Coordination of gene expression noise with cell size: extrinsic noise versus agent-based models of growing cell populations

Philipp Thomas[a)] and Vahid Shahrezaei
Department of Mathematics, Imperial College London, UK

The chemical master equation and the stochastic simulation algorithm are widely used to model the reaction kinetics inside living cells. It is thereby assumed that cell growth and division can be modelled for through effective dilution reactions and extrinsic noise sources. We here re-examine these paradigms through developing an analytical agent-based framework of growing and dividing cells accompanied by an exact simulation algorithm, which allows us to quantify the dynamics of virtually any intracellular reaction network affected by stochastic cell size control and division noise in a growing population. We find that the solution of the chemical master equation – including static extrinsic noise – exactly agrees with the one of the agent-based formulation when a simple condition on the network's topology is met. We illustrate this result for a range of common gene expression networks. When these conditions are not met, we demonstrate using analytical solutions of the agent-based models that the dependence of gene expression noise on cell size can qualitatively deviate from the effective master equation. Surprisingly, the latter distorts total noise in gene regulatory networks by at most 8% independently of network parameters. Our results highlight the accuracy of extrinsic noise modelling within the chemical master equation framework.

I. INTRODUCTION

Cells must continuously synthesise molecules in order to grow and divide. At a single cell level, gene expression and cell size are coordinated but highly heterogeneous which can drive of phenotypic variability and decision making in cell populations[1–5]. The interplay between these sources of cell-to-cell variability is however still not well understood since they have traditionally been studied separately. A general stochastic theory integrating size-dependent biochemical reactions with the dynamics of growing and dividing cells is hence still missing.

Many models of noisy gene expression and its regulation are based on the chemical master equation that describes the stochastic dynamics of biochemical reactions in a fixed reaction volume[6–8]. The small scale of compartmental sizes of cells implies that at any time only a small number of molecules is present leading to large variability of reaction rates from cell to cell, commonly referred to as gene expression noise[9–11]. Another factor contributing to gene expression noise is the fact that cells are continuously growing and dividing causing molecule numbers to (approximately) double over the course of a growth-division cycle. A common approach to account for these additional noise sources is to include extra degradation reactions that describe dilution of gene expression levels due to cell growth[9–13] akin to what is done in deterministic rate equation models[14,15]. We will refer to this approach as the *effective dilution model* (EDM, see Fig. 1a). However, little is known of how well this approach represents the dependence of gene expression noise on cell size observed in a growing population.

Cells achieve concentration homeostasis through coupling reaction rates to cell size via highly abundant upstream factors like cell cycle regulators, polymerases or ribosomes that approximately double over the division cycle[3,16,17]. Cell size fluctuates in single cells, however, providing a source of extrinsic noise in reaction rates that can be identified via noise decompositions[18,19]. A few studies combined effective dilution models with static cell size variations as an explanatory source of extrinsic noise[20–22]. In brief, the total noise in these models amounts to intrinsic fluctuations due to gene expression through balancing intracellular reactions and dilution at a certain size, and its extrinsic variation across cell sizes in the population. We refer to this class of models as *extrinsic noise models* (ENMs, see Fig. 1b). Yet it remains unclear how reliably these effective models can account for single cell dynamics whereby cells continuously synthesise molecules, grow and divide.

An increasing number of studies are investing efforts towards quantifying the dependence of gene expression noise on cell cycle progression and growth, either experimentally via ergodic principles or pseudo-time[23,24] and time-lapse imaging[22,25,26] or theoretically through noise decomposition[27–29] or agent-based approaches including age-structure and cell cycle dynamics[4,17,30–40]. The essence of *agent-based models* (ABMs) is that cells are represented by agents whose physiological state is tracked along with their molecular reaction networks. In principle, these models are able to predict gene expression distributions of cells progressing through well-defined cell cycle states as measured by time-lapse microscopy and snapshots of heterogeneous populations. These models

[a)] Electronic mail: p.thomas@imperial.ac.uk

must generally cast doubt on the predictions of master equation models in which growth and division are neglected. But these models presently cannot explain why previous master equation models including effective dilution reactions have fared reasonably well in predicting gene expression noise across cell populations as reported by current single-cell experiments[10,17,41].

Nevertheless, some ABM models still ignore cell size, a major physiological factor affecting both intracellular reactions and cell division dynamics alike. Since cell size varies at least two-fold as required by size homeostasis in a growing population, it is expected that size must significantly contribute to gene expression variation across a population. In this article, we bridge the gap between the chemical master equation and agent-based approaches by integrating cell size dynamics with the stochastic kinetics of molecular reaction networks in an exact manner. We provide rigorous conditions under which the chemical master equation including effective dilution reactions reproduces the dependence of the molecule number distribution on cell size of the agent-based model exactly, thus restoring trust in previous analysis (Sec. III A). Our condition is met by some but not all common models of gene expression. We show that in the later case the effective models merely work on average (Sec. III B). To overcome these limitations, we develop a comprehensive theoretical framework with which to quantify cell size scaling of gene expression in growing cells (Sec. III C). Our findings indicate that the effective dilution model often can qualitatively fail to predict this dependence and we provide analytical approximations that accurately describe gene expression noise even in the presence of cell size control variations and division errors (Sec. III D). In contrast, extrinsic noise models accurately capture noise statistics across population within a few percent of error (Sec. III E).

II. METHODS

We consider a biochemical reaction network of N molecular species $S = (S_1, S_2, \ldots, S_N)^T$ embedded in a cell of size s. The rth reaction then has the general form:

$$\sum_{i=1}^{N} \nu_{ir}^{-} S_i \xrightarrow{k_r} \sum_{i=1}^{N} \nu_{ir}^{+} S_i, \quad r = 1, \ldots, R, \tag{1}$$

where $\nu_r^{\pm} = (\nu_{1r}^{\pm}, \nu_{2r}^{\pm}, \ldots, \nu_{Nr}^{\pm})$ are the stoichiometric coefficients and k_r is the reaction rate constant. In the following, we outline deterministic, effective dilution and extrinsic models and develop a new agent-based approach coupling stochastic reaction dynamics to cell size in growing and dividing cells (Fig. 1).

A. Effective dilution models, extrinsic noise models and the chemical master equation

1. Rate equation models and balanced growth conditions

In balanced growth the vector of molecular concentrations $\bar{X} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_N)^T$ can be obtained from deterministic rate equation models. The balanced growth condition states that there exists a steady state between reaction and dilution rates

$$\alpha \bar{X} = \sum_{r=1}^{R} (\nu_r^{+} - \nu_r^{-}) f_r(\bar{X}). \tag{2}$$

Here, f_r are macroscopic rate function determining the reaction flux and α is the exponential growth rate of cells determining the dilution rate due to growth. A consequence of the balanced growth condition (2) is that the concentrations are independent of cell size, or equivalently that molecule numbers are proportional to size, i.e., $x = s\bar{X}$, ensuring concentration homeostasis.

2. Effective dilution model

The chemical master equation[6] and equivalently the stochastic simulation algorithm[7] are state-of-the-art stochastic models of reactions inside cells. Although well-established, they are strictly valid only when describing cellular fluctuations at constant cell size s. A straight-forward approach to circumvent this limitation is to supplement (1) by additional degradation reactions that model dilution of molecules due to cell growth with rate α:

$$S_i \xrightarrow{\alpha} \varnothing, \quad i = 1, 2, \ldots, N, \tag{3}$$

akin to what is traditionally for reaction rate equations (2). The chemical master equation of this *effective dilution model* (EDM) then takes the familiar form

$$0 = \frac{\partial \Pi_{\mathrm{EDM}}(x|s)}{\partial t} = [\mathbb{Q}(s) + \alpha \mathbb{D}] \Pi_{\mathrm{EDM}}(x|s), \tag{4}$$

governing the conditional probability of molecule numbers $x = (x_1, x_2, \ldots, x_N)^T$ in a cell of size s and where

$$\mathbb{Q}_{x,x'}(s) = \sum_{r=1}^{R} w_r(x', s)(\delta_{x, x' + \nu_r^{+} - \nu_r^{-}} - \delta_{x,x'}), \tag{5}$$

are the elements of the transition matrix of the molecular reactions (1) and we included the extra dilution reactions (3) via $\mathbb{D}_{x,x'}(s) = \sum_{i=1}^{N} x_i'(\delta_{x_i, x_i'+1} - \delta_{x_i, x_i'})$. We are here interested in the stationary solution and hence set the time-derivative in Eq. (4) to zero.

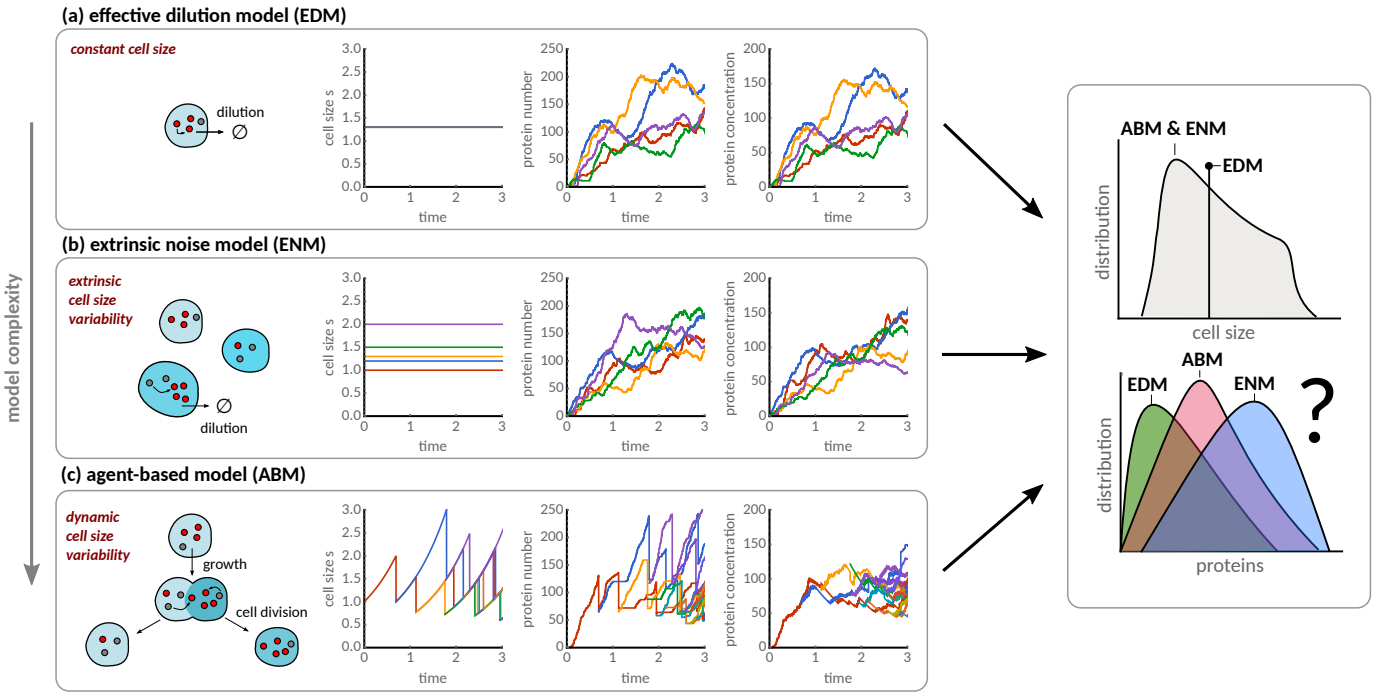

FIG. 1. **Modelling approaches for cell size dependence of gene expression.** **(a)** The *effective dilution model* describes cells at constant size with intracellular reactions coupled to effective dilution reactions. **(b)** The *extrinsic noise model* incorporates static cell size variability as a source of extrinsic noise coupled with effective dilution models **(c)** The *agent-based approach* models intracellular reactions occurring across a growing and dividing cell population without the need for effective dilution reactions.

Such effective models have been motivated by the fact that the mean concentrations follow the rate equations predictions $\bar{X}$, given by the solution of (2), when a large size limit is invoked, a standard result in the analysis of the chemical master equation[6,42]. For example, for the case of mass-action kinetics assuming fast intracellular diffusion[8], the propensities in (5) follow

$$w_r(x,s) = s^{1-|\nu_r^-|} k_r \prod_{i=1}^{N} \frac{x_i!}{(x_i-\nu_{ir}^-)!} \approx s f_r(X), \qquad (6)$$

where $X = x/s$ is the concentration, f_r the macroscopic rate in Eq. (2) and $|\nu_r^-| = \sum_{i=1}^N \nu_{ir}^-$. This correspondence is exact when w_r (and similarly f_r) are linear functions of the molecule numbers but it presents an approximation otherwise[6].

