Biophysical Modeling for Gene Expression and Evolution

Thesis by
Catherine Felce

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy in Physics

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2026
Defended December 10th, 2025

Catherine Felce
ORCID: 0009-0009-9909-6711

Dedicated to

Simeon Henry Brian Felce

In loving memory of

Marion Felce

ACKNOWLEDGEMENTS

Firstly, I would like to thank my advisor, Lior Pachter, who has been remarkably supportive and inspiring, always bringing a wealth of interesting ideas and knowledge to our discussions. I have deeply appreciated his perceptive academic guidance and genuine care. I thank my other committee members, Rob Phillips, Matt Pennell, and Matt Thomson for their helpful input and encouragement. Thank you to each and every member of the Pachter lab, past and present, especially my wonderful collaborators Meichen Fang and Gennady Gorin. I am particularly grateful to Laura Luebbert, Tara Chari, Taleen Dilanyan, Anne Kil, Rebekah Loving and Delaney Sullivan for welcoming me into the lab and showing me the ropes, and to Maria Carilli for inviting me. Thank you also to our fantastic lab manager, Charlene Kim. Outside of the lab, it was a pleasure working closely with Matt Pennell, Alex Cope, Joshua Schraiber, Madhumitha Krishnaswamy and Jim Fuller. Thank you to Annika Dugad, my PMA mentor, and to my undergraduate professors, who were kind and did much to inspire me in physics, especially Alex Schekochihin, James Binney, Alan Barr and Simon Hooker.

On a personal note, I want to thank my parents, whose generosity and guidance have been foundational. I am hugely grateful to my wider family, my cousins, aunts and uncles, whose warmth and strength inspire me and who are always rooting for me. Thank you especially to Heather and Jose Peña and their children, who have been my California home away from home. I thank all my friends in England, especially Jasmine, my sister through school and beyond, and Georgia, who got me through undergrad and somehow always understands how I feel. I am grateful to Philippa, for her word in time, and Grace, for her faithful love and her scientific contributions. Thank you Sophie for your encouragement; Sanela, Seeta, Olivia and Jack for your visits; and Cecilia for taking me in during the Eaton fire.

I am very grateful to my Caltech friends, who have given me much joy and invaluable support during my PhD. I met so many fantastic people who all helped to make grad school fun, and whom space does not permit to be named here. Thank you to Elvira Moreno, for listening to me a thousand times, to Elina Sendonaris, who braved wind and fire to bring me my car, and to Haroula Baliaka for her honesty and for housing me for six months. Special thanks to Elsie Loukiantchenko, Ailene Chan, Ina Sørensen, Su Direkci, Elie Bataille, Tom Henning, Sophia Tevosyan, Rohit Dilip and Jack Adeney. I will miss all of you and hope to keep in touch. Thank you to each of the members of the Graduate Christian Fellowship, Emily Sanger and the RA team, the Chapel at Pasadena and the Biola book group.

Finally, I am particularly grateful to my brothers and their families, especially little Simeon and Miriam, for all the joy and laughter they have brought. I also want to thank my grandparents on both sides, whom I still look up to and remember with fondness. I thank God for putting all of these wonderful people in my life.

ABSTRACT

Principled biophysical modeling is a necessary foundation for analyzing RNA sequencing data. In recent years, higher quality data for other data modalities at single-cell resolution have become available. I present joint biophysical models combining two of these modalities, chromatin accessibility measurements (ATAC-seq) and protein counts, individually with single-cell transcriptomic data, and give preliminary data results. I consider the extension of biophysically motivated models to the field of phylogenetics. I present competing mechanistic hypotheses for gene expression evolution and test them via parametrized single-cell cross-species data. I also consider a physics-inspired model for population-level evolution via maternal effects and interacting subpopulations.

PUBLISHED CONTENT AND CONTRIBUTIONS

Felce, Catherine, Alexander L. Cope, et al. (2025). "Biophysical Constraints on mRNA Decay Rates Shape Macroevolutionary Divergence in Steady-State Abundances". In: *bioRxiv*. DOI: 10.1101/2025.11.24.690267. eprint: 2025.11.24.690267. URL: https://doi.org/10.1101/2025.11.24.690267.
Catherine Felce performed the analyses and conceptualized the model.

Felce, Catherine, Meichen Fang, and Lior Pachter (2025). "Joint Biophysical Modeling of Paired Single-Cell RNA and Protein Measurements". In: *bioRxiv*. DOI: 10.1101/2025.11.14.688548. eprint: 2025.11.14.688548. URL: https://doi.org/10.1101/2025.11.14.688548.
Catherine Felce performed the analyses with Meichen Fang.

Felce, Catherine, Steinunn Liorsdóttir, and Lior Pachter (Nov. 2025). "Analogies between the virial theorem and the Price equation". In: *Phys. Rev. E* 112 (5), p. 054139. DOI: 10.1103/r8bd-lhm3. URL: https://link.aps.org/doi/10.1103/r8bd-lhm3.
Catherine Felce conceived the model and performed the analysis and writing of the sections on simple harmonic motion and spatial population growth.

Felce, C., G. Gorin, and L. Pachter (Dec. 2024). "Biophysical model for joint analysis of chromatin and RNA sequencing data". In: *Phys. Rev. E* 110 (6), p. 064405. DOI: 10.1103/PhysRevE.110.064405. URL: https://link.aps.org/doi/10.1103/PhysRevE.110.064405.
Catherine Felce conceived the model in collaboration with G.G., and performed the analyses.

Felce, Catherine and Jim Fuller (Oct. 2023). "Slowly Rotating Close Binary Stars in Cassini States". In: *Monthly Notices of the Royal Astronomical Society* 526.4, pp. 6168–6180. ISSN: 0035-8711. DOI: 10.1093/mnras/stad3053. eprint: https://academic.oup.com/mnras/article-pdf/526/4/6168/52633417/stad3053.pdf.
Catherine Felce contributed to the theory and performed the numerical integrations and data analysis.

Fuller, Jim and Catherine Felce (Oct. 2023). "Super Slowly Spinning Stars in Close Binaries". In: *Monthly Notices of the Royal Astronomical Society: Letters* 527.1, pp. L103–L109. ISSN: 1745-3925. DOI: 10.1093/mnrasl/slad150. eprint: https://academic.oup.com/mnrasl/article-pdf/527/1/L103/54610279/slad150.pdf.
Catherine Felce contributed to the theory and data collection.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

In physics, we celebrate finding the most universal, yet simplest principles which govern the behavior of complex inanimate systems. The search for such general principles underlying biology has historically been elusive (Dhar and Giuliani, 2010). Whilst scale separation often makes physics problems tractable (Modin and Viviani, 2022; Falkowski, 2023; Wilson, 2025), biologically interesting processes combine intricate webs of interaction spanning multiple length and time-scales (Bergelson et al., 2021; Dada and P. Mendes, 2011). The consistent laws and fungible fundamental particles of physics are far from the reality of biology, where organismal history and context have persistent effects (Ellis and Kopel, 2019), and variation between biological objects has been described as a 'fundamental theoretical principle' (Montévil et al., 2016).

However, as the abundance and resolution of biological data increases, more quantifiable biological models are within reach. Biologists are calling for the methodologies from physics to be brought to bear on fundamental questions in cellular biology (Phillips and Quake, 2006). Perhaps biology's most closely analogous field in physics is astrophysics, where a similar explosion of quantitative data has brought new opportunities and challenges (Schadt et al., 2010; Z. D. Stephens et al., 2015). In astrophysics, as in cellular biology, each observed phenomenom must be described via the integration of multiple physical theories. Neutron star mergers demand understanding of both general relativity, particle and dense matter physics (B. P. Abbott et al., 2017). Main sequence star binary descriptions may be incomplete without both classical mechanics and magnetohydronamics (Catherine Felce and Fuller, 2023). And similarly, the process of DNA transcription cannot be understood outside of an understanding of chromatin remodeling, transcription factor binding and polymerase recruitment (to name only the high level processes involved).

Since the primary foci of investigation in astrophysics and cellular biology are, in a way, fixed (i.e. the existing astronomical objects and organisms), both fields evade a completely reductionist treatment. Whereas particle physicists can create new phenomena by experiments at higher and higher energies, in astrophysics, as in cellular

biology, our path to greater insight is through greater resolution measurements of existing systems. As Phillips and Quake compare, increasingly comprehensive genomics assays could play a similar role to spectrometry, which allows astronomers to precisely ascertain both the speeds and chemical compositions of astrophysical bodies. Indeed, exact DNA sequencing and the precise quantification of RNA transcripts have already transformed the field of genomics and precipitated remarkable discoveries (Lander et al., 2001; Mortazavi et al., 2008; F. Tang et al., 2009; ENCODE Project Consortium, 2012; GTEx Consortium, 2015) .

In the next section, I will discuss how to build well-motivated models for transcriptomic data, moving closer to the level of rigor of physical theories available to analyze astrophysical observations. When considering such models, a difference between astrophysical and biological data is brought into focus, namely the difference in physical scale. This obvious but important discrepancy has implications for the kinds of models appropriate to each field. Whereas most astrophysical phenomena are on a large enough scale that we can 'average' over enough of the underlying randomness (quantum mechanics or individual particles within astrophysical fluids) that the dynamics appear deterministic, in biological systems, stochasticity is vitally important. In addition, the focus of biophysical modeling is on non-equilibrium systems. These differences imply that principled stochastic modeling, rather than deterministic laws, are the basis of a large class of models for cellular biology. We hope that such models can form the basis of a more 'bottom-up' approach to our understanding of cellular processes.

While 'top-down' research will remain essential in applications like medical research, with increasing resolution of genomic data, an improved theoretical understanding of the normal operation of cells and their gene expression is within reach. In biology, as in physics, although reductive models are not applicable to every regime, they can bring important insights with the correct application.

1.1 Biophysical modeling

Historically, most single-cell gene expression analysis has consisted in simply identifying differentially expressed genes between groups of cells (T. Wang et al., 2019; Das, A. Rai, and S. N. Rai, 2022; Dal Molin, Baruzzo, and Di Camillo, 2017). Model-agnostic mathematical tools, for example in dimension reduction and trajectory inference, have been applied to single-cell expression data with problematic results (Huang, Yam, and N. L. Tang, 2025; Gennady Gorin, Fang, et al., 2022; T.

Chari and Lior Pachter, 2023). Biophysical modeling attempts to replace heuristic descriptions of biological systems with principled foundational models. In particular, biophysical modeling of DNA transcription can allow us to distinguish between simple but distinct transcriptional mechanisms (Gennady Gorin, Vastola, et al., 2022). Analyzing single-cell RNA-seq datasets using these models can allow us to determine delineations between subpopulations that cannot be detected at the level of mean expression (Chari and Pachter, 2024). Precise biophysical modeling has proven useful in identifying and distinguishing technical, statistical and biological effects in transcriptional data (Cao, Yiling Wang, and Grima, 2025; Kim and Marioni, 2013; Bohrer and Roberts, 2016). It has been particularly effective in quantifying transcriptional bursting in terms of burst size and frequency (Larsson et al., 2019; Grima and Esmenjaud, 2024).

In (Phillips and Quake, 2006), the authors identify "understanding the collective effects that give rise to the exquisite organization in space and time revealed by cellular life", as an area for investigation in physics. They highlight that the 'final arrow' in the central dogma, which should connect the products of translation back to the DNA, is often neglected. They express the hope that, with increasingly quantitative data, these phenomena will become accessible to rigorous characterization. We have attempted to answer this call by integrating new experimental techniques into existing biophysical modeling frameworks. In particular, I investigate the analysis of data from two experimental techniques, ATAC-seq and protein sequencing (CITE-seq), which both shed light on 'closing the loop' of the central dogma.

ATAC-seq (Cusanovich et al., 2015), an assay for accessible chromatin regions, is a quantitative measure of chromatin configuration. Whilst chromatin configuration affects the transcriptional activity of different genes, it is also impacted by the proteins these genes produce. Chromatin binding factors play a key role in determining DNA configuration (Klemm, Shipony, and Greenleaf, 2019), and jointly modeling chromatin accessibility and gene expression therefore provides insight into the complex feedback loops that result in time-organized cellular processes. By considering gene-gene correlations at the DNA level in a statistical mechanical framework in Chapter 2, we cohere with the analogy given in (Phillips and Quake, 2006) between emergent phenomena in physics and collective organization in biology.

Direct measurement of expressed proteins represents another tool for investigation of the central dogma loop. CITE-seq (Stoeckius et al., 2017) gives simultaneous RNA and surface protein counts for individual cells, allowing for joint modeling of

the proteome and transcriptome at single-cell resolution. Since RNA counts have historically been used as a proxy for protein expression, disentangling the processes of transcription and translation has the potential for significant impact (Xiaojing Wang, Liu, and Zhang, 2014).

1.2 Biophysical modeling and evolution

Similarly to with gene expression, biological studies of evolution initially stopped short of rigorous quantification. As Felsenstein bemoans in his seminal work, 'Phylogenies and Quantitative Characters' (Felsenstein, 1988), species-level trait analysis was initially kept separate from the quantitative methods of evolutionary genetics. Whilst systematists grouped species by morphological traits, evolutionary geneticists restricted their precise gene-level observations to within-population studies. An entire school of thought (pattern cladism) rejected the notion of phylogenetic inference from similarity of traits between species. Before Felsenstein and others developed tools for their unification, the study of quantitative genetics and systematics were therefore entirely separate.

Felsenstein saw that, with the increasing availability of molecular data (e.g. DNA sequences) across species, evolutionary genetic models could now be extended 'across the species boundary'. His initial models, including the Brownian motion model for genetic drift, were perhaps overly simplistic, using restrictive assumptions and outright ignoring important dynamics such as selection itself. Nevertheless, the methodological tools developed by Felsenstein and others to bridge the gap between these top-down (systematist) and bottom-up (evolutionary genetics) approaches are now the foundations of the entire field of phylogenetic comparative methods (PCMs).

Whilst phylogenetic correlations were initially seen simply as a problem to be overcome to obtain independent trait measurements between species (Felsenstein, 1985), quantitative traits in combination with phylogenetic trees are now used to infer the dynamics of evolutionary processes (F. K. Mendes et al., 2018; Hadfield and Nakagawa, 2010; O'Meara, 2012). Many of these studies are based on Brownian motion, or Ornstein-Uhlenbeck (Brownian motion with the addition of mean-reversion), models of trait evolution (Butler and King, 2004; D. C. Adams, 2012; Dibán and Hinojosa, 2024). More complex models, including varying evolutionary optima across clades, have allowed us to identify and describe complicated phenomena such as convergent adaptive radiations (Mahler et al., 2013).

Many of these phylogenetic models use gene expression levels, determined from

mRNA measurements, as their quantitative trait of interest (Price et al., 2022; Hill, Vande Zande, and Wittkopp, 2021; Blekhman et al., 2008). However, using the level of gene expression itself as an evolving trait obscures the underlying biophysical mechanisms for transcription. The same problems with the normalization of gene expression values which motivate biophysical modeling (Bullard et al., 2010) have been known to affect PCMs using gene expression as continuous characters (Dimayacyac et al., 2023a). In Chapter 4, I explore the integration of biophysical modeling and phylogenetic comparative methods. Similarly to initial models for integrating systematics and evolutionary genetics, the proposed approach is simple, but I hope that it too may provide a useful bridge between these currently disjoint fields. Our approach can then be refined to capture more complicated histories and develop an increasingly nuanced picture of the evolution of transcriptional dynamics.

1.3 Physical models and ecology

The search for more quantitative descriptions for biological phenomena also extends to population genetics and ecology. As early as 1975, there have been calls for the use of more quantitative methods for ecology (Gates, 1975), under the name 'biophysical ecology'. A biologist and a physicist, Gates endorses the use of mathematical tools from physics and chemistry for building theoretical models in ecology. He emphasizes that these models should be predictive, and amenable to experimental investigation.

Ginzburg and Colyvan adhere closely to this philosophy in their book, 'Ecological orbits: How planets move and populations grow' (Ginzburg and Colyvan, 2004). In it, they propose a new ecological theory, which they believe more efficiently explains observed ecological allometries than pre-existing theories. In particular, they present a new foundational model for population cycles, inspired by analogy to inertial motion in physics. They especially draw a connection between ecological cycles and astronomical orbits, suggesting that environmental 'forces' might produce populational 'acceleration', rather than impacting the 'velocity' (population growth rate) directly. Using this framework, they are able to explain population cycles via maternal effects in a single species, without needing to posit interactions with any additional species, such as traditional predator-prey models (Lotka, 1920). Analogies between physical and evolutionary 'forces' have also been explored in the philosophy of ecology and evolution (C. Stephens, 2004; Justus, 2013; Sagoff, 2016). In Chapter 5 we introduce one such analogy, as well as extending a form of the model from (Ginzburg and Colyvan, 2004).

1.4 Outline

This thesis covers my work in developing biophysical models for RNA-seq data analysis and evolution. We begin with a joint model for integrating RNA-seq and ATAC-seq data, and exploring an extension of the telegraph model to correlated gene neighbors. This model draws inspiration from the Ising model in physics. Chapter 3 then explores the extension of current biophysical models to another modality, protein counts, by combining bursty transcription and constitutive translation. This is work carried out along with Meichen Fang, and we show theoretical results for this extended model, as well as fits to experimental data. Chapter 4 explores the integration of biophysical modeling into phylogenetic inference techniques. I explore the question of disentangling the different mechanisms involved in gene expression evolution. In the final chapter, I continue my focus on evolution, this time using the equivalence of two equations, the virial theorem in physics, and the Price equation in evolutionary biology, to inspire new models. I suggest that a gestational maternal effect should be included in our most basic second-order models for population growth, and join Ginzburg and Colyvan in suggesting that analogies between ecology and astrophysics could be illuminating in describing complex populational dynamics. Finally, I summarize the implications of this work, as well as possible future directions.

Chapter 2

BIOPHYSICAL MODEL FOR JOINT ANALYSIS OF CHROMATIN AND RNA SEQUENCING DATA

Felce, C., G. Gorin, and L. Pachter (Dec. 2024). "Biophysical model for joint analysis of chromatin and RNA sequencing data". In: *Phys. Rev. E* 110 (6), p. 064405. DOI: 10.1103/PhysRevE.110.064405. URL: https://link.aps.org/doi/10.1103/PhysRevE.110.064405.

2.1 Introduction

The Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), introduced in 2013, has become a widely adopted epigenetic discovery tool that can be used to identify regions of open chromatin (Jason D Buenrostro et al., 2013). In 2015, this method was extended to allow for measurements at single-cell resolution (scATAC-seq) (Cusanovich et al., 2015; Jason D. Buenrostro et al., 2015). These techniques were initially developed and applied independently from transcriptomics experiments. As a result, ATAC-seq data are still often analyzed in this manner, often by re-purposing gene-expression tools (Gontarz et al., 2020), and several methods have emerged for analyzing ATAC-seq data in isolation (H. Chen et al., 2019; Bravo González-Blas et al., 2019).

However, since the major interest in ATAC-seq data lies in elucidating the gene expression programs within cells, the data are best understood in combination with gene expression data. The integration of bulk ATAC-seq and RNA-seq data has been carried out in various forms (Jianfang Wang et al., 2022; F. Xu et al., 2023; Ding et al., 2023). However, approaches thus far have been based on heuristics and do not consider underlying biophysical processes such as transcription factor binding, transcriptional bursting, processing and splicing of transcripts, and degradation. ATAC-seq and RNA-seq data can, in principle, inform the specifics of models for these processes. These insights can then be fed back into mechanistic models, which can be used to analyze data and provide information about the dynamics of regulation and transcription.

Approaches for integrating unregistered scATAC-seq and scRNA-seq data, i.e. data collected from different cells from the same sample, include joint analysis of dif-

ferentially accessible regions (DARs) (X. Chen et al., 2022) and differential gene expression. Such methods have been used to compare gene regulation across tissues between healthy and pathological conditions (Jin Wang et al., 2018; Sahinyan et al., 2022; Yu et al., 2023; Nazzari et al., 2023; Dhara et al., 2021; Amarasinghe et al., 2023), and across tissue types (Nair et al., 2021), as well as to identify new gene-regulatory pathways (Z. Xu et al., 2022; S. Wang et al., 2022). Combined use of scRNA-seq and scATAC-seq data can also help to distinguish cell-types (Duren et al., 2018; J. Li et al., 2022) and scATAC data can be used to impute scRNA data (Raevskiy et al., 2023). An exciting recent development has been the ability to perform registered scATAC-seq and scRNA-seq, identifying DNA fragments and mRNA transcripts from the same individual cells (Cao et al., 2018; Ma et al., 2020). Such methods provide a direct avenue for joint analysis of the two modalities, thus avoiding the need for "integration" of scRNA-seq and scATAC-seq (Lee, Kaestner, and M. Li, 2023), albeit raising new challenges including sparsity of registered data (Booeshaghi, Gao, and Pachter, 2023). While both registered and unregistered data have been useful for a variety of applications, analysis methods have not incorporated biophysical modeling, and have instead relied on heuristics and phenomenological inferences (J. Li et al., 2022).

On the modeling side, progress has been made in building biophysically motivated and interpretable models for RNA sequencing data (Gayoso et al., 2021; Gligorijević and Pržulj, 2015; Hao et al., 2021). Gorin et al. have outlined a rigorous and general formulation for various models of transcription with intrinsic and extrinsic noise (Gorin, Vastola, and Pachter, 2023), which are potentially distinguishable using single-cell transcriptomics data. However, in that work, the derived distributions are marginalized over DNA-state, and the distinguishability of models using data including chromatin information is not explored. Currently, no such formulation has been applied to the analysis of ATAC-seq data, or to the joint analysis of ATAC-seq and RNA-seq data.

In this work, we lay the foundations for principled and biologically interpretable joint modeling of ATAC and RNA-seq data (see Figure 2.1). This approach leverages the additional information provided by multiomic data (scATAC and scRNA in the same cells) to give insight into transcriptional rates and mechanisms, improving on current ad hoc approaches to combining the two experimental modalities. It also builds on current theoretical frameworks, allowing us to begin considering which types of models can be distinguished using both data types.

We propose a model for chromatin dynamics, inspired by the Ising model from physics. Building on (Gorin, Vastola, and Pachter, 2023), we extend gene-gene correlations to linear chromatin regions of arbitrary length, and encode these relationships within the chemical master equation (CME) framework. Identifying these chromatin regions with ATAC-seq peak regions allows us to consider inference techniques which leverage both DNA and RNA information to parametrize these models. After reviewing the CME formulation with a simple example, we show how it can be used to model DNA and RNA dynamics for Ising-like loci of arbitrary lengths. In our model, nearest-neighbor coupling between chromatin sites modifies the stability of the available DNA-states, to a degree which can be tuned by model parameters.

We then use a mixture of theory and simulation to explore the implications of our proposed model. We derive first and second-order moments of the system and investigate the propagation of gene-gene correlations from the level of chromatin accessibility to the level of transcript counts. We extend (Gorin, Vastola, and Pachter, 2023) by considering model identifiability in the case where chromatin information is available. By simulating the full system within our modeling framework, and constructing an inference procedure leveraging both chromatin and transcriptomic information, we explore the distinguishability of this type of model using each of these modalities, as well as both combined. We consider inference under multiomic data taken from the same or different cells (registered vs un-registered), shedding light on the value of matched single-cell measurements for the development of biophysical models.

Finally, we fit our Ising-like model to three different single-cell ATAC-seq datasets, comparing it to a simpler model with chromatin-site independence. By positing a technical noise model, we demonstrate an inference procedure which allows the theory of the previous section to be applied directly to ATAC-seq data. We demonstrate the flexibility of our approach by using the chromatin-state transition rate matrix to describe DNA loci with six closely adjacent ATAC-seq peaks. This gives us enough candidate loci to meaningfully assess the performance of our model relative to a model with independent sites. Our analysis provides support for our characterization of chromatin dynamics, and the existence of nearest-neighbor correlations.

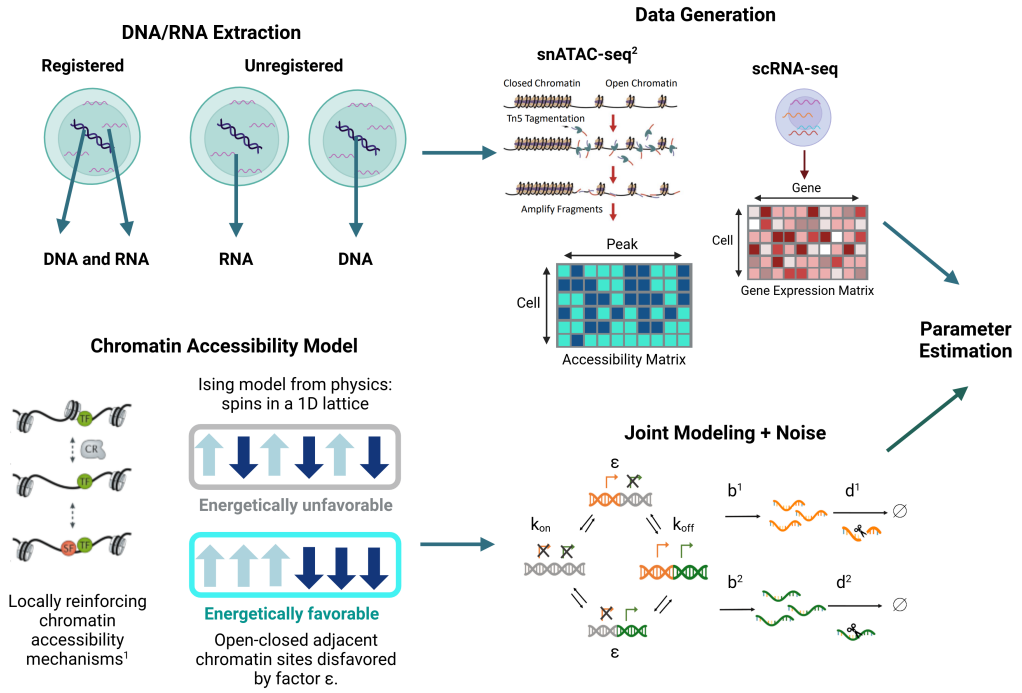


Figure 2.1: Joint modeling of gene accessibility and constitutive mRNA production allows for parameter estimation using scATAC-seq and scRNA-seq data. **DNA/RNA Extraction:** Registered scATAC-seq and scRNA-seq information can be taken from the same cells in multiome data. Alternatively, scATAC-seq and scRNA-seq can be performed on separate cells within the same sample. **Data Generation:** The transcripts captured during scRNA-seq are sequenced and processed to give a cell-by-gene expression matrix. The fragments of accessible chromatin captured during scATAC-seq are sequenced, and accessible regions are identified via peak calling on the bulk data. Fragments from each cell are then aligned to these peak regions, giving a binary cell-by-peak matrix, indicating whether or not each cell contained a fragment from each peak region. **Chromatin Accessibility Model:** We choose a model for chromatin accessibility where neighboring peak regions are correlated within each cell. This is reminiscent of the Ising model from physics. **Joint Modeling:** By combining a Markov chain model for chromatin-state switching, (illustrated, left), with assumptions for subsequent RNA dynamics, (e.g., constitutive transcription at accessible regions), we can estimate the joint distribution of transcript counts and chromatin accessibility across cells. The model is parameterized in terms of RNA characteristics for each transcript species (e.g., transcription rate: b^i, decay rate: d^i, for the i^{th} species), and features of the chromatin-state transition matrix (e.g. neighbor-neighbor correlations: ϵ, rate of gene switching on/off: k_{on}/k_{off}). **Parameter Estimation:** By comparing the observed data with predicted distributions from the joint model after adding noise, we can estimate parameters. References; 1: (Klemm, Shipony, and Greenleaf, 2019), 2: (Dillinger, 2021) Created with BioRender.com (Felce, 2023).

2.2 Model Details

Chromatin Neighbor Interactions

Recruitment of active chromatin re-modelers by pioneer transcription factors could lead to expanding regions of open chromatin (see stabilizing role of secondary transcription factors in (Klemm, Shipony, and Greenleaf, 2019)). These mechanisms could bias adjacent DNA regions towards sharing the same state of transcriptional accessibility. This motivates the use of correlation structure to describe chromatin, and we consider the simplest model that could yield such correlation: first-order neighbor interactions reminiscent of the Ising model from physics. We choose to parameterize this model by a parameter $\epsilon < 1$, which encodes the preference of neighboring chromatin sites to be aligned. $\epsilon = 1$ yields site independence, and allowing ϵ to differ from 1 gives us the next simplest model, where one misalignment in a configuration reduces the probability of the configuration by a factor ϵ. Note that in our parametrization, a smaller value of ϵ corresponds to a stronger correlation between neighboring regions. Such a formulation is supported by exploratory analysis, as shown in Section 2.5.

The Ising model from statistical mechanics describes magnetic behavior in terms of individual dipole spins which can each be in either of two opposite directions, denoted as up and down, or $+1/-1$. In ferromagnetic materials, it is energetically favorable for neighboring spins to be aligned. In the presence of an external magnetic field, it is also favorable for spins to align with the direction of the magnetic field, and the interplay of these two factors determines the probability distribution of different states of the system. Such systems can be simulated using Glauber dynamics, where spins are flipped with a probability proportional to $exp(\frac{-\Delta E}{T})$, where ΔE denotes the change in the energy of the system due to the flip. In our chromatin model, up and down spins are analogous to open and closed regions of chromatin, and our model is therefore effectively the 1-D Ising model.

Ising models have been successfully applied in several areas of computational biology. Mo et al. use an Ising model approach to analyze ChIP-chip data and discover transcription factor binding sites (Mo and Liang, 2010a; Mo and Liang, 2010b). They consider a hidden Markov model (HMM) treating chromosomally consecutive probes in a microarray as neighbors in a 1-dimensional Ising chain. Although their treatment and the connection to the Ising model is similar to ours in this work, the foundation for their Ising hypothesis is technical rather than biological, as it rests on the correspondence of many consecutive enriched probes to a single protein

binding event. The Ising model has also been applied to modeling collective cell organization (Weber and Buceta, 2016).

In the Ising model, the external magnetic field is assumed to be constant for all spins in the lattice. However, inspection of the first-order moments for chromatin accessibility from ATAC-seq data suggests that this feature of the model is not appropriate in our context (see Appendix A.1 for figures showing the mean openness at different ATAC peak sites). Therefore, we allow the ratio of chromatin opening and closing rates to vary between sites, giving a separate 'field strength' parameter per site, along with a common correlation parameter between sites.

CME Implementation

To build a statistical inference framework including this model of chromatin structure, we follow the chemical master equation (CME) approach of Gorin et al. (Gorin, Vastola, and Pachter, 2023). In this approach, the system is assumed to be Markovian, and we evolve the probability distribution across states in a deterministic manner. The state of the system is described by the number of transcripts present for each gene, and the state of chromatin openness for each gene. In our analysis, we have assumed a one-to-one mapping between ATAC-seq peak regions and genes. We assume constitutive transcription whenever a gene is in its open (or on-) state, and no transcription in the closed (or off-) state.

The processes of gene switching, transcription and decay occur at given rates, and these rates are the parameters of the system. These processes each contribute terms to the system of ordinary differential equations (ODEs) which govern the time evolution of the system. For example, decay processes contribute terms to the ODEs which move probability from states with higher transcript numbers to states with lower transcript numbers. These decay terms are straightforward and affect states independently from their chromatin configuration. In contrast, transcription rates at each gene can depend on the underlying state of the DNA.

CME Example

To illustrate the CME framework, we give the simple example of a single-gene system: the telegraph model introduced by Peccoud and Ycard (Peccoud and Ycart, 1995). The chromatin configuration of the system consists of the single gene being either open or closed for transcription, with corresponding transcription rates of b or 0 respectively. The transcript decay rate for this species is d, and the rates of

chromatin opening/closing are k_{on} and k_{off} respectively. The system would then be governed by the following ODEs:

$$\begin{aligned}\frac{dP_0(m,t)}{dt} &= -k_{\text{on}}P_0(m,t) + k_{\text{off}}P_1(m,t) \\ &\quad + d[(m+1)P_0(m+1,t) - mP_0(m,t)] \\ \frac{dP_1(m,t)}{dt} &= k_{\text{on}}P_0(m,t) - k_{\text{off}}P_1(m,t) + d[(m+1)P_1(m+1,t) - mP_1(m,t)] \\ &\quad + b[P_1(m-1,t) - P_1(m,t)],\end{aligned} \tag{2.1}$$

where $P_\alpha(m,t)$ represents the probability of the system to be in the gene state indexed by α, where here $\alpha = 0, 1$ corresponds to the gene being closed/open for transcription respectively, with an integer number, m, of transcripts, at time t. Note that, despite the use of a full derivative, the value of m is fixed in each equation, and Equation 2.1 represents a system of ODEs, one for each possible value of m.

The first terms encode switching between gene states, whilst the decay and transcription terms represent changes in transcript number. The transcription terms only affect the evolution of probabilities in state 1, since no transcription occurs in state 0.

These equations can be summarized in a matrix form. We introduce a system probability vector $\boldsymbol{P}(m,t)$, whose components are the $P_\alpha(m,t)$ described above. The gene-state dynamics can then be encapsulated in the transition matrix, H, given by:

$$H = \begin{pmatrix} -k_{\text{on}} & k_{\text{on}} \\ k_{\text{off}} & -k_{\text{off}} \end{pmatrix}, \tag{2.2}$$

and the evolution of the system can be succinctly described via:

$$\frac{d\boldsymbol{P}}{dt}(m,t) = H^T\boldsymbol{P}(m,t) + d[(m+1)\boldsymbol{P}(m+1,t) - m\boldsymbol{P}(m,t)] + \hat{B}\boldsymbol{P}(m,t), \tag{2.3}$$

where we have also introduced a diagonal transcription matrix, $\hat{B}$, defined such that $\hat{B}_{\alpha\alpha}$ is the transcription rate in the state indexed by α. In this case, $\hat{B}$ is given by:

$$\hat{B} = \begin{pmatrix} 0 & 0 \\ 0 & b \end{pmatrix}. \tag{2.4}$$

This construction can be extended to multiple transcript species and a greater number of gene states. Although we consider only those chromatin-state transitions which can be effected through the opening or closing of a single chromatin region, the framework allows for transitions between any pairs of chromatin states, via modification of the transition matrix.

Encoding Chromatin Dynamics

Working within this CME framework, we turn our attention to encoding the chromatin correlations discussed in Section 2.2.

Restricting Equation 2.3 to chromatin evolution by summing over all possible transcript numbers for each chromatin state, and generalizing to an arbitrary number of possible gene states, we arrive at the chromatin evolution equation:

$$\frac{d\boldsymbol{P}(t)}{dt} = H^T \boldsymbol{P}(t), \tag{2.5}$$

where, here, $\boldsymbol{P}$(t) is a vector whose components, $P_\alpha(t)$ represent the probability of the system to have chromatin configuration α at time t. We can then encode cooperation between neighboring chromatin regions within the chromatin-state transition matrix, H. To illustrate how this can be done, we consider a two-gene system, with correlations between the two genes.

We define a chromatin-state matrix S, such that ${S_\alpha}^i$ gives the openness of gene i in chromatin-state α, and 1/0 correspond to open/closed genes respectively. Note that here and in what follows, for matrices we use superscript Latin indices to reference DNA sites (genes), and subscript Greek indices to reference chromatin configurations. In this work, we will consider the chromatin-state indexed by α as a string of 1's and 0's, representing the openness/closedness of adjacent DNA sites. This gives 2^n configurations for n DNA sites. For the two-gene example system, we would have the state matrix:

$$S = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}. \tag{2.6}$$

For this two-gene system we can choose the gene-state transition matrix:

$$H = \begin{bmatrix} -k_{on,1} - k_{on,2} & k_{on,2} & k_{on,1} & 0 \\ \epsilon^{-1} k_{off} & -r_1 & 0 & \epsilon^{-1} k_{on,1} \\ \epsilon^{-1} k_{off} & 0 & -r_2 & \epsilon^{-1} k_{on,2} \\ 0 & k_{off} & k_{off} & -2k_{off} \end{bmatrix}, \tag{2.7}$$

where $r_i \equiv \epsilon^{-1}(k_{on,i} + k_{off})$, and ϵ is the correlation parameter introduced in Section 2.2. In the Ising model with positive correlations, $\epsilon < 1$, such that the two states that have adjacent chromatin with opposite openness values, (01 and 10), are destabilized relative to the other states, having decay rates and outward transition values which are divided by ϵ and so are large.

Note that to reproduce the Ising stationary distribution for the case of varying site-openness propensities, we need to set k_{off} as equal for all sites, and allow k_{on} to vary between sites. This gives a stationary distribution where the probability of each state is proportional to the product over sites of each k_{on} value, and ϵ to the power of the number of 'misaligned' neighbors in the state. (See Section 2.5.)

This formulation can be extended to systems with more than two adjacent genes, by again allowing all transitions between states differing by the openness value of a single gene, such that genes 'turn off' at a rate k_{off}, and turn on at a gene specific rate, $k_{on,i}$. The rows of the transition matrix should then be multiplied by a factor, $\epsilon^{-n_{mis}}$, where n_{mis} corresponds to the number of misalignments in the outgoing chromatin state (see Appendix A.2).

In our analysis of ATAC-seq data in Section 2.5, we consider a space of 64 (2^6) gene states corresponding to a binary openness value for six adjacent genes. By positing a kinetic relationship between each of these gene states and their RNA dynamics, we can describe the evolution of the system through a joint state space simultaneously encoding DNA configuration and transcriptome composition.

RNA Transcript Count Evolution

Now that we have encoded the desired chromatin dynamics into the chromatin-state transition matrix, we can leverage the machinery of the CME formulation to solve for the resulting RNA distributions in a simple model.

Returning to the evolution of the joint probability distribution for RNA transcripts and chromatin state, as in Equation 2.3, we now expand to the case of multiple

genes. We define the probability generating function (PGF) with gene-state index α as:

$$G_\alpha(\boldsymbol{z}, t) \equiv \sum_{\boldsymbol{m}} \boldsymbol{z}^{\boldsymbol{m}} P_\alpha(\boldsymbol{m}, t), \tag{2.8}$$

where $P_\alpha(\boldsymbol{m}, t)$ represents the probability of the system to be in the state indexed by α and have m^i RNA counts for each transcript i, and $\boldsymbol{z}$ represents the vector of PGF arguments, $[z^1, ..., z^n]$, for RNA species 1 to n. We define $\boldsymbol{z}^{\boldsymbol{m}} \equiv (z^1)^{m^1}(z^2)^{m^2}...(z^n)^{m^n}$.

Let $\boldsymbol{G}$ represent the vector of PGFs, $[G_0, ..., G_{N-1}]$, for gene states 0 to $N-1$. Let B be the production matrix, whose entries ${B_\alpha}^i$ represent the rate of production of species i in state α. For the simplest model, where transcript i is produced at rate b^i for states where gene i is active and at rate zero otherwise, we have the relation: $B = S\,\mathrm{diag}(\boldsymbol{b})$, for $\boldsymbol{b}$ the vector of transcription rates with components b^i. Let $\boldsymbol{d}$ be a vector representing the decay rates of each species i.

Following the procedure outlined in Gorin et al. (Gorin, Vastola, and Pachter, 2023), we then have the full evolution equation:

$$\partial_t \boldsymbol{G} = H^T \boldsymbol{G} + \Big[[\boldsymbol{d} \odot (1 - \boldsymbol{z})]^T \boldsymbol{\partial}\Big] \boldsymbol{G} - \mathrm{diag}[B(1 - \boldsymbol{z})]\boldsymbol{G}, \tag{2.9}$$

where the Hadamard product between two matrices is $(X \odot Y)_{\alpha\beta} \equiv X_{\alpha\beta} Y_{\alpha\beta}$ for any two matrices X and Y with the same dimensions (here dimension $n \times 1$ for n the number of RNA species). $(1-\boldsymbol{z})$ is the vector with components $1-z^1, 1-z^2, ..., 1-z^n$, and the second term on the right-hand side of Equation 2.9 involves a sum over partial derivatives with respect to PGF variables z^i.

The cooperation of adjacent chromatin regions only affects the evolution of the system via the transition matrix, which can be chosen to be of the form in Equation 2.7. We can then solve Equation 2.9 numerically (see Section 2.4), allowing us to characterize the impact of chromatin-level correlations on RNA transcript distributions.

2.3 Model Implications

Having encoded nearest-neighbor chromatin interactions into the CME framework, we can straightforwardly derive the statistical implications of our model. In particular, we consider the first and second-order moments of the system at both the

chromatin and RNA levels, and the behavior of correlations between genes at these two levels. The PGF evolution equation (2.9) above leads to the following moments and statistical properties, before considering noise.

Chromatin-State Moments

For considering moments, we introduce the vector $\boldsymbol{\pi}$, representing the steady-state probabilities of the chromatin states of the system. The vector $\boldsymbol{\pi}$ has the length of the entire state space, i.e. 2^n, where n is the number of sites in the locus. To obtain the steady state, we set the left-hand side of Equation 2.5 to zero. In terms of the chromatin-state transition matrix, H, this gives:

$$\boldsymbol{\pi} = \text{Norm}(\text{Kernel}[H^T]), \tag{2.10}$$

where $\boldsymbol{\pi}$ is normalized such that its entries sum to one. We denote the chromatin-state vector by $\boldsymbol{\sigma}$, where σ^i represents the openness of the i^{th} chromatin region. (For the gene state indexed by α, $\boldsymbol{\sigma}$ is equivalent to row α of the state matrix, S.) We consider a first moment of the system, the average openness value of an individual DNA site, i. From the state matrix S, defined above, we have:

$$\langle \sigma^i \rangle = \sum_\gamma \pi_\gamma S_\gamma{}^i = (\boldsymbol{\pi}^T S)^i, \tag{2.11}$$

where the $\langle\rangle$ denote expected values and σ^i represents the openness of the i^{th} region. The gene-state covariance in this notation is given by:

$$\text{Cov}(\sigma^i \sigma^j) = \sum_\beta \pi_\beta S_\beta{}^i S_\beta{}^j - \sum_{\alpha\beta} \pi_\alpha S_\alpha{}^i \pi_\beta S_\beta{}^j, \tag{2.12}$$

where the variance can be obtained by setting $i = j$ in this expression.

RNA Transcript Count Moments

To find moments concerning the i^{th} RNA transcript species, we differentiate Equation 2.9 with respect to z^i (see Appendix A.3 for details). This gives the transcript means for each species i:

$$\mu^i = \frac{1}{d^i} \sum_\alpha S_\alpha{}^i b^i \pi_\alpha = \frac{b^i}{d^i} \langle \sigma^i \rangle. \tag{2.13}$$

Defining the matrix:

$$M^i \equiv (d^i I - H^T)^{-1} \tag{2.14}$$

we arrive at an expression for the transcript covariances:

$$\mathrm{Cov}(m^i m^j) = \frac{b^i b^j}{d^i + d^j} \sum_{\beta\gamma} [S_\beta{}^i M^j_{\beta\gamma} S_\gamma{}^j + S_\beta{}^j M^i_{\beta\gamma} S_\gamma{}^i] \pi_\gamma - \mu^i \mu^j, \tag{2.15}$$

and variances:

$$\mathrm{Var}(m^i) = \frac{(b^i)^2}{d^i} \sum_{\alpha\beta} \left(S_\alpha{}^i M^i_{\alpha\beta} S_\beta{}^i \pi_\beta \right) + \mu^i - (\mu^i)^2. \tag{2.16}$$

Correlation Propagation

Now that we have derived the covariances between genes at both the chromatin and RNA levels, we can investigate the relationship between the gene-gene correlations at the two levels.

We define site-site correlations for accessibility and transcript count respectively via:

$$\begin{aligned} \rho^{ij}_\sigma &= \frac{\mathrm{Cov}(\sigma^i \sigma^j)}{\sqrt{\mathrm{Var}(\sigma^i)\mathrm{Var}(\sigma^j)}}, \\ \rho^{ij} &= \frac{\mathrm{Cov}(m^i m^j)}{\sqrt{\mathrm{Var}(m^i)\mathrm{Var}(m^j)}}. \end{aligned} \tag{2.17}$$

We define f as the ratio between these correlations:

$$f \equiv \frac{\rho^{ij}}{\rho^{ij}_\sigma}, \tag{2.18}$$

and an expression for f, along with its value plotted for a range of parameters, is given in Appendix A.4. We might assume that correlations between the transcript numbers for different genes would be weaker than the chromatin-level correlations between the parent genes, given that transcript dynamics are downstream of chromatin-state dynamics. We might also expect that the correlations at the transcript and chromatin levels would have the same sign. These constraints together would suggest that f should be contained within the range $0 < f < 1$. However, f can exceed 1 for certain parameter ranges in our model. We visualize this in Appendix A.4. We also

explore the intuition behind this result in Appendix A.5, constructing two illustrative toy systems, with $|f| > 1$ and $f < 0$ respectively, and confirming the behavior of their correlations.

2.4 Distinguishability Comparison: Multiome vs Unregistered

The joint biophysical model we have outlined can also be used to inform choices related to experimental design. In this section we explore the identifiability of the ϵ parameter using both registered and unregistered scRNA-seq and scATAC-seq data.

First we simulate the model described in Section 2.2 for a two-gene system of known ϵ, ($\epsilon = 0.3$), and sampled steady-state gene states and RNA counts for $20,000$ cells. We then solve for the joint steady-state distribution of the system analytically (using Equation 2.9) given different values of ϵ, whilst keeping the overall probabilities for each gene to be on constant (by adjusting k_{off}). In this exercise, we neglect noise.

The analytical probability mass function $P_{analytic}$ is obtained numerically. First, we define the auxiliary vector variable $\boldsymbol{u} := \boldsymbol{z} - 1$. Next, we apply the method of characteristics to Equation 2.9, obtaining a system of ordinary differential equations (cf. Eq. 51 of (Gorin, Vastola, and Pachter, 2023)):

$$\frac{d\boldsymbol{G}(\boldsymbol{U}(\mathsf{s}), T(\mathsf{s}))}{d\mathsf{s}} = -H^T \boldsymbol{G} - \mathrm{diag}[B\boldsymbol{U}(\mathsf{s})]\boldsymbol{G}. \tag{2.19}$$

In this notation, s is a characteristic variable, $T(\mathsf{s}) := t - \mathsf{s}$ is the characteristic time defined in terms of current process time t, and the RNA-specific entries of $\boldsymbol{U}(\boldsymbol{u}, \mathsf{s})$ solve the following system (cf. Eq. 52 of (Gorin, Vastola, and Pachter, 2023)):

$$\frac{d\boldsymbol{U}}{d\mathsf{s}} = -\mathrm{diag}[\mathrm{d}]\ \boldsymbol{U}. \tag{2.20}$$

The initial condition for Equation 2.20 is $\boldsymbol{U}(\mathsf{s} = 0) = \boldsymbol{u}$, implying that each entry $\boldsymbol{U}^i = u_i e^{-d_i \mathsf{s}}$.

To obtain the generating function at a particular time t and generating function coordinates $\boldsymbol{z}$, we directly solve the system of ordinary differential equations in Equation 2.19 using the *SciPy* function `integrate.solve_ivp` (Virtanen et al., 2020). We integrate from $\mathsf{s} = t$ to 0, yielding $\boldsymbol{G}(\boldsymbol{U}(0), T(0)) := \boldsymbol{G}(\boldsymbol{u}, t)$ by construction. The initial condition for this initial value problem is the overall PGF of the system at $t = 0$; to accelerate convergence, we begin with the (appropriately normalized) initial condition $\ker(H^T)$, i.e. the DNA states in their equilibrium distribution and

no molecules of RNA present. Finally, to obtain the overall distribution, we evaluate $\boldsymbol{G}$ over a grid of $\boldsymbol{z}$, then individually convert the generating function for each state σ to a probability mass function using an inverse fast Fourier transform (Singh and Bokes, 2012; Virtanen et al., 2020).

For each analytic solution, we assume that each sampled cell is an i.i.d. variable drawn from this distribution, and therefore that the probability of obtaining this set of simulated data is given by the product of the probabilities for each cell.

