Analytical solutions of the chemical master equation with bursty production and isomerization reactions

Gennady Gorin[1] and Lior Pachter[2]

[1]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, 91125
[2]Division of Biology and Biological Engineering & Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, 91125
[*]Address correspondence to Lior Pachter (lpachter@caltech.edu)

March 24, 2021

Abstract

Splicing cascades that alter gene products post-transcriptionally also affect expression dynamics. We study a class of processes and associated distributions which emerge from a bursty promoter coupled to a path graph of downstream mRNA splicing, and more generally the behavior of finite-activity jump distributions coupled to rooted directed acyclic graphs of splicing. These solutions provide full time-dependent joint distributions for an arbitrary number of species, offering qualitative and quantitative insights about how splicing can regulate expression dynamics.

1 Background

Recent advances in the analysis of single-cell RNA sequencing [1] and fluorescence microscopy [2] enable the quantification of pre-mRNA molecules alongside mature mRNA. These techniques provide an opportunity to infer the topologies and biophysical parameters governing the processes of mRNA transcription, splicing, export, and degradation in living cells. In particular, they provide a novel approach to inferring and studying the dynamics of splicing cascades [3].

However, drawing mechanistic conclusions from transcriptomics requires overcoming numerous statistical and computational challenges. Living cells contain mRNA in low copy numbers, and transient nascent species are even less abundant, leading to potential pitfalls if the discrete nature of the data is not appropriately modeled [4]. One approach to modeling dynamics from count data is to utilize detailed Markov models based on the chemical master equation (CME) [5–7]. Such modeling can, in principle, yield analytical solutions for the dynamics of genes affected by arbitrary splicing networks, thus circumventing tractability problems arising with matrix- and simulation-based methods that can be impractical for large numbers of species [8]. Several classes of analytical solutions to the CME are available [9], but their derivation tends to be *ad hoc*, with limited generalization to more complex systems.

Starting with the approach of Singh and Bokes to the problem of bursty transcription coupled to nuclear export and degradation of mRNA [10], we develop a class of solutions for gene dynamics affected by arbitrary splicing networks under the physiologically relevant assumption of bursty production of mRNA [11]. Fundamentally, the generating function of a Markov chain describing the evolution of molecules in a splicing cascade can be automatically computed, numerically integrated in time, and then inverted by Fourier transformation at an overall computational complexity of $O(\mathcal{N}\log\mathcal{N})$ in state space size. We begin with the example of splicing described by a path graph, where the order of intron removal is deterministic, demonstrate the existence of the solutions and their isomorphism to a class of moving average processes, and subsequently extend the procedure to a much broader class of splicing graphs and driving burst processes.

2 Path graph splicing

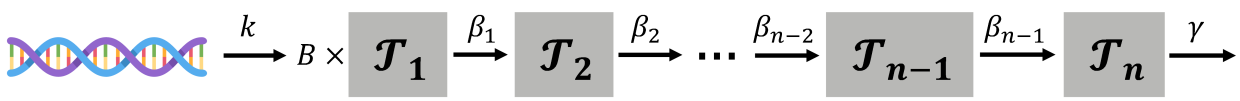

Figure 1: Graph representation of the generic path graph model. The source transcript $\mathcal{T}_1$ is synthesized at the gene locus in random geometrically distributed bursts (according to a distribution B with burst frequency k). Each molecule proceeds to isomerize in a chain of splicing reactions governed by successive rates $\beta_1, \beta_2, \ldots, \beta_{n-1}$, until reaching the form $\mathcal{T}_n$, which is ultimately degraded at rate γ.

Consider the system consisting of a bursting gene coupled to a n-step birth-death process, characterized by the path graph in Figure 1, where $B \sim Geom(b)$, and all reactions occur after exponentially-distributed waiting times. The bursts occur with rate k, the conversion of adjacent transcripts $\mathcal{T}_i$ to $\mathcal{T}_{i+1}$ occurs with rate β_i, and the degradation of $\mathcal{T}_n$ occurs with rate $\gamma := \beta_n$. We assume the rates of conversion and degradation are all distinct. The amount of species $\mathcal{T}_i$ can be described by the non-negative discrete random variable m_i.

2.1 Discrete formulation and algorithm

We would like to compute the probability mass function (PMF) of the count distribution, denoted by $P(m_1, ..., m_n, t)$; this corresponds to the probability of observing m_1 molecules of $\mathcal{T}_1$, m_2 molecules of $\mathcal{T}_2$, etc., at time t. Following a previous derivation [10], this problem can be reframed in terms of a partial differential equation involving the probability generating function (PGF) $G(x_1, ..., x_n, t)$:

$$G := \sum_{m_1,\ldots,m_n} x_1^{m_1} \ldots x_n^{m_n} P(m_1, ..., m_n, t)$$

$$\frac{\partial G}{\partial t} = k(F(x_1) - 1)G + \sum_{i=1}^{n-1} \beta_i (x_{i+1} - x_i) \frac{\partial G}{\partial x_i} + \gamma(1 - x_n)\frac{\partial G}{\partial x_n},$$

where F is the PGF of the burst distribution. Applying the transformations $u_i := x_i - 1$ and

$\phi := \ln G$ yields the equation:

$$\frac{\partial \phi}{\partial t} = k(M(u_1) - 1) + \sum_{i=1}^{n-1} \beta_i (u_{i+1} - u_i) \frac{\partial \phi}{\partial u_i} - \gamma u_n \frac{\partial \phi}{\partial u_n},$$

where $M(u) := F(1+u)$. This equation can be solved using the method of characteristics, with formal solution $\phi = \int_0^t [M(U_1(s)) - 1] ds$. The characteristics U_i, $i < n$ satisfy $\frac{dU_i}{ds} = \beta_i (U_{i+1} - U_i)$, with $U_n(u_1, ..., u_n; s) := U_n(s) = u_n e^{-\gamma s}$.
The functional form of $\frac{dU_i}{ds}$ implies that $U_1(s)$ is the weighted sum of exponentials $\sum_{i=1}^n A_i e^{-\beta_i s}$; the explicit analytical solution is provided in Section 8. The weights A_i can be computed through a simple iterative procedure, which proceeds from the terminal species and successively incorporates dependence on upstream rates:

$A_i \leftarrow 0 \; \forall \; i < n$
$A_n \leftarrow u_n$
$i \leftarrow n - 1$
while $i > 0$ **do**
  for $j > i$ **do**
    $A_j \leftarrow A_j \times \frac{\beta_i}{\beta_i - \beta_j}$
  end for
  $A_i \leftarrow u_i - \sum_{j>i} A_j$
  $i \leftarrow i - 1$
end while

Finally, given the physiologically plausible geometric burst distribution [12], $\phi(t; \cdot) = k \int_0^t \frac{bU_1(s)}{1 - bU_1(s)} ds$. The stationary PGF is e^ϕ evaluated at $t = \infty$; this quantity can be converted to a probability mass function via Fourier inversion. This is an entirely general procedure that can be used for *any* path graph of downstream processing, although regions of stiffness are to be expected, particularly near the degenerate cases of matching rates. In the most general case, any β_i may, in principle, be equal to any other, leading to a combinatorial explosion of auxiliary degenerate solutions. As the overarching motivation is the statistical inference of biophysical parameters, and the degenerate regions are all of measure zero, there appears to be little physical justification for considering them in further detail. However, we remark that these solution classes are likewise analytically tractable, with mixed polynomial-exponential functional forms entirely analogous to previous work [10].

3 Directed acyclic graph splicing

The simplicity of the ODEs governing the evolution of the CME lends itself to extensions to the broadest class of graphs representing the stochastic and incremental removal of discrete introns, the directed acyclic graphs.

3.1 Example: tree graphs

Suppose the downstream dynamics are given by a directed rooted tree. The solution procedure is entirely analogous to that of a path graph. First, starting at the leaves, the path subgraph solutions are produced by the procedure above, yielding a sum of exponentials. Then, at a node of degree

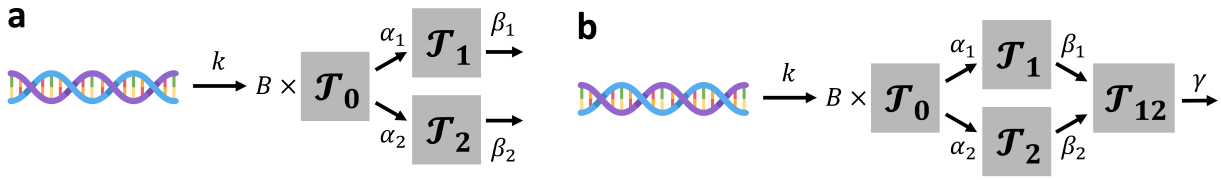

Figure 2: Graph representations of simple directed acyclic graph models. (a) Tree splicing graph with two terminal isoforms. (b) Convergent splicing graph with a single terminal isoform and two intermediate transcripts.

> 2 (i.e., molecular species with several potential products), the associated ODE has a functional form identical to that of a path graph. Therefore, the solutions are analogous.
As an illustration, consider the simplest tree graph, shown in Figure 2a, where the splicing reactions occur at rates α_1 and α_2 and degradation reactions occur at rates β_1 and β_2. Physically, this graph can be interpreted as a single source mRNA being directly and stochastically converted to one of two terminal isoforms by removal of intron 1 or intron 2. Clearly, $U_i = u_i e^{-\beta_i s}$ for $i \in \{1, 2\}$.
The ODE governing the source species is $\frac{dU_0}{ds} = \alpha_1(U_1 - U_0) + \alpha_2(U_2 - U_0) = \alpha_1 U_1 + \alpha_2 U_2 - (\alpha_1 + \alpha_2)U_0$. The solution is $U_0 = \frac{\alpha_1}{\alpha_1+\alpha_2-\beta_1} u_1 e^{-\beta_1 s} + \frac{\alpha_2}{\alpha_1+\alpha_2-\beta_2} u_2 e^{-\beta_2 s} + K e^{-(\beta_1+\beta_2)s}$, with $K = u_0 - \frac{\alpha_1}{\alpha_1+\alpha_2-\beta_1} u_1 - \frac{\alpha_2}{\alpha_1+\alpha_2-\beta_2}$. Finally, the expression for U_0 can be directly plugged into the desired burst distribution generating function.
The distributional properties are equivalent to those derived for the path graph solutions. This generalization follows immediately from considering a gene locus with constitutive dynamics coupled to a directed rooted tree.

3.2 Example: two-intron mRNA

Consider the same tree graph as in the example above, and suppose $\mathcal{T}_1$ and $\mathcal{T}_2$ are converted to product $\mathcal{T}_{12}$ at rates β_1 and β_2, as shown in Figure 2b. Afterward, $\mathcal{T}_{12}$ is degraded at rate γ. Physically, this graph can be interpreted as a single source mRNA being converted to one of two intermediate isoforms by the removal of one of two introns, then to a single terminal isoform by the removal of the other intron. Clearly, $U_{12} = u_{12} e^{-\gamma s}$. Further, setting $f_i := \frac{\beta_i}{\beta_i - \gamma}$, $U_i = (u_i - f_i u_{12}) e^{-\beta_i s} + f_i u_{12} e^{-\gamma s}$. Finally, the dynamics of the source molecule S are governed by the following ODE:

$$\frac{dU_0}{ds} = \alpha_1(u_1 - f_1 u_{12}) e^{-\beta_1 s} + \alpha_2(u_2 - f_2 u_{12}) e^{-\beta_2 s} + (\alpha_1 f_1 + \alpha_2 f_2) u_{12} e^{-\gamma s} - (\alpha_1 + \alpha_2) U_0$$

Yet again, the functional form affords a straightforward analytical solution:

$$U_0 = K e^{-cs} + \frac{C_1}{c - \beta_1} e^{-\beta_1 s} + \frac{C_2}{c - \beta_2} e^{-\beta_2 s} + \frac{C_3}{c - \gamma} e^{-\gamma s},$$

where $c := \alpha_1 + \alpha_2$, $C_1 := \alpha_1(u_1 - f_1 u_{12})$, $C_2 := \alpha_2(u_2 - f_2 u_{12})$, and $C_3 := \alpha_1 f_1 + \alpha_2 f_2$. From the initial condition $U_0(s = 0) = u_0$, we yield $K = u_0 - \frac{C_1}{c-\beta_1} - \frac{C_2}{c-\beta_2} - \frac{C_3}{c-\gamma}$. The details of the ODE solution are described in Section 8 and the computation procedure is demonstrated in Figure 3.