3. Extrinsic noise model

A common way to incorporate static size variability between cells in the model is to consider cell size s to be distributed across cells according to a cell size distribution $\Pi(s)$. We will refer to this approach as the *extrinsic noise model* (ENM), which leads to a mixture model of concentrations $X = x/s$,

$$\Pi_{\mathrm{ENM}}(X) = \int_0^\infty \mathrm{d}s\, \Pi_{\mathrm{EDM}}(x = Xs|s)\Pi(s) \qquad (7)$$

and analogous expressions for the molecule number distributions.

4. Linear Noise Approximation

The advantage of the EDM and ENM is that its noise statistics can be approximated in closed-form using the linear noise approximation[6,43,44]. In the same limit, the covariance matrix can be decomposed into intrinsic and extrinsic components using the law of total variance[18,19]

$$\Sigma_Y = \underbrace{\Sigma_Y^{\mathrm{int}}}_{\text{gene expression}} + \underbrace{\Sigma_Y^{\mathrm{ext}}}_{\text{cell size variation}}, \qquad (8)$$

which correspond to molecular fluctuations due to gene expression and cell size variation, respectively, for $Y \in \{\mathrm{EDM}, \mathrm{ENM}\}$. Specifically, for molecule numbers x, we have $\Sigma_Y^{\mathrm{int}} = E_\Pi[\mathrm{Cov}_Y[x|s]]$ and $\Sigma_Y^{\mathrm{ext}} = \mathrm{Cov}_\Pi[E_Y[x|s]]$;

and analogously for concentrations. The intrinsic components $\Sigma_Y^{\rm int}$ satisfy a Lyapunov equation called the linear noise approximation:

$$0 = \mathcal{J}_d \Sigma_Y^{\rm int} + \Sigma_Y^{\rm int} \mathcal{J}_d^T + \Omega_Y^{-1} \mathcal{D}_d(\bar{X}), \tag{9}$$

where Ω_Y has to be chosen depending on whether concentration or number covariances are of interest

Ω_Y	concentration	numbers
EDM	s	s^{-1}
ENM	$E_\Pi[s^{-1}]^{-1}$	$E_\Pi[s]^{-1}$

. (10)

The matrix $\mathcal{J}_d$ is the Jacobian of rate equations (2) and $\mathcal{D}_d$ denotes the diffusion matrix obeying

$$\mathcal{J}_d(\bar{X}) = \mathcal{J}(\bar{X}) - \alpha \underline{1}, \quad \mathcal{D}_d(\bar{X}) = \mathcal{D}(\bar{X}) + \alpha\, {\rm diag}(\bar{X}), \tag{11}$$

where $\mathcal{J}(\bar{X}) = \sum_{r=1}^R (\nu_r^+ - \nu_r^-) \nabla_{\bar{X}}^T f_r(\bar{X})$ and $\mathcal{D}(\bar{X}) = \sum_{r=1}^R f_r (\nu_r^+ - \nu_r^-)(\nu_r^+ - \nu_r^-)^T$. The extrinsic components follow from the dependence of the mean on cell size, which features only in the molecule number noise of the ENM:

$\Sigma_Y^{\rm ext}$	concentration	numbers
EDM	0	0
ENM	0	${\rm Var}_\Pi(s) \bar{X} \bar{X}^T$

(12)

As a concrete example, we consider transcription of mRNAs with a size-dependent transcription rate which are then translated into (non-degrading) proteins:

$$\varnothing \xrightarrow{k_0 s} M \xrightarrow{k_{\rm dm}} \varnothing, \quad M \xrightarrow{k_s} M + P. \tag{13}$$

We then account for dilution through the additional reactions

$$M \xrightarrow{\alpha} \varnothing, \quad P \xrightarrow{\alpha} \varnothing. \tag{14}$$

The mean protein concentration is given by $\bar{P} = k_0 b/\alpha$, and the coefficient of variation predicted by the EDM and ENM models follow the familiar expression[10]

$$\mathrm{CV}_\mathrm{Y}^2 = \frac{1}{\Omega_Y \bar{P}} \left(1 + b \frac{\gamma}{\gamma + 1}\right) + \frac{\Sigma_Y^{\rm ext}}{\bar{P}^2}, \tag{15}$$

where Ω_Y and $\Sigma_Y^{\rm ext}$ are given by Eqs. (10) and (12), respectively, and the parameters

$$\gamma = 1 + \frac{k_{\rm dm}}{\alpha}, \quad b = \frac{k_s}{k_{\rm dm} + \alpha}, \tag{16}$$

correspond to the ratio of mRNA and protein degradation/dilution rates and the translational burst size, respectively.

B. Agent-based modelling

Little is known about the validity and accuracy of these effective models to predict cellular noise in growing populations. To this end, we introduce an agent-based modelling approach representing cells as agents that progressively synthesise molecules via intracellular reactions (1), grow in size and undergo cell division. Every division gives rise to two daughter cells of varying birth sizes, each of which inherits a proportion of molecules from the mother cell via stochastic size-dependent partitioning at division. An exact simulation algorithm of the ABM is given in Box 1. Our algorithm combines the First-Division algorithm, previously introduced for agent-based cell populations[36], with the Extrande method adapted to simulate reaction networks embedded in a growing cell[45].

In the following, we introduce the theoretical framework with which we characterise the population snapshot distributions analytically. We assume that the population dynamics establishes long-term stationary distribution $\Pi(s, s_0, x)$ characterising the fraction of cells with molecule numbers x, cell size s and birth size s_0 that is invariant in time. We focus on characterising the marginal cell size distribution $\Pi(s, s_0)$ and the molecule number distribution $\Pi(x|s, s_0)$ via Bayes' formula

$$\Pi(s, s_0) = \textstyle\sum_x \Pi(s, s_0, x), \quad \Pi(x|s, s_0) = \dfrac{\Pi(s, s_0, x)}{\Pi(s, s_0)}. \tag{17}$$

Together, they provide the full information about the stationary population process.

Cell size distribution

We assume that cells grow exponentially $s(\tau) = s_0 e^{\alpha \tau}$ and divide with a size-dependent rate, $\alpha s \gamma(s, s_0)$, that depends on the cell growth rate α, the birth size s_0 and cell size s. For example, if the division size is linearly related to birth size

$$s_d = a s_0 + \Delta, \tag{18}$$

the division rate $\gamma(s, s_0) = \gamma(s - a s_0)$ follows from the distribution $\varphi(\Delta)$ of the noise term Δ in (18) via $\gamma(\Delta) = \varphi(\Delta)/(\int_\Delta^\infty {\rm d}u \varphi(u))$. The model generalises the sizer ($a = 0$) to concerted cell size controls such as the adder ($a = 1$) and timer-like ($2 > a > 1$) models[46–50]. More generally, the division rate γ determines the division size distribution $\varphi(s_d|s_0)$ via

$$\varphi(s_d|s_0) = \gamma(s_d, s_0) e^{-\int_{s_0}^{s_d} {\rm d}s \gamma(s, s_0)}. \tag{19}$$

After cell division, size is partitioned between cells and the birth size of the two daughter cells is obtained from

$s_0' = \theta s_d$ and $s_0'' = (1-\theta)s_d$ where θ is the inherited size fraction, a random variable between 0 and 1 with distribution $\pi(\theta)$ (see Box 1).

The general agent-based framework to characterise the size distribution $\Pi(s, s_0)$ has been given by Thomas[51]. They showed that, in the long term, the total number of cells increases exponentially with rate α[52,53] and that the corresponding size distribution satisfies

$$\left(\alpha + \frac{\partial}{\partial s}\alpha s + \alpha s\gamma(s, s_0)\right)\Pi(s, s_0) = 0 \tag{20a}$$

$$s_0\Pi(s_0, s_0) = 2\int_0^\infty \mathrm{d}s' \int_0^{s'} \mathrm{d}s_0'\, B(s_0|s')s'\gamma(s', s_0')\Pi(s, s_0'), \tag{20b}$$

where $B(s_0|s') = \int_0^1 \mathrm{d}\theta\, \pi(\theta)\delta(\theta - s_0/s')$. Eqs. (20) can be solved analytically[51]

$$\Pi(s, s_0) = \frac{2}{Z}\psi_{\mathrm{bw}}(s_0)\Phi(s|s_0)\frac{1}{s^2}, \tag{21}$$

where $\psi_{\mathrm{bw}}(s_0)$ is the birth size distribution in a backward lineage (see[51] for details), $\Phi(s|s_0) = \exp(-\int_{s_0}^{s} \mathrm{d}s'\gamma(s', s_0))$ is the probability that a cell born at size s_0 has not divided before reaching size s, and $Z = E_{\psi_{\mathrm{bw}}}[s_0^{-1}]$ is a normalising constant.

Molecule number distributions conditioned on cell size

The starting point of this study is the stationary distribution $\Pi(s, s_0, x)$ combining gene expression and cell size dynamics, which follows

$$\left(\alpha + \frac{\partial}{\partial s}\alpha s + \alpha s\gamma(s, s_0)\right)\Pi(s, s_0, x) = \mathbb{Q}(s)\Pi(s_0, s_0, x) \tag{22a}$$

$$s_0\Pi(s_0, s_0, x) = 2\sum_{x'}\int_0^\infty \mathrm{d}s' \int_0^{s'} \mathrm{d}s_0'\, B(x|x', s', s_0)B(s_0|s')s'\gamma(s', s_0')\Pi(s', s_0', x'). \tag{22b}$$

The division kernel $B(x|x', s', s_0)$ denotes the probability that a cell born at size s_0 inherits x molecules from a total of x' molecules from its mother, which divided at size s'.

By summing (22) over x, we see that the cell size distribution $\Pi(s, s_0)$ follows Eq. (20) and its solution is therefore given by Eq. (21). It then follows using Eq. (17) in (22) that the conditional molecule number distribution $\Pi(x|s, s_0)$ for cells of size s that were born at size s_0 satisfies

$$\alpha s\frac{\partial}{\partial s}\Pi(x|s, s_0) = \mathbb{Q}(s)\Pi(x|s, s_0), \tag{23a}$$

where $\mathbb{Q}$ is the transition matrix of the molecular reactions defined after Eq. (4). The equation is to be solved subject to the boundary condition

$$\Pi(x|s_0, s_0) = \sum_{x'}\int_0^\infty \mathrm{d}s' \int_0^{s'} \mathrm{d}s_0'\, B(x|x', s', s_0)\rho(s', s_0'|s_0)\Pi(x'|s', s_0'), \tag{23b}$$

which follows from combing the boundary condition (22b) with Eq. (19) and Eq. (21). The solution of these equations depends implicitly on the ancestral cell size distribution ρ,

$$\rho(s', s_0'|s_0) = \frac{1}{\psi_{\mathrm{bw}}(s_0)}\frac{s_0}{s'}B(s_0|s')\varphi(s'|s_0')\psi_{\mathrm{bw}}(s_0'), \tag{23c}$$

that gives the probability of a cell born at size s_0 having an ancestor with division size s' and birth size s_0'. The main difference between the molecule number distributions of the ABM and the EDM/ENM is the boundary condition at cell division, which as we shall see can have a significant effect on the reaction dynamics.

III. RESULTS

A. The effective dilution model is valid for a class of reaction networks

A central result of our analysis (Theorem 1 in Appendix A) is the fact that, if the EDM (4) admits a stationary solution with generating function of the form

$$G_{\mathrm{EDM}}(z|s) = \sum_x z^x \Pi_{\mathrm{EDM}}(x|s) = F(s(z-1)), \tag{24}$$

for any function F independent of s, then it is also a solution of the ABM (23):

$$\Pi(x|s, s_0) = \Pi_{\mathrm{EDM}}(x|s), \tag{25}$$

and this solution is independent of the birth size s_0 and any details of the size distribution and cell size control. The condition is sufficient and necessary assuming independent binomial partitioning at cell division and can be checked in practice without explicitly solving for $G_{\mathrm{EDM}}(z|s)$ (or $\Pi_{\mathrm{EDM}}(x|s)$). For example, assuming mass-action kinetics (6), the validity of the EDM requires a reaction network to be composed of only mono-molecular reactions (see Appendix A):

$$\varnothing \xrightarrow{s} S_1, \text{ or } S_1 \to \varnothing, \text{ or } S_1 \to S_2, \tag{26}$$

for any pair of species S_1 and S_2 that are partitioned at cell division. It is well known, that the solutions of (26)

Box 1: First-Division Algorithm for agent-based simulations of size-dependent gene regulatory networks

Exact simulation algorithm of general stochastic reaction networks within growing cells (agents) undergoing binary cell division according to cell size control rules[46,48,49]. The algorithm combines the Extrande method[45] for simulating reaction networks embedded in a growing cell and the First-Division algorithm[36] for the population dynamics. The state of each cell is given by birth time t_0, birth size s_0, present cell size s and the vector of molecule numbers x.

