Mechanistic modeling with a variational autoencoder for multimodal 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

January 13, 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. In simulated benchmarking, *biVI* accurately recapitulates key properties of interest, including cell type structure, parameter values, and copy number distributions. In biological datasets, *biVI* provides a route for the identification of the biophysical mechanisms underlying differential expression. The analytical approach outlines a generalizable strategy for representing multimodal datasets generated by single-cell RNA sequencing.

1 Introduction

Advances in experimental methods for single-cell RNA sequencing (scRNA-seq) allow for the simultaneous quantification of multiple cellular species at a time, such as nascent and mature transcriptomes [1,2], surface [3–5] and nuclear [6] proteomes, and chromatin accessibility [7,8]. The biophysical "integration" of such datasets requires the parameterization of interpretable mechanistic models [9]. Such modeling is challenging because mechanistic models tend to be intractable for more than a few co-regulated genes; in practice, single-cell genomics measurements are made for thousands of genes in tens of thousands of cells [10].

One approach to integration of multimodal single-cell genomics data has been to leverage recent advances in machine learning [11–13] and matrix factorization techniques [14]. This approach aims to summarize multimodal single cell measurements in a common low-dimensional space, which purports to reflect the cell state structure. Specifically, single-cell variational inference (*scVI*), a popular machine learning framework for analyzing scRNA-seq data [12,15], has previously been adapted to multimodal data with protein [11] and chromatin [16] measurements. However, methods based on this approach are neither informed by nor lead to specific biophysical interpretations, and omit the intrinsic causal relationships between "upstream" and "downstream" parts of the central dogma [17,18].

Consider, for example, the joint modeling of nascent and mature RNA counts, which are readily available by realigning existing scRNA-seq reads [1,2], via the *scVI* framework. While *scVI* has not been considered for such integration, one can conceive of utilizing it for this purpose as follows: each cell could be encoded by a neural network to a latent low-dimensional vector z, which could be decoded into a set of cell- and gene-specific parameters by another neural network. These parameters would define a negative binomial distribution used to evaluate and optimize the likelihoods of RNA count matrices. In principle, nascent and mature RNA molecules can be "integrated" with *scVI* by concatenating the two matrices and treating nascent and mature transcripts as distinct "genes." Latent representations of cells would then be decoded to produce parameters for independent distributions of nascent and mature molecules, as diagrammed in Figure 1a. However, the likelihoods and inferred parameters would have no biophysical interpretation and would merely be represented as part of a neural "black box" used to reduce data dimensionality.

Here, we introduce *biVI*, a strategy that adapts the variational autoencoder (VAE) framework to work with well-characterized stochastic models of transcription, and use it to fit simulated and biological datasets that include nascent and mature RNA counts. First, we propose several stochastic models of transcription, formalized by chemical master equations (CMEs), that could give rise to count distributions consistent with *scVI*. For example, Figure 1b illustrates the bursty model of transcription, in which nascent RNA molecules are produced in geometrically distributed bursts with mean b, which arrive at a constant rate k. Nascent molecules undergo splicing at rate β to produce mature molecules; these are, in turn, degraded with constant rate γ. The bursty model has been extensively experimentally validated in mammalian cells [19–21], and gives rise to a joint distribution of nascent and mature molecules parameterized by the burst size and the relative rates of transcription, splicing, and degradation. While the joint steady-state distribution induced by the bursty model is analytically intractable [22], we have previously shown that it can be efficiently and accurately approximated by the combination of judiciously chosen basis functions and a pre-trained neural network designed to predict their weights [23]. *biVI* replaces *scVI*'s independent nascent and mature likelihoods with mechanistically motivated, *b*iologically *i*nterpretable *bi*variate likelihoods. The method encodes a concatenated matrix of nascent and