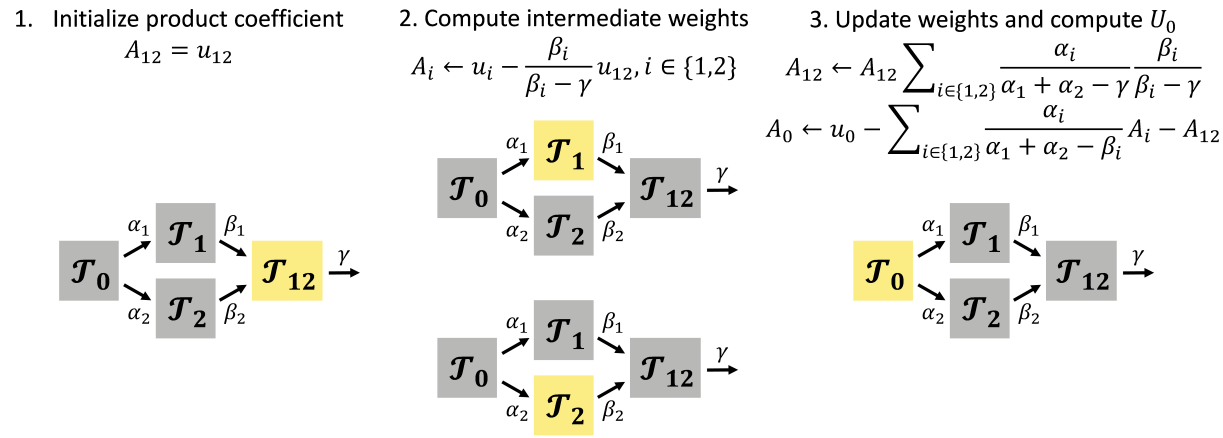

Figure 3: Illustration of the solution algorithm. The differential equation structure requires the backward propagation of downstream species' solutions, weighted by ratios of rates.

3.3 Algorithm

This procedure is facile to extend to an arbitrary directed acyclic graph with a unique root. The reaction system is fully characterized by two arrays, the stoichiometric matrix matrix S and the rate vector r used in the stochastic simulation algorithm [13]. Thus, for example, the path graph can be represented by the full rate vector $[k, \beta_1, ..., \beta_{n-1}, \gamma]$ and the following stoichiometric matrix:

$$\begin{bmatrix} B & 0 & 0 & ... & 0 & 0 \\ -1 & 1 & 0 & ... & 0 & 0 \\ 0 & -1 & 1 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & -1 & 1 \\ 0 & 0 & 0 & ... & 0 & -1 \end{bmatrix}$$

A well-defined system has n species and $q + 1$ reactions. We assume unbounded accumulation does not occur and the graph is connected, so each species has influx and efflux reaction pathways, implying $q \geq n$. Further, we suppose that each species is associated with at most one degradation reaction; if several channels are available, they can be added by the superposition property of a Poisson process. Finally, we assume that all isomerization and degradation rates are distinct. Since the underlying graph is acyclic, there exists at least one terminal species with a single efflux reaction. Thus, consider a single production reaction with rate k coupled to a set of isomerization reactions with rates c_{ji} and degradation reactions with rates c_{j0}.

The downstream dynamics are determined by U_1, an exponential sum with n terms. In this generic case, *lumped* rate exponents r_i must be computed as the net efflux rate from each molecule. A simple implementation of the routine is outlined below:

$r_i \leftarrow \sum_j c_{ij} \; \forall \; i$ ▷ Compute all exponents
$A_{ij} \leftarrow 0 \; \forall \; i, j \in [1, n]$ ▷ Initialize coefficient array
$R_D \leftarrow \{k \mid \sum_i S_{ki} = -1\}$ ▷ Identify reaction channels that correspond to degradation

```
D ← {i | S_ki = −1 for k ∈ R_D}                          ▷ Identify degraded species
T ← {j ∈ D | ∄ (S_kj = −1 ∩ S_ki = 1)}                   ▷ Identify terminal species
R_T ← {k | S_ki = −1 for i ∈ T}                          ▷ Identify terminal reactions
A_ii ← u_i ∀ i ∈ T                                       ▷ Assign terminal degradation coefficients
C ← {∅}                                                  ▷ Initialize set of computed species
while |C| < n do
    for j ∉ C do                                         ▷ Iterate over all species that have not yet been computed
        R_I = {k | (S_kj = −1 ∩ S_ki = 1)}               ▷ Identify isomerization reactions originating at j
        P = {i | (S_ki = 1 for k ∈ R_I)}                 ▷ Identify corresponding isomerization products
        if P ⊂ C then                                    ▷ If all products have been computed
            for i ∈ P do                                 ▷ For each isomerization product
                A_jk ← A_jk + A_ik × c_ji / (r_j − r_i)  ▷ Propagate results from downstream solutions
            end for
            A_jj ← u_j − Σ_k A_jk                        ▷ Apply initial condition, compute coefficient for e^{−r_j s}
            C ← C ∪ j                                    ▷ Adjust set of computed species
        end if
    end for
end while
```

The terminal exponential sum is given by $U_1 = \sum_{i=1}^{n} A_{1i} e^{-r_i s}$. As before, the full time-dependent solution for any burst distribution with a well-defined moment-generating function is computable by quadrature. The case of the geometric burst distribution yet again corresponds to $\phi(t; u_1, ..., u_n) = k \int_0^t \frac{bU_1(s)}{1-bU_1(s)} ds$.