Algorithm 1: First-division algorithm simulating agent-based population dynamics

Input: Cell states $\{t_{0,i}, s_{0,i}, s_i, x_i\}_{i=1,\dots,M}$ for population of M cells.
Output: Cell states at time $T > 0$.
Require: Growth rate α, cell size control model (18) and division error distribution π.
Initialise $t \leftarrow 0$;
while $t < T$ **do**
  Compute division times $t_{d,i} = t_{0,i} + \alpha \ln(s_{d,i}/s_{0,i})$ for $i = 1, .., M$ where $s_{d,i}$ is computed from (18);
  Compute the first division time $t_d^* = \min_i t_{d,i}$ and index $i^* = \text{argmin}_i t_{d,i}$;
  if $t_d^* < T$ **then**
    Grow all cells until $t = t_d^*$ using Algorithm 2;
    Draw a random number $\theta \sim \pi$ and divide the cell i^* according to $s_{0,i^*}, s_{i^*} \leftarrow \theta s_{i^*}$ and $s_{0,M+1}, s_{M+1} \leftarrow (1-\theta)s_{i^*}$;
    Partition molecules binomially $x_{i^*}^{\text{new}} \sim \mathcal{B}(x_{i^*}^{\text{old}}, \theta)$ and assign the rest to the other daughter cell $x_{M+1} \leftarrow (x_{i^*}^{\text{old}} - x_{i^*}^{\text{new}})$;
    Set $M \leftarrow M + 1$;
  else
    Grow all cells until $t = T$ using Algorithm 2;
  end
end

Algorithm 2: Extrande algorithm simulating reaction networks in a growing cell

Input: Cell state (t_0, s_0, s, x) at time T_0.
Output: Cell state at time $T' > T_0$.
Require: Growth rate α, stoichiometric vectors $(\nu_r)_{r=1,..,R}$ and propensities $(w_r)_{r=1,..,R}$.
Initialise $t \leftarrow T_0$ and let $a_0(x, s) = \sum_{r=1}^{R} w_r(x, s)$;
while $t < T'$ **do**
  Compute bound $B = \max_{t' \in [T_0, T']} a_0(x, s_0 e^{\alpha t'})$;
  Generate putative reaction time $\tau \sim \exp(1/B)$;
  if $t + \tau \geq T'$ **then**
    Set time $t \leftarrow T'$ and cell size $s \leftarrow s_0 e^{\alpha(T'-t_0)}$;
  else
    Update time $t \leftarrow t + \tau$ and cell size $s \leftarrow s_0 e^{\alpha(t-t_0)}$;
    Generate random number $u \sim U_{(0,1)}$;
    if $a_0(x, s) \geq Bu$ **then**
      Choose reaction associated with the smallest positive integer j less than or equal to R satisfying $\sum_{r=1}^{j} w_r(x, s) \geq Bu$ and update molecular state $x \leftarrow x + \nu_j$;
    end
  end
end

are multi-variate Poisson distributions[57–59] due to extra dilution reactions present. Our result is however not limited to this case but also applies to including exogenous genetic states D, which are not partitioned at cell division and do not scale with cell size, as commonly included in gene expression models via reactions $D \xrightarrow{s} D + S_1$. For example, it includes general multistate gene expression models; i.e., compare Eq. (7) in Ham et al.[60] with our condition (24) when transcription rates are assumed to scale with cell size.

We illustrate the predictive power of this result for three common gene expression models (Fig. 2a-c) that admit explicit solutions of the chemical master equation. For bursty mRNA expression involving a two-state promoter[54] (Fig. 2a), condition (24) is met whenever the transcription rate scales with cell size. For bursty protein expression, condition (24) is met when translation rate, and hence burst size, scales with cell size for both the two-stage (Fig. 2b) and three-stage gene expression models (Fig. 2c)[55]. We further observe excellent numerical agreement between the ABM simulations and analytical EDM solutions in all these cases.

On the other hand, discrepancies between the EDM and ABM solutions will be apparent when reactions violate our network conditions. To illustrate this point, we return to the simple gene expression model (13). Note that in the EDM extra reactions are added for the dilution of mRNAs and proteins, while for the ABM proteins are explicitly diluted through growth and divisions. Using our condition (24), it is straight-forward to verify that the Poissonian mRNA distributions of the effective dilution model coincide exactly with the distributions of the ABM (Fig. 2d). However, this condition is not met for the protein distribution since the translation reaction is not a monomolecular reaction of the form (26). To demonstrate the breakdown of the EDM, we compare the analyt-

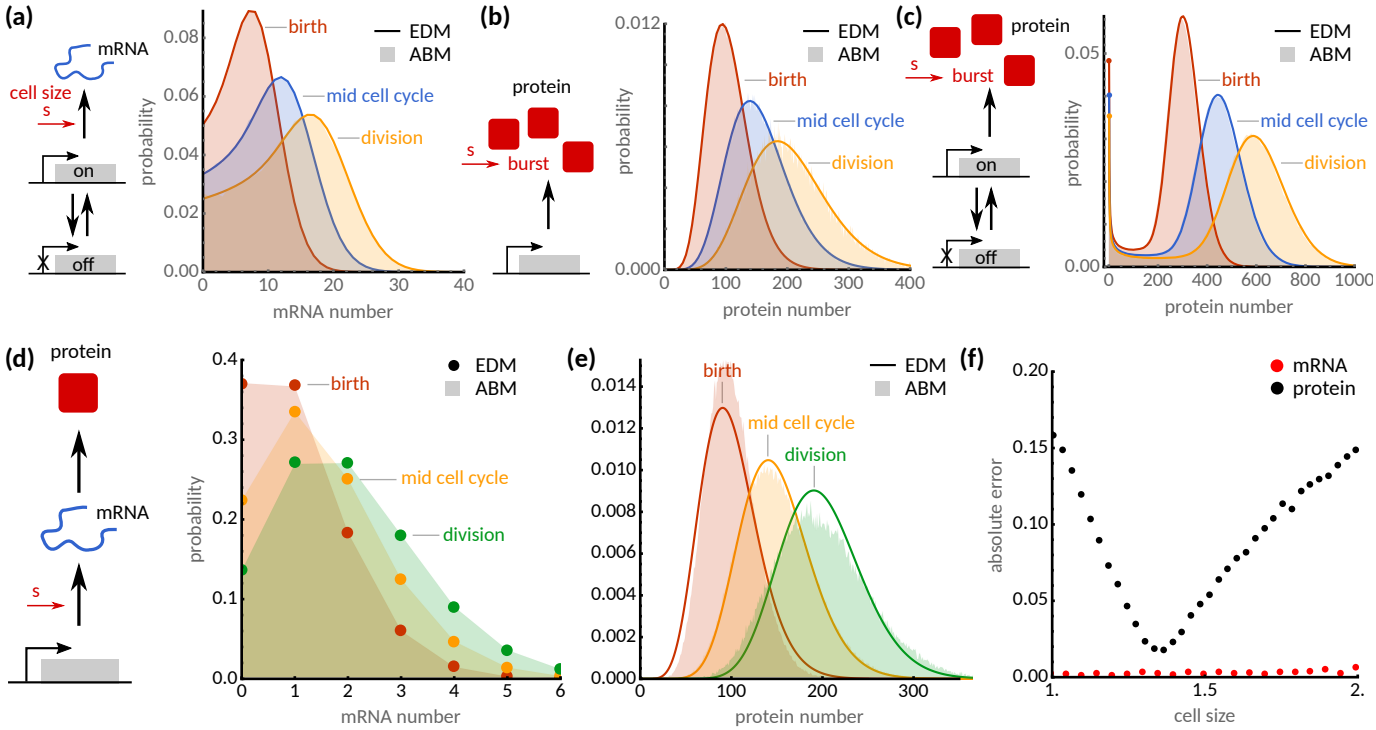

FIG. 2. **Distributions of CME and agent-based models agree for mono-molecular reaction networks.** **(a-c)** The EDM (solid lines, analytical solution[54,55]) agrees with agent-based simulations (shaded areas) for a range of gene expression models. **(d)** mRNA distributions simulated using the ABM (shaded areas) are shown for cells of sizes $s = s_0$ (red), $s = 1.5s_0$ (orange) and $s = 2s_0$ (green), which agree with the effective dilution model (dots, Poisson distribution). **(e)** Simulated protein distributions (shaded areas) disagree with the effective dilution model (solid lines, solution in Ref.[56]). **(f)** Absolute error (ℓ_1) of the effective dilution model as a function of cell size for mRNA (teal) and protein (red) distributions. ABM simulations were obtained using the First-Division Algorithm (Box 1) assuming an adder model ($a = 1$) and parameters $k_0 = 10$, $k_{\rm dm} = 9$, $k_s = 100$, $\alpha = 1$. Cell cycle noise Δ assumes gamma distribution $\varphi(\Delta)$ with unit mean and standard deviation 0.1, while division noise assumes symmetric beta distribution with $\mathrm{CV}[\theta] = 0.01$.

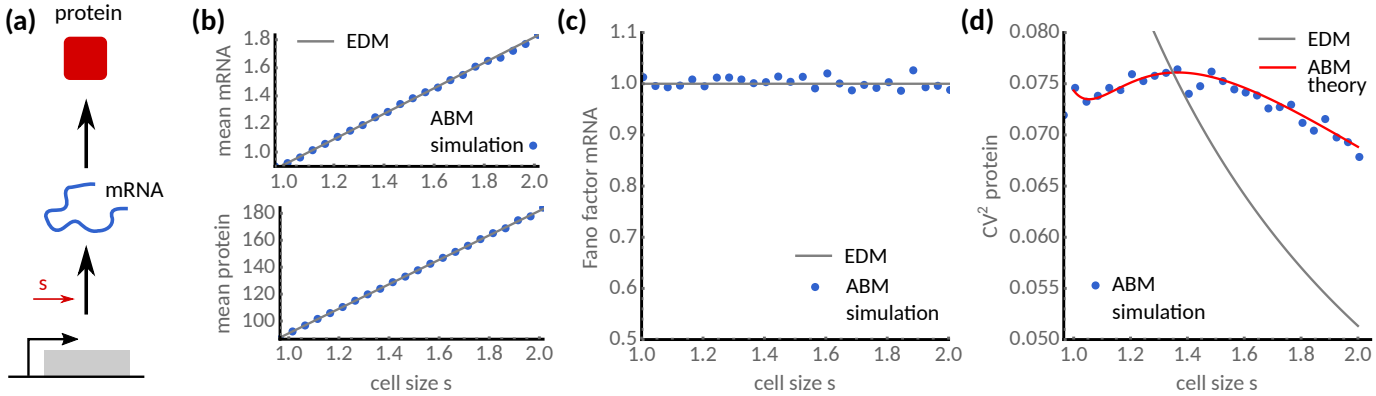

FIG. 3. **Comparing the statistics of the effective dilution and agent-based models.** **(a)** Simple model of mRNA transcription and protein translation (13). **(b)** Mean mRNA (top) and protein levels (bottom) agree with the EDM (solid grey lines) and ABM simulations (blue dots). **(c)** mRNA statistics display unit Fano factor indicating Poisson statistics in agreement with EDM. **(d)** ABM simulations (dots, Box 1) display non-monotonic cell size scaling of protein noise, which are predicted by the agent-based theory (solid red) but not by the EDM (solid grey). Parameters are $k_0 = 10$, $k_{\rm dm} = 10$, $k_s = 100$, $\alpha = 1$. Cell size control parameters are as in Fig. 2.

ical steady state distributions obtained by Bokes et al.[56] against ABM simulations at various cell sizes (Fig. 2e).

We observe that the error of the EDM (as quantified by the ℓ_1-distance of the two distributions, Fig. 2f) is particularly pronounced for newborn and dividing cells. The remainder of this article is dedicated to investigate the sources and consequences of these discrepancies.

B. The effective dilution model approximates the mean concentrations

The conditions given above provide general criteria for the validity of the EDM probability distributions. In practice, however, it is often sufficient to analyse the first few moments such as mean and variances. Here, we establish that under the mass-action scaling assumption (6) the mean concentrations of the ABM agree with the EDM.

Using Eq. (23) it can be shown the mean numbers satisfy

$$\alpha s \frac{\partial}{\partial s} E_\Pi[x|s,s_0] = \sum_{r=1}^{R} \nu_r w_r(E_\Pi[x|s,s_0]), \tag{27a}$$

and the boundary condition

$$E_\Pi[x|s_0,s_0] = E_\rho\left[\frac{s_0}{s'} E_\Pi[x|s',s_0']\Big|s_0\right]. \tag{27b}$$

Using a linear noise approximation, we set $E[x|s,s_0] = s\bar{X}$ and $w_r(E_\Pi[x|s,s_0]) \approx s f_r(\bar{X})$ and insert the resulting expression into Eqs. (27) from which it follows that $\bar{X}$ is independent of size and satisfies the rate equations (2). We conclude that, for mass-action kinetics, the effective dilution model holds on average. While this result is exact for networks with linear propensities, it represents an approximation for networks with nonlinear propensities akin to the system size expansion[6] valid in the limit of large cell size and molecule numbers.