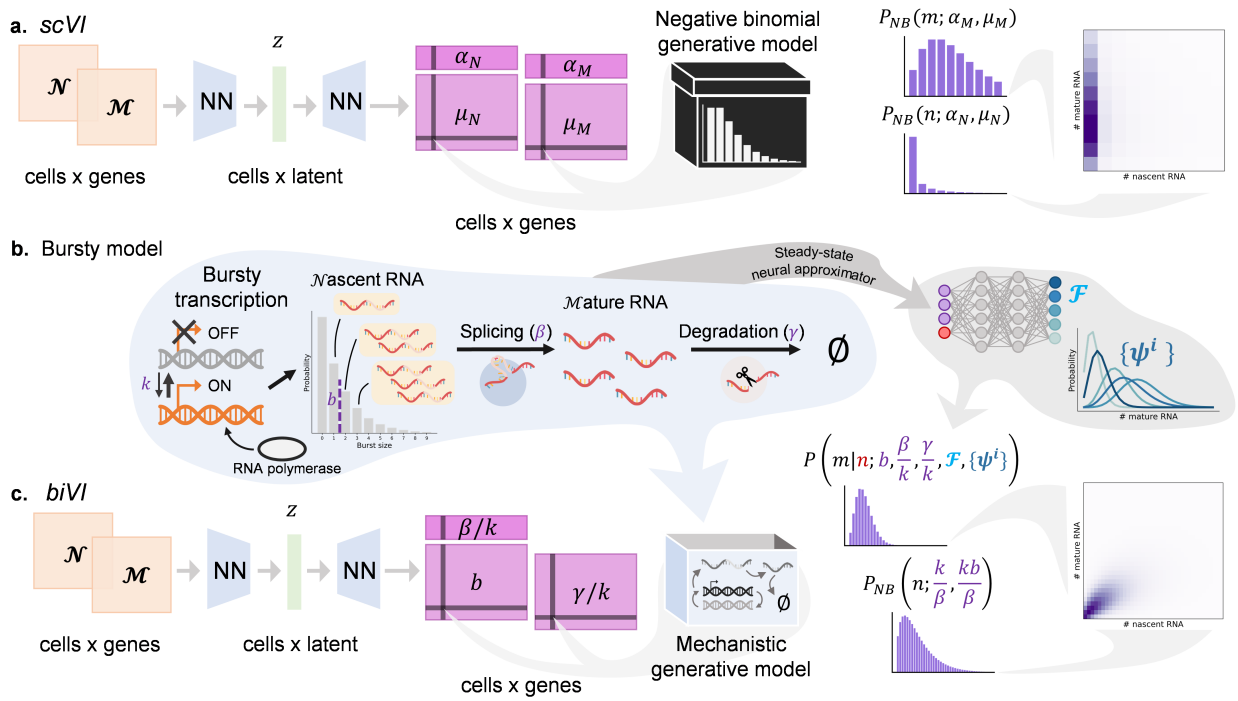

Figure 1: *biVI* reinterprets and extends *scVI* to infer biophysical parameters rather than statistical summaries. **a.** *scVI* can take in concatenated nascent ($\mathcal{N}$) and mature ($\mathcal{M}$) RNA count matrices, encode each cell with a common low-dimensional representation z, and learn the per-cell parameters μ_N and μ_M and per-gene parameters α_N and α_M that encode formally independent nascent and mature count distributions. This approach is descriptive and not motivated by any specific model of physiology. **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}. In the bursty regime, where k_{RNAP} and k_{off} are relatively high, nascent RNA molecules are produced in geometrically distributed bursts with mean $b = k_{\text{RNAP}}/k_{off}$. After production, molecules are spliced at a constant rate β and degraded at a constant rate γ. Although the two-stage telegraph model does not have a closed-form solution, its steady-state distribution can be approximated with the combination of 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 common 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.

mature counts to a low-dimensional representation, which in turn decodes to parameters which have biophysically meaningful interpretations under a particular model (Figure 1c). Using *biVI*, we can thus infer biophysical parameters for mechanistic models of transcription for tens of thousands of single cells and genes. While we focus our analysis on the bursty model, *biVI* implements two alternative descriptions of single-molecule variation, the constitutive and extrinsic noise models previously discussed in the literature [9,24,25]. The descriptions and derivations of each biophysical model are available in Section S1, with diagrams for the constitutive model in Figure S1 and extrinsic model in Figure S2.

2 Results and discussion

Before using *biVI* to infer biophysical parameters with experimental data, we sought to evaluate its ability to reconstruct distributions from a known mechanistic ground truth. By showing that *biVI* can accurately recapitulate simulated data distributions, we can increase confidence in the validity of biophysical parameters inferred from experimental data. The simulation procedure is outlined in Section 4.5, with further details in Section S2. For the bursty model, we generated parameters b, β/k, and γ/k for 2,000 genes, increasing burst sizes for marker genes to produce twenty distinct cell types (Section S2). Next, we sampled nascent and mature RNA counts for 10,000 cells, approximately equally distributed across cell types. We trained *biVI* with the bivariate bursty likelihood and *scVI* with independent negative binomial likelihoods on 80% of the simulated cells (Section 4.5), holding out 20% for testing. To reconstruct distributions for a given gene in a specific cell type, we took the average of the probability laws obtained for that cell type (Section 4.5.3). We reproduced the analysis for the constitutive and extrinsic models, with results shown in Figures S4 and S5.

On simulated data generated with the bursty model, the negative log-likelihood, or reconstruction loss, of the 2,000 held-out testing cells was lower for *biVI* (3,885.8) than for *scVI* (3,953.6) (Figure S3). Applying *biVI* to fit the model used to generate the data set gave lower reconstruction losses than *scVI* for all three simulated data sets (Figure S4, Figure S5). Furthermore, the Kullback-Liebler divergence (KLD) between true and reconstructed count distributions, separately computed for each cell type, was lower for *biVI* (0.014 on average) than *scVI* (0.212 on average) (Figure S3). Figure 2a qualitatively illustrates *biVI*'s improved ability to reconstruct distributions based on noisy sampled observations. If the model is correct, *biVI* achieves better reconstruction performance on several metrics. Conversely, the suboptimal results under the *incorrect* models suggest that *biVI* may provide a route for model selection.

A primary use of *scVI* is to reduce the dimensionality of similar cells, such as multiple observations of the same cell type, to similar low-dimensional latent representations, removing noise and redundancies. *biVI* facilitates this application for multimodal data. To characterize how well cell types are preserved in the latent space, we applied several clustering metrics (Section S4) to the latent representation of 2,000 cells (simulated, bursty model, held-out set) encoded by *biVI* and *scVI* (Section S2), using the ground truth cell types as cluster assignments. The average percent of latent-space nearest neighbors in the same cell type as each cell, computed over all cells, was similar for *biVI* (90.0%) and *scVI* (90.4%). The silhouette score, a measure of how similar each cell is to other cells in its cluster compared to other clusters, achieved analogous performance, yielding 0.383 for *biVI* and 0.387 for *scVI*. Finally, scaled inter-cluster distance (average inter-cluster distance divided by average intra-cluster distance, with larger values corresponding to better cluster

