a memoryless Poisson arrival process. This process, in turn, arises as the approximation of a timescale-separated process; for example, it is well-understood [233, 267] that the reaction schema

$$\begin{aligned} \mathcal{S}_{\text{off}} &\xrightarrow{k_{\text{on}}} \mathcal{S}_{\text{on}} \\ \mathcal{S}_{\text{on}} &\xrightarrow{k_{\text{off}}} \mathcal{S}_{\text{off}} \\ \mathcal{S}_{\text{on}} &\xrightarrow{k_{\text{init}}} \mathcal{S}_{\text{on}} + \mathcal{X} \end{aligned} \tag{A.64}$$

is equivalent to

$$\varnothing \xrightarrow{k_{\text{on}}} B \times \mathcal{X} \tag{A.65}$$

with B a geometrically-distributed random variable with mean b, whenever k_{off}, $k_{\text{init}} \to \infty$ with $\frac{k_{\text{init}}}{k_{\text{off}}} := b$ finite. There are a number of ways to prove this. For example, it is straightforward to consider the case with degradation of $\mathcal{X}$, find its distribution [143], then take the relevant limit and show that the transient distributions match the bursty case. However, this procedure relies on somewhat tedious manipulation of special functions.

The easiest approach being with noticing that this process affords a representation in terms of the instantaneous transcription rate $K(t)$, which is equal to zero when $s = \mathcal{S}_{\text{off}}$ and k_{init} when $s = \mathcal{S}_{\text{on}}$. This process's value depends on its past, implying that it is not a subordinator[10] (contradicting [7, 8]). The duration of each on period is exponential with scale k_{off}^{-1}. The total transcriptional intensity of each on period is exponential with scale $k_{\text{off}}^{-1} k_{\text{init}} = b$. The number of molecules generated per on period is Poisson, with a mean given by the intensity, i.e., geometric with scale b. As $k_{\text{off}} \to \infty$, the on periods become infinitesimally short, and the process becomes memoryless, producing a jump subordinator.

A.8.2 The CIR process converges to the inverse Gaussian subordinator

This section adapts a portion of [113] by G.G.*, J.J.V.*, M.F., and L.P. This analysis was performed by J.J.V. and G.G.

Consider, now, the case of the Cox–Ingersoll–Ross transcriptional driver (Equations 7.3 and 7.4) coupled to unspecified downstream dynamics. We have some characteristic U corresponding to downstream species and the following ODE for the CIR characteristic $U_K(\mathsf{s})$:

$$\frac{dU_K(\mathsf{s})}{d\mathsf{s}} = -\kappa U_K + \kappa\theta U_K^2 + U. \tag{A.66}$$

For $\kappa \to \infty$, both sides of the equation are approximately zero: U_K rapidly equilibrates. Applying the quadratic formula:

$$\begin{aligned} 0 &\approx \kappa\theta U_K^2 - \kappa U_K + U \\ U_K &\approx \frac{1}{2\kappa\theta}\left[\kappa \pm \sqrt{\kappa^2 - 4\kappa\theta U}\right] \\ &= \frac{1}{2\theta}\left[1 \pm \sqrt{1 - 4\frac{\theta}{\kappa}U}\right] := \frac{1}{2\theta}\left[1 \pm \sqrt{1 - 4bU}\right]. \end{aligned} \tag{A.67}$$

We have assumed b is finite, so $\theta \to \infty$. Therefore, we find that the transcription operator $\mathcal{A}(U_K(\mathsf{s}))$ takes the form

$$\begin{aligned} \frac{a\theta}{2\theta}\left[1 \pm \sqrt{1 - 4bU}\right] &= \frac{a}{2}\left[1 \pm \sqrt{1 - 4bU}\right] \\ &= \frac{a}{2}\left[1 - \sqrt{1 - 4bU}\right]. \end{aligned} \tag{A.68}$$

We have chosen the negative sign because otherwise $U_K(\mathsf{s})$ does not converge to zero, and does not produces a steady-state solution when integrated. This expression is the moment-generating function of the inverse Gaussian subordinator. Interestingly, even though this process is memoryless, it is not a compound Poisson process: it has infinitely many jumps in each finite interval. Although this limit is somewhat degenerate, it is useful to consider, as it fills an apparent lacuna in the finance literature [20, 22]: when $U = e^{-\gamma \mathsf{s}}$, we find that the stationary distribution has the relatively simple closed-form log-PGF:

$$\log G(u) = \frac{a}{\gamma}\left(1 - \sqrt{1 - 4bu}\right) + \frac{a}{\gamma}\log\left(\frac{1 + \sqrt{1 - 4bu}}{2}\right). \tag{A.69}$$

In our understanding, this is the solution to the "OU-IG" case listed as "Not known" in Table 2 of [22].

A.8.3 The Poisson representation facilitates adaptation of finance results

A.8.3.1 Time-dependent bursty processes

This section adapts a portion of [105] by G.G. and L.P. This derivation was performed by G.G.

We can use the isomorphisms between continuous and discrete processes to bypass tedious calculations. For example, mixture models are fairly popular for representing differentiation trajectories: each latent time is associated with a set of parameters for the negative binomial distribution; these parameters smoothly evolve throughout the trajectory [77, 216], representing modulation of bursty transcription. We can reasonably ask whether this framework can be used to represent "RNA velocity"-like trajectories, with meaningful transient effects (Section 6.1.1). This question is interesting given that this holds true for non-bursty processes: the distribution of a process with constitutive production, coupled to some isomerization and degradation reactions, is Poisson with a time-dependent mean (a trivial consequence of Equation 4.25, but explored in further detail in [146]). Is it possible that a time-dependent negative binomial can represent a transient process in the same fashion?

This intuition turns out to be incorrect even in the simplest case with $n = 1$. The discrete bursty process is a Poisson mixture of the Γ-OU process (Equations 7.1 and 7.2). Therefore, its transient PGF coincides with the transient MGF of the Γ-OU process, which is well-known [241]:

$$G(u,t) = \left(\frac{1 - bue^{-\gamma t}}{1 - bu}\right)^{k/\gamma}. \tag{A.70}$$

For finite t, Equation A.70 is not the PGF of a negative binomial distribution. In other words, a "pseudotime"-dependent negative binomial distribution simply emulates a collection of local steady states, rather than any transient processes, in the vein of Equation 10.2.

A.8.3.2 Autocatalysis

This section adapts a portion of [115] by G.G., J.J.V., and L.P. This derivation was performed by G.G.