C. Scaling of fluctuations with size in individual cells manifests the breakdown of effective dilution model

Next we will investigate the scaling of fluctuations with cell size. Under the linear noise approximation the covariance matrix $\Sigma(s,s_0) = \mathrm{Cov}_\Pi[x|s,s_0]$ evolves according to

$$\alpha s \frac{\partial}{\partial s}\Sigma(s,s_0) = \mathcal{J}\Sigma(s,s_0) + \Sigma(s,s_0)\mathcal{J}^T + s\mathcal{D}(\bar{X}), \tag{28a}$$

where $\mathcal{J}(\bar{X})$ and $\mathcal{D}(\bar{X})$ are the Jacobian and diffusion matrices defined after Eq. (11). To make analytical progress we assume for now that cell division is deterministic, which implies the following boundary condition

$$4\Sigma(s_0,s_0) = 2s_0\mathrm{diag}(\bar{X}) + \Sigma(2s_0,s_0). \tag{28b}$$

The first term is due to binomial partitioning and the second stems from gene expression noise at division. It is implicit in the deterministic division assumption that the birth size s_0 across cells is fixed and that the size distribution in Eq. (21) reduces to

$$\Pi(s) = \Pi(s|s_0) = \frac{2s_0}{s^2} \tag{29}$$

for $s_0 \le s \le 2s_0$ and zero otherwise, in agreement with previous results[61,62]. Similarly, the ancestral distribution (23c) reduces to $\rho(s',s_0'|s_0) = \delta(s'-2s_0)\delta(s_0'-s_0)$.

Eqs. (28) can be solved in closed form using the eigendecomposition of the Jacobian $\mathcal{J}$. The solution to (28a) that respects the boundary condition (28b) is (Appendix C)

$$\Sigma(s,s_0) = \sum_{ij} \frac{s\,\hat{u}_i\hat{u}_j^\dagger}{(\alpha-\lambda_i-\lambda_j^*)} \times \left[\tilde{\mathcal{D}}_{ij} + \frac{\tilde{\mathcal{D}}_{ij} + \tilde{X}_{ij}(\lambda_i+\lambda_j^*-\alpha)}{2^{\frac{\lambda_i+\lambda_j^*}{\alpha}-1}-2}\left(\frac{s_0}{s}\right)^{1-\frac{\lambda_i+\lambda_j^*}{\alpha}}\right], \tag{30}$$

where † denotes the conjugate-transpose and we defined the matrices $\tilde{\mathcal{D}} = U^{-1}\mathcal{D}U^{-\dagger}$, $\tilde{X} = U^{-1}\mathrm{diag}(\bar{X})U^{-\dagger}$ and $U = (\hat{u}_1,\ldots,\hat{u}_N)$ whose columns are the eigenvectors of $\mathcal{J}$ such that $U^{-1}\mathcal{J}U = \mathrm{diag}(\lambda)$.

We demonstrate the implications of this assumption using example (13). The mean of mRNA numbers m and protein numbers p are

$$E_\Pi[m|s] = s\frac{k_0}{\alpha\gamma}, \quad E_\Pi[p|s] = s\frac{bk_0}{\alpha}, \tag{31}$$

where the constants are defined in Eq. (16). These expressions hold both for the EDM and the ABM as shown in the previous section and are in excellent agreement with the ABM simulations (Fig. 3a,b).

Using the linear noise approximation, Eq. (28), we find that the cell size dependent fluctuations satisfy

$$\begin{aligned}
&\mathrm{Var}_\Pi[m|s,s_0] = E_\Pi[m|s],\\
&\mathrm{Cov}_\Pi[m,p|s,s_0] = \frac{sbk_0}{\alpha\gamma}\left(1+\left(\frac{s_0}{s}\right)^\gamma \frac{2^\gamma}{1-2^{\gamma+1}}\right),\\
&\mathrm{Var}_\Pi[p|s,s_0] = E_\Pi[p|s]\times\\
&\qquad\left(1+2b-\frac{s_0}{s}\frac{4b\gamma}{3(\gamma-1)}+\left(\frac{s_0}{s}\right)^\gamma\frac{b}{\gamma-1}\frac{2^{\gamma+1}}{(2^{\gamma+1}-1)}\right),
\end{aligned} \tag{32}$$

a result which is exact in the absence of cell cycle variations since our example involves only linear reactions.

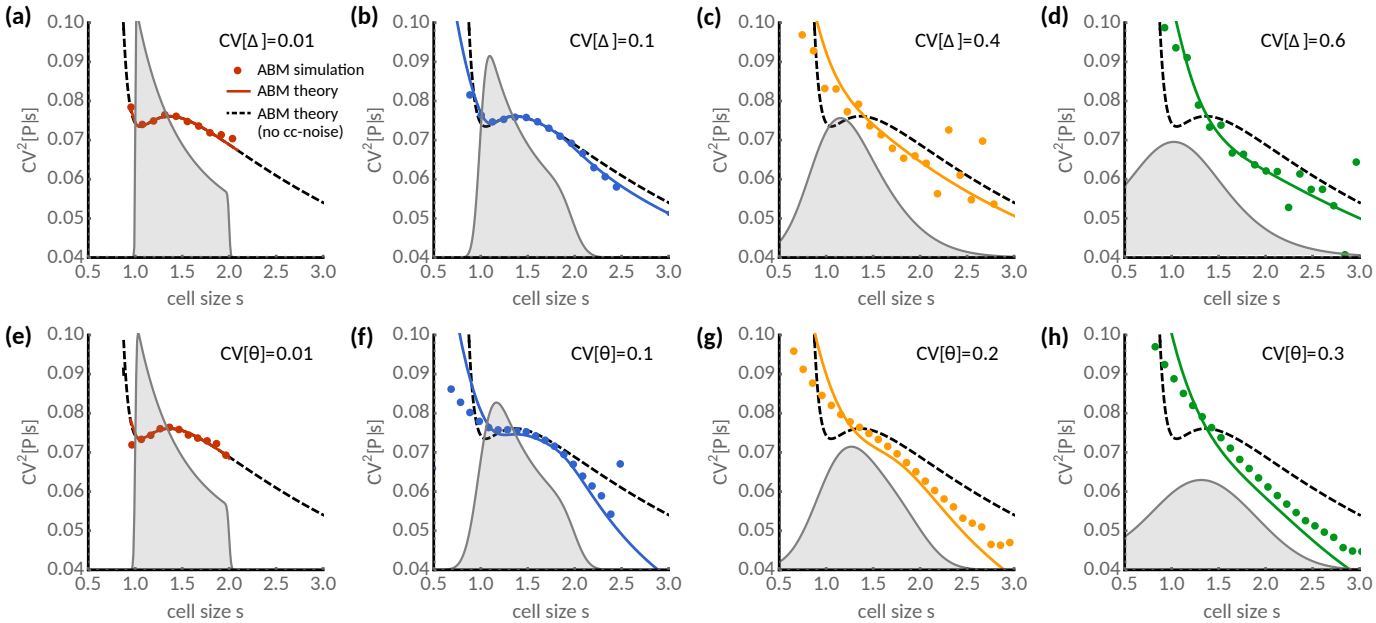

FIG. 4. **Effect of noise in cell size control and division on gene expression noise in single cells. (a-d)** Protein noise as a function of cell size s for various noise levels in added size CV^2_Δ. For comparison, the prediction without cell cycle noise (dashed black line, Eq. (32)) and the cell size distributions (shaded grey) are shown. **(e-h)** Same as (a-d) but with division noise affecting the inherited size fraction CV^2_θ. Analytical predictions (solid lines, Eq. (34a) with (32) and (34b)) and ABM simulations (dots) using the First-Division Algorithm (Box 1) are shown. Gene expression model and all other parameters are as given in Fig. 3. Added size Δ assumes a gamma distribution with unit mean and $\mathrm{CV}[\Delta] = 0.01$ (e-f) while division errors θ followed a symmetric beta distribution with $\mathrm{CV}[\theta] = 0.01$ (a-d).

First, we note that the mRNA variance of the ABM agrees precisely with the effective dilution model (Fig. 3c). The agreement is a direct consequence of the network conditions (26). However, the expressions for the predicted mRNA-protein covariance and protein variance disagree. To explore this dependence, we compare the corresponding coefficients of variation of both models (Fig. 3d). The EDM overestimates cell-to-cell variation of small cells but overestimates it for large cells. Moreover, the EDM's coefficient of variation decreases monotonically with cell size, but this is not the case for the ABM.

Strikingly, the dependence of the coefficient of variation displays a peak as cells progress through the cell cycle (Fig. 3d, solid red line) that is not seen in the EDM (solid grey) but is in excellent agreement with the ABM simulations (blue dots). This can be seen directly from Eqs. (32) for which protein fluctuations can be approximated in the limit of fast mRNA degradation ($\gamma \gg 1$) as

$$\mathrm{CV}^2_{\mathrm{ABM}}[p|s, s_0] \approx \frac{1}{E_\Pi[p|s]}\left(1 + b\left(2 - \frac{s_0}{s}\frac{4}{3}\right)\right), \quad (33)$$

which has a maximum at a cell size of $s = s_0 \frac{8b}{3(2b+1)}$ as confirmed by agent-based simulations (Fig. 3d). Depending on the burst size, the peak shifts from $s = s_0$ for $b = 3/2$ to $s = 4/3 s_0$ for $b \gg 1$. The qualitatively different scaling of gene expression noise with cell size manifests the breakdown of the EDM, similarly to what has been previously observed in the noise dependence on cell age in ABMs[37].

D. Effect of cell size control on gene expression dynamics

Next, we ask how fluctuations in the cell size control affect gene expression noise. It may be intuitively expected that noise in division timings and partitioning causes variable birth sizes and hence variable expression levels. While it is clear that from the conditions given in Sec. III A this noise source cannot impact mRNA fluctuations, its effect on protein noise remains to be elucidated.

To this end, we assume small birth-size variations such that their impact can be quantified by replacing the actual birth size with an averaged estimate $E_\Pi[s_0|s]$ of the retrospective birth size for a cell of size s. The covariance matrix (or any other moment) can then be approximated as

$$\Sigma(s) = E_\Pi[\Sigma(s, s_0)|s] \approx \Sigma(s, E_\Pi[s_0|s])). \quad (34a)$$

This simplification can formally be justified through a saddle point approximation since the joint distribution $\Pi(s, s_0)$ has a maximum at $\Pi(s, E_\Pi[s_0|s]))$. Since generally no analytical expression of $E_\Pi[s_0|s]$ can be derived from Eq. (21) in the presence of cell size control fluctuations, we approximate $E_\Pi[s_0|s]$ by a matched asymptotic expansion (Appendix B):

$$E_\Pi[s_0|s] \approx \bar{s}_0 - \underbrace{\sigma \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{(s-\bar{s}_0)^2}{2\sigma^2}}}{1 + \mathrm{erf}\left(\frac{s-\bar{s}_0}{\sqrt{2}\sigma}\right)}}_{\text{small cells}} + \underbrace{2a\sigma^2\gamma(s - a\bar{s}_0)}_{\text{large cells}}, \tag{34b}$$

which holds for the linear cell size control model (18). The first term is the average birth size in the absence of cell size control fluctuations, the second term denotes the contributions from small cells, while the third term stems from large cells. The parameters in Eq. (34b) are given by the mean birth size $\bar{s}_0$ and variance σ^2 in a backward lineage tracing the ancestors of a random cell in the population (see Ref.[51] for details):

$$\bar{s}_0 = \frac{(2-a)\left(1+\mathrm{CV}_\theta^2\right)}{2 - a(1+\mathrm{CV}_\theta^2)},$$
$$\sigma^2 = \bar{s}_0^2 \frac{\mathrm{CV}_\Delta^2\left(1+3\mathrm{CV}_\theta^2\right)\left(2-a(1+\mathrm{CV}_\theta^2)\right)^2 + 4\mathrm{CV}_\theta^2\left(1-\mathrm{CV}_\theta^2\right)}{\left(\mathrm{CV}_\theta^2+1\right)^2\left(4-a^2(1+3\mathrm{CV}_\theta^2)\right)}. \tag{34c}$$

Eqs. (34) provide a closed form approximation of the cell size dependence of any given moment accurate to order $O(\sigma^3)$.

To test the accuracy of proposed approximation, we first evaluate the dependence against cell size control noise. We observe that increasing noise leads to a more monotonic decrease gene expression noise with cell size (Fig. 4a-d) in good agreement with ABM simulations, even for large cell size fluctuations. We further ask about the effects of partitioning noise, which shows a similar dependence but agrees less well with the ABM simulations for cells smaller than the average birth size (Fig. 4e-h), presumably since the effect of large variability in birth sizes is not captured in our small noise approximation. Nevertheless, the present approximation qualitatively captures the overall cell size dependence of the ABM simulations (Fig. 4). Our findings confirm that birth size variation contributes significantly to the cell size dependence of gene expression noise.