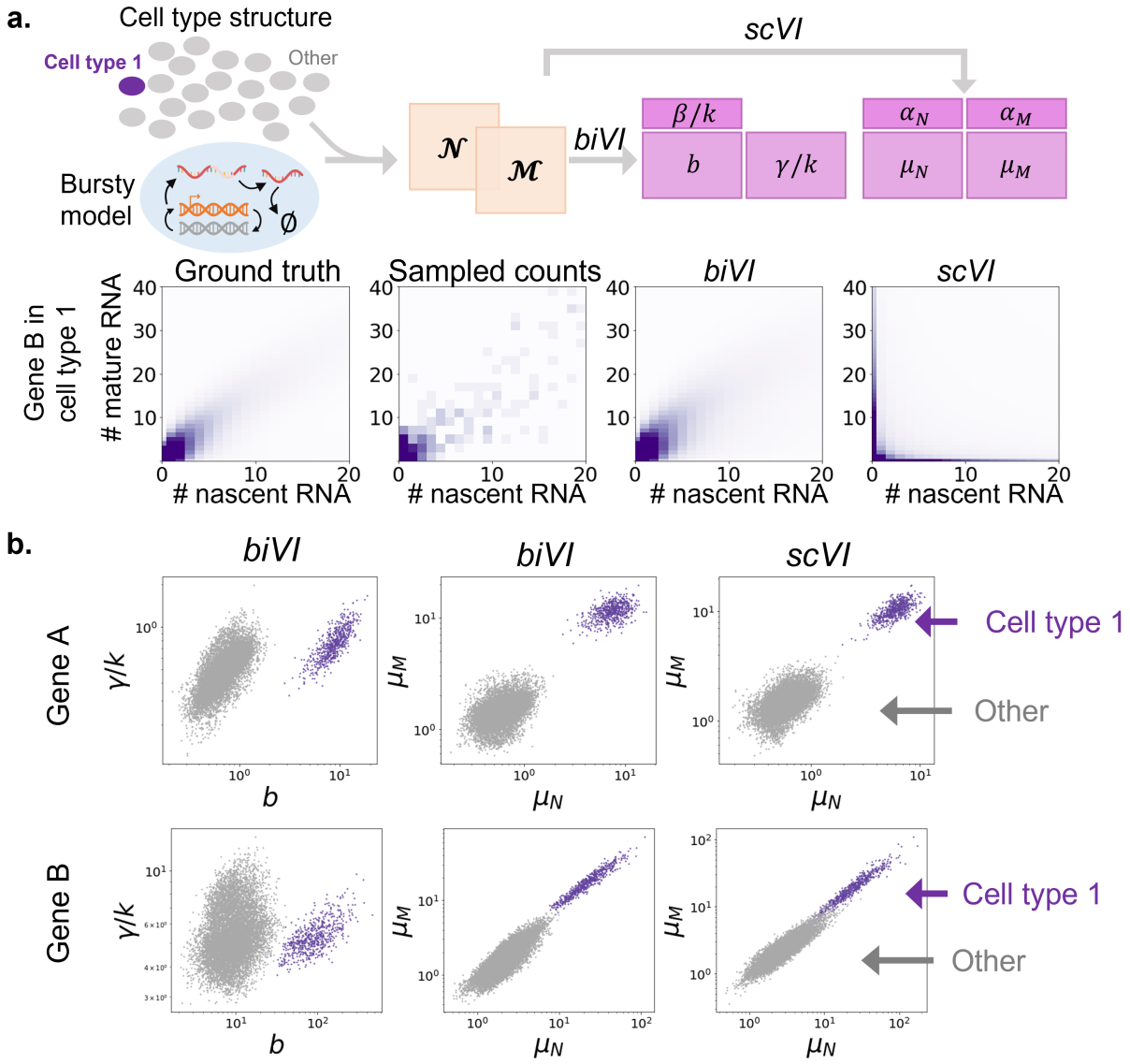

Figure 2: *biVI* reflects the global structure and stochasticity of simulated bursty data. **a.** We generated parameters for 2,000 genes under the bursty model of transcription, and sampled nascent and mature counts from the resulting joint distribution to yield 10,000 cells across 20 cell types. Next, we trained *biVI* with the bivariate bursty likelihood and *scVI* with independent negative binomial likelihoods. For a sample gene, *biVI* accurately reconstructs the underlying bivariate distribution (darker color: higher; lighter color: lower probability mass or number of observations). **b.** Cell-specific parameters inferred by *biVI* and *scVI* for two marker genes for cell type 1, Gene A and Gene B. *biVI* yields nascent and mature means, as well as parameter values, which are related to the means through a transformation. *scVI* only yields the means.

separation), was 2.75 for *biVI* and 2.60 for *scVI*. These clustering metrics are evidence that cells can be similarly well represented in a low dimensional space using *biVI* and *scVI*, and that *biVI* modifications do not impede the process of variational inference or downstream analyses.

By incorporating nascent data, *biVI* enables the detection of interesting differences in gene expression that may be subtle or poorly identifiable from mature RNA counts alone. By using a specific biophysical model, it provides an explanatory framework. Figure 2b shows inferred parameters across all simulated cells for two marker genes, which have higher burst sizes in the simulated cell type 1. The first column plots the *biVI* inferred burst size b against relative degradation rate γ/k for every cell, the second column displays *biVI* inferred mature means μ_M and nascent means μ_N, whereas the third displays the *scVI* inferred means. The *biVI* means and parameters are related through the transformations outlined in Section 4.9. Gene A shows a clear separation of cell type 1 in all workflows. In contrast, Gene B shows a clearer separation of parameter values than averages, and correctly ascribes it to a change in burst size.

With these simulation benchmarks suggesting that *biVI* can perform as well as or better than *scVI*, we proceeded to use the two pipelines to analyze experimental data. We selected a single scRNA-seq library generated from mouse brain tissue [26] and processed it with *kallisto*|*bustools* [2] to obtain spliced and unspliced count matrices (Section 4.7). These count matrices contained the data for 6,418 previously annotated cells across 19 cell subtypes with expression measurements for the 2,000 most highly variable genes (Section 4.7). Next, we fit the counts with *biVI* (Section 4.8), making the key assumption that unspliced and spliced molecules may be treated as the nascent and mature species of the bursty generative model. 4,622 cells were used to train both VAE models with 513 validation cells, with 1,283 testing cells held out for assessing performance.

Figure 3a-b shows the observed and reconstructed distributions of *Foxp2* (an L6 CT neuron marker), and *Rorb* (an L5 IT neuron marker), restricted to their respective cell types. Consistently with simulations, qualitative visual inspection suggests that *biVI* recapitulates distribution shapes better than *scVI*.

In addition to accurately capturing observed distributions, we can interpret the inferred parameters to determine *how* genes are regulated. Figure 3c-d shows the decoded averages and parameters for *Foxp2* and *Rorb* across all cells. The average nascent and mature counts are higher in the cell types they mark, highlighted in corresponding colors. However, additionally, the parametric *biVI* results suggest these genes are upregulated at the transcriptional level, by increasing the burst size rather than decreasing the degradation rate.