Interestingly, this approach generalizes to cases with $D \neq 0$. Suppose we are interested in a 1-species system with $N = 1$, $n = 1$, $m = 0$, involving birth at α,

death at γ, and autocatalysis at q:

$$\begin{aligned} \varnothing &\xrightarrow{\alpha} X \\ X &\xrightarrow{q} X + X \\ X &\xrightarrow{\gamma} \varnothing. \end{aligned} \tag{A.71}$$

This system was introduced, but not treated, by Jahnke and Huisinga [146], and, to our knowledge, first solved with master equation and generating function calculations in [300]. However, we can also solve it merely by matching terms, without any new calculations.

This reaction schema yields the following PDE terms:

$$\begin{aligned} \mathcal{A}(u) &= \alpha u \\ C &= -\gamma + q \\ D &= q. \end{aligned} \tag{A.72}$$

The same form can be obtained by defining a system with $N = 1$, $n = 0$, and $m = 1$, with drift α, mean-reversion at rate $\gamma - q$, and diffusion q. This system matches the functional form of a Cox–Ingersoll–Ross (CIR) process with drift ab, mean-reversion rate a and square-root noise with intensity σ [62]:

$$\begin{aligned} dy_t &= a(b - y_t)dt + \sigma\sqrt{y_t}dW_t \\ D = \frac{1}{2}\sigma^2 = q &\implies \sigma = \sqrt{2q} \\ ab &= \alpha \\ C = -a &= -\gamma + q. \end{aligned} \tag{A.73}$$

As $t \to \infty$, the CIR process approaches the gamma distribution with shape ν and scale θ:

$$\begin{aligned} P(y; \nu, \theta) &= \frac{1}{\Gamma(\nu)\theta^\nu} y^\nu e^{-y/\theta} \\ \nu &= \frac{2ab}{\sigma^2} = \frac{\alpha}{q} \\ \theta &= \frac{\sigma^2}{2a} = \frac{q}{\gamma - q}, \end{aligned} \tag{A.74}$$

which follows from standard identities [62]. The distribution has the MGF

$$\begin{aligned} M &= \left(\frac{1}{1 - \theta h}\right)^\nu, \\ \text{implying the Poisson mixture PGF } G &= \left(\frac{1}{1 - \theta u}\right)^\nu. \end{aligned} \tag{A.75}$$

This is the probability generating function of a negative binomial distribution with the shape $\nu = \alpha/q$ and mean $\nu\theta = \frac{\alpha}{\gamma - q}$. With some algebra, we can rewrite the PGF as

$$\left(\frac{\gamma - c}{\gamma}\right)^{\alpha/q} \left(\frac{1}{1 - \frac{qg}{\gamma}}\right)^{\alpha/q}, \tag{A.76}$$

which is the analytical solution reported in the second line of Eq. 4.47 of [299] under the assumption $\gamma > q$. We find, then, that autocatalysis with constitutive transcription yields a stationary distribution equivalent to bursty transcription with no autocatalysis.

A.8.3.3 Autocatalysis with bursty production

This section adapts a portion of [115] by G.G., J.J.V., and L.P. This derivation was performed by G.G.

Obtaining this result, we may ask how the distribution changes if the molecules are produced in geometric bursts B with mean size b:

$$\begin{aligned} \varnothing &\xrightarrow{\alpha} B \times X \\ X &\xrightarrow{q} X + X \\ X &\xrightarrow{\gamma} \varnothing. \end{aligned} \tag{A.77}$$

These reactions yield the following PDE terms:

$$\begin{aligned} \mathcal{A}(u) &= \alpha \left[\frac{1}{1 - bu} - 1\right] \\ C &= -\gamma + q \\ D &= q. \end{aligned} \tag{A.78}$$

To solve the system, we first find the solution to the Bernoulli-type differential equation, defining $c = -C = \gamma - q$ for convenience:

$$\begin{aligned} \frac{dU}{d\mathsf{s}} &= -cU + qU^2 \text{ such that } U(\mathsf{s} = 0) = u \text{ yields} \\ U(\mathsf{s}) &= \frac{cue^{-c\mathsf{s}}}{c + qu(e^{-c\mathsf{s}} - 1)}. \end{aligned} \tag{A.79}$$

Then the log-generating function of the stationary distribution is given by the integral of $\mathcal{A}(U(\mathsf{s}))$:

$$\begin{aligned}
\frac{1}{1-bU} - 1 &= \frac{b\frac{cue^{-c\mathsf{s}}}{c+qu(e^{-c\mathsf{s}}-1)}}{1-b\frac{cue^{-c\mathsf{s}}}{c+qu(e^{-c\mathsf{s}}-1)}} = \frac{bcue^{-c\mathsf{s}}}{c+qu(e^{-c\mathsf{s}}-1)-bcue^{-c\mathsf{s}}} \\
&= \frac{bcue^{-c\mathsf{s}}}{(c-qu)-(bc-q)ue^{-c\mathsf{s}}} = \frac{bcu}{c-qu}\frac{e^{-c\mathsf{s}}}{1-\frac{(bc-q)u}{c-qu}e^{-c\mathsf{s}}} \\
\log G &= \int_0^\infty \mathcal{A}(U(\mathsf{s}))d\mathsf{s} = \frac{\alpha bcu}{c-qu}\int_0^\infty \frac{e^{-c\mathsf{s}}}{1-\frac{(bc-q)u}{c-qu}e^{-c\mathsf{s}}}d\mathsf{s} \\
&= -\frac{\alpha b}{bc-q}\log\left(\frac{c-qu-(bc-q)u}{c-qu}\right) = \frac{\alpha b}{bc-q}\log\left(\frac{c-qu}{c-bcu}\right) \\
&= \frac{\alpha b}{bc-q}\log\left(\frac{1-qc^{-1}u}{1-bu}\right) = \frac{\alpha b}{bc-q}\log\left(\frac{b^{-1}-(bc)^{-1}qu}{b^{-1}-u}\right) \\
G &= \left(\frac{b^{-1}-(bc)^{-1}qu}{b^{-1}-u}\right)^\nu = \left(\frac{1-qc^{-1}u}{1-bu}\right)^\nu, \text{ such that} \\
\nu &= \frac{\alpha b}{bc-q}.
\end{aligned} \tag{A.80}$$

To achieve a positive ν, we must have

$$\begin{aligned}
bc - q &> 0 \\
b(\gamma - q) &> q \\
b\gamma &> q(1+b) \\
\gamma &> q\frac{1+b}{b},
\end{aligned} \tag{A.81}$$

i.e., whereas in the non-bursty case, a steady state was guaranteed by having $\gamma > q$, in the bursty case we must impose a more restrictive condition. The second inequality implies that the coefficient of u in the numerator

$$\frac{q}{bc} = \frac{q}{b(\gamma-q)} < 1; \tag{A.82}$$

in other words, $(bc)^{-1}q \in (0,1)$ can be represented by $e^{-\kappa\tau}$ for some positive κ and τ. Therefore, the GF

$$G = \left(\frac{b^{-1}-ue^{-\kappa\tau}}{b^{-1}-u}\right)^\nu \tag{A.83}$$

matches the functional form of the time-dependent MGF of the gamma Ornstein–Uhlenbeck process started at $y = 0$ [241], and the PGF of the bursty transcription/degradation process started at $x = 0$, precisely as discussed in Section A.8.3.1.