E. Extrinsic noise models provide surprisingly accurate approximations of total noise in agent-based populations

So far, we explored whether the EDM captures the cell-size dependence of distributions in single cells. Instead, here we investigate if the ENM (7) can capture the total variation across a population irrespective of cell size thus including extrinsic noise. It is implicit in our analysis that the ENM follows the size distribution $\Pi(s)$ of the ABM. We first aim to evaluate the total noise statistics of concentrations

$$X = \frac{x}{s}. \tag{35}$$

Since the mean concentration is independent of size, $E_\Pi[X|s, s_0] = \bar{X}$, we have $\mathrm{Cov}_\Pi[E[X|s, s_0]] = 0$ and

$$E_\Pi[X] = \bar{X}, \quad \mathrm{Cov}_\Pi[X] = E_\Pi[\bar{\Sigma}(s, s_0)], \tag{36}$$

where $\bar{\Sigma}(s, s_0) = \mathrm{Cov}_\Pi[X|s, s_0]$ given by the solution of Eqs. (28). The concentration covariance matrix can be obtained analytically in the limit of deterministic cell division by integrating (30) over the cell size distribution $\Pi(s)$ in (29), for which we obtain

$$\mathrm{Cov}_\Pi[X] = \frac{1}{\Omega}\sum_{ij} \hat{u}_i \hat{u}_j^\dagger \frac{\beta_{ij}\left(\coth\left(\frac{1}{2}\beta_{ij}\ln 2\right)+3\right)}{3(\beta_{ij}+1)} \frac{\alpha \tilde{X}_{ij}}{\xi_{ij}} + \frac{\beta_{ij}\left(3\beta_{ij} - \coth\left(\frac{1}{2}\beta_{ij}\ln 2\right)\right)}{3\left(\beta_{ij}^2 - 1\right)} \frac{\tilde{\mathcal{D}}_{ij}}{\xi_{ij}}, \tag{37}$$

where $\xi_{ij} = 2\alpha - \lambda_i - \lambda_j^*$, $\beta_{ij} = \frac{\xi_{ij}}{\alpha}$, $\Omega^{-1} = E_\Pi[s^{-1}] = \frac{3}{4}\frac{1}{s_0}$, and $\hat{u}_i$ are the eigenvectors of the Jacobian $\mathcal{J}$ introduced before Eq. (30). Similarly, considering molecule number fluctuations, $\mathrm{Cov}_\Pi[x] = \Sigma_{\mathrm{ABM}}^{\mathrm{int}} + \Sigma_{\mathrm{ABM}}^{\mathrm{ext}}$, the intrinsic noise contribution is given by

$$\Sigma_{\mathrm{ABM}}^{\mathrm{int}} = \Omega \sum_{ij} \hat{u}_i \hat{u}_j^\dagger \left[\frac{\left(2^{\beta_{ij}} - 2\right)\beta_{ij}}{\left(2^{\beta_{ij}} - 1\right)\left(\beta_{ij} - 1\right)\ln 4} \frac{\alpha \tilde{X}_{ij}}{\xi_{ij}} + \frac{\beta_{ij}\left(\left(2^{\beta_{ij}} - 1\right)\beta_{ij}\ln 4 - 2^{\beta_{ij}}(1+\ln 4) + 2 + \ln 4\right)}{\left(2^{\beta_{ij}} - 1\right)\left(\beta_{ij} - 1\right)^2 \ln 4} \frac{\tilde{\mathcal{D}}_{ij}}{\xi_{ij}}\right], \tag{38}$$

with $\Omega = E_\Pi[s] = s_0 \ln 4$, and the non-zero extrinsic noise contribution $\Sigma_{\mathrm{ABM}}^{\mathrm{ext}} = \mathrm{Var}_\Pi(s)\bar{X}\bar{X}^T$ as in the ENM presented in Sec. II A 4.

Total noise of agent-based and extrinsic noise models

We go on to compare the ENM introduced in Sec. III E with the ABM. In the case of a single variable, Eq. (37) greatly simplifies (see Appendix C), which shows that the relative error of the ENM can be bounded:

$$\frac{6}{7} \leq \frac{\mathrm{CV}_{\mathrm{ENM}}^2[X]}{\mathrm{CV}_{\mathrm{ABM}}^2[X]} \leq \frac{3}{2}\ln 2, \tag{39}$$

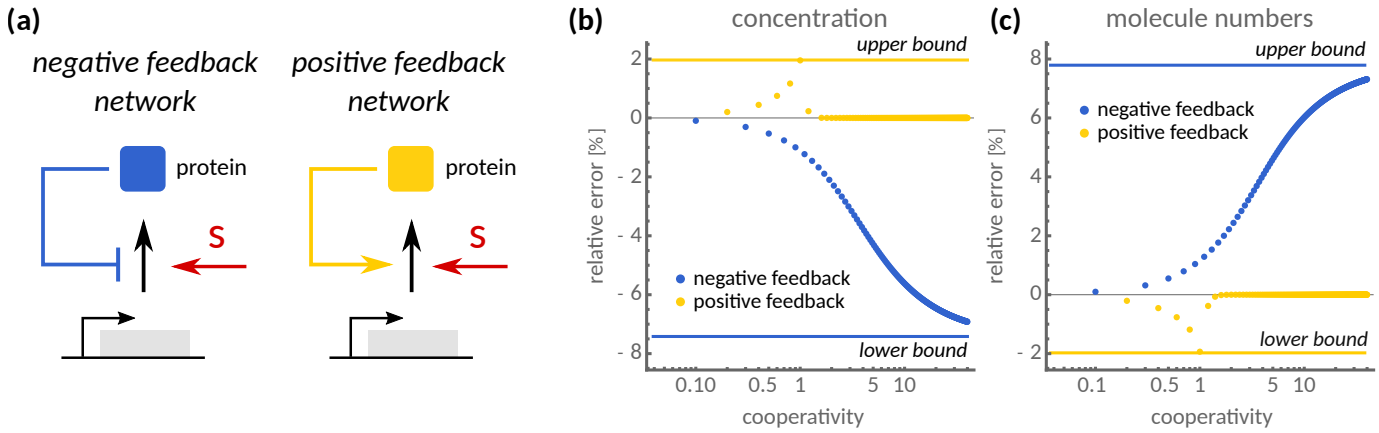

FIG. 5. **Genetic feedback regulation limits the validity of extrinsic noise models.** **(a)** Common autoregulatory network motifs including negative and positive feedback activation (see Appendix C for details). Percentage error $[100\% \times (\mathrm{CV}_{\mathrm{ENM}}/\mathrm{CV}_{\mathrm{ABM}} - 1)]$ of the ENM relative to the ABM is shown for **(b)** the coefficient of variation of concentration fluctuations and **(c)** the coefficient of variation of molecule number fluctuations as a function of the Hill exponent measuring cooperativity of gene autoregulation. For negative feedback regulation, the ENM underestimates the concentration fluctuations of the ABM (blue dots in (b)), while it overestimates molecule number fluctuations (blue dots in (c)). The maximum deviation is achieved in the limit of ultrasensitive gene regulation ($\approx 8\%$), in agreement with the theoretical bounds (solid blue lines, Eqs. (39) and (40)). For positive feedback regulation, the ENM always overestimates concentration fluctuations (yellow dots in (b)), while it underestimates the molecule number fluctuations (yellow dots in (c)). Maximum deviation is attained for a Hill coefficient of 1 in agreement with the theoretical bounds (solid yellow lines). Parameters are $k_0 = \alpha = 1$ and $K = 0.2$ for the negative feedback network with activation function (C4) and $k_0 = K = \alpha = 1$, $\delta = 10^{-4}$ for the positive feedback network with activation function (C5).

implying that the ENM overestimates the coefficient of variation by at most 8% but underestimates it by at most 2%. Conversely, an analogous calculation (Appendix C) for the coefficient of variation in molecule numbers x shows that

$$2\ln^2(2) \leq \frac{\mathrm{CV}^2_{\mathrm{ENM}}[x]}{\mathrm{CV}^2_{\mathrm{ABM}}[x]} \leq \frac{4\ln 2}{1+\ln 4}, \tag{40}$$

implying that the ENM underestimates the coefficient of variation of the ABM by at most 8% but underestimates it by at most 2%, opposite to what we found to concentration noise statistics.

We apply these error bounds to gene regulatory networks with positive and negative transcriptional feedback activation (see Appendix C for details). The positive feedback loop (Fig. 5a, yellow), achieves the upper bound for the coefficient of variation of concentration (Fig. 5b, yellow) depending on parameters and cooperativity, and it achieves the lower bound for the coefficient of variation of molecule numbers (Fig. 5c, yellow). Conversely, the network with negative feedback architecture (Fig. 5a, blue) achieves the lower bound for the coefficient of variation of concentration (Fig. 5b, blue) but achieves the upper bound for the coefficient of variation of molecule numbers (Fig. 5c, blue). Nevertheless, the discrepancy between the EDM and ABM are generally less than 8% independent of the gene expression network considered.

The prescribed error bounds strictly hold for networks involving only a single species and must be considered as error estimates for general multi-species networks. For example, for the simple two-species gene expression network (13), our theory, Eq. (37), provides an analytical expression for the protein concentration fluctuations

$$\mathrm{CV}^2_{\mathrm{ABM}}[P] = \frac{1}{\Omega\bar{P}}\left[1 + \frac{2b\left(12 - \frac{24(2^{\gamma-1}-1)}{(2^{\gamma+1}-1)(\gamma-1)} + 13\gamma\right)}{27(\gamma+2)}\right], \tag{41}$$

where $\Omega = E_{\Pi}[s^{-1}]^{-1}$ and γ, b are defined as in Eq. (16). Similarly, using Eq. (38) we obtain the corresponding molecule number fluctuations

$$\mathrm{CV}^2_{\mathrm{ABM}}[p] = \mathrm{CV}^2_{\Pi}[s] + \frac{1}{\Omega\bar{P}}\left[1 + \frac{2b\left(\frac{2(2^{\gamma-1}-1)}{(2^{\gamma+1}-1)(\gamma-1)} + \gamma(\ln(8)-1) - 1\right)}{3\gamma\ln(2)}\right] \tag{42}$$

with $\Omega = E_{\Pi}[s]$. These expressions involve an intricate dependence on γ. It can be verified that, in this case,