We can leverage these differences to compare gene regulation trends across numerous cell types. We identified genes with significant differences in the parameters of one cell type versus all others using a two-sided t-test (Section 4.9), using a Bonferroni-corrected p-value threshold of 0.05 and a $\log_2$ fold change threshold of 1. To stay consistent with *scVI* [15], and avoid the conflation of "systematic" and "sequencing depth"-like effects, we evaluated differences in the compositional, "normalized" versions of inferred burst sizes, nascent means, and mature means, independent of the *scVI* cell size factor. We note that Figure 3 demonstrates the "scaled," or non-normalized versions of these quantities. The relative degradation rate is independent of sequencing depth, so differential expression analysis was performed directly on the inferred values. To avoid the potential pitfalls of small sample sizes, we excluded all cell types with fewer than 10 barcodes (L5 ET, L6 IT Car3, VLMC, SMC) from this analysis. Figure 3e shows the fraction of identified genes in each cell subclass that exhibited differences in burst size, relative degradation rate, or both. While cell subclasses modify expression of genes through a variety of regulatory strategies (modifying

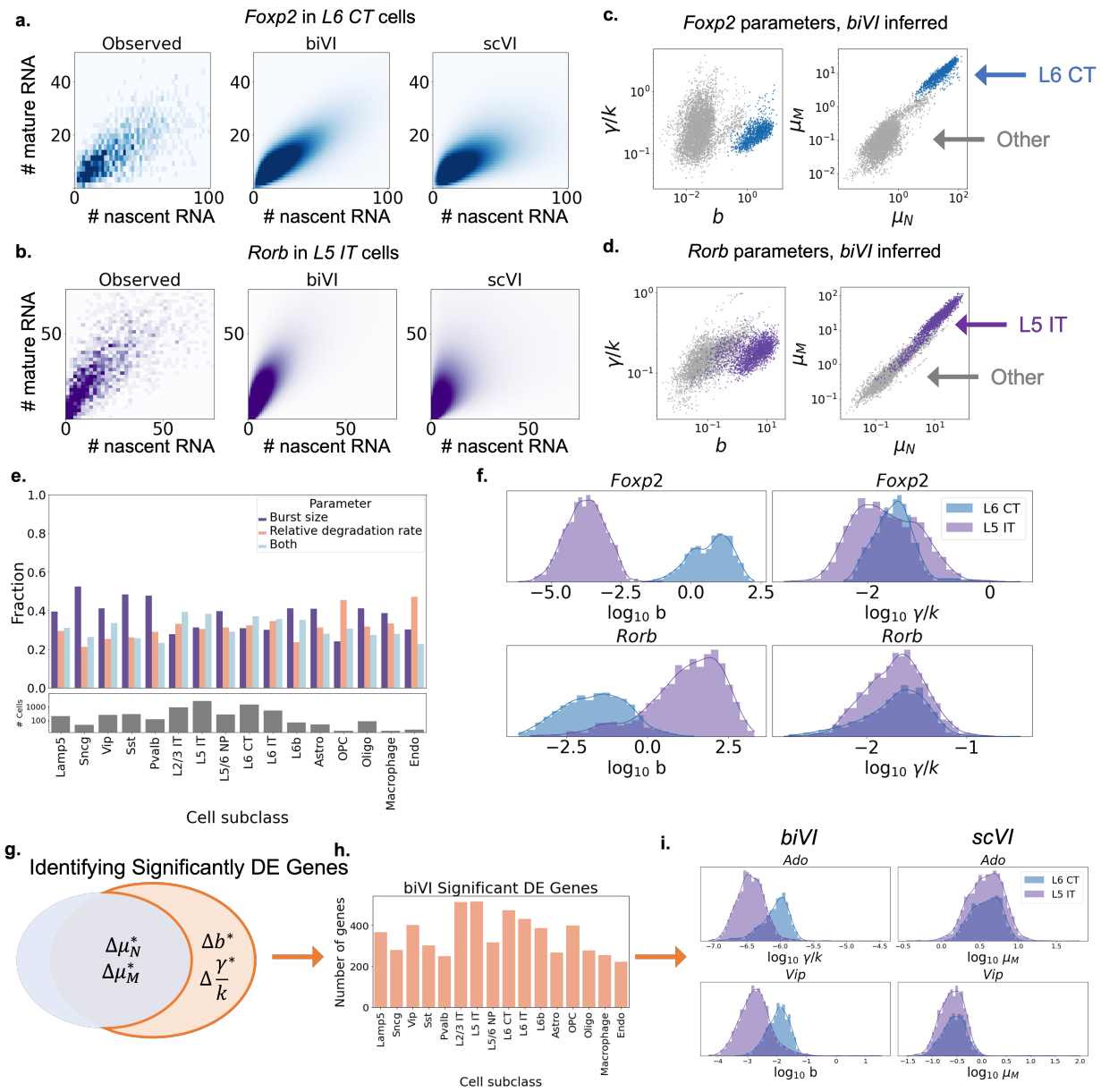

Figure 3: *biVI* successfully fits single-cell neuron data and suggests the biophysical basis for expression differences. **a.-b.** Observed and inferred distributions of *Foxp2*, a marker gene for L6 CT cells, and *Rorb*, a marker gene for L5 IT cells (conventions as in Figure 2a). **c.-d.** Cell-specific 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.** Inferred parameters for *Foxp2* and *Rorb* in L5 IT (purple) and L6 CT (blue) cells show consistent differences in burst size. **g.** *biVI* allows the identification of cells which exhibit differences in burst size or relative degradation rate, without necessarily demonstrating differences in mean expression. **h.** Hundreds of genes demonstrate this modulation behavior, albeit with variation across cell subclasses. **i.** Histograms of *biVI* parameters and *scVI* mature means for two genes that exhibit parameter modulation without identifiable mean modulation. *Ado* appears to show differences in the degradation rate, whereas *Vip* appears to show differences in the burst size (top: results for *Ado*; bottom: results for *Vip*; purple: L5 IT cells; blue: L6 CT cells).

burst size, degradation rate, or both), some interesting patterns can be observed. For example, the GABAergic subclasses (first five categories) appeared to more consistently exhibit deviation in burst size than in degradation rate. These coarse-grained differences may suggest common regulatory behaviors are present in similar cell types, although more sophisticated statistical methods are necessary to draw rigorous conclusions about these trends.

