Biophysical modeling with variational autoencoders for bimodal, single-cell RNA sequencing data

Maria Carilli[1,*], Gennady Gorin[2,*], Yongin Choi[3,4], Tara Chari[1], and Lior Pachter[1,5,**]

[1]Division of Biology and Biological Engineering, California Institute of Technology
[2]Division of Chemistry and Chemical Engineering, California Institute of Technology
[3]Biomedical Engineering Graduate Group, University of California, Davis
[4]Genome Center, University of California, Davis
[5]Department of Computing and Mathematical Sciences, California Institute of Technology
[*]These two authors contributed equally to this work.
[**]lpachter@caltech.edu

May 2, 2023

Abstract

We motivate and present *biVI*, which combines the variational autoencoder framework of *scVI* with biophysically motivated, bivariate models for nascent and mature RNA distributions. While previous approaches to integrate bimodal data via the variational autoencoder framework ignore the causal relationship between measurements, *biVI* models the biophysical processes that give rise to observations. We demonstrate through simulated benchmarking that *biVI* captures cell type structure in a low-dimensional space and accurately recapitulates parameter values and copy number distributions. On biological data, *biVI* provides a scalable route for identifying the biophysical mechanisms underlying gene expression. This analytical approach outlines a generalizable strategy for treating multimodal datasets generated by high-throughput, single-cell genomic assays.

1 Main

Advances in experimental methods for single-cell RNA sequencing (scRNA-seq) allow for the simultaneous quantification of multiple cellular species, such as nascent and mature transcriptomes [1,2], surface [3–5] and nuclear [6] proteomes, and chromatin accessibility [7,8]. While these rich datasets have the potential to enable unprecedented insight into cell type and state in development and disease, joint analyses of distinct modalities remain challenging. We show that principled biophysical "integration" of multimodal datasets can be achieved through parameterization of interpretable mechanistic models [9], scalable to measurements made for thousands of genes in tens of thousands of cells [10].

Recent approaches to integrate and reduce the dimensionality of multimodal single-cell genomics data have leveraged advances in machine learning [11–13]. For example, the popular tool *scVI* is a variational autoencoder (VAE) that uses neural networks to encode scRNA-seq counts to a low-dimensional representation. This is decoded by another neural network to a set of cell- and gene- specific parameters for conditional likelihood distributions of observed counts. These distributions are chosen *post hoc* to be consistent with the discrete, over-dispersed nature of scRNA-seq counts, but can be derived from biophysical models (Section S1). Extensions of *scVI* to bimodal data have been attempted for protein [11] and chromatin measurements [14] by jointly encoding data modalities to a single latent space, then employing two decoding networks to produce parameters for *independent* conditional likelihoods specific to each datatype. Nascent and mature transcripts, available by realigning existing scRNA-seq reads [1,2], could be similarly treated (Figure 1a). However, using independent conditional likelihoods for bimodal measurements derived from the same gene ignores the inherent causality between observations and has no biophysical basis: the generative model is merely part of a neural "black box" used to summarize data.

Nevertheless, good causal model candidates are available: for example, Figure 1b illustrates the extensively validated [15–17] bursty model of transcription. Nascent RNA molecules are produced in geometrically distributed bursts with mean b at constant rate k and spliced at rate β to produce mature molecules, which are degraded with constant rate γ. While the joint steady-state distribution induced by the bursty model is analytically intractable [18], we have previously shown that it can be approximated by a set of basis functions with neural-network learned weights [19].

We introduce *biVI*, a strategy that adapts *scVI* to work with well-characterized stochastic models of transcription. First, we propose several models, formalized by chemical master equations (CMEs), that could give rise to bivariate count distributions for nascent and mature transcripts. We then use the bivariate, CME-derived distribution as the conditional data likelihood distribution for nascent and mature counts (Figure 1c). The inferred conditional likelihood parameters thus have biophysical interpretations as part of a mechanistic model of transcriptional dynamics. Although we focus on the bursty model, *biVI* implements the closed-form constitutive and extrinsic noise models previously discussed in the literature [9,20,21] (derivations in Section S1 and diagrams in Figures S1 and S2).

After using simulations to show that *biVI* models, when compared to *scVI*, better recapitulate ground-truth distributions and achieve similar clustering of cell types' latent representations (Figures S3, S4, S5, and Section S6.7), we applied *biVI* and *scVI* to experimental data [22] (Section 2.6) from mouse brain tissue [22]. As shown in Figure 2a-b, *biVI* recapitulates empirical distribution shapes better than *scVI* (Section 3) while allowing for interpretation of cell-specific parameters to determine *how* genes are regulated (Section S8). For example, in Figure 2c-d, we illustrate that the upregulation of markers *Foxp2* and *Rorb* can be ascribed to an increase in burst size. We

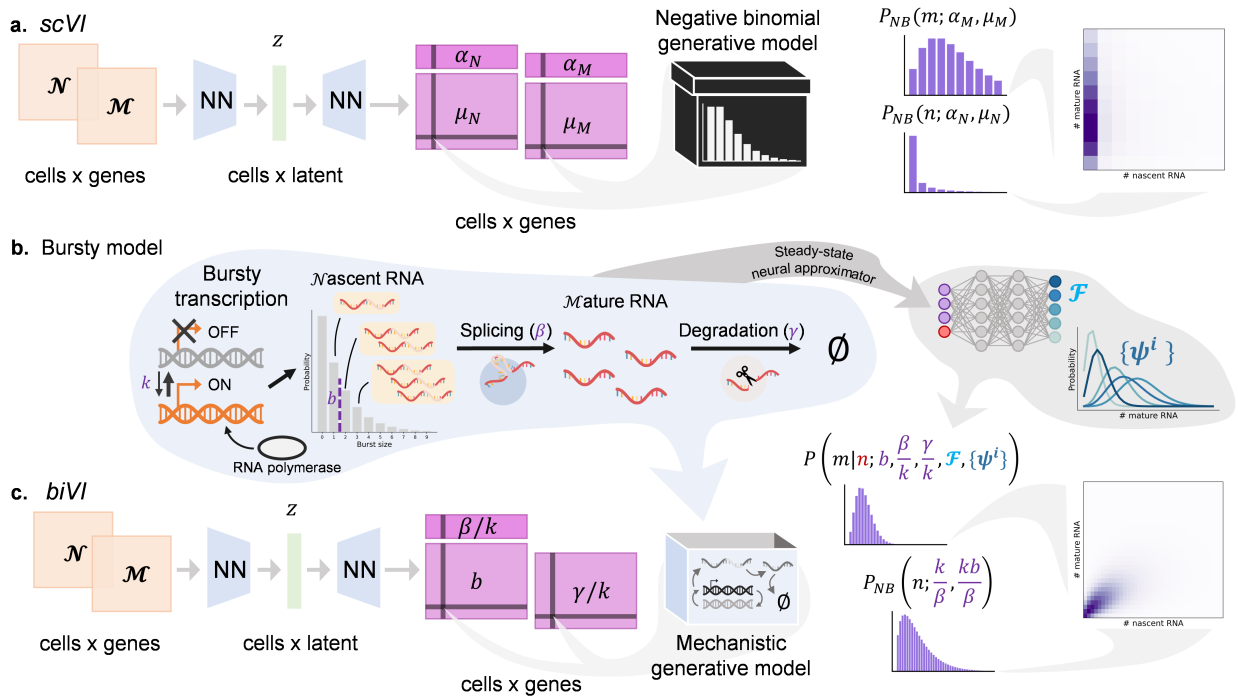

Figure 1: *biVI* reinterprets and extends *scVI* to infer biophysical parameters. **a.** *scVI* can take in concatenated nascent ($\mathcal{N}$) and mature ($\mathcal{M}$) RNA count matrices, encode each cell to a low-dimensional space z, and learn per-cell parameters μ_N and μ_M and per-gene parameters α_N and α_M for independent nascent and mature count distributions. This approach is not motivated by any specific biophysical model. **b.** A schematic of the telegraph model of transcription: a gene locus has the on rate k, the off rate k_{off}, and the RNA polymerase binding rate k_{RNAP}. Nascent RNA molecules are produced in geometrically distributed bursts with mean $b = k_{\text{RNAP}}/k_{off}$, which are spliced at a constant rate β and degraded at a constant rate γ. Although there is no closed-form solution, this model's steady-state distribution can be approximated by a pre-trained neural network $\mathcal{F}$ and a set of basis functions $\{\psi^i\}$. **c.** *biVI* can take in nascent and mature count matrices, produce a low-dimensional representation for each cell, and output per-cell parameters b and γ/k, as well as the per-gene parameters β/k, for a mechanistically motivated joint distribution of nascent and mature counts.