Finally, we define $a = e^{-\kappa\tau}$ and rewrite G again:

$$\begin{aligned} G &= \left(\frac{1-abu}{1-bu}\right)^{\nu} = \left(\frac{a(1-abu)}{a(1-bu)}\right)^{\nu} = \left(\frac{a(1-abu)}{1-abu-(1-a))}\right)^{\nu} \\ &= \left(\frac{a}{1-(1-a)\frac{1}{1-abu}}\right)^{\nu} \\ &= G_{\mathrm{NB}}\left(\frac{1}{1-abu}\right) \\ &= \sum_{k=0}^{\infty} P_{\mathrm{NB}}(k)\left(\frac{1}{1-abu}\right)^{k}. \end{aligned} \tag{A.84}$$

This is a negative binomial–negative binomial mixture. In other words, the distribution is equivalent to that of a negative binomial distribution with scale parameter ab and stochastic shape parameter k, with k in turn drawn from a negative binomial distribution with the shape ν and success probability a. We can confirm this result through a direct calculation:

$$\begin{aligned} P(x) &= \sum_{k=0}^{\infty} P(k)P(x|k), \text{ with PGF} \\ G(g) &= \sum_{x=0}^{\infty} g^x \sum_{k=0}^{\infty} P(k)P(x|k) = \sum_{k=0}^{\infty}\sum_{x=0}^{\infty} g^x P(k)P(x|k), \text{ and assuming NB } P, \\ &= \sum_{k=0}^{\infty}\sum_{x=0}^{\infty} g^x \frac{\Gamma(\nu+k)}{k!\Gamma(\nu)}(a)^{\nu}(1-a)^k \times \frac{\Gamma(k+x)}{x!\Gamma(k)}\left(\frac{k}{k+kab}\right)^k\left(\frac{kab}{k+kab}\right)^x \\ &= \sum_{k=0}^{\infty}\left[\sum_{x=0}^{\infty}\frac{\Gamma(k+x)}{x!\Gamma(k)}\left(\frac{gab}{1+ab}\right)^x\right] \times \frac{\Gamma(\nu+k)}{k!\Gamma(\nu)}(a)^{\nu}(1-a)^k\left(\frac{1}{1+ab}\right)^k \\ &= \sum_{k=0}^{\infty}\left(\frac{1}{1-\frac{gab}{1+ab}}\right)^k \frac{\Gamma(\nu+k)}{k!\Gamma(\nu)}(a)^{\nu}(1-a)^k\left(\frac{1}{1+ab}\right)^k \\ &= \sum_{k=0}^{\infty}\left(\frac{1}{1+ab-gab}\right)^k \frac{\Gamma(\nu+k)}{k!\Gamma(\nu)}(a)^{\nu}(1-a)^k \\ &= \sum_{k=0}^{\infty}\left(\frac{1}{1-abu}\right)^k \frac{\Gamma(\nu+k)}{k!\Gamma(\nu)}(a)^{\nu}(1-a)^k = \sum_{k=0}^{\infty} P_{\mathrm{NB}}(k)\left(\frac{1}{1-abu}\right)^k. \end{aligned} \tag{A.85}$$

This expression uses the shape-mean parametrization for the conditional probability $P(x|k)$ and the shape-probability parametrization for the mixing probability $P(k)$.

However, is is considerably easier to notice that the analysis of the Γ-OU model by Sabino and Petroni [241] states that the transient process law is equivalent to that

of an Erlang distribution with scale parameter ab and stochastic shape parameter k, with k in turn drawn from a negative binomial distribution with the shape ν and success probability a, or scale $\frac{1}{1+a}$. This immediately implies that the distribution of the corresponding discrete process is a negative binomial-negative binomial mixture with equivalent parameters.

Although this distribution cannot be expressed in closed form, its construction makes the simulation of the bursty transient and stationary autocatalytic processes trivial, and suggests that simple finite approximations (i.e., up to a modest k) may be developed.

A.8.4 Many processes are closed under species-independent, sequestering sampling

This section adapts and extends a portion of [115] by G.G., J.J.V., and L.P. This derivation was performed by G.G.

Consider a sequestering, species-independent technical noise model, such that the probability of retaining a molecules of any species X_i is p. Assume $D = 0$. The set of downstream characteristics $\mathbf{U}(\mathsf{s})$ is a linear combination of the entries of $\mathbf{u}$. Since integrating sampling amounts to substituting $u_i \leftarrow pu_i$, this procedure effectively rescales the entire function $\mathbf{U}$ by p. If the upstream generation process has a scale parameter, such that $\mathcal{A}(\mathbf{u})$ involves multiplication by θu_i, sampling is equivalent to rescaling θ by p. For the common processes, we obtain the rescaling

$$\begin{aligned} &kpU \text{ for constitutive production or drift and} \\ &k\left[\frac{1}{1-bpU} - 1\right] \text{ for bursting or jumps,} \end{aligned} \tag{A.86}$$

for each component indexed by ω. This holds if the constitutive production parameter k is stochastic; for example, sampling amounts to rescaling the parameter θ of the usual extrinsic noise model. This also holds with no loss of generality if the burst or drift processes are state-dependent.

It is unclear whether this result generally holds for $D \neq 0$. It very well might: the CIR driver coupled to two downstream species has this property, which we exploit in Section 7.1. However, the derivation is somewhat subtle and requires the direct manipulation of differential equations, so the generalization will require further investigation.

For the bursty model with $n = 1$, the resulting distribution is negative binomial with shape k/γ and scale pb. This justifies treating the parameter ν in Equation 2.4 as

purely biological: the negative binomial shape is invariant under downsampling. However, there is no particular reason to think it is *uniform*, because the burst size may differ between cell subpopulations[11].

This result does not hold in its full generality when the sampling probabilities are species-dependent; for example, when p_N and p_M are distinct, we can identify bp_N and p_N/p_M. The result does not hold when non-sequestering noise models are used.

Appendix B

QUALITATIVE DISCUSSION OF SEQUENCING PROCEDURES AND THEIR CAVEATS

B.1 Notes on nomenclature and binary assignment

This section summarizes and unifies a portion of [112] by G.G., M.F., T.C., and L.P., as well as [44] by M.C.*, G.G.*, Y.C., T.C., and L.P. This theoretical discussion was written by G.G.