Finally, *biVI* identified novel genes that were significantly differential between cell subclasses in burst size or relative degradation rate but not nascent or mature inferred means (Section 4.9, Figure 3g-h). For some cell types, there were several hundred such genes, potentially interesting targets for follow-up experimental investigation. For example, the gene *Ado* demonstrated statistically significant differences in the *biVI* inferred relative degradation rate, but not *scVI* inferred mature means, in the neuronal subclass L6 CT (Figure 3i, top row). *Ado* has been associated with gliablastoma metabolism in non-neuronal cells [27]; the observation that it shows significant differences in inferred relative degradation rates in a neuronal cell subclass could lead to a broader understanding of the role its noise behaviors play in the growth and progression of malignant gliomas. As another example, burst size of *Vip*, the gene that codes for the peptide hormone vasoactive intestinal peptide (VIP), is statistically significantly different in neuronal cells of subclass L5 IT (Figure 3i, bottom row). While VIP is known to act as a hormone receptor and marks a GABAergic neuronal subclass (Vip cells) [28], the regulation of its burst size in (glutamatergic) L5 IT cells suggests functions in neurons outside of the subclass it marks, a suggestive area for follow-up investigation. Such mechanistic description provides a framework for characterizing the connection between the gene's role and the cell's regulatory strategies beyond the mean expression [29,30].

3 Conclusion

We have demonstrated that a mechanistically interpretable autoencoder can be defined in a self-consistent way for bivariate data. This autoencoder, implemented in the *biVI* framework, combines the functional and distributional assumptions of *scVI* with the mechanistic hypotheses of transcriptional biophysics. When applied to simulated data, it correctly recovers three properties of interest. (1) Recapitulation of the gene-specific distributions, as characterized quantitatively through the KLD and qualitatively through the distribution shape. (2) Maintenance of cell type structure, as quantified by a variety of clustering metrics. (3) Recovery of the physical parameters used to generate count matrices, including the biophysical basis of differences in marker genes. When applied to real sequencing data, it achieves similar performance, and attains a degree of interpretability: we can begin to attribute differences in expression to specific regulatory mechanisms.

This interpretability is both enabled and limited by the modeling decisions: the biological findings are only as meaningful as the model. For simplicity and consistency with *scVI*, we have made several key assumptions. Most crucially, we propose that the model has been correctly specified, and that only the burst sizes and degradation rates are modulated between different cell types or states. It is likely that the latter assumption can be made somewhat more physically realistic by proposing that regulation primarily occurs at the gene locus, i.e., allowing the burst size b and burst frequency k to vary while keeping the processing rates β and γ constant. At this time, the most appropriate way to mathematically formulate this model while maintaining compatibility with neural networks is not yet clear. On the other hand, controlling for misspecification of the generative model appears to be a much more challenging problem, whether in terms of establishing a goodness-of-fit criterion or in terms of performing model selection simultaneously with inference.

Preliminary comparisons, wherein we use multiple models to fit simulated datasets, suggest that the correct model may be identifiable, but the appropriate statistical procedures are not yet clear. Relaxing model assumptions and implementing these extensions are the next natural steps for the interpretable summary of large datasets.

More fundamentally, we have assumed that a variational autoencoder framework is *a priori* appropriate for single-cell RNA sequencing datasets. However, this framework is incompatible with certain common axioms [31]. For example, *scVI* and *biVI* cannot represent truly discrete cell types, i.e., sets of cells that have an identical latent representation but may have distinct count distributions due to single-molecule stochasticity. In addition, steady-state probability distributions are intrinsically incompatible with *transient* phenomena, such as differentiation (as discussed in [31]) and cell cycling, limiting the scope of this approach to stationary systems. Finally, we have merely asserted that the *biVI* mean parameters can be represented in a neural, compositional form, and demonstrated that this formulation produces acceptably faithful data summaries; however, we have not taken the next step of justifying or critiquing this assertion. To improve the prediction of nascent and mature count matrices, we have imposed biophysical constraints on the species' joint distributions. Although the generative model is interpretable, the neural component is still a computational heuristic. In a similar spirit, but the orthogonal direction, previous work in the summary of single-cell data has used *linear* decoders, which constrain the compositional gene abundance parameters to be linear transformations of the latent embedding vector [32]. We anticipate that broader improvements in interpretability will require considerable further investigation and formalization, potentially with analogous, hypothesis-driven constraints on the variational autoencoder structure.

In spite of the limitations imposed by these assumptions, the approach we present in this report provides a fairly general, principled toolbox for the treatment of multimodal datasets. As we discuss in the mathematical treatment, the analytical procedure involves instantiating a set of biophysical hypotheses, analyzing the resulting stochastic system to obtain its distribution, and using this distribution as a multivariate generative model. As single-cell technologies evolve to provide larger-scale and more precise measurements of biomolecules, we anticipate that this approach can be self-consistently extended to provide a more comprehensive picture of biophysical processes in living cells.

4 Methods

In order to extend the *scVI* method to work with multimodal molecule count data in a way that is coherent with biology as well as the autoencoder, 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 4.2), 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 4.3). 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. 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.

4.1 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 $\mathbb{N}_0$:

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

4.2 *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. Each cell is represented by a low-dimensional vector z, which is formally a random variable. *scVI* uses the "decoder" neural network to generate the quantities ρ_{cg}, which describe the compositional abundance of gene g in cell c as a function of z, such that $\sum_g \rho_{cg} = 1$. Furthermore, it either fits or uses a plug-in estimate for ℓ_c, a cell-specific "size factor," 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.

4.3 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 [24,25]) or technical variability in the sequencing process (in the spirit of [33]). 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.

4.3.1 *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.

4.3.2 *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.