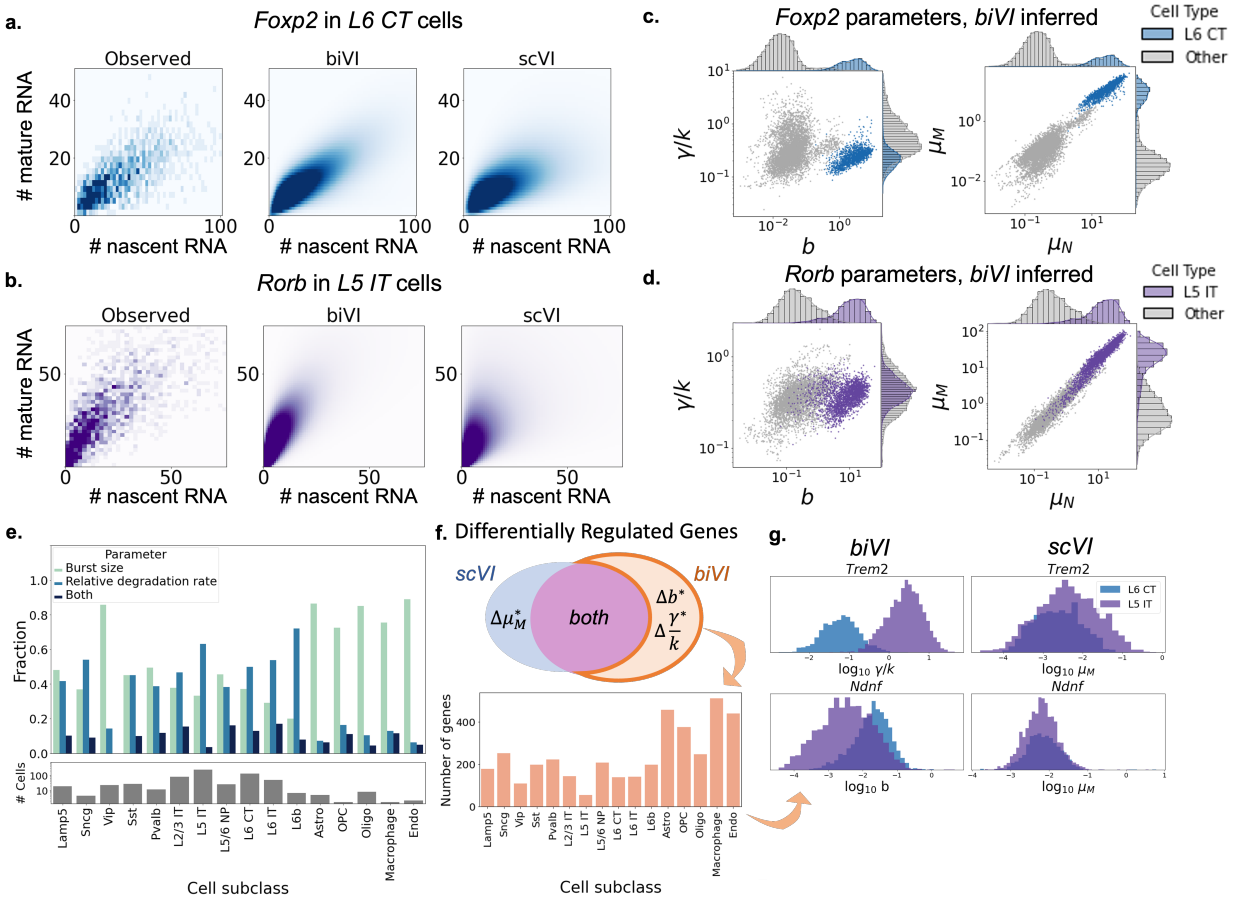

Figure 2: *biVI* successfully fits single-cell neuron data and suggests the biophysical basis for expression differences. **a.-b.** Observed, *scVI*, and *biVI* reconstructed distributions of *Foxp2*, a marker gene for L6 CT (layer 6 corticothalamic) cells, and *Rorb*, a marker gene for L5 IT (layer 5 intratelencephalic) cells, restricted to respective cell type. **c.-d.** Cell-specific parameters inferred for *Foxp2* and *Rorb* demonstrate identifiable differences in means and parameters in the marked cell types. **e.** Cell subclasses show different modulation patterns, with especially pronounced distinctions in non-neuronal cells (top: fractions of genes exhibiting differences in each parameter; bottom: number of cells in each subclass). **f.** *biVI* allows the identification of cells which exhibit differences in burst size or relative degradation rate, without necessarily demonstrating differences in mature mean expression. Hundreds of genes demonstrate this modulation behavior, with variation across cell subclasses. **g.** Histograms of *biVI* parameters and *scVI* mature means for two genes that exhibit parameter modulation without identifiable mature mean modulation. *Trem2* (top) shows differences in the degradation rate in *L5 IT* cells, whereas *Ndnf* (bottom) shows differences in burst size in *L6 CT* cells.

generalize this approach in Figure 2e, which shows the fraction of identified genes in each cell subclass that exhibited significant differences in burst size, relative degradation rate, or both (Section 2.8). Interesting trends across cell subclasses begin to emerge: neuronal cells appear to regulate gene expression via a mix of regulatory strategies, while non-neuronal cells seem to preferentially modulate burst size.

Finally, *biVI* can identify distributional differences which do not result in mean expression changes (Section 2.8, Figure 2g-h). For some cell subclasses, there were several hundred such genes, interesting targets for follow-up experimental investigation. For example, the gene *Ndnf*, which codes for the neuron derived neurotrophic factor NDNF, demonstrated a statistically significant difference in the *biVI* inferred burst size, but not *scVI* inferred mature mean, in the neuronal subclass L6 CT (Figure 2i, top row). NDNF promotes the growth, migration, and survival of neurons [23]; characterizing its regulatory patterns could help elucidate its role in neuronal maintenance. As another example, the relative degradation rate of the gene coding for the triggering receptor expressed on myeloid cells-2 (TREM2), variants of which are strongly associated with increased risk of Alzheimer's disease [24], was found to be greater in the neuronal L5 IT subclass than in other subclasses (Figure 2i, bottom row). While known to be highly expressed in microglia [24], understanding its modulation in other cell subclasses could yield a better understanding of its cell type specific effects on the development of Alzheimer's disease. Such mechanistic description provides a framework for characterizing the connection between a gene's role and a cell's regulatory strategies beyond a mere change in mean expression [25, 26].

We have demonstrated that bivariate distributions arising from mechanistic models can be used in variational autoencoders for principled integration of unspliced and spliced RNA-seq data. This improves model interpretability: conditional parameter estimates give insight into the mechanisms of gene regulation that result in differences in expression. While we impose biophysical constraints on species' conditional joint distributions, orthogonal improvements in interpretability can be made by changing the decoder architecture. *biVI* models can be instantiated with single-layer linear decoders [27] to directly link latent variables with gene mean parameters via layer weights (Section S9 and Figure S9).