In the field of microbiology, "nascent" RNA is often, but not always, used to characterize the mRNA molecules in the process of synthesis, associated to a DNA strand via an RNA polymerase complex [53, 54, 239, 319]. In this framework, the "mature" transcriptome is simply the complement of the nascent transcriptome, i.e., all molecules that are not chemically associated to a DNA strand. Therefore, the canonical definition of "nascent" RNA is equivalent to *transcribing* RNA, which is a polymeric structure with a particular sequence.

Transcribing RNA can be observed directly through electron micrography [54]. However, more typically, it is investigated through more or less direct experimental proxies that can be scaled to many genes and cells at a time. In the single-cell fluorescence subfield, DNA or membrane staining can be used to identify bright spots localized to the nucleus, which is treated as signal from RNA at the transcription site [121, 206]; this signal may include contributions from RNA incidentally, or mechanistically, retained at a DNA locus [319]. In this strategy, "nascent" molecules are DNA-associated. Alternatively, and perhaps more commonly, transcribing molecules have been studied by using probes targeted to the 5′ and 3′ regions [251, 309, 318, 327], or to intronic and exonic regions [16, 252, 262, 307]. In this strategy, "nascent" molecules contain a particular region, either synthesized earlier or removed later in the RNA life-cycle.

The use of intron data as a proxy for active transcription is reminiscent of, but distinct from sequence census [316] strategies that directly study RNA sequences. These strategies, in turn, typically use chemical methods to enrich for newly transcribed RNA. For example, Reimer et al. isolate chromatin, then deplete sequences that have been post-transcriptionally poly(A) tailed [235]. Analogously, Drexler et al. use 4-thiouridine (4sU) labeling to enrich for newly synthesized molecules [76].

These approaches may produce conflicting results; for example, introns may be rich both in poly(A) handles [168] and 4sU targets [235], giving rise to obscure technical effects. Therefore, these "processed" or "temporally labeled" proxies are coarsely representative of transcriptional dynamics, and their quantitative interpretability is unclear as of yet.

The sequence content may be used more directly, by conceding that DNA association or localization are not easily accessible by sequence census methods, and treating splicing *per se*. This approach has a fairly long history. Intronic quantification has been used to characterize transcriptional mechanisms in microarray datasets [326], and to characterize differentiation programs in RNA sequencing [223, 224]. In single-cell RNA sequencing, intronic content has been leveraged to identify transient behaviors from snapshot data [168], albeit with some outstanding theoretical concerns and caveats (Section 6.1). Briefly, it is, in principle, possible to coarsely classify molecules with intronic content as "unspliced" or "pre-mRNA" and aggregate all others as "spliced," "mature," or simply "mRNA."

The quantification of transcripts so classified is a relatively straightforward genomic alignment problem. The multiple available implementations [80, 168, 197, 264] tend to disagree on the appropriate assignment of ambiguous sequencing reads [112, 264], obscuring a more fundamental problem: the binary classification is somewhat arbitrary [33, 76, 158, 194], and it is likely that detailed splicing graph models will be necessary in the future (as proposed Section 10.2).

We can illustrate the problem using the simplest example of a three-exon, two-intron gene, with a "parent" transcript $E_1I_1E_2I_2E_3$. It seems reasonable enough to call $E_1I_1E_2I_2E_3$ "unspliced" and to call "terminal" transcript $E_1E_2E_3$ "spliced." But what of the "intermediate" transcripts $E_1E_2I_2E_3$ and $E_1I_1E_2E_3$? Even if we have perfect information about the sequence content, by placing intronic reads into the "unspliced" category, we conflate the parent and intermediate transcripts. On the other hand, if we place all barcodes with splice junctions into the "spliced" category, we conflate the intermediate and terminal transcripts. Adding more complexity, some isoforms may retain introns through alternative splicing mechanisms; for example, the intermediate transcripts may be exported, translated, and degraded alongside the terminal one. Of course, in practice, the "parent" transcript may not actually exist as a distinct species if I_1 is removed before the transcription of E_3 is completed. The focus on sequence is yet another step removed from the transcriptional dynamics, particularly since some of the splicing processes may

occur after transcriptional elongation is complete [60].

Adding yet more complexity to the modeling, "mature" — whether "off-template," "spliced," or "processed" — molecules are not immediately available for degradation; first, the process of nuclear export must take place. Studies that presuppose access to imaging data tend to model it explicitly [27, 86, 127, 206, 261]. However, this approach has not been applied in sequencing assays, as current technologies do not distinguish nuclear and cytoplasmic molecules. Furthermore, comparisons of paired single-cell and single-nucleus datasets are hampered by the limited characterization of the noise sources in the latter technology.

Pending the development of more sophisticated sequencing and alignment technologies, as well as the implementation of tractable models of biology, the data exploration portion of our study focuses on the "spliced" and "unspliced" matrices generated by *kallisto* | *bustools* [197]. This choice is a compromise, and we adopt it after considering the following factors:

- Availability of quantification workflows: spliced and unspliced matrices are straightforward to generate.
- Model tractability: the two-stage models can be evaluated; more sophisticated models require new algorithms, because they involve underspecified, high-dimensional distributions (as alluded to in Chapter 5 and Section 10.2).
- The scope of sequencing data: single-cell protocols do not yet give access to sub-cellular information, so inference of elongation or nuclear retention dynamics is acutely underspecified.

We use the terms "nascent" and "mature" to identify the unspliced and spliced RNA matrices. This choice of nomenclature is deliberate. Although it somewhat conflicts with the established microbiology literature, this terminology is intended to emphasize the models' generality. The two-stage Markovian process is axiomatic. The specific identities assigned to the mathematical objects may range beyond counts identified by sequence census methods. They may represent the discretized and subtracted intensities of $3'$, $5'$, intron, or exon fluorescent probes, the counts of molecules within and outside the nuclear envelope, or polymerase counts obtained by micrography. Therefore, the terminology should be taken in the sense used for similar non-delayed models in [42, 86, 90, 91].

In sum, we cannot justify the two-stage model from first principles: the biology is far too complicated, and we cannot possibly assert that this simplistic model accurately represents the complex polymeric phenomena occurring in living cells. Instead, we emphasize two points.

1. The theoretical framework does not depend on these assumptions and identifications, and can be extended to account for other phenomena.
2. The two-stage model typically produces at least fair fits to the data.

This appears to be sufficient to provide a fair first-order treatment of the data at hand.

B.2 Notes on ambiguity

This section reproduces a portion of the supplement to [115] by G.G., J.J.V., and L.P. This theoretical discussion was largely written by G.G., with some essential background research by J.J.V.

The binary assignment problems outlined above arise even with perfect data — but we do not typically have perfect data. In Section 4.4.3, we mathematically formalized potential ambiguities in the quantification of different transcripts, mentioning two special cases of perfect identifiability and perfect ambiguity. We did not elaborate on this model component further, as it is, at this time, less immediately actionable than other components, and requires fairly considerable bioinformatic infrastructure to integrate with analysis. In the current section, we explore this model in more detail.