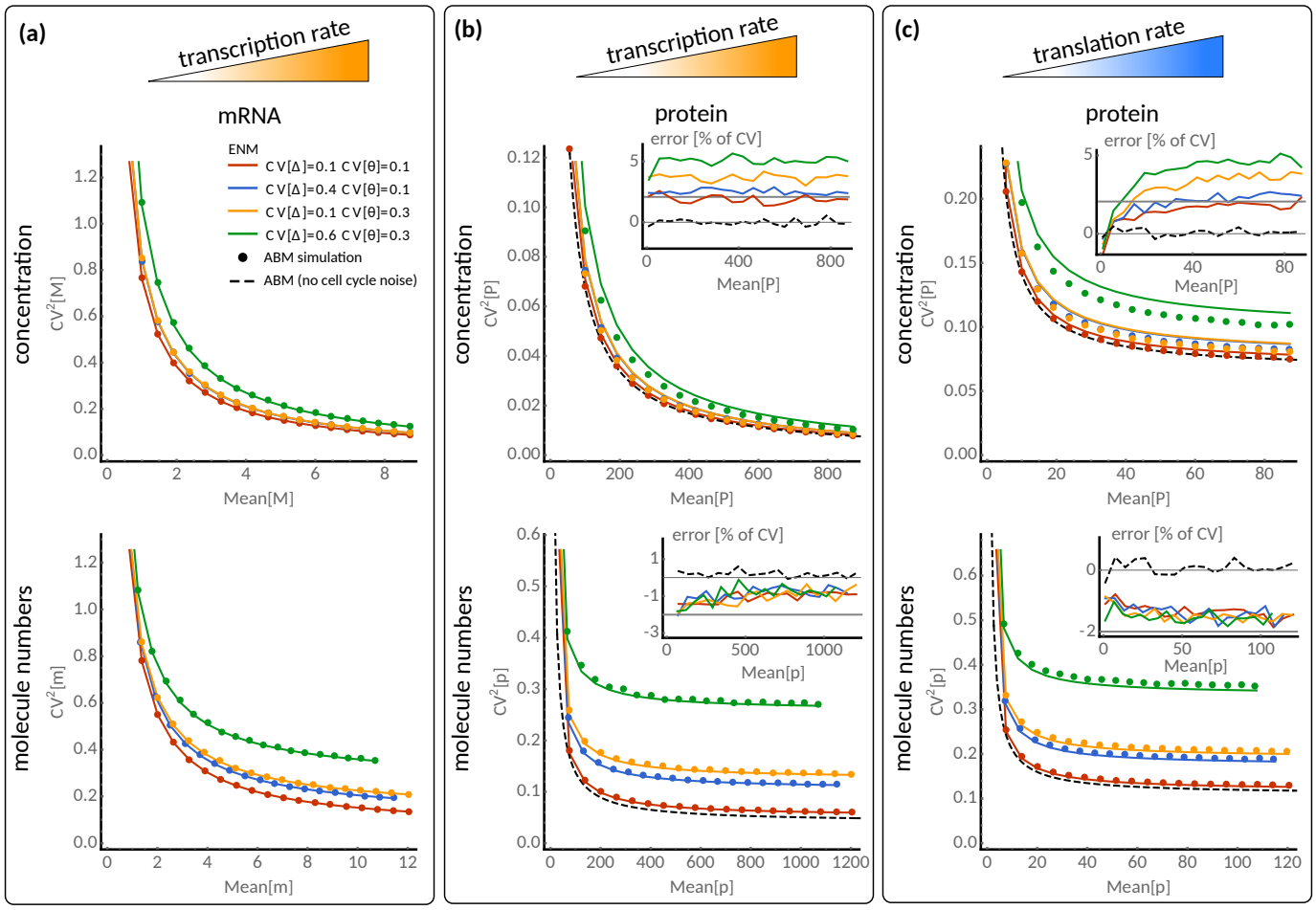

FIG. 6. **The extrinsic noise model approximates gene expression noise with size control and division errors. (a)** Scaling of mRNA concentration noise with mean concentrations for various noise levels in added size CV[Δ] and partition noise CV[θ] when the transcription rate k_0 is varied (top). Corresponding scaling is shown for mRNA numbers (bottom). Analytical predictions of the ENM (solid lines, Eq. (8) with (10) and (12)) and ABM simulations using the First-Division Algorithm (dots, Box 1) are shown. The inset shows the relative error in CV of the effective dilution compared to ABM simulations [$100\% \times (\mathrm{CV}_{\mathrm{ENM}}/\mathrm{CV}_{\mathrm{ABM}} - 1)$]. **(b)** Scaling of protein noise with mean protein concentration (top) and numbers (bottom) when the translation rate k_s is varied. **(c)** Same as (b) but varying the transcription rate k_s. For comparison, the ABM predictions without cell cycle noise are shown (dashed black lines, Eq. (41) for concentrations and (42) for molecule numbers) and the implied error bounds (solid grey). See caption of Fig. 3 for the remaining parameters.

the ENM (Eq. (15)) overestimates the coefficient of variation of concentration and underestimates the coefficient of variation of molecule numbers by at most 2% (irrespectively of γ). These bounds are in agreement with the theoretical estimates obtained for single-species networks and we conjecture that these bounds also provide accurate error estimates for multi-species networks. Our theoretical error bounds (39) and (40) thus argue that the ENM generally provides surprisingly accurate approximations of the ABM statistics.

To conclude this section we ask which cell size distributions make the ENMs and ABMs agree exactly, independently of the reaction network considered. To this end, we make the ansatz $\Pi(s) \sim s^{-\phi}$ for deterministic growth in the interval $s \in [s_0, 2s_0)$ and we integrate (30) over this hypothetical size distribution. Interestingly, we find different exponents for concentrations ($\phi = 1$) and for molecule number fluctuations ($\phi = 3$). It is intriguing that the case $\phi = 1$ corresponds to a size distribution observed in mother-machine lineage[51] (assuming deterministic growth), and hence the total noise statistics observed in this microfluidic device will mirror the ENM only when concentrations are considered but not for molecule numbers. On the contrary, molecule number

fluctuations of the ENM seem to lack a physical interpretation as we are unaware of a realistic counterpart for the case $\phi = 3$. Nevertheless, these theoretical exponents are close to $\phi = 2$ observed in population snapshots, which justifies why ENMs fair relatively well in approximating the total noise in ABMs.

Effect of cell size control fluctuations and division errors

Stochastic fluctuations in cell size control and division errors provide additional noise sources affecting protein levels across single cells. Since these sources are beyond the scope of our analytical theory, we assess their impact on the predictions of the ENM and ABM for the simple two-species gene expression model (13) using detailed agent-based simulations.

We observe that, for varying transcription rate, the mRNA noise-mean relationship of the ABM follows exactly the ENM predictions for various strength of cell size control fluctuations and division errors (Fig. 6a), both in concentration fluctuations and numbers. The observation confirms our theoretical predictions (Condition (26)) that the ENM is exact in this case. However, the protein noise-mean relationships of the ABM and ENM differ (Fig. 6b). This discrepancy, albeit small, increases with cell size control noise for concentration measures but appears independent of cell size control noise for protein number fluctuations (Fig. 6b, insets). Specifically, we observe that the ENM underestimates ABM noise of protein numbers, while the ENM overestimates protein concentration fluctuations in qualitative agreement with the theoretical bounds (39)-(40). With increasing cell size control noise, our bounds derived for deterministic divisions are accurate for molecule number noise but not for concentration noise which increases beyond the maximum deviations of 2% implied by (41) and (42). The same conclusions hold when varying translation rate in the noise-mean relationship (Fig. 6c). Presumably, robustness of the molecule number bounds is due the fact that the ENM and ABM predictions are generally dominated by extrinsic noise, which has the same effect in both models.

IV. DISCUSSION

We presented an agent-based framework comprising an exact simulation algorithm and an analytical approach to study the cell size dependence of gene expression noise across growing and dividing cell populations. The key feature distinguishing our ABM approach from previous agent-based studies is that we explicitly account for the coupling of intracellular reaction rates with cell size enabling concentration homeostasis. We find that while gene expression noise depends on cell size, it also features an intricate history dependence on birth size, a fact that is captured by ABMs but not always by their effective counterparts. Alongside the theory, we provided an exact simulation framework (Box 1) generalising previous single-cell simulation algorithms[4,17,34,63,64] towards snapshot-distributions across a growing cell population with an arbitrary size dependence of the gene regulatory interactions.

Our theory provides explicit conditions ((24) and Theorem 1) for traditional modelling approaches based on the chemical master equation, such as EDMs and ENMs, to be valid also in the agent-based case. These conditions rest on network topology and size-scaling of the reaction rates and they do not require solving the chemical master equation or stochastic simulations. Specifically, our conditions guarantee that certain network topologies generate gene expression distributions that, when conditioned on cell size, are entirely independent of extrinsic noise sources such as cell size control and division noise. They thus reveal whether a network embedded in a growing cell can be insulated against cell cycle and growth noise, an important feature that can guide the design of synthetic circuits.

Although we focused on population snapshots, our network conditions apply to isolated lineages as observed in the mother machine[65] as well. In particular, they ensure that size-dependent gene expression statistics of specific network architectures agree between lineages and across populations despite different levels of cell size variability, cell cycle heterogeneity and whether cells compete for growth or not. We have shown that in the absence of these conditions, cell size scaling of expression noise differs dramatically between EDMs and ABMs of cell populations, both quantitatively and qualitatively. Interestingly, the ENM remains valid for total concentration noise in mother-machine lineages despite its failure to accurately predict its cell size scaling and the corresponding total noise in molecule numbers.

A limitation of our study is that we assumed growth rate to be independent of gene expression dynamics. Previous studies[3], however, found that total noise of concentrations in mother-machine-like lineages follows closely the ENM prediction with expression-dependent growth. Nevertheless, it may be expected that selection plays a pronounced role in populations where cells compete for growth unlike in mother-machine lineages, which in turn may lead to significant deviations of ENMs from ABMs[36,66,67].

Despite the limitations of our study, we found that effective master equation models closely approximate the total noise statistics of ABMs, such as the noise-mean relationships of proteins, under the mass-action scaling

assumption (6). In fact, ENMs deviate at most 8% from ABM's total noise prediction in terms of coefficients of variation. To achieve such modest deviations, however, one requires feedback regulation with ultrasensitivity or regimes close to critical points, but ENMs fare much better for simple gene expression networks as the one in (13). Our results therefore reinstate the validity of effective models, and thus significantly extend the scope of state-of-the-art master equation methods to a broad range of single-cell analyses in growing cell populations.

CODE AVAILABILITY

The implementation of the First-Division Algorithm (Box 1) in Julia is available at `github.com/pthomaslab/fda`.

FUNDING

This work has been supported by a UKRI Future Leaders Fellowship (MR/T018429/1) and the EPSRC Centre for Mathematics of Precision Healthcare (EP/N014529/1).

REFERENCES

[1] D. J. Kiviet, P. Nghe, N. Walker, S. Boulineau, V. Sunderlikova, and S. J. Tans, "Stochasticity of metabolism and growth at the single-cell level," Nature **514**, 376–379 (2014).
[2] F. J. Bruggeman and B. Teusink, "Living with noise: on the propagation of noise from molecules to phenotype and fitness," Curr Opin Syst Biol **8**, 144–150 (2018).
[3] P. Thomas, G. Terradot, V. Danos, and A. Y. Weiße, "Sources, propagation and consequences of stochasticity in cellular growth," Nat Commun **9**, 1–11 (2018).
[4] F. Bertaux, S. Marguerat, and V. Shahrezaei, "Division rate, cell size and proteome allocation: impact on gene expression noise and implications for the dynamics of genetic circuits," Royal Soc Open Sci **5**, 172234 (2018).
[5] C. A. Vargas-Garcia, K. R. Ghusinga, and A. Singh, "Cell size control and gene expression homeostasis in single-cells," Curr Opin Syst Biol **8**, 109–116 (2018).
[6] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, Vol. 1 (Elsevier, 1992).
[7] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," J Phys Chem **81**, 2340–2361 (1977).
[8] D. T. Gillespie, "A rigorous derivation of the chemical master equation," Physica A **188**, 404–425 (1992).
[9] H. H. McAdams and A. Arkin, "Stochastic mechanisms in gene expression," Proc Natl Acad Sci **94**, 814–819 (1997).
[10] M. Thattai and A. Van Oudenaarden, "Intrinsic noise in gene regulatory networks," Proc Natl Acad Sci **98**, 8614–8619 (2001).
[11] P. S. Swain, M. B. Elowitz, and E. D. Siggia, "Intrinsic and extrinsic contributions to stochasticity in gene expression," Proc Natl Acad Sci **99**, 12795–12800 (2002).
[12] X. Lei, W. Tian, H. Zhu, T. Chen, and P. Ao, "Biological sources of intrinsic and extrinsic noise in cI expression of lysogenic phage lambda," Sci Rep **5**, 13597 (2015).
[13] M. K. Tonn, P. Thomas, M. Barahona, and D. A. Oyarzún, "Stochastic modelling reveals mechanisms of metabolic heterogeneity," Commun Biol **2**, 1–9 (2019).
[14] B. P. Ingalls, *Mathematical modeling in systems biology: an introduction* (MIT press, 2013).
[15] D. A. Charlebois and G. Balázsi, "Modeling cell population dynamics," In Silico Biol **13**, 21–39 (2019).
[16] J. Lin and A. Amir, "Homeostasis of protein and mRNA concentrations in growing cells," Nat Commun **9**, 1–11 (2018).
[17] X.-M. Sun, A. Bowman, M. Priestman, F. Bertaux, A. Martinez-Segura, W. Tang, C. Whilding, D. Dormann, V. Shahrezaei, and S. Marguerat, "Size-dependent increase in RNA Polymerase II initiation rates mediates gene expression scaling with cell size," Curr Biol **30**, 1217–1230 (2020).
[18] A. Hilfinger and J. Paulsson, "Separating intrinsic from extrinsic fluctuations in dynamic biological systems," Proc Natl Acad Sci **108**, 12167–12172 (2011).
[19] C. G. Bowsher and P. S. Swain, "Identifying sources of variation and the flow of information in biochemical networks," Proc Natl Acad Sci **109**, E1320–E1328 (2012).
[20] H. Kempe, A. Schwabe, F. Crémazy, P. J. Verschure, and F. J. Bruggeman, "The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise," Mol Biol Cell **26**, 797–804 (2015).
[21] T. Blasi, F. Buettner, M. K. Strasser, C. Marr, and F. J. Theis, "cgcorrect: a method to correct for confounding cell–cell variation due to cell growth in single-cell transcriptomics," Phys Biol **14**, 036001 (2017).
[22] N. Nordholt, J. Van Heerden, R. Kort, and F. J. Bruggeman, "Effects of growth rate and promoter activity on single-cell protein expression," Sci Rep **7**, 1–11 (2017).
[23] R. Kafri, J. Levy, M. B. Ginzberg, S. Oh, G. Lahav, and M. W. Kirschner, "Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle," Nature **494**, 480–483 (2013).
[24] K. Kuritz, D. Stöhr, D. S. Maichl, N. Pollak, M. Rehm, and F. Allgöwer, "Reconstructing temporal and spatial dynamics from single-cell pseudotime using prior knowledge of real scale cell densities," Sci Rep **10**, 1–10 (2020).
[25] C. Zopf, K. Quinn, J. Zeidman, and N. Maheshri, "Cell-cycle dependence of transcription dominates noise in gene expression," PLoS Comput Biol **9**, e1003161 (2013).
[26] N. Walker, P. Nghe, and S. J. Tans, "Generation and filtering of gene expression noise by the bacterial cell cycle," BMC Biol **14**, 11 (2016).
[27] A. Schwabe and F. J. Bruggeman, "Contributions of cell growth and biochemical reactions to nongenetic variability of cells," Biophys J **107**, 301–313 (2014).
[28] D. Antunes and A. Singh, "Quantifying gene expression variability arising from randomness in cell division times," J Math Biol **71**, 437–463 (2015).
[29] M. Soltani, C. A. Vargas-Garcia, D. Antunes, and A. Singh, "Intercellular variability in protein levels from stochastic expression and noisy cell cycle processes," PLoS Comput Biol **12**, e1004972 (2016).
[30] O. G. Berg, "A model for the statistical fluctuations of protein numbers in a microbial population," J Theor Biol **71**, 587–603 (1978).
[31] N. Brenner, K. Farkash, and E. Braun, "Dynamics of protein distributions in cell populations," Phys Biol **3**, 172 (2006).
[32] I. G. Johnston, B. Gaal, R. P. das Neves, T. Enver, F. J. Iborra, and N. S. Jones, "Mitochondrial variability as a source of extrinsic cellular noise," PLoS Comput Biol **8**, e1002416 (2012).