Relaxing assumptions and modeling more molecular modalities (e.g., protein counts and chromatin accessibility) are natural extensions. As single-cell technologies evolve to provide larger-scale, more precise measurements of biomolecules, we anticipate that our approach can be applied and extended for a more comprehensive picture of biophysical processes in living cells.

2 Methods

In order to extend the *scVI* method to work with multimodal molecule count data in a way that is coherent with biology, we define bivariate likelihood functions that (i) encode a specific, precedented mechanistic model of transcriptional regulation and (ii) are admissible under the assumptions made in the standard *scVI* pipeline. On a high level, our method entails the following steps:

1. Choose one of the *scVI* univariate generative models (Section 2.2), including the functional form of its likelihood and any assumptions about its distributional parameters.
2. Identify a one-species chemical master equation (CME) that produces this distribution as its steady state, and translate assumptions about distributional parameters into assumptions about the biophysical quantities that parameterize the CME (Section 2.3). The one-species system and its assumptions will typically not be uniquely determined.
3. Identify a two-species CME and derive assumptions about parameter values consistent with the one-species system (Section 2.3). There will typically be multiple ways to preserve the assumptions but only a single CME.
4. Modify the autoencoder architecture to output the variables that parameterize the CME solution under the foregoing assumptions, and use this solution as the generative model (Section 2.5).

2.1 Statistical preliminaries

We use the standard parameterization of the Poisson distribution:

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

We use the shape-mean parameterization of the univariate negative binomial distribution:

$$P_{\text{NB}}(x;\alpha,\mu) = \frac{\Gamma(\alpha+x)}{x!\Gamma(\alpha)}\left(\frac{\alpha}{\alpha+\mu}\right)^{\alpha}\left(\frac{\mu}{\alpha+\mu}\right)^{x}. \tag{2}$$

We use mean parameterization of the geometric distribution on $\mathbb{N}_0$:

$$P_{\text{Geo}}(x;b) = \left(\frac{b}{b+1}\right)^{x}\left(\frac{1}{b+1}\right). \tag{3}$$

2.2 *scVI* models

A brief summary of the generative process of the standard, univariate *scVI* pipeline is useful to contextualize the options and constraints of the bivariate model. In the Bayesian model, each cell has some posterior probability $p_c(z_c)$ over a low-dimensional space and can be represented as a sample z_c from that posterior. *scVI* uses the "decoder" neural network to map from realizations z_c to quantities ρ_{cg}, which describe the compositional abundance of gene g in cell c as a function of z_c, such that $\sum_g \rho_{cg} = 1$. Furthermore, a cell-specific "size factor" ℓ_c is sampled from a lognormal distribution parameterized by either fit or plug-in estimates of mean and variance such that the mean expression of a gene in a given cell is $\mu_{cg} = \rho_{cg}\ell_c$.

The univariate workflow provides the options of three discrete generative models: Poisson with mean μ_{cg}, negative binomial with mean μ_{cg} and gene-specific dispersion parameter α_g, and zero-inflated negative binomial, with an additional Bernoulli mixture parameter. We report the master equation models consistent with the first two generative laws below, and discuss a potential basis for and reservations about the zero-inflated model in Section S1.4.

Due to the intractability of the posterior probability $p_c(z_c)$, *scVI* uses variational inference to infer an approximate posterior $q_z(z_c)$, which is in form a multivariate Gaussian. Models are trained via stochastic optimization of the Evidence Lower Bound, or ELBO, which minimizes the Kullback-Leibler divergence between the approximate posterior and a prior and maximizes the expectation value of conditional likelihood over the approximate posterior. The Gaussian form of the approximate posterior makes possible a reparameterization trick to calculate gradients of the ELBO over expectation estimates made by Monte Carlo sampling from the approximate posterior. Further, the encoding network amortizes inference by learning a map between data to parameters of the approximate posterior [12, 28].

2.3 Master equation models

The one-species CMEs encode reaction schema of the following type:

$$\varnothing \to \mathcal{X} \xrightarrow{\gamma} \varnothing, \tag{4}$$

where $\mathcal{X}$ is a generic transcript species used to instantiate a univariate *scVI* generative model, γ is the transcript's Markovian degradation rate, and the specific dynamics of the transcription process (first arrow) are deliberately left unspecified for now. Such systems induce univariate probability laws of the form $P(x)$.

The two-species CMEs encode reaction schema of the following type:

$$\varnothing \to \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \varnothing, \tag{5}$$

where $\mathcal{N}$ denotes a *n*ascent species, $\mathcal{M}$ denotes a *m*ature species, and β denotes the nascent species' Markovian conversion rate. Such systems induce bivariate probability laws of the form $P(n, m)$. We typically identify the nascent species with unspliced transcripts and the mature species with spliced transcripts. We use the nascent/mature nomenclature to simplify notation and emphasize that this identification is natural for scRNA-seq data, but not mandatory in general.

Formalizing a model in terms of the CME requires specifying the precise mechanistic meaning of ρ_{cg} and ℓ_c. Previous reports equivocate regarding the latter [11], appealing either to cell-wide effects on the biology (in the spirit of [20,21]) or technical variability in the sequencing process (in the spirit of [29]). For completeness, we treat both cases.

Below, we present the theoretical results, including the biophysical models, the functional forms of bivariate distributions consistent with the standard *scVI* models, and the consequences of introducing further assumptions. The full derivations are given in Section S1.

2.3.1 *Constitutive*: The Poisson model and its mechanistic basis

The Poisson generative model can be recapitulated by the following schema:

$$\varnothing \xrightarrow{k} \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \varnothing, \tag{6}$$

where k is a constant transcription rate. This process converges to the bivariate Poisson stationary distribution, with the following likelihood:

$$P(n, m; \mu_N, \mu_M) = P_{\text{Poiss}}(n; \mu_N) P_{\text{Poiss}}(m, \mu_M), \tag{7}$$

where $\mu_N = k/\beta$ and $\mu_M = k/\gamma$. If we suppose each gene's β and γ are constant across cell types, the likelihoods involve a single compositional parameter ρ_{cg}, such that

$$\begin{aligned} \mu_N &= \frac{\gamma_g}{\beta_g} \rho_{cg} \ell_c \\ \mu_M &= \rho_{cg} \ell_c, \end{aligned} \tag{8}$$

where $\gamma_g/\beta_g \in \mathbb{R}^+$ is a gene-specific parameter that can be fit or naïvely estimated by the ratio of the unspliced and spliced averages. On the other hand, if the downstream processes' kinetics can also change between cell types, we must use two compositional parameters:

$$\begin{aligned} \mu_N &= \rho_{cg}^{(N)} \ell_c \\ \mu_M &= \rho_{cg}^{(M)} \ell_c. \end{aligned} \tag{9}$$

We refer to this model as "Poisson," reflecting its functional form, or "constitutive," reflecting its biophysical basis.

2.3.2 *Extrinsic*: The negative binomial model and a possible mixture basis

The negative binomial generative model can be recapitulated by the following schema:

$$\varnothing \xrightarrow{k \sim K} \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \varnothing, \tag{10}$$

where k is the transcription rate, a realization of K, a gamma random variable with shape α, scale η, and mean $\langle K \rangle = \alpha\eta$. This process converges to the bivariate negative binomial (BVNB) stationary distribution, with the following likelihood:

$$P_{\text{extrinsic}}(n, m; \alpha, \mu_N, \mu_M) = \frac{\Gamma(\alpha + n + m)}{n!m!\Gamma(\alpha)} \left(\frac{1}{\alpha + \mu_N + \mu_M} \right)^{\alpha+n+m} \alpha^\alpha \mu_N^n \mu_M^m, \tag{11}$$

where $\mu_N = \langle K \rangle / \beta$ and $\mu_M = \langle K \rangle / \gamma$. If we suppose that cell type differences only involve changes in the transcription rate scaling factor η, with constant α, β, and γ, the likelihoods involve a single compositional parameter ρ_{cg}. The mean parameters are identical to Equation 8, with an analogous parameter γ_g/β_g, as well as a gene-specific shape parameter α_g. On the other hand, if the downstream processes' kinetics can also change between cell types, we must use two compositional parameters, as in Equation 9.