Even the simplest system, shown in Figure B.1a, can contain ambiguity that limits or prevents the identification of transcriptional dynamics. In this illustrative example, we consider a gene with only one intron and two flanking exonic sequences. We suppose that quantification and assignment only consider whether the read overlaps an intron or splice junction. An intron-containing read uniquely identifies the transcript as nascent, whereas a junction-spanning read uniquely identifies the transcript as mature. On the other hand, a fully exonic read does not provide any information about the source molecule. For the sake of completeness, we use "read" as a shorthand for the union of all reads corresponding to a particular UMI: since fragmentation is random, a given UMI will be associated with reads that cover slightly different regions.

The abundance of fully exonic reads depends on the structure and poly(A) content of the source transcript, as well as the sequencing technology. For example, in Figure

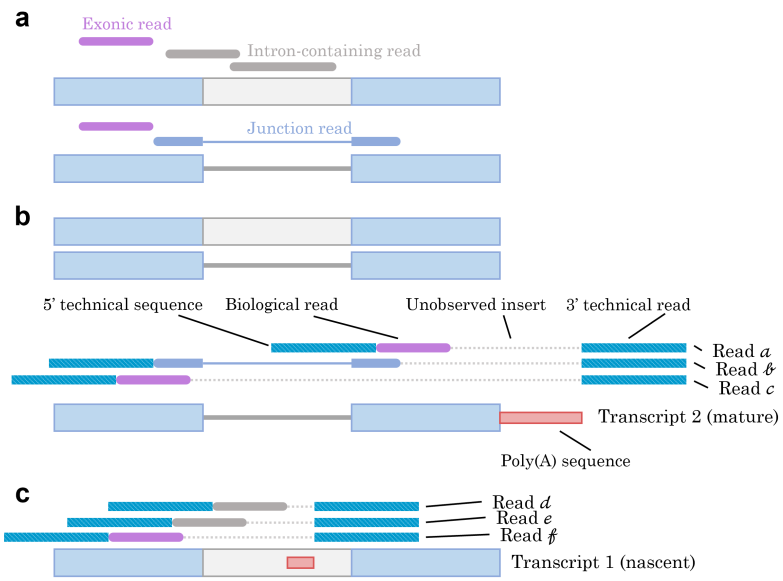

Figure B.1: Potential sources of short-read sequencing ambiguity in a hypothetical one-intron, two-exon transcript.
a. Possible splicing information conveyed by reads in the hypothetical transcript (magenta: reads that only contain exonic information; dark gray: reads that contain intronic information; dark blue: reads that overlap a splice junction. Blue block: exon; gray block: present intron; line: excised intron. 3′ end is toward the left).
b. Categories of reads that can be obtained by sequencing the transcript, assuming no endogenous poly(A) content (cyan block: technical reads and indices; dotted lines: residual inserts not observed by sequencing; red block: poly(A) sequence).
c. Categories of reads that can be obtained by capturing a transcript at an endogenous, intronic poly(A) sequence (conventions as in **a** and **b**).

B.1b, we consider reads that can be obtained from a gene that has little to no genomic poly(A) content (red). The unspliced molecules, as well as spliced molecules that have not yet been capped (top), cannot be observed at all. Therefore, their kinetics are not identifiable. On the other hand, the fully mature, poly(A)-tailed transcript (bottom) can be captured at the tail. This capture pattern can give rise to junction reads or formally non-identifiable exonic reads. If the 3′ exon is particularly long relative to fragment length, sequenced fragments will be enriched for purely exonic reads in the 3′ exon. Analogously, if fragment length and the 3′ exon are long relative to read length, sequenced fragments will be enriched for exonic reads in the 3′ exon. In Figure B.1c, we illustrate the analogous patterns that can emerge if the unspliced molecule has a single intronic poly(A) region. If the intron and read length are short relative to fragment length, sequencing will produce reads in the 3′ exon.

To characterize transcriptional kinetics, we seek to quantify transient transcripts,

which may or may not be mutually identifiable. This is infeasible to optimize on an experimental level. Even in the simple example we provided for illustration, to characterize the source transcript, the transcript region, fragment, and read lengths need to reside in a regime that produces unambiguous reads. To identify the splice junction in Figure B.1b, we require fragments that are slightly longer than the 3′ exon and reads that can cover the distance to the junction. However, read length cannot be changed without switching technologies; longer reads typically mean sacrificing the number of sampled cells [122, 226, 328]. In the same vein, fragmentation protocols cannot be easily interchanged, as they are optimized for a particular sequencing chemistry. Finally, even if these technical constraints were no object, it would *still* be impossible to optimize for unambiguous capture genome-wide: intron and exon length vary over many orders of magnitude [184, 335] and require different read and fragment length regimes for different genes.

Hypothetically, it may be possible to parametrize the $\mathscr{P}^a$ matrix as in Section 4.4.3, and fit it alongside the biological noise parameters. This approach may not be entirely futile. For example, Equation 4.46 demonstrates the relevant generating function for two biological species and three identifiable equivalence classes: 1, unambiguous nascent, 2, unambiguous mature, 3, ambiguous. The marginal of the nascent species has a functional form distinct from the marginal of the mature species [261], which immediately implies that $\mathscr{P}^a_{1,3} = 0$, $\mathscr{P}^a_{2,3} > 0$ produces distributions functionally distinct from $\mathscr{P}^a_{2,3} = 0$, $\mathscr{P}^a_{1,3} > 0$. In other words, we ought to be able to distinguish the case where all ambiguous counts originate from nascent RNA from the case where they originate from mature RNA, at least in the limit of immaculate and infinite data. We do not, however, expect this approach to be practical for real datasets.

We speculate that it may be more productive to use genomic information to constrain the ambiguity properties, in a similar spirit to [131]. For example, if a read lies in the 3′ untranslated region, *and* we know there is little endogenous poly(A) content 3′ of the read, then we should conclude the read is generated by priming at the poly(A) tail of a capped molecule. In other words, it may be possible to exploit the base information from the genome annotation, the fragment size distributions from orthogonal experiments, and the read size characteristic of the technology to directly construct the $\mathscr{P}^a$ matrix for each transcript.

We may illustrate this point in a more quantitative way. Consider the simplest case shown in Figure B.1, and further assume that poly(A) capping is rapid. Using

the notation in Equation 4.46, we find that $\mathscr{P}^{a}_{1,3} = p(\mathscr{f}|1)$, i.e., the probability of sequencing a nascent molecule to obtain a read with an insert from the 3′ exon. Analogously, $\mathscr{P}^{a}_{2,3} = p(\mathscr{a}|2) + p(\mathscr{c}|2)$, i.e., the probability of sequencing a mature molecule to obtain a read with an insert from the 3′ or 3′ exon. It appears legitimate to propose that these probabilities are only dependent on the sequence and experimental conditions, and may be effectively approximated by exploiting polymer physics or long-read data.