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

The negative binomial generative model may be recapitulated by the alternative schema [22]:

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

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 [22]. However, this strategy is inefficient in the variational autoencoder framework, where each observation is associated with a distinct set of parameters. Furthermore, it is likely incompatible with automatic differentiation.

In [23], 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.

4.4 *biVI* modifications

We built modifications upon *scVI* version 0.18.0 [34]. The *scVI* framework already supports the constitutive model. By setting conditional likelihood to "poisson," no modification of *scVI* architecture is necessary. The conditional likelihood is the product of two Poisson distributions (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 $\alpha \in \mathbb{R}_{\geq 0}^{N_G}$ rather than $\alpha \in \mathbb{R}_{\geq 0}^{2N_G}$, where N_G is the number of genes. For the extrinsic model, the conditional likelihood is set to the extrinsic likelihood $P_{\text{extrinsic}}(n, m; \alpha, \mu_N, \mu_M)$ (Equation 11). For the bursty model, the conditional likelihood is set to the bursty likelihood $P_{\text{bursty}}(n, m; \alpha, \mu_N, \mu_M)$ (Equation 13).

4.5 Simulated data

To validate our implementation of *biVI*, and to understand potential pitfalls inherited from the standard *scVI* autoencoder workflow, we generated ground truth data by simulation. Simulated ground truth is limited in important ways, as it omits known and unknown sources of biological and technical variability. However, to begin to understand the limitations of integrating descriptive autoencoders with mechanistic models, it is helpful to have simple and well-understood simulated datasets. By judiciously choosing simulation scenarios, we can characterize the relative and absolute accuracy of the procedures under ideal-case conditions, which describe a natural upper bound on their performance.

The simulation procedures are described in Section S2. For each of the three models (constitutive, extrinsic, and bursty), we simulated a data set of nascent and mature RNA counts for 2,000 genes across 10,000 cells, spread across 20 cell types. The cell types were distinguished by low-magnitude variation in all genes' parameters, as well as higher-magnitude upregulation of expression in a small set of cell type-specific marker genes. In the extrinsic noise model, upregulation was effected by increasing the transcription rate scale η. In the analogous constitutive case, it was effected by increasing the transcription rate k. In the bursty model, it was effected by increasing the burst size b.

4.5.1 Fitting simulated data

We fit the three simulated data sets using *biVI* with the three generative physical models, as well as standard *scVI* with the negative binomial distribution as its conditional likelihood. All models were trained on 7,200 cells with 800 validation cells for 400 epochs with a learning rate of 0.001. The encoder and decoder networks consisted of 3 layers with 128 nodes each, with a latent space of dimension 10. As in standard *scVI*, a standard normal prior was used for the latent space.

4.5.2 Performance metrics on simulated data

Performance of the trained models was tested on 2,000 held-out cells for each simulated data set. In order to characterize the absolute reconstruction of ground truth cell types, the mean squared error (MSE) between *biVI* inferred means, μ_N and μ_M, and ground truth simulated means was calculated for each cell. Furthermore, we assessed how well different models reduced cells of the same cell type to similar latent spaces while maintaining good separation of cells in different cell types. To compare clustering accuracy for different models, we applied several metrics to the latent spaces obtained for testing data, using simulated ground truth cell type as cluster assignment. We calculated average intra-cluster distance (ICD), or the Euclidean distance between each cell's latent representation and the mean of its assigned cluster in the latent space averaged over all cells. We also calculated average intra-cluster variance (ICV), or the variance of ICDs for each cell in a cluster averaged over all clusters. Inter-cluster distance was also calculated: distance between cluster means averaged over all pair-wise clusters (excluding the cluster's distance to its own mean). Finally, scaled inter-cluster distance, or inter-cluster distance divided by average intra-cluster distance, was found. We also report nearest neighbor percentages (Figures S3, S4, S5), or the percent of the n_κ nearest neighbors in the same cell type as each cell, where n_κ is the number of cells in the given cell's cell type. A quantitative description of clustering metrics is provided in Section S4. Broadly speaking, these metrics characterize the methods' utility for discovering discrete cell populations using typical clustering algorithms, which perform best when low-dimensional clusters are relatively condensed, homogeneous, and reflect the high-dimensional structure.

4.5.3 Reconstructing gene distributions

A simulated cell type is defined by cell type-specific parameters for each gene, $\theta_{\kappa g}$, where κ indexes over cell types, g indexes over genes, and θ contains the model parameters. Each cell in a cell type has identical parameters. *biVI* and *scVI*, however, infer parameters for every cell and gene: θ_{cg}, where c indexes over cells and g indexes over genes. To compare reconstructed distributions to simulated ground truth distributions for a given gene of a given cell type, we first select all cells of that cell type. Next, we define the distribution under a given model for each cell using *biVI* or *scVI* inferred parameters. We average over the cell-specific probabilities for a given gene to produce a cell-type distribution for that gene:

$$\hat{P}_{\kappa g}(n,m) = \frac{1}{n_\kappa} \sum_{c_\kappa=1}^{n_\kappa} P(n,m;\theta_{c_\kappa g}), \tag{14}$$

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:

$$P(n,m;\theta_{c_\kappa g}) = P_{\mathrm{NB}}(n;\alpha_g^N,\mu_N)P_{\mathrm{NB}}(m;\alpha_g^M,\mu_M), \tag{15}$$

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 4.4). 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.

4.5.4 Kullback-Leibler divergence between simulated and reconstructed distributions

We compare *biVI* and *scVI*'s ability to accurately reconstruct ground truth gene distributions by calculating the truncated Kullback-Leibler divergence (KLD) between ground truth simulated distributions and reconstructed distributions for all genes in all cell types. We reconstruct gene and cell-type specific *biVI* and *scVI* distributions $\hat{P}(n,m)$ as described above in Section 4.5.3. We also calculate true distributions $Q_{\kappa g}(n,m;\theta_{\kappa g})$ using the gene and cell-type specific ground truth simulated parameters $\theta_{\kappa g}$ (parameter generation and sampling are described in Section S2).