We refer to this model as "extrinsic" to reflect its biophysical basis in extrinsically stochastic rates of transcriptional initiation.

2.3.3 *Bursty*: The negative binomial model and a possible bursty basis

The negative binomial generative model may be recapitulated by the alternative schema [18]:

$$\varnothing \xrightarrow{k} B \times \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \varnothing, \tag{12}$$

where k is the burst frequency and B is a geometric random variable with mean b (Equation 3). This system converges to the following stationary distribution:

$$P_{\text{bursty}}(n, m; \alpha, \mu_N, \mu_M) = P_{\text{NB}}(n; \alpha, \mu_N) P(m|n; \alpha, \mu_N, \mu_M), \tag{13}$$

where $\mu_N = kb/\beta$, $\mu_M = kb/\gamma$, and α is arbitrarily set to k/β for simplicity.

Although the nascent marginal is known to be negative binomial, the joint $P(n, m)$ and conditional $P(m|n)$ distributions are not available in closed form. For a given set of parameters, the joint distribution can be approximated over a finite microstate domain $n, m \in [0, ß_N - 1] \times [0, ß_M - 1]$, with total *s*tate *s*pace size $ß_N ß_M$. This approach is occasionally useful, if intensive, for evaluating the likelihoods of many independent and identically distributed samples. The numerical procedure entails using quadrature to calculate values of the generating function on the complex unit sphere, then performing a Fourier inversion to obtain a probability distribution [18]. However, this strategy is inefficient in the variational autoencoder framework, where each observation is associated with a distinct set of parameters. Furthermore, it is incompatible with automatic differentiation.

In [19], we demonstrated that the numerical approach can be simplified by approximating $P(m|n)$ with a learned mixture of negative binomial distributions: the weights are given by the outputs of a neural network, whereas the negative binomial bases are constructed analytically. The neural network is trained on the outputs of the generating function procedure. Although the generative model does not have a simple closed-form expression, it is represented by a partially neural, pre-trained function that is *a priori* compatible with the VAE.

If we suppose cell type differences only involve changes in the burst size b, with constant k, β and γ, we use Equation 13 to evaluate likelihoods. These likelihoods involve a single compositional parameter ρ_{cg}, with mean parameters identical to Equation 8, with an analogous parameter γ_g/β_g, as well as a gene-specific shape parameter α_g. On the other hand, if kinetics of the degradation process can also change between cell types, we must use two compositional parameters, as in Equation 9. There is no admissible way to allow modulation in the burst frequency.

We refer to this model as "bursty," reflecting its biophysical basis.

2.4 *biVI* bursty generative model

Following the notation of *scVI* [28], *biVI*'s generative process for the bursty hypothesis models expression values of x_{cn} and x_{cm} of nascent and mature counts, respectively, in cell c as:

$$\begin{aligned} z_c &\sim \text{Normal}(0, I) \\ \ell_c &\sim \text{log normal}(\ell_\mu, \ell_{\sigma^2}) \\ \rho_{cg}^{(N)}, \rho_{cg}^{(M)} &= f(z_c, s_c) \\ \mu_{cn}, \mu_{cm} &= \rho_{cg}^{(N)} \ell_c, \rho_{cg}^{(M)} \ell_c \\ x_{cn}, x_{cm} &\sim P_{\text{bursty}}(n, m; \alpha, \mu_{cn}, \mu_{cm}), \end{aligned} \tag{14}$$

with a standard, multivariate normal prior on the latent space z vector. Here, $\ell_\mu, \ell_{\sigma^2}$ are by default observed mean and variance in log-sequencing depth ('log-library size' in *scVI*) across a cell's batch, although they can be learned. Further, as in *scVI*, f is neural network that produces fraction of sequencing depth parameters $\rho_{cg}^{(N)}, \rho_{cg}^{(M)}$ for nascent and mature counts. The sum of nascent and mature fractions is constrained to be 1 over a cell c by a softmax applied to the network output: $\sum_{n,m=0}^{G}(\rho_{cg}^{(N)}, \rho_{cg}^{(M)}) = 1$, where G is the number of genes. $\alpha \in \mathbb{R}^G$ is a network parameter jointly optimized across all cells during the variational inference procedure. To recover biophysical parameters, α is arbitrarily set to k/β. Burst size b and relative degradation rate k/γ can be recovered according to the following conversions:

$$\begin{aligned} b &= \frac{\mu_{cn}}{\alpha} \\ \gamma/k &= \frac{\mu_{nc}}{\mu_{cm}\alpha} \end{aligned} \tag{15}$$

We further set $k = 1$ with no loss of generality at steady-state. Generative processes for constitutive and extrinsic noise models are discussed in Sections S2 and S3.

2.5 *biVI* modifications to *scVI*

Our code is built upon *scVI* version 0.18.0 [30]; the following outlines the modifications we made for *biVI*. The *scVI* framework already supports the constitutive model. By setting conditional likelihood to "poisson," no modification of *scVI* architecture is necessary. The conditional data likelihood distribution is the product of two Poisson distributions (Equation 7). Explicitly, unspliced and spliced count matrices can be concatenated along the cell axis to produce a matrix of shape C by $2G$, where C is the number of cells and G the number of genes. *scVI* will then produce $2G$ Poisson mean parameters for the two Poisson distributions of Equation 7.

For the extrinsic and bursty models, mean parameters for nascent and mature counts, μ_N and μ_M, and a single shape parameter α are necessary. The default *scVI* architecture returns two independent parameters for nascent and mature counts of the same gene. *biVI* thus modifies the *scVI* architecture to update vectors $\alpha \in \mathbb{R}_{\geq 0}^{G}$ rather than $\alpha \in \mathbb{R}_{\geq 0}^{2G}$, where G is the number of genes. For the extrinsic model, the conditional data likelihood distribution is set to the extrinsic likelihood $P_{\text{extrinsic}}(n, m; \alpha, \mu_N, \mu_M)$ (Equation 11). For the bursty model, the conditional data likelihood distribution is set to the bursty likelihood $P_{\text{bursty}}(n, m; \alpha, \mu_N, \mu_M)$ (Equation 13). These models also intake concatenated unspliced and spliced matrices of shape C by $2G$.

2.6 Preprocessing Allen data

Raw 10x v3 single-cell data were originally generated by the Allen Institute for Brain Science [22]. The raw reads in FASTQ format [31] and cluster metadata [32] were obtained from the NeMO Archive. We selected mouse library B08 (donor ID 457911) for analysis.

To obtain spliced and unspliced counts, we first obtained the pre-built mm10 mouse genome released by 10x Genomics (`https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest`, version 2020-A). We used *kallisto*|*bustools* 0.26.0 [2] to build an intronic/exonic reference (`kb ref` with the option `--lamanno`). Next, we pseudoaligned the reads to this reference (`kb count` with the option `--lamanno`) to produce unspliced and spliced count matrices. We used the outputs produced by the standard *bustools* filter. This filter was relatively permissive: all (8,424) barcodes given cell type annotations in the Allen metadata were present in the output count matrix (10,975 barcodes).