For registered data, we calculate the log-likelihood function given by:

$$LL_{reg} = \sum_{c} \log\left[P_{analytic}(\boldsymbol{\sigma_c}, \boldsymbol{m_c})\right], \tag{2.21}$$

where c indexes cells in the simulated data, $\boldsymbol{\sigma_c}$ is the gene state observed in cell c (as defined in Section 2.3), and $\boldsymbol{m_c}$ is the vector of transcript numbers for cell c (as defined in Section 2.2).

To show the distinguishability achievable with scATAC-seq or scRNA-seq data alone, we repeat the same process, using the respective marginal probability distributions:

$$\begin{aligned} \tilde{P}(\boldsymbol{\sigma}) &= \sum_{\boldsymbol{m}} P(\boldsymbol{\sigma}, \boldsymbol{m}), \\ \tilde{P}(\boldsymbol{m}) &= \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}, \boldsymbol{m}). \end{aligned} \tag{2.22}$$

This gives the log-likelihoods:

$$\begin{aligned} LL_{ATAC} &= \sum_{c} \log\left[\tilde{P}(\boldsymbol{\sigma_c})\right], \\ LL_{RNA} &= \sum_{c} \log\left[\tilde{P}(\boldsymbol{m_c})\right]. \end{aligned} \tag{2.23}$$

To compare to the unregistered RNA-seq and ATAC-seq data scenario, we sum over log-likelihood contributions from both of the marginal distributions:

$$LL_{unreg} = \sum_{c} \log\left[\tilde{P}(\boldsymbol{\sigma_c})\right] + \sum_{c'} \log\left[\tilde{P}(\boldsymbol{m_{c'}})\right], \tag{2.24}$$

where these two sums are over separate groups of sampled cells.

To assess the information gained from each modality, we make two different comparisons. Firstly, we consider the case where the total number of sampled cells is kept constant. The log-likelihoods in eq:reg-LL,eq:ATAC-RNA-LL,eq:unreg-LL are calculated for the same $10,000$ sampled cells in each case. For the unregistered case (Equation 2.24), the first term is a sum over half of these cells, and the second term is a sum over the remaining half. These results are shown for a simulation with $\epsilon = 0.3$ in the top panel of Figure 2.2. The other parameters used were $k_{on} = 0.1$, $b_1 = 10$, $b_2 = 15$, $d_1 = 5$, $d_2 = 7$, and a value of k_{off} calculated to give an overall probability $p_{on} = 0.4$ for each gene to be open. The simulation was performed to a maximum simulation time of 10 divided by the slowest biological rate (here k_{on}), to ensure that equilibrium had been reached.

Secondly, we consider the case of fixed experimental cost. We use publicly available prices from 10x Genomics for separate scATAC-seq and scRNA-seq vs registered multiome per cell (Genomics, n.d.[c]), to estimate the ratio of prices for the different assay types. Although the exact ratio of costs depends on the number of reactions desired, the results that follow are not sensitive to the exact ratio. We use the rounded ratio given by pricing for a single sample of 10,000 cells:

$$\mathrm{s}cATAC : \mathrm{s}cRNA : \mathrm{m}ultiome = 1 : 1.2 : 1.8, \tag{2.25}$$

and we take the cost for unregistered scATAC and scRNA as the average of the scATAC and scRNA prices per cell ($(1 + 1.2)/2 = 1.1$). We then weight the number of cells for each modality by the inverse of its price, using 10,000 sampled cells for the multiomic case, and proportionally more, from the remaining pool of simulated cells, for the other modalities. The log-likelihood curves obtained via this process are shown in the bottom panel of Figure 2.2.

The results in Figure 2.2 reflect the fact that, whilst multiomic data provides the most distinguishing power per cell for models of this type (shown by the multiomic log-likelihood curve giving the sharpest peak around the true ϵ value in the top panel), after adjusting for cost, the results are more complicated. Since the price of obtaining multiomic data is close to the sum of performing scATAC and scRNA separately, unregistered data can sample almost twice as many cells for a similar cost. However, information is also lost in the lack of registered RNA and ATAC information from the same cells. The bottom panel of Figure 2.2 shows that unregistered and registered multiomic data are comparable in their ability to distinguish

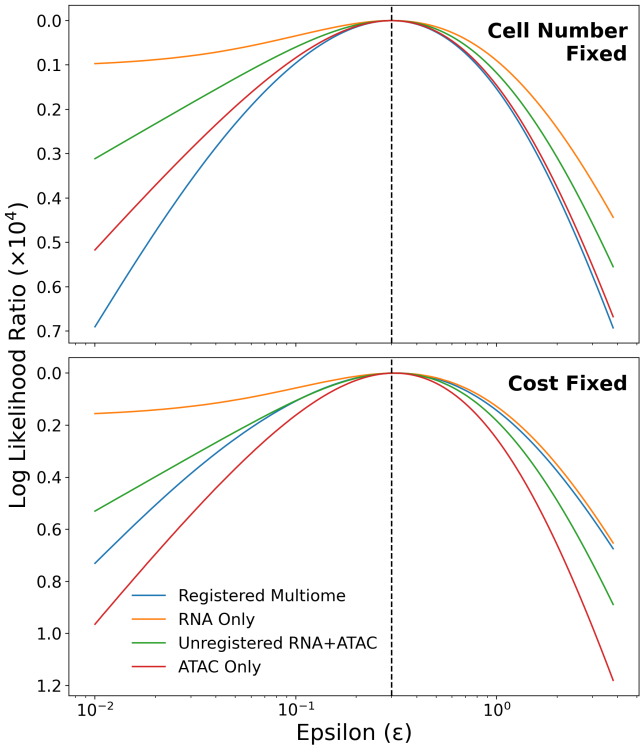


Figure 2.2: Distinguishability of parameter ϵ under our correlative model using simulated cells at $\epsilon = 0.3$ (dotted black line). The relative log-likelihood of the simulated data at values of ϵ differing from the true value are plotted for four different emulated scenarios: scATAC-seq only, scRNA-seq only, registered multiome and unregistered scATAC-seq and scRNA-seq. These scenarios are compared keeping a fixed number of cells (**top**), or a fixed experimental cost (**bottom**).

ϵ in these kinds of models. scATAC-seq data alone would be the most cost-efficient method for determining ϵ.

2.5 Application to ATAC-seq Data

Finally, we apply our model to single-cell ATAC-seq data, and compare it to a model where adjacent chromatin loci are constrained to be uncorrelated.

Data Preparation

We use three 10x Genomics single-cell ATAC-seq datasets to perform our analysis:

- 10k Human Peripheral Blood Mononuclear Cells, ATAC v2, Chromium Controller, analyzed using Cell Ranger ATAC 2.1.0, (2020, September 9) (Genomics, n.d.[b])

- 8k Adult Mouse Cortex Cells, ATAC v2, Chromium Controller, analyzed using Cell Ranger ATAC 2.1.0, (2022, March 29) (Genomics, n.d.[d])
- 10k 1:1 Mixture of Human GM12878 and Mouse EL4 Cells, ATAC v2, Chromium Controller, analyzed using Cell Ranger ATAC 2.1.0, (2022, March 29) (Genomics, n.d.[a])

We pre-processed these datasets using the snATAK pipeline (Booeshaghi, Gao, and Pachter, 2023). For the mixed mouse-human dataset, we first separated the human and mice cells and removed doublets. The snATAK pipeline provides a standardized process with a few parameters specifying the structure of the input FASTQ files (see Appendix A.8), and includes peak calling and pseudo-alignment. The output is a cell-by-peak matrix indicating how many read fragments were aligned to each peak region in each cell. After filtering for high quality cells, we binarized the cell-by-peak matrix. This means that if a cell had at least one read aligning to a certain peak, that peak region was considered open for that cell.

Since our model considers regions which are adjacent in space, we then restricted our focus to those peaks which were within 1.5kbp of their adjacent peak. This distance was chosen somewhat arbitrarily to admit a reasonable number of allowed groups of peaks. In particular, we examined groups of six consecutive ATAC peaks, where each was at most 1.5kbp from the next. Here and in what follows, we refer to these groups of six contiguous peak regions as 'loci'.

The distribution at each locus was then calculated by counting the frequency of different configurations across cells. The possible configurations are given by the length-six binary strings, e.g. 110000, which would represent two consecutive open regions followed by four consecutive closed regions.

Exploratory Analysis for Ising-like Model

After selecting suitable genetic loci for analysis, we investigated whether positive correlations are observed between neighboring chromatin sites. We calculated the Pearson correlation coefficient, r_{xy}, between sites in each pair of adjacent ATAC-seq peak regions in the selected loci. (See Appendix A.9 for the details of this calculation.) In Figure 2.3, we plot these correlation coefficients for each pair of adjacent sites across the three datasets. This exploratory analysis lends support to the use of an Ising-like model, since a consistently positive correlation coefficient between adjacent sites is indeed observed. The proportions of pairs with positive

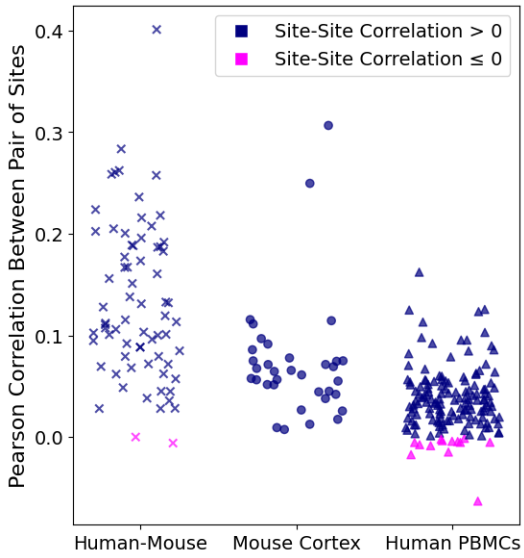


Figure 2.3: Pearson correlation coefficients between sites in each pair, for each of the three datasets (x positions within each column are random). Each point represents a pair of adjacent ATAC-seq peak regions included in our selected loci. All pairs with a positive Pearson correlation are colored in dark blue, and those with a negative or zero correlation in magenta. The number of pairs with positive correlation is 68/70, 35/35, and 137/150, for the human-mouse mixture, adult mouse cortex, and PBMC datasets respectively. The preponderance of positive correlations lends support to an Ising-like model.

Pearson correlation are 137/150, 35/35, and 68/70 for the PBMC, adult mouse cortex and human-mouse mixture datasets respectively.

ATAC-seq Noise Treatment

For data analysis, we added noise to the result of the above CME dynamics in the following way. We start from the steady-state gene-state distribution $\boldsymbol{\pi}$ defined in Equation 2.10. Then, with the addition of technical noise, we reach a final steady-state distribution, $\tilde{\boldsymbol{\pi}}$, arrived at via:

$$\mathcal{N}\boldsymbol{\pi} = \tilde{\boldsymbol{\pi}}, \tag{2.26}$$

where we have defined a noise matrix, $\mathcal{N}$. We consider the properties of this noise matrix for dropout noise which symmetrically affects all sites in a locus, giving them a probability $p_{\mathrm{d}rop}$ of being lost. Although this technical noise model is, in a sense, symmetrical, there is an asymmetry in the space of binary strings, since transition

from a state with fewer on-sites to a state with more on-sites is impossible. In this sense, the possible states of a locus form a partially ordered set. For instance,

$$010100 \prec 010101, \text{ yet } 010100 \not\prec 011001, \tag{2.27}$$

where the partial order, $a \prec b$, indicates that sequence a can be obtained, starting from b, via technical noise. In this example, the first two strings are connected by a transition encoded in $\mathcal{N}$, but the second two are not: the technical noise process cannot create false positives.

In the case of a three-site locus the matrix $\mathcal{N}$ is eight-dimensional. Using the basis defined by the state matrix, S:

$$S = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \tag{2.28}$$

we would have a noise matrix $\mathcal{N}$ given by:

$$\begin{pmatrix} 1 & p & p & p & p^2 & p^2 & p^2 & p^3 \\ 0 & (1-p) & 0 & 0 & p(1-p) & p(1-p) & 0 & p^2(1-p) \\ 0 & 0 & (1-p) & 0 & p(1-p) & 0 & p(1-p) & p^2(1-p) \\ 0 & 0 & 0 & (1-p) & 0 & p(1-p) & p(1-p) & p^2(1-p) \\ 0 & 0 & 0 & 0 & (1-p)^2 & 0 & 0 & p(1-p)^2 \\ 0 & 0 & 0 & 0 & 0 & (1-p)^2 & 0 & p(1-p)^2 \\ 0 & 0 & 0 & 0 & 0 & 0 & (1-p)^2 & p(1-p)^2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & (1-p)^3 \end{pmatrix}, \tag{2.29}$$

where, here, $\mathrm{p} = \mathrm{p}_{\mathrm{d}rop}$, and we have assumed that dropout at each site is independent. This is a triangular matrix. See Appendix A.6 for a visualization of the effect of this type of noise.

Model Inference

To assess the Ising model for chromatin, we compare a model fixing ϵ to be one, and a model allowing ϵ to vary as a parameter. We compare the models using the Bayesian Information Criterion (BIC), which quantifies goodness-of-fit whilst penalizing extra parameters (see the exact definition in Equation A.35).

Note that, from above, the probabilities for a chromatin configuration $\boldsymbol{\sigma}$ in this model, before technical dropout, are given by:

$$P(\boldsymbol{\sigma}) \propto \epsilon^{n_{mis}} \prod_{i=1}^{6} \left(\frac{k_{on,i}}{k_{off}} \right)^{\sigma_i}, \tag{2.30}$$

where n_{mis} is the number of pairs of misaligned neighbors, and again $\sigma_i = 0, 1$ for site i closed/open. For example,

$$P(110001) \propto \frac{1}{k_{off}^3} \epsilon^2 k_{on,1} k_{on,2} k_{on,6}. \tag{2.31}$$

We then add technical noise in the form of binomial dropout, where each site independently has probability p_{drop} of flipping from $\sigma_i = 1$ to $\sigma_i = 0$. However, in the $\epsilon = 1$ case, when fitting based on ATAC-seq data alone, and assuming the system is at steady state, independent binomial dropout is equivalent to a change in k_{on} values. This is because we can express the probability of a site being accessible in terms of its biological $k_{on,i}$ and k_{off} rates, and multiply this value by $(1 - p_{drop})$ to account for technical dropout. The resulting probability of measuring accessibility at a site i, given dropout, is then equivalent to the probability without dropout, given the substitution $k_{on,i} \rightarrow k'_{on,i}$, where:

$$\frac{k_{on,i}}{k_{on,i} + k_{off}} (1 - p_{drop}) = \frac{k'_{on,i}}{k'_{on,i} + k_{off}}. \tag{2.32}$$

Therefore, after constraining $\epsilon = 1$, the reduced model can be fully described using only six parameters ($k_{on,i}$ for $i = [1, 6]$).

So we are left with two models. The simple six-parameter model has independent sites, ignoring gene-gene correlations, and the effect of loss of reads is absorbed into the 'field strength' at each site. The eight-parameter model includes site-site correlations parameterized by ϵ, and the probabilities from Equation 2.30 are adjusted by adding binomial loss at each site (see Appendix A.6). This model does

capture correlations between genes, and is reminiscent of the Ising model from physics.

Results

We apply this inference procedure to the six-site loci identified in each dataset (see Section 2.5). We fit both the six-parameter and eight-parameter models at each locus, using stochastic global optimization to minimize log-likelihood, and repeated each fit ten times to check convergence.

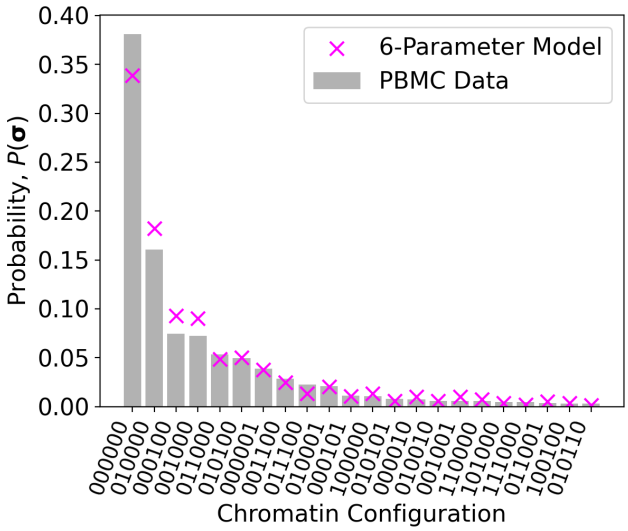
We show example fits for the six and eight-parameter models in Figure 2.4, for one six-site locus. The locus is from chromosome 18 of the 10k Human PBMCs, ATAC v2, Chromium Controller dataset.

The bar chart shows the observed fraction of the most frequently observed chromatin configurations at this locus. The magenta and blue markers show the analytic distribution using the best-fit parameters for the six/eight-parameter models respectively. See Appendix A.7 for fits at other loci. There we also verified the reliability of the result of the optimization algorithm with an MCMC fit for one locus.

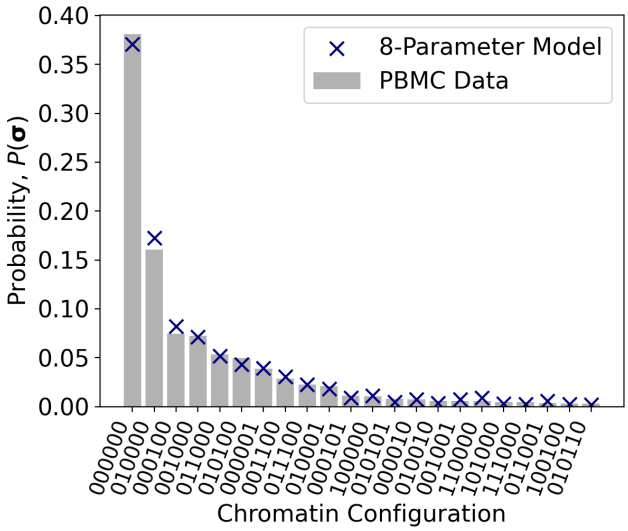
To analyze the effectiveness of the Ising hypothesis, we looked at all of the contiguous six-site loci, as defined above, across the three different 10x datasets. We calculated the BIC score for the six and eight-parameter models at each locus, and record the BIC score difference in favor of the extra parameter ϵ (combined with $p_{\mathrm{d}rop}$). We show the distribution of BIC score differences for each dataset in Figure 2.5. Almost all of the loci provide support for using the eight-parameter Ising-like model over a model with independent sites. The proportions of loci with a BIC score favoring the Ising-like model are 26/30, 7/7, and 14/14 for the PBMC, mouse cortex, and human-mouse mixture datasets respectively. The particularly striking positive outlier in the PBMC plot is locus 10:69043845-69054726. The high BIC difference is due to the much better fit of the eight-parameter model for this locus, as shown in Appendix A.10.

2.6 Discussion

In this work we have presented for the first time a minimal biophysical model suitable for joint analysis of RNA-seq and ATAC-seq data. These two modalities are increasingly assayed in single-cell genomics experiments, and our work explores and compares two models which, although relatively simplistic, offer a basic framework for quantitative analysis of multiomic data. We expect that these models will



(a)



(b)

Figure 2.4: Six-parameter (**top**) and eight-parameter (**bottom**) fits to observed distribution at locus chrom 18:77059667-77072738 from human PBMCs. N.B. The 22 most frequently observed strings of the total 64 are shown on the x-axis. The gray bars show the empirical distribution, and the magenta and blue markers the fitted distribution for the six and eight-parameter models respectively.

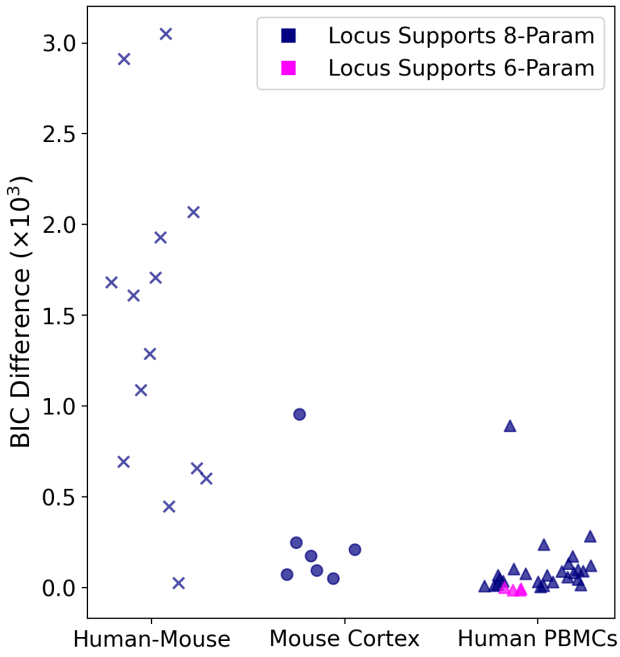


Figure 2.5: BIC score improvement with inclusion of correlations, i.e. six-parameter BIC score minus eight-parameter for each chromatin locus across three datasets. Positive values are plotted in blue, and represent loci at which the BIC score gives support for the full eight-parameter model, whilst negative values are plotted in magenta.

be elaborated on and improved, for example by the inclusion of transcription factor binding. We have already shown that the models help sharpen quantitative questions about gene expression and its relationship to chromatin dynamics, by exploring the ratio of correlations at the chromatin and transcriptome levels.

The framework we have outlined allows us to investigate the effects of DNA-level mechanistic assumptions on the expected distributions of both chromatin configuration and transcriptomics. In this case, we have investigated the effect of nearest-neighbor correlations on the system dynamics. To undertake our analysis, we have made use of a form of the Ising model from physics, which has been studied at length. Inspired by its use in other areas of computational biology, we have successfully applied the Ising model in the novel context of ATAC-seq data.

After hypothesizing an Ising model for chromatin dynamics, we followed Gorin et al. (Gorin, Vastola, and Pachter, 2023), and encoded these dynamics using chemical master equations. Making simple assumptions about the corresponding RNA dynamics, we were able to formulate the behavior of both modalities in a

tractable and biophysically meaningful way. The first and second order moments of the system were simple to derive from the resulting matrix equations, and allowed us to explore the relationship between correlations at the DNA and at the transcript level. In particular, the mathematical tractability of this class of models allowed us to uncover the unintuitive result that downstream correlations between transcripts of different species can theoretically be higher than the correlations of their parent genes at the level of chromatin configuration.

We then used our model to explore experimental design considerations for combining scATAC and scRNA data. In the context of a two-gene system, we have illustrated the ambiguity around the relative merits of multiomic vs unregistered data at fixed cost for parameter identifiability. This brings into question the rationale for choosing multiome assays over unregistered data, depending on the question being asked, and also highlights the importance of increased data quality for refining biophysical models. The results are also resilient to small variations in pricing, and depend only on the insight that multiomic RNA-seq + ATAC-seq assays intrinsically sample on the order of half as many cells as single-modality measurements. Although this numerical experiment has important limitations - for instance, it omits the impact of dropout noise - it provides a principled foundation for multiomic experiment design. As we obtain a better understanding of technical factors, we can extend this design approach to include noise considerations.

Finally, we were able to directly compare the steady-state distributions derived from the CME formulation to real scATAC-seq datasets, using a binomial dropout model for scATAC-seq technical noise. As a proof of concept, we limited our analysis to loci with six contiguous, close ATAC peak sites. We compared a simple six-parameter model of site-independence, with our eight-parameter Ising-like model plus dropout, and found that our model was almost always preferable in terms of the Bayesian Information Criterion. Our analysis highlights the importance of site-site accessibility correlations, despite the sparsity of ATAC-seq data, and informs our understanding of relations between neighboring DNA regions. Our work does not shed light on the biological details underlying these correlations, but the assumptions implicit in our models point towards experiments that may provide more insight into mechanistic underpinnings. For example, loci with BIC scores that highly favor the model including gene correlations could be candidates for more targeted chromatin remodeling or transcription factor binding experiments. Our approach could also be extended by relaxing the criteria we have used for selecting DNA loci, and assessing

nearest-neighbor correlations in a more genome-wide manner.

The techniques that we've introduced in this paper could be usefully extended to other data types, such as Hi-C data. Due to the allowed flexibility in specifying the transition matrix, the one-dimensional Ising model discussed in this work could be generalized to an Ising model with arbitrary underlying graph structure. Coupling of chromatin sites which are not linearly adjacent could be introduced via extra factors of the correlation parameter ϵ. In addition, since our framework couples RNA and chromatin dynamics in a simple and extendable way, our model can be incorporated directly into current biophysical modeling tools such as biVI (Carilli et al., 2023). By straightforward modification of our choices for chromatin and RNA dynamics (e.g. constitutive transcription), our approach can be used to explore many more complex systems in a tractable way. Our postulation of an Ising-like structure for the chromatin-state transition matrix could also be extended to analysis of other multiomic assays. Following the prescription outlined in Gorin et al. (Gorin, Vastola, and Pachter, 2023), we could further test the assumptions of this model using single-cell RNA-seq data.

Overall, although sparsity of single-cell ATAC-seq data limits the power of the conclusions reached, we have outlined a promising approach to biophysical modeling of the data type, including its use in conjunction with single-cell RNA-seq data.

2.7 Data Availability

Scripts implementing these analyses and simulations and reproducing fig:distingfig:BIC-scatter are available at `https://github.com/pachterlab/FGP_2023`.

References

Amarasinghe, Harindra E. et al. (Oct. 2023). "Mapping the epigenomic landscape of human monocytes following innate immune activation reveals context-specific mechanisms driving endotoxin tolerance". In: *BMC Genomics* 24.1, p. 595. ISSN: 1471-2164. DOI: 10.1186/s12864-023-09663-0.

Booeshaghi, A. Sina, Fan Gao, and Lior Pachter (2023). "Assessing the multimodal tradeoff". In: *bioRxiv*. DOI: 10.1101/2021.12.08.471788. URL: https://www.biorxiv.org/content/early/2023/04/18/2021.12.08.471788.

Bravo González-Blas, Carmen et al. (May 2019). "cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data". In: *Nature Methods* 16.5, pp. 397–400. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-019-0367-1.

Buenrostro, Jason D et al. (Oct. 2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In: *Nature Methods* 10.12, pp. 1213–1218. DOI: 10.1038/nmeth.2688. URL: https://doi.org/10.1038/nmeth.2688.

Buenrostro, Jason D. et al. (June 2015). "Single-cell chromatin accessibility reveals principles of regulatory variation". In: *Nature* 523.7561, pp. 486–490. DOI: 10.1038/nature14590. URL: https://doi.org/10.1038/nature14590.

Cao, Junyue et al. (Sept. 2018). "Joint profiling of chromatin accessibility and gene expression in thousands of single cells". In: *Science* 361.6409, pp. 1380–1385. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aau0730.

Carilli, Maria et al. (Jan. 2023). "Biophysical modeling with variational autoencoders for bimodal, single-cell RNA sequencing data". In: *Nature Methods*. DOI: 10.1101/2023.01.13.523995. URL: http://biorxiv.org/lookup/doi/10.1101/2023.01.13.523995.

Chen, Huidong et al. (Dec. 2019). "Assessment of computational methods for the analysis of single-cell ATAC-seq data". In: *Genome Biology* 20.1, p. 241. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1854-5.

Chen, Xin et al. (Nov. 2022). "MSR1 characterized by chromatin accessibility mediates M2 macrophage polarization to promote gastric cancer progression". In: *International Immunopharmacology* 112, p. 109217. ISSN: 15675769. DOI: 10.1016/j.intimp.2022.109217.

Cusanovich, Darren A. et al. (May 2015). "Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing". In: *Science* 348.6237, pp. 910–914. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aab1601.

Dhara, S. et al. (May 2021). "Pancreatic cancer prognosis is predicted by an ATAC-array technology for assessing chromatin accessibility". In: *Nature Communications* 12.1, p. 3044. ISSN: 2041-1723. DOI: 10.1038/s41467-021-23237-2.

Dillinger, Stefan (2021). *Complete Guide to Understanding and Using ATAC-Seq — activemotif.com*. https://www.activemotif.com/blog-atac-seq. [Accessed 20-11-2023].

Ding, Min et al. (May 2023). "Integration of ATAC-Seq and RNA-Seq reveals FOSL2 drives human liver progenitor-like cell aging by regulating inflammatory factors". In: *BMC Genomics* 24.1, p. 260. ISSN: 1471-2164. DOI: 10.1186/s12864-023-09349-7.

Duren, Zhana et al. (July 2018). "Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations". In: *Proc Natl Acad Sci U S A* 115.30, pp. 7723–7728.

Felce, C. (2023). *Created in BioRender*. BioRender.com/r42x127.

Gayoso, Adam et al. (Mar. 2021). "Joint probabilistic modeling of single-cell multiomic data with totalVI". In: *Nature Methods* 18.3, pp. 272–282. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-020-01050-x.

Genomics, 10x (n.d.[a]). *10k 1:1 Mixture of Human GM12878 and Mouse EL4 Cells, ATAC v2, Chromium Controller, Single Cell ATAC dataset analyzed using Cell Ranger ATAC 2.1.0*. Version cellranger-atac-2.1.0. (2022, March 29). URL: https://www.10xgenomics.com/datasets/10k-1-1-mixture-of-human-gm12878-and-mouse-el4-cells-atac-v2-chromium-controller-2-standard.

– (n.d.[b]). *10k Human PBMCs, ATAC v2, Chromium Controller, Single Cell ATAC dataset analyzed using Cell Ranger ATAC 2.1.0*. Version Cell Ranger ATAC v2.1.0. (2020, September 9). URL: https://www.10xgenomics.com/datasets/10k-human-pbmcs-atac-v2-chromium-controller-2-standard.

– (n.d.[c]). *10x Genomics Store*. https://www.10xgenomics.com/store. Accessed: 2024-10-21.

– (n.d.[d]). *8k Adult Mouse Cortex Cells, ATAC v2, Chromium Controller, Single Cell ATAC dataset analyzed using Cell Ranger ATAC 2.1.0*. Version cellranger-atac-2.1.0. (2022, March 29). URL: https://www.10xgenomics.com/datasets/8k-adult-mouse-cortex-cells-atac-v2-chromium-controller-2-standard.

Gligorijević, Vladimir and Nataša Pržulj (Nov. 2015). "Methods for biological data integration: perspectives and challenges". In: *Journal of The Royal Society Interface* 12.112, p. 20150571. ISSN: 1742-5689, 1742-5662. DOI: 10.1098/rsif.2015.0571.

Gontarz, Paul et al. (June 2020). "Comparison of differential accessibility analysis strategies for ATAC-seq data". In: *Scientific Reports* 10.1, p. 10150. ISSN: 2045-2322. DOI: 10.1038/s41598-020-66998-4.

Gorin, Gennady, John J. Vastola, and Lior Pachter (Oct. 2023). "Studying stochastic systems biology of the cell with single-cell genomics data". In: *Cell Systems* 14.10, 822–843.e22. ISSN: 24054712. DOI: 10.1016/j.cels.2023.08.004.

Hao, Yuhan et al. (June 2021). "Integrated analysis of multimodal single-cell data". In: *Cell* 184.13, 3573–3587.e29. ISSN: 00928674. DOI: 10.1016/j.cell.2021.04.048.

Klemm, Sandy L., Zohar Shipony, and William J. Greenleaf (Jan. 2019). "Chromatin accessibility and the regulatory epigenome". In: *Nature Reviews Genetics* 20.4, pp. 207–220. DOI: 10.1038/s41576-018-0089-8. URL: https://doi.org/10.1038/s41576-018-0089-8.

Lee, Michelle Y. Y., Klaus H. Kaestner, and Mingyao Li (Oct. 2023). "Benchmarking algorithms for joint integration of unpaired and paired single-cell RNA-seq and ATAC-seq data". In: *Genome Biology* 24.1, p. 244. ISSN: 1474-760X. DOI: 10.1186/s13059-023-03073-x. URL: https://doi.org/10.1186/s13059-023-03073-x.

Li, Jinlu et al. (Apr. 2022). "Integrative Single-Cell RNA-seq and ATAC-seq Analysis of Mesenchymal Stem/Stromal Cells Derived from Human Placenta". In: *Frontiers in Cell and Developmental Biology* 10, p. 836887. ISSN: 2296-634X. DOI: 10.3389/fcell.2022.836887.

Ma, Sai et al. (Nov. 2020). "Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin". In: *Cell* 183.4, 1103–1116.e20. ISSN: 00928674. DOI: 10.1016/j.cell.2020.09.056.

Mo, Qianxing and Faming Liang (Jan. 2010a). "A hidden Ising model for ChIP-chip data analysis". In: *Bioinformatics* 26.6, pp. 777–783.

– (Jan. 2010b). *Bayesian Modeling of ChIP-chip Data Through a High-Order Ising Model*. DOI: 10.1111/j.1541-0420.2009.01379.x. URL: http://dx.doi.org/10.1111/j.1541-0420.2009.01379.x.

Nair, Venugopalan D. et al. (2021). "Differential analysis of chromatin accessibility and gene expression profiles identifies cis-regulatory elements in rat adipose and muscle". In: *Genomics* 113.6, pp. 3827–3841. ISSN: 0888-7543. DOI: https://doi.org/10.1016/j.ygeno.2021.09.013.

Nazzari, Marta et al. (Sept. 2023). "Investigation of the effects of phthalates on in vitro thyroid models with RNA-Seq and ATAC-Seq". In: *Frontiers in Endocrinology* 14, p. 1200211. ISSN: 1664-2392. DOI: 10.3389/fendo.2023.1200211.

Peccoud, J. and B. Ycart (Oct. 1995). "Markovian Modeling of Gene-Product Synthesis". In: *Theoretical Population Biology* 48.2, pp. 222–234. ISSN: 00405809. DOI: 10.1006/tpbi.1995.1027.

Raevskiy, Mikhail et al. (Mar. 2023). "Epi-Impute: Single-Cell RNA-seq Imputation via Integration with Single-Cell ATAC-seq". In: *International Journal of Molecular Sciences* 24.7, p. 6229. ISSN: 1422-0067. DOI: 10.3390/ijms24076229.

Sahinyan, Korin et al. (Feb. 2022). "Application of ATAC-Seq for genome-wide analysis of the chromatin state at single myofiber resolution". In: *eLife* 11. Ed. by YM Dennis Lo and Nora Yucel, e72792. ISSN: 2050-084X. DOI: 10.7554/eLife.72792.

Singh, Abhyudai and Pavol Bokes (Sept. 2012). "Consequences of mRNA transport on stochastic variability in protein levels". In: *Biophys J* 103.5, pp. 1087–1096.

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Wang, Jianfang et al. (Aug. 2022). "Integration of RNA-seq and ATAC-seq identifies muscle-regulated hub genes in cattle". In: *Frontiers in Veterinary Science* 9. ISSN: 2297-1769. DOI: 10.3389/fvets.2022.925590. URL: https://www.frontiersin.org/journals/veterinary-science/articles/10.3389/fvets.2022.925590/full.

Wang, Jin et al. (2018). "ATAC-Seq analysis reveals a widespread decrease of chromatin accessibility in age-related macular degeneration." In: *Nature communications* 9(1), 1364. DOI: https://doi.org/10.1038/s41467-018-03856-y.

Wang, Shicong et al. (Sept. 2022). "Integrating ATAC-seq and RNA-seq Reveals the Dynamics of Chromatin Accessibility and Gene Expression in Apple Response to Drought". In: *International Journal of Molecular Sciences* 23.19, p. 11191. ISSN: 1422-0067. DOI: 10.3390/ijms231911191.

Weber, Marc and Javier Buceta (June 2016). "The cellular Ising model: a framework for phase transitions in multicellular environments". In: *Journal of the Royal Society Interface* 13.119, p. 20151092. ISSN: 1742-5689. DOI: 10.1098/rsif.2015.1092.

Xu, Feifei et al. (2023). "Integration of ATAC-Seq and RNA-Seq identifies key genes and pathways involved in the neuroprotection of S-adenosylmethionine against perioperative neurocognitive disorder". In: *Computational and Structural Biotechnology Journal* 21, pp. 1942–1954. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2023.03.001.

Xu, Zhong et al. (Oct. 2022). "Integration of ATAC-seq and RNA-seq analysis identifies key genes affecting intramuscular fat content in pigs". In: *Frontiers in Nutrition* 9, p. 1016956. ISSN: 2296-861X. DOI: 10.3389/fnut.2022.1016956.

Yu, Dongdong et al. (Apr. 2023). "Co-profiling reveals distinct patterns of genomic chromatin accessibility and gene expression in pulmonary hypertension caused by chronic hypoxia". In: *Respiratory Research* 24.1, p. 104. ISSN: 1465-993X. DOI: 10.1186/s12931-023-02389-3.

Chapter 3

INTEGRATING PROTEIN COUNTS INTO SINGLE-CELL RNA-SEQ ANALYSIS

Felce, Catherine, Meichen Fang, and Lior Pachter (2025). "Joint Biophysical Modeling of Paired Single-Cell RNA and Protein Measurements". In: *bioRxiv*. DOI: 10.1101/2025.11.14.688548. eprint: 2025.11.14.688548. URL: https://doi.org/10.1101/2025.11.14.688548.

3.1 Abstract

Surface protein measurements can supplement gene expression information from single-cell RNA sequencing to provide a more complete assessment of cell identity and function. Recently developed multiomic assays facilitate such measurements, and can, in principle, be utilized to understand the dynamics of transcription and translation. We develop a framework for biophysical modeling of transcription jointly with translation from single-cell data, along with a suitable technical noise model for sequencing data. We demonstrate its efficacy using simulations, and illustrate how it can be useful in practice with 10x multiomic data. Our proof-of-principle highlights the potential for jointly modeling transcription and translation as data quality and measurement accuracy improves.

3.2 Background

Although messenger RNA has historically been used as a proxy for protein expression (Junker et al., 2014; Fu et al., 2007), many studies have highlighted the role of post-transcriptional regulation in decoupling RNA and protein levels (Wei et al., 2015; Franks, Airoldi, and Slavov, 2017), especially in embryogenesis and other developmental processes (Kuersten and Goodwin, 2003). The poor correlation between mRNA and protein counts across the genome has been extensively investigated (Maier, Güell, and Serrano, 2009), with biologists calling for principled modeling (McManus, Zhe Cheng, and Vogel, 2015; Greenbaum et al., 2003), including technical noise models (Yansheng Liu, Beyer, and Aebersold, 2016), to shed light on these discrepancies. While differentially expressed RNA has been shown to correlate more strongly with its associated proteins than non-differentially expressed RNA (Koussounadis et al., 2015), the range of correlations for individual

genes is still broad ($r = -0.95$ to 0.94) and differentially expressed proteins have been shown not to reliably correlate with differentially expressed RNA (Huber et al., 2004; Ideker, 2001).

Given that combined RNA expression and proteomic data are helpful for analyzing the immune response (Wu et al., 2020) and bacterial growth (Nobori et al., 2020), there is a clear benefit to investigating the separate regulatory contributions of transcription and translation using multiomic data. Recognizing this, statistical tools have been created to disentangle RNA and protein-level regulation using time-course data (Teo et al., 2014), and studies have confirmed that RNA and protein-level regulation are both important, with some genes being regulated in opposite directions simultaneously (Zhandong Cheng et al., 2016). Cross-species studies have also highlighted the potential decoupling of mRNA and protein levels, and have tried to unravel the effects of transcriptional, translational, and post-translational mechanisms on expression divergence (J. Wang et al., 2018).

However, time-course or cross-species measurements are difficult to obtain, and increasingly ubiquitous single-cell data offer the opportunity, in principle, to study the dynamics of transcription and translation in a complementary way. We propose a biophysical model based on single-cell multiomic data that can simultaneously describe transcriptional and translational mechanisms and untangle their contributions to controlling protein expression.

Single-cell resolution RNA measurements have already enabled biophysical modeling of nascent and mature RNA and the inference of transcriptional rate parameters (Gorin, Vastola, and L. Pachter, 2023). In that setting, the modeling relies on multimodal datasets, specifically spliced and unspliced counts measured from mature and nascent RNAs (Gorin, Vastola, Fang, et al., 2022). The inclusion of chromatin information from ATAC-seq has also been considered (Felce, Gorin, and L. S. Pachter, 2024). In this work we extend these approaches to include protein count data.

High quality single-cell protein quantification, including isoform differentiation, has become available with single-cell western blotting (Hughes et al., 2014). Since 2017, with the advent of high-throughput simultaneous proteomic and transcriptomic measurements in single-cells with REAP-seq (Peterson et al., 2017) and CITE-seq (Stoeckius et al., 2017), principled biophysical modeling of the joint modalities is now within reach.

Although models combining RNA dynamics with constitutive protein translation of

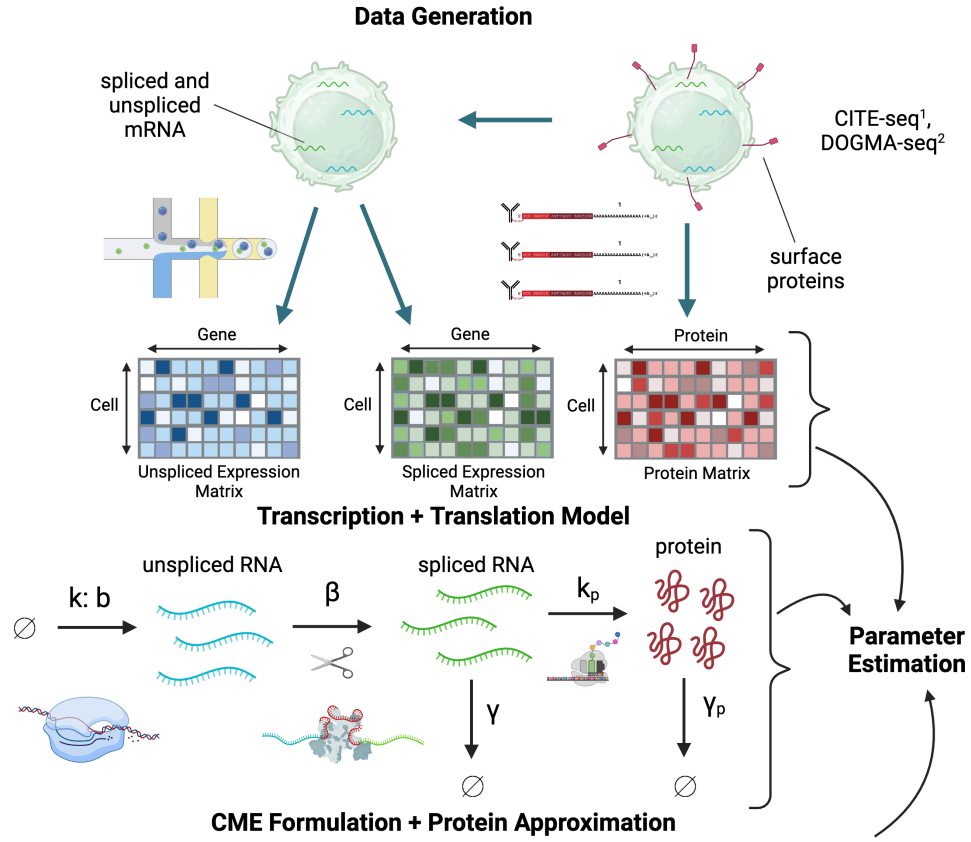


Figure 3.1: **Fitting a joint biophysical model for RNA and protein**. Single-cell transcriptomics and surface protein expression can be simultaneously measured using methods such as [1]CITE-seq (Stoeckius et al., 2017) and [2]DOGMA-seq (Mimitou et al., 2021). Transcriptomic data can be used to generate unspliced and spliced count matrices. We assume a bursty transcription and constitutive translations with constitutive splicing and degradation. We derive the chemical master equation (CME) following Singh et al. (Singh and Bokes, 2012) and apply continuous approximation to protein when necessary. The resultant joint probability distribution of unspliced mRNA, spliced mRNA, and protein is solved numerically and used for parameter estimation.

mature RNA transcripts have been suggested (Singh and Bokes, 2012), we solve numerically for the probability distribution for such a system and demonstrate that this can be used to reliably infer biophysically meaningful parameters on real datasets. Our approach, summarized in Figure 3.1, provides a proof of principle for fitting biophysical models to multiomic data.

3.3 Results

Model

Symbol	Meaning
$k \in \mathbb{R}_+$	Burst frequency
$B \in \mathbb{N}$	Burst size
$b \in \mathbb{R}$	Mean burst size
$\beta \in \mathbb{R}$	Splicing rate of unspliced transcripts
$\gamma \in \mathbb{R}$	Decay rate of spliced transcripts
$k_p \in \mathbb{R}$	Translation rate of spliced transcripts
$\gamma_p \in \mathbb{R}$	Decay rate of proteins
$X_u \in \mathbb{N}$	Unspliced RNA copy number
$X_s \in \mathbb{N}$	Spliced RNA copy number
$X_p \in \mathbb{N}$	Protein copy number
$P(x_u, x_s, x_p, t)$	Probability density of state $(x_u, x_s, x_p) \in \mathbb{N}^3$ at time t
$G(z_u, z_s, z_p, t)$	Generating function (GF) of $P(x_u, x_s, x_p, t)$
$\phi(u_u, u_s, u_p, t)$	Factorial-cumulant GF $\log G(u_u + 1, u_s + 1, u_p + 1, t)$
$F(z_u) = \frac{1}{b+1-bz_u}$	Generating function of B, i.e., $\sum_{n=0}^{\infty} z_u^n P(B = n)$
$M(u_u) = \frac{bu_u}{1-bu_u}$	Transformation of the burst PGF, i.e., $M(u_u) = F(1 + u_u) - 1$

Table 3.1: Notation for the joint biophysical model, and expressions used for numerical solving of the steady-state probability distribution over unspliced, spliced, and protein counts.

We consider the bursting limit of the telegraph model as in Singh and Bokes (Singh and Bokes, 2012), with RNA processing from nascent to mature transcripts at a rate β, and translation rate per RNA molecule k_p (see Table 3.1 for all parameter identifications). This system can be summarized in the following reactions.

$$\emptyset \xrightarrow{k} B \times U \quad U \xrightarrow{\beta} S \quad S \xrightarrow{\gamma} \emptyset \quad S \xrightarrow{k_p} S + P \quad P \xrightarrow{\gamma_p} \emptyset \tag{3.1}$$

As in Singh et al. (Singh and Bokes, 2012), we define the probability mass function $P(x_u, x_s, x_p, t)$, which evolves in this system according to:

$$\begin{aligned}\frac{dP}{dt} = k\left(\sum_{n=0}^{x_u} P(B=n)P(x_u-n,x_s,x_p,t) - P(x_u,x_s,x_p,t)\right)\\ + \beta\left((x_u+1)P(x_u+1,x_s-1,x_p,t) - x_u P(x_u,x_s,x_p,t)\right)\\ + k_p\left(x_s P(x_u,x_s,x_p-1,t) - x_s P(x_u,x_s,x_p,t)\right)\\ + \gamma\left((x_s+1)P(x_u,x_s+1,x_p,t) - x_s P(x_u,x_s,x_p,t)\right)\\ + \gamma_p\left((x_p+1)P(x_u,x_s,x_p+1,t) - x_p P(x_u,x_s,x_p,t)\right). \qquad (3.2)\end{aligned}$$

We then use generating function methods to calculate the stationary distribution. As in (Gorin, Vastola, Fang, et al., 2022; Gans, 1960), we define the probability generating function, G, via

$$G(\boldsymbol{z},t) \equiv \sum_{x_u=0}^{\infty}\sum_{x_s=0}^{\infty}\sum_{x_p=0}^{\infty} z_u^{x_u} z_s^{x_s} z_p^{x_p} P(x_u,x_s,x_p,t), \qquad (3.3)$$

where $\boldsymbol{z}$ represents the vector of arguments (z_u, z_s, z_p). We then define ϕ via

$$\phi(u_u,u_s,u_p,t) \equiv \log G(u_u+1,u_s+1,u_p+1,t). \qquad (3.4)$$

Using the method of characteristics (see Appendix B.1), we arrive at the following system of equations:

$$\frac{d\tilde{u}_s}{ds} = -\gamma\tilde{u}_s + k_p u_p e^{-\gamma_p s}(\tilde{u}_s+1) \qquad (3.5)$$

$$\frac{d\tilde{u}_u}{ds} = \beta(\tilde{u}_s - \tilde{u}_u) \qquad (3.6)$$

$$\phi(u_u,u_s,u_p,\infty) = \int_0^{\infty} \frac{kb\tilde{u}_u(s)}{1-b\tilde{u}_u(s)}ds \qquad (3.7)$$

$$\tilde{u}_s(0) = u_s \qquad \tilde{u}_u(0) = u_u, \qquad (3.8)$$

which can be solved numerically (see Appendix B.2).