On one hand, this example is somewhat trivial by design. On the other, even this simplified picture of splicing omits important features. First, we presuppose that annotations exist for all downstream transcripts. As alluded to in Section 10.2, this is not typically the case, and the identities of and causal relationships between intermediate transcripts are obscure without dedicated study. Second, we presuppose that transcripts can be described as some combination of introns and exons, which transform by the excision of introns. However, even this seemingly reasonable latent assumption ignores elongation, which has been the subject of considerable study elsewhere. For example, a "nascent" transcript may not exist as a physical object, because splicing may complete before the 3′ exon is fully transcribed. A more biophysically realistic picture should take into account the fact that splicing occurs during and after elongation. In addition to these biological challenges, there is a variety of technical ones. For example, introns that have already been spliced out may, in principle, be captured and sequenced. If splicing is Markovian, this would be represented as a splitting reaction $\mathcal{X} \to \mathcal{Y} + \mathcal{Z}$, which is a splitting reaction not immediately tractable using our framework. Finally, due to a variety of technical effects, the reads themselves may have a more complex relationship to the source transcripts. These effects include strand invasion and aberrant priming, and may lead to reads containing antisense and template switch oligo sequences [1]. Formally, all of these effects can be integrated into a sufficiently complicated stochastic model. In practice, we recommend introducing complexity only when simpler models fail.

B.3 Notes on imputation and reconstruction

This section reproduces a portion of the supplement to [112] by G.G., M.F., T.C., and L.P. This theoretical discussion was written by G.G.

In the current supplement, we point out that count "correction" through imputation can produce arbitrarily incorrect results. Although this result is fairly elementary, it does not appear to have been applied in the single-cell sequencing field, and raises

questions regarding standard imputation methods. Suppose we have a data point $\mathcal{D}_{\mathsf{cg}}$ and the corresponding true mRNA abundance x_{cg} for a particular molecular species, cell c, and gene g. Sequencing is not perfect: the data point $\mathcal{D}_{\mathsf{cg}}$ is generated from x_{cg} according to a non-deterministic schema, with an unknown probability law $P(\mathcal{D}_{\mathsf{cg}}|x_{\mathsf{cg}})$.

Two problems emerge. First, a point estimate of x_{cg} based on observed $\mathcal{D}_{\mathsf{cg}}$ is necessarily incomplete: the sequencing process induces an entire distribution of possible x_{cg}. This conditional distribution is given by Bayes' formula:

$$P(x_{\mathsf{cg}}|\mathcal{D}_{\mathsf{cg}}) = \frac{P(\mathcal{D}_{\mathsf{cg}}|x_{\mathsf{cg}})P(n_{\mathsf{cg}})}{P(\mathcal{D}_{\mathsf{cg}})}. \tag{B.1}$$

Assigning a single value is questionable, and downplays the effects of uncertainty. This remains a problem even if a theoretically optimal choice is taken, such as the point estimate

$$\hat{x}_{\mathsf{cg}} = \text{argmax}_{x_{\mathsf{cg}}} P(x_{\mathsf{cg}}|\mathcal{D}_{\mathsf{cg}}). \tag{B.2}$$

Second, the conditional distribution depends on $P(x_{\mathsf{cg}})$, the actual ground truth distribution. This distribution is unknown and needs to be identified and fit based on the data. Therefore, any imputation procedure that assigns a point estimate without considering the underlying distribution is *a priori* distortive.

In other words, this Bayesian argument illustrates that meaningful count correction is impossible without identifying and fitting the data-generating model, which encodes biological effects in $P(x_{\mathsf{cg}})$ and technical effects in $P(x_{\mathsf{cg}}|\mathcal{D}_{\mathsf{cg}})$. Count correction is strictly less powerful than parameter estimation for the biological and technical models, because count correction requires those parameters, whereas knowledge of the parameters immediately implies the entire distribution of the biological and observed variables.

This critique does not appear to apply to, e.g., the probabilistic "imputation" in *scVI*, as the autoencoder framework reports a distribution of μ_{cg}, rather than a point estimate x_{cg}. This approach is somewhat more coherent, as it explicitly represents stochasticity.

B.4 Notes on graph methods

This section reproduces a portion of the supplement of [112] by G.G., M.F., T.C., and L.P. This theoretical discussion was written by G.G.

Throughout Section 6.1, we have discussed k-nearest neighbor (k-NN) graphs in the context of RNA velocity and embeddings. As k-NN is ubiquitous in scRNA-seq, and its applications are manifold, a full analysis is not feasible. In this supplement, we discuss a set of purely theoretical pitfalls which may limit the utility or interpretability of such graphs, leaving validation on simulated data to future work.

A k-NN graph purports to reflect relationships between cells based on similarity between their transcriptomic "states." These states are typically mature RNA copy numbers that have undergone several steps of count processing, including size-normalization, log-transformation, filtering, and projection onto the top few principal components. The determination of neighbors in this space represents an uncomfortable compromise: if there are too few dimensions, the projection may be unrepresentative of the underlying data matrix; if there are too many, it may be skewed by the "curse of dimensionality." Such distortions are evident in Figure 6.1d.

More subtly, it is unclear that observed transcriptomic similarity between barcodes should imply similarity between cells. *In silico* UMI counts have been filtered through the random process of sequencing; the true underlying transcriptomic state is unknown, and cannot be precisely reconstructed (as outlined above, in Section B.3). We anticipate that certain narrow problems, such as cell type identification, may be insensitive to this source of error. For example, it is possible that simulated benchmarks can provide empirical results in the vein of "assuming transcriptional dynamics are bursty, cell types are distinguished by at least ten marker genes, these marker genes have an expression differential of at least one order of magnitude, and each cell type comprises 10–20% of the entire dataset, a community detection-based algorithm has a 80% classification accuracy according to a particular metric, which falls to 75% if a particular noise model is imposed." However, at this time, constructing undirected k-NN graphs based on imperfectly observed data appears to have limited theoretical or empirical justification.

The construction of directed k-NN graphs, as proposed in the original RNA velocity publication [168] and extended elsewhere [171], is considerably more problematic. A directed graph implies a causal relationship between observed cell states; in the implementations of RNA velocity, this causal relationship is Markovian, with a transition rate governed by the alignment between velocities and neighbor directions. We can analyze potential issues in this reasoning by gradually increasing the complexity of the system under analysis.