Based on previous clustering results, we selected cells that were given cell type annotations, and omitted "low quality" or "doublet" barcodes [22], for a total of 6,418 cells. Although any choice to retain or omit cells from analysis is arbitrary, our work models the generating process that produced cells' nascent and mature counts by presupposing each barcode corresponds to a single cell. Therefore, we propose that cells identified as low-quality (empty cells) or as doublets (two cells measured in one observation) [22] have a fundamentally different data-generating process than individual single cells, and therefore remove them before fitting VAE models. However, we stress that the stochastic nature of transcription and sequencing, the intrinsic uncertainties associated with read alignment, and the numerical compromises made in clustering large datasets mean that previous annotations are not "perfect," merely a reasonable starting point for comparing alternative methods.

We used Scanpy [33] to restrict our analysis to the most variable genes, which presumably reflect the cell type signatures of interest. The spliced count matrix for the 6,418 retained cells was normalized to sum to 10,000

counts per cell, then transformed with log1p. The top 2,000 most highly variable genes were identified using `scanpy.pp.highly_variable_genes` on spliced matrices with minimum mean of 0.0125, maximum mean of 3, and minimum dispersion of 0.5 [33]. Spliced and unspliced matrices were subset to include only the 2,000 identified highly variable genes, then concatenated along the cell axis in the order unspliced, spliced to produce a count matrix of size 6,418 by 4,000.

2.7 Fitting Allen data

We applied *biVI* with the three generative models (bursty, constitutive, and extrinsic) and *scVI* with negative binomial likelihoods to the concatenated unspliced and spliced count matrix obtained by the filtering procedures outlined above. We made the key assumption that unspliced and spliced counts could be treated as the nascent and mature species of the bursty generative model (see discussion in Section S5). 4,622 cells were used for training with 513 validation cells, and 1,283 cells were held out for testing performance. All models were trained for 400 epochs with a learning rate of 0.001. Encoders and decoder consisted of 3 layers of 128 nodes, and each model employed a latent dimension of 10.

2.8 Bayes factor hypothesis testing for differential expression

After fitting the VAE models, we sought to identify meaningful statistical differences that distinguish cell types. We excluded cell subclasses "L6 IT Car3," "L5 ET," "VLMC," and "SMC" from this analysis, as they contained fewer than ten annotated cells and may require more sophisticated statistical models to account for small sample sizes. The following analysis thus considers 6,398 cells in 16 unique subclasses. We only computed differential expression metrics under the bursty model.

Differential parameter values were tested for each assigned subclass label (as annotated in [22]) versus all others using a Bayes factor hypothesis test following [11]. We reproduce Equations (18) - (21) of [11] below for clarity.

Estimating differential values of any parameter θ^g of gene g in cells a and b can be done according to the following Bayesian framework. First, as in Equation (18) of [11], the log fold change (LFC) of θ^g between two cells a and b can be calculated as follows:

$$\mathrm{LFC}^g_{a,b} := \log_2\theta^g_b - \log_2\theta^g_a. \tag{16}$$

Then, as in Equation (19) of [11], the probability that the magnitude of the LFC is greater than some effect threshold T can be found by evaluating $\mathrm{LFC}^g_{a,b}$ over the posterior distributions of each cell:

$$P(|\mathrm{LFC}^g_{a,b}| \geq T|a,b) \approx \int \mathbb{I}(|\mathrm{LFC}^g_{a,b}| \geq T) q_a(z_a) q_b(z_b) dz_a dz_b, \tag{17}$$

where, in practice, the integral is approximated with many Monte Carlo samples from the two cells' posteriors. Two hypotheses are tested: H_1, or that the magnitude of the LFC is greater than or equal to threshold T, and H_0, or the null hypothesis that the magnitude of the LFC is less than T. A Bayes factor for gene g between cells a and b ($BF^g_{a,b}$) is calculated to compare the two hypotheses, as in Equation (20) of [11]:

$$\mathrm{BF}^g_{a,b} = \frac{P(|\mathrm{LFC}^g_{a,b}| \geq T|a,b)}{P(|\mathrm{LFC}^g_{a,b}| < T|a,b)}. \tag{18}$$

Extending this to test differential expression between two groups of cells A and B amounts to "aggregating the posterior," as in Equation (21) of [11], or evaluating the same $P(|\mathrm{LFC}^g_{A,B}| \geq T\|A,B)$ over

$$\frac{1}{|A|}\frac{1}{|B|}\sum_{a\in A} q_a(z_a) \sum_{b\in B} q_b(z_b). \tag{19}$$

In other words, a random sample z_a can be be taken from the approximate posterior of any cell belonging to group A and decoded to produce parameter θ^g_a; likewise a random sample z_b can be taken from the approximate posterior of any cell belonging to group B and decoded to produce parameter θ^b_a. The LFC between the two parameters can then be calculated. Repeating this for many Monte Carlo samples over the aggregate posteriors allows estimation of the Bayes factor between two groups.

For the results shown in Figure 1, we used cutoffs of $T \geq 1.0$, or a magnitude LFC of ≥ 2, and a Bayes factor threshold of 1.5. The Bayes factors were calculated on normalized burst size and means for *biVI*, i.e., the fractional inferred burst size or inferred means (before scaling by sampled sequencing depth for that cell), and normalized

means for *scVI*. This controlled for differences in parameters due to sequencing depth that were not biologically meaningful. Relative degradation rate γ/k is independent of sequencing depth: hypothesis tests were performed directly on inferred relative degradation rates. While batch identity can also be integrated over to compare groups of cells from different batches, our analysis did not require this as all cells were from the same batch.

3 Reconstructing gene distributions

Let $\theta_{\kappa g}$ be mechanistic model parameters for gene g in cell type κ. While parameters for a given gene are identical across all cells in a specific cell type, *biVI* and *scVI* infer unique parameters for every cell and gene: θ_{cg}, where c indexes over cells and g indexes over genes. To reconstruct distributions for a given gene in a specific cell type κ, we sample once from the posterior distribution $q_c(z)$ of each cell $c \in \kappa$ to obtain point-estimates of conditional parameters $\theta_{c_\kappa g}$, where conditional refers to a single sampling from a cell's posterior, or a particular realization of z_c. We then average over the cell-specific conditional probabilities for the gene to produce a cell type marginal distribution:

$$\hat{P}_{\kappa g}(n, m) = \frac{1}{n_\kappa} \sum_{c_\kappa=1}^{n_\kappa} P(n, m; \theta_{c_\kappa g}), \tag{20}$$

where n_κ is the total number of cells in cell type κ, and c_κ indexes over all cells in that cell type. This identity follows immediately from defining the cell type's distribution as the mixture of the distributions of its constituent cells. In the case of *biVI*, we plug in Equation 7, 11, or 13 for $P(n, m; \theta_{c_\kappa g})$. In the case of *scVI*, we use a product of two independent negative binomial laws:

$$P(n, m; \theta_{c_\kappa g}) = P_{\text{NB}}(n; \alpha_g^N, \mu_N) P_{\text{NB}}(m; \alpha_g^M, \mu_M), \tag{21}$$

where μ_N and μ_M are cell- and gene-specific, whereas α^N and α^M are fit separately and take on different values (Section 2.5). For simplicity, this comparison omits uncertainty associated with θ_{cg}, which is formally inherited from the uncertainty in the latent representation z for each cell c.

Thus, Equation 21 is an approximation to the posterior predictive distribution, or marginal distribution of data given the approximated posterior, if we assume Monte Carlo sampling from the approximate posterior distributions of cells within that cell type as a reasonable proxy for sampling from the cell type's posterior distribution. The posterior predictive, or marginal, distribution is:

$$P_{\kappa,g}(n, m) \approx \int P(n, m|z) q_\kappa(z) dz$$

where $q(z)$ is the approximate posterior. We further note that conditional data likelihood and the marginal distribution are *not* necessarily of the same form (for example, if the conditional data likelihood distribution is negative binomial, the marginal distribution of genes is not necessarily negative binomial).