On top of the biological model, which is summarized in Figure 3.1 **Transcription + Translation Model**, we also incorporate a technical noise model to account for noise in the sequencing process following (Gorin and L. Pachter, 2022), with both RNA and protein counts assumed to undergo Poissonian sampling, whose parameters are optimized via grid search.

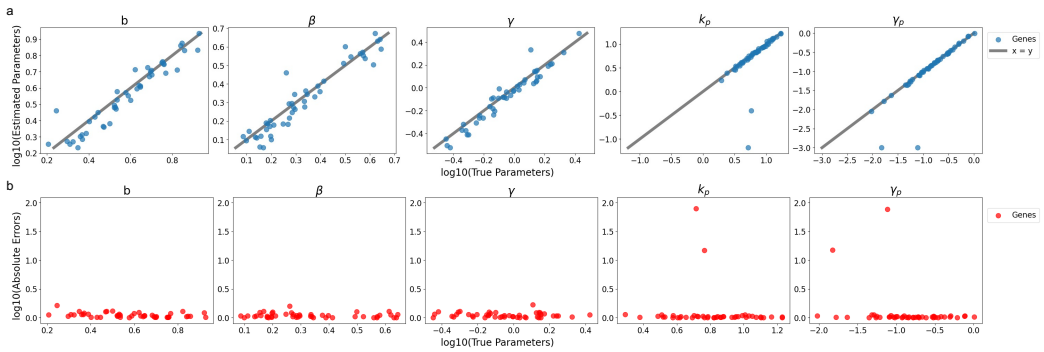


Figure 3.2: **Inference accuracy on simulations** a) Estimated parameters versus true parameters. b) Absolute errors versus true parameters.

Simulation

First, we use simulated data to assess the accuracy of estimated parameters under our model. We simulated 100 genes by first drawing their true parameters from a biologically plausible range, then generating steady-state data using the Gillespie algorithm. The application of our mean-expression filter resulted in a final set of 46 genes. We applied our inference framework to these 46 simulated genes and were able to recover all of the parameters in the full, joint model with unspliced, spliced and protein counts. The inference accuracy is shown in Figure 3.2. Parameter estimates were generally accurate, except for two outliers whose protein degradation rates converged to the optimization lower bound. The two outliers reside in a regime of low mRNA and high protein expression, where absolute parameter values are less identifiable, and only the ratio of translation to protein degradation was correctly inferred. The simulation results suggest that we should exclude parameter estimates found at the optimization bounds when fitting to real datasets.

Fits to Real Data

Given the simulation results, we first sought to apply the full model to the unspliced RNA, spliced RNA and protein counts from real datasets. However, we observed that unspliced counts are generally low and noisy, making it difficult to identify genes with high expression across all modalities. Therefore, we asked whether protein and spliced RNA counts are informative enough for inference. We refit the model using only the spliced mRNA and protein counts from the 46 simulated genes and found all parameters to be identifiable, though with a loss of accuracy in the splicing rate estimates (Figure 3.3). As we are more interested in the kinetic parameters related to translation, we decided to fit our model on the protein and spliced RNA counts

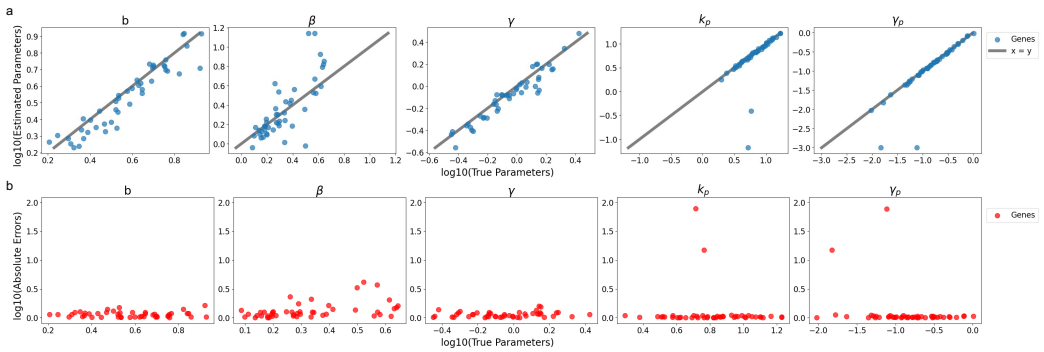


Figure 3.3: **Inference accuracy with only spliced mRNA and protein counts on simulations** a) Estimated parameters versus true parameters b) Absolute errors versus true parameters.

of real datasets for now. The full model can be used when data quality permits in the future.

In particular, we fit our model on the protein and spliced RNA counts from the following human PBMC datasets:

- 10k Human PBMCs Stained with TotalSeq™-B Human TBNK Cocktail, Chromium GEM-X Single Cell 3' Universal 3' Gene Expression dataset analyzed using Cell Ranger 8.0.0 (2024, March 13) (Genomics, 2024)
- 10K Human PBMCs, Gene Expression with a Panel of TotalSeq™-B Antibodies, analyzed using Cell Ranger 3.0.0, (2018, November 19) (Genomics, 2018)

We processed the raw transcript reads using kb-python (kallisto and bustools) to obtain spliced and unspliced RNA counts. We used the cell-by-protein count matrix provided by 10x Genomics. For protein complexes with subunits, we assumed that the presence of the complex indicates the presence of all constituent subunits, so we chose from among the subunits when mapping to RNA transcripts. More details on data processing can be found in Appendix B.3.

In Figure 3.4, we show the fits of our model to a few marker genes. The technical sampling parameters to which these fits correspond, and the optimal fitted biophysical parameters, are given in Appendix (B.3). The cells were subsetted to monocytes and T cells for the 2024 and 2018 datasets respectively.

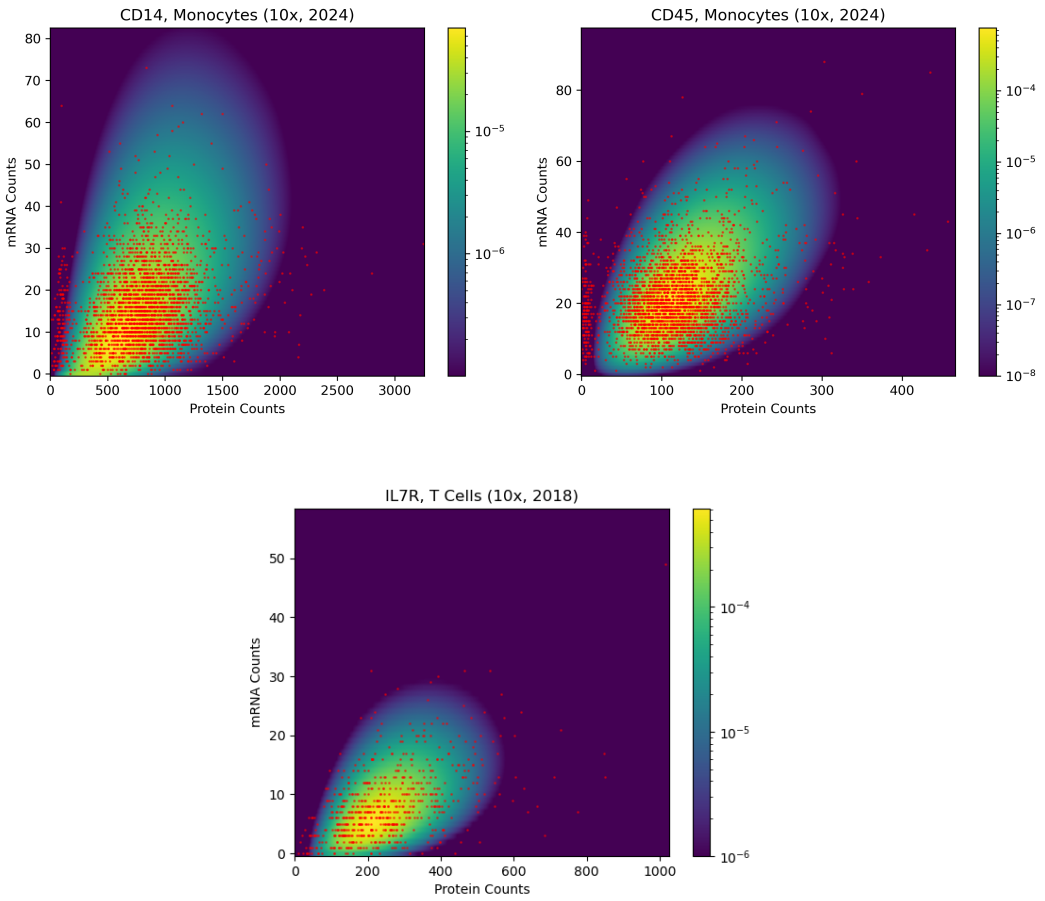


Figure 3.4: Fits to genes. The color map shows the calculated probability density at the optimal biological parameters, on a logarithmic scale. The red scatter points are the observed distribution of mRNA vs protein counts over cells. The cells were subsetted to monocytes and T cells for the 10x 2024 (Genomics, 2024) and 2018 (Genomics, 2018) datasets respectively.

3.4 Discussion

In this work we have presented and explored the power of fitting a joint biophysical model for RNA and protein to single-cell data, and extracted biophysically meaningful parameters. We have demonstrated the robustness of our fitting procedure for the full joint biophysical model via simulation. We have then shown that most of the biophysical parameters can also be inferred to some accuracy using only spliced and protein counts. This approach could be appropriate until the available data includes sufficiently comprehensive unspliced RNA measurements. We have used this reduced model to fit transcriptional rate parameters to CITE-seq data, and have shown that we can successfully reproduce the observed spliced-protein distributions.

Our work highlights the compatibility of current inference frameworks (Gorin, Vastola, and L. Pachter, 2023) with biophysical models including additional modalities, and we hope that others will continue our approach, perhaps by including chromatin accessibility measurements as an additional model layer. The method described here could also be straightforwardly extended to include principled biophysical clustering using meK-Means (Chari, Gorin, and L. Pachter, 2023). One limitation of our analysis is the simplification that surface proteins are a proxy for proteins produced within the cell. A non-instantaneous process of surface protein export would add another layer to the model, and the inclusion of protein export is a potential extension of this investigation. Additionally, length-dependent translation (Rogers et al., 2017) could be added to build on the picture provided here.

Overall, the approach outlined here represents a first foray into using biophysical modeling to disentangle transcriptional and translational parameters. With development, this method could provide a convenient avenue for determining metabolic rates in systems where direct experimental measurements are difficult to obtain. The full model presented here also provides a framework for the principled integration of single-cell unspliced, spliced, and protein count information. The power of this methodology will increase with the improved reliability of these measurement techniques.

Data and Code Availability

A github repository with all of the datasets, scripts for fitting parameters, and the simulations presented here is available at: `https://github.com/pachterlab/FFP_2025`.

Acknowledgments

We thank Andrew LeDuc for helpful discussions. We also thank Charles Trimble for generously funding part of C.F.'s research.

References

Chari, Tushar, Gennady Gorin, and Lior Pachter (2023). "Biophysically Interpretable Inference of Cell Types from Multimodal Sequencing Data". In: *bioRxiv : the preprint server for biology*. 2023.09.17.558131. DOI: 10.1101/2023.09.17.558131. URL: https://doi.org/10.1101/2023.09.17.558131.

Cheng, Zhandong et al. (2016). "Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress". In: *Molecular Systems Biology* 12.1, p. 855. DOI: 10.15252/msb.20156423.

Felce, Catherine, Gennady Gorin, and Lior S. Pachter (Dec. 2024). "Biophysical model for joint analysis of chromatin and RNA sequencing data". In: *Physical Review E* 110.6, p. 064405. DOI: 10.1103/PhysRevE.110.064405. URL: https://doi.org/10.1103/PhysRevE.110.064405.

Franks, Alexander, Edoardo Airoldi, and Nikolai Slavov (May 2017). "Post-transcriptional regulation across human tissues". In: *PLoS Computational Biology* 13.5, e1005535. DOI: 10.1371/journal.pcbi.1005535.

Fu, N. et al. (2007). "Comparison of Protein and mRNA Expression Evolution in Humans and Chimpanzees". In: *PLoS ONE* 2.2, e216. DOI: 10.1371/journal.pone.0000216.

Gans, P. J. (1960). "Open first order Stochastic processes". In: *The Journal of Chemical Physics* 33, pp. 691–694. DOI: 10.1063/1.1731239.

Genomics, 10x (Nov. 2018). *10k PBMCs from a Healthy Donor - Gene Expression with a Panel of TotalSeq-B Antibodies Universal 3' Gene Expression*. 10x Genomics Datasets. Version Cell Ranger v3.0.0. Dataset analyzed using Cell Ranger v3.0.0. URL: https://www.10xgenomics.com/datasets/10-k-pbm-cs-from-a-healthy-donor-gene-expression-and-cell-surface-protein-3-standard-3-0-0.

– (Oct. 2024). *10k Human PBMCs Stained with TotalSeq-B Human TBNK Cocktail, Chromium GEM-X Single Cell 3' Gene Expression*. 10x Genomics Datasets. Version Cell Ranger v8.0.0. Dataset analyzed using Cell Ranger v8.0.0. URL: https://www.10xgenomics.com/datasets/10k-human-pbmcs-stained-with-totalseq-B-human-TBNK-cocktail-GEM-X.

Gorin, Gennady and Lior Pachter (2022). "Monod: mechanistic analysis of single-cell RNA sequencing count data". In: *bioRxivorg*, p. 2022.06.11.495771. DOI: 10.1101/2022.06.11.495771.

Gorin, Gennady, John J. Vastola, Mingyu Fang, et al. (2022). "Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments". In: *Nature Communications* 13, p. 7620. DOI: 10.1038/s41467-022-34857-7. URL: https://doi.org/10.1038/s41467-022-34857-7.

Gorin, Gennady, John J. Vastola, and Lior Pachter (Oct. 2023). "Studying stochastic systems biology of the cell with single-cell genomics data". In: *Cell Systems* 14.10, 822–843.e22. ISSN: 24054712. DOI: 10.1016/j.cels.2023.08.004.

Greenbaum, Dov et al. (Aug. 2003). "Comparing Protein Abundance and mRNA Expression Levels on a Genomic Scale". In: *Genome Biology* 4.9, p. 117. ISSN: 1474-760X. DOI: 10.1186/gb-2003-4-9-117.

Huber, Martin et al. (2004). "Comparison of proteomic and genomic analyses of the human breast cancer cell line T47D and the antiestrogen-resistant derivative T47D-r". In: *Molecular & Cellular Proteomics* 3, pp. 43–55.

Hughes, Alex J et al. (July 2014). “Single-Cell Western Blotting”. In: *Nature Methods* 11.7, pp. 749–755. ISSN: 1548-7105. DOI: 10.1038/nmeth.2992.

Ideker, Trey (2001). “Integrated genomic and proteomic analyses of a systematically perturbed metabolic network”. In: *Science* 292, pp. 929–934.

Junker, Jan Philipp et al. (Nov. 2014). “A Predictive Model of Bifunctional Transcription Factor Signaling during Embryonic Tissue Patterning”. In: *Developmental Cell* 31.4, pp. 448–460. ISSN: 1534-5807. DOI: 10.1016/j.devcel.2014.10.017. (Visited on 09/23/2025).

Koussounadis, Antonis et al. (2015). “Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system”. In: *Scientific Reports* 5, p. 10775. DOI: 10.1038/srep10775.

Kuersten, Sebastian and Elizabeth B. Goodwin (2003). “The power of the 3 UTR: translational control and development”. In: *Nature Reviews Genetics* 4, pp. 626–637. DOI: 10.1038/nrg1125.

Liu, Yansheng, Andreas Beyer, and Ruedi Aebersold (2016). “On the Dependency of Cellular Protein Levels on mRNA Abundance”. In: *Cell* 165.3, pp. 535–550. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.03.014.

Maier, Tobias, Marc Güell, and Luis Serrano (2009). “Correlation of mRNA and protein in complex biological samples”. In: *FEBS Letters* 583.24, pp. 3966–3973. DOI: 10.1016/j.febslet.2009.10.036.

McManus, Joel, Zhe Cheng, and Christine Vogel (2015). “Next-generation analysis of gene expression regulation – comparing the roles of synthesis and degradation”. In: *Molecular BioSystems*. DOI: 10.1039/C5MB00310E.

Mimitou, Eleni P et al. (2021). “Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells”. In: *Nat. Biotechnol.* 39, pp. 1246–1258. DOI: 10.1038/s41587-021-00927-2.

Nobori, Takanori et al. (2020). “Multidimensional gene regulatory landscape of a bacterial pathogen in plants”. In: *Nature Plants* 6, pp. 883–896. DOI: 10.1038/s41477-020-0690-7.

Peterson, Vanessa M et al. (Oct. 2017). “Multiplexed Quantification of Proteins and Transcripts in Single Cells”. In: *Nature Biotechnology* 35.10, pp. 936–939. ISSN: 1546-1696. DOI: 10.1038/nbt.3973.

Rogers, David W. et al. (June 2017). “Ribosome reinitiation can explain length-dependent translation of messenger RNA”. In: *PLoS Computational Biology* 13.6, e1005592. DOI: 10.1371/journal.pcbi.1005592. URL: https://doi.org/10.1371/journal.pcbi.1005592.

Singh, Abhyudai and Pavol Bokes (Sept. 2012). “Consequences of mRNA transport on stochastic variability in protein levels”. en. In: *Biophys. J.* 103.5, pp. 1087–1096.

Stoeckius, Marlon et al. (2017). "Simultaneous epitope and transcriptome measurement in single cells". In: *Nature Methods* 14, pp. 865–868. DOI: 10.1038/nmeth.4380.

Teo, Guangyu et al. (2014). "PECA: a novel statistical tool for deconvoluting time-dependent gene expression regulation". In: *Journal of Proteome Research* 13.1, pp. 29–37. DOI: 10.1021/pr400855q.

Wang, Jin et al. (2018). "ATAC-Seq analysis reveals a widespread decrease of chromatin accessibility in age-related macular degeneration." In: *Nature communications* 9(1), 1364. DOI: https://doi.org/10.1038/s41467-018-03856-y.

Wei, Y. N. et al. (Feb. 2015). "Transcript and protein expression decoupling reveals RNA binding proteins and miRNAs as potential modulators of human aging". In: *Genome Biology* 16.1, p. 41. DOI: 10.1186/s13059-015-0608-2.

Wu, Michael et al. (2020). "Transcriptional and proteomic insights into the host response in fatal COVID-19 cases". In: *Proceedings of the National Academy of Sciences of the United States of America* 117.45, pp. 28336–28343. DOI: 10.1073/pnas.2018030117.

Chapter 4

BIOPHYSICS OF GENE EXPRESSION EVOLUTION

Felce, Catherine et al. (2025). "Biophysical Constraints on mRNA Decay Rates Shape Macroevolutionary Divergence in Steady-State Abundances". In: *bioRxiv*. DOI: 10.1101/2025.11.24.690267. eprint: 2025.11.24.690267. URL: https://doi.org/10.1101/2025.11.24.690267.

Abstract

Evolutionary changes to gene expression are understood to be a major driver of phenotypic divergence between species. Researchers have investigated the drivers of this divergence by fitting evolutionary models to multi-species 'omic' datasets. It is now apparent that steady-state mRNA expression levels show patterns consistent with evolutionary constraints, likely as a consequence of stabilizing selection. However, as all previous work has used bulk RNA measurements, it has been impossible to determine which of the many cellular processes that contribute to steady-state abundances underlie the divergence between species. Here we develop a novel paradigm for addressing this open problem. Using multi-species single-cell expression data and biophysical models, we estimate mRNA transcriptional burst sizes, splicing rates, and decay rates across multiple species. We then derive phylogenetic models that describe the divergence of these rates under alternative evolutionary scenarios and fit these to the comparative data. We find evidence for biophysical constraints on the rates of mRNA decay, such that macroevolutionary divergence in expression is primarily a consequence of variation in transcriptional bursting.

Keywords: Biophysical modeling, single-cell transcriptomics, phylogenetics, evolutionary theory

Introduction

Gene expression divergence is understood to be a key determinant of phenotypic divergence between species (King and Wilson, 1975; Wray et al., 2003; Carroll, 2008). Omics technologies enabled comparative analyses of gene expression across individuals and species, providing insights into the molecular and evolutionary mechanisms shaping gene expression evolution, with most studies focusing on mRNA expression evolution measured via RNA-seq. Numerous studies have investigated gene expression evolution across species, revealing both widespread stabilizing selection and lineage-specific adaptive shifts in mRNA expression levels (Gilad, Oshlack, and Rifkin, 2006; Blekhman, Oshlack, and Gilad, 2008; Brawand et al., 2011; Joshua G Schraiber et al., 2013; Barr, Rhodes, and Gilad, 2023). Furthermore, this data has been used to identify complex evolutionary patterns such as organ-specific evolution (Brawand et al., 2011; Chen et al., 2019), conserved gene regulatory modules (S. Roy et al., 2013), differential expression correlated with the emergence of complex phenotypes (Bastide et al., 2022), and gene-by-gene coevolution of gene expression (Cope, O'Meara, and Gilchrist, 2020). Recently, Cope et al. (Cope, Joshua G. Schraiber, and Pennell, 2025) developed a phylogenetic framework to explicitly model the coevolution of mRNA and protein levels, revealing the mutational and selective coupling between these two layers of gene expression, and finding that natural selection is generally stronger on mean protein expression levels.

Despite the progress made in identifying patterns of gene expression evolution and the processes that drive them, these studies have been limited by their use of bulk mRNA measurements. While mean expression levels are a useful measure for studying gene expression evolution, these obfuscate *how* expression levels evolve and which mechanisms of gene expression are most dynamic or most constrained by natural selection. Steady-state mRNA abundances are determined by a large array of cellular processes (Furlan, de Pretis, and Pelizzola, 2020; Tippmann et al., 2012; Steinbrecht et al., 2024; Park et al., 2012; Alemu et al., 2014; Raj and van Oudenaarden, 2008), which can make different contributions to between-species expression divergence. For example, experimental work using a two-species yeast hybrid found that changes to mRNA degradation rates were often accompanied by opposite-effect changes in transcription rate (Dori-Bachash, Shema, and Tirosh, 2011). This implies that the evolution of regulatory elements has a multifaceted effect on gene expression levels (Hill, Vande Zande, and Wittkopp, 2021; Sarropoulos et al., 2021; Liu, Mosti, and Silver, 2021) that is strongly dependent on interactions between these elements across the genome, suggesting coevolution of regulatory

mechanisms (Barrière, Gordon, and Ruvinsky, 2012; Brown et al., 2014; Zrimec et al., 2020).

To access information about these important regulatory mechanisms from transcriptomic data, we need new approaches that explicitly consider biophysical parameters. Principled biophysical modeling is crucial for extracting biological information from RNA-seq data (Gorin, Vastola, and Pachter, 2023). The emergence of single-cell RNA-seq technologies has allowed for the estimation of key biophysical parameters, such as transcriptional burst size and frequency (Luo et al., 2022; Mahat et al., 2024; Tang et al., 2023). Single-cell snapshot experiments can effectively provide two 'time points', via the analysis of nascent and mature transcripts (Fang, Gorin, and Pachter, 2025; Zeisel et al., 2011; Gorin, Fang, et al., 2022). Such models have been used to estimate the relationship of transcriptional bursting to cell-cycle stage (Sukys and Grima, 2025), for principled cell-type clustering (Chari, Gorin, and Pachter, 2024), and to disentangle biological and technical noise (Gorin, Vastola, Fang, et al., 2022). The quantification of biophysical parameters from single-cell data opens up new routes for investigating mRNA expression evolution at the levels of transcriptional and post-transcriptional dynamics. With the increased availability of cross-species single-cell datasets, comparative analyses of biophysical parameters could reveal new insights into the biophysics of gene expression evolution on macroevolutionary timescales.

In this work, we introduce a new paradigm for investigating the evolutionary interplay of different regulatory 'levers'. Instead of bulk RNA measurements, we use single-cell data, fitting them with biophysically meaningful parameters. Specifically, we consider evolutionary hypotheses involving the coevolution of transcription and mRNA degradation across vertebrates (Figure 4.1). Motivated by experimental results, (Dori-Bachash, Shema, and Tirosh, 2011), we posit a fitness landscape that gives rise to constraining selection on the mean level of spliced mRNA expression. However, when selection acts only on mean expression, the underlying biophysical processes that result in mean expression may evolve as if unconstrained, a process known as systems drift(True and Haag, 2001; Schiffman and Ralph, 2022; Jiang et al., 2023; Veller and Muralidhar, 2025). Thus, we investigate models in which one of the biophysical parameters is under constraining selection, whilst the other is free to adapt to maintain the optimal level of spliced mRNA, to test the hypothesis that selection acts both at the level of gene expression and at the level of the biophysical determinants of gene expression. Building upon the methods of Cope et al. (Cope,

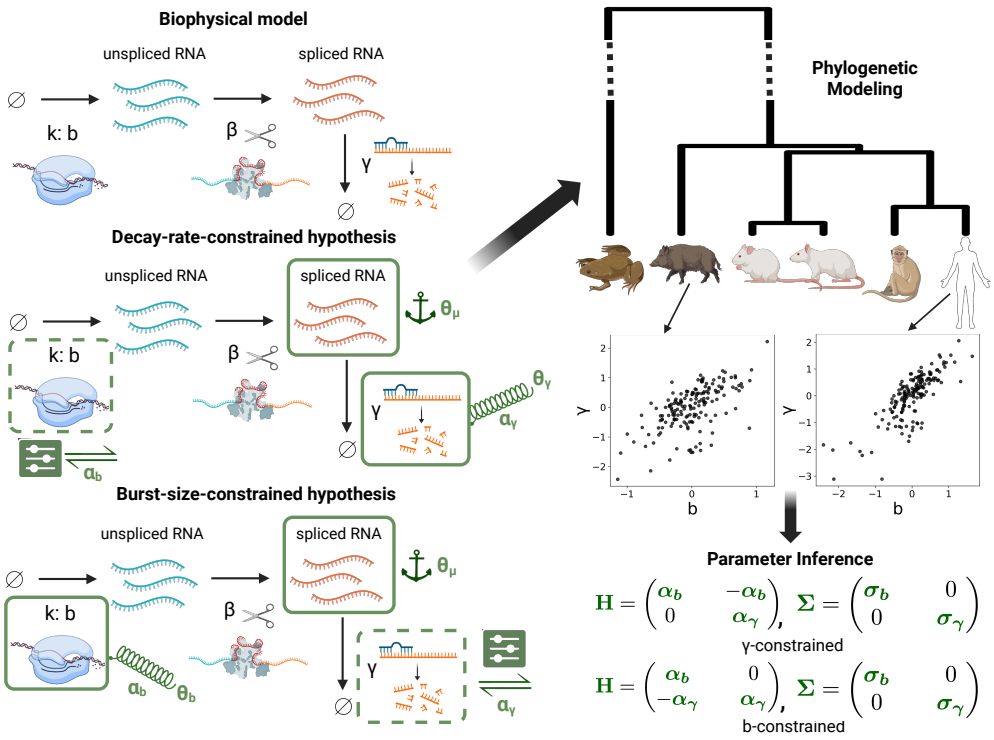


Figure 4.1: Our evolutionary model selection framework. The competing hypotheses impose different constraints on the evolution of the bursting size b and degradation rate γ. The model corresponding to each hypothesis is fit to the derived b and γ values for the species on the phylogenetic tree. Fitted selection matrices are obtained, and AICs suggest support for the decay-rate-constrained model.

Joshua G. Schraiber, and Pennell, 2025), we investigate the joint adaptation of transcriptional burst size and mRNA decay rate, finding support for the model with stabilizing selection on mRNA decay rates and mean spliced expression. This suggests that between-species differences in mRNA levels can be largely attributed to the flexible adaptation of transcriptional burst sizes.

Results

For our basic biophysical model, we chose bursty transcription with the following dynamics:

$$\emptyset \xrightarrow{k} B \times U \quad U \xrightarrow{\beta} S \quad S \xrightarrow{\gamma} \emptyset, \tag{4.1}$$

where transcriptional bursts occur at a rate k, producing bursts of unspliced (U) transcripts with sizes distributed according to B, a geometric distribution with mean

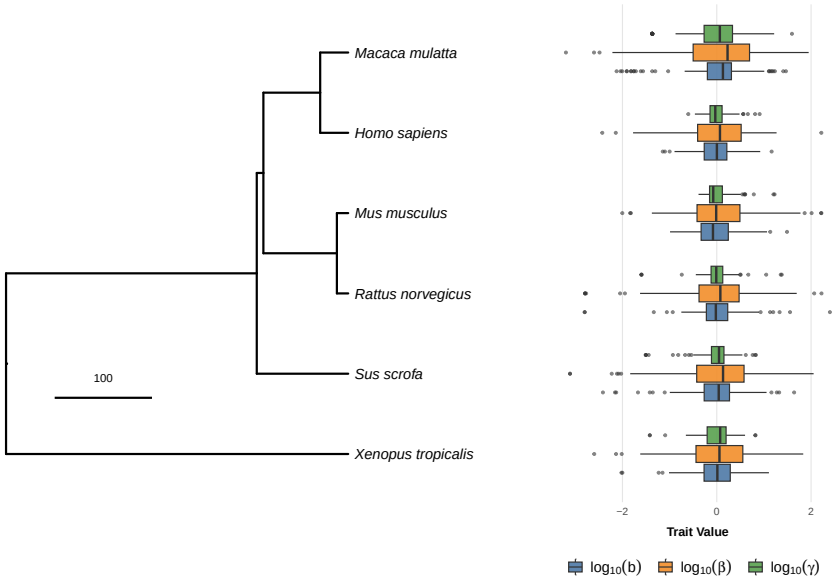


Figure 4.2: Values of the biophysical parameters across the phylogeny. The phylogeny is taken from TimeTree(Kumar et al., 2022) (scale bar in Myr). The boxplots at each tip show the mean-centered parameter distributions over genes for mean burst size, b, splicing rate, β, and RNA decay rate, γ, in the corresponding species. These trait values are the inputs into our phylogenetic model.

size b. The unspliced transcripts are converted to spliced mRNA (S) at a rate β, and the spliced transcripts then decay at a rate γ. The $\emptyset$ symbol indicates that the transcript before transcription and after decay does not appear in the model.

The biophysical parameters, b, β and γ, can be estimated from single-cell data using Monod (Gorin, Chari, et al., 2025), which optimizes the likelihood of the observed counts under the bursty model. We used Monod to infer these biophysical model parameters for single-cell transcriptomics data from Jiao et al.(Jiao et al., 2024) across the spleens of individuals from six different species. The output from this procedure were per-gene, per-species values of the biophysical parameters (b, β, γ) across 167 orthologous genes (Figure 4.2).

We then developed a combined biophysical and phylogenetic model describing the evolution of log burst size ($\log(b)$) and log mRNA decay rate ($\log(\gamma)$) along a tree. We chose these two parameters because together they determine the log mean spliced mRNA level

$$\log(\mu) = \log(b) - \log(\gamma), \tag{4.2}$$

which is independent of the splicing rate, β (note that γ is given in units of the burst initiation rate, k). Considering these two parameters allows us to test for

a constrained version of quantitative systems drift (Veller and Muralidhar, 2025) (recall that the log burst size and log mRNA decay rate would be free to drift if selection only acted on log mean expression). In Appendix C.2, we show that a two dimensional Ornstein-Uhlenbeck model captures the coevolution of burst size and mRNA decay rate due to selection on the mean spliced expression and an additional constraint on one of the two biophysical parameters. Thus, we can determine the biophysical mechanism through which gene expression evolution is mediated, and test which of the contributing biophysical processes is most constrained, alongside selection on mean expression.

We adopt a hierarchical model across genes, to exploit the large number of measured genes and compensate for the limited number of species in our dataset. In particular, while we assume that evolutionary rates are shared among genes, we allow the *optimal* biophysical parameters to vary between genes; in Appendix C.5 we show that we can analytically integrate over a Gaussian prior on the optima. When sharing information across genes, due to lineage specific adaptation as well as biological and technical noise, some genes may not have any phylogenetic signal (Eng, Bravo, and Keleş, 2009; Blomberg, Garland, and Ives, 2003; Freckleton, Harvey, and Pagel, 2002). Thus, we assume that with probability p_{wn} the biophysical parameters for a gene are taken from a white-noise distribution (see the outlier model in Chaix et al.(Chaix et al., 2008)), and with probability $1 - p_{\mathrm{wn}}$ that they evolve corresponding to our coevolutionary model (see Methods for more details on the full model).

We used the logarithms of the biophysical parameters as continuous characters in our two hypothesized OU models (Figure 4.1; see Appendix C.2 for the derivation of these models from fitness landscapes). The first model, which we henceforth refer to as the decay-rate-constrained (γ-constrained) model, assumes that the primary form of selection is stabilizing selection on the mRNA decay rates. The burst size is then assumed to adjust in response, to achieve an optimal mean level of spliced mRNA expression. In this model, the selection matrix, H, takes the form:

$$H = \begin{pmatrix} \alpha_b & -\alpha_b \\ 0 & \alpha_\gamma \end{pmatrix}. \tag{4.3}$$

The second model, which we refer to as the burst-size-constrained (b-constrained) model, assumes that the fitness is most sensitive to the average value of the transcriptional burst size, b. The value of b therefore undergoes strong constraining

selection, whilst the decay rate is assumed to adapt to maintain the optimal mean level of spliced mRNA expression. In this model, the selection matrix takes the form:

$$H = \begin{pmatrix} \alpha_b & 0 \\ -\alpha_\gamma & \alpha_\gamma \end{pmatrix}. \tag{4.4}$$

We used simulations to assess the suitability of our model for inferring evolutionary parameters from biophysical parameters. The simulation results confirm that we can reliably distinguish between the two models described above (see Appendix C.2, Figures C.1-C.4).

The data support a decay-rate-constrained model of transcriptional evolution

After fitting both phylogenetic models to the average burst sizes, b, and decay rates, γ, extracted from the Jiao et al.(Jiao et al., 2024) data, we compared AIC values and found support for the decay-rate-constrained model (AIC = 2,124), over both the independent (AIC = 2,726) and burst-size-constrained (AIC = 3,186) models. The fit parameters are shown in Table 4.1, and the AIC values are shown in Figure 4.3a. The out-performance of the γ-constrained model over the independent model confirms the importance of modeling the *coevolution* of biophysical parameters.

Model	α_b	α_γ	σ_b	σ_γ	p_{wn}
γ-constrained	31.5	2.33	0.923	1.06	0.149
b-constrained	0.022	3.56	3.01	0.169	0.900
Independent	1.36	1.59	0.651	0.831	0.329

Table 4.1: Phylogenetic model fitted parameters: selection rates, $\alpha_{b,\gamma}$, and mutations rates, $\sigma_{b,\gamma}$, on average transcriptional burst size, b, and mRNA decay rate, γ, along with the probability for a gene's biophysical parameters to be drawn from a white-noise distribution, p_{wn}.

There are numerous lines of evidence indicating that more highly-expressed genes generally experience stronger selection pressures, such as stronger purifying selection on amino acid substitutions (Drummond and Wilke, 2008; Managadze et al., 2011), stronger bias towards fast/accurate codons (Bénitière, Lefébure, and Duret, 2025; Cope and Shah, 2025) , and more conserved *cis* regulatory elements (Berthelot et al., 2018). In previous work, Cope et al. found that high-expression genes exhibited stronger selection on mRNA levels compared to low-expression genes. We

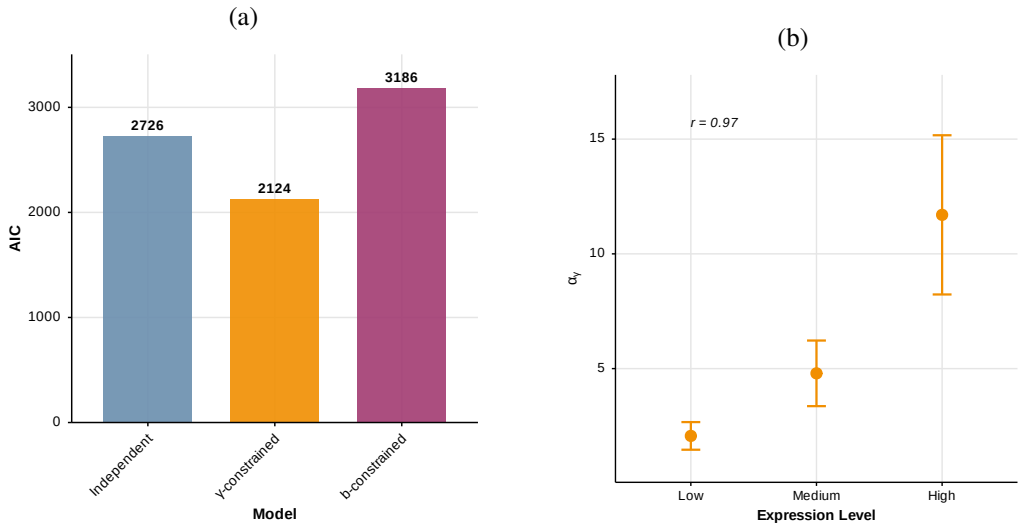


Figure 4.3: **(a):** AIC comparison across the tested phylogenetic models. A lower AIC value indicates a better model fit. **(b):** Selection coefficient for decay rate, α_γ, in the decay-rate-constrained model, for gene sub-groups ($n = 56, 55, 56$) binned by expression level.

decide to verify this observation by using the AIC-preferred γ-constrained model to compare the driving selection rate, α_γ, across genes with varying expression levels. To test whether selection on mRNA decay rates is stronger in more highly expressed genes, we binned the genes under investigation based on their median expression levels in human spleens (GTEx Consortium, 2013) and fit our phylogenetic mixture model separately to the genes in each bin, obtaining three sets of evolutionary parameters (Figure 4.3b). Additionally, we computed dN/dS values for the genes across the six species tree and found that, as expected, the genes with higher expression are subject to stronger purifying selection than those with lower expression (See Appendix C.2, Figure C.5). These results suggest that our fitted selection rate correlates with mRNA expression level, consistent with other patterns observed in protein-coding sequence and gene regulatory evolution.

Discussion

Our results reveal that the best model for mRNA expression evolution explicitly models the coevolution of mRNA burst size and decay rate. This accords with computational and experimental evidence that transcriptional evolution is a coordinated process across all regions of the genome (Zrimec et al., 2020), and that burst size and decay rates evolve in a compensatory manner (Andrie, Wakefield, and Akey, 2014;

Schaefke et al., 2018; Dori-Bachash, Shema, and Tirosh, 2011). We further show that this coordinated model should involve constraining selection on the decay rate, with burst sizes then adapting to maintain the desired overall expression level. This is consistent with the pleiotropy of the mechanisms of decay-rate adaptation, which suggest that decay rates may be under stabilizing selection (Agarwal and Kelley, 2022) independently from transcriptional burst sizes. For example, there is a close connection between mRNA decay and translation (Wu et al., 2019; Bae and Coller, 2022; Hanson and Coller, 2018; Carneiro et al., 2019; B. Roy and Jacobson, 2013; Chan et al., 2018; Bicknell et al., 2024). This implies that protein-level constraints may also cause stabilizing selection on RNA decay rates. In addition, alternative 3'UTRs, another determinant of RNA stability, simultaneously affect membrane protein localization (Berkovits and Mayr, 2015), mRNA localization, and translational efficiency (Mayr, 2016). This suggests that decay rates cannot adapt freely to achieve a certain level of expression, without also affecting other cellular processes.

In contrast, regulatory elements responsible for transcription are known to evolve rapidly (McQuarrie et al., 2024). For example, flexibly evolving promoter regions are thought to underlie a significant proportion of phenotypic diversity in humans (R. S. Young et al., 2022), and the frequent complete turnover of functional promoters has been observed in both humans and mice (Robert S. Young et al., 2015). These promoter regions are directly related to transcriptional burst size (Larsson et al., 2019; Hendy et al., 2017). Perhaps more so than promoters regions, comparative analysis reveal enhancer regions to experience rapid evolution (Villar et al., 2015; Uebbing et al., 2024), with multiple studies indicating that enhancers play an important role in regulating the frequency of transcriptional bursting (Bartman et al., 2016; Fukaya, Lim, and Levine, 2016; Larsson et al., 2019; Tünnermann et al., 2025). Chromatin state in regulatory regions can also be modified to influence transcriptional dynamics (Ernst et al., 2011), and has been shown to vary widely across human individuals (Kasowski et al., 2013). Chromatin state therefore represents another free 'tuning knob' which can affect transcriptional burst size.

Overall, this work represents a new paradigm for probing the regulatory mechanisms underlying macroevolutionary mRNA expression divergence. For the first time, we combine principled biophysical modeling on single-cell data across species with phylogenetic comparative modeling. By selecting a multivariate OU model for biophysically meaningful parameters, we have investigated the selective coupling of these traits, giving insight into how transcription and decay rates coevolve. We

have expanded on existing phylogenetic techniques, incorporating a multivariate OU model into a phylogenetic mixture model accounting for genes with no phylogenetic signal, as well as applying biophysical single-cell modeling in a coherent cross-species framework.

The approach outlined here can be extended to incorporate more complicated dynamics from both the biophysical and phylogenetic perspectives. As additional single-cell data modalities become available, with their corresponding joint biophysical models, these can be straightforwardly integrated into our method. For example, joint models for single-cell RNA with protein counts (Felce, Fang, and Pachter, 2025) could be used to provide insight into the coevolution of translation and protein decay rates, along with the existing transcriptional rates. This would represent another avenue for corroborating the coevolutionary model proposed by Cope et al. (Cope, Joshua G. Schraiber, and Pennell, 2025). In addition, including an integrated biophysical model for RNA and chromatin accessibility measurements (Felce, Gorin, and Pachter, 2024) could allow us to investigate the contribution of evolving on/off rates to gene expression evolution in a full telegraph model. The power of these approaches will increase as higher quality single-cell data, including more combined modalities across a greater number of species, become available.

On the phylogenetic side, our framework could be adapted to include more complicated evolutionary models, for example including mutational coupling between biophysical parameters, or expanding the dimensionality of the OU model to include coevolution between additional traits. With the flexibility to adapt and extend the two halves of our combined approach, researchers will be able to further dissect the contributions of different regulatory processes to the evolution of gene expression. This will provide new insights into how gene regulation has been shaped over the tree of life.

Data and code availability

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The average expression data used for the gene binning described in this manuscript were obtained from the GTEx Portal, V10 spleen tissue, on 10/15/2025. The transcriptomic data used in our analysis is taken from Jiao et al. (Jiao et al., 2024). The code and data for the phylogenetic model used to perform these analyses, and scripts to reproduce Figures 4.2, 4.3, and

the supplementary figures are available at `https://github.com/pachterlab/FCSKPP_2025/tree/main` (Pachter Lab, 2025).

Acknowledgements

M.K. and M.P. were supported by NIGMS award R35GM151348 and startup funds from Cornell University. We also thank Charles Trimble for generously funding part of C.F.'s research through Caltech's CI2 grant program.

Methods

Data processing and biophysical modeling

For this study we use single-cell RNA-seq data from Jiao et al. (Jiao et al., 2024), extracted from the spleen of seven different species. We processed the data using kallisto (Bray et al., 2016; Melsted et al., 2021; Delaney K Sullivan et al., 2025a; Delaney K. Sullivan et al., 2025b) to obtain spliced and unspliced count matrices. After clustering the data from each species by cell-type, we excluded the fish sample from further analysis because of an indistinct and low-count T-cell cluster, leaving six remaining species, which were filtered for T-cells. We searched for genes which had orthologs in all six species using Ensembl BioMart (Kinsella et al., 2011).

We then fit transcriptional rates for these genes in each species separately using Monod (Gorin, Chari, et al., 2025), using the bursty transcription model with Poisson technical noise. Monod is an inference framework which fits per-gene biophysical parameters for a selection of transcriptional models. The dynamics of the model are encapsulated in a chemical master equation, which can then be numerically solved to give steady-state distributions for spliced and unspliced RNA counts, including the impact of technical noise. The likelihood of the observed spliced/unspliced count matrices can then be maximized over biophysical parameters. In practice, this process is repeated over a grid of technical parameters, and the combined values which maximize the likelihood of the data are outputted.

The output of this procedure is a per-gene burst size, b, splicing rate β, and decay rate, γ, with the rates given in units of the transcription initiation rate, k, all in log space. We then subtracted the mean of each parameter across genes from each species. After fitting with Monod, which filters some genes, and removing genes without a fitted ortholog in all six species, we were left with 167 genes. The mean-centered log values of b, β, and γ for each of these genes were used as the traits for the following analysis.

Phylogenetic modeling and parameter inference

We consider the two-dimensional Ornstein-Uhlenbeck models for burst size, b, and decay rate, γ described in Results. Under these models, the logarithms of b and γ follow the following evolution equations:

$$d\boldsymbol{X}_t \;=\; -H(\boldsymbol{X}_t - \hat{\boldsymbol{X}})dt + \Sigma d\boldsymbol{W}_t, \tag{4.5}$$

where $\boldsymbol{W}_t$ is a Wiener process. In our models, we set Σ as a diagonal matrix with entries σ_b and σ_γ. $\boldsymbol{X}_t$ represents the logarithms of the two biophysical rates:

$$\boldsymbol{X}_t \;=\; \begin{pmatrix} \log b_t \\ \log \gamma_t \end{pmatrix}. \tag{4.6}$$

The forms for the selection matrices in each model:

$$H \;=\; \begin{pmatrix} \alpha_b & -\alpha_b \\ 0 & \alpha_\gamma \end{pmatrix} \qquad \text{and} \qquad H \;=\; \begin{pmatrix} \alpha_b & 0 \\ -\alpha_\gamma & \alpha_\gamma \end{pmatrix}, \tag{4.7}$$

are derived from the following forms for the fitness function, w:

$$\begin{aligned} w(b,\gamma) \;&\propto\; \exp\left(-\frac{(\hat{\gamma}-\theta_\gamma)^2}{2V_\gamma} - \frac{((\hat{b}-\hat{\gamma})-\theta_\mu)^2}{2V_\mu}\right), \\ &\propto \exp\left(-\frac{(\hat{b}-\theta_b)^2}{2V_b} - \frac{((\hat{b}-\hat{\gamma})-\theta_\mu)^2}{2V_\mu}\right), \end{aligned} \tag{4.8}$$

where we use $\hat{b}, \hat{\gamma}$ for the logarithms of b, γ, and for optimal log parameter values, $\theta_{b,\gamma}$ and an optimal log spliced mean expression level, θ_μ. Note that, since the optima and parameter values are in log space, $\hat{b} - \hat{\gamma}$ represents the ratio of burst size to mRNA decay rate, which is proportional to the mean spliced expression level. See Appendix C.2, for the full derivation of these models, which follows Cope et al. (Cope, Joshua G. Schraiber, and Pennell, 2025).

For parameter inference, we take the values of $\log b$ and $\log \gamma$ for each species, along with the phylogenetic tree. We fit a mixture model, where each gene is

generated from a white noise distribution with probability p_{wn}, and from the relevant phylogenetic model with probability $1 - p_{\text{wn}}$. The evolutionary selection strengths, $\alpha_{b,\gamma}$, and stochastic rates, $\sigma_{1,2}$, are constrained to be equal across genes, whereas the optima θ_μ and $\theta_{b,\gamma}$ are assumed to be drawn from normal distributions whose parameters are optimized.

We also fitted an independent OU model using MCMC for all of the biophysical parameters (b, β and γ), whose details and results are included in Appendix C.3 (Figures C.6-C.8). In addition, we fit two three-parameter-H versions of the coevolution model, whose results are included in Appendix C.4, Figures C.9-C.12.