4 Continuous formulation

We have defined a series of discrete joint distributions induced by a graph governing splicing and degradation. However, we can equivalently recast the problem in terms of the stochastic differential equations governing the distributions' Poisson intensities Λ_i. Specifically, the following identity can be used to relate a set of stochastic processes $\Lambda_1, ..., \Lambda_n$ with joint distribution $F_{\Lambda_1,...,\Lambda_n}$ to the solution of the CME [9, 14–16]:

$$P(m_1, ..., m_n, t) = \int \prod_{i=1}^{n} \frac{e^{-\Lambda_i} \Lambda_i^{m_i}}{m_i!} dF_{\Lambda_1,...,\Lambda_n}$$

As an illustration, we can consider the intensity of an n-step isomerization process driven by a finite-activity compound Poisson Lévy subordinator L_t:

$$\begin{aligned} d\Lambda_1 &= -\beta_1 \Lambda_1 dt + dL_t \\ d\Lambda_2 &= -\beta_2 \Lambda_2 dt + \beta_1 \Lambda_1 dt \\ &\dots \\ d\Lambda_n &= -\gamma \Lambda_n dt + \beta_{n-1} \Lambda_n dt, \end{aligned}$$

This formulation has an intimate connection with the theory of moving average processes, which can be immediately seen by applying variation of parameters to the Poisson representation:

$$
\begin{aligned}
\Lambda_i^p &= Ce^{-\beta_i t} \\
\Lambda_i &= C_t e^{-\beta_i t} \\
d\Lambda_i &= -\beta_i C_t e^{-\beta_i t} dt + e^{-\beta_i t} dC_t \\
&= -\beta_i C_t e^{-\beta_i t} dt + \beta_i \Lambda_{i-1} dt \\
C_t &= \int_0^t \beta_i \Lambda_{i-1}(s) e^{\beta_i s} ds \\
\Lambda_i(t) = e^{-\beta_i t} C_t &= \int_0^t \beta_i \Lambda_{i-1}(s) e^{-\beta_i (t-s)} ds
\end{aligned}
$$

By convention, $\beta_n := \gamma$ and $\Lambda_0 ds := dL_s$, the underlying driving subordinator.
Thus, each Λ_i is the exponentially-weighted moving average of Λ_{i-1}; Λ_1 is the moving average of the Poisson shot noise introduced by the compound Poisson subordinator. Although ample literature exists on the topic of moving average processes [17], it generally considers the problem of inference and prediction from discrete-time observations. Furthermore, discussions in the context of the chemical systems only tend to consider the first-order moving average of shot noise [9].
This formulation implies that any n-step isomerization process is identical to an $(n-i)$-step isomerization process driven by an order i iterated moving average process. This identity affords a route for the analytical computation of hybrid continuous-discrete system solutions merely by marginalizing the first i species.
The Poisson representation enables the simultaneous discussion of the properties of $F_{m_1,...,m_n}$, the stationary distribution of the continuous-time Markov chain, and $F_{\Lambda_1,...,\Lambda_n}$, the stationary distribution of the underlying series of stochastic differential equations. In fact, from standard properties of Poisson mixtures [18], the PGF $G(x_1, ..., x_n)$ of the former evaluated at $u_i = x_i - 1$, precisely yields the moment-generating function (MGF) of the latter.
This connection also allows the evaluation of the solution for a broad class of compound Poisson driving processes. Specifically, $M(u) = \frac{1}{1-bu}$ is the MGF associated with exponentially distributed jumps. In the discrete domain, this translates to the geometric burst distribution. However, the downstream dynamics, represented by $U_1(s)$, are independent of the burst distribution. Therefore, the procedure affords computable solutions for any finite-activity pure-jump Lévy driving process. Extensions to generic directed acyclic graphs are entirely analogous. These solutions, identical to the CME case, can be used to exactly solve arbitrary DAGs representing *continuous*-valued molecular concentrations. This representation is conventional for high-abundance species [19].

5 Distribution properties

Using the algorithms above, we can compute the generating functions corresponding to the stochastic processes of interest. However, it is not yet clear that these generating functions are everywhere well-defined. For example, certain physiologically plausible noise models do not have all moments [20]; their generating functions fail to converge in certain regimes. By analyzing the functional form of the downstream process and assuming geometric-distributed bursts, we demonstrate that this class of processes is guaranteed to yield convergent generating functions and finite moments.

5.1 Positivity of the exponential sum

First, we demonstrate that the downstream processes yield a strictly positive functional form of time dependence, or $\psi_i(s) > 0$ for all $s > 0$. Consider $F(x) = x$, corresponding to constitutive production of the source species (i.e. a Poisson birth process), with no molecules present at $t = 0$. Focusing on the marginal of species i, this assumption yields $\phi(u_i; t) = k \int_0^t U_1(u_i; s)ds = k \int_0^t u_n \psi_n(s)ds$. Evaluating e^ϕ at $x_i = 0$, i.e. $u_i = -1$, marginalizes over all $j \neq i$ and yields the probability of observing zero counts of species i: $G(1 - \delta_{ij}; t) = P(m_i = 0, t) = P_0(t) = \exp(-k \int_0^t \psi_i(s)ds)$. The corresponding time derivative is $\frac{dP_0}{dt} = -k\psi_n(t) \exp(-k \int_0^t \psi_i(s)ds)$. Simultaneously, we know that $P_0(t) = e^{-\lambda_i(t)}$, where $\lambda_i(t)$ is the solution of the reaction rate equation for species i [14]. Clearly, $\frac{dP_0}{dt} = -P_0(t)\frac{d\lambda_i(t)}{dt}$. The reaction rate $\frac{d\lambda_i(t)}{dt} > 0$ at $t = 0$ under the given initial conditions.
Furthermore, $\frac{d\lambda_i(t)}{dt}$ is strictly positive. This follows from the reaction rate equations. By the continuous formulation, λ_i is a weighted moving average of some set of processes $\{\lambda_k\}$. λ_1 is a strictly increasing function governed by $\frac{k}{r_1}(1 - e^{-r_1 s})$. The property of being strictly increasing is retained under moving average and rescaling. Therefore, each successive moving average must be strictly increasing.
Finally, $P_0 \in (0, 1) > 0$, because the solution for m_i is given by a Poisson distribution, which has support on $\mathbb{N}$. Therefore, $\frac{dP_0}{dt}$ is strictly negative. As the exponential term and k are positive, this implies $\psi_i(s)$ is strictly positive for all $s > 0$.