First, suppose that the RNA dynamics and the chemistry of sequencing are non-random, whereas the cell observations are independent draws of $y_N(t), y_M(t)$ from an underlying (deterministic and perfectly known) trajectory, using the notation in Section 6.1.1. Intuitively, building a directed graph raises questions: one cell is not the "descendant" of another, as both cells were captured and sequenced simultaneously. In this model formulation, the observed cells do not have causal relationships at all. We can build a graph that connects each cell to the neighbors of its position at Δt; at this point, the function used to define the graph is deliberately left generic. Even in this ideal-case scenario, this graph is highly dependent on the value of Δt and strictly less informative than the dynamical system parameters, as its construction *requires* those parameters.

Next, suppose that RNA dynamics are still deterministic, but sequencing injects noise into the observations. In the ideal case, where the dynamical system and sequencing noise parameters are perfectly known, extrapolating from the current state is impossible: the true RNA abundance is unknown. At best, we may instantiate a set of trajectories conditional on all possible unobserved biological states. As before, these trajectories depend on the time horizon Δt. Aggregating the trajectories by assigning a weight to a single graph edge is strictly less informative than reporting the trajectories; that, in turn, is less informative than reporting the system parameters. The question of the amount of error incurred by this approach is coarsely equivalent to the question of a hidden Markov model's approximability by a Markov model. Omitting uncertainty due to technical noise is equivalent to assuming a hidden Markov model can be effectively described without latent states. This assumption may be approximately valid (e.g., in the limit of perfect sequencing), or grossly incorrect, with no apparent *a priori* way of constraining error.

Suppose now that RNA dynamics of n molecular species are stochastic on the state space $\Omega = \mathbb{N}_0^n$, with no observation noise. If we sample $\mathsf{N_c}$ cells, the microstates corresponding to these data make up a subset Ω_D, such that $|\Omega_D| \leq \mathsf{N_c}$. If the data are generated by a biological Markov process on Ω, a process truncated to Ω_D will either not be Markov, or fail to recapitulate features of the biological process. This generally holds when $|\Omega_D| < |\Omega|$. The extent of incurred error may vary from minimal to egregious, and cannot be constrained without further knowledge of the system.

This argument has fairly severe consequences and foundations. Defining a directed graph on Ω_D is superfluous: Ω_D is itself constructed out of samples from the traversal

of a directed graph isomorphic to the biological continuous-time Markov chain. This CTMC is isomorphic to the CME. If states $\mathbf{x}_i, \mathbf{x}_j \subseteq \Omega_D \subseteq \Omega$ have the nonzero rate k_{ij} for the $\mathbf{x}_i \rightarrow \mathbf{x}_j$ transition, it is possible to construct an approximating CTMC on Ω_D merely by setting the rate of the corresponding transition to k_{ij}. However, if either one of those states is not in Ω_D, some dynamics are lost. The question of the amount of error incurred by truncation is roughly equivalent to the question of a given infinite CTMC's approximability on a finite subdomain. For broad classes of CME models, finite approximations can do arbitrarily well: truncation to a finite subdomain Ω_D incurs an error governed by the amount of probability flux to and from states outside this domain, and converges (in distribution) to the true CTMC as $|\Omega_D| \rightarrow \infty$. The finite state projection algorithm [203] exploits this approach to evaluate CME solutions (Section 4.2.2.2). However, the FSP is adaptive, and expands Ω_D on a grid until a desired precision is achieved, rather than using a relatively small set of points which do not necessarily have transitions in the underlying CTMC.

The lack of these direct transitions between observed states in Ω_D implies that a CTMC on Ω cannot be projected down to Ω_D. This principle can be demonstrated using a striking trivial case. Consider a three-state Markov chain:

$$0 \xrightarrow{k_{01}} 1 \xrightarrow{k_{12}} 2. \tag{B.3}$$

The full state space is $\Omega = \{0, 1, 2\}$. Consider a case where states 0 and 2 are observed in multiple independent chains at some time t, i.e., $\Omega_D = \{0, 2\}$. We wish to define a neighborhood relationship between observations in state 0 and those in state 2, and summarize it as a CTMC on Ω_D. The following results emerge immediately.

Even with perfect knowledge of the original CTMC, a truncated CTMC will not fully recapitulate its dynamics. The residence time in state 0 is exponentially distributed:

$$f(\tau) = k_{01} e^{-\tau k_{01}}. \tag{B.4}$$

The true transition from 0 to 2 has a hitting time distribution described by a hypo-exponential law:

$$f(\tau) = \frac{k_{01} k_{12}}{k_{01} - k_{12}} \left[e^{-\tau k_{12}} - e^{-\tau k_{01}} \right]. \tag{B.5}$$

The transition from 0 to 2 on Ω_D — equivalent to the residence time in state 0 — is constrained to be exponentially distributed:

$$f(\tau) = k_{02} e^{-\tau k_{02}}. \tag{B.6}$$

As the exponential distribution has one parameter, it can match the true residence time distribution, or a single moment of the hitting time distribution, but not both. If we match the residence time distribution, the approximation becomes arbitrarily good as $k_{12} \to \infty$ and arbitrarily poor as $k_{12} \to 0$. The latter case makes not observing the long-lived state 1 somewhat improbable. However, as system dimensionality grows — e.g., if multiple independent CTMCs are started — the intermediate state will be unobserved in at least one of those chains almost surely.

Even with perfect knowledge of the original CTMC, a generic stochastic process will not fully recapitulate its dynamics. We can define a non-Markovian process on Ω_D that will have a hitting time distribution given by Equation B.5. However, its residence time will fail to be distributed per Equation B.4. By constraining the process to traverse only observed states, the contributions from unobserved intermediate states are omitted, with error that cannot be easily bounded.

This inability to "compress" CTMCs into a smaller domain can also be treated in a more generic way. To define transitions between states, we must assign a single number — the rate — to the transition. Intuitively, we expect that the rates of CTMCs on Ω_D should reflect the relative probabilities of transitions between states in the original CTMC on Ω. Thus, given three states $\mathbf{x}_i$, $\mathbf{x}_j$, and $\mathbf{x}_l$, we would like to impose the following criterion:

$$P(\mathbf{x}_j, t; \mathbf{x}_i, 0) > P(\mathbf{x}_l, t; \mathbf{x}_i, 0) \; \forall \; t \in \mathbb{R}_+ \implies k_{ij} > k_{il}, \tag{B.7}$$

where P refers to full CTMC's probability of being in a state $\mathbf{x}_j$ or $\mathbf{x}_l$ at time t, conditional on being in state $\mathbf{x}_i$ at time 0, and k are rates in the "compressed" CTMC. This criterion appears to be the only "natural" one, and it induces a partially ordered set, which is insufficient to even order the transition rates in the CTMC on Ω_D.