References

Agarwal, Vikram and David R. Kelley (Nov. 2022). "The Genetic and Biochemical Determinants of mRNA Degradation Rates in Mammals". In: *Genome Biology* 23.1, p. 245. ISSN: 1474-760X. DOI: 10.1186/s13059-022-02811-x.

Alemu, Elfalem Y. et al. (Jan. 2014). "Determinants of Expression Variability". In: *Nucleic Acids Research* 42.6, pp. 3503–3514. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1364.

Andrie, J. M., J. Wakefield, and J. M. Akey (2014). "Heritable variation of mRNA decay rates in yeast". In: *Genome Research* 24.12, pp. 2000–2010. DOI: 10.1101/gr.175802.114.

Bae, Haneui and Jeff Coller (2022). "Codon Optimality-Mediated mRNA Degradation: Linking Translational Elongation to mRNA Stability". In: *Molecular Cell* 82.8, pp. 1467–1476. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2022.03.032.

Barr, Kenneth A., Katherine L. Rhodes, and Yoav Gilad (Sept. 2023). "The relationship between regulatory changes in cis and trans and the evolution of gene expression in humans and chimpanzees". In: *Genome Biology* 24, p. 207. ISSN: 1474-7596. DOI: 10.1186/s13059-023-03019-3. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10496171/ (visited on 11/19/2025).

Barrière, Antoine, Kacy L. Gordon, and Ilya Ruvinsky (Sept. 2012). "Coevolution within and between Regulatory Loci Can Preserve Promoter Function Despite Evolutionary Rate Acceleration". In: *PLOS Genetics* 8.9, e1002961. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002961. (Visited on 10/31/2025).

Bartman, Caroline R. et al. (Apr. 2016). "Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping". en. In: *Molecular Cell* 62.2, pp. 237–247. ISSN: 10972765. DOI: 10.1016/j.molcel.2016.03.007. (Visited on 11/24/2025).

Bastide, Paul et al. (Dec. 2022). "A Phylogenetic Framework to Simulate Synthetic Interspecies RNA-Seq Data". In: *Molecular Biology and Evolution* 40.1,

msac269. ISSN: 0737-4038. DOI: 10.1093/molbev/msac269. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11249980/ (visited on 11/17/2025).

Bénitière, Florian, Tristan Lefébure, and Laurent Duret (Jan. 2025). "Variation in the Fitness Impact of Translationally Optimal Codons among Animals". In: *Genome Research* 35.3, pp. 446–458. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.279837.124. (Visited on 11/20/2025).

Berkovits, Binyamin D. and Christine Mayr (June 2015). "Alternative 3'UTRs Act as Scaffolds to Regulate Membrane Protein Localization". In: *Nature* 522.7556, pp. 363–367. ISSN: 0028-0836. DOI: 10.1038/nature14321. (Visited on 10/17/2025).

Berthelot, Camille et al. (Jan. 2018). "Complexity and Conservation of Regulatory Landscapes Underlie Evolutionary Resilience of Mammalian Gene Expression". In: *Nature Ecology & Evolution* 2.1, pp. 152–163. ISSN: 2397-334X. DOI: 10.1038/s41559-017-0377-2.

Bicknell, Alicia A. et al. (Apr. 2024). "Attenuating Ribosome Load Improves Protein Output from mRNA by Limiting Translation-Dependent mRNA Decay". In: *Cell Reports* 43.4. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2024.114098. (Visited on 11/18/2025).

Blekhman, Ran, Alicia Oshlack, and Yoav Gilad (2008). "Segmented shifts in gene expression between human and chimpanzee". In: *Genome Research* 18.9, pp. 1514–1522. DOI: 10.1101/gr.075713.107.

Blomberg, Simon P., JR. Garland Theodore, and Anthony R. Ives (Apr. 2003). "Testing for Phylogenetic Signal in Comparative Data: Behavioral Traits Are More Labile". In: *Evolution; international journal of organic evolution* 57.4, pp. 717–745. ISSN: 0014-3820. DOI: 10.1111/j.0014-3820.2003.tb00285.x.

Brawand, D. et al. (2011). "The evolution of gene expression levels in mammalian organs". In: *Nature* 478, pp. 343–348. DOI: 10.1038/nature10532.

Bray, Nicolas L et al. (2016). "Near-optimal probabilistic RNA-seq quantification". In: *Nature biotechnology* 34.5, pp. 525–527.

Brown, Andrew Anand et al. (Apr. 2014). "Genetic Interactions Affecting Human Gene Expression Identified by Variance Association Mapping". In: *eLife* 3. Ed. by Philipp Khaitovich, e01381. ISSN: 2050-084X. DOI: 10.7554/eLife.01381.

Carneiro, Rodolfo L et al. (Jan. 2019). "Codon Stabilization Coefficient as a Metric to Gain Insights into mRNA Stability and Codon Bias and Their Relationships with Translation". In: *Nucleic Acids Research* 47.5, pp. 2216–2228. ISSN: 0305-1048. DOI: 10.1093/nar/gkz033.

Carroll, Sean B. (2008). "Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution". In: *Cell* 134.1, pp. 25–36. DOI: 10.1016/j.cell.2008.06.030.

Chaix, R. et al. (Nov. 2008). "Evolution of Primate Gene Expression: Drift and Corrective Sweeps?" In: *Genetics* 180.3, pp. 1379–1389. DOI: 10.1534/genetics.108.089623. URL: https://doi.org/10.1534/genetics.108.089623.

Chan, Leon Y et al. (Sept. 2018). "Non-Invasive Measurement of mRNA Decay Reveals Translation Initiation as the Major Determinant of mRNA Stability". In: *eLife* 7. Ed. by Alan G Hinnebusch, James L Manley, and Roy Parker, e32536. ISSN: 2050-084X. DOI: 10.7554/eLife.32536. (Visited on 11/18/2025).

Chari, Tara, Gennady Gorin, and Lior Pachter (Sept. 2024). "Biophysically Interpretable Inference of Cell Types from Multimodal Sequencing Data". In: *Nature Computational Science* 4.9, pp. 677–689. ISSN: 2662-8457. DOI: 10.1038/s43588-024-00689-2.

Chen, J. et al. (2019). "A quantitative framework for characterizing the evolutionary history of mammalian gene expression". In: *Genome Research* 29, pp. 53–63. DOI: 10.1101/gr.238873.118.

Cope, Alexander L., Brian C. O'Meara, and Michael A. Gilchrist (May 2020). "Gene expression of functionally-related genes coevolves across fungal species: detecting coevolution of gene expression using phylogenetic comparative methods". In: *BMC Genomics* 21.1, p. 370. ISSN: 1471-2164. DOI: 10.1186/s12864-020-6761-3. URL: https://doi.org/10.1186/s12864-020-6761-3 (visited on 11/19/2025).

Cope, Alexander L., Joshua G. Schraiber, and Matt Pennell (2025). "Macroevolutionary Divergence of Gene Expression Driven by Selection on Protein Abundance". In: *Science* 387.6738. DOI: 10.1126/science.ads2658.

Cope, Alexander L. and Premal Shah (July 2025). "Macroevolutionary Changes in Natural Selection on Codon Usage Reflect Evolution of the tRNA Pool across a Budding Yeast Subphylum". In: *Proceedings of the National Academy of Sciences* 122.27, e2419889122. DOI: 10.1073/pnas.2419889122. (Visited on 11/20/2025).

Dori-Bachash, Mally, Efrat Shema, and Itay Tirosh (July 2011). "Coupled Evolution of Transcription and mRNA Degradation". In: *PLoS Biology* 9.7. Ed. by Jürg Bähler, e1001106. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001106. (Visited on 10/16/2025).

Drummond, Daniel A. and Claus O. Wilke (2008). "Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution". In: *Cell* 134.2, pp. 341–352. DOI: 10.1016/j.cell.2008.05.042.

Eng, Kevin H., Héctor Corrada Bravo, and Sündüz Keleş (Oct. 2009). "A Phylogenetic Mixture Model for the Evolution of Gene Expression". In: *Molecular Biology and Evolution* 26.10, pp. 2363–2372. ISSN: 0737-4038. DOI: 10.1093/molbev/msp149. (Visited on 11/19/2025).

Ernst, Jason et al. (May 2011). "Systematic Analysis of Chromatin State Dynamics in Nine Human Cell Types". In: *Nature* 473.7345, pp. 43–49. ISSN: 0028-0836. DOI: 10.1038/nature09906. (Visited on 10/17/2025).

Fang, Meichen, Gennady Gorin, and Lior Pachter (2025). "Trajectory inference from single-cell genomics data with a process time model". In: *PLoS Computational Biology* 21.1. Version 2, published 21 January 2025, e1012752. DOI: 10.1371/journal.pcbi.1012752.

Felce, Catherine, Meichen Fang, and Lior Pachter (2025). "Joint Biophysical Modeling of Paired Single-Cell RNA and Protein Measurements". In: *bioRxiv : the preprint server for biology*. DOI: 10.1101/2025.11.14.688548.

Felce, Catherine, Gennady Gorin, and Lior Pachter (2024). "A Biophysical Model for ATAC-seq Data Analysis". In: *bioRxiv*. DOI: 10.1101/2024.01.25.577262. URL: https://www.biorxiv.org/content/early/2024/01/29/2024.01.25.577262.

Freckleton, R. P., P. H. Harvey, and M. Pagel (Dec. 2002). "Phylogenetic Analysis and Comparative Data: A Test and Review of Evidence." In: *The American Naturalist* 160.6, pp. 712–726. ISSN: 0003-0147. DOI: 10.1086/343873. (Visited on 11/19/2025).

Fukaya, Takashi, Bomyi Lim, and Michael Levine (July 2016). "Enhancer Control of Transcriptional Bursting". en. In: *Cell* 166.2, pp. 358–368. ISSN: 00928674. DOI: 10.1016/j.cell.2016.05.025. (Visited on 11/24/2025).

Furlan, Mattia, Stefano de Pretis, and Mattia Pelizzola (Dec. 2020). "Dynamics of Transcriptional and Post-Transcriptional Regulation". In: *Briefings in Bioinformatics* 22.4, bbaa389. ISSN: 1477-4054. DOI: 10.1093/bib/bbaa389.

Gilad, Yoav, Alicia Oshlack, and Scott A. Rifkin (2006). "Natural selection on gene expression". In: *Trends in Genetics* 22.8, pp. 456–461. DOI: 10.1016/j.tig.2006.06.002.

Gorin, Gennady, Tara Chari, et al. (Nov. 2025). "Monod: model-based discovery and integration through fitting stochastic transcriptional dynamics to single-cell sequencing data". en. In: *Nature Methods* 22.11. Publisher: Nature Publishing Group, pp. 2286–2300. ISSN: 1548-7105. DOI: 10.1038/s41592-025-02832-x. URL: https://www.nature.com/articles/s41592-025-02832-x (visited on 11/23/2025).

Gorin, Gennady, Meichen Fang, et al. (2022). "RNA velocity unraveled". In: *PLoS Computational Biology* 18.9. Version 2, published September 12, 2022, e1010492. DOI: 10.1371/journal.pcbi.1010492. URL: https://doi.org/10.1371/journal.pcbi.1010492.

Gorin, Gennady, John J. Vastola, Meichen Fang, et al. (Dec. 2022). "Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments". In: *Nature Communications* 13.1, p. 7620.

ISSN: 2041-1723. DOI: 10.1038/s41467-022-34857-7. URL: https://doi.org/10.1038/s41467-022-34857-7.

Gorin, Gennady, John J. Vastola, and Lior Pachter (2023). "Studying stochastic systems biology of the cell with single-cell genomics data". In: *Cell Systems* 14.10, 822–843.e22. ISSN: 2405-4712. DOI: https://doi.org/10.1016/j.cels.2023.08.004. URL: https://www.sciencedirect.com/science/article/pii/S2405471223002442.

GTEx Consortium (2013). "The Genotype-Tissue Expression (GTEx) project". In: *Nature Genetics* 45.6, pp. 580–585. DOI: 10.1038/ng.2653.

Hanson, Gavin and Jeff Coller (Jan. 2018). "Codon Optimality, Bias and Usage in Translation and mRNA Decay". In: *Nature Reviews Molecular Cell Biology* 19.1, pp. 20–30. ISSN: 1471-0072, 1471-0080. DOI: 10.1038/nrm.2017.91. (Visited on 10/16/2025).

Hendy, Oliver et al. (Nov. 2017). "Differential Context-Specific Impact of Individual Core Promoter Elements on Transcriptional Dynamics". In: *Molecular Biology of the Cell* 28.23, pp. 3360–3370. ISSN: 1059-1524. DOI: 10.1091/mbc.E17-06-0408. (Visited on 10/17/2025).

Hill, Mark S., Pétra Vande Zande, and Patricia J. Wittkopp (Apr. 2021). "Molecular and Evolutionary Processes Generating Variation in Gene Expression". In: *Nature Reviews Genetics* 22.4, pp. 203–215. ISSN: 1471-0064. DOI: 10.1038/s41576-020-00304-w.

Jiang, Daohan et al. (July 2023). "On the Decoupling of Evolutionary Changes in mRNA and Protein Levels". In: *Molecular Biology and Evolution* 40.8, msad169. ISSN: 0737-4038. DOI: 10.1093/molbev/msad169. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10411491/ (visited on 11/19/2025).

Jiao, Anjun et al. (Jan. 2024). "Single-cell sequencing reveals the evolution of immune molecules across multiple vertebrate species". In: *Journal of Advanced Research* 55. Epub 2023 Mar 4., pp. 73–87. ISSN: 2090-1224. DOI: 10.1016/j.jare.2023.02.017. URL: https://doi.org/10.1016/j.jare.2023.02.017.

Kasowski, Maya et al. (Nov. 2013). "Extensive Variation in Chromatin States across Humans". In: *Science (New York, N.Y.)* 342.6159, pp. 750–752. ISSN: 1095-9203. DOI: 10.1126/science.1242510.

King, Mary–Claire and Allan C. Wilson (1975). "Evolution at two levels in humans and chimpanzees". In: *Science* 188.4184, pp. 107–116. DOI: 10.1126/science.1090005.

Kinsella, Rhoda J. et al. (2011). "Ensembl BioMarts: A Hub for Data Retrieval across Taxonomic Space". In: *Database: The Journal of Biological Databases and Curation* 2011, bar030. ISSN: 1758-0463. DOI: 10.1093/database/bar030.

Kumar, Sudhir et al. (Aug. 2022). "TimeTree 5: An Expanded Resource for Species Divergence Times". In: *Molecular Biology and Evolution* 39.8, msac174. ISSN: 1537-1719. DOI: 10.1093/molbev/msac174. (Visited on 10/31/2025).

Larsson, Anton J. M. et al. (Jan. 2019). "Genomic Encoding of Transcriptional Burst Kinetics". In: *Nature* 565.7738, pp. 251–254. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0836-1.

Liu, Jing, Federica Mosti, and Debra L. Silver (2021). "Human Brain Evolution: Emerging Roles for Regulatory DNA and RNA". In: *Current Opinion in Neurobiology* 71, pp. 170–177. ISSN: 0959-4388. DOI: 10.1016/j.conb.2021.11.005.

Luo, Songhao et al. (Dec. 2022). "Genome-Wide Inference Reveals That Feedback Regulations Constrain Promoter-Dependent Transcriptional Burst Kinetics". In: *Nucleic Acids Research* 51.1, pp. 68–83. ISSN: 0305-1048. DOI: 10.1093/nar/gkac1204.

Mahat, Dig B. et al. (July 2024). "Single-Cell Nascent RNA Sequencing Unveils Coordinated Global Transcription". In: *Nature* 631.8019, pp. 216–223. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07517-7.

Managadze, D. et al. (2011). "Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs". In: *Genome Biology and Evolution* 3, pp. 1390–1404. DOI: 10.1093/gbe/evr116.

Mayr, Christine (Mar. 2016). "Evolution and Biological Roles of Alternative 3′UTRs". In: *Trends in cell biology* 26.3, pp. 227–237. ISSN: 0962-8924. DOI: 10.1016/j.tcb.2015.10.012. (Visited on 10/17/2025).

McQuarrie, David W. J. et al. (July 2024). "Rapid Evolution of Promoters from Germline-Specifically Expressed Genes Including Transposon Silencing Factors". In: *BMC Genomics* 25.1, p. 678. ISSN: 1471-2164. DOI: 10.1186/s12864-024-10584-9. (Visited on 10/17/2025).

Melsted, Páll et al. (2021). "Modular, efficient and constant-memory single-cell RNA-seq preprocessing". In: *Nature biotechnology* 39.7, pp. 813–818.

Pachter Lab (2025). *FCSKPP_2025: Code repository*. https://github.com/pachterlab/FCSKPP_2025. GitHub repository, accessed 22 November 2025.

Park, Jungsun et al. (Jan. 2012). "What Are the Determinants of Gene Expression Levels and Breadths in the Human Genome?" In: *Human Molecular Genetics* 21.1, pp. 46–56. ISSN: 0964-6906. DOI: 10.1093/hmg/ddr436. (Visited on 11/20/2025).

Raj, Arjun and Alexander van Oudenaarden (Oct. 2008). "Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences". In: *Cell* 135.2, pp. 216–226. ISSN: 0092-8674. DOI: 10.1016/j.cell.2008.09.050. (Visited on 11/20/2025).

Roy, B. and A. Jacobson (2013). "The intimate relationships of mRNA decay and translation". In: *Trends in Genetics* 29.12, pp. 691–699. DOI: 10.1016/j.tig.2013.09.002.

Roy, Sushmita et al. (June 2013). "Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules". eng. In: *Genome Research* 23.6, pp. 1039–1050. ISSN: 1549-5469. DOI: 10.1101/gr.146233.112.

Sarropoulos, Ioannis et al. (2021). "Developmental and Evolutionary Dynamics of Cis-Regulatory Elements in Mouse Cerebellar Cells". In: *Science* 373.6558, eabg4696. DOI: 10.1126/science.abg4696.

Schaefke, Bernhard et al. (2018). "The Evolution of Posttranscriptional Regulation". In: *WIREs RNA* 9.5, e1485. DOI: 10.1002/wrna.1485.

Schiffman, Joshua S and Peter L Ralph (2022). "System drift and speciation". In: *Evolution* 76.2, pp. 236–251.

Schraiber, Joshua G et al. (2013). "Inferring evolutionary histories of pathway regulation from transcriptional profiling data". In: *PLoS computational biology* 9.10, e1003255.

Steinbrecht, David et al. (Dec. 2024). "Subcellular mRNA Kinetic Modeling Reveals Nuclear Retention as Rate-Limiting". In: *Molecular Systems Biology* 20.12, pp. 1346–1371. ISSN: 1744-4292. DOI: 10.1038/s44320-024-00073-2.

Sukys, Augustinas and Ramon Grima (Apr. 2025). "Cell-Cycle Dependence of Bursty Gene Expression: Insights from Fitting Mechanistic Models to Single-Cell RNA-seq Data". In: *Nucleic Acids Research* 53.7, gkaf295. ISSN: 1362-4962. DOI: 10.1093/nar/gkaf295.

Sullivan, Delaney K et al. (2025a). "Accurate quantification of nascent and mature RNAs from single-cell and single-nucleus RNA-seq". In: *Nucleic acids research* 53.1, gkae1137.

Sullivan, Delaney K. et al. (2025b). "kallisto, bustools and kb-python for quantifying bulk, single-cell and single-nucleus RNA-seq". In: *Nature Protocols* 20.3, pp. 587–607. DOI: 10.1038/s41596-024-01057-0.

Tang, Wenhao et al. (June 2023). "Modelling Capture Efficiency of Single-Cell RNA-sequencing Data Improves Inference of Transcriptome-Wide Burst Kinetics". In: *Bioinformatics (Oxford, England)* 39.7, btad395. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad395.

Tippmann, Sylvia C et al. (2012). "Chromatin Measurements Reveal Contributions of Synthesis and Decay to Steady-state mRNA Levels". In: *Molecular Systems Biology* 8.1, p. 593. DOI: 10.1038/msb.2012.23.

True, John R and Eric S Haag (2001). "Developmental system drift and flexibility in evolutionary trajectories". In: *Evolution & development* 3.2, pp. 109–119.

Tünnermann, Jana et al. (Mar. 2025). *Enhancer control of promoter activity and variability via frequency modulation of clustered transcriptional bursts*. en. Pages: 2025.03.26.645410 Section: New Results. DOI: 10.1101/2025.03.26.645410. (Visited on 11/24/2025).

Uebbing, Severin et al. (Oct. 2024). "Evolutionary Innovations in Conserved Regulatory Elements Associate With Developmental Genes in Mammals". In: *Molecular Biology and Evolution* 41.10, msae199. ISSN: 1537-1719. DOI: 10.1093/molbev/msae199. (Visited on 11/24/2025).

Veller, Carl and Pavitra Muralidhar (2025). "Quantitative System Drift". In: *bioRxiv : the preprint server for biology*. DOI: 10.1101/2025.09.17.676933.

Villar, Diego et al. (Jan. 2015). "Enhancer Evolution across 20 Mammalian Species". en. In: *Cell* 160.3, pp. 554–566. ISSN: 00928674. DOI: 10.1016/j.cell.2015.01.006. (Visited on 11/24/2025).

Wray, Gregory A. et al. (2003). "The evolution of transcriptional regulation in eukaryotes". In: *Molecular Biology and Evolution* 20.9, pp. 1377–1419. DOI: 10.1093/molbev/msg140.

Wu, Qiushuang et al. (2019). "Translation affects mRNA stability in a codon-dependent manner in human cells". In: *eLife* 8, e45396. DOI: 10.7554/eLife.45396.

Young, R. S. et al. (2022). "The contribution of evolutionarily volatile promoters to molecular phenotypes and human trait variation". In: *Genome Biology* 23.1, p. 89. DOI: 10.1186/s13059-022-02634-w. URL: https://doi.org/10.1186/s13059-022-02634-w.

Young, Robert S. et al. (Oct. 2015). "The Frequent Evolutionary Birth and Death of Functional Promoters in Mouse and Human". In: *Genome Research* 25.10, pp. 1546–1557. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.190546.115. (Visited on 10/17/2025).

Zeisel, Amit et al. (2011). "Coupled pre-mRNA and mRNA Dynamics Unveil Operational Strategies Underlying Transcriptional Responses to Stimuli". In: *Molecular Systems Biology* 7.1, p. 529. DOI: 10.1038/msb.2011.62.

Zrimec, Jan et al. (Dec. 2020). "Deep Learning Suggests That Gene Expression Is Encoded in All Parts of a Co-Evolving Interacting Gene Regulatory Structure". In: *Nature Communications* 11.1, p. 6141. ISSN: 2041-1723. DOI: 10.1038/s41467-020-19921-4.

Chapter 5

PHYSICAL MODELS IN POPULATION EVOLUTION

Felce, Catherine, Steinunn Liorsdóttir, and Lior Pachter (Nov. 2025). "Analogies between the virial theorem and the Price equation". In: *Phys. Rev. E* 112 (5), p. 054139. DOI: 10.1103/r8bd-lhm3. URL: https://link.aps.org/doi/10.1103/r8bd-lhm3.

5.1 Introduction

We observe that the time-averaged continuous Price equation is identical to the positive momentum virial theorem, and we discuss the applications and implications of this connection. We also introduce ecological models, using a maternal effect, which can describe arbitrary population size cycles and trait evolution across spatially separated populations. The virial theorem sheds light on the time-averaged behavior of these models.

5.2 The virial theorem

The virial theorem was first described by Rudolf Clausius in connection with his studies on heat transfer (Clausius, 1870). In its simplest form, it relates the time-averaged kinetic energy $\langle T\rangle_\tau = \langle \frac{1}{2}\sum_{i=1}^n m_i v_i(t)^2\rangle_\tau = \frac{1}{\tau}\int_0^\tau \frac{1}{2}\sum_{i=1}^n m_i v_i(t)^2 dt$ of n objects with masses $m_1,\ldots,m_n$ and velocities $v_1(t),\ldots,v_n(t)$, to their time averaged potential energy $\langle U\rangle_\tau = \langle \sum_{i=1}^n F_i(t) z_i(t)\rangle_\tau$, where $F_1(t),\ldots,F_n(t)$ are the forces acting on the n objects and $z_1(t),\ldots,z_n(t)$ are their respective positions:

Theorem 1 *(Virial theorem, 1870).* For stably bound gravitational systems:

$$\langle T\rangle_\tau = -\frac{1}{2}\langle U\rangle_\tau. \tag{5.1}$$

The mathematical underpinning of the virial theorem is the product rule from calculus. The derivative of the Clausius virial $S(t) = \sum_{i=1}^n p_i(t) z_i(t)$ where $p_i(t) = m_i v_i(t)$ is:

$$\begin{aligned}\frac{dS(t)}{dt} &= \sum_{i=1}^{n} m_i \frac{dv_i(t)}{dt} z_i(t) + \sum_{i=1}^{n} p_i(t) \frac{dz_i(t)}{dt} \\ &= \sum_{i=1}^{n} F_i(t) z_i(t) + \sum_{i=1}^{n} m_i v_i(t)^2 \end{aligned} \tag{5.2}$$

Since for stably bound systems the velocities and positions of objects have upper and lower bounds, the average of the derivative of $S(t)$ over a period of time τ will be zero in the limit of large τ, i.e. $\langle \frac{dS(t)}{dt} \rangle_\tau \approx 0$. Therefore, equation (5.2) implies:

$$\left\langle \sum_{i=1}^{n} F_i(t) z_i(t) \right\rangle_\tau + \left\langle \sum_{i=1}^{n} m_i v_i(t)^2 \right\rangle_\tau = 0. \tag{5.3}$$

In the specific case of a gravitationally bound system, we have the identification $\sum_i F_i z_i = U$, for U the potential energy, and we obtain from (5.3) the gravitational virial theorem (5.1):

$$\begin{aligned}\left\langle \frac{dS(t)}{dt} \right\rangle_\tau &= \langle U \rangle_\tau + 2\langle T \rangle_\tau \\ \implies \langle T \rangle_\tau &= -\frac{1}{2} \langle U \rangle_\tau,\end{aligned}$$

where T is the kinetic energy of the system. In general, for forces with a potential of the form $U \propto r^n$, we have $\langle T \rangle = \frac{n}{2} \langle U \rangle$.

The virial theorem was well known to physicists in the late 19th and early 20th centuries (Rayleigh, 1905; Einstein, 1922), however, its power as a discovery tool for astrophysics was first highlighted by Fritz Zwicky (Zwicky, 1933). Zwicky used the virial theorem to estimate the mass of the Coma cluster, thereby identifying a mass deficit in comparison to luminosity estimates, leading him to posit the existence of what he called *dunkle materie* (dark matter) (Zwicky, 1933). Although Zwicky's mass estimates were inaccurate (The and White, 1986; Merritt, 1987), the principle of using the virial theorem to identify a measurement gap was sound, and the virial theorem has become widely used in physics and astrophysics. For example, it has been used to deduce the strength of magnetic fields in stars (Mukhopadhyay, Sarkar, and Tout, 2025) and the radii of elliptical galaxies (Binney and Merrifield, 1998).

It can also be used to derive classic laws such as the ideal gas law (Fowler, 1929; Hanson, 1995), and extensions are applicable in many settings, including quantum mechanics (Georgescu and Gérard, 1999), astrophysical hydrodynamics (Shore, 2012), and fluid mechanics (Oguz and Prosperetti, 1990; Alazard and Zuily, 2023).

5.3 The Price equation

The Price equation (Price et al., 1970) pertains to selection in evolutionary processes. It was motivated by a desire to understand the evolution of altruism (Harman, 2011), and has been described as a "fundamental theorem of evolution" (Queller, 2017) due to its generalization and unification of many results in evolutionary biology. For example, Fisher's fundamental theorem of natural selection (Fisher, 1930) is a special case of the Price equation (Queller, 2017; Frank, 1997).

The Price equation relates the change in a trait in a population over time, to fitness values in subpopulations. Formally, the (discrete) Price equation as published in (Price et al., 1970) (we follow notation from (Frank, 1997)) considers a numerical trait in n subpopulations at time t denoted $\mathbf{z}(t) = (z_1(t), \ldots, z_n(t))$. The subpopulations have sizes $p_1(t), \ldots, p_n(t)$, and have (Wrightian) fitness $\mathbf{w}(t) = (w_1(t), \ldots, w_n(t))$ defined by $w_i(t) = \frac{p_i(t+\Delta t)}{p_i(t)}$ where Δt denotes the time interval of one generation (Wagner, 2010). The Wrightian fitness is the average number of offspring an individual contributes to the next generation. Let $q_i(t) = \frac{p_i(t)}{\sum_{j=1}^{n} p_j(t)}$ be the relative size of the i^{th} population, and define the population average fitness to be $\overline{\mathbf{w}}(t) = \sum_{i=1}^{n} q_i(t) w_i(t)$. Note that $\mathbf{q}(t)$ forms a probability distribution for $\mathbf{w}(t)$ viewed as a random variable, and $\mathbb{E}(\mathbf{w}(t)) = \overline{\mathbf{w}}(t)$. Let $\Delta z_i(t) = z_i(t + \Delta t) - z_i(t)$, $\Delta \mathbf{z}(t) = \mathbf{z}(t + \Delta t) - \mathbf{z}(t)$, and $\overline{\mathbf{z}}(t) = \sum_{i=1}^{n} q_i(t) z_i(t)$ with $\Delta \overline{\mathbf{z}}(t) = \overline{\mathbf{z}}(t + \Delta t) - \overline{\mathbf{z}}(t)$.

Theorem 2 *(The Price equation, 1970)***.**

$$\Delta \overline{\mathbf{z}}(t) = \frac{1}{\overline{\mathbf{w}}(t)} \mathrm{cov}(\mathbf{w}(t), \mathbf{z}(t)) + \frac{1}{\overline{\mathbf{w}}(t)} \mathbb{E}(\mathbf{w}(t) \odot \Delta \mathbf{z}(t)), \tag{5.4}$$

where $\mathbb{E}(\mathbf{w}(t) \odot \Delta \mathbf{z}(t))$ is the expected value of the Hadamard product of $\mathbf{w}(t)$ and $\Delta \mathbf{z}(t)$ with respect to the relative subpopulation sizes, and $\mathrm{cov}(\mathbf{w}(t), \mathbf{z}(t)) = \mathbb{E}(\mathbf{w} \odot \mathbf{z}) - \mathbb{E}(\mathbf{w})\mathbb{E}(\mathbf{z})$ is the covariance between the subpopulation fitnesses and trait values with respect to the relative subpopulation sizes.

Intuitively, if subpopulation fitness has positive covariance with trait values, then the trait is beneficial, and the trait value, averaged across populations, will increase after a generation. However, if the covariance between subpopulation fitness and trait

values is negative, higher trait values are detrimental and the trait value averaged across populations will decrease after a generation.

The Price equation as published in (Price et al., 1970) is discrete in time, and proof of the identity uses basic properties of expectation and covariance along with the fact that $q_i(t + \Delta t) = \frac{q_i(t)w_i(t)}{\overline{\mathbf{w}}(t)}$, which we leave as an exercise for the reader. Note that:

$$\begin{aligned}
\overline{\mathbf{w}}(t)\Delta\overline{\mathbf{z}}(t) &= \overline{\mathbf{w}}(t)\overline{\mathbf{z}}(t + \Delta t) - \overline{\mathbf{w}}(t)\overline{\mathbf{z}}(t) \\
&= \overline{\mathbf{w}}(t) \sum_{i=1}^{n} q_i(t + \Delta t)z_i(t + \Delta t) - \overline{\mathbf{w}}(t)\overline{\mathbf{z}}(t) \\
&= \sum_{i=1}^{n} q_i(t)w_i(t)z_i(t + \Delta t) - \overline{\mathbf{w}}(t)\overline{\mathbf{z}}(t) \\
&= \sum_{i=1}^{n} q_i(t)w_i(t)z_i(t) - \overline{\mathbf{w}}(t)\overline{\mathbf{z}}(t) \\
&+ \sum_{i=1}^{n} q_i(t)w_i(t)z_i(t + \Delta t) \\
&- \sum_{i=1}^{n} q_i(t)w_i(t)z_i(t) \\
&= \mathrm{cov}(\mathbf{w}(t), \mathbf{z}(t)) + \mathbb{E}(\mathbf{w}(t) \odot \Delta\mathbf{z}(t)), \\
\implies \Delta\overline{\mathbf{z}}(t) &= \frac{1}{\overline{\mathbf{w}}(t)}\mathrm{cov}(\mathbf{w}(t), \mathbf{z}(t)) \\
&+ \frac{1}{\overline{\mathbf{w}}(t)}\mathbb{E}(\mathbf{w}(t) \odot \Delta\mathbf{z}(t)).
\end{aligned}$$

The discrete-time Price equation has a continuous-time analog (Price, 1972; Ellner, Geber, and Hairston Jr, 2011). It is formulated using the Malthusian fitness $\mathbf{r}(t) = r_1, \ldots, r_n$ given by $r_i(t) = \frac{1}{p_i(t)}\frac{dp_i(t)}{dt}$ instead of the Wrightian fitness $\mathbf{w}(t)$. The Malthusian fitness is the per capita rate at which individuals contribute offspring to the next generation.

Theorem 3 *(The continuous Price equation, 1972)***.**

$$\frac{d}{dt}\mathbb{E}(\mathbf{z}(t)) = \mathrm{cov}(\mathbf{r}(t), \mathbf{z}(t)) + \mathbb{E}\left(\frac{d\mathbf{z}(t)}{dt}\right). \tag{5.5}$$

The continuous-time Price equation (5.5) is the continuum limit of the discrete Price equation (5.4). To see this, we begin by multiplying the Price equation by $\frac{\overline{\mathbf{w}}(t)}{\Delta t}$:

$$\frac{\overline{\mathbf{w}}(t)\Delta\overline{\mathbf{z}}(t)}{\Delta t} = \frac{1}{\Delta t}\text{cov}(\mathbf{w}(t), \mathbf{z}(t)) + \frac{1}{\Delta t}\mathbb{E}(\mathbf{w}(t) \odot \Delta\mathbf{z}(t)).$$

We will now see why, contrary to convention, we have indexed the variables in equation (5.4) with time. Starting with the left hand side, we observe that:

$$\begin{aligned}
\lim_{\Delta t\to 0} \overline{\mathbf{w}}(t) &= \lim_{\Delta t\to 0} \sum_{i=1}^{n} q_i(t) w_i(t) \\
&= \lim_{\Delta t\to 0} \sum_{i=1}^{n} \frac{p_i(t) p_i(t+\Delta t)}{\left(\sum_{j=1}^{n} p_j(t)\right) p_i(t)} \\
&= \frac{1}{\sum_{j=1}^{n} p_j(t)} \lim_{\Delta t\to 0} \sum_{i=1}^{n} p_i(t+\Delta t) \;=\; 1.
\end{aligned}$$

Therefore:

$$\begin{aligned}
&\lim_{\Delta t\to 0} \frac{\overline{\mathbf{w}}(t)\Delta\overline{\mathbf{z}}(t)}{\Delta t} \\
&= \lim_{\Delta t\to 0} \frac{\sum_{i=1}^{n} q_i(t+\Delta t) z_i(t+\Delta t) - \sum_{i=1}^{n} q_i(t) z_i(t)}{\Delta t} \\
&= \frac{d}{dt}\mathbb{E}(\mathbf{z}(t)).
\end{aligned}$$

The covariance term, in the limit as $\Delta t \to 0$, is given by:

$$\begin{aligned}
&\lim_{\Delta t\to 0} \frac{1}{\Delta t}\text{cov}(\mathbf{w}(t), \mathbf{z}(t)) \\
&\quad = \lim_{\Delta t\to 0} \frac{\sum_{i=1}^{n} q_i(t) w_i(t) z_i(t) - \overline{\mathbf{w}}(t)\overline{\mathbf{z}}(t)}{\Delta t}.
\end{aligned}$$

Let $g_i(t) = \frac{1}{\Delta t} ln(w_i(t))$. Note that $w_i(t) = e^{g_i(t)\Delta t}$ and that $\lim_{\Delta t\to 0} g_i(t) = r_i(t)$.

Substituting $e^{g_i(t)\Delta t}$ for $w_i(t)$ yields:

$$\begin{aligned}
&\lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{cov}(\mathbf{w}(t), \mathbf{z}(t)) \\
&= \lim_{\Delta t \to 0} \frac{\sum_{i=1}^{n} q_i(t) e^{g_i(t)\Delta t} z_i(t) - \sum_{i=1}^{n} q_i(t) e^{g_i(t)\Delta t} \overline{\mathbf{z}}(t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\sum_{i=1}^{n} q_i(t) e^{g_i(t)\Delta t} (z_i(t) - \overline{\mathbf{z}}(t))}{\Delta t} \\
&= \lim_{\Delta t \to 0} \sum_{i=1}^{n} q_i(t) g_i(t) e^{g_i(t)\Delta t} (z_i(t) - \overline{\mathbf{z}}(t)), \\
&\text{by the L'Hôpital-Bernoulli rule (L'Hôpital, 1696),} \\
&= \sum_{i=1}^{n} q_i(t) r_i(t) (z_i(t) - \overline{\mathbf{z}}(t)) \\
&= \sum_{i=1}^{n} q_i(t) r_i(t) z_i(t) - \sum_{i=1}^{n} q_i(t) r_i(t) \overline{\mathbf{z}}(t) \\
&= \mathrm{cov}(\mathbf{r}(t), \mathbf{z}(t)).
\end{aligned}$$

Finally, we have that:

$$\begin{aligned}
\lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{E}(\mathbf{w}(t) \odot \Delta \mathbf{z}(t)) &= \lim_{\Delta t \to 0} \frac{\sum_{i=1}^{n} q_i(t) e^{g_i(t)\Delta t} \Delta z_i(t)}{\Delta t} \\
&= \sum_{i=1}^{n} q_i(t) \frac{dz_i(t)}{dt} \\
&= \mathbb{E}\left(\frac{d\mathbf{z}(t)}{dt}\right).
\end{aligned}$$

In summmary:

$$\frac{\overline{\mathbf{w}}(t)\Delta \mathbf{z}(t)}{\Delta t} = \frac{1}{\Delta t} \mathrm{cov}(\mathbf{w}(t), \mathbf{z}(t)) + \frac{1}{\Delta t} \mathbb{E}(\mathbf{w}(t) \odot \Delta \mathbf{z}(t))$$

(discrete Price equation (5.4))

$$\begin{array}{cc}
\Big\downarrow \lim_{\Delta t \to 0} & \Big\downarrow \lim_{\Delta t \to 0} \\
\frac{d}{dt} \mathbb{E}(\mathbf{z}(t)) = \mathrm{cov}(\mathbf{r}(t), \mathbf{z}(t)) + \mathbb{E}\left(\frac{d\mathbf{z}(t)}{dt}\right) &
\end{array}$$

(continuous Price equation (5.5)).

5.4 The Price equation from the virial theorem

In the physics setting, recall that the momentum $p_i(t) = m_i v_i(t) = m_i \frac{dz_i(t)}{dt}$. Let $r_i = \frac{1}{p_i(t)} \frac{dp_i(t)}{dt}$, i.e. acceleration divided by velocity. If all the momenta are

greater than zero, i.e., $p_i(t) > 0$ for all i, we can define relative momentum as $q_i(t) = \frac{p_i(t)}{\sum_{j=1}^n p_j(t)}$. Consider the virial density $\tilde{S}(t) = \sum_{i=1}^n q_i(t) z_i(t)$ (Englert, 2014), whose derivative is $\frac{d\tilde{S}(t)}{dt} = \frac{d}{dt}\mathbb{E}(\mathbf{z}(t))$. The product rule applied to the virial density is:

$$
\begin{aligned}
&\frac{d}{dt}\mathbb{E}(\mathbf{z}(t)) \\
&\quad = \sum_{i=1}^n \frac{d}{dt}\left(\frac{p_i(t)}{\sum_{j=1}^n p_j(t)}\right) z_i(t) + \sum_{i=1}^n q_i(t)\frac{dz_i(t)}{dt} \\
&\quad = \sum_{i=1}^n \frac{\frac{dp_i(t)}{dt}\sum_{j=1}^n p_j(t) - p_i(t)\sum_{j=1}^n \frac{dp_j(t)}{dt}}{(\sum_{j=1}^n p_j(t))^2} z_i(t) \\
&\qquad + \mathbb{E}\left(\frac{d\mathbf{z}(t)}{dt}\right) \\
&\quad = \sum_{i=1}^n \frac{p_i(t)}{\sum_{j=1}^n p_j(t)}\left(r_i(t) - \frac{\sum_{j=1}^n r_j(t)p_j(t)}{\sum_{j=1}^n p_j(t)}\right) z_i(t) \\
&\qquad + \mathbb{E}\left(\frac{d\mathbf{z}(t)}{dt}\right) \\
&\quad = \sum_{i=1}^n q_i(t) r_i(t) z_i(t) - \sum_{j=1}^n q_j(t) r_j(t) \sum_{i=1}^n q_i(t) z_i(t) \\
&\qquad + \mathbb{E}\left(\frac{d\mathbf{z}(t)}{dt}\right) \\
&\quad = \mathbb{E}(\mathbf{r}(t) \odot \mathbf{z}(t)) - \mathbb{E}(\mathbf{r}(t))\mathbb{E}(\mathbf{z}(t)) + \mathbb{E}\left(\frac{d\mathbf{z}(t)}{dt}\right), \\
\implies \quad &\frac{d}{dt}\mathbb{E}(\mathbf{z}(t)) = \operatorname{cov}(\mathbf{r}(t), \mathbf{z}(t)) + \mathbb{E}\left(\frac{d\mathbf{z}(t)}{dt}\right). \qquad (5.6)
\end{aligned}
$$

This shows that the virial (density) equation (5.6) and the continuous Price equation (5.5) are mathematically identical. Therefore, the relationships between traits and fitness in evolutionary biology are not only reminiscent of the relationships between physical quantities like distance, velocity, and acceleration; they are the same. It is therefore not surprising to find a direct analog of the virial theorem in genetics (Frank and Slatkin, 1990):

Theorem 4 *(Frank and Slatkin, 1990).* For a population in equilibrium, with $\mathbf{z}(t)^2 = \mathbf{z}(t) \odot \mathbf{z}(t)$:

$$\operatorname{cov}(\mathbf{w}(t), \mathbf{z}(t)^2) = -\mathbb{E}(\mathbf{w} \odot \Delta[\mathbf{z}(t)^2]).$$

In other words, the rate at which selection removes phenotypic variance from the population (left-hand side) is equal to the rate at which mutation adds variance (right-hand side). This result is a special case of the discrete Price equation (5.4), where the traits whose evolution are being considered are the squares of z_i. Because the population is assumed to be in equilibrium, we can take the left-hand side to be zero. The time variable, t, takes discrete values, separated by the length of a generation, as in equation (5.4). However, the result also has a continuous-time analog similar to Theorem 3. Frank and Slatkin considered the specific case where the z_i are proportional to the allelic states of the haploid genotypes, but their result is independent of the form of z.

In summary, we have the following relationships:

Biology		**Physics**
Price equation		virial theorem
$\downarrow \lim_{\Delta t \to 0}$		$\downarrow p_i > 0$
continuous Price equation	=	positive momentum virial theorem

In genetics, the natural time increment to consider is discrete (generation), whereas in physics continuous-time is more natural. Thus, the discrete Price equation pertains to change in a trait after a single generation, whereas the virial theorem is formulated with continuous-time, and is additionally time averaged. However, the less intuitive forms of these equations that arise from the correspondences derived above may yield important insights. For example, the perspective of the virial theorem as a special case of the equipartition theorem (Podio-Guidugli, 2019) may be fruitful in evolutionary biology (Nourmohammad, Held, and Lässig, 2013). From the other direction, it might be interesting to consider physical systems with positive-momentum constituents in terms of means and covariances over a momentum-weighted distribution. Translation between biology and physics via the virial theorem and the Price equation may also accelerate discovery of generalizations. While the stochastic Price equation in evolution (Rice, 2008) and the stochastic virial theorem in astronomy (Cresson, Nottale, and Lehner, 2021) were discovered independently, their similarity suggests other generalizations could similarly parallel each other. Moreover, the virial theorem has been applied in a variety of fields, meaning that understanding its relationship to the Price equation could be relevant beyond physics and biology.

For example, Anderson (Andersen, 2004) discusses the utility of the Price equation in decomposing short-term economic growth into selection and innovation effects. He emphasizes that the general form of the Price equation allows complex systems to be described in terms of nested selection effects in a 'multi-level population' (e.g. corporations which are made up of constituent plants). In the following section, we likewise consider how the composition of subpopulations could explain complicated dynamics in ecology. Although Anderson identifies interpopulation interactions as a limitation of the Price equation for describing long-term evolutionary change, we show that a spatial population growth model can be used with the Price equation to study separate but interacting subpopulations.

5.5 Simple Harmonic Motion

The connection between the Price equation and the virial theorem is evident in the study of a model motivated by work of Ginzburg and Colyvan (L. Ginzburg and Colyvan, 2004). In their book "Ecological Orbits: How Planets Move and Populations Grow", they show that a maternal effect can generate population cycles (L. R. Ginzburg and Taneyhill, 1994), obviating the need for predator-prey models to explain such dynamics. A maternal effect is defined as "the causal influence of the maternal genotype or phenotype on the offspring phenotype" (Wolf and Wade, 2009). Note that these maternal attributes can be genetic (e.g. snail shell chirality determined by maternal genotype (Boycott et al., 1931)), or due to environmental effects (Fox, Thakar, and Mousseau, 1999). The Price equation motivates an analysis of an extension of the Ginzburg and Colyvan model to subpopulations, while the virial theorem motivates a study of time-averaged behavior, which we show leads to a relationship between population-size entropy and trait variance. Formally, (L. Ginzburg and Colyvan, 2004) posit that a trait $z(t)$ that changes over time is linked to the size $p(t)$ of a population as follows:

$$p(t+\Delta t) = p(t)\tilde{f}(z(t)) \tag{5.7}$$

$$z(t+\Delta t) = z(t)\tilde{g}\left(\frac{R}{p(t+\Delta t)}\right), \tag{5.8}$$

where $\tilde{f}$ and $\tilde{g}$ are monotonically increasing functions, and Δt again represents the time interval for a single generation. The maternal effect is captured by $z(t)$ on the right-hand side of the trait evolution equation (5.8), indicating that a trait associated with individuals in a generation depends on the trait of mothers in the

current generation, as well as the fraction of the total resources, R, available to each individual in the next generation, i.e. $\frac{R}{p(t+\Delta t)}$. We extend this model to include an additional function $\tilde{h}$ that captures the potentially dominant impact of transgenerational effects as seen in matrotrophic species (Bian et al., 2015; Harding, 2001; Reznick, Callahan, and Llauredo, 2015; Roseboom, Rooij, and Painter, 2006):

$$z(t+\Delta t) = z(t)\tilde{g}\left(\frac{R}{p(t+\Delta t)}\right)\tilde{h}\left(\frac{R}{p(t)}\right). \tag{5.9}$$

This formulation captures environmental impacts on the mother which affect her offspring directly, such as nutrition during gestation (Thorne, Dean, and Hepworth, 1976). We can further simplify (5.9) to the case where transgenerational effects dominate, i.e. we assume that $\tilde{g} = 1$, giving:

$$z(t+\Delta t) = z(t)\tilde{h}\left(\frac{R}{p(t)}\right). \tag{5.10}$$

To derive a continuum limit from (5.10), as in our derivation of (5.5), we consider the limit where Δt becomes an infinitesimal time increment. We have:

$$\begin{aligned}\frac{dz(t)}{dt} &= \lim_{\Delta t\to 0}\frac{z(t+\Delta t)-z(t)}{\Delta t}\\ &= \lim_{\Delta t\to 0}\frac{z(t)\left[\tilde{h}\left(\frac{R}{p(t)}\right)-1\right]}{\Delta t}\\ &= z(t)h\left(\frac{R}{p(t)}\right),\end{aligned} \tag{5.11}$$

where we have defined the infinitesimal growth rate, h, via $h\left(\frac{R}{p(t)}\right) \equiv \lim_{\Delta t\to 0}\frac{\tilde{h}(R/p(t))-1}{\Delta t}$. This gives, equivalently:

$$\frac{d\ln z(t)}{dt} = h\left(\frac{R}{p(t)}\right). \tag{5.12}$$

Similarly, equation (5.7) has a continuum limit given by:

$$\frac{d\ln p(t)}{dt} = f(z(t)), \tag{5.13}$$

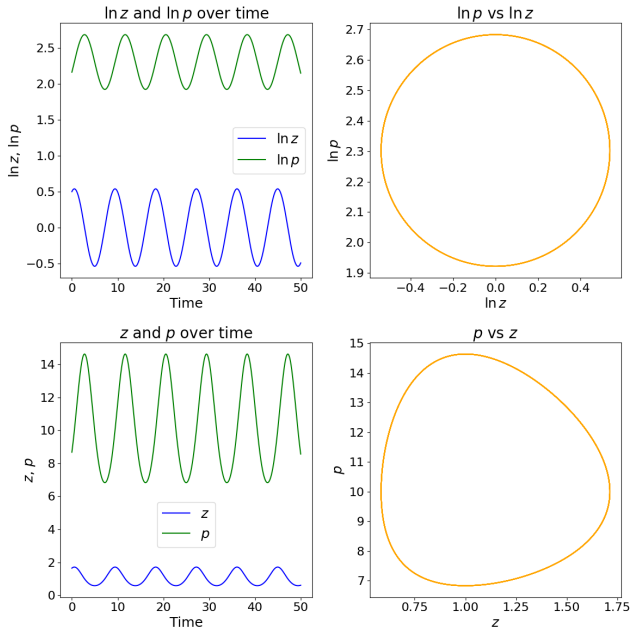


Figure 5.1: Simple harmonic motion of $\ln z$ and $\ln p$.

where, again, we define the infinitesimal growth rate f via $f\left(z(t)\right) \equiv \lim_{\Delta t \to 0} \frac{\tilde{f}(z(t))-1}{\Delta t}$.