5.2 Existence of generating functions and moments

Next, we show that $G(1 - \delta_{ij}; t)$, the generating function of the ith marginal, is finite for the geometric burst system. This follows from the construction of the original PGF: the marginal PGF is guaranteed to be finite if $1 - bu_i\psi_i$ is never zero. But for the relevant domain $\Re(u_n) \leq 0$, on the shifted complex unit circle, $\Re(1 - bu_i\psi_i) \geq 1$, except at the degenerate initial case. The existence of the marginal moments of m_i is implied by the existence of the generating function. The existence of all cross moments follows from the Cauchy-Schwartz inequality. Per standard properties, this existence property holds for both m_i and Λ_i.
The tails of the stationary discrete marginals decay no slower than the geometric distribution. This follows immediately from the lower bound on $\Re(1 - bu_i\psi_i)$, which in turn gives an upper bound on x_i [10]. Equivalently, this follows from the existence of all moments [20]. An analytical radius of convergence has been given previously for $n = 2$ [10], but numerical optimization is necessary to establish rates of tail decay for $n > 2$.

5.3 Statistical properties

All marginals are infinitely divisible. This follows from the functional form of the PGF: the random variable corresponding to any marginal distribution can be written as as a sum of q random variables with burst frequency k/q.
Only the first marginal is self-decomposable. This follows from the condition that a random variable has a self-decomposable (sd) law if and only if it offers a representation of the form $Y = \int_0^\infty e^{-t} dX_t$, with Lévy X_t [21]. However, only Λ_0 is Lévy. All downstream intensity processes have nontrivial, almost-everywhere C^∞ trajectories, which implies they cannot be represented by a Lévy triplet: the only permitted continuous Lévy processes are linear combinations of the (non-differentiable) Brownian motion W_t and the trivial process t. Therefore, Λ_i is sd for $i = 1$ and non-sd for all $i > 1$.

All marginals are unimodal. Multimodality in the distribution of the moving average over time is contingent on time-inhomogeneity in the trajectory process. However, the underlying driving process is defined to be time-homogeneous, with uniformly distributed jump times; therefore, each downstream process has a unimodal distribution over time. By ergodicity, this distribution is equivalent to the ensemble distribution. Therefore, all marginals of downstream species are unimodal.

5.4 Moments

The moments of the marginals can be computed directly from the derivatives of the marginal MGF $e^{\phi(u_i)}$ at $u_i = 0$:

$$\frac{de^{\phi}}{du_i} = e^{\phi}\frac{d\phi}{du_i} = e^{\phi}\frac{d}{du_i}k\int_0^{\infty}\frac{bu_i\psi_i(s)}{1-bu_i\psi_i(s)}ds$$
$$= kb\int_0^{\infty}\psi_i(s)ds = kb\int_0^{\infty}\sum_{i=1}^{n}a_ie^{-r_is}ds = kb\sum_{i=1}^{n}\frac{a_i}{r_i}.$$

For the path graph system, the CME immediately yields $\mu_i = \frac{kb}{r_i}$. Considering the previously discussed constitutive solution $\phi^0(u_i) = ku_i\int_0^{\infty}\psi_i(s)ds$, we yield the identity $\int_0^{\infty}\psi_i(s)ds = \sum_{i=1}^{n}\frac{a_i}{r_i}$, which is equal to $\frac{1}{r_i}$ for the path graph system. Per the standard properties of mixed Poisson distributions [22], the value of μ_i is identical for the underlying continuous process and the derived discrete process.
The stationary second moments can be found analogously:

$$\frac{d^2e^{\phi}}{du_i^2} = e^{\phi}\frac{d^2\phi}{du_i^2} + e^{\phi}\left(\frac{d\phi}{du_i}\right)^2$$
$$= 2kb^2\int_0^{\infty}\psi_i^2(s)ds + \left(\frac{kb}{r_i}\right)^2$$
$$\psi_n^2(s) = \sum_{j,k=1}^{n}a_ja_ke^{-(r_k+r_k)s} = \sum_{j=1}^{n}a_j^2e^{-2r_js} + \sum_{j,k=1,\ j\neq k}^{n}a_ja_ke^{-(r_j+r_k)s}$$
$$\frac{d^2e^{\phi}}{du_i^2} = 2kb^2\sum_{j,k=1}^{n}\frac{a_ja_k}{r_j+r_j} + \mu_i^2 = \mathbb{E}[\Lambda_i^2]$$
$$\mathbb{V}[\Lambda_i] = 2kb^2\sum_{j,k=1}^{n}\frac{a_ja_k}{r_j+r_k},$$

which is straightforward to compute, but does not appear to have an easily amenable analytical form beyond the first few cases. Naturally, the standard properties of Poisson mixtures [22] allow conversion to the discrete domain, with $\mathbb{V}[m_i] = \mathbb{V}[\Lambda_i] + \mu_i$.
The covariances can be computed directly from the derivatives of the MGF $e^{\phi(u_l,u_i)}$, i.e., the marginalized MGF for $u_q = 0$ for all $q \neq l, i$. The particular functional form of $\phi(u_l, u_i)$ is given by $k\int_0^{\infty}\frac{bU_1(u_i,u_l;s)}{1-bU_1(u_i,u_l;s)}$. By construction, $U_1(u_l, u_i; s) = u_l\psi_l(s)+u_i\psi_i(s)$, where each ψ is the exponential sum corresponding to the marginal of the species in question. This yields:

$$\frac{d^2 e^{\phi}}{du_l du_i} = \frac{d}{du_l} e^{\phi} k \int_0^{\infty} \frac{b\psi_i ds}{(1 - bu_l\psi_l - bu_i\psi_i)^2}$$

$$= e^{\phi} k^2 \int_0^{\infty} \frac{b\psi_l ds}{(1 - bu_l\psi_l - bu_i\psi_i)^2} \int_0^{\infty} \frac{b\psi_i ds}{(1 - bu_l\psi_l - bu_i\psi_i)^2} + 2e^{\phi} k \int_0^{\infty} \frac{b^2\psi_l\psi_i ds}{(1 - bu_l\psi_l - bu_i\psi_i)^3}$$

$$= k^2 b^2 \int_0^{\infty} \psi_l ds \int_0^{\infty} \psi_i ds + 2kb^2 \int_0^{\infty} \psi_l\psi_i ds$$

$$= \mu_l\mu_i + 2kb^2 \int_0^{\infty} \psi_l\psi_i ds = \mathbb{E}[\Lambda_l\Lambda_i] = \mathrm{Cov}(\Lambda_l, \Lambda_i) + \mu_l\mu_i,$$

which implies that

$$\mathrm{Cov}(\Lambda_l, \Lambda_i) = 2kb^2 \int_0^{\infty} \psi_l\psi_i ds,$$

As above,

$$\psi_l(s)\psi_i(s) = \sum_{j,k=1}^{n} a_j c_k e^{-(r_j+r_k)s}$$

$$\int_0^{\infty} \psi_l(s)\psi_i(s) ds = \sum_{j,k=1}^{n} \frac{a_j c_k}{r_j + r_k}$$

$$\mathrm{Cov}(\Lambda_l, \Lambda_i) = 2kb^2 \sum_{j,k=1}^{n} \frac{a_j c_k}{r_j + r_k} = 2kb^2 \sum_{j,k=1}^{n} \frac{a_j c_k}{r_j + r_k},$$

where a_j are the weights associated with ψ_l and c_k are the weights associated with ψ_i. This form of the summation is very general; for example, in case of the path graph system with $l > i$, it can be equivalently represented as $2kb^2 \sum_{j=1}^{l} \sum_{k=1}^{i} \frac{a_j c_k}{r_j + r_k}$. From standard identities, the covariance of a mixed bivariate Poisson distribution with no intrinsic covariance forcing is identical to the covariance of the mixing distribution [22]. Therefore, this result holds for both the CME and the underlying SDE.

The Pearson correlation coefficient follows immediately from the result above:

$$\rho = \frac{\sum_{j,k=1}^{n} \frac{a_j c_k}{r_j + r_k}}{\sqrt{\sum_{j,k=1}^{n} \frac{a_j a_k}{r_j + r_k} \times \sum_{j,k=1}^{n} \frac{c_j c_k}{r_j + r_k}}}$$

Since mixing decreases variance but not covariance, the correlation coefficient of the discrete system will always be lower than that of its continuous or hybrid analog.

6 Validation

We generated a random directed acyclic graph, shown in Figure 4. The numbers of species (7) and isomerization reactions (11) were chosen arbitrarily. We enforced the existence of a single unique source node (a) and the weakly connected property to ensure only a single source mRNA would be

present and all isoforms would be reachable from it, but did not impose any other conditions. The number of degraded species (3) was chosen arbitrarily; we assigned degradation reactions to the two sink species (c, e) and randomly chose a degraded intermediate (b) from a uniform distribution over the molecular species.

All reaction rates were drawn from a log-uniform distribution on $[10^{-0.5}, 10^{0.5}]$; we chose to sample them from a single order of magnitude to avoid the trivial degenerate cases that occur in cases of very slow or very fast export [10]. This process produced the parameter values $k = 0.44$, $\beta = [0.48, 2.12, 1.31, 2.21, 1.16, 2.41, 0.4, 1.19, 0.37, 1.19, 0.53]$, and $\gamma = [0.94, 2.38, 0.72]$, with the indices corresponding to those in Figure 4. Finally, we chose the geometric burst model with $b = 10$.