[33] R. Luo, L. Ye, C. Tao, and K. Wang, "Simulation of E. coli gene regulation including overlapping cell cycles, growth, division, time delays and noise," PloS One **8**, e62380 (2013).

[34] D. Gomez, R. Marathe, V. Bierbaum, and S. Klumpp, "Modeling stochastic gene expression in growing cells," J Theor Biol **348**, 1–11 (2014).

[35] I. G. Johnston and N. S. Jones, "Closed-form stochastic solutions for non-equilibrium dynamics and inheritance of cellular components over many cell divisions," Proc R Soc A **471**, 20150050 (2015).

[36] P. Thomas, "Making sense of snapshot data: ergodic principle for clonal cell populations," J Royal Soc Interface **14**, 20170467 (2017).

[37] P. Thomas, "Intrinsic and extrinsic noise of gene expression in lineage trees," Sci Rep **9**, 474 (2019).

[38] C. A. Nieto, C. A. V. Garcia, C. Sanchez, J. C. Arias-Castro, and J. M. Pedraza, "Correlation between protein concentration and bacterial cell size can reveal strategies of gene expression," Phys Biol **17**, 045002 (2020).

[39] J. Jedrak and A. Ochab-Marcinek, "Contributions to the 'noise floor' in gene expression in a population of dividing cells," Sci Rep **10**, 1–13 (2020).

[40] R. Perez-Carrasco, C. Beentjes, and R. Grima, "Effects of cell cycle variability on lineage and population measurements of messenger RNA abundance," J Royal Soc Interface **17**, 20200360 (2020).

[41] R. Dessalles, V. Fromion, and P. Robert, "Models of protein production along the cell cycle: An investigation of possible sources of noise," Plos One **15**, e0226016 (2020).

[42] T. G. Kurtz, "Strong approximation theorems for density dependent Markov chains," Stoch Process Their Appl **6**, 223–240 (1978).

[43] J. Elf and M. Ehrenberg, "Fast evaluation of fluctuations in biochemical networks with the linear noise approximation," Genome Res **13**, 2475–2484 (2003).

[44] P. Thomas, H. Matuschek, and R. Grima, "Computation of biochemical pathway fluctuations beyond the linear noise approximation using iNA," in *2012 IEEE International Conference on Bioinformatics and Biomedicine* (IEEE, 2012) pp. 1–5.

[45] M. Voliotis, P. Thomas, R. Grima, and C. G. Bowsher, "Stochastic simulation of biomolecular networks in dynamic environments," PLoS Comput Biol **12**, e1004923 (2016).

[46] M. Osella, E. Nugent, and M. C. Lagomarsino, "Concerted control of Escherichia coli cell division," Proc Natl Acad Sci **111**, 3431–3435 (2014).

[47] S. Taheri-Araghi, S. Bradde, J. T. Sauls, N. S. Hill, P. A. Levin, J. Paulsson, M. Vergassola, and S. Jun, "Cell-size control and homeostasis in bacteria," Curr Biol **25**, 385–391 (2015).

[48] Y. Tanouchi, A. Pai, H. Park, S. Huang, R. Stamatov, N. E. Buchler, and L. You, "A noisy linear map underlies oscillations in cell size and gene expression in bacteria," Nature **523**, 357–360 (2015).

[49] J. T. Sauls, D. Li, and S. Jun, "Adder and a coarse-grained approach to cell size homeostasis in bacteria," Curr Opin Cell Biol **38**, 38–44 (2016).

[50] M. Priestman, P. Thomas, B. D. Robertson, and V. Shahrezaei, "Mycobacteria modify their cell size control under sub-optimal carbon sources," Front Cell Develop Biol **5**, 64 (2017).

[51] P. Thomas, "Analysis of cell size homeostasis at the single-cell and population level," Front Phys **6**, 64 (2018).

[52] H. J. Heijmans, "On the stable size distribution of populations reproducing by fission into two unequal parts," Math Biosci **72**, 19–50 (1984).

[53] P. Thomas, "Single-cell histories in growing populations: relating physiological variability to population growth," bioRxiv , 100495 (2017).

[54] J. Peccoud and B. Ycart, "Markovian modeling of gene-product synthesis," Theor Popul Biol **48**, 222–234 (1995).

[55] V. Shahrezaei and P. S. Swain, "Analytical distributions for stochastic gene expression," Proc Natl Acad Sci **105**, 17256–17261 (2008).

[56] P. Bokes, J. R. King, A. T. Wood, and M. Loose, "Exact and approximate distributions of protein and mrna levels in the low-copy regime of gene expression," J Math Biol **64**, 829–854 (2012).

[57] C. W. Gardiner and S. Chaturvedi, "The Poisson representation. I. A new technique for chemical master equations," J Stat Phys **17**, 429–468 (1977).

[58] C. Gadgil, C. H. Lee, and H. G. Othmer, "A stochastic analysis of first-order reaction networks," Bull Math Biol **67**, 901–946 (2005).

[59] T. Jahnke and W. Huisinga, "Solving the chemical master equation for monomolecular reaction systems analytically," J Math Biol **54**, 1–26 (2007).

[60] L. Ham, D. Schnoerr, R. D. Brackston, and M. P. Stumpf, "Exactly solvable models of stochastic gene expression," J Chem Phys **152**, 144106 (2020).

[61] F. Maclean and R. Munson, "Some environmental factors affecting the length of Escherichia coli organisms in continuous cultures," Microbiology **25**, 17–27 (1961).

[62] A. Koch and M. Schaechter, "A model for statistics of the cell division process," Microbiology **29**, 435–454 (1962).

[63] T. Lu, D. Volfson, L. Tsimring, and J. Hasty, "Cellular growth and division in the Gillespie algorithm," IET Syst Biol **1**, 121–128 (2004).

[64] C. Blanco, C. Nieto, C. Vargas, and J. Pedraza, "PyEcoLib: a python library for simulating E. coli stochastic size dynamics," bioRxiv , 319152 (2020).

[65] P. Wang, L. Robert, J. Pelletier, W. L. Dang, F. Taddei, A. Wright, and S. Jun, "Robust growth of Escherichia coli," Curr Biol **20**, 1099–1103 (2010).

[66] T. Nozoe, E. Kussell, and Y. Wakamoto, "Inferring fitness landscapes and selection on phenotypic states from single-cell genealogical data," PLoS Genet **13**, e1006653 (2017).

[67] M. Ciechonska, M. Sturrock, A. Grob, G. Larrouy-Maumus, V. Shahrezaei, and M. Isalan, "Ohm's law for emergent gene expression under fitness pressure," bioRxiv , 693234 (2019).

[68] C. Gardiner, *Stochastic methods*, Vol. 4 (Springer Berlin, 2009).

Appendix A: Validity conditions for EDMs and ENMs

Theorem 1. *Assume that the partitioning kernel $B(x|x',\theta)$ is binomial with probability θ given by the ratio of daughter birth size and mother division size. The stationary solution of the EDM (4) is also a solution of the ABM (23):*

$$\Pi(x|s,s_0) = \Pi_{EDM}(x|s), \tag{A1}$$

if and only the EDM (4) admits a solution with generating function of the form

$$G_{EDM}(z|s) = \sum_x z^x \Pi_{EDM}(x|s) = F(s(z-1)), \tag{A2}$$

for any function F independent of s and the solution is independent of the birth size s_0.

The utility of the theorem is that its condition (A2) can be checked without solving the chemical master equation. We demonstrate this aspect for a general reaction network of the form (1) with mass-action propensities (6) whose generating function (see Chapter 7 in[68]) obeys:

$$\begin{aligned}&\alpha s \frac{\partial}{\partial s} G(z|s,s_0)\\ &= \sum_{r=1}^{R} k_r s^{1-|\nu_r^-|}\left(z^{\nu_r^+} - z^{\nu_r^-}\right)\partial_z^{\nu_r^-} G(z|s,s_0)\end{aligned} \tag{A3}$$

Substituting $G(z|s,s_0) = F(s(z-1))$ gives

$$\begin{aligned}&\alpha x \cdot \nabla F(x)\\ &= \sum_{r=1}^{R} k_r s\left((x+s)^{\nu_r^+} s^{-|\nu_r^+|} - (x+s)^{\nu_r^-} s^{-|\nu_r^-|}\right)\partial_x^{\nu_r^-} F(x).\end{aligned}$$

It can now be seen that the right-hand side of the above equation is independent of s if either (i) $|\nu_r^-| = 0$ and $|\nu_r^+| = 1$, (ii) $|\nu_r^-| = 1$ and $|\nu_r^+| = 0$, or (iii) $|\nu_r^-| = |\nu_r^+| = 1$. Thus the EDM and the ABM solutions coincide for mass-action reaction networks (1) when they only comprise only the monomolecular reactions given in (26).

The proof of the theorem is divided in three steps below, which shows that the binomial partitioning assumption cannot be removed due to the biological constraint of molecule number conservation at cell division.

1. Dynamics invariant of birth size

Assume that, the partition kernel depends only the inherited size fraction

$$B(x|x',s',s_0) = B(x|x',\theta),$$

where $\theta = s_0/s'$. Then we say that Π is invariant against partitioning, if

$$\Pi(x|\theta s, s_0) = \sum_{x'} B(x|x',\theta)\Pi(x'|s,s_0), \tag{A4}$$

a central condition that can be checked from the reaction kinetics. It then follows that the conditional distribution $\Pi(x|s,s_0)$ is independent of birth size s_0. This can be verified using (A4) in the boundary condition (23b), which leads to

$$\Pi(x|s_0,s_0) = \int_0^\infty \mathrm{d}s' \int_0^{s'} \mathrm{d}s_0'\, \rho(s',s_0'|s_0)\Pi(x|s_0,s_0').$$

This implies that $\Pi(x|s,s_0)$ must be independent of birth size

$$\Pi(x|s,s_0) = \Pi(x|s).$$

In the following, we show that under condition (A4) $\Pi(x|s)$ coincides with the EDM solution.

2. Effective dilution model

Let us denote the generating function of the partitioning kernel by $G_B(z|x',\theta) = \sum_x z^x B(x|x',\theta)$ such that the invariance condition (A4) becomes

$$G(z|\theta s) = \sum_{x'} G_B(z|x',\theta)\Pi(x'|s). \tag{A5}$$

Assume that additionally it holds that

$$\theta\partial_\theta G_B(z|x',\theta) = \sum_{i=1}^{N}(z_i-1)\partial_{z_i} G_B(z|x',\theta). \tag{A6}$$

Differentiating Eq. (A5) with respect to θ then gives

$$\begin{aligned}\theta\partial_\theta G(z|\theta s) &= \sum_{x'} \theta\partial_\theta G_B(z|x',\theta)\Pi(x'|s)\\ &= \sum_{i=1}^{N}(z_i-1)\partial_{z_i} G(z|\theta s),\end{aligned} \tag{A7}$$

where in the last line we have used assumption (A6), i.e., $G_B(z|x',\theta) = (1-\theta+\theta z)^{x'}$. Changing variables ($\theta s \to s$) in (A7) yields

$$s\partial_s G(z|s) = \sum_{i=1}^{N}(z_i-1)\frac{\partial}{\partial z_i} G(z|s),$$

or equivalently the effective dilution model

$$\begin{aligned}&\left(\alpha s\frac{\partial}{\partial s}\right)\Pi(x|s)\\ &= -\alpha\sum_{i=1}^{N}\left[(x_i+1)\Pi(x_1,...,x_i+1,...,x_N|s) - x_i\Pi(x|s)\right]\\ &= -\mathbb{D}\Pi(x|s).\end{aligned} \tag{A8}$$

Using the above relation, we see that (23a) coincides with (4) and (A1) follows.