We note that it makes biological sense that h (and $\tilde{h}$) should be a monotonically increasing concave function, since it captures the diminishing returns of increasing food per individual mother. This motivates the following specific functional form for h:

$$h\left(\frac{R}{p(t)}\right) = \frac{1}{m} \ln\left(\frac{R}{p(t)}\right), \tag{5.14}$$

where $\frac{1}{m}$ is a scaling factor representing the strength of the gestational maternal effect. Note that in what follows h can be more general and include an additive constant, although we omit it for simplicity of presentation. Substituting (5.14) into (5.12) and defining $p_0 := R$ yields:

$$\frac{d \ln z(t)}{dt} = -\frac{1}{m} \left(\ln p(t) - \ln p_0\right). \tag{5.15}$$

Now, as suggested by the covariance term in (5.6), we consider the relationship between Malthusian fitness $r(t) = \frac{d \ln p(t)}{dt}$ and trait value $z(t)$. If we assume that fitness is linearly related to the logarithm of the trait value, i.e. $f(z(t)) = k\ln(z(t)) + c$, for some constants, c and k, equation (5.13) becomes:

$$\frac{d\ln(p(t))}{dt} = k\ln(z(t)) + c. \tag{5.16}$$

The form of equations (5.15-5.16) leads to the well-studied dynamics of simple harmonic motion for the logarithms of z and p. In particular, differentiating (5.15) and setting $z_0 := e^{\frac{-c}{k}}$ we find that:

$$\frac{d^2\ln(z(t))}{dt^2} = -\frac{k}{m}(\ln(z(t)) - \ln(z_0)). \tag{5.17}$$

Note that this second order differential equation resembles the acceleration equation for a mass on a spring, where k is the analog of a spring constant. The "stiffness" of the spring, k, is related to the strength of the trait's effect on fitness. The angular frequency of the motion is determined by the product of k with the strength of the maternal effect, $\frac{1}{m}$, via $\omega = \sqrt{\frac{k}{m}}$. The explicit solution for $\ln(z)$ is given by:

$$\ln z(t) = A\cos(\omega t) + B\sin(\omega t) + \ln z_0,$$

where $A = \ln z(0) - \ln z_0$ and $B = \frac{1}{\omega}\left(\frac{d\ln z(t)}{dt}(0)\right)$.

The logarithmic population size also oscillates with simple harmonic motion according to:

$$\frac{d^2\ln(p(t))}{dt^2} = -\frac{k}{m}(\ln(p(t)) - \ln(p_0)),$$

so that:

$$\ln p(t) = \left(-\frac{kB}{\omega}\right)\cos(\omega t) + \left(\frac{kA}{\omega}\right)\sin(\omega t) + \ln p_0.$$

Equation (5.15) shows that, for a certain form of maternal effect, there is a natural relationship between the time derivative of (the logarithm of) the trait value and the logarithm of the population size. By choosing the form of the selection "force" via $f(z(t))$, we can consider different kinds of "bound motion" of which simple harmonic motion is a fundamental example.

The evolution equations (5.7,5.8) and the extension (5.9) deal with single populations but can be readily extended to multiple subpopulations, which can then be aggregated to shed light on the behavior of a full system. Consider the case in which the i^{th} subpopulation from among the n subpopulations has a value for a trait represented by $z_i(t)$, a population size $p_i(t)$, a fitness scaling k_i, and individual maternal effects m_i. Suppose, in addition, that each subpopulation follows simple harmonic motion as described above. The virial equation (5.6) applied to $\ln \mathbf{z}(t)$ is:

$$\frac{d}{dt}\mathbb{E}(\ln \mathbf{z}(t)) = \mathrm{cov}(\mathbf{r}(t), \ln \mathbf{z}(t)) + \mathbb{E}\left(\frac{d\ln \mathbf{z}(t)}{dt}\right). \tag{5.18}$$

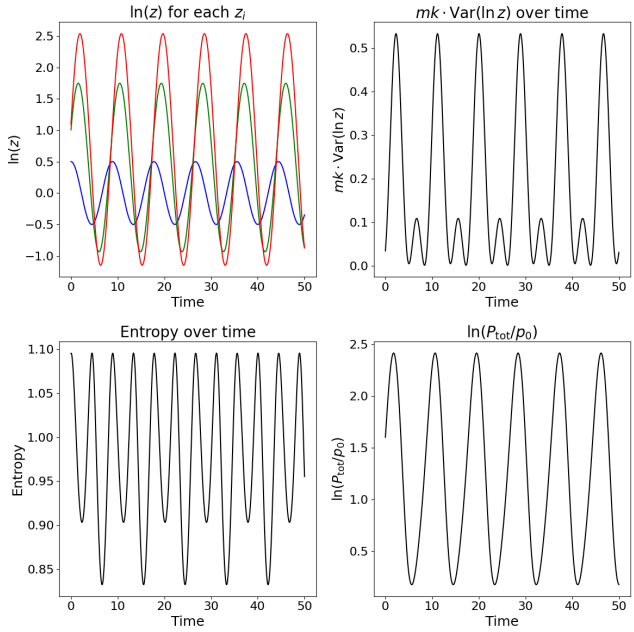


Figure 5.2: Behavior of three subpopulations and the quantities in (5.20).

Since each subpopulation performs simple harmonic motion in $\ln z(t)$, $\mathbb{E}[\ln(z)] = \sum_{i=1}^{n} q_i \ln z_i$ is bounded, so the time average of $\frac{d}{dt}\mathbb{E}(\ln \mathbf{z}(t))$ will go to zero. From equations (5.15) and (5.16) , we therefore have:

$$
\begin{aligned}
0 &= \langle \mathrm{cov}(\mathbf{k} \odot \ln \mathbf{z}(t) + \mathbf{c}, \ln \mathbf{z}(t)) \rangle_\tau \\
&\quad + \left\langle -\sum_i q_i(t) \frac{1}{m_i} (\ln(p_i(t)) - \ln(p_{0i}(t))) \right\rangle_\tau \\
&= \langle \mathrm{cov}(\mathbf{k} \odot \ln \mathbf{z}(t) + \mathbf{c}, \ln \mathbf{z}(t)) \rangle_\tau \\
&\quad + \left\langle -\sum_{i=1}^{n} \frac{1}{m_i} q_i(t) \ln (q_i(t)) \right\rangle_\tau \\
&\quad - \left\langle \sum_{i=1}^{n} \frac{1}{m_i} q_i(t) \ln \left(\frac{P_{\mathrm{tot}}(t)}{p_{0i}} \right) \right\rangle_\tau ,
\end{aligned}
\tag{5.19}
$$

where we have defined the total population, $P_{tot}(t) = \sum_i p_i(t)$. In the special case where the values k_i, m_i, c_i, and p_{0i} are the same between subpopulations and given by k, m, c, and p_0, respectively, the above simplifies to:

$$mk\,\langle \mathrm{var}(\ln \mathbf{z}(t))\rangle_\tau + \left\langle -\sum_{i=1}^{n} q_i(t)\ln(q_i(t)) \right\rangle_\tau = \left\langle \ln\left(\frac{P_{\mathrm{tot}}(t)}{p_0}\right)\right\rangle_\tau . \tag{5.20}$$

This equation describes the balance between variation in the trait between subpopulations and the entropy in the distribution of subpopulation sizes, when subpopulations are following simple harmonic motion in the way we have described. When subpopulations have different trait values, selection acts to create a non-uniform distribution of populations sizes. This illustrates the essence of the general case (5.19), where the Shannon entropy is replaced by a weighted entropy (Suhov et al., 2016).

Notably, as a simple consequence of the Fourier theorem (Rudin et al., 1964), the combination of multiple subpopulations exhibiting simple harmonic motion can give rise to almost completely general dynamics for the overall population size:

Theorem 5 *(Universality of ecological orbits)***.** Maternal effects driving simple harmonic motion in subpopulations are sufficient to generate any periodic population dynamics that satisfy the Dirichlet conditions.

5.6 Spatial Population Growth

Now, we consider a situation where a trait z affects the population growth rate of a subpopulation via competition with neighboring subpopulations. We first consider subpopulations that are spatially separated along one direction, such that each subpopulation is centered at a unique spatial coordinate, x, a distance Δx away from its two nearest neighbors. Whereas before, the growth rate of each subpopulation was given by a function, $f(z)$, of its average trait value, now the growth rate is determined by how much an 'intrinsic fitness' function, $f_{int}(z)$, exceeds that of the subpopulation's neighbors. This represents the competitive fitness advantage over the neighbors conferred by this trait. This situation would give rise to the following form for the population growth rate:

$$\frac{\partial[\ln p(x,t)]}{\partial t} = \Phi\big(f_{int}[z(x,t)] - f_{int}[z(x+\Delta x,t)]\big) + \Phi\big(f_{int}[z(x,t)] - f_{int}[z(x-\Delta x,t)]\big), \tag{5.21}$$

where $z(x,t)$ is the average value of a trait in the subpopulation at position x, at time t, and Φ is a function relating the excess intrinsic fitness of a subpopulation over its neighbor to an additive contribution to its growth rate.

Since the function Φ represents the effect of competition over resources, it should increase with the difference in intrinsic fitness between neighboring populations, and decrease with their distance apart. If we posit that Φ is linear in the fitness differential between neighboring populations, and inversely proportional to the square of the distance between populations, we have that:

$$\begin{aligned}\frac{\partial[\ln p(x,t)]}{\partial t} &= \frac{\tilde{\phi}}{(\Delta x)^2}\big(-f_{int}[z(x+\Delta x,t)] \\ &\qquad + 2f_{int}[z(x,t)] - f_{int}[z(x-\Delta x,t)]\big) \\ &= -\tilde{\phi}\frac{\partial^2[f_{int}(z(x,t))]}{\partial x^2},\end{aligned} \tag{5.22}$$

for some constant $\tilde{\phi}$, where in the second line we have taken the limit as $\Delta x \to 0$. This represents the case where the distance between subpopulations is negligible compared with the distances covered by the overall population, and the spatial dimension becomes effectively continuous.

If we choose the intrinsic fitness function, f_{int}, to be linear in the logarithm of the trait: $f_{int} = \frac{\phi}{\tilde{\phi}}\ln(z) + const$, for ϕ some new constant, then (5.22) implies:

$$r(x,t) = \frac{\partial \ln(p(x,t))}{\partial t} = -\phi\frac{\partial^2 \ln(z(x,t))}{\partial x^2}. \tag{5.23}$$

If, within each subpopulation, (i.e. at each spatial coordinate), we retain the maternal effect relation from above (5.15), the dynamics for z are described by:

$$m\frac{\partial^2[\ln z(x,t)]}{\partial t^2} = \phi\frac{\partial^2[\ln z(x,t)]}{\partial x^2}. \tag{5.24}$$

This is the one-dimensional wave equation for $\ln(z)$. The speed of the waves in this system would be given by $c^2 = \frac{\phi}{m}$. The speed of propagation of these waves increases with the strength of the maternal effect ($\frac{1}{m}$) and the size of the effect of relative intrinsic fitness on the growth rate for a subpopulation (ϕ). This formulation

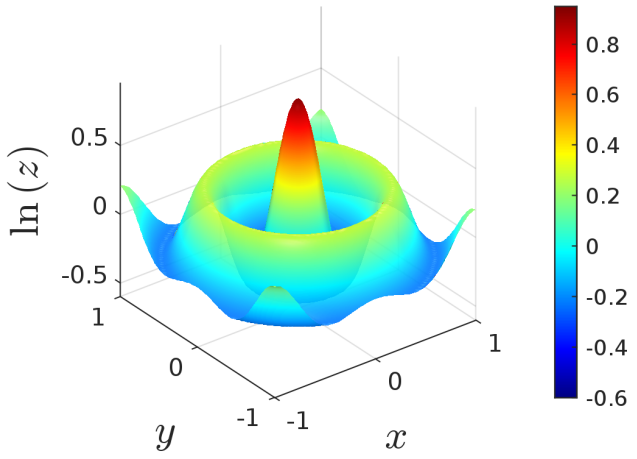


Figure 5.3: A snapshot of the evolution of $\ln(z)$ via the wave equation with two spatial coordinates (units are arbitrary).

could be straightforwardly extended to the more realistic two-dimensional case, with the substitution of the spatial gradient operator, ∇^2, in place of the one-dimensional derivative. Dynamics for $\ln(z)$ in the two-dimensional case are illustrated in Figure 5.3.

A solution to the one-dimensional homogeneous wave equation (5.24) is given by the sum of modes of the form:

$$\ln(z) = \sum_n \alpha_n \cos(k_n x - \omega_n t), \tag{5.25}$$

where α_n are constants and $\frac{\omega_n}{k_n} = \sqrt{\frac{\phi}{m}}$ for all modes, n. This corresponds to a solution for $\ln(p)$ of the form:

$$\begin{aligned} \ln(p) &= -\phi \sum_n \frac{\alpha_n}{\omega_n} k_n^2 \sin(k_n x - \omega_n t) + \ln(p_0) \\ &= -m \sum_n \alpha_n \omega_n \sin(k_n x - \omega_n t) + \ln(p_0), \end{aligned} \tag{5.26}$$

where, again, p_0 is defined as R in (5.14). An example of these solutions at a single point in time is shown in Figure 5.4.

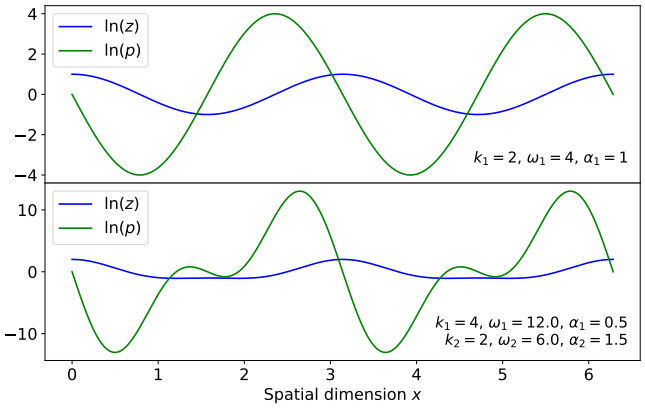


Figure 5.4: Spatial oscillations of the logarithmic trait value ($\ln(z)$) and population size ($\ln(p)$) for solutions of the one-dimensional wave equation of the form (5.25, 5.26). The top panel depicts the spatial pattern for a single mode, and the bottom panel the sum of two modes, with amplitudes and wavenumbers shown.

We then return to the virial-Price equation (5.5), again taking $\ln(z)$ as our trait. The $\mathbb{E}\left(\frac{d\ln(z(t))}{dt}\right)$ term evaluates to zero as in the simple harmonic motion example above. For the covariance term, using the explicit forms of the solutions in (5.25, 5.26), and considering the case of a single wave mode with amplitude α and frequencies k, ω, we have:

$$\begin{aligned}
&\mathrm{cov}(\boldsymbol{r}(t), \ln \boldsymbol{z}(t)) \\
&\quad = \mathrm{cov}\left(\alpha m\omega^2 \cos(k\boldsymbol{x} - \omega t), \alpha \cos(k\boldsymbol{x} - \omega t)\right) \\
&\quad = m\omega^2 \mathrm{var}(\ln \boldsymbol{z}),
\end{aligned} \tag{5.27}$$

where, as above, the expectations in the variance are taken over the population fractions, q_i, for each spatially separated subpopulation. We therefore have, analogously to equation (5.20), that:

$$\begin{aligned}
\left\langle m^2\omega^2 \mathrm{var}(\ln \boldsymbol{z})\right\rangle_\tau + \left\langle -\sum_{i=1}^{n} q_i(t)\ln(q_i(t))\right\rangle_\tau \\
= \left\langle \ln\left(\frac{P_{\mathrm{tot}}(t)}{p_0}\right)\right\rangle_\tau,
\end{aligned} \tag{5.28}$$

where P_{tot} is given by summing the subpopulation sizes at each spatial position. The second term is related to the entropy of the population distribution, as before,

and the first term is related to the amplitude and frequency of the oscillatory mode of the system, and is reminiscent of mass multiplied by the energy within a single wave. A large, high frequency wave oscillation in $\ln(z)$ implies a large variation in $\frac{d\ln(z)}{dt}$, and hence large variations in $\ln(p)$ (5.15), leading to a lower entropy in the distribution of q.

5.7 Evolutionary theory and Newtonian mechanics

The *dynamical interpretation* of evolutionary theory posits a correspondence between theories of evolution and Newtonian mechanics (Sober, 1984; Hitchcock and Velasco, 2014). In this framework, notions such as selection or mutation in biology are associated to forces in physics (Sober, 1984). For example, the dynamical interpretation of evolutionary theory posits that directional selection is a constant force that accelerates allele frequency change, whereas mutation provides a diffusive force introducing variability into evolutionary trajectories. The identical form of equations (5.5) and (5.6) can constrain such associations and clarify subsequent analogies (Table 5.1). For example, although the standard form of the virial theorem (5.3) is an energy equation, equation (5.6) is a velocity equation, which in biology translates to rates of change of biological quantities. Furthermore, rate of change of a trait or phenotype, i.e., $\frac{d}{dt}\mathbb{E}(\mathbf{z}(t))$ in equation (5.5) or the finite difference $\Delta\bar{\mathbf{z}}(t)$ in equation (5.4), corresponds to the momentum-averaged bulk velocity of a collection of physical objects. The momentum-averaged positions $\mathbb{E}(\mathbf{z}(t))$ and velocities $\frac{d}{dt}\mathbb{E}(\mathbf{z}(t))$ are discrete analogs of momentum-averaged position and momentum velocity in electromagnetism, where they emerge from the virial density in an application of the virial theorem to electromagnetic pulses (Englert, 2014). Whereas $\mathrm{cov}(\mathbf{r}(t), \mathbf{z}(t))$ is frequently referred to as the *selection term* in the Price equation (Bourrat et al., 2023), the connection to the virial theorem suggests that it is better described as a selection *rate*. Similarly, the *transmission* term $\mathbb{E}\left(\frac{d\mathbf{z}(t)}{dt}\right)$ is more accurately a transmission *rate*. Most significantly, while the dynamical interpretation typically relies on associating force to natural selection, drift, migration, or mutation (Hitchcock and Velasco, 2014), the equivalence between the virial theorem and the Price equation, suggests that force is more naturally associated to fitness. The correspondence of force to a rate of change is not surprising, since force is the rate of change of momentum. This stands more in line with the *statistical interpretation* of evolutionary theory (Dennis M Walsh, 2000; Denis M Walsh, Lewens, and Ariew, 2002; Matthen and Ariew, 2002), which, among several critiques of the dynamical interpretation, finds fault with the analogies of biological

processes such as mutation with forces in physics, arguing that the physical forces are causal in a way that processes such as selection or mutation are not (Dennis M Walsh, 2000). However, the analogy of population growth with force can be viewed as consistent with the dynamical interpretation; for example, population growth can directly affect DNA polymorphism patterns (Williamson et al., 2005). Moreover, the virial theorem in the setting of the ecological simple harmonic oscillator affirms (Hitchcock and Velasco, 2014) in noting that "natural selection turns out to be more similar to forces such as friction and elastic forces rather than the more canonical gravitation."

Table 5.1: Glossary of Terms

Variable	Biology	Physics	Dim.
i	subpopulation	object	-
$z_i(t)$	trait / phenotype	position	L
$\frac{dz_i(t)}{dt}$	evolutionary rate	velocity	LT^{-1}
$r_i(t)$	Malthusian fitness	acceleration ÷ velocity	T^{-1}
$p_i(t)$	population size	momentum	MLT^{-1}
$q_i(t)$	relative population size	relative momentum	1
$\frac{dp_i(t)}{dt}$	population growth rate	force	MLT^{-2}
$\mathbb{E}(\mathbf{z}(t))$	population-averaged trait/phenotype	momentum-averaged position	L
$\frac{d}{dt}\mathbb{E}(\mathbf{z}(t))$	group evolutionary rate	bulk momentum velocity	LT^{-1}
$\mathrm{cov}(\mathbf{r}(t), \mathbf{z}(t))$	selection rate	extrinsic momentum velocity	LT^{-1}
$\mathbb{E}\left(\frac{d\mathbf{z}(t)}{dt}\right)$	transmission rate	intrinsic momentum velocity	LT^{-1}

In particular, considerations of the analogies between biology and physics via the virial theorem led us to generalize the work of (L. Ginzburg and Colyvan, 2004) and to derive the ecological simple harmonic oscillator, which to our knowledge is the first example of such an oscillator that emerges solely from maternal effects and does not require a predator-prey or other more sophisticated model. The extension of (5.8) to (5.9) is interesting in its own right, and should be fruitful to develop in future work. Moreover, Theorem 5 shows that subpopulations subject to distinct maternal effects can generate arbitrarily complex population dynamics, thereby affirming the main thesis of (L. Ginzburg and Colyvan, 2004). We have further extended Ginzburg and Colyvan's work (L. Ginzburg and Colyvan, 2004) by considering

an ecological model for spatial population growth. In this model, competition between neighboring populations, combined with maternal effects, gives a spatial wave equation for trait evolution (5.24). In both the simple harmonic motion and spatial models, we have shown that simple ecological mechanisms can reproduce the fundamental modes of motion in physics. Whilst the similarity between the ecological equations and physical equations of motion (e.g. the resemblence of (5.17) to a mass on a spring) is optical only, the usefulness of such equations in deriving fundamental modes of ecological motion highlights the utility of alternative physical-biological analogies to those chosen in Table 5.1.

Ultimately, analogies between the Price equation and the virial theorem point towards potentially productive directions for exploration in both biology and physics. The statistical framing of the virial theorem in (5.6) highlights phenomena that may have been overlooked in the physics realm. For example, the first term on the right-hand side of (5.6), namely $\mathrm{cov}(\mathbf{r}(t), \mathbf{z}(t))$, can be understood to quantify the extent of the Yule-Simpson effect (Pearson, Lee, and Bramley-Moore, 1899; Yule, 1903; Simpson, 1951), which describes a situation where within-group trends can be reversed upon averaging. In biology, the Price equation has the potential to be used more widely as a tool. Although it has been hailed as a unifying framework for researchers (Luque, 2017), one that "can serve as a heuristic principle to formulate and systematize different theories and models in evolutionary biology" (Luque and Baravalle, 2021), the emphasis on its use has been more oriented toward understanding how it generalizes specific equations, rather than applying it for biological discovery. For example, the Price equation can be used to derive the breeder's equation (Bijma, 2020; Zhang and Hill, 2010), Fisher's fundamental theorem (Queller, 2017; Price, 1972), the high rare mutation large effect "house of cards approximation" regime for genetic variance at mutation-selection balance (Zhang and Hill, 2010; Turelli, 1984), and many other formulas and identities in genetics (Zhang and Hill, 2010; Rice, 2004). However, it has been referred to as a tautology and a vacuous statement without application. In (Van Veelen et al., 2012) the Price equation is described as a theorem that establishes that "If the left-hand side is computed as suggested in (Price et al., 1970), and the right-hand side too, then they are equal." This critique of the Price equation, namely that it does not and cannot serve as a *tool*, stands in contradiction to evidence from physics, where the mathematically equivalent virial theorem has been understood as a powerful tool since its use to discover dark matter in 1933 (Zwicky, 1933). In fact, the Price equation has already been used to deduce the existence of specific environmental

effects in evolutionary biology. Grant and Grant (P. R. Grant and B. R. Grant, 1995) conducted an experiment on Daphne Major, a Galápagos island, capturing and labeling mature finches and their offspring. They were able to measure the change in mean trait value between successive generations. They also compared the trait measurements in adult populations before and after selection events to determine selection rates. They used the breeder's equation, a restricted form of the Price equation, to predict the measured mean trait change from the measured selection rates and known heritabilities. Specifically, they start with the breeder's equation, $R = h^2 s$, where R, the response, is equivalent to the change in the mean, $\Delta\bar{z}$. The heritablity, h^2, and selection strength, s, combine to give a linear approximation to the covariance term in the Price equation. They then expand to a multivariate version of this term, which takes into account the correlation of different traits. For the six traits they considered, the equations, $i = 1, \ldots, 6$, become:

$$\Delta\bar{z}_i = \beta_i h_i^2 + \sum_{j \neq i} \beta_j h_j h_i r_{ij}, \tag{5.29}$$

where the β_i are direct selection coefficients on each trait, i, and the r_{ji} represent the genetic correlations between traits i and j. Note that the final term of the Price equation, $\propto \mathbb{E}(\boldsymbol{w} \odot \Delta \boldsymbol{z})$, is ignored in this framework. Where there was a gap between their predictions and the measured mean trait change, they were able to identify environmental effects which made this non-selective term of the Price equation significant. They predicted a restricted food supply which affected the growth of the adult finches, which corresponded to a drought period. The gap identified by Grant and Grant is analogous not only to the way the virial theorem is used in physics and astrophysics (Marc and McMillan, 1985), but also to the missing heritability in human genetics where heritability for complex traits estimated from twin studies do not match heritability estimates derived from genome-wide association studies. In other words, the equivalence we have demonstrated between the Price equation and the virial theorem shows that the description of missing heritability as dark matter (Manolio et al., 2009) may be understood to be more than just an informal analogy between mysteries in genetics and astronomy.

Author Contributions and Acknowledgements

SL studied the virial theorem while participating in the "Introduction to Astrophysics" cluster in the COS- MOS summer program held at UC Irvine from July 9,

2023 to August 4, 2023. Specifically, she used the virial theorem to repeat Zwicky's Coma cluster mass estimates using modern measurements of velocity dispersion and galaxy positions. LP learned of the virial theorem from SL and, in discussing its proof with SL, realized that it must be related to the Price equation. SL and LP explored the applications and implications of the connection. CF identified and developed the connection to ecological orbits and simple harmonic motion via the maternal effect. LP drafted the initial manuscript; both SL and LP edited the first version posted on arXiv [24]. CF, SL, and LP edited the final version and code [12]. The authors are ordered alphabetically.

SL thanks Manoj Kaplinghat and Gopolang Mohlabeng who led the "Introduction to Astrophysics" cluster (cluster 4) at the 2023 UC Irvine COSMOS program. LP relied in part on notes about Fisher's theorem of natural selection from his April 22, 2008 lecture for UC Berkeley course Math 239: Discrete Mathematics for the Life Sciences that were transcribed and edited by Cynthia Vinzant and Caroline Uhler. LP thanks Junhyong Kim for suggesting a possible connection of this work to the ideas in the book Ecological Orbits: How Planets Move and Populations Grow by Lev Ginzburg and Mark Colyvan, a discussion that led to CF's extension of the work described in the book and the sections on simple harmonic motion and spatial population growth. CF was partially supported by funding from Charles Trimble.

References

Alazard, Thomas and Claude Zuily (2023). "Virial theorems and equipartition of energy for water-waves". In: *arXiv preprint arXiv:2304.07872*.

Andersen, Esben S (2004). "Population thinking, Price's equation and the analysis of economic evolution". In: *Evolutionary and Institutional Economics Review* 1, pp. 127–148.

Bian, Jiang-Hui et al. (2015). "Maternal effects and population regulation: maternal density-induced reproduction suppression impairs offspring capacity in response to immediate environment in root voles Microtus oeconomus". In: *Journal of Animal Ecology* 84.2, pp. 326–336. DOI: https://doi.org/10.1111/1365-2656.12307. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2656.12307. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2656.12307.

Bijma, P (2020). "The Price equation as a bridge between animal breeding and evolutionary biology". In: *Philosophical Transactions of the Royal Society B* 375.1797, p. 20190360.

Binney, James and Michael Merrifield (1998). *Galactic Astronomy*.

Bourrat, Pierrick et al. (2023). “What is the price of using the Price equation in ecology?” In: *Oikos*, e10024.

Boycott, Arthur Edwin et al. (1931). “II. The inheritance of sinistrality in *Limnæa peregra* (Mollusca, Pulmonata)”. In: *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character* 219, pp. 51–131. DOI: 10.1098/rstb.1931.0002. URL: http://doi.org/10.1098/rstb.1931.0002.

Clausius, Rudolf (1870). “XVI. On a mechanical theorem applicable to heat”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 40.265, pp. 122–127.

Cresson, Jacky, Laurent Nottale, and Thierry Lehner (2021). “Stochastic modification of Newtonian dynamics and induced potential—Application to spiral galaxies and the dark potential”. In: *Journal of Mathematical Physics* 62.7.

Einstein, Albert (1922). “Die Grundlage der allgemeinen Relativitätstheorie”. In: *Annalen der Physik* 49.7.

Ellner, Stephen P, Monica A Geber, and Nelson G Hairston Jr (2011). “Does rapid evolution matter? Measuring the rate of contemporary evolution and its impacts on ecological dynamics”. In: *Ecology letters* 14.6, pp. 603–614.

Englert, Berthold-Georg (2014). *lectures on classical electrodynamics*. World Scientific Publishing Company.

Fisher, Ronald A (1930). *The genetical theory of natural selection*. Clarendon Press, Oxford, Valorium edition, Bennett JH (Editor), 1999, Oxford University Press, Oxford, UK].

Fowler, Ralph H (1929). *Statistical Mechanics*. Cambridge University Press, Cambridge, UK.

Fox, Charles W., Meghna S. Thakar, and Timothy A. Mousseau (1999). “The evolutionary genetics of an adaptive maternal effect: egg size plasticity in a seed beetle”. In: *Evolution* 53.2, pp. 552–560. DOI: 10.1111/j.1558-5646.1999.tb03790.x. URL: https://doi.org/10.1111/j.1558-5646.1999.tb03790.x.

Frank, Steven A (1997). “The Price equation, Fisher’s fundamental theorem, kin selection, and causal analysis”. In: *Evolution* 51.6, pp. 1712–1729.

Frank, Steven A and Montgomery Slatkin (1990). “The distribution of allelic effects under mutation and selection”. In: *Genetics Research* 55.2, pp. 111–117.

Georgescu, Vladimir and Christian Gérard (1999). “On the virial theorem in quantum mechanics”. In: *Communications in Mathematical Physics* 208.2, pp. 275–281.

Ginzburg, Lev and Mark Colyvan (2004). *Ecological orbits: How planets move and populations grow*. Oxford University Press.

Ginzburg, Lev R. and Dale E. Taneyhill (1994). "Population Cycles of Forest Lepidoptera: A Maternal Effect Hypothesis". In: *Journal of Animal Ecology* 63.1, pp. 79–92. ISSN: 00218790, 13652656. URL: http://www.jstor.org/stable/5585 (visited on 12/06/2024).

Grant, Peter R. and B. Rosemary Grant (Apr. 1995). "Predicting microevolutionary responses to directional selection on heritable variation". In: *Evolution* 49.2, pp. 241–251. DOI: 10.1111/j.1558-5646.1995.tb02236.x.

Hanson, Mervin P (1995). "The virial theorem, perfect gases, and the second virial coefficient". In: *Journal of chemical education* 72.4, p. 311.

Harding, JE (Feb. 2001). "The nutritional basis of the fetal origins of adult disease". In: *International Journal of Epidemiology* 30.1, pp. 15–23. ISSN: 0300-5771. DOI: 10.1093/ije/30.1.15. eprint: https://academic.oup.com/ije/article-pdf/30/1/15/18478235/300015.pdf. URL: https://doi.org/10.1093/ije/30.1.15.

Harman, Oren (2011). *The price of altruism: George Price and the search for the origins of kindness*. WW Norton & Company, New York, NY.

Hitchcock, Christopher and Joel D Velasco (2014). "Evolutionary and Newtonian forces". In: *Ergo* 1.2, p. 39.

L'Hôpital, Guillame de (1696). *Analyse Des Infiniment Petits Pour L'Intelligence Des Lignes Courbes*. Chez Montalant, Paris, France.

Luque, Victor J (2017). "One equation to rule them all: a philosophical analysis of the Price equation". In: *Biology & Philosophy* 32.1, pp. 97–125.

Luque, Victor J and Lorenzo Baravalle (2021). "The mirror of physics: on how the Price equation can unify evolutionary biology". In: *Synthese* 199.5-6, pp. 12439–12462.

Manolio, Teri A et al. (2009). "Finding the missing heritability of complex diseases". In: *Nature* 461.7265, pp. 747–753.

Marc, Guilhem and William G McMillan (1985). "The Virial Theorem". In: *Advances in Chemical Physics*, pp. 209–361.

Matthen, Mohan and André Ariew (2002). "Two ways of thinking about fitness and natural selection". In: *The Journal of Philosophy* 99.2, pp. 55–83.

Merritt, David (1987). "The distribution of dark matter in the Coma cluster". In: *The Astrophysical Journal* 313, pp. 121–135.

Mukhopadhyay, Banibrata, Arnab Sarkar, and Christopher A. Tout (2025). "Modified Virial Theorem for Highly Magnetized White Dwarfs". In: (). (Visited on 09/12/2025).

Nourmohammad, Armita, Torsten Held, and Michael Lässig (2013). "Universality and predictability in molecular quantitative genetics". In: *Current opinion in genetics & development* 23.6, pp. 684–693.

Oguz, Hasan N and Andrea Prosperetti (1990). "A generalization of the impulse and virial theorems with an application to bubble oscillations". In: *Journal of Fluid Mechanics* 218, pp. 143–162.

Pearson, Karl, Alice Lee, and Leslie Bramley-Moore (1899). "VI. Mathematical contributions to the theory of evolution.—VI. Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 6.192, pp. 257–330.

Podio-Guidugli, Paolo (2019). "The virial theorem: A pocket primer". In: *Journal of Elasticity* 137.2, pp. 219–235.

Price, George R et al. (1970). "Selection and covariance." In: *Nature* 227, pp. 520–521.

Price, George R (1972). "Fisher's 'fundamental theorem' made clear". In: *Annals of human genetics* 36.2, pp. 129–140.

Queller, David C (2017). "Fundamental theorems of evolution". In: *The American Naturalist* 189.4, pp. 345–353.

Rayleigh, Lord (1905). "XLII. On the momentum and pressure of gaseous vibrations, and on the connexion with the virial theorem". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 10.57, pp. 364–374.

Reznick, D., H. Callahan, and R. Llauredo (2015). "Maternal Effects on Offspring Quality in Poeciliid Fishes". In: *American Zoologist* 36.2, pp. 147–156. DOI: 10.1093/icb/36.2.147.

Rice, Sean H (2004). *Evolutionary theory: mathematical and conceptual foundations*. Sinauer Associates, Sunderland MA.

– (2008). "A stochastic version of the Price equation reveals the interplay of deterministic and stochastic processes in evolution". In: *BMC evolutionary biology* 8, pp. 1–16.

Roseboom, Tessa, Susanne de Rooij, and Rebecca Painter (Aug. 2006). "The Dutch famine and its long-term consequences for adult health". In: *Early Human Development* 82.8. Epub 2006 Jul 28, pp. 485–491. DOI: 10.1016/j.earlhumdev.2006.07.001. URL: https://doi.org/10.1016/j.earlhumdev.2006.07.001.

Rudin, Walter et al. (1964). *Principles of mathematical analysis*. Vol. 3. McGraw-hill New York.

Shore, Steven N (2012). *An introduction to astrophysical hydrodynamics*. Academic Press, Cambridge, MA.

Simpson, Edward H (1951). "The interpretation of interaction in contingency tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 13.2, pp. 238–241.

Sober, Elliott (1984). *The nature of selection*. MIT Press, Cambridge, MA.

Suhov, Yuri et al. (2016). "Basic inequalities for weighted entropies". In: *Aequationes mathematicae* 90, pp. 817–848.

The, Lih S and Simon D M White (1986). "The mass of the Coma cluster". In: *Astronomical Journal* 92.6, pp. 1248–1253.

Thorne, E Tom, Ron E Dean, and William G Hepworth (1976). "Nutrition during gestation in relation to successful reproduction in elk". In: *The Journal of Wildlife Management*, pp. 330–335.

Turelli, Michael (1984). "Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle". In: *Theoretical population biology* 25.2, pp. 138–193.

Van Veelen, Matthijs et al. (2012). "Group selection and inclusive fitness are not equivalent; the Price equation vs. models and statistics". In: *Journal of theoretical biology* 299, pp. 64–80.

Wagner, Günter P (2010). "The measurement theory of fitness". In: *Evolution* 64.5, pp. 1358–1376.

Walsh, Denis M, Tim Lewens, and André Ariew (2002). "The trials of life: Natural selection and random drift". In: *Philosophy of Science* 69.3, pp. 452–473.

Walsh, Dennis M (2000). "Chasing shadows: natural selection and adaptation". In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 31.1, pp. 135–153.

Williamson, Scott H et al. (2005). "Simultaneous inference of selection and population growth from patterns of variation in the human genome". In: *Proceedings of the National Academy of Sciences* 102.22, pp. 7882–7887.

Wolf, Jason B. and Michael J. Wade (Apr. 2009). "What are maternal effects (and what are they not)?" eng. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1520. Place: London Publisher: Royal Society, pp. 1107–1115. ISSN: 0962-8436. DOI: 10.1098/rstb.2008.0238.

Yule, G Udny (1903). "Notes on the theory of association of attributes in statistics". In: *Biometrika* 2.2, pp. 121–134.

Zhang, Xu-Sheng and William G Hill (2010). "Change and maintenance of variation in quantitative traits in the context of the Price equation". In: *Theoretical population biology* 77.1, pp. 14–22.

Zwicky, Fritz (1933). "Die Rotverschiebung von extragalaktischen Nebeln". In: *Helvetica Physica Acta, Vol. 6, p. 110-127* 6, pp. 110–127.

Chapter 6

FUTURE DIRECTIONS

Overall, I hope that this thesis exemplifies the utility of using inspiration from physics in approaching biological questions. Whilst the different kinds of complexity encountered in biological processes render description in terms of fundamental laws more difficult, there is still value in considering how much can be explained via simple, foundational models. I have used the most basic concepts from condensed matter (C. Felce, G. Gorin, and L. Pachter, 2024), and mechanics (Catherine Felce, Liorsdóttir, and Lior Pachter, 2025), to suggest new models for biological data and ecological cycles. I hope that both of these approaches will be expanded upon, as their value lies partly in the existing physics machinery which has been developed to characterize systems obeying these dynamics. For example, the linear Ising-like model introduced for chromatin in Chapter 2 could be extended to a non-linear model of chromatin connectivity informed by Hi-C data. Ising model behavior for general graph structures has been widely studied in physics (Dorogovtsev, Goltsev, and J. F. F. Mendes, 2002; Dembo and Montanari, 2010), potentially allowing us to leverage existing theory to reach new insights.

I have built on the Pachter lab's mission to expand the purview of principled biophysical models to explain new datatypes (Catherine Felce, Fang, and Lior Pachter, 2025). I hope that this represents a contribution to building on our understanding of biology 'from the ground up'. Whilst this is not the approach that, for example, gives the most immediate clinical applications, I think that it should be valued from the perspective of pure understanding. A biologist's mindset towards transcription tends to focus on the causes and effects of transcription *in flux*. Their interest lies in cases where transcriptional or translational rates change due to regulation, rather than when stochastic noise dominates in basal transcription. In contrast, physics reduces processes down to the normal, with some of the greatest advances coming from noting how objects behave *without* external intervention (Galilei, 1957). Along with other biologists (Milo and Phillips, 2015), I believe in the value of 'recognizing regularities' in these fundamental biological processes, and precisely quantifying the typical transcriptional and translational function of cells using biophysical modeling.

Once the 'normal' behavior of a system has been precisely quantified, it is then that we can use anomalous behaviors to identify new biology. This methodology is the standard in physics, for example in astrophysics. As discussed in Chapter 5, deviations from the Virial theorem were used to discover dark matter. Deviations from expected spin-orbit dynamics in binary star systems can be used to infer the presence of third bodies (Catherine Felce and Fuller, 2023; Fuller and Catherine Felce, 2023). Along with Ginzburg and Colyvan (Ginzburg and Colyvan, 2004), I am optimistic that some of these approaches can be expanded to biological and ecological systems. If an inertial model with maternal effects can explain population cycles within a single species, deviations from these trajectories could help to recognize more complicated interactions. In line with this, extensions of the work in Chapter 5 include incorporating maternal effects into existing predator-prey graph models (Çelik et al., 2025).

One potentially powerful extension of the work described in this thesis lies in the combination of biophysically modeled parameters for protein expression, and phylogenetic comparative methods. As noted by (Dimayacyac et al., 2023b), whilst many PCM studies have effectively used mRNA levels as a proxy for protein expression, protein levels are closer to phenotype and the coupling of the two modalities can be complex. This idea motivated (Cope, Schraiber, and Pennell, 2024) in their study of the coevolution of protein and RNA levels. The framework introduced in (Catherine Felce, Cope, et al., 2025) could therefore be combined with biophysical parameters obtained from the inference procedure introduced in (Catherine Felce, Fang, and Lior Pachter, 2025), to corroborate and extend the insights from (Cope, Schraiber, and Pennell, 2024).

The models introduced in this thesis could also be extended to leverage more powerful machine learning techniques. Neural networks for PDE solving (Li et al., 2021) could be used to approximate the solutions of the extended chemical master equations described in Chapters 2 and 3, giving faster inference. Neural networks have also been shown to be effective at choosing between evolutionary models and constructing phylogenies, improving on the speed of maximum likelihood methods (Kulikov, Derakhshandeh, and Mayer, 2024). These and other techniques could be straightforwardly applied to the biophysical phylogenetics framework outlined in Chapter 4. With increased interest in model-based machine learning, the bayesian hierarchical approach described in Chapter 4 is well-suited for generalization to more complicated graphical models (Bishop, 2013). This is just one example

of how principled biophysical modeling, using inspiration and methodology from physics, allows us to create foundational frameworks which can easily be built upon to incorporate new techniques and harness new data types. These biophysical models balance interpretability and statistical power, and bring us closer to a more quantitative understanding of biology.

BIBLIOGRAPHY

Abbott, B. P. et al. (Oct. 2017). "GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral". In: *Phys. Rev. Lett.* 119 (16), p. 161101. DOI: 10.1103/PhysRevLett.119.161101. URL: https://link.aps.org/doi/10.1103/PhysRevLett.119.161101.

Adams, Dean C. (Nov. 2012). "Comparing Evolutionary Rates for Different Phenotypic Traits on a Phylogeny Using Likelihood". In: *Systematic Biology* 62.2, pp. 181–192. ISSN: 1063-5157. DOI: 10.1093/sysbio/sys083. eprint: https://academic.oup.com/sysbio/article-pdf/62/2/181/24577504/sys083.pdf.

Bergelson, Joy et al. (June 2021). "Functional Biology in Its Natural Context: A Search for Emergent Simplicity". In: *eLife* 10, e67646. ISSN: 2050-084X. DOI: 10.7554/eLife.67646.

Bishop, Christopher M. (Feb. 2013). "Model-Based Machine Learning". In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 371.1984, p. 20120222. ISSN: 1364-503X. DOI: 10.1098/rsta.2012.0222. (Visited on 11/05/2025).

Blekhman, Ran et al. (Nov. 2008). "Gene Regulation in Primates Evolves under Tissue-Specific Selection Pressures". In: *PLOS Genetics* 4.11, pp. 1–13. DOI: 10.1371/journal.pgen.1000271.

Bohrer, Christopher H. and Elijah Roberts (Feb. 2016). "A Biophysical Model of Supercoiling Dependent Transcription Predicts a Structural Aspect to Gene Regulation". In: *BMC Biophysics* 9.1, p. 2. ISSN: 2046-1682. DOI: 10.1186/s13628-016-0027-0.

Bullard, James H. et al. (Feb. 2010). "Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments". In: *BMC Bioinformatics* 11.1, p. 94. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-94.

Butler, Marguerite A. and Aaron A. King (2004). "Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution." In: *The American Naturalist* 164.6, pp. 683–695. DOI: 10.1086/426002. eprint: https://doi.org/10.1086/426002.

Cao, Zhixing, Yiling Wang, and Ramon Grima (Jan. 2025). "Deterministic Patterns in Single-Cell Transcriptomic Data". In: *npj Systems Biology and Applications* 11.1, p. 6. ISSN: 2056-7189. DOI: 10.1038/s41540-025-00490-5.

Çelik, Türkü Özlüm et al. (2025). *Strata of Ecological Coexistence via Grassmannians*. arXiv: 2509.00165 [math.AG]. URL: https://arxiv.org/abs/2509.00165.

Chari and Pachter (2024). "Biophysically interpretable inference of cell types from multimodal sequencing data". In: *Nature Computational Science* 4.9, pp. 677–689. DOI: 10.1038/s43588-024-00689-2.

Chari, Tara and Lior Pachter (Aug. 2023). "The Specious Art of Single-Cell Genomics". In: *PLOS Computational Biology* 19.8, e1011288. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011288. (Visited on 11/05/2025).

Cope, Alexander L., Joshua G. Schraiber, and Matthew Pennell (2024). "Macroevolutionary divergence of gene expression shaped by selection on protein abundance". In: *bioRxiv*. Preprint. DOI: 10.1101/2024.07.08.602411. URL: https://doi.org/10.1101/2024.07.08.602411.

Cusanovich, Darren A. et al. (May 2015). "Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing". In: *Science* 348.6237, pp. 910–914. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aab1601.

Dada, Joseph O. and Pedro Mendes (Feb. 2011). "Multi-Scale Modelling and Simulation in Systems Biology". In: *Integrative Biology* 3.2, pp. 86–96. ISSN: 1757-9708. DOI: 10.1039/c0ib00075b. (Visited on 11/08/2025).

Dal Molin, Alessandra, Giacomo Baruzzo, and Barbara Di Camillo (2017). "Single-Cell RNA-Sequencing: Assessment of Differential Expression Analysis Methods". In: *Frontiers in Genetics* 8, p. 62. ISSN: 1664-8021. DOI: 10.3389/fgene.2017.00062.

Das, Samarendra, Anil Rai, and Shesh N. Rai (July 2022). "Differential Expression Analysis of Single-Cell RNA-Seq Data: Current Statistical Approaches and Outstanding Challenges". In: *Entropy (Basel, Switzerland)* 24.7, p. 995. ISSN: 1099-4300. DOI: 10.3390/e24070995.

Dembo, Amir and Andrea Montanari (Apr. 2010). "Ising Models on Locally Tree-like Graphs". In: *The Annals of Applied Probability* 20.2, pp. 565–592. ISSN: 1050-5164, 2168-8737. DOI: 10.1214/09-AAP627. (Visited on 11/08/2025).

Dhar, Pawan K. and Alessandro Giuliani (2010). "Laws of biology: why so few?" In: *Systems and Synthetic Biology* 4.1, pp. 7–13. DOI: 10.1007/s11693-009-9049-0.