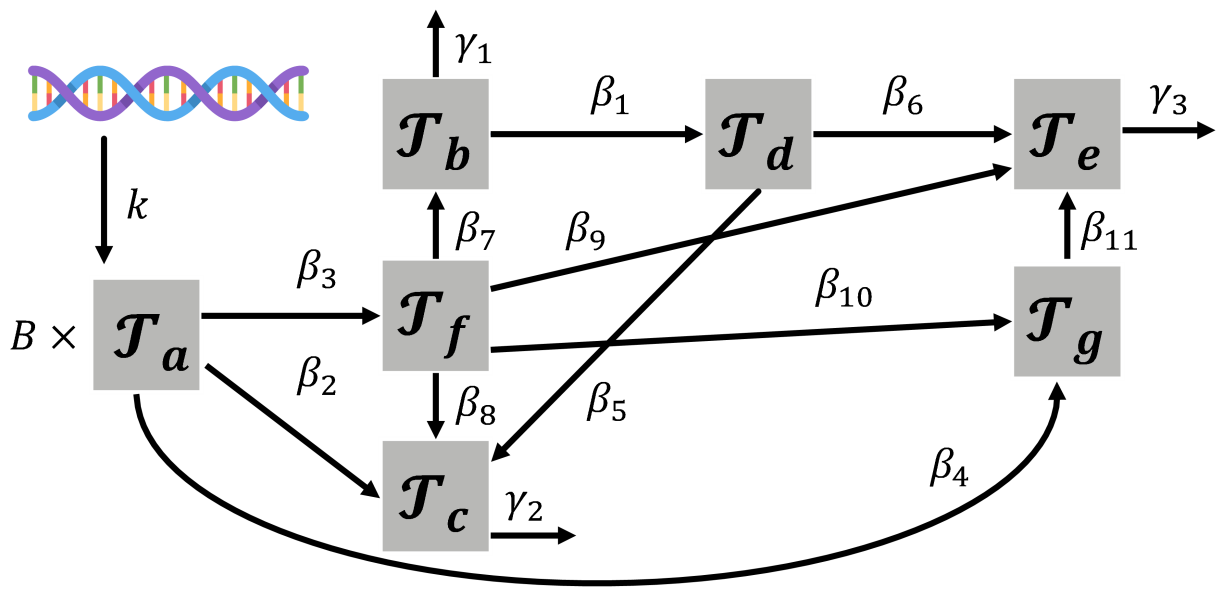

Figure 4: Graph representation of the randomly generated transcription, splicing, and degradation model. A single source isoform $\mathcal{T}_a$ is converted to a variety of downstream isoforms $\mathcal{T}_b, ..., \mathcal{T}_g$, which isomerize according to a directed acyclic graph.

We applied the algorithm to compute the exponents and coefficients, and computed the stationary distributions of all species. The simulated distributions match the quantitative results for the marginals, as shown in Figure 5. Furthermore, the 49 entries of the covariance matrix are likewise effectively predicted by the procedure for moment calculation.

7 Discussion

We have described a broad extension of previous work concerning the solutions of monomolecular reaction networks coupled to a gene locus. In particular, by exploiting the standard properties of reaction rate equations, we have demonstrated that all moments and cross-moments exist for each species in a generic directed acyclic graph. Further, we have derived the analytical expressions for the generating functions and demonstrated their existence.

Although we primarily focus on distributions, with an eye to inference from atemporal data, the

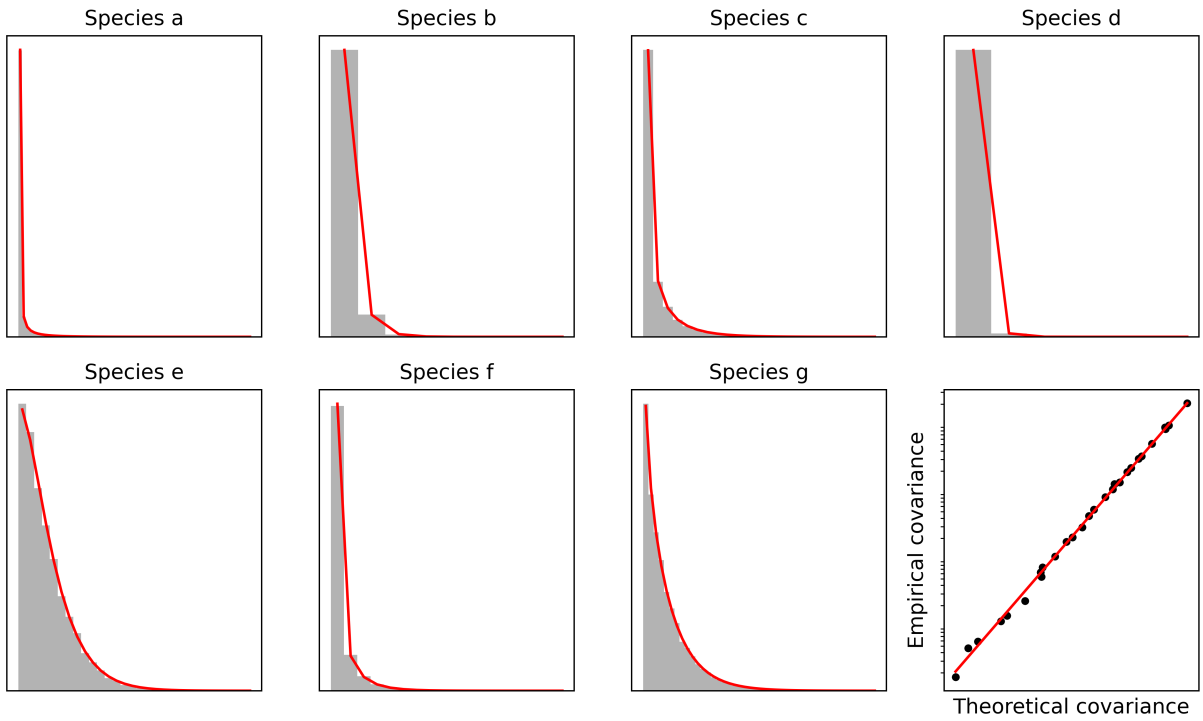

Figure 5: Validation of the randomly generated graph model. Histograms: empirical results (10,000 simulations). Black dots: theoretical and empirical covariance. Red lines: theory.

solution is robust in several complementary dimensions. Firstly, the upper limit of the integral $k\int_0^t[M(U_1(s))-1]ds$ is arbitrary, and can be evaluated up to a finite time horizon to yield a transient distribution. Secondly, the formulation of the problem presupposes $m_i(t=0)=0$ for all i. However, if nonzero molecule counts are present at $t=0$, it is straightforward to compute the resulting log-PGF via $\phi=\phi^h+\sum_{i=1}^n m_i(t=0)\ln(1+U_i)$, where ϕ^h is the result for $m_i(t=0)=0$. This relation produces arbitrary *conditional* distributions, as derived elsewhere [23]. Therefore, the current approach can be used to compute the likelihoods of entire trajectories of observations, and thus perform parameter estimation on live-cell data.

The computational complexity of this procedure is $O(\mathcal{N}\log\mathcal{N})$ in state space size. However, this complexity ostensibly arises from the n-dimensional inverse Fourier transform. We expect the time complexity to be $O(\mathcal{N})$ in the practical regime, and irrelevant outside due to the *space* complexity of holding an array of size $\mathcal{N}$ in memory for the inverse Fourier transform.