We calculate both true and reconstructed model (*biVI* or *scVI*) distributions over a 50 by 50 grid of nascent and mature values, then normalize both to sum to 1.0 over the grid by dividing all probabilities by their sum over the grid. Normalization ensures that KL divergences are well-defined. We then calculate the KL divergence over the restricted grid:

$$\text{KLD} = \sum_{m=0}^{49} \sum_{n=0}^{49} \hat{P}_{\kappa g}(n,m) \times \log \frac{\hat{P}_{\kappa g}(n,m)}{Q_{\kappa g}(n,m;\theta_{\kappa g})}. \tag{16}$$

4.6 Connecting the models to transcriptome data

Thus far, we have used "nascent" and "mature" as shorthands for the discrete species in a two-stage model of RNA processing. In other words, Equation 5 is axiomatic for this nomenclature. We have gone one step further and named the rate of conversion β the *splicing rate*, explicitly identifying the nascent species with unspliced RNA and the mature species with spliced RNA. This identification is a modeling decision that elides the considerable simplification of biological complexities. In the current section, we elaborate on the assumptions.

In the field of microbiology, "nascent" RNA is often, but not always, used to characterize the mRNA molecules in the process of synthesis, associated to a DNA strand via an RNA polymerase complex [35–38]. In this framework, the "mature" transcriptome is simply the complement of the nascent transcriptome, i.e., all molecules that are not chemically associated to a DNA strand. Therefore, the canonical definition of "nascent" RNA is equivalent to *transcribing* RNA, which is a polymeric structure with a particular sequence.

Transcribing RNA can be observed directly through electron micrography [37]. However, more typically, it is investigated through more or less direct experimental proxies that can be scaled to many genes and cells at a time. In the single-cell fluorescence subfield, DNA or membrane staining can be used to identify bright spots localized to the nucleus, which is treated as signal from RNA at the transcription site [29,39]; this signal may include contributions from RNA incidentally, or mechanistically, retained at a DNA locus [36]. In this strategy, "nascent" molecules are DNA-associated. Alternatively, and more commonly, transcribing molecules have been studied by using probes targeted to the $5'$ and $3'$ regions [40–43], or to intronic and exonic regions [44–47]. In this strategy, "nascent" molecules contain a particular region, either synthesized earlier or removed later in the RNA life-cycle.

The use of intron data as a proxy for active transcription is reminiscent of, but distinct from sequence census [48] strategies that directly study RNA sequences. These strategies, in turn, typically use chemical methods to enrich for newly transcribed RNA. For example, Reimer et al. isolate chromatin, then deplete sequences that have been post-transcriptionally poly(A) tailed [49]. Analogously, Drexler et al. use 4-thiouridine (4sU) labeling to enrich for newly synthesized molecules [50]. These approaches may produce conflicting results; for example, introns may be rich both in poly(A) handles [1] and 4sU targets [49], giving rise to obscure technical effects. Therefore, these "processed" or "temporally labeled" proxies are coarsely representative of transcriptional dynamics, and their quantitative interpretability is unclear as of yet.

The sequence content may be used more directly, by conceding that DNA association or localization are not easily accessible by sequence census methods, and treating splicing *per se*. This approach has a fairly long history. Intronic quantification has been used to characterize transcriptional mechanisms in microarray datasets [51], and to characterize differentiation programs in RNA sequencing [52,53]. In single-cell RNA sequencing, intronic content has been leveraged to identify transient behaviors from snapshot data [1], albeit with some outstanding theoretical concerns and caveats [54]. Briefly, it is, in principle, possible to coarsely classify molecules with intronic content as "unspliced" or "pre-mRNA" and aggregate all others as "spliced," "mature," or simply "mRNA." By applying this binary classification, and defining a simple first-order model of splicing, data may be successfully fit. The quantification of transcripts so classified is a relatively straightforward genomic alignment problem. The multiple available implementations [1,2,55,56] tend to disagree on the appropriate assignment of ambiguous sequencing reads [54,56], obscuring a more fundamental problem: the binary classification is mechanistically limited [50,57–59], and it is likely that detailed splicing graph models will be necessary in the future [31,54]. The focus on sequence is yet another step removed from the transcriptional dynamics, particularly since some of the splicing processes occur after transcriptional elongation is complete [60]. However, in spite of its limitations and assumptions, the simple Markovian two-stage model has been successful in the past [61,62]. Mechanistic evidence does not suggest that, e.g., deterministic delayed elongation is necessary to represent "unspliced" distributions [63]: under this model, we appear to be able to treat splicing and degradation as Markovian, without representing the elongation process at all, and obtain reasonable fits to the data. Somewhat surprisingly [64], the geometric-Poisson distribution, which describes bursty transcription coupled to deterministic elongation [65], is a particularly poor fit for unspliced RNA counts [63].

Adding yet more complexity to the modeling, "mature" – whether "off-template," "spliced," or "processed" – molecules are not immediately available for degradation; first, the process of nuclear export must take place. Studies that presuppose access to imaging data tend to model it explicitly [22, 29, 66–68]. However, this approach has not been applied in sequencing assays, as current technologies do not distinguish nuclear and cytoplasmic molecules. Furthermore, comparisons of paired single-cell and single-nucleus datasets are hampered by the limited characterization of the noise sources in the latter technology. Yet again, we have generally found that omitting this effect in single-cell data produces acceptable fits [61–63].

Pending the development of more sophisticated sequencing and alignment technologies, as well as the implementation of tractable models of biology, the data exploration portion of our study focuses on the "spliced" and "unspliced" matrices generated by *kallisto* |*bustools* [2]. This choice is a compromise, and we adopt it after considering the following factors:

- Availability of quantification workflows: spliced and unspliced matrices are straightforward to generate.
- Model tractability: the two-stage models can be evaluated [23]; more sophisticated models require new algorithms to be integrated with variational autoencoders.
- The scope of sequencing data: single-cell protocols do not yet give access to sub-cellular information, so inference of elongation or nuclear retention dynamics is acutely underspecified.
- Self-consistency and compatibility: *a priori*, we seek to recapitulate the data types (spliced RNA counts) and distributions (negative binomial and Poisson) already implemented in *scVI*, leaving improvements to further investigation.