Dibán, María José and Luis Felipe Hinojosa (Jan. 2024). "Testing the Tropical Niche Conservatism Hypothesis: Climatic Niche Evolution of Escallonia Mutis Ex L. F. (Escalloniaceae)". In: *Plants (Basel, Switzerland)* 13.1, p. 133. ISSN: 2223-7747. DOI: 10.3390/plants13010133.

Dimayacyac, Jose Rafael et al. (Dec. 2023a). "Evaluating the Performance of Widely Used Phylogenetic Models for Gene Expression Evolution". In: *Genome Biology and Evolution* 15.12, evad211. DOI: 10.1093/gbe/evad211. URL: https://doi.org/10.1093/gbe/evad211.

Dimayacyac, Jose Rafael et al. (Nov. 2023b). "Evaluating the Performance of Widely Used Phylogenetic Models for Gene Expression Evolution". In: *Genome Biology and Evolution* 15.12, evad211. ISSN: 1759-6653. DOI: 10.1093/gbe/evad211. eprint: https://academic.oup.com/gbe/article-pdf/15/12/evad211/54912015/evad211.pdf.

Dorogovtsev, S. N., A. V. Goltsev, and J. F. F. Mendes (July 2002). "Ising model on networks with an arbitrary distribution of connections". In: *Phys. Rev. E* 66 (1), p. 016104. DOI: 10.1103/PhysRevE.66.016104. URL: https://link.aps.org/doi/10.1103/PhysRevE.66.016104.

Ellis, George F. R. and Jonathan Kopel (Jan. 2019). "The Dynamical Emergence of Biology From Physics: Branching Causation via Biomolecules". In: *Frontiers in Physiology* 9, p. 1966. ISSN: 1664-042X. DOI: 10.3389/fphys.2018.01966. (Visited on 11/08/2025).

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74. DOI: 10.1038/nature11247.

Falkowski, Adam (2023). "Lectures on SMEFT". In: *European Physical Journal C* 83.656. DOI: 10.1140/epjc/s10052-023-11821-3. URL: https://doi.org/10.1140/epjc/s10052-023-11821-3.

Felce, C., G. Gorin, and L. Pachter (Dec. 2024). "Biophysical model for joint analysis of chromatin and RNA sequencing data". In: *Phys. Rev. E* 110 (6), p. 064405. DOI: 10.1103/PhysRevE.110.064405. URL: https://link.aps.org/doi/10.1103/PhysRevE.110.064405.

Felce, Catherine, Alexander L. Cope, et al. (2025). "Biophysical Constraints on mRNA Decay Rates Shape Macroevolutionary Divergence in Steady-State Abundances". In: *bioRxiv*. DOI: 10.1101/2025.11.24.690267. eprint: 2025.11.24.690267. URL: https://doi.org/10.1101/2025.11.24.690267.

Felce, Catherine, Meichen Fang, and Lior Pachter (2025). "Joint Biophysical Modeling of Paired Single-Cell RNA and Protein Measurements". In: *bioRxiv*. DOI: 10.1101/2025.11.14.688548. eprint: 2025.11.14.688548. URL: https://doi.org/10.1101/2025.11.14.688548.

Felce, Catherine and Jim Fuller (Oct. 2023). "Slowly Rotating Close Binary Stars in Cassini States". In: *Monthly Notices of the Royal Astronomical Society* 526.4, pp. 6168–6180. ISSN: 0035-8711. DOI: 10.1093/mnras/stad3053. eprint: https://academic.oup.com/mnras/article-pdf/526/4/6168/52633417/stad3053.pdf.

Felce, Catherine, Steinunn Liorsdóttir, and Lior Pachter (Nov. 2025). "Analogies between the virial theorem and the Price equation". In: *Phys. Rev. E* 112 (5), p. 054139. DOI: 10.1103/r8bd-lhm3. URL: https://link.aps.org/doi/10.1103/r8bd-lhm3.

Felsenstein, Joseph (1985). "Phylogenies and the Comparative Method". In: *The American Naturalist* 125.1, pp. 1–15. ISSN: 00030147, 15375323. URL: http://www.jstor.org/stable/2461605 (visited on 11/04/2025).

– (1988). "Phylogenies and Quantitative Characters". In: *Annual Review of Ecology and Systematics* 19, pp. 445–471. DOI: 10.1146/annurev.es.19.110188.002305. URL: https://www.jstor.org/stable/2097162.

Fuller, Jim and Catherine Felce (Oct. 2023). "Super Slowly Spinning Stars in Close Binaries". In: *Monthly Notices of the Royal Astronomical Society: Letters* 527.1, pp. L103–L109. ISSN: 1745-3925. DOI: 10.1093/mnrasl/slad150. eprint: https://academic.oup.com/mnrasl/article-pdf/527/1/L103/54610279/slad150.pdf.

Galilei, Galileo (1957). "On Motion". In: *Discoveries and Opinions of Galileo*. Ed. and trans. by Stillman Drake. New York: Doubleday, pp. 1–109.

Gates, David M. (1975). "Introduction: Biophysical Ecology". In: *Perspectives of Biophysical Ecology*. Ed. by David M. Gates and Rudolf B. Schmerl. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–28. ISBN: 978-3-642-87810-7. DOI: 10.1007/978-3-642-87810-7_1.

Ginzburg, Lev and Mark Colyvan (2004). *Ecological orbits: How planets move and populations grow*. Oxford University Press.

Gorin, Gennady, Meichen Fang, et al. (2022). "RNA velocity unraveled". In: *PLoS Computational Biology* 18.9. Version 2, published September 12, 2022, e1010492. DOI: 10.1371/journal.pcbi.1010492. URL: https://doi.org/10.1371/journal.pcbi.1010492.

Gorin, Gennady, John J Vastola, et al. (2022). "Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments". In: *Nature Communications* 13.1, p. 7620. DOI: 10.1038/s41467-022-34857-7.

Grima, Ramon and Pierre-Marie Esmenjaud (2024). "Quantifying and Correcting Bias in Transcriptional Parameter Inference from Single-Cell Data". In: *Biophysical Journal* 123.1, pp. 4–30. ISSN: 0006-3495. DOI: 10.1016/j.bpj.2023.10.021.

GTEx Consortium (2015). "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". In: *Science* 348.6235, pp. 648–660. DOI: 10.1126/science.1262110.

Hadfield, J. D. and S. Nakagawa (Mar. 2010). "General Quantitative Genetic Methods for Comparative Biology: Phylogenies, Taxonomies and Multi-trait Models for Continuous and Categorical Characters". In: *Journal of Evolutionary Biology* 23.3, pp. 494–508. ISSN: 1010-061X. DOI: 10.1111/j.1420-9101.2009.01915.x. eprint: https://academic.oup.com/jeb/article-pdf/23/3/494/54530878/jevbio0494.pdf.

Hill, Mark S., Pétra Vande Zande, and Patricia J. Wittkopp (Apr. 2021). "Molecular and Evolutionary Processes Generating Variation in Gene Expression". In: *Nature Reviews. Genetics* 22.4, pp. 203–215. ISSN: 1471-0064. DOI: 10.1038/s41576-020-00304-w.

Huang, Jinghan, Phillip S.C. Yam, and Nelson L.S. Tang (2025). "'Trans-differentiation of Neutrophils from Plasmablast' Is an Artefact Caused by over-Reliance on Machine Algorithms in Single Cell RNA Sequencing Analysis: Lesson Learnt and Steps Ahead". In: *bioRxiv : the preprint server for biology*. DOI: 10.1101/2025.02.07.636761. eprint: https://www.biorxiv.org/content/early/2025/02/07/2025.02.07.636761.full.pdf.

Justus, James (2013). "Philosophical Issues in Ecology". In: *The Philosophy of Biology*. Ed. by Kostas Kampourakis. Vol. 1. Dordrecht: Springer Netherlands, pp. 343–371. ISBN: 978-94-007-6536-8 978-94-007-6537-5. DOI: 10.1007/978-94-007-6537-5_17. (Visited on 11/09/2025).

Kim, Jong Kyoung and John C. Marioni (Jan. 2013). "Inferring the Kinetics of Stochastic Gene Expression from Single-Cell RNA-sequencing Data". In: *Genome Biology* 14.1, R7. ISSN: 1474-760X. DOI: 10.1186/gb-2013-14-1-r7.

Klemm, Sandy L., Zohar Shipony, and William J. Greenleaf (Jan. 2019). "Chromatin accessibility and the regulatory epigenome". In: *Nature Reviews Genetics* 20.4, pp. 207–220. DOI: 10.1038/s41576-018-0089-8. URL: https://doi.org/10.1038/s41576-018-0089-8.

Kulikov, Nikita, Fatemeh Derakhshandeh, and Christoph Mayer (2024). "Machine Learning Can Be as Good as Maximum Likelihood When Reconstructing Phylogenetic Trees and Determining the Best Evolutionary Model on Four Taxon Alignments". In: *Molecular Phylogenetics and Evolution* 200, p. 108181. ISSN: 1055-7903. DOI: 10.1016/j.ympev.2024.108181.

Lander, E. S. et al. (2001). "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822, pp. 860–921. DOI: 10.1038/35057062.

Larsson, Anton J. M. et al. (Jan. 2019). "Genomic Encoding of Transcriptional Burst Kinetics". In: *Nature* 565.7738, pp. 251–254. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0836-1.

Li, Zongyi et al. (2021). *Fourier Neural Operator for Parametric Partial Differential Equations*. arXiv: 2010.08895 [cs.LG]. URL: https://arxiv.org/abs/2010.08895.

Lotka, Alfred J. (1920). "Analytical Note on Certain Rhythmic Relations in Organic Systems". In: *Proceedings of the National Academy of Sciences* 6.7, pp. 410–415. DOI: 10.1073/pnas.6.7.410. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.6.7.410.

Mahler, D. Luke et al. (2013). "Exceptional convergence on the macroevolutionary landscape in island lizard radiations". In: *Science* 341.6143, pp. 292–295. ISSN: 0036-8075. DOI: 10.1126/science.1232392.

Mendes, Fábio K et al. (July 2018). "A Multispecies Coalescent Model for Quantitative Traits". In: *eLife* 7. Ed. by Patricia J Wittkopp, Antonis Rokas, and Matthew Pennell, e36482. ISSN: 2050-084X. DOI: 10.7554/eLife.36482.

Milo, Ron and Rob Phillips (2015). *Cell Biology by the Numbers*. 1st. Illustrated by Nigel Orme. New York: Garland Science (Taylor & Francis Group), p. 400. ISBN: 9780815345374. DOI: 10.1201/9780429258770.

Modin, Klas and Milo Viviani (2022). "Canonical scale separation in two-dimensional incompressible hydrodynamics". In: *Journal of Fluid Mechanics* 943, A36. DOI: 10.1017/jfm.2022.457.

Montévil, Maël et al. (Oct. 2016). "Theoretical Principles for Biology: Variation". In: *Progress in Biophysics and Molecular Biology* 122.1, pp. 36–50. ISSN: 1873-1732. DOI: 10.1016/j.pbiomolbio.2016.08.005.

Mortazavi, Ali et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7, pp. 621–628. DOI: 10.1038/nmeth.1226.

O'Meara, Brian C. (2012). "Evolutionary Inferences from Phylogenies: A Review of Methods". In: *Annual Review of Ecology, Evolution, and Systematics* 43.Volume 43, 2012, pp. 267–285. ISSN: 1545-2069. DOI: 10.1146/annurev-ecolsys-110411-160331.

Phillips, Rob and Stephen R. Quake (May 2006). "The Biological Frontier of Physics". In: *Physics Today* 59.5, pp. 38–43. DOI: 10.1063/1.2216960. URL: https://physicstoday.aip.org/features/the-biological-frontier-of-physics.

Price, Peter D. et al. (July 2022). "Detecting Signatures of Selection on Gene Expression". In: *Nature Ecology & Evolution* 6.7, pp. 1035–1045. ISSN: 2397-334X. DOI: 10.1038/s41559-022-01761-8.

Sagoff, Mark (2016). "Are There General Causal Forces in Ecology?" In: *Synthese* 193.9. DOI: 10.1007/s11229-015-0907-x.

Schadt, Eric E. et al. (Sept. 2010). "Computational Solutions to Large-Scale Data Management and Analysis". In: *Nature Reviews Genetics* 11.9, pp. 647–657. ISSN: 1471-0064. DOI: 10.1038/nrg2857.

Stephens, Christopher (2004). "Selection, Drift, and the ?Forces? Of Evolution". In: *Philosophy of Science* 71.4, pp. 550–570. DOI: 10.1086/423751.

Stephens, Zachary D. et al. (2015). "Big Data: Astronomical or Genomical?" In: *PLOS Biology* 13.7, e1002195. DOI: 10.1371/journal.pbio.1002195. URL: https://doi.org/10.1371/journal.pbio.1002195.

Stoeckius, Marlon et al. (2017). "Simultaneous epitope and transcriptome measurement in single cells". In: *Nature Methods* 14, pp. 865–868. DOI: 10.1038/nmeth.4380.

Tang, Fuchou et al. (2009). "mRNA-Seq whole-transcriptome analysis of a single cell". In: *Nature Methods* 6.5, pp. 377–382. DOI: 10.1038/nmeth.1315.

Wang, Tianyu et al. (Jan. 2019). "Comparative Analysis of Differential Gene Expression Analysis Tools for Single-Cell RNA Sequencing Data". In: *BMC Bioinformatics* 20.1, p. 40. ISSN: 1471-2105. DOI: 10.1186/s12859-019-2599-6.

Wang, Xiaojing, Qi Liu, and Bing Zhang (Dec. 2014). "Leveraging the Complementary Nature of RNA-Seq and Shotgun Proteomics Data". In: *Proteomics* 14.0, pp. 2676–2687. ISSN: 1615-9853. DOI: 10.1002/pmic.201400184. (Visited on 11/04/2025).

Wilson, Kenneth G. (2025). *Kenneth G. Wilson – Nobel Lecture*. NobelPrize.org. Accessed 2025-12-12. URL: https://www.nobelprize.org/prizes/physics/1982/wilson/lecture/.

Appendix A

ATAC SUPPLEMENTARY INFORMATION

A.1 Site Inhomogeneity

We consider the mean chromatin openness values at each ATAC-seq peak site within the six-site loci we selected from each dataset. For a single ATAC-seq peak region x, we have a mean openness value, $\langle x \rangle$, given by:

$$\langle x \rangle = \frac{1}{N_c} \sum_{i=1}^{N_c} x_i, \tag{A.1}$$

where N_c is the number of cells in the dataset, and x_i denotes the openess value (0 or 1) of the chromatin measured at site x in cell i.

Figures A.1-A.3 show the mean site openness values at each site, within each locus, for all three datasets. These results indicate that, even within a single six-site locus, different ATAC-seq peak regions have significantly different average openness measurements.

A.2 Transition Matrix Generalization

The transition matrix given for a two-gene system in the main text can be extended naturally to a gene locus with an arbitrary number of chromatin sites. We allow transitions only to states which are accessible from the current state by changing the openness value of a single site, with transition rates $k_{on,i}$ for site i, and k_{off} for all sites. To express the preference for aligned adjacent site cooperation, we multiply these transition rates by a factor of $\epsilon^{-n_{mis}}$, where n_{mis} represents the number of 'misaligned' adjacent gene pairs in the current state.

For example, a DNA locus with three adjacent chromatin sites, and the chromatin-state basis given in the state matrix, S, of Equation 2.29, would have the following chromatin-state transition matrix, H:

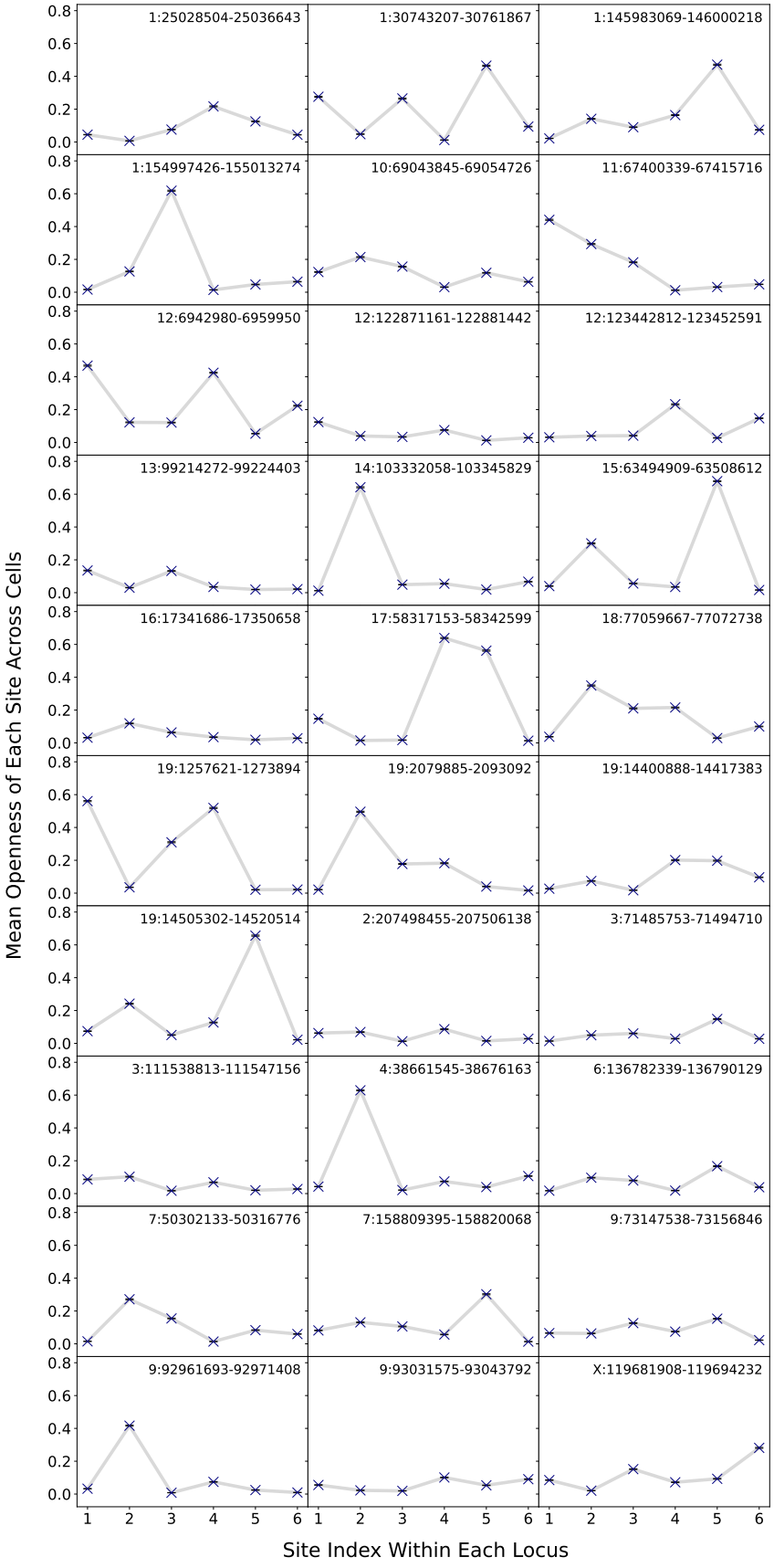


Figure A.1: The mean chromatin openness for the six adjacent ATAC-seq peak sites at each of the 30 loci from the PBMC dataset.

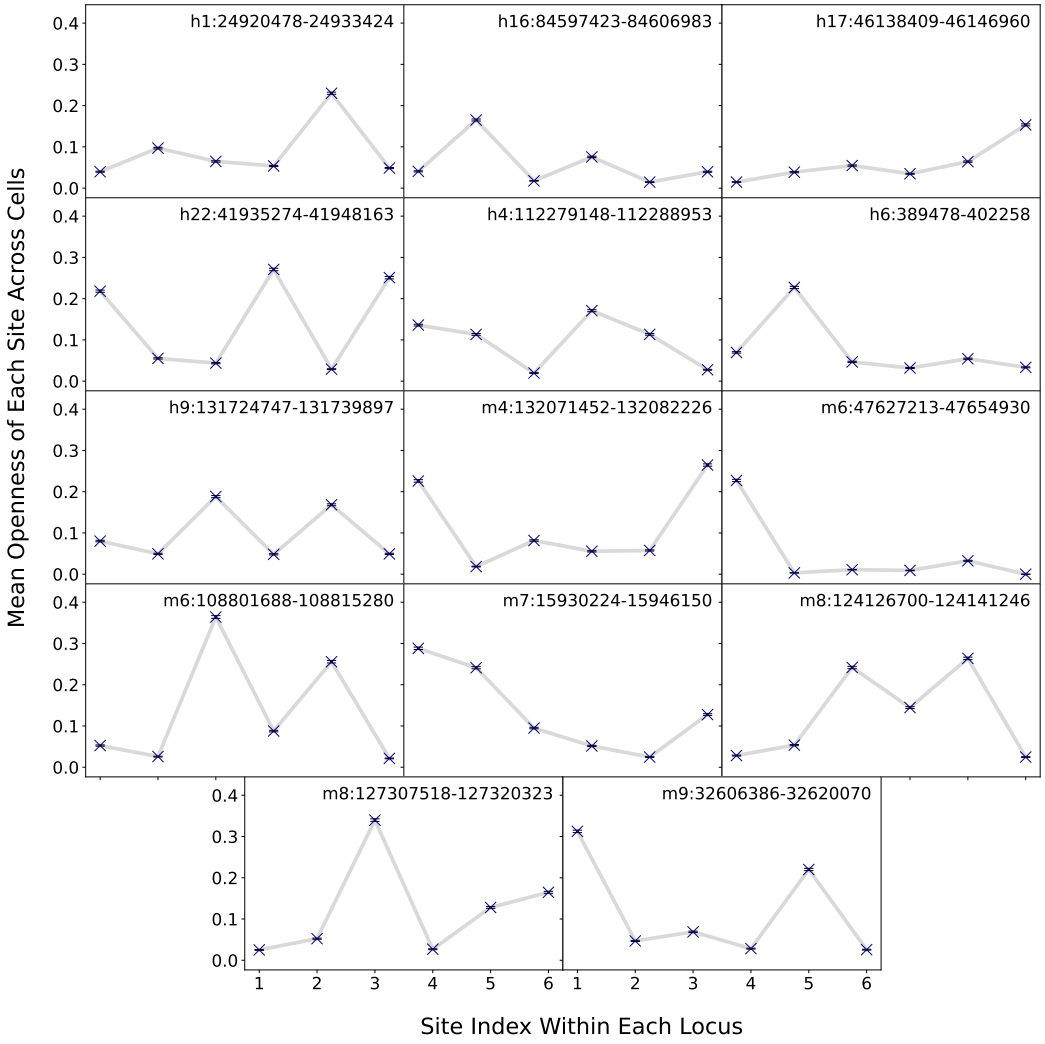


Figure A.2: The mean chromatin openness for the six adjacent ATAC-seq peak sites at each of the 14 loci from the human-mouse mixture dataset.

$$\operatorname{diag}\begin{bmatrix} 1 \\ \epsilon^{-1} \\ \epsilon^{-2} \\ \epsilon^{-1} \\ \epsilon^{-1} \\ \epsilon^{-2} \\ \epsilon^{-1} \\ 1 \end{bmatrix} \begin{pmatrix} -\Sigma_{000} & k_{\text{on},1} & k_{\text{on},2} & k_{\text{on},3} & 0 & 0 & 0 & 0 \\ k_{\text{off}} & -\Sigma_{100} & 0 & 0 & k_{\text{on},2} & k_{\text{on},3} & 0 & 0 \\ k_{\text{off}} & 0 & -\Sigma_{010} & 0 & k_{\text{on},1} & 0 & k_{\text{on},3} & 0 \\ k_{\text{off}} & 0 & 0 & -\Sigma_{001} & 0 & k_{\text{on},1} & k_{\text{on},2} & 0 \\ 0 & k_{\text{off}} & k_{\text{off}} & 0 & -\Sigma_{110} & 0 & 0 & k_{\text{on},3} \\ 0 & k_{\text{off}} & 0 & k_{\text{off}} & 0 & -\Sigma_{101} & 0 & k_{\text{on},2} \\ 0 & 0 & k_{\text{off}} & k_{\text{off}} & 0 & 0 & -\Sigma_{011} & k_{\text{on},1} \\ 0 & 0 & 0 & 0 & k_{\text{off}} & k_{\text{off}} & k_{\text{off}} & -\Sigma_{111} \end{pmatrix}, \tag{A.2}$$

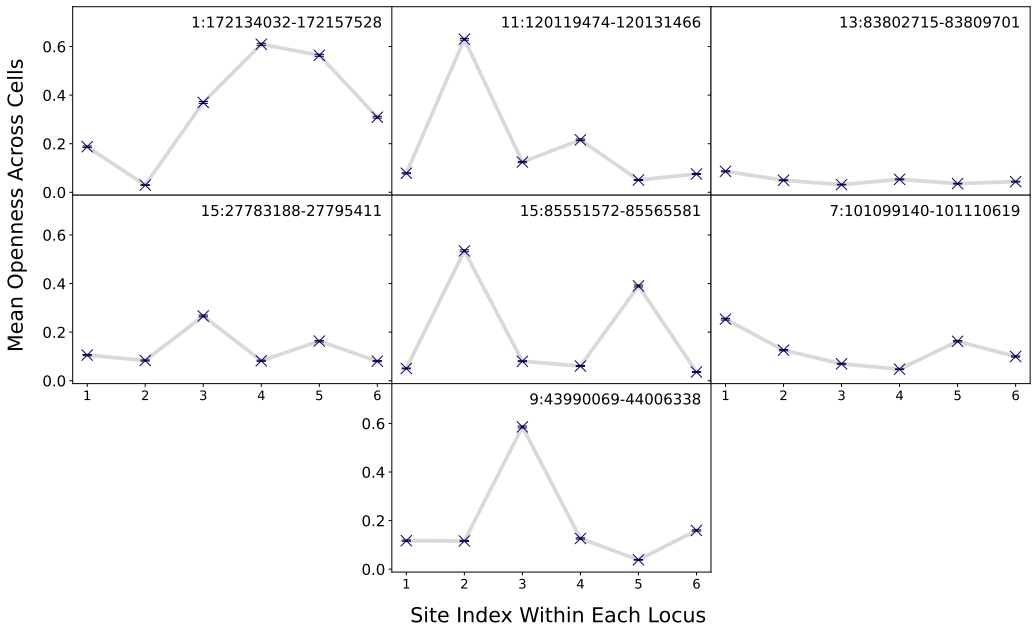


Figure A.3: The mean chromatin openness for the six adjacent ATAC-seq peak sites at each of the 7 loci from the mouse cortex dataset.

where Σ_α is the decay rate of state α, and is equal to the sum of the remaining values in the row ($\Sigma_{000} = k_{on,3}+k_{on,2}+k_{on,1}$, $\Sigma_{001} = k_{off}+k_{on,2}+k_{on,1}$, etc.), such that each row sums to zero. The first matrix is the diagonal of a vector $\tilde{\epsilon}$, where each entry $\tilde{\epsilon}_\alpha$ is equal to $\epsilon^{-n_{mis}}$, where n_{mis} is the number of opposite openness adjacent chromatin sites in state α. This generalization allows biophysical gene-state switching between similar configurations, at rates which encode the preference of neighboring DNA regions to have the same openness value, and can be extended straightforwardly to loci of arbitrary length.

A.3 RNA Transcript Count Moments

To find moments concerning the i^{th} species, we differentiate Equation 2.9 with respect to z^i. This allows us to find conditional means, $\tilde{\mu}^i_\alpha$, which represent the expected number of transcripts of species i, given that the system is in chromatin-state α. From the definition of G_α given in Equation 2.8, we note that:

$$\tilde{\mu}^i_\alpha = \frac{1}{\mathcal{P}_\alpha}\frac{\partial \boldsymbol{G}_\alpha}{\partial z_i}\Big|_{\boldsymbol{z}=\boldsymbol{1}}, \tag{A.3}$$

where $\mathcal{P}_\alpha$ is defined to be the probability that the system is in chromatin-state α. We define the vector $\boldsymbol{\mathcal{P}}$ to be the vector with the $\mathcal{P}_\alpha$ as compenents. Noting that

$\boldsymbol{G} = \boldsymbol{\mathcal{P}}$ for $\boldsymbol{z} = \mathbf{1}$, setting the LHS of Equation 2.9 to zero and taking a derivative with respect to z^i, we reach steady-state conditional means given by:

$$\tilde{\boldsymbol{\mu}}^{\boldsymbol{i}} = [(d^i I - H^T)\hat{\Pi}]^{-1}\hat{B}^i\boldsymbol{\pi}, \tag{A.4}$$

where $\hat{\Pi}$ is the diagonal matrix of $\boldsymbol{\pi}$, and $\hat{B}_i$ is the diagonal matrix of the i^{th} column of B. For our model of transcription, in index notation, this becomes:

$$[\hat{B}^i]_{\alpha\beta} = \delta_{\alpha\beta}S^i_\alpha b^i, \tag{A.5}$$

where no summation over α is implied. $\tilde{\boldsymbol{\mu}}^{\boldsymbol{i}}$ is the vector whose components α are defined in Equation A.3.

For the unconditional means, $\mu^i = \sum_\alpha \hat{\Pi}\tilde{\mu}^i_\alpha$, we find:

$$\mu^i = \sum_\alpha \left[(d^i I - H^T)^{-1}\hat{B}^i\boldsymbol{\pi}\right]_\alpha = \sum_\alpha \left[M^i\hat{B}^i\boldsymbol{\pi}\right]_\alpha, \tag{A.6}$$

where, for ease of this and future calculations, we define:

$$\begin{aligned} M^i &\equiv (d^i I - H^T)^{-1}, \\ \text{and } M^{ij} &\equiv [(d^i + d^j)I - H^T]^{-1}. \end{aligned} \tag{A.7}$$

We also make use of the Neumann series approximation for the matrix inverse, for a matrix $\tilde{H}$ with spectral radius less than one:

$$(I - \tilde{H})^{-1} = \sum_{k=0}^{\infty}\tilde{H}^k, \tag{A.8}$$

where for our purposes we define $\tilde{H}$ via:

$$H^T = d^i\tilde{H}^i, \tag{A.9}$$

for a scalar, d^i. Then, if the spectral radius of $H^T/d^i < 1$, we have:

$$M^i = (d^i I - H^T)^{-1} = \frac{1}{d^i} \sum_{k=0}^{\infty} (\tilde{H}^i)^k. \tag{A.10}$$

Recalling the symmetry of H^T and hence $\tilde{H}$, we have:

$$\sum_{\alpha} \tilde{H}_{\alpha\beta} = 0. \tag{A.11}$$

Inserting the Neumann expansion for M^i in Equation A.6, and using the symmetry in Equation A.11, we see that:

$$\mu^i = \frac{1}{d^i} \sum_{\alpha} S_\alpha{}^i b^i \pi_\alpha = \frac{b^i}{d^i} \langle \sigma^i \rangle. \tag{A.12}$$

For correlations between species, we use the equation:

$$\frac{1}{\mathcal{P}_\alpha} \frac{\partial^2 G_\alpha}{\partial z^i \partial z^j} = \langle m^i m^j \rangle_\alpha, \tag{A.13}$$

noting that here and in what follows, we require $i \neq j$. Differentiating Equation 2.9 twice, we get, in steady state:

$$\begin{aligned} 0 = {} & H^T \hat{\Pi} \tilde{\boldsymbol{\mu}}^{\boldsymbol{ij}} - d^i \hat{\Pi} \tilde{\boldsymbol{\mu}}^{\boldsymbol{ij}} - d^j \hat{\Pi} \tilde{\boldsymbol{\mu}}^{\boldsymbol{ij}} \\ & + \hat{B}^i \hat{\Pi} \tilde{\boldsymbol{\mu}}^{\boldsymbol{j}} + \hat{B}^j \hat{\Pi} \tilde{\boldsymbol{\mu}}^{\boldsymbol{i}}, \end{aligned} \tag{A.14}$$

where $(\tilde{\mu}^{ij})_\alpha = \langle m^i m^j \rangle_\alpha$ is the conditional expectation of the product of transcripts i and j given gene state α.

Solving for the mean product vector over gene states, we get:

$$\begin{aligned} \tilde{\boldsymbol{\mu}}^{\boldsymbol{ij}} = {} & [((d^i + d^j) I - H^T) \hat{\Pi}]^{-1} \\ & (\hat{B}^i \hat{\Pi} \tilde{\boldsymbol{\mu}}^{\boldsymbol{j}} + \hat{B}^j \hat{\Pi} \tilde{\boldsymbol{\mu}}^{\boldsymbol{i}}. \end{aligned} \tag{A.15}$$

Convolving this with the steady-state probability distribution for states, we can find the unconditional expectation value of the product of two genes. Subtracting the product of the means of each gene, we arrive at the correlation between the two gene counts.

The unconditional mean product is given by:

$$\begin{aligned}\mu^{ij} &= \sum_{\alpha} \left([(d^i + d^j)I - H^T]^{-1} [\hat{B}^i \hat{\Pi} \tilde{\boldsymbol{\mu}}^{\boldsymbol{j}} + \hat{B}^j \hat{\Pi} \tilde{\boldsymbol{\mu}}^{\boldsymbol{i}}) \right)_\alpha \\ &= \sum_{\alpha} \left(M^{ij} [\hat{B}^i \hat{\Pi} \tilde{\boldsymbol{\mu}}^{\boldsymbol{j}} + \hat{B}^j \hat{\Pi} \tilde{\boldsymbol{\mu}}^{\boldsymbol{i}}) \right)_\alpha, \end{aligned} \tag{A.16}$$

with M^{ij} as defined above (A.7).

Converting into index notation and using the relation between B and S (A.5), we find:

$$\begin{aligned}\mu^{ij} = b^i b^j \sum_{\alpha\beta} M^{ij}_{\alpha\beta} \sum_{\gamma} [&S_\beta{}^i M^j_{\beta\gamma} S_\gamma{}^j \\ &+ S_\beta{}^j M^i_{\beta\gamma} S_\gamma{}^i] \pi_\gamma. \end{aligned} \tag{A.17}$$

Using another symmetry argument related to A.11 (see Appendix A.3), this is always equivalent to:

$$\mu^{ij} = \frac{b^i b^j}{d^i + d^j} \sum_{\beta\gamma} [S_\beta{}^i M^j_{\beta\gamma} S_\gamma{}^j + S_\beta{}^j M^i_{\beta\gamma} S_\gamma{}^i] \pi_\gamma. \tag{A.18}$$

Note that, since to zeroth order in $\tilde{H}$, (i.e. in the limit of very slow switching, $k_{off}, k_{on} \ll d^i, d^j$ and ϵ not too small), we have:

$$\begin{aligned} M^{i(0)} &= \frac{1}{d^i} I, \\ M^{ij(0)} &= \frac{1}{d^i + d^j} I, \end{aligned} \tag{A.19}$$

giving, to zeroth order in $\tilde{H}$:

$$\mu^{ij(0)} = \frac{b^i b^j}{d^i d^j}\langle \sigma^i \sigma^j \rangle, \tag{A.20}$$

as expected.

From Equation A.16 and the expansion of M^i, in the case of bounded $\tilde{H}$ we have the exact expression for the covariance:

$$\mathrm{Cov}(m^i m^j) = \frac{b^i b^j}{d^i d^j}\mathrm{Cov}(\sigma^i \sigma^j) + \frac{b^i b^j}{d^i + d^j} \sum_{\beta\gamma} \sum_{k=1}^{\infty} \pi_\gamma [\frac{1}{d^j} S^i_\beta (\tilde{H}^j)^k_{\beta\gamma} S_\gamma{}^j + \frac{1}{d^i} S^j_\beta (\tilde{H}^i)^k_{\beta\gamma} S_\gamma{}^i]. \tag{A.21}$$

If $\tilde{H}$ is not bounded, the series will not converge and we need to return to the expression in Equation A.17.

Note that, due to the structure of $\tilde{H}$, the first-order term in $\tilde{H}$ always evaluates to zero (see Appendix A.11). We also show that both terms in the sum are identical, except for factors of $d^{i,j}$, giving:

$$\mathrm{Cov}(m^i m^j) = \frac{b^i b^j}{d^i d^j}\mathrm{Cov}(\sigma^i \sigma^j) + \frac{b^i b^j}{d^i + d^j} \sum_{\beta\gamma} \sum_{k=2}^{\infty} \pi_\gamma S^i_\beta (H^T)^k_{\beta\gamma} S_\gamma{}^j \left[\left(\frac{1}{d^j}\right)^{k+1} + \left(\frac{1}{d^i}\right)^{k+1} \right]. \tag{A.22}$$

Again, we see that for switching which is slow compared to the RNA decay rates, i.e. $\tilde{H} << I$ for all entries of the matrix, $\mathrm{Cov}(m^i m^j) \sim \frac{b^i b^j}{d^i d^j}\mathrm{Cov}(\sigma^i \sigma^j)$, as expected. Following the procedure above, we now consider the variance of transcript i. Defining $\tilde{\mu} \equiv \mu^i - (\mu^i)^2$, we have:

$$\begin{aligned} \mathrm{Var}(m^i) &= \sum_\alpha [(2d^i I - H^T)^{-1} 2\hat{B}^i \hat{\Pi} \tilde{\mu}^i]_\alpha + \tilde{\mu} \\ &= \sum_{\alpha\gamma\nu} 2M^{[2i]}_{\alpha\gamma} b^i S_\gamma{}^i M^i_{\gamma\nu} b^i S^i_\nu \pi_\nu + \tilde{\mu} \\ &= \frac{(b^i)^2}{d^i} \sum_{\alpha\beta} \left(S_\alpha{}^i M^i_{\alpha\beta} S_\beta{}^i \pi_\beta \right) + \tilde{\mu}, \end{aligned} \tag{A.23}$$

where we have defined matrix $M^{[2i]}$ via $M^{[2i]} \equiv (2d^i I - H^T)^{-1}$, and used the explicit decompositions of $\hat{B}$ and $\hat{\Pi}$. For the third equality we have used the argument in section A.3, to give the relation:

$$\sum_{\alpha} M^{[2i]}_{\alpha\gamma} = \frac{1}{2d^i} \sum_{\alpha} I_{\alpha\beta}. \tag{A.24}$$

For the case of bounded $\tilde{H}$, this becomes:

$$\begin{aligned} \mathrm{Var}(m^i) =& \left(\frac{b^i}{d^i}\right)^2 [\langle \sigma^i \sigma^i \rangle + \sum_{\gamma\beta} \sum_{k=1}^{\infty} S_\gamma{}^i \tilde{H}^k_{\gamma\beta} S_\beta{}^i \pi_\beta] + \frac{b^i}{d^i} \langle \sigma^i \rangle - \left(\frac{b^i}{d^i}\right)^2 \langle \sigma^i \rangle^2 \\ =& \left(\frac{b^i}{d^i}\right)^2 \left[\mathrm{Var}(\sigma^i) + \sum_{\gamma\beta} \sum_{k=1}^{\infty} S_\gamma{}^i \tilde{H}^k_{\gamma\beta} S_\beta{}^i \pi_\beta\right] + \frac{b^i}{d^i} \langle \sigma^i \rangle. \end{aligned} \tag{A.25}$$

A.4 Correlation Propagation

We define site-site correlations for accessibility and transcript count respectively via:

$$\rho^{ij}_{\sigma} = \frac{\mathrm{Cov}(\sigma^i \sigma^j)}{\sqrt{\mathrm{Var}(\sigma^i)\mathrm{Var}(\sigma^j)}}, \tag{A.26}$$

$$\rho^{ij} = \frac{\mathrm{Cov}(m^i m^j)}{\sqrt{\mathrm{Var}(m^i)\mathrm{Var}(m^j)}}. \tag{A.27}$$

We define f, the ratio between these quantities via:

$$\rho^{ij} = f \rho^{ij}_{\sigma}. \tag{A.28}$$

Returning to matrix notation, and dropping the assumption of bounded $\tilde{H}$, we recall the transcript covariances and variances:

$$\begin{aligned} \mathrm{Cov}(m^i m^j) = \frac{b^i b^j}{d^i + d^j} \sum_{\beta\gamma} [& S_\beta{}^i M^j_{\beta\gamma} S_\gamma{}^j \\ & + S_\beta{}^j M^i_{\beta\gamma} S_\gamma{}^i] \pi_\gamma - \mu^i \mu^j, \end{aligned} \tag{A.29}$$

$$\mathrm{Var}(m^i) = \frac{(b^i)^2}{d^i} \sum_{\alpha\beta} \left({S_\alpha}^i M^i_{\alpha\beta} {S_\beta}^i \pi_\beta \right) + \mu^i - (\mu^i)^2. \tag{A.30}$$

Compare these with the gene-state covariances and variances:

$$\mathrm{Cov}(\sigma^i \sigma^j) = \sum_\alpha \pi_\alpha {S_\alpha}^i {S_\alpha}^j - \sum_{\alpha\beta} \pi_\alpha {S_\alpha}^i \pi_\beta {S_\beta}^j, \tag{A.31}$$

where the variance is obtained by setting $i = j$ in this expression. An expression for f is obtained by taking the appropriate ratios of these moments:

$$f = \left[\frac{\frac{b^i b^j}{d^i + d^j} \sum_{\beta\gamma} [{S_\beta}^i M^j_{\beta\gamma} {S_\gamma}^j + {S_\beta}^j M^i_{\beta\gamma} {S_\gamma}^i] \pi_\gamma - \mu^i \mu^j}{\sqrt{\frac{(b^i)^2}{d^i} \sum_{\alpha\beta} \left({S_\alpha}^i M^i_{\alpha\beta} {S_\beta}^i \pi_\beta \right) + \mu^i - (\mu^i)^2} \sqrt{\frac{(b^j)^2}{d^j} \sum_{\alpha\beta} \left({S_\alpha}^j M^j_{\alpha\beta} {S_\beta}^j \pi_\beta \right) + \mu^j - (\mu^j)^2}} \right] \left[\frac{\sum_\alpha \pi_\alpha {S_\alpha}^i {S_\alpha}^j - \sum_{\alpha\beta} \pi_\alpha {S_\alpha}^i \pi_\beta {S_\beta}^j}{\sqrt{\sum_\alpha \pi_\alpha {S_\alpha}^i {S_\alpha}^i - \sum_{\alpha\beta} \pi_\alpha {S_\alpha}^i \pi_\beta {S_\beta}^i} \sqrt{\sum_\alpha \pi_\alpha {S_\alpha}^j {S_\alpha}^j - \sum_{\alpha\beta} \pi_\alpha {S_\alpha}^j \pi_\beta {S_\beta}^j}} \right]^{-1}. \tag{A.32}$$

We show the value of f for a two-gene system with varying $k_{\mathrm{on},1}, k_{on,2}$ and two sets of remaining parameters in Figure A.4. The other parameters are: $b_1 = b_2 = 10, d_1 = d_2 = 1$, and $\epsilon = 0.7, k_{\mathrm{o}ff} = 30$; $\epsilon = 0.5, k_{\mathrm{o}ff} = 0.5$ for the left and right panels respectively.

Since RNA counts are downstream of chromatin dynamics and add an additional layer of stochasticity, we might expect that f would be constrained within the range: $0 < f < 1$. Whilst the left panel in Figure A.4 shows a parameter regime where this constraint is observed, the right panel provides an example where $f > 1$. The intuition behind this surprising result is explored further in section A.5.

A.5 Illustrative Toy Systems

Here we consider illustrative toy systems where the fraction, f, of transcript correlation to gene-state correlation, falls outside of the naively expected range, $0 < f < 1$. To emphasize how this can be achieved, we demonstrate the behavior of two different two-gene systems.

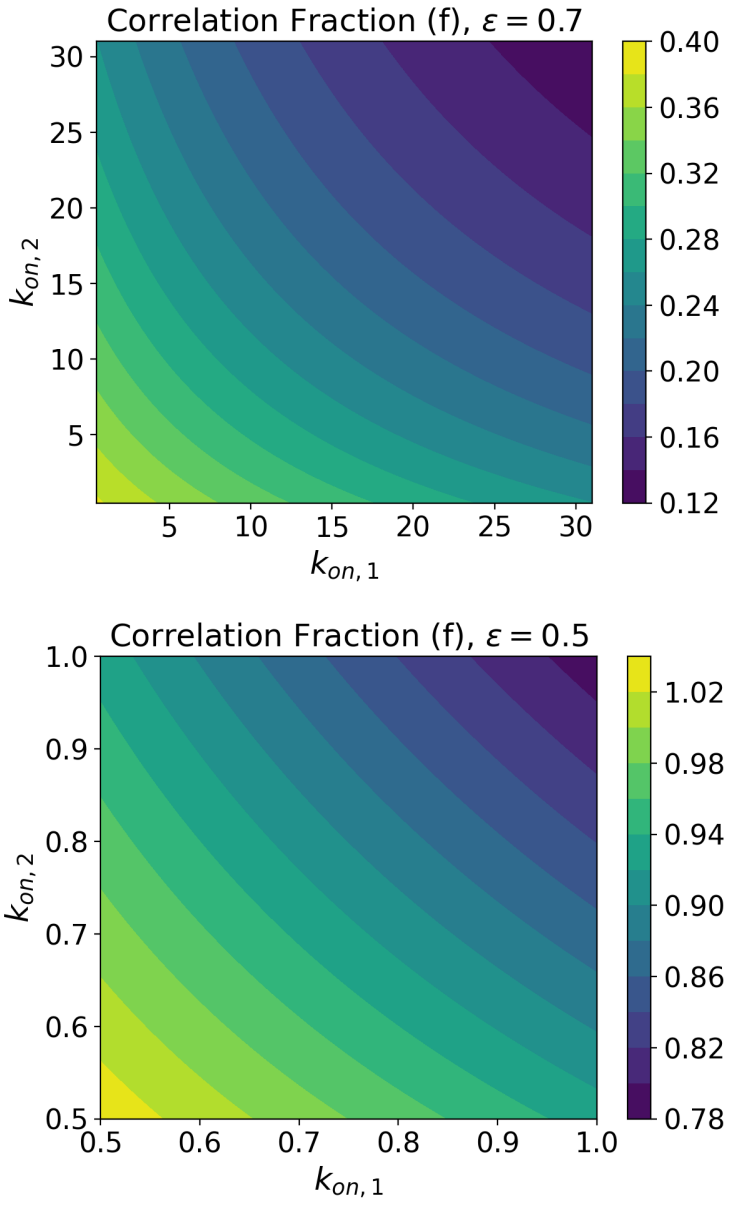


Figure A.4: Ratio, f, between gene-gene correlations at the transcript and at the chromatin level, calculated for different values of $k_{on,1}$, $k_{on,2}$ in a two-gene system. The other parameters are, **Top:** $\epsilon = 0.7$, $k_{off} = 30$, $b_1 = b_2 = 10$, $d_1 = d_2 = 1$; **Bottom:** $\epsilon = 0.5$, $k_{off} = 0.5$, $b_1 = b_2 = 10$, $d_1 = d_2 = 1$.

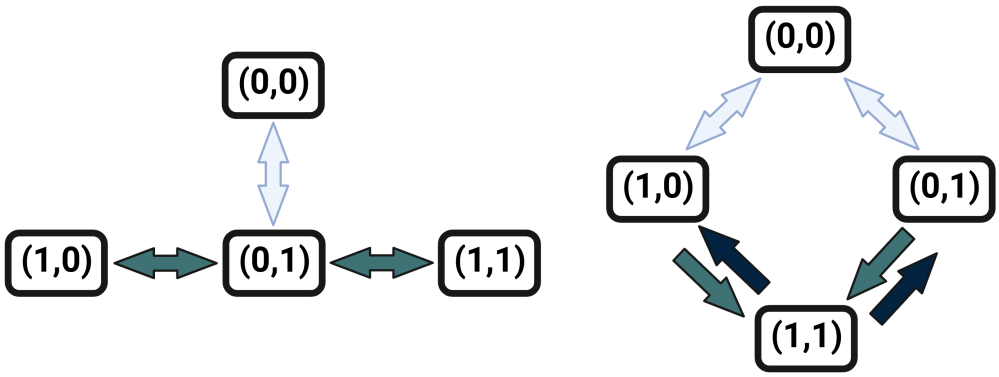


Figure A.5: Simple toy systems for which the value of f falls outside of the naively expected range. The states (i, j) correspond to the openness of sites i and j in that state, where $0, 1$ indicate closed and open chromatin respectively. The arrows between chromatin states represent allowed transitions, and their weights correspond to the rates of these transitions. **Left:** A system with $|f| > 1$. **Right:** A system with $f < 0$, i.e. reversed correlations between the transcript and chromatin-state levels.