Considering the hypothetical case of fitting multidimensional data associated with a single gene across thousands of cells – e.g., several terminal isoforms and tens of intermediate species – these memory constraints may motivate the application of generating function metrics instead of likelihood-based inference. Given a relatively sparse n-dimensional histogram, it is straightforward to compute the associated dense Fourier transform incrementally, storing directly to disk. Afterward, it is possible to parallelize over small regions of the Fourier space, compute a distance between a proposal and the empirical generating function, and optimize to estimate parameters.

Curiously, this class of analytical solutions to reaction networks can be adapted to a subset of

diffusion problems. General diffusion on a multidimensional lattice is not directly solvable, because it violates the assumption of acyclic graph structure. However, percolation through a directed acyclic graph coupled to a source and a set of sinks can be described using the same mathematical formalism. Further, such a percolation can represent the incremental movement of RNA polymerase along a DNA strand, integrating discrete copy number statistics with submolecular details in a single analytical framework [24, 25].

Finally, we do not treat the aforementioned auxiliary degenerate solutions that arise when $\beta_i = \beta_j$ for some i, j. However, Section 5 guarantees the existence of these solutions and all moments: the procedure does not rely on a particular functional form of ψ_i, only its existence, which is implied by the test subordinator $F(x) = x$ matching the constitutive case [14].

8 Appendix: ODE solution

The solution to the generic ODE $\frac{dy}{dx} = \sum_i a_i e^{-b_i x} - ry$ is $y = \sum_i \frac{a_i}{r-b_i} e^{-b_i x} + Ke^{-rx}$. Using the initial condition $y(0) = \xi$ yields $K = \xi - \sum_i \frac{a_i}{r-b_i} a_i$. Consider an efflux rate $r = \sum_j \beta_j + \gamma$, i.e., a set of isomerization pathways and a single degradation pathways from the species in question. Care must be taken when the downstream paths converge (as in the case of the two-intron system): duplicate terms in product characteristics U_j need to be aggregated, with $\sum_j \beta_j U_j$ rewritten as $\sum_i a_i e^{-b_i x}$. This is computationally straightforward to do by choosing appropriate data structures, as with the matrix A in the implementation of the DAG algorithm.

9 Code availability

Python code that can be used to reproduce Figure 5, including the graph generation, simulation, and plotting routines, is available at `https://github.com/pachterlab/GP_2021_2`.

10 Acknowledgments

The DNA illustration used in Figures 1 and 4, modified from [23], is a derivative of the DNA Twemoji by Twitter, Inc., used under CC-BY 4.0. The directed acyclic graph generation code was adapted from the IPython Parallel reference documentation: `https://ipyparallel.readthedocs.io/en/latest/dag_dependencies.html`. The yellow and gray colors used in Figures 1, 3, and 4 are the Pantone Colors of the Year 2021, PANTONE 17-5104 Ultimate Gray and PANTONE 13-0647 Illuminating. G.G. and L.P. were partially funded by NIH U19MH114830.

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] 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.

[3] Alex K. Shalek, Rahul Satija, Xian Adiconis, Rona S. Gertner, Jellert T. Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, John J. Trombetta, Dave Gennert, Andreas Gnirke, Alon Goren, Nir Hacohen, Joshua Z. Levin, Hongkun Park, and Aviv Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, June 2013.

[4] Carlos F Buen Abad Najar, Nir Yosef, and Liana F Lareau. Coverage-dependent bias creates the appearance of binary splicing in single cells. *eLife*, 9:e54603, June 2020.

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

[6] Brian Munsky, Zachary Fox, and Gregor Neuert. Integrating single-molecule experiments and discrete stochastic models to understand heterogeneous gene transcription dynamics. *Methods*, 85:12–21, 2015.

[7] Gennady Gorin and Lior Pachter. Intrinsic and extrinsic noise are distinguishable in a synthesis – export – degradation model of mRNA production. Preprint, bioRxiv: 10.1101/2020.09.25.312868, September 2020.

[8] M. Ullah and O. Wolkenhauer. Family tree of Markov models in systems biology. *IET Systems Biology*, 1(4):247–254, July 2007.

[9] Crispin Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. Springer, 3 edition, 2004.

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

[11] 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.

[12] Ido Golding, Johan Paulsson, Scott M. Zawilski, and Edward C. Cox. Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell*, 123(6):1025–1036, December 2005.

[13] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, December 1976.

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

[15] Lisa Amrhein, Kumar Harsha, and Christiane Fuchs. A mechanistic model for the negative binomial distribution of single-cell mRNA counts. Preprint, bioRxiv: 657619, June 2019.

[16] C. W. Gardiner and S. Chaturvedi. The poisson representation. I. A new technique for chemical master equations. *Journal of Statistical Physics*, 17(6):429–468, December 1977.

[17] Chris Chatfield, Anne B. Koehler, J. K. Ord, and Ralph D. Snyder. A New Look at Models For Exponential Smoothing. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 50(2):147–159, July 2001.

[18] Harry H Panjer. Mixed Poisson Distributions. In *Encyclopedia of Actuarial Science*. John Wiley & Sons, Ltd, 2004.

[19] Nir Friedman, Long Cai, and X. Sunney Xie. Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. *Physical Review Letters*, 97(16):168302, October 2006.

[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] Ole. E. Barndorff-Nielsen and Steen Thorbjørnsen. Self-Decomposability and Lévy Processes in Free Probability. *Bernoulli*, 8(3):323–366, 2002. Publisher: International Statistical Institute (ISI) and Bernoulli Society for Mathematical Statistics and Probability.

[22] Dimitris Karlis and Evdokia Xekalaki. Mixed Poisson Distributions. *International Statistical Review / Revue Internationale de Statistique*, 73(1):35–58, 2005. Publisher: [Wiley, International Statistical Institute (ISI)].

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

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

[25] Gennady Gorin, Mengyu Wang, Ido Golding, and Heng Xu. Stochastic simulation and statistical inference platform for visualization and estimation of transcriptional kinetics. *PLOS ONE*, 15(3):e0230736, March 2020.