We use the terms "nascent" and "mature" to identify the unspliced and spliced RNA matrices. This choice of nomenclature is deliberate. Although it somewhat conflicts with the established microbiology literature, this terminology is intended to emphasize the models' generality, both here and elsewhere [9, 23, 69]. The two-stage Markovian process is axiomatic. The specific identities assigned to the mathematical objects may range beyond counts identified by sequence census methods. They may represent the discretized and subtracted intensities of 3', 5', intron, or exon fluorescent probes, the counts of molecules within and outside the nuclear envelope, or polymerase counts obtained by micrography. Therefore, the terminology should be taken in the sense used for similar non-delayed models in [64, 66, 70, 71].

4.7 Preprocessing Allen data

Raw 10x v3 single-cell data were originally generated by the Allen Institute for Brain Science [26]. The raw reads in FASTQ format [72] and cluster metadata [73] 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`). 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 [26], 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) [26] 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 [74] to restrict our analysis to several thousand variable genes, which presumably reflect the cell type signatures of interest. The mature 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` with minimum mean of 0.0125, maximum mean of 3, and minimum dispersion of 0.5 [74].

4.8 Fitting Allen data

We applied *biVI* with the three generative models (bursty, constitutive, and extrinsic) and *scVI* with negative binomial likelihoods to the count matrices obtained by the filtering procedures outlined above. 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.

4.9 Differential parameter values and gene expression

After fitting the VAE models, we sought to identify meaningful statistical differences that distinguish cell types. In order to avoid instabilities, we excluded cell subclasses "L6 IT Car3," "L5 ET," "VLMC," and "SMC" from this analysis, as they contained fewer than ten annotated cells, necessitating more sophisticated statistical models that take into account their 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 and gene expression were tested for each assigned subclass label (as annotated in [26]) versus all others using two-sided t-tests implemented in `scipy.stats.ttest_ind` [75] under the null hypothesis that there was no difference in the average of the tested values. We applied Bonferroni corrections for the two-tailed test, the 16 subclass labels, and the 2,000 genes: values were considered significantly differential if the p-value under the t-test was below $0.05 \times (2 \times 16 \times 2,000)^{-1} = 7.813 \times 10^{-7}$. To assign significance, we further required that the magnitude of $\log_2$ fold change to be greater than 1, i.e., the arithmetic mean of the tested value was doubled or halved in one cell subclass versus the rest. We adopted this effect size threshold due to the relatively large number of cells and corresponding prevalence of genes with significant p-values.

The t-tests were performed on normalized burst size and means, i.e., the inferred burst size or inferred means divided by observed sequencing depth for that cell. The observed sequencing depth of a cell ℓ_c was set to the sum of all observed counts, nascent and mature, in that cell. Normalization was thus done to ensure that identified differences in parameters was not due to sequencing depth but was biologically meaningful. Relative degradation rate is independent of sequencing depth: t-tests were performed directly on inferred relative degradation rates.

As discussed in Section 4.4, *scVI* fits the cell-specific parameters $\rho_{cg}^{(N)}$ and $\rho_{cg}^{(M)}$, as well as gene-specific parameters α_g^N and α_g^M. To compute effect sizes and p-values, we simply compared the distributions of ρ_{cg} in distinct cell populations, using $\rho_{cg}^{(N)}$ for differences in nascent expression and $\rho_{cg}^{(M)}$ for differences in mature expression.

As discussed in Section S1.3, *biVI* fits $\rho_{cg}^{(N)}$, $\rho_{cg}^{(M)}$ and α_g under the bursty generative model. To calculate parameters, we use the definitions outlined in the section, enforcing $\alpha_g = k_g/\beta_g$. This yields the following relationships:

$$\begin{aligned} \beta_g &= \frac{k_g}{\alpha_g}, \\ \gamma_z &= \frac{\rho_{cg}^{(N)}}{\rho_{cg}^{(M)}} \frac{k_g}{\alpha_g}, \\ b_z &= \frac{\rho_{cg}^{(N)}}{\alpha_g} C, \end{aligned} \tag{17}$$

where C is a large multiplicative constant shared across cells and genes. We set k_g to unity with no loss of generality at steady state. Finally, we compare the inferred values of γ_z/k_g and b_z/C to identify differentially expressed genes.

5 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.

6 Acknowledgments

M.C., G.G., T.C., and L.P. were partially funded by IGVF-1-UCI.IGVF 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, 2, 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] R. Argelaguet, D. Arnol, D. Bredikhin, Y. Deloro, B. Velten, J. C. Marioni, and O. Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21, 2020.

[15] 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.

[16] 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.

[17] Francis Crick. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, August 1970.

[18] Gene-Wei Li and X. Sunney Xie. Central dogma at the single-molecule level in living cells. *Nature*, 475(7356):308–315, July 2011.

[19] 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.

[20] 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.

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

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] D. Shen, L. Tian, F. Yang, J. Li, X. Li, Y. Yao, E. W. F. Lam, P. Gao, B. Jin, and R. Wang. Ado/hypotaurine: a novel metabolic pathway contributing to glioblastoma development. *Cell Death Discovery*, 7, 2021.

[28] I. Gozes and D. E. Brenneman. Vip: molecular biology and neurobiological function. *Molecular Neurobiology*, 3:201–236, 1989.

[29] 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.

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

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

[32] 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.

[33] 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.

[34] 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.

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

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

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

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

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

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

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

[61] Gennady Gorin and Lior Pachter. Length Biases in Single-Cell RNA Sequencing of pre-mRNA. Preprint, bioRxiv: 2021.07.30.454514, July 2021.

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

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

[64] 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.

[65] 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.

[66] 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.

[67] 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.

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

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

[70] 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.

[71] 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.

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

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

[74] 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.

[75] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1.0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, March 2020.

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

[77] 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.

[78] 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.

[79] 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.

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

[81] 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.

[82] 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.