The first system, although un-biological (including instantaneous changes of two genes at the same time), shows how Markov state transitions combined with constitutive transcript production can give transcript correlations without any correlations in the underlying chromatin openness. This corresponds to an infinite value of $|f|$.

The chromatin-state structure of this imagined system is shown in the left panel of Figure A.5, where the chromatin-states are labeled (i, j), for $i, j = 0, 1$ depending on the openness of each of the two chromatin sites. The arrows in the diagram represent transitions between chromatin configurations, and their weight represents the rate of each transition.

We could encode such a system using the following transition matrix:

$$H = \begin{pmatrix} -\delta k & 0 & \delta k & 0 \\ 0 & -k & k & 0 \\ \delta k & k & -(2+\delta)k & k \\ 0 & 0 & k & -k \end{pmatrix} \tag{A.33}$$

with $\delta < 1$, which has a uniform stationary distribution. However, due to the structure of the graph in Figure A.5, whilst there is no correlation between the openness of genes 0 and 1 in the steady state, there is clearly a correlation between their time averaged histories. This is because, if, for instance, gene 1 is on, that

indicates that the system is in the lower half of the state graph diagram, making it more likely that it has also been in the lower half for its recent history. This means that knowing that gene 1 is on increases the probability that gene 2 has been on in the system's recent history. In our gene model, this corresponds to a correlation between transcript numbers, despite there being no correlation between the openness of genes 1 and 2 in steady state. The correlations resulting from simulations of this system are shown on the top row of Figure A.6. Whilst correlations between transcripts are always significant and positive, the correlations between sites are around zero, resulting in f values with large magnitude.

The second example system, (no longer including un-biological transitions), has reversed sign correlations (i.e. $f < 0$). We consider chromatin-state evolution defined by the transition matrix:

$$H = \begin{pmatrix} -2\delta k & \delta k & \delta k & 0 \\ \delta k & -k(1+\delta) & 0 & k \\ \delta k & 0 & -k(1+\delta) & k \\ 0 & k\chi & k\chi & -2k\chi \end{pmatrix} \tag{A.34}$$

for $\delta < 1, \chi > 1$, and illustrated in the right hand panel of Figure A.5. The correlations observed for such a system from simulations are shown in the bottom row of Figure A.6. Whilst the correlations between transcripts are significant and positive as in the previous example, due to slow transitions from the $(0,0)$ state, there is a negative correlation between site openness values, due to the relative stability of the $(1,0)$ and $(0,1)$ states.

A.6 Noise

We consider binomial dropout with probability $p_{\mathrm{d}rop}$ applied independently at each site in a locus. Figure A.7 shows how this affects various distributions over three-site configurations in the Ising-like model we have described. In the case of uncorrelated (including uniformly distributed) sites, the binomial dropout effects all configurations with the same total number of open sites equally. However, for the case of correlated neighbors, ($\epsilon < 1$), the effect of dropout varies even between configurations with the same total number of open sites. For example, between $p_{\mathrm{d}rop} = 0.1$ and $p_{\mathrm{d}rop} = 0.5$, the probability of the $(0,1,0)$ configuration is reduced less than the $(0,0,1)$ and $(1,0,0)$ configurations. This is because the $(0,1,0)$ configura-

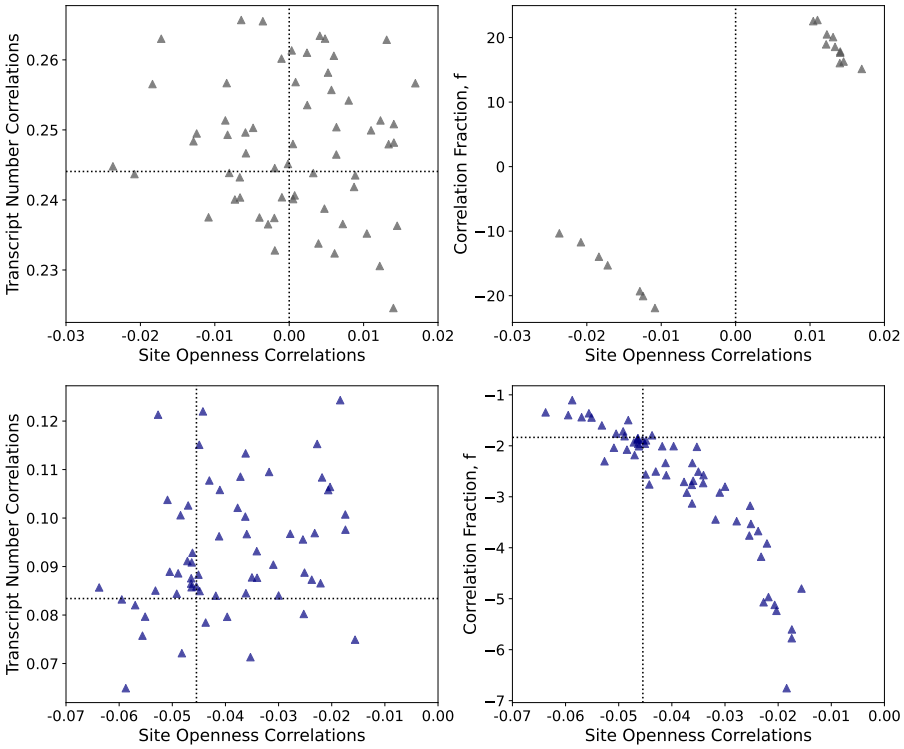


Figure A.6: Correlations for simulated toy systems, with analytic solutions shown as dotted lines. **Left**: Transcript correlations versus site openness correlations. **Right:** Correlation ratio, f, versus site openness correlations. **Top:** First toy system described, with $|f| > 1$. **Bottom:** Second toy system described, with $f < 0$. The parameters used were: $\delta = 0.1, \chi = 1.2, k = 1, b_1 = b_2 = 2, d_1 = d_2 = 0.5$, and results shown are for 5,000 simulated cells.

tion 'receives' probability from both the $(1, 1, 0)$ and $(0, 1, 1)$ configurations under dropout, which are favored by a factor $1/\epsilon$ compared to the $(1, 0, 1)$ configuration.

A.7 Model Fitting

Fits at the first 5 loci in dataset 1 are included in Figure A.8. The models are as described in the main text, and the parameters were fitted using Python's `differential_evolution` package. The algorithm was run 10 times per locus per model, to verify that the method converged closely to the same parameters each time. We also fit a locus from the 10x PBMC data using MCMC (see Figure A.9). The parameters found by the `differential_evolution` algorithm matched the most

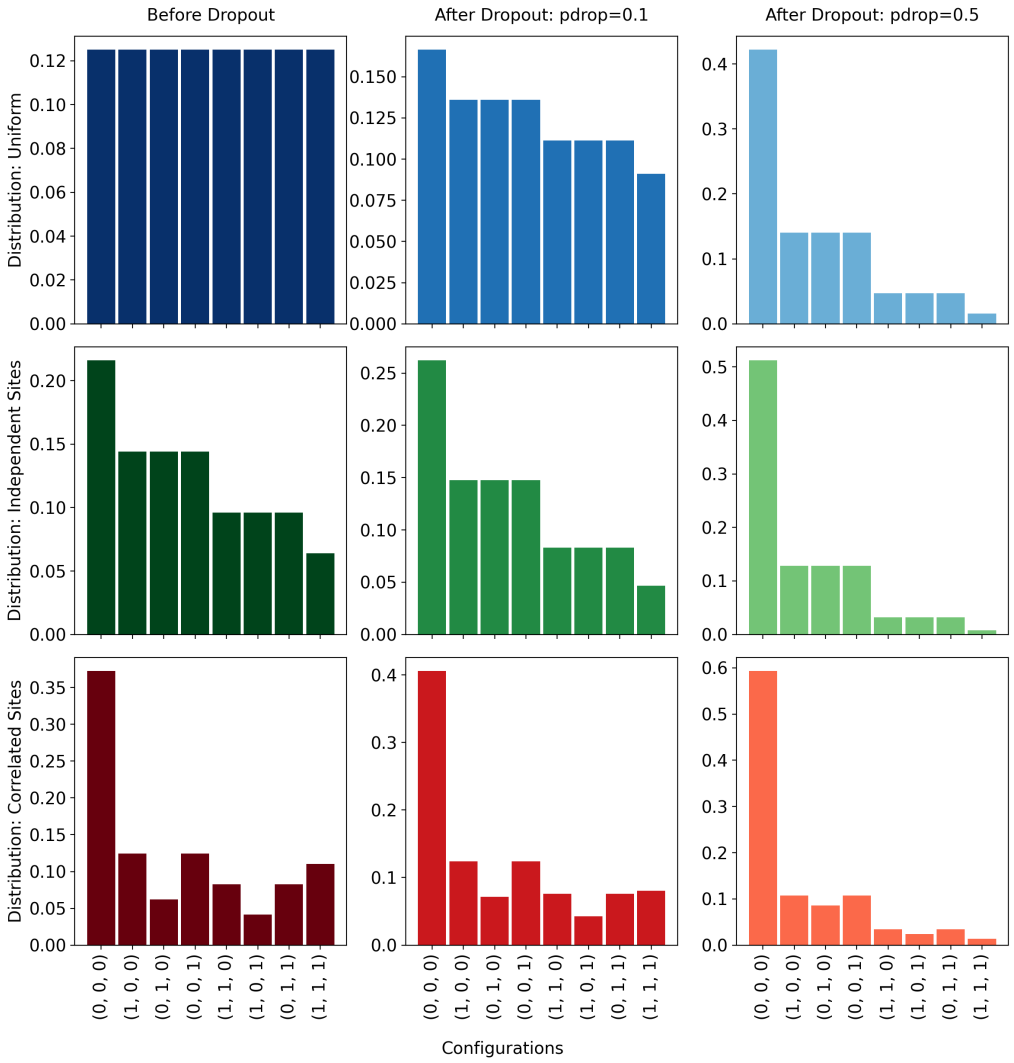


Figure A.7: The effect of binomial dropout on various three-site distributions. **Top:** Uniform distribution. **Middle:** Distribution from Ising model transition matrix with independent sites. Parameters: $k_{on} = 1$ for all sites, $k_{off} = 1.5$, $\epsilon = 1$. **Bottom:** Distribution from Ising model transition matrix with correlated adjacent sites. Parameters: $k_{on} = 1$ for all sites, $k_{off} = 1.5$, $\epsilon = 0.5$.

likely values from the MCMC results. The Bayesian Information Criteria (BIC) for the fits at each locus are shown in the main text. The BIC is defined as:

$$\mathrm{B}IC = k\ln(n) - 2\ln\hat{L}, \tag{A.35}$$

where k is the number of parameters in the model, n is the number of data-points, and $\hat{L}$ is the likelihood of the data under the fitted parameters.

A.8 snATAK Input

For pre-processing using snATAK, we used the technology string (argument option -x): `0,0,0:-1,0,0:1,0,0,2,0,0`, where files 0, 1, and 2 correspond to the R2, R1, and R3 FASTQ files respectively. This technology string indicates that the cell barcodes are found in the R2 file, that there are no UMIs for these ATAC-seq data, and that the paired-end reads are found in files R1 and R3. The entirety of each file is used. Note that for the whitelist we used the reverse complement of the whitelist provided by 10x.

A.9 Nearest-Neighbor Correlations

For each pair of adjacent ATAC-seq peak sites in the loci we selected, we calculated the Pearson correlation coefficient between the two sites in the pair. The sample Pearson correlation coefficient between two variables x and y is defined via:

$$\begin{aligned} r_{xy} =& \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ =& \frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\sqrt{\langle x^2\rangle - \langle x\rangle^2}\sqrt{\langle y^2\rangle - \langle y\rangle^2}}, \end{aligned} \tag{A.36}$$

where n is the size of the sample of joint observations of x and y, and $\langle\rangle$ denotes the sample averages of the enclosed quantities across all cells. In this case, we are considering each pair of adjacent sites, (x, y), which have openness values 0 or 1. Each cell in the dataset gives another joint observation of the value of the sites. For a pair of sites, we consider the number of observations across cells of each possible openness configuration, $((0,0),(0,1),(1,0),(1,1))$, and label the proportion of such observations p_{00}, p_{01}, p_{10}, and p_{11} respectively. Note that, since the openness values at sites x and y take binary values, we have:

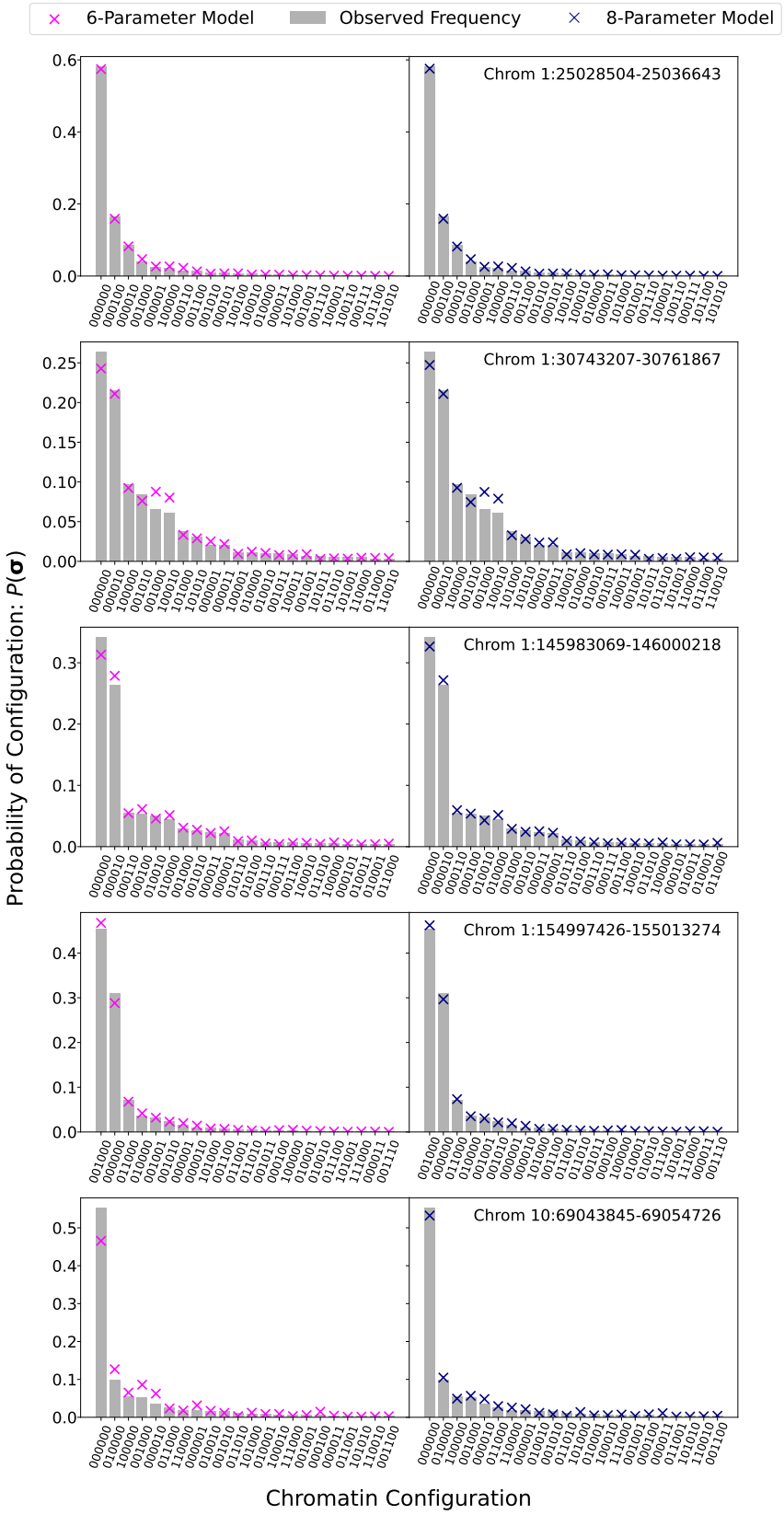


Figure A.8: Six and eight-parameter model fits to 5 loci in the human PBMC dataset. The scatter plots show the analytic distribution at the best fitting parameters, and the bar chart shows the empirical distribution.

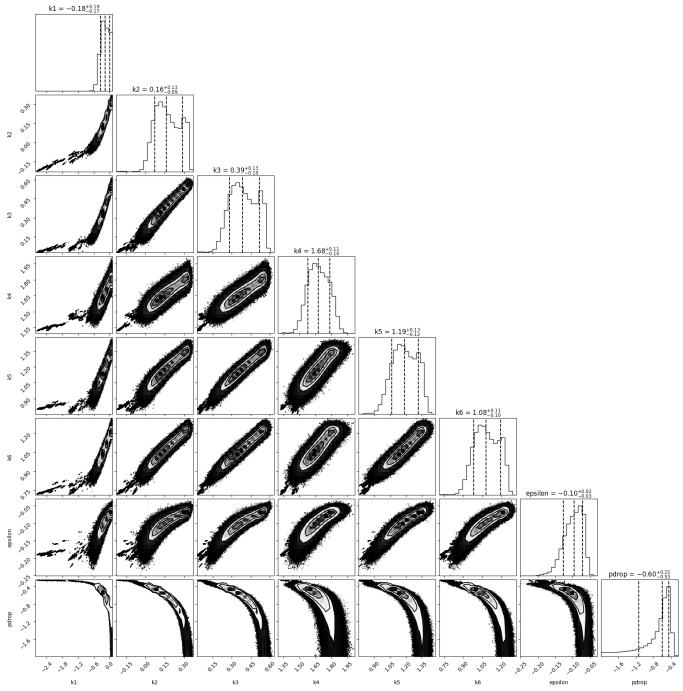


Figure A.9: MCMC methods used to fit the eight-parameter model on locus 11:67400339-67415716 of the PBMC dataset. The most probable parameter values matched with those found using the global minimization algorithm. 512 walkers were used, with initial values of $k_{on,i} = 1$ for all sites, $\epsilon = 1$ and $p_{drop} = 0.7$. The parameters were fit in log-space.

$$\begin{aligned} \langle x^2 \rangle &= \langle x \rangle = p_{11} + p_{10}, \\ \langle y^2 \rangle &= \langle y \rangle = p_{11} + p_{01}. \end{aligned} \tag{A.37}$$

From Equation A.36, we then have:

$$r_{xy} = \frac{p_{11} - (p_{10} + p_{11})(p_{01} + p_{11})}{\sqrt{p_{11} + p_{10} - (p_{11} + p_{10})^2}\sqrt{p_{11} + p_{01} - (p_{11} + p_{01})^2}}. \tag{A.38}$$

We calculate the Pearson coefficient in this way from each pair of adjacent sites in the selected loci. We can also calculate the average openness of the pair of sites, via:

$$\mu_{av} = \frac{\langle x \rangle + \langle y \rangle}{2}. \tag{A.39}$$

A.10 Positive Outlier: Locus 10:69043845-69054726

In the PBMC dataset, locus 10:69043845-69054726 shows a significantly larger BIC difference in favor of the eight-parameter model than any other locus. The fits of each model are shown in the bottom row of Figure A.8. Figure A.1 (the middle of the second row) also shows that this locus seems to have mean openness values that are correlated between neighboring sites.

A.11 Symmetries

Transition Matrix Symmetry

Let us consider terms of the form:

$$\sum_{\alpha\beta} {S_\alpha}^i \tilde{H}_{\alpha\beta} {S_\beta}^j \pi_\beta. \tag{A.40}$$

Recalling that $\tilde{H} \propto H^T$, we note that the components $\tilde{H}_{\alpha\beta}$ are only non-zero for states which are connected (denoted $\alpha \sim \beta$), meaning that α can be reached from β by flipping the openness of a single site. Recall also that ${S_\alpha}^i = 0, 1$ indicates that site i is closed/open in state α. Since ${S_\alpha}^i$ and ${S_\beta}^j$ must both be 1 for non-zero contributions to A.40, and $\alpha \sim \beta$, we can re-write Expression A.40 as:

$$\sum_{\alpha} \sum_{\beta\sim\alpha} {S_\alpha}^i {S_\beta}^j \Big({S_\beta}^i {S_\alpha}^j + (1 - {S_\beta}^i) {S_\alpha}^j + (1 - {S_\alpha}^j) {S_\beta}^i \Big) \tilde{H}_{\alpha\beta} \pi_\beta. \tag{A.41}$$

This is because, unless ${S_\beta}^i$ and ${S_\alpha}^j$ are both zero, the multiplier within the parentheses evaluates to 1. But, as discussed above, for $\beta \sim \alpha$, we cannot have ${S_\alpha}^i = {S_\beta}^j = 1$, and ${S_\beta}^i = {S_\alpha}^j = 0$, since this would require both sites i and j to change openness value between states α and β

We can think of the multiplier within the parentheses of Expression A.41 as dividing the expression into the different possible transitions between connected states. The first term represents transitions between states where both sites i and j are open in both α and β, the second from i closed and j open to both open, and the last from both open to i open and j closed. Then, noting that $\boldsymbol{\pi}$ is the steady-state vector of the Markovian transition matrix H^T, we have that:

$$\sum_{\alpha,\alpha\sim\beta,\alpha\neq\beta} \tilde{H}_{\alpha\beta}\pi_\beta = \sum_{\alpha,\alpha\sim\beta,\alpha\neq\beta} \tilde{H}_{\beta\alpha}\pi_\alpha, \tag{A.42}$$

allowing us to re-write the final term of Equation A.41 as:

$$\begin{aligned}
&\sum_\alpha \sum_{\beta\sim\alpha} {S_\alpha}^i {S_\beta}^j (1 - {S_\alpha}^j){S_\beta}^i \tilde{H}_{\alpha\beta}\pi_\beta \\
&= \sum_\alpha \sum_{\beta\sim\alpha} {S_\alpha}^i {S_\beta}^j (1 - {S_\alpha}^j){S_\beta}^i \tilde{H}_{\beta\alpha}\pi_\alpha \\
&= \sum_\alpha \sum_{\beta\sim\alpha} {S_\beta}^i {S_\alpha}^j (1 - {S_\beta}^j){S_\alpha}^i \tilde{H}_{\alpha\beta}\pi_\beta,
\end{aligned} \tag{A.43}$$

where in the last line we have simply relabeled the indices (note the symmetry of the relation $\beta \sim \alpha$). This allows us to re-write the total expression in A.41 as:

$$\begin{aligned}
\sum_\alpha {S_\alpha}^i {S_\alpha}^j \sum_{\beta\sim\alpha} \Big({S_\beta}^j {S_\beta}^i + (1 - {S_\beta}^i){S_\beta}^j + (1 - {S_\beta}^j){S_\beta}^i \Big) \tilde{H}_{\alpha\beta}\pi_\beta \\
= \sum_\alpha {S_\alpha}^i {S_\alpha}^j \sum_{\beta\sim\alpha} \tilde{H}_{\alpha\beta}\pi_\beta = 0,
\end{aligned} \tag{A.44}$$

where the first equality comes from the fact that ${S_\alpha}^i {S_\alpha}^j$ selects for transitions into states with both sites i and j open. The connected origin states β must either have both states i and j open, or exactly one of the i and j sites open. Hence the sum in terms of ${S_\beta}^{i,j}$ expresses all of the non-zero transitions between states β and α, making the expression redundant with the specification that all $\beta \sim \alpha$, (i.e. the multiplier in parentheses must evaluate to 1, similarly to before). The second equality comes from the symmetry of $\tilde{H}$ (A.11), and the fact that only $\tilde{H}_{\alpha\beta}$ are only non-zero for $\beta \sim \alpha$ (or alternatively from the definition of $\boldsymbol{\pi}$).

Thus we have shown that:

$$\sum_{\alpha\beta} {S_\alpha}^i \tilde{H}_{\alpha\beta} {S_\beta}^j \pi_\beta = 0. \tag{A.45}$$

Inverse Matrix Symmetry

We consider the inverse matrix defined via:

$$M \equiv (I - \tilde{H})^{-1}. \tag{A.46}$$

Note that, by the definition of M:

$$(I - \tilde{H})M = I, \tag{A.47}$$

and that, also:

$$(I - \tilde{H})M = M - \tilde{H}M. \tag{A.48}$$

Considering the RHS of Equation A.48 component-wise, and summing over α, we note that:

$$\begin{aligned} \sum_{\alpha}[M_{\alpha\beta} - \sum_{\gamma}\tilde{H}_{\alpha\gamma}M_{\gamma\beta}] &= \sum_{\alpha} M_{\alpha\beta} - \sum_{\gamma}\sum_{\alpha}\tilde{H}_{\alpha\gamma}M_{\gamma\beta} \\ &= \sum_{\alpha} M_{\alpha\beta}, \end{aligned} \tag{A.49}$$

where for the last equality we have used the symmetry of $\tilde{H}$:

$$\sum_{\alpha}\tilde{H}_{\alpha\beta} = 0. \tag{A.50}$$

Then, equating this with the component-wise version of Equation A.47, we arrive at:

$$\sum_{\alpha} M_{\alpha\beta} = \sum_{\alpha} I_{\alpha\beta}, \tag{A.51}$$

as desired. □

Appendix B

PROMONOD SUPPLEMENTARY INFORMATION

B.1 Stationary distribution of mRNA and protein via generating function methods

The generating functions, as defined in Equations 3.3-3.4 evolve as follows:

$$\begin{aligned}
\frac{\partial G(z_u, z_s, z_p, t)}{\partial t} &= k(F(z_u) - 1)G + \beta(z_s - z_u)\frac{\partial G}{\partial z_u} + \gamma(1 - z_s)\frac{\partial G}{\partial z_s} \\
&\quad + k_p(z_p - 1)z_s\frac{\partial G}{\partial z_s} + \gamma_p(1 - z_p)\frac{\partial G}{\partial z_p}, \\
\implies \frac{\partial \phi(u_u, u_s, u_p, t)}{\partial t} &= kM(u_u) + \beta(u_s - u_u)\frac{\partial \phi}{\partial u_u} - \gamma u_s\frac{\partial \phi}{\partial u_s} \\
&\quad + k_p u_p(u_s + 1)\frac{\partial \phi}{\partial u_s} - \gamma_p u_p\frac{\partial \phi}{\partial u_p}.
\end{aligned}$$

When $b \to 0, F(z_u) \approx 1 + bz_u$, and this reduces to the constitutive model. For the steady-state solution, we set $\frac{\partial \phi}{\partial t} = 0$ and solve the resulting first order PDE using the method of characteristics (Courant and Hilbert, 1962). We write all the variables as functions of s, i.e., $\phi(\tilde{u_u}(s), \tilde{u_s}(s), \tilde{u_p}(s), t(s))$, giving the characteristic ODEs:

$$\begin{aligned}
\frac{d\tilde{t}}{ds} &= -1 & \tilde{t}(0) &= t \\
\frac{d\tilde{u}_u}{ds} &= \beta(\tilde{u}_s - \tilde{u}_u) & \tilde{u}_u(0) &= u_u \\
\frac{d\tilde{u}_s}{ds} &= -\gamma\tilde{u}_s + k_p\tilde{u}_p(\tilde{u}_s + 1) & \tilde{u}_s(0) &= u_s \\
\frac{d\tilde{u}_p}{ds} &= -\gamma_p\tilde{u}_p & \tilde{u}_p(0) &= u_p
\end{aligned} \tag{B.1}$$

and

$$\phi(u_u, u_s, u_p, t) = \phi(\tilde{u}_u(0), \tilde{u}_s(0), \tilde{u}_p(0), 0) + \int_0^t kM(\tilde{u}_u(s))ds. \tag{B.2}$$

For the bursty model with geometrically distributed burst sizes, we have $M(u) = \frac{bu}{1-bu}$. We have immediately that

$$\tilde{u}_p = u_p e^{-\gamma_p s}. \tag{B.3}$$

In summary, we need to solve the following:

$$\begin{aligned}
\frac{d\tilde{u}_s}{ds} &= -\gamma\tilde{u}_s + k_p u_p e^{-\gamma_p s}(\tilde{u}_s + 1) & \tilde{u}_s(0) &= u_s \\
\frac{d\tilde{u}_u}{ds} &= \beta(\tilde{u}_s - \tilde{u}_u) & \tilde{u}_u(0) &= u_u \\
\phi(u_u, u_s, u_p, \infty)) &= \int_0^\infty \frac{kb\tilde{u}_u(s)}{1 - b\tilde{u}_u(s)} ds.
\end{aligned}$$

B.2 Solving probability distribution numerically

We used Runge-Kutta 4 (Runge, 1895) to numerically integrate these equations.

B.3 Biological data analysis

Data processing

We obtained our sequencing data from the following human PBMC datasets:

- 10k Human PBMCs Stained with TotalSeq™-B Human TBNK Cocktail, Chromium GEM-X Single Cell 3' Universal 3' Gene Expression dataset analyzed using Cell Ranger 8.0.0 (2024, March 13) (Genomics, 2024)
- 10K Human PBMCs, Gene Expression with a Panel of TotalSeq™-B Antibodies, analyzed using Cell Ranger 3.0.0, (2018, November 19) (Genomics, 2018)

The raw RNA-seq fastq files were processed with `kb-python` (Sullivan et al., 2025); we used the `nac` workflow with the appropriate 10x technology string. The call for processing the 2024 dataset is shown as an example here:

```
kb count \
--overwrite \
--h5ad \
--workflow=nac \
-i /home/cfelce/proMonod/data/ref/human/GRCh38.110/index.idx \
-g /home/cfelce/proMonod/data/ref/human/GRCh38.110/t2g.txt \
-x 10xv4 \
```

```
 -o /home/cfelce/proMonod/data/RNA_S2 \
-c1 /home/cfelce/proMonod/data/ref/human/GRCh38.110/cdna.txt \
-c2 /home/cfelce/proMonod/data/ref/human/GRCh38.110/nascent.txt \
-m 16G \
--verbose \
--filter bustools \
10k_Human_PBMC_TotalSeqB_3p_gemx_fastqs/gex/
 10k_Human_PBMC_TotalSeqB_3p_gemx_gex1_S2_L001_R1_001.fastq.gz \
10k_Human_PBMC_TotalSeqB_3p_gemx_fastqs/gex/
 10k_Human_PBMC_TotalSeqB_3p_gemx_gex1_S2_L001_R2_001.fastq.gz \
10k_Human_PBMC_TotalSeqB_3p_gemx_fastqs/gex/
 10k_Human_PBMC_TotalSeqB_3p_gemx_gex1_S2_L002_R1_001.fastq.gz \
10k_Human_PBMC_TotalSeqB_3p_gemx_fastqs/gex/
 10k_Human_PBMC_TotalSeqB_3p_gemx_gex1_S2_L002_R2_001.fastq.gz \
10k_Human_PBMC_TotalSeqB_3p_gemx_fastqs/gex/
 10k_Human_PBMC_TotalSeqB_3p_gemx_gex1_S2_L003_R1_001.fastq.gz
10k_Human_PBMC_TotalSeqB_3p_gemx_fastqs/gex/
 10k_Human_PBMC_TotalSeqB_3p_gemx_gex1_S2_L003_R2_001.fastq.gz
10k_Human_PBMC_TotalSeqB_3p_gemx_fastqs/gex/
 10k_Human_PBMC_TotalSeqB_3p_gemx_gex1_S2_L004_R1_001.fastq.gz
10k_Human_PBMC_TotalSeqB_3p_gemx_fastqs/gex/
 10k_Human_PBMC_TotalSeqB_3p_gemx_gex1_S2_L004_R2_001.fastq.gz
```

Some proteins, or protein complexes, corresponded to multiple RNA transcripts in the data. The identifications in Table B.1 (10x, 2024) and Table B.2 (10x, 2018), were used to map protein counts to RNA counts. The cells were clustered using Leiden clustering (see the scripts at `https://github.com/pachterlab/FFP_2025`), and subsetted to monocytes for the 10x, 2024 dataset, and to a group of T-cells for the 10x, 2018 dataset.

Fitted parameters

We fit the bursty transcription with translation model, with Poissonian count sampling, described in the main text, Section 3.3. Note that, since we fit only spliced RNA and protein counts, the technical sampling rate for unspliced counts only affected the method of moments initialization values.

We show the optimal biological parameters for each fit in Table B.3, along with

Protein	Subunits (alternate forms)	Ensembls
CD3 (complex)	CD3γ (x1) CD3δ (x1) CD3ϵ (x2)	ENSG00000160654 ENSG00000167286 **ENSG0000019885**
CD4	-	ENSG00000010610
CD8	CD8a (CD8b)	ENSG00000153563 (ENSG00000172116)
CD11C	-	ENSG00000140678
CD14	-	ENSG00000170458
CD16	CD16a (CD16b)	ENSG00000203747 (ENSG00000162747)
CD19	-	ENSG00000177455
CD56	-	ENSG00000149294
CD45	-	ENSG00000081237 (ENSG00000262418)

Table B.1: Proteins (complexes) with their constituent subunits or alternative forms, and the corresponding Ensembl IDs, for the 10x 2024 dataset. Where alternate forms are given in parentheses, the non-parenthesized Ensembl ID was used. Where subchains of a complex are shown, the bolded ensembl was used.

the technical parameters used (capture rates $\lambda_{u,s,p}$ for unspliced, spliced and protein counts respectively), and the initialization method used for fitting.

References

Courant, Richard and David Hilbert (1962). *Methods of Mathematical Physics, Volume II*. Wiley-Interscience.

Genomics, 10x (Nov. 2018). *10k PBMCs from a Healthy Donor - Gene Expression with a Panel of TotalSeq-B Antibodies Universal 3' Gene Expression*. 10x Genomics Datasets. Version Cell Ranger v3.0.0. Dataset analyzed using Cell Ranger v3.0.0. URL: https://www.10xgenomics.com/datasets/10-k-pbm-cs-from-a-healthy-donor-gene-expression-and-cell-surface-protein-3-standard-3-0-0.

– (Oct. 2024). *10k Human PBMCs Stained with TotalSeq-B Human TBNK Cocktail, Chromium GEM-X Single Cell 3' Gene Expression*. 10x Genomics Datasets. Version Cell Ranger v8.0.0. Dataset analyzed using Cell Ranger v8.0.0. URL: https://www.10xgenomics.com/datasets/10k-human-pbmcs-stained-with-totalseq-B-human-TBNK-cocktail-GEM-X.

Gene Symbol	Ensembl ID
PTPRC	ENSG00000081237
FCGR3A	ENSG00000203747
CD247	ENSG00000198821
CD3E	ENSG00000198851
CD3D	ENSG00000167286
CD3G	ENSG00000160654
FUT4	ENSG00000196371
NCAM1	ENSG00000149294
CD4	ENSG00000010610
IGHG1	ENSG00000211896
IGHG2	ENSG00000211893
ISG20	ENSG00000172183
CD19	ENSG00000177455
PDCD1	ENSG00000188389
CD8A	ENSG00000153563
TIGIT	ENSG00000181847
IL7R	ENSG00000168685
CD14	ENSG00000170458

Table B.2: Final one-to-one mapping, used for the 10x 2018 dataset, between protein (gene) symbols and Ensembl IDs.

Runge, C. (June 1895). “Ueber Die Numerische Auflösung von Differentialgleichungen”. In: *Mathematische Annalen* 46.2, pp. 167–178. ISSN: 1432-1807. DOI: 10.1007/BF01446807.

Sullivan, Delaney K. et al. (2025). “kallisto, bustools and kb-python for quantifying bulk, single-cell and single-nucleus RNA-seq”. In: *Nature Protocols* 20.3, pp. 587–607. DOI: 10.1038/s41596-024-01057-0.

	2024, CD14	**2024, CD45**	**2018, IL7R**
Cell Type	Monocytes	Monocytes	T-cell cluster
$\log_{10} b$	3.04	1.96	4.2
$\log_{10} \beta$	-0.403	-0.993	-0.665
$\log_{10} \gamma$	1.77	-0.406	1.31
$\log_{10} k_p$	3.5	-1.14	1.24
$\log_{10} \gamma_p$	-0.634	-0.862	-0.764
$\log_{10} \lambda_u$	0.0	-	0.0
$\log_{10} \lambda_s$	0.0	-1	-2.0
$\log_{10} \lambda_p$	-2.5	0	-2.5
Initialization	Moments estimate	10 random restarts	Moments estimate

Table B.3: Optimal biophysical parameters for each of the fits shown in the main text, and the corresponding technical capture rates. The cell type and initialization method for the fit are also given.

Appendix C

MECHANISMS OF GENE EXPRESSION EVOLUTION: SUPPLEMENTARY INFORMATION

C.1 Data processing

Pre-processing

For this study we use single-cell RNA-seq data from Jiao et al. Jiao et al., 2024, extracted from the spleen of seven different species. We processed the data using kallisto Sullivan et al., 2025 to obtain spliced and unspliced count matrices, and filtered out low UMI cells. An example kallisto call is given here:

```
kb count --verbose -i ./frog/index.idx -g ./frog/t2g_mm10.txt
   -x 10xv2 -o ./frog/output -t 24 -m 8G -c1 ./frog/cdna_t2c.txt
   -c2 ./frog/intron_t2c.txt --workflow=nac --filter bustools
   --strand=unstranded --sum=cell
   ../SRR16490736_1.fastq ../SRR16490736_2.fastq,
```

with example output summary:

```
 "n_targets": 76913,
 "n_bootstraps": 0,
 "n_processed": 289062079,
 "n_pseudoaligned": 190915538,
 "n_unique": 35185121,
 "p_pseudoaligned": 66.0,
 "p_unique": 12.2,
 "kallisto_version": "0.50.1",
 "index_version": 13,
 "start_time": "Wed Jun 12 15:04:30 2024"
```

After clustering the data from each species by cell-type, we excluded the fish sample from further analysis because of an indistinct and low-count T-cell cluster, leaving six remaining species, which were filtered for T-cells.

Fitting biophysical parameters

We searched for genes which had orthologs in all six species using Ensembl BioMart Kinsella et al., 2011. We then fit transcriptional rates for these genes in each species separately using Monod Gorin and Pachter, 2023, using the bursty transcription model with Poisson technical noise. After fitting with Monod, which filters some genes, and removing genes without a fitted ortholog in all six species, we were left with 167 genes. An example of the Monod run is given below:

```
fitmodel = cme_toolbox.CMEModel('Bursty','Poisson')
filt_param = {'min_means':[0.01, 0.01], 'max_maxes':[350, 350],
'min_maxes':[1,3]}

lb = [-1.0, -1.8, -1.8 ]
ub = [4.2, 2.5, 3.5]
# samp_lb, samp_ub = [-8, -3],[-5, 0]
samp_lb, samp_ub = [-11, -6],[-5, 0]

grid = [6,7]

fitted_adata = inference.perform_inference(combined_adata,
   fitmodel, n_genes=5000, seed=5, phys_lb=lb, phys_ub=ub,
   gridsize=grid, samp_lb=samp_lb, samp_ub=samp_ub,
   filt_param=filt_param, gradient_param={'max_iterations':5,
   'init_pattern':'moments','num_restarts':1},
   dataset_string=dataset_string, viz=True,num_cores=32)
```

The output of this procedure is a per-gene burst size, b, splicing rate β, and decay rate, γ, with the rates given in units of the transcription initiation rate, k, all in log space. We then subtracted the mean of each parameter across genes from each species, and used the resulting values as the traits for the phylogenetic analysis.

Phylogenetic tree

For the phylogenetic tree, we used the following tree, from TimeTree Kumar et al., 2022, in Newick format:

```
(Frog:351.68654000,(Pig:94.00000000,((Rat:11.64917000,
```

```
Mouse:11.64917000) '14':75.55083000 (Human:28.82000000,
Macaque:28.82000000)'13':58.38000000) '25':6.80000000)
'37':257.68654000);
```

C.2 Two-dimensional evolution model

Phylogenetic model derivation from fitness landscape

Following Cope et al. Cope, Schraiber, and Pennell, 2025, we consider a fitness function of the form:

$$w(b,\gamma) \propto \exp\left(-\frac{(\log\gamma - \theta_\gamma - \phi_\gamma \log b)^2}{2V_\gamma} - \frac{(\log b - \theta_b - \phi_b \log\gamma)^2}{2V_b}\right). \quad \text{(C.1)}$$

Since transcriptional rates are the mechanisms by which cells control the level of transcription, we assume that mutations can affect b and γ independently. This corresponds to setting $c = 0$ in the model from Cope et al. Cope, Schraiber, and Pennell, 2025. Then, using their expression for the evolution matrix, (their F), we have:

$$H = \begin{pmatrix} \alpha_b(1+\phi_\gamma^2\omega) & -\alpha_b(\phi_b + \phi_\gamma\omega) \\ -\alpha_\gamma\left(\frac{\phi_b}{\omega} + \phi_\gamma\right) & \alpha_\gamma\left(1 + \frac{\phi_b^2}{\omega}\right) \end{pmatrix}, \quad \text{(C.2)}$$

We then consider two different models. In the first model, we assume that selection acts first on the decay rate, γ, and then on the mean spliced RNA value, μ_s, via adaptation of the burst size, b. This is equivalent to setting $\phi_\gamma = 0$ and $\phi_b = 1$, recalling that, in log space, $\log\mu_s = \log b - \log\gamma$. Note that, since the unspliced mean, μ_u, is given by $\frac{b}{\beta}$, and the spliced mean is given by $\frac{\mu_u\beta}{\gamma}$, the splicing rate β does not influenced the mean spliced counts, μ_s (recall that rates are in fitted in units of the transcriptional initiation rate). In this model, we also relabel $V_b \to V_\mu$ and $\theta_b \to \theta_\mu$, to emphasize that the pressure on b to adjust is equivalent to selection pressure on the mean spliced RNA level. So we have, for the fist, γ-constrained model:

$$w(b,\gamma) \propto \exp\left(-\frac{(\log\gamma - \theta_\gamma)^2}{2V_\gamma} - \frac{(\log b - \log\gamma - \theta_\mu)^2}{2V_\mu}\right). \quad \text{(C.3)}$$

For the γ-constrained model, H becomes:

$$H = \begin{pmatrix} \alpha_b & -\alpha_b \\ -\frac{\alpha_\gamma}{\omega} & \alpha_\gamma \left(1 + \frac{1}{\omega}\right) \end{pmatrix}, \tag{C.4}$$

and $\omega = \frac{V_\mu}{V_\gamma}$. In this γ-constrained model, we further assume that $V_\gamma \ll V_\mu$, such that we can neglect terms $O\left(\frac{1}{\omega}\right)$, giving:

$$H = \begin{pmatrix} \alpha_b & -\alpha_b \\ 0 & \alpha_\gamma \end{pmatrix}. \tag{C.5}$$

For our second model, we assume that b is tightly constrained, and γ is more free to vary and adjust to the optimum mean RNA level. This corresponds to setting $\phi_b = 0$ and $\phi_\gamma = 1$. The fitness function then becomes:

$$w(b, \gamma) \propto \exp\left(-\frac{(\log b - \log \gamma - \theta_\mu)^2}{2V_\mu} - \frac{(\log b - \theta_b)^2}{2V_b}\right), \tag{C.6}$$

where we have relabeled $V_\gamma \to V_\mu$ and $\theta_\gamma \to -\theta_\mu$ (note the change of sign in the first term has no effect due to the squaring). This gives, as above:

$$H = \begin{pmatrix} \alpha_b(1 + \omega) & -\alpha_b\omega \\ -\alpha_\gamma & \alpha_\gamma \end{pmatrix}. \tag{C.7}$$

For this, b-constrained, model, $\omega \equiv \frac{V_b}{V_\mu}$, and we assume $V_b \ll V_\mu$, giving:

$$H = \begin{pmatrix} \alpha_b & 0 \\ -\alpha_\gamma & \alpha_\gamma \end{pmatrix}. \tag{C.8}$$

We compare these models with a fully independent model (diagonal H), as well as more generic models. For all models, we fit a diagonal matrix for the stochastic term:

$$\Sigma = \begin{pmatrix} \sigma_b & 0 \\ 0 & \sigma_\gamma \end{pmatrix}, \tag{C.9}$$

which depends on the relative mutation rates of b and γ.

Recall Cope et al. also have:

$$\hat{\boldsymbol{X}} = \begin{pmatrix} \frac{\theta_b + \phi_b \theta_\gamma}{1 - \phi_b \phi_\gamma} \\ \frac{\theta_\gamma + \phi_\gamma \theta_b}{1 - \phi_b \phi_\gamma} \end{pmatrix}, \tag{C.10}$$

where, again, we have switched to our notation. Since either ϕ_b or ϕ_γ is zero in each of our models, the denominator is always equal to one. In particular, we have:

$$\hat{\boldsymbol{X}} = \begin{pmatrix} \theta_\mu + \theta_\gamma \\ \theta_\gamma \end{pmatrix}, \qquad \begin{pmatrix} \theta_b \\ -\theta_\mu + \theta_b \end{pmatrix}, \tag{C.11}$$

for the first and second models respectively. We then assume that the logarithms of b and γ evolve along the tree according to

$$d\boldsymbol{X}_t = -H(\boldsymbol{X}_t - \hat{\boldsymbol{X}}) + \Sigma d\boldsymbol{W}_t, \tag{C.12}$$

as in the independent case, where now

$$d\boldsymbol{X}_t = \begin{pmatrix} \log b \\ \log \gamma \end{pmatrix}, \tag{C.13}$$

and H, $\hat{\boldsymbol{X}}$ and Σ are as specified above. Note that the selection matrix H is the only structural difference between the two models.

Evolutionary parameter inference

We fit a mixture model where a fraction p_{wn} of genes are assumed to be drawn from a white-noise distribution. The rest of the genes are assumed to be generated from the relevant phylogenetic model, described in C.2.

The evolutionary parameters in H and Σ are assumed constant across genes, but the optimal values $\theta_\mu, \theta_{b,\gamma}$ are allowed to vary per gene, and are assumed to be drawn from a normal distribution centered at $\bar{\theta}^P_\mu$ and $\bar{\theta}^P_{b,\gamma}$ respectively, where the P denotes the population of optima over genes. The standard deviations of these populations of optima are also fit as part of the likelihood maximization ($\tau^P_{b,\gamma}, \tau^P_\mu$). We then analytically integrate over the possible values of θ_μ and $\theta_{b,\gamma}$.

The white-noise distribution is a two-dimensional Gaussian with mean $\hat{\boldsymbol{X}}$, with the values of $\theta_\mu, \theta_{b,\gamma}$ assumed to be drawn from the same distributions as for the phylogenetic model, and diagonal covariance matrix with entries $\sigma^{b,\gamma}_{wn}$. Since we analytically integrate over possible values of $\theta_{b,\gamma}$, this amounts to a new two-dimensional Gaussian distribution with means $\bar{\theta}^P_\mu$ and $\bar{\theta}^P_{b,\gamma}$, with covariance matrix given by $\mathrm{Cov}(\hat{\boldsymbol{X}})$ (see section C.5).

Overall, this gives an 11-parameter model, with 4 evolutionary parameters, 3 white-noise parameters (the two standard deviations and the mixture probability p_{wn}), and 4 optima distribution parameters. The likelihood of the data under the full model is optimized using `optimx`Nash and Varadhan, 2011.

Simulation results: Two-dimensional model

For the model comparison, we simulate 100 datasets under each of hypotheses 1 and 2 (decay rate and burst size-driven), using PCMBase Mitov et al., 2020. We draw random sets of true parameters for each simulation uniformly between the bounds. We then fit the simulated datasets using the procedure described above. The results for the decay-rate-constrained model are shown in Figure C.1, and the results for the burst-size-constrained model are shown in Figure C.2.

We also performed a model comparison by simulating data under both models, and fitting both datasets under each model. We show the distribution of AIC differences in favor of the correct model in Figure C.3. We show the corresponding accuracy of differentiating between models using the AIC value in Figure C.4. This is given by the fraction of model fits with AIC differences above each cutoff which would be attributed to the correct model.