4 Data availability

Simulated datasets, simulated parameters used to generate them, and Allen dataset B08 and its associated metadata are available in the Zenodo package 7497222. All analysis scripts and notebooks are available at `https://github.com/pachterlab/CGCCP_2023`. The repository also contains a Google Colaboratory demonstration notebook applying the methods to a small human blood cell dataset.

5 Acknowledgments

M.C., G.G., T.C., and L.P. were partially funded by NIH 5UM1HG012077-02 and NIH U19MH114830. Y.C. was partially funded by T32 GM007377. G.G. thanks Drs. Ido Golding and Heng Xu for the inspiration leading to the explanatory model for the zero-inflated negative binomial distribution in Section S1.4. The RNA illustrations used in Figures 1, S1, and S2 were derived from the DNA

Twemoji by Twitter, Inc., used under the CC-BY 4.0 license. We thank the Caltech Bioinformatics Resource Center for GPU resources that helped in performing the analyses.

References

[1] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, August 2018.

[2] Páll Melsted, A. Sina Booeshaghi, Lauren Liu, Fan Gao, Lambda Lu, Kyung Hoi Min, Eduardo da Veiga Beltrame, Kristján Eldjárn Hjörleifsson, Jase Gehring, and Lior Pachter. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature Biotechnology*, 39(7):813–818, July 2021.

[3] Vanessa M Peterson, Kelvin Xi Zhang, Namit Kumar, Jerelyn Wong, Lixia Li, Douglas C Wilson, Renee Moore, Terrill K McClanahan, Svetlana Sadekova, and Joel A Klappenbach. Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, 35(10):936–939, October 2017.

[4] Eleni P. Mimitou, Anthony Cheng, Antonino Montalbano, Stephanie Hao, Marlon Stoeckius, Mateusz Legut, Timothy Roush, Alberto Herrera, Efthymia Papalexi, Zhengqing Ouyang, Rahul Satija, Neville E. Sanjana, Sergei B. Koralov, and Peter Smibert. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature Methods*, 16(5):409–412, May 2019.

[5] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, September 2017.

[6] Hattie Chung, Christopher N. Parkhurst, Emma M. Magee, Devan Phillips, Ehsan Habibi, Fei Chen, Bertrand Z. Yeung, Julia Waldman, David Artis, and Aviv Regev. Joint single-cell measurements of nuclear proteins and RNA in vivo. *Nature Methods*, 18(10):1204–1212, October 2021.

[7] M. Reyes, K. Billman, N. Hacohen, and P.C. Blainey. Simultaneous profiling of gene expression and chromatin accessibility in single cells. *Advanced Biosystems*, 3,11, 2019.

[8] Florian De Rop, Joy N Ismail, Carmen Bravo González-Blas, Gert J Hulselmans, Christopher Campbell Flerin, Jasper Janssens, Koen Theunis, Valerie M Christiaens, Jasper Wouters, Gabriele Marcassa, Joris de Wit, Suresh Poovathingal, and Stein Aerts. HyDrop enables droplet based single-cell ATAC-seq and single-cell RNA-seq using dissolvable hydrogel beads. *eLife*, 11:e73971, February 2022.

[9] Gennady Gorin, John J. Vastola, Meichen Fang, and Lior Pachter. Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments. *Nature Communications*, 13(1):7620, December 2022.

[10] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604, April 2018.

[11] Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L. Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3):272–282, March 2021.

[12] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Katherine Wu, Michael Jayasuriya, Edouard Melhman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Carlos Talavera-López, Lior Pachter, Fabian J. Theis, Aaron Streets, Michael I. Jordan, Jeffrey Regier, and Nir Yosef. scvi-tools: a library for deep probabilistic analysis of single-cell omics data. Preprint, bioRxiv: 2021.04.28.441833, April 2021.

[13] Xiang Lin, Tian Tian, Zhi Wei, and Hakon Hakonarson. Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nature Communications*, 13(1):7705, December 2022.

[14] Tal Ashuach, Daniel A. Reidenbach, Adam Gayoso, and Nir Yosef. PeakVI: A deep generative model for single-cell chromatin accessibility analysis. *Cell Reports Methods*, 2(3):100182, March 2022.

[15] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biology*, 4(10):e309, September 2006.

[16] R. D. Dar, B. S. Razooky, A. Singh, T. V. Trimeloni, J. M. McCollum, C. D. Cox, M. L. Simpson, and L. S. Weinberger. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences*, 109(43):17454–17459, October 2012.

[17] A. Sanchez and I. Golding. Genetic Determinants and Cellular Constraints in Noisy Gene Expression. *Science*, 342(6163):1188–1193, December 2013.

[18] Abhyudai Singh and Pavol Bokes. Consequences of mRNA Transport on Stochastic Variability in Protein Levels. *Biophysical Journal*, 103(5):1087–1096, September 2012.

[19] Gennady Gorin, Maria Carilli, Tara Chari, and Lior Pachter. Spectral neural approximations for models of transcriptional dynamics. Preprint, bioRxiv: 2022.06.16.496448, June 2022.

[20] Lucy Ham, Rowan D. Brackston, and Michael P. H. Stumpf. Extrinsic Noise and Heavy-Tailed Laws in Gene Expression. *Physical Review Letters*, 124(10):108101, March 2020.

[21] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic Gene Expression in a Single Cell. *Science*, 297(5584):1183–1186, 2002.

[22] Zizhen Yao, Hanqing Liu, Fangming Xie, Stephan Fischer, Ricky S. Adkins, Andrew I. Aldridge, Seth A. Ament, Anna Bartlett, M. Margarita Behrens, Koen Van den Berge, Darren Bertagnolli, Hector Roux de Bézieux, Tommaso Biancalani, A. Sina Booeshaghi, Héctor Corrada Bravo, Tamara Casper, Carlo Colantuoni, Jonathan Crabtree, Heather Creasy, Kirsten Crichton, Megan Crow, Nick Dee, Elizabeth L. Dougherty, Wayne I. Doyle, Sandrine Dudoit,

Rongxin Fang, Victor Felix, Olivia Fong, Michelle Giglio, Jeff Goldy, Mike Hawrylycz, Brian R. Herb, Ronna Hertzano, Xiaomeng Hou, Qiwen Hu, Jayaram Kancherla, Matthew Kroll, Kanan Lathia, Yang Eric Li, Jacinta D. Lucero, Chongyuan Luo, Anup Mahurkar, Delissa McMillen, Naeem M. Nadaf, Joseph R. Nery, Thuc Nghi Nguyen, Sheng-Yong Niu, Vasilis Ntranos, Joshua Orvis, Julia K. Osteen, Thanh Pham, Antonio Pinto-Duarte, Olivier Poirion, Sebastian Preissl, Elizabeth Purdom, Christine Rimorin, Davide Risso, Angeline C. Rivkin, Kimberly Smith, Kelly Street, Josef Sulc, Valentine Svensson, Michael Tieu, Amy Torkelson, Herman Tung, Eeshit Dhaval Vaishnav, Charles R. Vanderburg, Cindy van Velthoven, Xinxin Wang, Owen R. White, Z. Josh Huang, Peter V. Kharchenko, Lior Pachter, John Ngai, Aviv Regev, Bosiljka Tasic, Joshua D. Welch, Jesse Gillis, Evan Z. Macosko, Bing Ren, Joseph R. Ecker, Hongkui Zeng, and Eran A. Mukamel. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*, 598(7879):103–110, October 2021.

[23] XL. Kuang, XM. Zhao, HF. Xu, YY. Shi, JB. Deng, and GT. Sun. Spatio-temporal expression of a novel neuron-derived neurotrophic factor (ndnf) in mouse brains during development. *BMC Neurosci*, 11, 2010.

[24] T. K. Ulland and M. Colonna. Trem2 — a key player in microglial biology and alzheimer disease. *Nature Reviews Neurology*, 14:667–675, 2018.