3. Necessity of binomial partitioning

Finally, we show that condition (A6) required for the validity of the EDM implies independent binomial partitioning. (A6) is a linear PDE that can be solved using the method of characteristics, which leads to

$$\theta\frac{\partial z_i}{\partial\theta} = (1-z_i), \quad \theta\frac{\partial G_B}{\partial\theta} = 0.$$

The general solution is $G_B(z|x',\theta) = J(1-\theta+\theta z)$ where the function J is fixed by the condition that for $\theta = 1$ all molecules are partitioned deterministically, i.e., $J(z) = z^{x'}$. Hence, we obtain

$$G_B(z|x',\theta) = (1-\theta+\theta z)^{x'},$$

which corresponds to independent binomial partitioning, and (A5) is then equivalent to

$$G(z|\theta s) = G((1-\theta)+\theta z|s). \tag{A9}$$

Finally, we show that condition (A9) is equivalent to $G(z|\theta s) = F(s(z-1))$. Specifically, expanding (A9) around $z = 1$ and identifying the series coefficients with the factorial moments $\mu_n(s) = E_\Pi[x(x-1)\dots(x-n)|s]$, we find that the factorial moments are homogeneous functions of order $|n| = \sum_i n_i$: $\mu_n(\theta s) = \theta^{|n|}\mu_n(s)$. Then by Euler's homogeneous function theorem, it follows that the factorial moments with index n, satisfy $s\frac{\partial}{\partial s}\mu_n(s) = |n|\mu_n(s)$ and hence $\mu_n(s) = s^{|n|}\mu_n(1)$. This implies that the generating function is

$$G(z|s) = \sum_n s^{|n|}\mu_n(1)(z-1)^n = F(s(z-1)),$$

with $F(x) = \sum_n x^n\mu_n(1)$ independent of s which concludes the proof of Theorem 1.

Appendix B: Approximation of birth size moments

Here derive an analytical approximation (34b) for the conditional birth size moments. We start by rewriting $E_\Pi[s_0|s]$ in terms of the backward lineage distribution $\psi_{\rm bw}$ using Eq. (21):

$$\begin{aligned} E_\Pi[s_0|s] &= \int_0^s \mathrm{d}s_0\, s_0\Pi(s_0|s) \\ &= \frac{\int_0^s \mathrm{d}s_0\, s_0\Pi(s_0,s)}{\int_0^s \mathrm{d}s_0\,\Pi(s_0,s)} = \frac{E_\psi[s_0\Phi(s|s_0)1_{s_0\le s}]}{E_\psi[\Phi(s|s_0)1_{s_0\le s}]}. \end{aligned} \tag{B1}$$

We now apply matched asymptotic expansion to this expression.

1. Large cell asymptotics

For large cells $s \gg s_0$, we can extend the range of integration in Eq. (B1) and compute the expectation value as follows

$$\begin{aligned} E_\psi[f(s_0,s)] &= \int_0^\infty \mathrm{d}s_0\psi_{bw}(s_0)f(s_0,s) \\ &= \int_0^\infty \mathrm{d}s_0\int_{-\infty}^\infty \frac{\mathrm{d}k}{2\pi}e^{-ik(s_0-\bar{s}_0)}\left(1-\frac{k^2\sigma^2}{2}\right)f(s_0,s) + O(\sigma^3) \\ &= f(\bar{s}_0,s) + \frac{\sigma^2}{2}\frac{\partial^2 f(\bar{s}_0,s)}{\partial\bar{s}_0^2} + O(\sigma^3). \end{aligned}$$

Using $f(s_0,s) = s_0\Phi(s|s_0)$ and $f(s_0,s) = \Phi(s|s_0)$ in Eq. (B1), the conditional moments of birth size can be approximated by

$$\begin{aligned} E_\Pi^{\rm large}[s_0|s] &= \bar{s}_0\left\{1+\frac{2\sigma^2}{\bar{s}_0}\frac{\partial\ln\Phi(s|\bar{s}_0)}{\partial\bar{s}_0}\right\} + O(\sigma^3) \\ &= \left\{\bar{s}_0 + 2a\sigma^2\gamma(s-a\bar{s}_0)\right\} + O(\sigma^3), \end{aligned}$$

where the last equality follows from $\gamma(s,s_0) = \gamma(s-as_0)$ for the linear cell size control model (18), and $\bar{s}_0$ and σ are the mean and standard deviation of the backward lineage distribution ψ_{bw} given by Eqs. (34c).

2. Small cell asymptotics

Next we consider small cells by noting that $\Phi(s|s_0)$ is practically constant when $s \approx s_0$, the integral in (B1) can be approximated by

$$E_\Pi[s_0|s] \approx \frac{E_\psi[s_0 1_{s_0\le s}]}{E_\psi[1_{s_0\le s}]}, \tag{B2}$$

Assuming that ψ, is approximately Gaussian with mean $\bar{s}_0$ and variance σ^2, we find that near $s \approx \bar{s}_0$, we have find

$$\begin{aligned} E_\psi[s_0 1_{s_0\le s}] &\approx \frac{1}{2}\bar{s}_0\left(1+\mathrm{erf}\left(\frac{s-\bar{s}_0}{\sqrt{2}\sigma}\right)\right) + \frac{\sigma e^{-\frac{(s-\bar{s}_0)^2}{2\sigma^2}}}{\sqrt{2\pi}} \\ E_\psi[1_{\bar{s}_0\le s}] &\approx \frac{1}{2}\left(1+\mathrm{erf}\left(\frac{s-\bar{s}_0}{\sqrt{2}\sigma}\right)\right) \end{aligned} \tag{B3}$$

and hence

$$E_\Pi^{\rm small}[s_0|s] = \bar{s}_0 - \frac{\sqrt{\frac{2}{\pi}}\sigma e^{-\frac{(s-\bar{s}_0)^2}{2\sigma^2}}}{1+\mathrm{erf}\left(\frac{s-\bar{s}_0}{\sqrt{2}\sigma}\right)} + O\left(\sigma^3\right),$$

which is accurate to order σ^3.

3. Global asymptotics

The two asymptotic solutions can be matched at the boundary layer. Since

$$\lim_{s\to\infty} E_\Pi^{\text{small}}[s_0|s] = \lim_{s\to s_0} E_\Pi^{\text{large}}[s_0|s] = \bar{s}_0,$$

the uniformly valid matched asymptotic expansion is

$$E_\Pi[s_0|s] \approx E_\Pi^{\text{small}}[s_0|s] + E_\Pi^{\text{large}}[s_0|s] - \bar{s}_0,$$

which gives Eq. (34b).

Appendix C: Analytical solutions and error bounds using the linear noise approximation

We begin by outlining the solution of (28). Defining $\tilde{\Sigma}(s,s_0) = U^{-1}\Sigma(s,s_0)U^{-\dagger}$, (28) becomes

$$\alpha s\partial_s\tilde{\Sigma}_{ij} = (\lambda_i+\lambda_j^*)\tilde{\Sigma}_{ij} + s\tilde{\mathcal{D}}_{ij} \tag{C1}$$

$$4\tilde{\Sigma}_{ij}(s_0,s_0) = \tilde{\Sigma}_{ij}(2s_0,s_0) + 2s_0\tilde{X}_{ij}. \tag{C2}$$

Eq. (C1) has the solutions

$$\tilde{\Sigma}_{ij}(s,s_0) = c_{ij}s^{\frac{\lambda_i+\lambda_j^*}{\alpha}} + \frac{\tilde{\mathcal{D}}_{ij}s}{\alpha\left(1-\frac{\lambda_i+\lambda_j^*}{\alpha}\right)},$$

where the constants c_{ij} are fixed using the boundary condition (C2) which gives Eq. (30) of the main text.

Next, we will deduce some results implied by Eq. (28) for reaction networks (1) with a single species. In this case, the solution (30) reduces to

$$\Sigma(s,s_0) = \frac{s}{(\alpha-2\mathcal{J})}\left(\frac{2\left(\mathcal{D}+2\mathcal{J}\bar{X}-\alpha\bar{X}\right)}{\left(2^{\frac{2\mathcal{J}}{\alpha}}-4\right)}\left(\frac{s_0}{s}\right)^{1-\frac{2\mathcal{J}}{\alpha}} + \mathcal{D}\right).$$

Using this result in Eq. (37), we obtain the concentration variance

$$\bar{\Sigma} = \frac{\beta\left(\coth\left(\frac{1}{2}\beta\ln 2\right)+3\right)}{3(\beta+1)}\frac{\alpha\bar{X}}{\Omega\xi} + \frac{\beta\left(3\beta-\coth\left(\frac{1}{2}\beta\ln 2\right)\right)}{3\left(\beta^2-1\right)}\frac{\mathcal{D}}{\Omega\xi}, \tag{C3}$$

where $\beta = \frac{\xi}{\alpha}$ and $\xi = 2\alpha - 2\mathcal{J} = -2\mathcal{J}_d$.

Noting that (C3) increases monotonically with x and that the ENM yields $\bar{\Sigma}_{\text{EDM}} = \frac{\alpha\bar{X}}{\Omega\xi} + \frac{\mathcal{D}}{\Omega\xi}$, we obtain the bounds

$$\frac{2}{3\ln 2}\bar{\Sigma}_{\text{ENM}} \le \bar{\Sigma} \le \bar{\Sigma}_{\text{ENM}}\left(1+\frac{1}{3}\frac{\alpha\bar{X}}{\mathcal{D}+\alpha\bar{X}}\right) \le \frac{7}{6}\bar{\Sigma}_{\text{ENM}},$$

where the last inequality arises from the fact that $\mathcal{D}(\bar{X}) \ge \alpha\bar{X}$, which implies Eq. (39). The minimum is achieved when $\beta\to 0$ ($\mathcal{J}=\alpha$) and the maximum is achieved when $\mathcal{D}=\alpha\bar{X}$ and $\beta\to\infty$ ($|\mathcal{J}|\gg\alpha$).

Similarly, we see that the intrinsic molecule number variance is

$$\Sigma_{\text{ABM}}^{\text{int}} = \frac{\left(2^\beta-2\right)\beta}{\left(2^\beta-1\right)(\beta-1)\ln 4}\frac{\Omega\alpha\bar{X}}{\xi} + \frac{x\left(\left(2^\beta-1\right)\beta\ln 4 - 2^\beta(1+\ln 4)+2+\ln 4\right)}{\left(2^\beta-1\right)(\beta-1)^2\ln 4}\frac{\Omega\mathcal{D}}{\xi}.$$

Since the above equation decreases monotonically with x, we obtain

$$\frac{1+\ln 4}{\ln 16}\Sigma_{\text{ENM}}^{\text{int}} \le \Sigma_{\text{ABM}}^{\text{int}} \le \Sigma_{\text{ENM}}^{\text{int}}\frac{1}{2\ln^2(2)}.$$

Furthermore, taking also extrinsic variance Σ^{ENM} into account, it follows that $\frac{1+\ln 4}{\ln 16}\Sigma_{\text{ENM}} \le \text{Cov}_\Pi[x] \le \Sigma_{\text{ENM}}\frac{1}{2\ln^2(2)}$ and hence Eq. (40).

In the following, we give two examples of common autoregulatory circuits of the form

$$\varnothing \xrightarrow{sf(\bar{X})} b\times P,$$

where $\bar{X}$ is the protein concentration and b is the burst size following an arbitrary discrete distribution β, that explicitly achieve these bounds.

a. Negative feedback We consider bursty protein expression with negative feedback with rate

$$f(\bar{X}) = k_0/((\bar{X}/K)^n+1), \tag{C4}$$

where k_0 is the burst frequency, K is the saturation constant and n is the Hill exponent. The model satisfies the rate equation $\alpha\bar{X} = E_\beta[b]f(\bar{X})$. The Jacobian of this model is $\mathcal{J}(\bar{X}) = -n\alpha(k_0-\alpha\bar{X})$ and the diffusion coefficient satisfies $\frac{\mathcal{D}}{\alpha\bar{X}} = E_\beta[b](1+\text{CV}_\beta[b]^2) \ge 1$, since b has a discrete distribution β. Thus this example achieves the maximum amplification of the concentration CV of around 8% when $n\to\infty$ ($|\mathcal{J}|\gg\alpha$).

b. Positive feedback Consider instead an example of positive feedback with rate equation

$$f(\bar{X}) = \delta + \frac{k_0\bar{X}^n}{\bar{X}^n+K^n}. \tag{C5}$$

We find that the parameter choice $n=1$, $\delta\ll k_0$, $K\alpha\approx E_\beta[b]k_0$, achieves the lower error bound of around 2%.