Data fit details

We include the fitted parameter values for the two-parameter H, two-dimensional OU model in Table C.1. Note that the burst-size-constrained model hits the upper bound for p_{wn}, the probability for each gene to be pure noise in our mixture model.

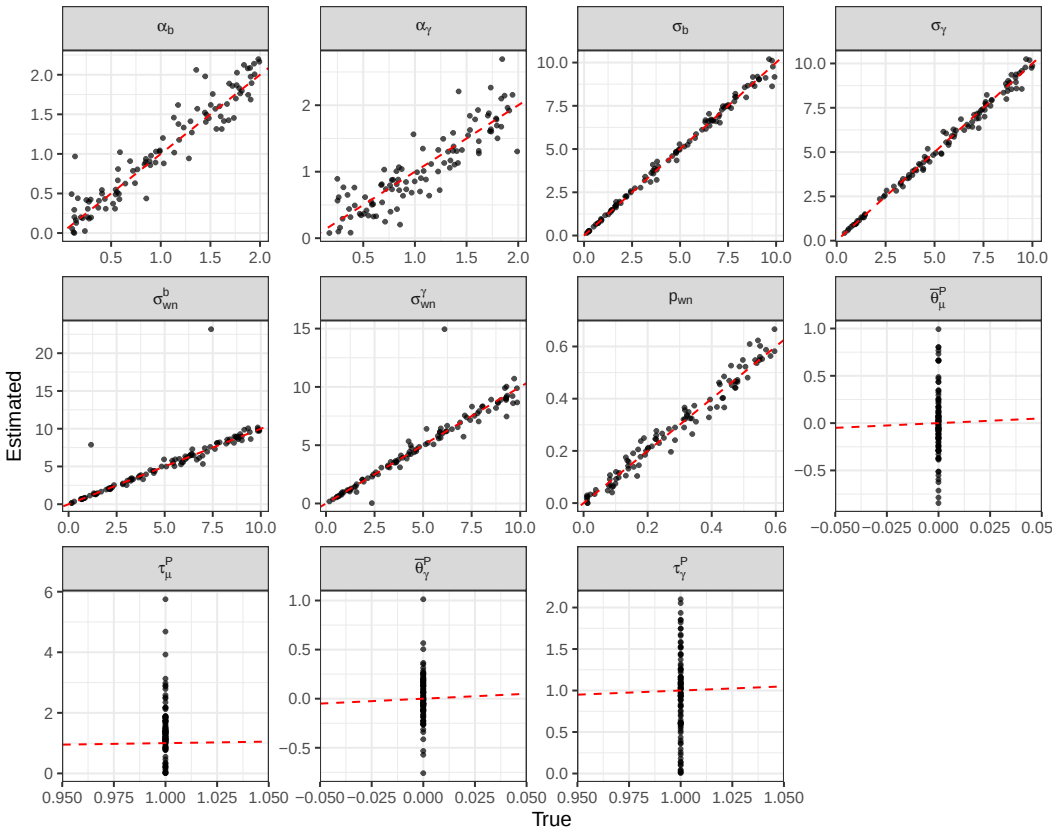


Figure C.1: Fitted vs true parameters for the decay-rate-constrained model.

Table C.1: Two-dimensional OU model fitted parameters

Model	α_b	α_γ	σ_b	σ_γ	p_{wn}	AIC
γ-constrained	31.5	2.33	0.923	1.06	0.149	2124
b-constrained	0.022	3.56	3.01	0.169	0.900	3186
Independent	1.36	1.59	0.651	0.831	0.329	2726

dN/dS calculations

To look at the signatures of selection for the 167 genes in our dataset across the six species phylogeny, we computed dN/dS values. We adopted a standard bioinformatic pipeline, using protein and cDNA sequences from Ensembl Dyer et al., 2024, protein alignment using MAFFT Katoh and Standley, 2013, codon based alignment with Pal2Nal Suyama, Torrents, and Bork, 2006 and dN/dS calculations using CODEML from PAML Yang, 2007; Álvarez-Carretero, Kapli, and Yang, 2023. Out of 167 genes, we obtained values for 159 gene ortholog groups which we plotted based on the bins of gene expression, shown in Figure C.5.

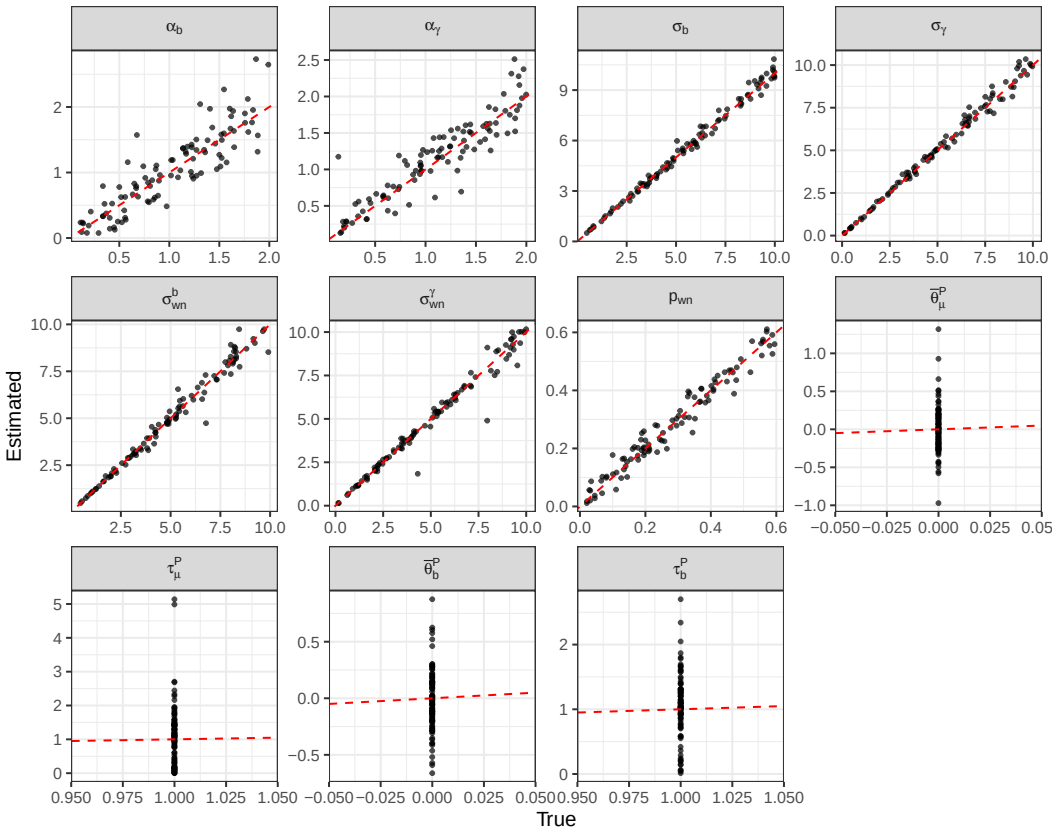
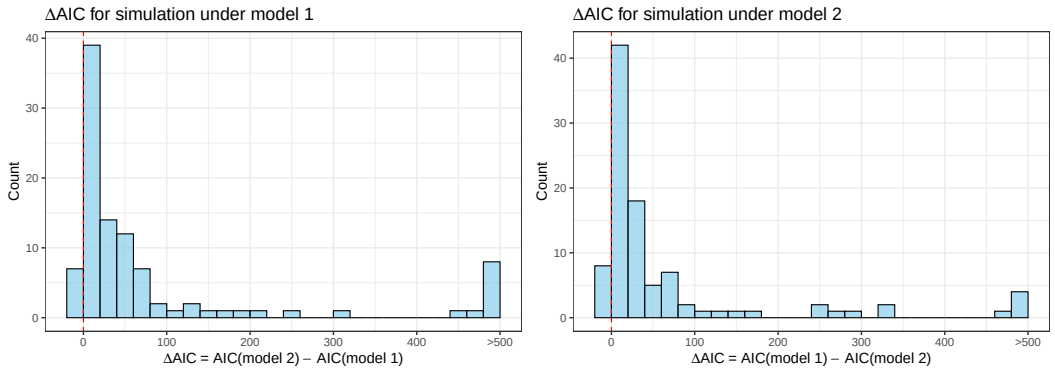


Figure C.2: Fitted vs true parameters for the burst-size-constrained model.

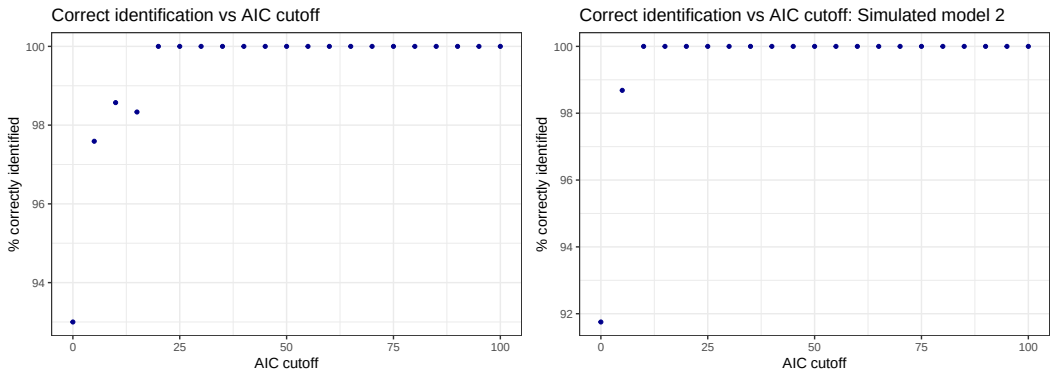


(a) (b)

Figure C.3: Distributions of AIC differences in favor of the correct model, for datasets simulated under the decay-rate-constrained model **(a)** and the burst-size-constrained model **(b)**, fitted under both models.

C.3 Independent evolution model

We also considered the simple case where, for each gene, the biophysical parameters, b, β, and γ, evolve independently. We assume that each gene also evolves



(a) (b)

Figure C.4: Accuracy of model identifications made at various AIC cutoffs, for datasets simulated under the decay-rate-constrained model **(a)** and the burst-size-constrained model **(b)**, fitted under both models. Accuracy is defined as the fraction of model fits with AIC differences above each cutoff that would be attributed to the correct model.

independently, but that all genes evolve according to the same evolutionary dynamics Chaix et al., 2008 Cope, Schraiber, and Pennell, 2025. This amounts to a shared selection matrix, H, and mutation matrix, Σ, in the OU evolution equation:

$$d\boldsymbol{X}_t = -H(\boldsymbol{X}_t - \hat{\boldsymbol{X}}) + \Sigma d\boldsymbol{W}_t, \tag{C.14}$$

where $\boldsymbol{X}$ is given by the logarithms of the three biophysical rates,

$$\boldsymbol{X}_t = \begin{pmatrix} \log b_t \\ \log \beta_t \\ \log \gamma_t \end{pmatrix}. \tag{C.15}$$

The assumed independence of the biophysical rates is enforced by diagonality of H and Σ. We therefore define parameter-specific selections rates, α, and stochastic standard deviations, σ, via:

$$H = \begin{pmatrix} \alpha_b & 0 & 0 \\ 0 & \alpha_\beta & 0 \\ 0 & 0 & \alpha_\gamma \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_b & 0 & 0 \\ 0 & \sigma_\beta & 0 \\ 0 & 0 & \sigma_\gamma \end{pmatrix}. \tag{C.16}$$

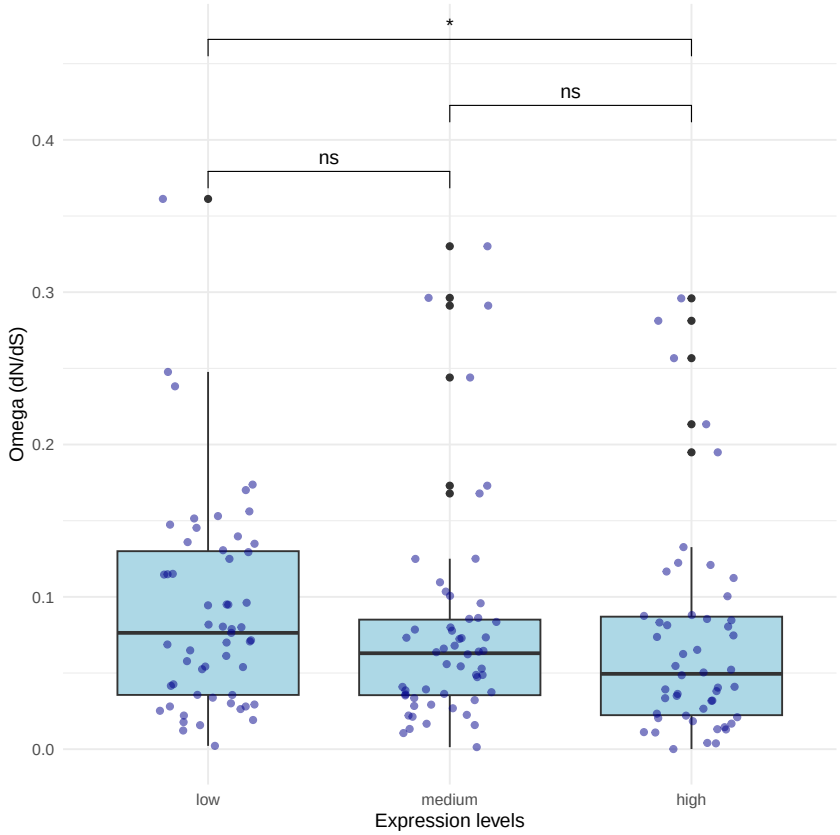


Figure C.5: dN/dS values for 159 genes across the six species tree, binned by expression level

If we assume that the root value is drawn from the stationary distribution, we have that:

$$(V_{\text{phylog}})_{ij} = \frac{\sigma^2}{2\alpha}\left(1 - e^{-2\alpha t_{ij}}\right). \tag{C.17}$$

Each gene is assumed to have its own optimum value, given by θ_g, drawn from a normal distribution, $N(\bar{\theta}, \tau^2)$. The parameters of this normal distribution are optimized during fitting, and the hyper-prior is specified along with the priors for the other model parameters. We can integrate over the possible values of θ_g for each gene, to give a new variance-covariance matrix given by:

$$V_{ij} = (V_{phylog})_{ij} + V_{\theta}. \tag{C.18}$$

Since initial fits gave high values for the diagonal elements of H (indicating low phylogenetic signal for many genes), we fitted a mixture model, following Chaix et al. Chaix et al., 2008. For the 'outlier' distribution, we used a white-noise (wn) (normal) distribution, centered at $\bar{\theta}$, with variance $\sigma^2_{wn} + \tau^2$, to represent a process of rapid mean-reversion with negligible phylogenetic signal. As in Chaix et al., 2008, we then optimized for a total likelihood given by:

$$\mathrm{LL} = p_{\mathrm{wn}}\mathrm{LL}_{out} + (1 - p_{\mathrm{wn}})\mathrm{LL}_{in}, \tag{C.19}$$

where LL_{in} is the likelihood of the original model. We implemented this via an MCMC in rstan. The simulation results for MCMC fits to the selection strength, standard deviations, and white noise probabilities are shown in Figure C.6.

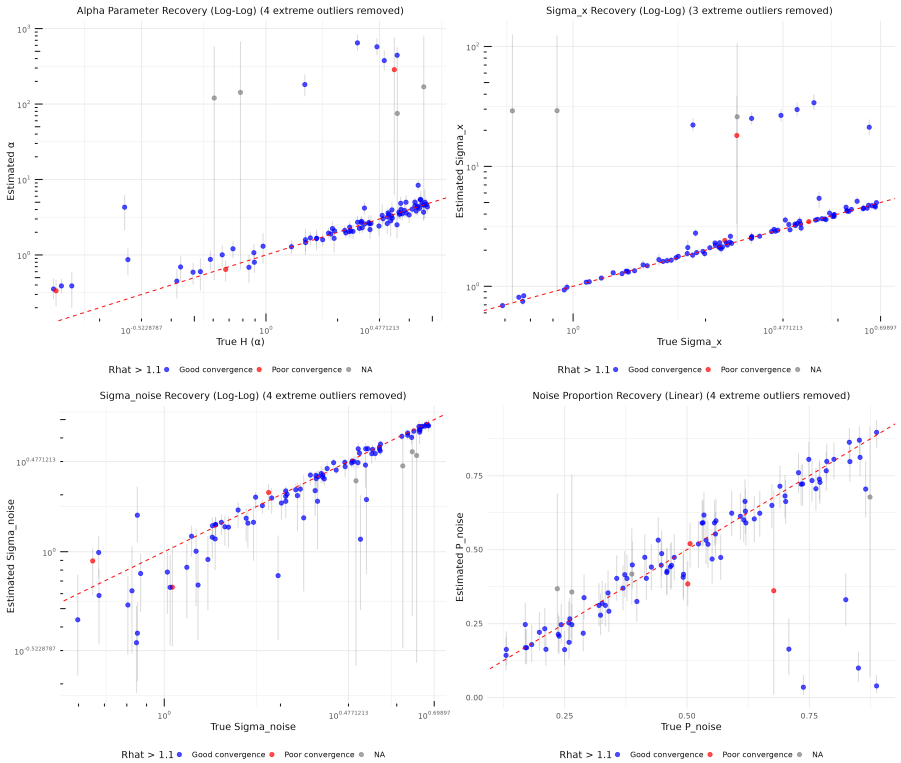


Figure C.6: Estimated vs true parameters for simulated data under the independent model described in Section C.3. The mean posterior values for the MCMC, are plotted against the known, simulated parameters. The red points showed poor convergence.

Independent model results

We applied the model to the data described in Section C.1. The posterior distributions for the most significant parameters are shown in Figure C.7. The posteriors for all of the parameters are shown in Figure C.8.

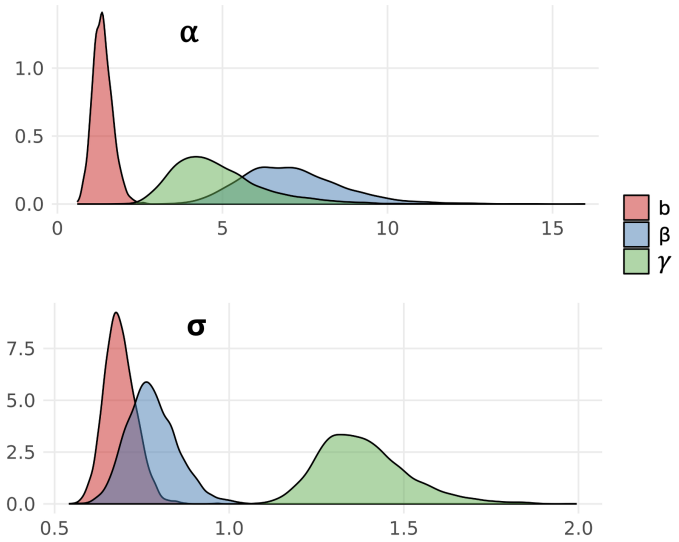


Figure C.7: Posterior distributions for the selection strength (α) and the mutation strength (σ), for the independent evolution of the logarithms of the biophysical parameters: burst size (b, red), splicing rate (β, blue), and decay rate (γ, green).

C.4 Three-parameter H models

We also investigated models of the above form, but with the off-diagonal H terms allowed to vary from the corresponding diagonal value. This is equivalent to allowing ϕ_b and ϕ_γ respectively to differ from one (see Section C.2). This gives selection matrices:

$$H_b = \begin{pmatrix} \alpha_b & -\phi_b\alpha_b \\ 0 & \alpha_\gamma \end{pmatrix}, \qquad H_\gamma = \begin{pmatrix} \alpha_b & 0 \\ -\phi_\gamma\alpha_\gamma & \alpha_\gamma \end{pmatrix}, \tag{C.20}$$

and $\hat{X}$ values:

$$\hat{X}_b = \begin{pmatrix} \theta_b + \phi_b\theta_\gamma \\ \theta_\gamma \end{pmatrix}, \qquad \hat{X}_\gamma = \begin{pmatrix} \theta_b \\ \theta_\gamma + \phi_\gamma\theta_b \end{pmatrix}, \tag{C.21}$$

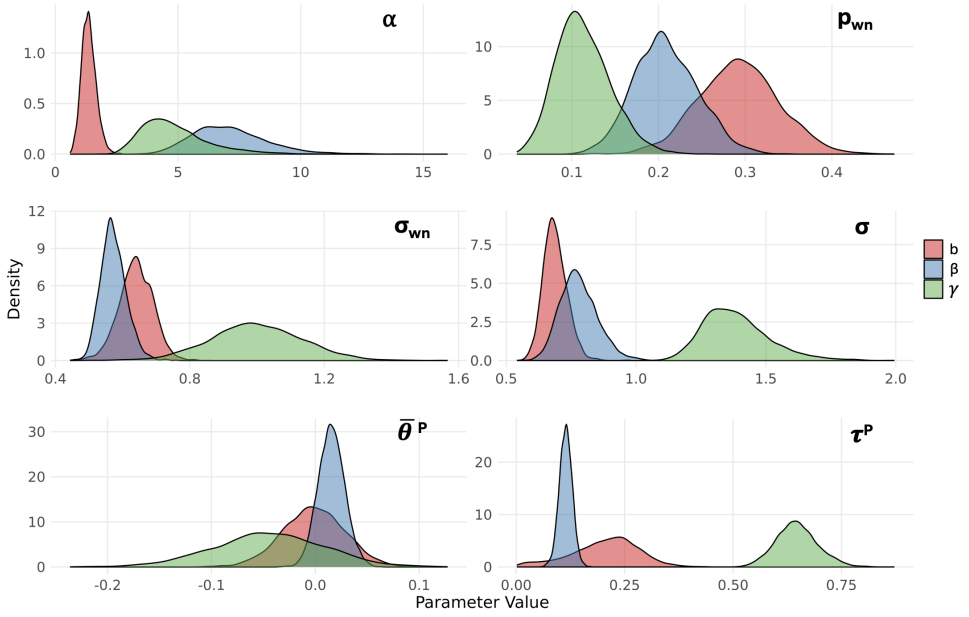


Figure C.8: Posterior distributions for the selection strength (α) and mutation strength (σ) in the independent OU model. The mixture model probability, p_{wn}, the white-noise distribution standard deviation, σ_{wn}, and the parameters of the assumed underlying distribution for the gene optima and white-noise means ($\bar{\theta}^P, \tau^P$) are also shown. These are the parameters governing the independent evolution of the logarithms of the biophysical parameters: burst size (b, red), splicing rate (β, blue), and decay rate (γ, green).

for the two models respectively. The other details of the models were identical to Section C.2. Although these models had better AIC scores than those discussed in the main text, because of the required additional parameter, and the superior interpretability of the two-dimensional models, we have discussed those more at length.

Simulations results: three-parameter models

As for the two-parameter H case, we simulate 100 datasets under each of hypotheses 1 and 2 (decay rate and burst size-driven), this time with the extra $\phi_{b,\gamma}$ parameters. The results for the three-parameter decay-rate-constrained model are shown in Figure C.9, and the results for the three-parameter burst-size-constrained model are shown in Figure C.10.

As before, we perform a simulated model comparison for the three-parameter H case. We show the distribution of AIC differences in favor of the true simulated model in Figure C.11. We show the corresponding accuracy of differentiating

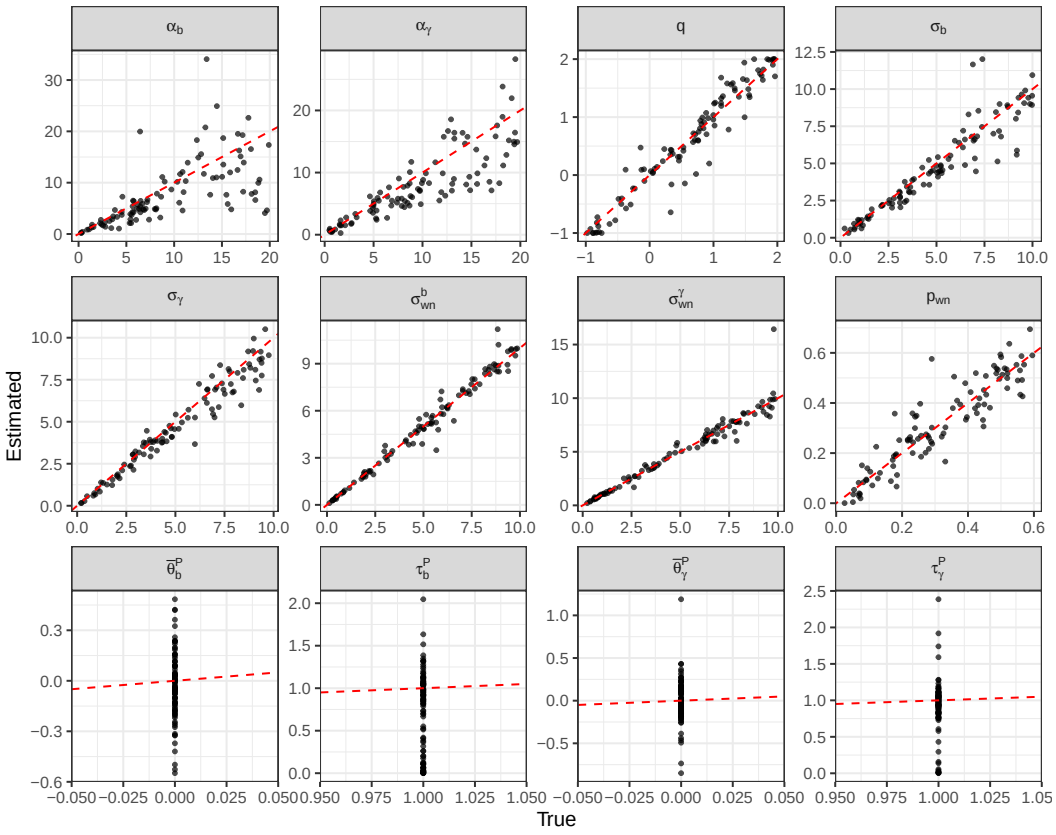


Figure C.9: Fitted vs true parameters for the decay-rate-constrained, three-parameter H model.

between models using the AIC value in Figure C.12.

Data fits: three-parameter models

We include the fitted parameter values for the three-parameter H models in Table C.2. Note that the AIC is again much better for the decay-rate-constrained model, and p_{wn} goes to the upper bound in the burst-size-constrained model.

Table C.2: Three-parameter model fitted parameters

Model	α_b	α_γ	$\phi_{b,\gamma}$	σ_b	σ_γ	p_{wn}	AIC
γ-constrained	130.2	4.21	0.717	2.20	1.33	0.226	1972
b-constrained	1.91	4.68	1.84	0.567	0.00676	0.9	3166

C.5 Theta integration

For a vector of traits (e.g. (b, β, γ)), we consider the distribution of values for a single species given a constant optimum, θ. We define:

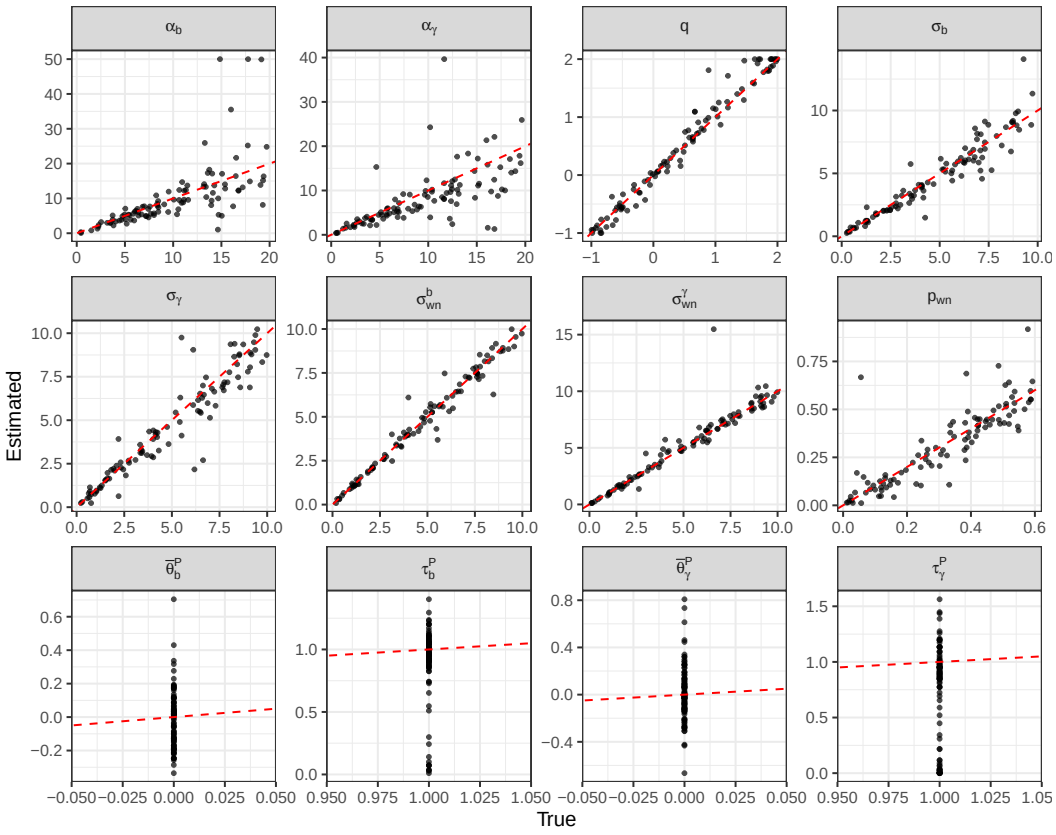
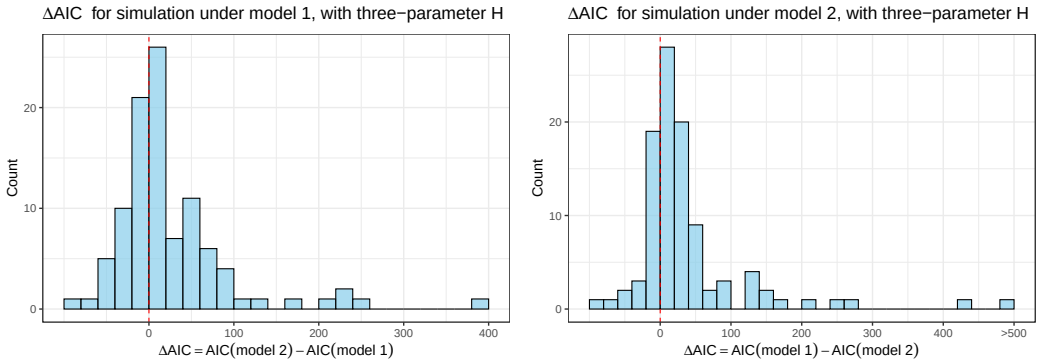
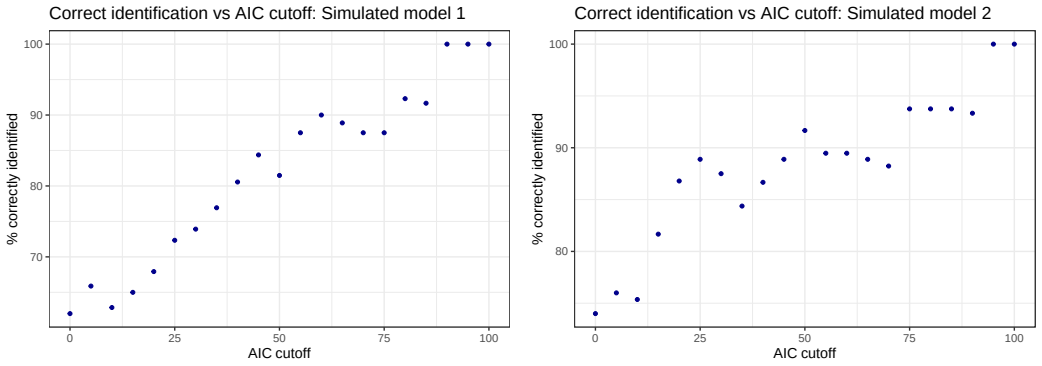


Figure C.10: Fitted vs true parameters for the burst-size-constrained, three-parameter *H* model.



(a) (b)

Figure C.11: Distributions of AIC differences in favor of the simulated model for datasets simulated under the three-parameter *H* versions of the decay-rate-constrained model (**left**) and the burst-size-constrained model (**right**), fitted under both models.



(a) (b)

Figure C.12: Accuracy of model identifications made at various AIC cutoffs for datasets simulated under the three-parameter H versions of the decay-rate-constrained model (**left**) and the burst-size-constrained model (**right**), fitted under both models. The accuracy is defined as the fraction of model fits with AIC differences above each cutoff which would be attributed to the correct model.

$$\boldsymbol{f}(\boldsymbol{x}_t, t) \equiv e^{\hat{H}t}\boldsymbol{x}_t, \tag{C.22}$$

where $\hat{H}$ is the selection matrix. As in the one-trait case, we have:

$$d\boldsymbol{f} = e^{\hat{H}t}d\boldsymbol{x}_t + \hat{H}e^{\hat{H}t}\boldsymbol{x}_t. \tag{C.23}$$

Then, using the OU equation:

$$d\boldsymbol{x}_t = -\hat{H}(\boldsymbol{x}_t - \boldsymbol{\theta})dt + \hat{\Sigma}d\boldsymbol{W}_t, \tag{C.24}$$

where $d\boldsymbol{W}$ is a vector whose components are independent instantiations of the random variable, dW (Wiener process), and correlations between changes in the different variables of $\boldsymbol{x}_t$ can be introduced via off-diagonal elements in $\boldsymbol{\Sigma}$. This gives:

$$\begin{aligned} d\boldsymbol{f} &= e^{\hat{H}t}\left(-\hat{H}(\boldsymbol{x}_t - \boldsymbol{\theta})dt + \hat{\Sigma}d\boldsymbol{W}_t\right) + \hat{H}e^{\hat{H}t}\boldsymbol{x}_t \\ &= e^{\hat{H}t}\hat{H}\boldsymbol{\theta}dt + e^{\hat{H}t}\hat{\Sigma}d\boldsymbol{W}_t, \end{aligned} \tag{C.25}$$

where, for the second equality, we have used the commutation of $e^{\hat{H}t}$ and $\hat{H}$. Using the convenient properties of the matrix exponential, we can solve this via:

$$\begin{aligned}\boldsymbol{f} &= \boldsymbol{f}(0) + \int_0^t d\boldsymbol{f} \\ &= \boldsymbol{x_0} + \left[\hat{H}\hat{H}^{-1}e^{\hat{H}s}\boldsymbol{\theta}\right]_0^t + \int_0^t e^{\hat{H}s}\hat{\Sigma}\boldsymbol{dW}_s \\ &= \boldsymbol{x_0} + \left(e^{\hat{H}t} - \hat{I}\right)\boldsymbol{\theta} + \int_0^t e^{\hat{H}s}\hat{\Sigma}\boldsymbol{dW}_s, \end{aligned} \tag{C.26}$$

for s our dummy time variable. Then, by the definition of $\boldsymbol{f}$, (and the commutation of $\hat{H}$ with itself and its inverse), we have:

$$\begin{aligned}&\boldsymbol{x}_t = e^{-\hat{H}t}\boldsymbol{f}(\boldsymbol{x}_t, t) \\ &\implies \boldsymbol{x}_t = e^{-\hat{H}t}\boldsymbol{x_0} + \left(\hat{I} - e^{-\hat{H}t}\right)\boldsymbol{\theta} + \int_0^t e^{-\hat{H}(t-s)}\hat{\Sigma}\boldsymbol{dW}_s. \end{aligned} \tag{C.27}$$

Note that in the limit $t \to \infty$, for positive definite $\hat{H}$, the first term vanishes, and we are left with a linear sum of the independent normal random variables, dW_s, which make up $\boldsymbol{dW}_s$. Since each component of $\boldsymbol{x}_t$ is a linear comination of the dW values, the vector $\boldsymbol{x}_t$ has a multivariate normal distribution. We can proceed to calculate the moments of this distribution. We find, for a single tip, i, at time $t = t_i$,

$$\mathbb{E}(\boldsymbol{x}_{t_i}) = e^{-\hat{H}t_i}\mathbb{E}(\boldsymbol{x_0}) + \left(\hat{I} - e^{-\hat{H}t_i}\right)\boldsymbol{\theta}, \tag{C.28}$$

since $\mathbb{E}(\boldsymbol{dW_t}) = \boldsymbol{0}$ for all t, and $\boldsymbol{\theta}$ and $\hat{H}$ are fixed parameters of the system. We now consider the variance-covariance of the different traits at a single tip. We use Greek indices to indicate the different traits at a tip, (and Latin indices to indicate the tip (extant species)). Since $\boldsymbol{x_0}$ and the stochastic path integral are independent random variables (and the middle term is constant), we have:

$$\mathrm{Var}(\boldsymbol{x}_{t_i}) = \mathrm{Var}\left(e^{-\hat{H}t_i}\boldsymbol{x_0}\right) + \mathrm{Var}\left(\int_0^{t_i} e^{-\hat{H}(t_i-s)}\hat{\Sigma}\boldsymbol{dW}_s\right). \tag{C.29}$$

First we focus on the $\boldsymbol{x_0}$ term. By simple properties of matrices we have:

$$\text{Var}\left(e^{-\hat{H}t_i}\boldsymbol{x_0}\right) = e^{-\hat{H}t_i}\hat{V}_{x_0}\left(e^{-\hat{H}t_i}\right)^T, \tag{C.30}$$

for $\hat{V}_{x_0}$ the variance-covariance matrix for $\boldsymbol{x_0}$. Then we focus on the integral term. As above, the expectations of all components of $\boldsymbol{dW}$ are zero. For generic traits α and β we therefore have:

$$\text{Cov}_i^{integral}(\alpha,\beta) = \mathbb{E}\left[\left(\int_0^{t_i} e^{-\hat{H}(t_i-s)}\hat{\Sigma}\boldsymbol{dW}_s\right)_\alpha\left(\int_0^{t_i} e^{-\hat{H}(t_i-u)}\hat{\Sigma}\boldsymbol{dW}_u\right)_\beta\right]. \tag{C.31}$$

Since each individual dW_s is assumed independent from the last, and has zero expectation, we have:

$$\begin{aligned}
\text{Cov}_i^{integral}(\alpha,\beta) &= \mathbb{E}\left[\int_0^{t_i}\left(e^{-\hat{H}(t_i-s)}\Sigma\boldsymbol{dW}_s\right)_\alpha\left(e^{-\hat{H}(t_i-s)}\Sigma\boldsymbol{dW}_s\right)_\beta\right] \\
&= \mathbb{E}\left[\int_0^{t_i}\sum_{\gamma\nu\epsilon\mu} e^{-\hat{H}(t_i-s)}{}_{\alpha\gamma}\hat{\Sigma}_{\gamma\nu}(\boldsymbol{dW}_s)_\nu e^{-\hat{H}(t_i-s)}{}_{\beta\epsilon}\hat{\Sigma}_{\epsilon\mu}(\boldsymbol{dW}_s)_\mu\right] \\
&= \int_0^{t_i}\sum_{\gamma\nu\epsilon\mu} e^{-\hat{H}(t_i-s)}{}_{\alpha\gamma}\hat{\Sigma}_{\gamma\nu}e^{-\hat{H}(t_i-s)}{}_{\beta\epsilon}\hat{\Sigma}_{\epsilon\mu}\delta_{\mu\nu}ds \\
&= \int_0^{t_i}\sum_{\gamma\epsilon\mu} e^{-\hat{H}(t_i-s)}{}_{\alpha\gamma}\hat{\Sigma}_{\gamma\mu}e^{-\hat{H}(t_i-s)}{}_{\beta\epsilon}\hat{\Sigma}_{\epsilon\mu}ds \\
&= \sum_{\gamma\epsilon\mu}\hat{\Sigma}_{\gamma\mu}\hat{\Sigma}_{\epsilon\mu}\int_0^{t_i} e^{-\hat{H}(t_i-s)}{}_{\alpha\gamma}e^{-\hat{H}(t_i-s)}{}_{\beta\epsilon}ds \\
&= \int_0^{t_i}\left[e^{-\hat{H}(t_i-s)}\hat{\Sigma}\hat{\Sigma}^T(e^{-\hat{H}(t_i-s)})^T\right]_{\alpha\beta}ds.
\end{aligned} \tag{C.32}$$

For notes on solving this, see Jonathan Goodman, NYU n.d. Next, we consider the covariance between traits α and β, between two *different* species i and j. We have:

$$\text{Cov}_{ij\alpha\beta} = \text{Cov}_{ij}^{integral}(\alpha,\beta) + \left[e^{-\hat{H}t_i}\hat{V}_{x_0}\left(e^{-\hat{H}t_j}\right)^T\right]_{\alpha\beta}. \tag{C.33}$$

Recall that the stationary variance satisfies:

$$\hat{V}_{stat}\hat{H}^T + \hat{H}\hat{V}_{stat} = \hat{\Sigma}\hat{\Sigma}^T. \tag{C.34}$$

We have already shown that the stationary distribution is multivariate Gaussian. If $\boldsymbol{x_0}$ is drawn from this distribution, it will therefore be distributed via:

$$\boldsymbol{x_0} \sim N(\boldsymbol{\theta}, \hat{V}_{stat}), \tag{C.35}$$

and the entire process is also multivariate normal given by:

$$\boldsymbol{x}_t \sim N(\boldsymbol{\theta}, \hat{V}), \tag{C.36}$$

using again that $\hat{H}^{-1}$ and $e^{-\hat{H}t}$ commute to calculate the mean, and $\hat{V}$ is the full covariance matrix. If $\boldsymbol{\theta}$ is another random variable chosen independently from a prior, and the previous distribution of $\boldsymbol{x}_t$ is in fact $\boldsymbol{x}_t|\boldsymbol{\theta}$, we consider again:

$$\boldsymbol{x}_t = e^{-\hat{H}t}\boldsymbol{x_0} + \left(\hat{I} - e^{-\hat{H}t}\right)\boldsymbol{\theta} + \int_0^t e^{-\hat{H}(t-s)}\hat{\Sigma}\boldsymbol{dW}_s. \tag{C.37}$$

Since the last term is linear combinations of the same dW, it is clearly multivariate normal, and independent from the other two terms. $\boldsymbol{\theta}$ is drawn from a normal prior, $N(\mu_\theta, \hat{V}_\theta)$. This determines the distribution of the r.v. $\boldsymbol{x_0}$. This distribution is determined by imagining the stationary version of the equation, where $t \to \infty$:

$$\boldsymbol{x}_{stat} = \boldsymbol{\theta} + \lim_{t\to\infty} \int_0^t e^{-\hat{H}(t-s)}\hat{\Sigma}\boldsymbol{dW}_s. \tag{C.38}$$

This stationary distribution clearly would be multivariate normal with expectation $\boldsymbol{\theta}$ and some stationary variance $\hat{V}_{stat}$, which depends on the model parameters. To show that $\boldsymbol{x_0}$ is multivariate normal with $\boldsymbol{\theta}$, consider the sum of the two variables. Consider $\boldsymbol{x_0} = \boldsymbol{\theta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim N(0, \hat{V}_{stat})$. For deterministic α, β and *independent* $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$, we have $\alpha\boldsymbol{x_0} + \beta\boldsymbol{\theta} = (\alpha + \beta)\boldsymbol{\theta} + \alpha\boldsymbol{\epsilon}$. Clearly, the whole of $\boldsymbol{x}_t$ is still a multivariate normal random variable. The expected value is given by:

$$\begin{aligned}\mathbb{E}(\boldsymbol{x}_t) &= e^{-\hat{H}t}\mathbb{E}(\boldsymbol{x_0}) + \left(\hat{I} - e^{-\hat{H}t}\right)\boldsymbol{\theta} \\ &= e^{-\hat{H}t}\mu_\theta + \left(\hat{I} - e^{-\hat{H}t}\right)\mu_\theta \\ &= \mu_\theta, \end{aligned} \tag{C.39}$$

where we have used that $\mathbb{E}(\boldsymbol{dW}) = \mathbf{0}$. Considering $\boldsymbol{\theta}$ as a random variable changes the variance-covariance matrix of the distribution, since now we have to consider:

$$\mathrm{Var}_i\left(e^{-\hat{H}t_i}\boldsymbol{x_0} + \left(\hat{I} - e^{-\hat{H}t_i}\right)\boldsymbol{\theta}\right), \tag{C.40}$$

retaining the separation of this term from the integral term, since the two are independent. Considering the decomposition above, we get:

$$\mathrm{Var}_i\left(e^{-\hat{H}t}(\boldsymbol{\theta} + \boldsymbol{\epsilon}) + \left(\hat{I} - e^{-\hat{H}t}\right)\boldsymbol{\theta}\right) = \mathrm{Var}_i\left(e^{-\hat{H}t}\boldsymbol{\epsilon} + \boldsymbol{\theta}\right). \tag{C.41}$$

Considering this, trait-component-wise, we have:

$$\begin{aligned}\left[\mathrm{Var}_i\left(e^{-\hat{H}t}\boldsymbol{\epsilon} + \boldsymbol{\theta}\right)\right]_{\alpha\beta} &= \left\langle\left(e^{-\hat{H}t}\boldsymbol{\epsilon} + \boldsymbol{\theta}\right)_\alpha\left(e^{-\hat{H}t}\boldsymbol{\epsilon} + \boldsymbol{\theta}\right)_\beta\right\rangle \\ &\quad - \left\langle\left(e^{-\hat{H}t}\boldsymbol{\epsilon} + \boldsymbol{\theta}\right)_\alpha\right\rangle\left\langle\left(e^{-\hat{H}t}\boldsymbol{\epsilon} + \boldsymbol{\theta}\right)_\beta\right\rangle \\ &= \left\langle\left(\sum_\gamma (e^{-\hat{H}t})_{\alpha\gamma}\boldsymbol{\epsilon}_\gamma + \boldsymbol{\theta}_\alpha\right)\left(\sum_\mu (e^{-\hat{H}t})_{\beta\mu}\boldsymbol{\epsilon}_\mu + \boldsymbol{\theta}_\beta\right)\right\rangle - \mu_\theta^2\mathbf{1} \\ &= \left\langle\sum_\gamma (e^{-\hat{H}t})_{\alpha\gamma}\boldsymbol{\epsilon}_\gamma \sum_\mu (e^{-\hat{H}t})_{\beta\mu}\boldsymbol{\epsilon}_\mu + \boldsymbol{\theta}_\alpha\boldsymbol{\theta}_\beta\right\rangle - \mu_\theta^2\mathbf{1} \\ &= \sum_{\gamma\mu}(e^{-\hat{H}t})_{\alpha\gamma}(e^{-\hat{H}t})_{\beta\mu}\langle\boldsymbol{\epsilon}_\gamma\boldsymbol{\epsilon}_\mu\rangle + \hat{V}_{\boldsymbol{\theta}} \\ &= (e^{-\hat{H}t_i})\hat{V}_{stat}(e^{-\hat{H}t_i})^T + \hat{V}_{\boldsymbol{\theta}}. \end{aligned} \tag{C.42}$$

Since the integral term is unchanged, this simply represents summing the original variance-covariance matrix with the variance-covariance matrix for $\boldsymbol{\theta}$. Since the $\boldsymbol{x_0}$ and $\boldsymbol{\theta}$ values are common between species, the components of $\mathrm{Cov}_{ij\alpha\beta}$, between

different species i and j, will also pick up these same covariance terms. This will give the full, overall formula:

$$\mathrm{Cov}_{ij\alpha\beta} = \mathrm{Cov}_{ij}^{integral}(\alpha,\beta) + \left[e^{-\hat{H}t_i}\hat{V}_{x_0}\left(e^{-\hat{H}t_j}\right)^T\right]_{\alpha\beta} + \left[\hat{V}_{\boldsymbol{\theta}}\right]_{\alpha\beta}. \tag{C.43}$$

This allows us to effectively integrate over a Gaussian prior on the optima, by using the prior distribution covariance in $\hat{V}_\theta$, and considering a new Gaussian distribution for $\boldsymbol{x_t}$ with a covariance given by Eq. C.43.