[25] Brian Munsky, Guoliang Li, Zachary R. Fox, Douglas P. Shepherd, and Gregor Neuert. Distribution shapes govern the discovery of predictive models for gene regulation. *Proceedings of the National Academy of Sciences*, 115(29):7533–7538, 2018.

[26] Brian Munsky, Brooke Trinh, and Mustafa Khammash. Listening to the noise: random fluctuations reveal gene network parameters. *Molecular Systems Biology*, 5:318, October 2009.

[27] Valentine Svensson, Adam Gayoso, Nir Yosef, and Lior Pachter. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421, June 2020.

[28] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018.

[29] Jingshu Wang, Mo Huang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, John Murray, Arjun Raj, Mingyao Li, and Nancy R. Zhang. Gene expression distribution deconvolution in single-cell RNA sequencing. *Proceedings of the National Academy of Sciences*, 115(28):E6437–E6446, July 2018.

[30] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mariano Gabitto, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Vitalii Kleshchevnikov, Carlos Talavera-López, Lior Pachter, Fabian J. Theis, Aaron Streets, Michael I. Jordan, Jeffrey Regier, and Nir Yosef. A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, February 2022.

[31] Allen Institute for Brain Science. FASTQ files for Allen v3 mouse MOp samples, February 2020.

[32] Allen Institute for Brain Science. Analyses for Allen v3 mouse MOp samples, February 2020.

[33] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, December 2018.

[34] Tobias Jahnke and Wilhelm Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54:1–26, September 2006.

[35] Ruben Perez-Carrasco, Casper Beentjes, and Ramon Grima. Effects of cell cycle variability on lineage and population measurements of messenger RNA abundance. *Journal of The Royal Society Interface*, 17(168):20200360, July 2020.

[36] Gennady Gorin and Lior Pachter. Length biases in single-cell RNA sequencing of pre-mRNA. *Biophysical Reports*, 3(1):100097, March 2023.

[37] Gennady Gorin and Lior Pachter. *Monod*: mechanistic analysis of single-cell RNA sequencing count data. Preprint, bioRxiv: 2022.06.11.495771, June 2022.

[38] Gennady Gorin and Lior Pachter. Intrinsic and extrinsic noise are distinguishable in a synthesis – export – degradation model of mRNA production. Preprint, bioRxiv: 2020.09.25.312868, September 2020.

[39] Ruochen Jiang, Tianyi Sun, Dongyuan Song, and Jingyi Jessica Li. Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biology*, 23:31, January 2022.

[40] Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, February 2020.

[41] Chen Jia. Kinetic Foundation of the Zero-Inflated Negative Binomial Model for Single-Cell RNA Sequencing Data. *SIAM Journal on Applied Mathematics*, 80(3):1336–1355, January 2020.

[42] Heng Xu, Leonardo A Sepúlveda, Lauren Figard, Anna Marie Sokac, and Ido Golding. Combining protein and mRNA quantification to decipher transcriptional regulation. *Nature Methods*, 12(8):739–742, August 2015.

[43] Gennady Gorin and Lior Pachter. Modeling bursty transcription and splicing with the chemical master equation. *Biophysical Journal*, 121(6):1056–1069, February 2022.

[44] Joseph Rodriguez and Daniel R. Larson. Transcription in Living Cells: Molecular Mechanisms of Bursting. *Annual Review of Biochemistry*, 89(1):189–212, June 2020.

[45] Heng Xu, Samuel O. Skinner, Anna Marie Sokac, and Ido Golding. Stochastic Kinetics of Nascent RNA. *Physical Review Letters*, 117(12):128101, 2016.

[46] Sandeep Choubey, Jane Kondev, and Alvaro Sanchez. Deciphering Transcriptional Dynamics In Vivo by Counting Nascent RNA Molecules. *PLOS Computational Biology*, 11(11):e1004345, 2015.

[47] Sandeep Choubey. Nascent RNA kinetics: Transient and steady state behavior of models of transcription. *Physical Review E*, 97(2):022402, 2018.

[48] Mariana Gómez-Schiavon, Liang-Fu Chen, Anne E. West, and Nicolas E. Buchler. BayFish: Bayesian inference of transcription dynamics from population snapshots of single-molecule RNA FISH in single cells. *Genome Biology*, 18(1):164, December 2017.

[49] Mengyu Wang, Jing Zhang, Heng Xu, and Ido Golding. Measuring transcription at a single gene copy reveals hidden drivers of bacterial individuality. *Nature Microbiology*, 4:2118–2127, September 2019.

[50] Daniel Zenklusen, Daniel R Larson, and Robert H Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology*, 15(12):1263–1271, 2008.

[51] Adrien Senecal, Brian Munsky, Florence Proux, Nathalie Ly, Floriane E. Braye, Christophe Zimmer, Florian Mueller, and Xavier Darzacq. Transcription Factors Modulate c-Fos Transcriptional Bursts. *Cell Reports*, 8(1):75–83, July 2014.

[52] Keren Bahar Halpern, Sivan Tanami, Shanie Landen, Michal Chapal, Liran Szlak, Anat Hutzler, Anna Nizhberg, and Shalev Itzkovitz. Bursty Gene Expression in the Intact Mammalian Liver. *Molecular Cell*, 58(1):147–156, April 2015.

[53] Samuel O Skinner, Heng Xu, Sonal Nagarkar-Jaiswal, Pablo R Freire, Thomas P Zwaka, and Ido Golding. Single-cell analysis of transcription kinetics across the cell cycle. *eLife*, 5:e12175, January 2016.

[54] Sheel Shah, Yodai Takei, Wen Zhou, Eric Lubeck, Jina Yun, Chee-Huat Linus Eng, Noushin Koulena, Christopher Cronin, Christoph Karp, Eric J. Liaw, Mina Amin, and Long Cai. Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. *Cell*, 174(2):363–376.e16, July 2018.

[55] Yihan Wan, Dimitrios G. Anastasakis, Joseph Rodriguez, Murali Palangat, Prabhakar Gudla, George Zaki, Mayank Tandon, Gianluca Pegoraro, Carson C. Chow, Markus Hafner, and Daniel R. Larson. Dynamic imaging of nascent RNA reveals general principles of transcription dynamics and stochastic splice site selection. *Cell*, 184(11):2878–2895.e20, May 2021.

[56] Barbara Wold and Richard M Myers. Sequence census methods for functional genomics. *Nature Methods*, 5(1):19–21, January 2008.

[57] Kirsten A. Reimer, Claudia A. Mimoso, Karen Adelman, and Karla M. Neugebauer. Co-transcriptional splicing regulates 3' end cleavage during mammalian erythropoiesis. *Molecular Cell*, 81(5):998–1012.e7, March 2021.

[58] Heather L. Drexler, Karine Choquet, and L. Stirling Churchman. Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Molecular Cell*, 77(5):985–998.e8, March 2020.

[59] A. Zeisel, W. J. Kostler, N. Molotski, J. M. Tsai, R. Krauthgamer, J. Jacob-Hirsch, G. Rechavi, Y. Soen, S. Jung, Y. Yarden, and E. Domany. Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Molecular Systems Biology*, 7(1):529–529, September 2011.

[60] Harold Pimentel, John G. Conboy, and Lior Pachter. Keep Me Around: Intron Retention Detection and Analysis. Preprint, arXiv: 1510.00696, October 2015.

[61] Harold Pimentel, Marilyn Parra, Sherry L. Gee, Narla Mohandas, Lior Pachter, and John G. Conboy. A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Research*, 44(2):838–851, January 2016.

[62] Gennady Gorin, Meichen Fang, Tara Chari, and Lior Pachter. RNA velocity unraveled. *PLOS Computational Biology*, 18(9):e1010492, September 2022.