In conclusion, graph-based methods are problematic for representing relationships between cells. They can represent certain aspects of dynamics, but inevitably contradict the underlying graph that governs the biophysical CTMC. It is possible to make them agree, albeit only by considerably expanding the graph beyond observed

states, recapitulating the CME. In other words, the only cell–cell graph which can quantitatively summarize Markovian biological processes is the graph underlying the CME, with an infinite number of states, remaining forever out of reach and recalling Borges's and Carroll's dichotomy of the map and the territory [35, 45]: "...we now use the country itself, as its own map, and I assure you it does nearly as well."

INDEX

[1]Note that we do not explicitly allow competing reaction pathways for species with non-Markovian efflux. The precise justification for this is somewhat subtle. In the Markovian case, setting up multiple reaction channels and choosing between them based on a categorical distribution over reactions is equivalent to simply seeing which reaction fires first. But in the non-Markovian case, these things are distinct: obviously, a reaction with deterministic waiting time $\tau_1 < \tau_2$ will always fire before a reaction with waiting time τ_2. This is, however, no obstacle: we can simply prepend a virtual Markovian species that is rapidly converted into species that follow one of the two waiting time distributions, and use a common generating function argument for all three species to obtain the sum of these species.

[2]To encode the dynamics of non-catalytic, non-Markovian processes, we exploit the isomorphism between characteristics and survival functions. This is most relevant for discrete systems, but Equation 4.30 suggests the correct way to connect this framework to continuous systems: Markovian interconversion is essentially an exponentially weighted moving average; deterministically-timed interconversion is a simple delay (a degenerate single-point "moving average"); generic non-Markovian interconversion maps to a generic moving average kernel. However, at this point, this framing is largely a mathematical curiosity.

[3]This nomenclature is somewhat at odds with the usual usage in [81, 277, 298]. Here, we use it to mean this particular noise model, in the spirit of, but with a narrow meaning than [124]. Elsewhere throughout the thesis, we somewhat loosely use it to indicate biological variability attributable to cell-to-cell differences, as in [136]. The nomenclature has evolved somewhat in the constituent reports, and by, e.g., [113], we converged on the following arbitrary but intuitive convention: in a distribution induced by a purely biological process, "intrinsic" noise is the Poisson component of variance or CV^2, "extrinsic" noise is everything else.

[4]Here, a particularly imaginative reader of [112] and [115] will exclaim: "Hold on, if this entire formalization just amounts to fitting Equation 4.33, why do we need joint nascent and mature RNA data at all?" They will be correct: this even works for $n = 1$. The key idea is that distributions of stochastic systems exhibit exponential convergence to their steady state, so near-equilibrium states will be more "condensed" and far-from-equilibrium states will be more "dispersed," because they are less stable, hence fewer cells will be sampled from them. Having multivariate data provides the advantage of disambiguating the dynamics when multiple attractors exist, e.g., the trajectories above and below the equilibrium line showcased in [168], which encode this exponential convergence to the high- and low-expression attractors, and would be difficult to impossible to distinguish based on a single modality.

[5]Here, an extraordinarily careful and attentive reader of [113] may raise a very natural question. The CIR model is derived under the assumption the regulator is present at high concentrations and largely resides near the mean. Why should we consider this limit, which is typically near zero and exhibits jump behavior? The justification is fourfold. First, there may be other systems that lead to CIR driving behavior without the approximation we have made, perhaps relating to membrane transport rather than production and degradation of regulators. Second, although the limit is nonphysical, the *convergence* is relatively rapid, so there are large portions of parameter

space where the limiting behavior is an approximately valid description of the dynamics without violating the assumptions too harshly. Interestingly, the convergence to the extrinsic noise model is considerably slower. Third, it motivates the consideration of relatively "exotic" drivers that are otherwise ignored throughout this thesis: elsewhere, we assume all jump processes are compound Poisson. Fourth, the the continuous CIR process is equivalent to discrete autocatalysis, per Section A.8.3.2, so there may be an analogous fully discrete multi-stage process that follows the same statistics; however, we have not investigated this direction further.

[6]A careful reader of [107] may raise the concern: if we observe high expression of short nascent transcripts, does this mean these transcripts are present at *even higher* abundance in the cell, potentially in the thousands or tens of thousands of copies? The answer is unclear, and will require dedicated study of specific genes. This is likely a combination of effects: real high biological expression, limitations in the choice of reference, limitations of the constant-C_N assumption, obscure technical artifacts like priming of spliced-out introns.

[7]More strictly, the *zero-inflated* negative binomial (ZINB) distribution is most popular for *scVI*. However, its meaning is unclear; the distribution appears to be an obsolete [274] holdover from pre-single-cell and single-molecule sequencing technology analyses. I am aware of one publication that attempts to provide a basis for this model [149], but it conflates a technical effect (dropout) with a biological one (promoter state switching) without any apparent justification. Ultimately, the problem with using a ZINB model to describe technical variation is that there is no mechanistic motivation for a process that leads to the loss of all of a gene's molecules, nor an explanation for why it should choose *that* gene without depleting others. That said, the ZINB distribution can be immediately obtained in the slow limit of an $N = 2$ model with a bursty and an inactive state.

[8]Here, a subtlety emerges. I do not believe that, e.g., landscapes or gradients are a helpful way to conceptualize *discrete molecule counts*, because forcing a discrete object into a Procrustean bed of continuous models appears to be questionable modeling practice. That said, they may be effective for the underlying *continuous parameters*. For example, the model I simulate from in Section 6.2 is precisely a (non-Markovian) graph traversal that represents transitions between cell types, and the time-varying $\mathcal{H}$ operator introduced in Chapter 4 indicates some generic transient process, which may, in turn, be reasonably well-approximated by a simple function. This function could be axiomatic (i.e., "a cell type is a point mass with respect to a set of parameter distributions") or reflect a specific biophysical process. The key idea is that we have to operate on a slightly higher level of abstraction if we want to use these models, because the layers of single-molecule biological and technical stochasticity are inherent and non-negotiable. Considerable further work remains to determine whether this approach is promising, whether it leads to intractable problems, or whether it amounts to kicking the can down the road without providing any useful scientific insights.

[9]We can, of course, define dedicated bound states that are only accessible through molecule interactions, but this makes no difference to the mathematical challenges.

[10]This point is explored in substantial detail in [108].

[11]A naïve reading of this result might suggest that we have spent a considerable amount of theoretical effort to recapitulate something that is already fairly accepted practice. This reading is grossly incorrect, and the fact that the terms happen to match in this simplest of cases should only

encourage us to be more vigilant and thorough about developing and disclosing explicit mechanistic models, as they can provide insights about the basis for the success of procedures which are, at first glance, *ad hoc*. Further, the result *does not* fully agree with the standard description: for example, the interpretation of the Poisson variance term as purely technical does not hold under this model.