[63] Kristján Eldjárn Hjörleifsson, Delaney K. Sullivan, Guillaume Holley, Páll Melsted, and Lior Pachter. Accurate quantification of single-nucleus and single-cell RNA-seq transcripts. Preprint, bioRxiv: 2022.12.02.518832, December 2022.

[64] Charlotte Soneson, Avi Srivastava, Rob Patro, and Michael B. Stadler. Preprocessing choices affect RNA velocity results for droplet scRNA-seq data. *PLOS Computational Biology*, 17(1):e1008585, January 2021.

[65] Maxime Mazille, Katarzyna Buczak, Peter Scheiffele, and Oriane Mauger. Stimulus-specific remodeling of the neuronal transcriptome through nuclear intron-retaining transcripts. *The EMBO Journal*, 41(21):e110192, 2022.

[66] A. Sina Booeshaghi, Zizhen Yao, Cindy van Velthoven, Kimberly Smith, Bosiljka Tasic, Hongkui Zeng, and Lior Pachter. Isoform cell-type specificity in the mouse primary motor cortex. *Nature*, 598(7879):195–199, October 2021.

[67] O Kessler, Y Jiang, and L A Chasin. Order of intron removal during splicing of endogenous adenine phosphoribosyltransferase and dihydrofolate reductase pre-mRNA. *Molecular and Cellular Biology*, 13(10):6211–6222, October 1993.

[68] Allison Coté, Chris Coté, Sareh Bayatpour, Heather L Drexler, Katherine A Alexander, Fei Chen, Asmamaw T Wassie, Edward S Boyden, Shelley Berger, L Stirling Churchman, and Arjun Raj. pre-mRNA spatial distributions suggest that splicing can occur post-transcriptionally. Preprint, bioRxiv: 2020.04.06.028092, June 2021.

[69] Gennady Gorin, Shawn Yoshida, and Lior Pachter. Transient and delay chemical master equations. Preprint, bioRxiv: 2022.10.17.512599, October 2022.

[70] Zhixing Cao and Ramon Grima. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proceedings of the National Academy of Sciences*, 117(9):4682–4692, March 2020.

[71] Qingchao Jiang, Xiaoming Fu, Shifu Yan, Runlai Li, Wenli Du, Zhixing Cao, Feng Qian, and Ramon Grima. Neural network aided approximation and parameter inference of non-Markovian models of gene expression. *Nature Communications*, 12(1):2618, December 2021.

[72] Tatiana Filatova, Nikola Popović, and Ramon Grima. Modulation of nuclear and cytoplasmic mRNA fluctuations by time-dependent stimuli: Analytical distributions. *Mathematical Biosciences*, 347:108828, May 2022.

[73] Maike M.K. Hansen, Ravi V. Desai, Michael L. Simpson, and Leor S. Weinberger. Cytoplasmic Amplification of Transcriptional Noise Generates Substantial Cell-to-Cell Variability. *Cell Systems*, 7(4):384–397.e6, October 2018.

[74] Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Control of Transcript Variability in Single Mammalian Cells. *Cell*, 163(7):1596–1610, December 2015.

[75] Gennady Gorin and Lior Pachter. Special function methods for bursty models of transcription. *Physical Review E*, 102(2):022409, August 2020.

[76] Xiaoming Fu, Heta P Patel, Stefano Coppola, Libin Xu, Zhixing Cao, Tineke L Lenstra, and Ramon Grima. Quantifying how post-transcriptional noise and gene copy number variation bias transcriptional parameter inference from mRNA distributions. *eLife*, 11:e82493, October 2022.

[77] Xiaoming Fu, Heta P. Patel, Stefano Coppola, Libin Xu, Zhixing Cao, Tineke L. Lenstra, and Ramon Grima. Accurate inference of stochastic gene expression from nascent transcript heterogeneity. Preprint, bioRxiv: 2021.11.09.467882, November 2021.

[78] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, October 2011.

[79] The Gene Ontology Consortium, Seth Carbon, Eric Douglass, Benjamin M Good, Deepak R Unni, Nomi L Harris, Christopher J Mungall, Siddartha Basu, Rex L Chisholm, Robert J Dodson, Eric Hartline, Petra Fey, Paul D Thomas, Laurent-Philippe Albou, Dustin Ebert, Michael J Kesling, Huaiyu Mi, Anushya Muruganujan, Xiaosong Huang, Tremayne Mushayahama, Sandra A LaBonte, Deborah A Siegele, Giulia Antonazzo, Helen Attrill, Nick H Brown, Phani Garapati, Steven J Marygold, Vitor Trovisco, Gil dos Santos, Kathleen Falls, Christopher Tabone, Pinglei Zhou, Joshua L Goodman, Victor B Strelets, Jim Thurmond, Penelope Garmiri, Rizwan Ishtiaq, Milagros Rodríguez-López, Marcio L Acencio, Martin Kuiper, Astrid Lægreid, Colin Logie, Ruth C Lovering, Barbara Kramarz, Shirin C C Saverimuttu, Sandra M Pinheiro, Heather Gunn, Renzhi Su, Katherine E Thurlow, Marcus Chibucos, Michelle Giglio, Suvarna Nadendla, James Munro, Rebecca Jackson, Margaret J Duesbury, Noemi Del-Toro, Birgit H M Meldal, Kalpana Paneerselvam, Livia Perfetto, Pablo Porras, Sandra Orchard, Anjali Shrivastava, Hsin-Yu Chang, Robert Daniel Finn, Alexander Lawson Mitchell, Neil David Rawlings, Lorna Richardson, Amaia Sangrador-Vegas, Judith A Blake, Karen R Christie, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry M Sitnikov, Midori A Harris, Stephen G Oliver, Kim Rutherford, Valerie Wood, Jaqueline Hayles, Jürg Bähler, Elizabeth R Bolton, Jeffery L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Cody Plasterer, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lisa Matthews, James P Balhoff, Suzi A Aleksander, Michael J Alexander, J Michael Cherry, Stacia R Engel, Felix Gondwe,

Kalpana Karra, Stuart R Miyasato, Robert S Nash, Matt Simison, Marek S Skrzypek, Shuai Weng, Edith D Wong, Marc Feuermann, Pascale Gaudet, Anne Morgat, Erica Bakker, Tanya Z Berardini, Leonore Reiser, Shabari Subramaniam, Eva Huala, Cecilia N Arighi, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Alex Bateman, Marie-Claude Blatter, Emmanuel Boutet, Emily Bowler, Lionel Breuza, Alan Bridge, Ramona Britto, Hema Bye-A-Jee, Cristina Casals Casas, Elisabeth Coudert, Paul Denny, Anne Estreicher, Maria Livia Famiglietti, George Georghiou, Arnaud Gos, Nadine Gruaz-Gumowski, Emma Hatton-Ellis, Chantal Hulo, Alexandr Ignatchenko, Florence Jungo, Kati Laiho, Philippe Le Mercier, Damien Lieberherr, Antonia Lock, Yvonne Lussi, Alistair MacDougall, Michele Magrane, Maria J Martin, Patrick Masson, Darren A Natale, Nevila Hyka-Nouspikel, Sandra Orchard, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Sangya Pundir, Catherine Rivoire, Elena Speretta, Shyamala Sundaram, Nidhi Tyagi, Kate Warner, Rossana Zaru, Cathy H Wu, Alexander D Diehl, Juancarlos N Chan, Christian Grove, Raymond Y N Lee, Hans-Michael Muller, Daniela Raciti, Kimberly Van Auken, Paul W Sternberg, Matthew Berriman, Michael Paulini, Kevin Howe, Sibyl Gao, Adam Wright, Lincoln Stein, Douglas G Howe, Sabrina Toro, Monte Westerfield, Pankaj Jaiswal, Laurel Cooper, and Justin Elser. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1):D325–D334, January 2021.