Coordination of gene expression noise with cell size: analytical results for agent-based models of growing cell populations

Philipp Thomas[1, a)] and Vahid Shahrezaei[1]
Department of Mathematics, Imperial College London, UK

The chemical master equation and the Gillespie algorithm are widely used to model the reaction kinetics inside living cells. It is thereby assumed that cell growth and division can be modelled 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. We find that the solution of the chemical master equation – including static extrinsic noise – exactly agrees with the agent-based formulation when the network under study exhibits *stochastic concentration homeostasis*, a novel condition that generalises concentration homeostasis in deterministic systems to higher order moments and distributions. We illustrate stochastic concentration homeostasis for a range of common gene expression networks. When this condition is not met, we demonstrate by extending the linear noise approximation to agent-based models that the dependence of gene expression noise on cell size can qualitatively deviate from the chemical master equation. Surprisingly, the total noise of the agent-based approach can still be well approximated by extrinsic noise models.

I. INTRODUCTION

Cells must continuously synthesise molecules to grow and divide. At a single cell level, gene expression and cell size are coordinated but heterogeneous which can drive phenotypic variability and decision making in cell populations[1–5]. The interplay between these sources of cell-to-cell variability is 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 only a small number of molecules is present at any time 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 cell growth 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 EDMs 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 and dilution, and 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 describe cells that 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 master equations including cell cycle dynamics[4,17,30–35] and agent-based approaches including age-structure of growing populations[35–40]. The essence of *agent-based models* (ABMs) is that each cell in a population is represented by an agent 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. The unprecedented detail of these models must cast doubt on the predictions of master equation models (EDMs and ENMs) in which growth and division are modelled by effective dilution reactions. Yet it is presently unclear why these effective models have fared reasonably well in predicting gene expression noise reported by single-cell experiments[10,17,41].

Nevertheless, most ABMs still ignore cell size, a major physiological factor affecting both intracellular reactions

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

and cell division dynamics alike. Since cell size varies at least two-fold as required by size homeostasis in a growing population, and it scales some reaction rates as required by concentration homeostasis, it is expected that cell 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.

The outline of the paper is as follows. First, we explain the analytical framework for EDMs, ENMs and ABMs (II). Then we introduce the concept of *stochastic concentration homeostasis*, a rigorous condition under which the chemical master equations of the EDM and ENM agree exactly with the ABM (Sec. III A). This new condition is met by some but not all common models of gene expression. We show that when these conditions are not met, the effective models agree with the ABM only on average (Sec. III B). To address this problem, we propose a comprehensive theoretical framework extending the linear noise approximation to agent-based dynamics with which we quantify cell size scaling of gene expression in growing cells (Sec. III C). Our findings indicate that the EDM can qualitatively fail to predict this dependence but our novel approximation method accurately describes gene expression noise in the presence of cell size control variations and division errors. We further show that ENMs present surprisingly accurate approximations for the total noise statistics (Sec. III D).

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 network 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 of the r^{th} reaction. In the following, we outline deterministic, effective dilution and extrinsic noise 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 concentration homeostasis

Deterministically, the vector of molecular concentrations $\bar{X} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_N)^T$ is governed by rate equation models in balanced growth conditions. 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, $f_r(\bar{X})$ are macroscopic reaction-rate functions and α is the exponential growth rate of cells determining the dilution rate due to growth. Since these quantities are independent of cell size, the balanced growth condition (2) implies concentration homeostasis in rate equation models.

2. Effective dilution model

The chemical master equation[6] and equivalently the stochastic simulation algorithm[7] are state-of-the-art stochastic models of reaction kinetics 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 of rate α that model dilution of molecules due to cell growth:

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

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 $x = (x_1, x_2, \ldots, x_N)^T$ of the species S 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. Such effective models are motivated through the fact[6,42] that when the microscopic propensities w_r are linked to the macroscopic rate functions f_r of the rate equation models via mass-action kinetics

$$w_r(x, s) \approx s f_r(X), \tag{6}$$

where $X = x/s$ is the concentration, the mean concentrations of EDMs follow the concentrations $\bar{X}$ of the rate equations (2) (see Sec. II A 4).

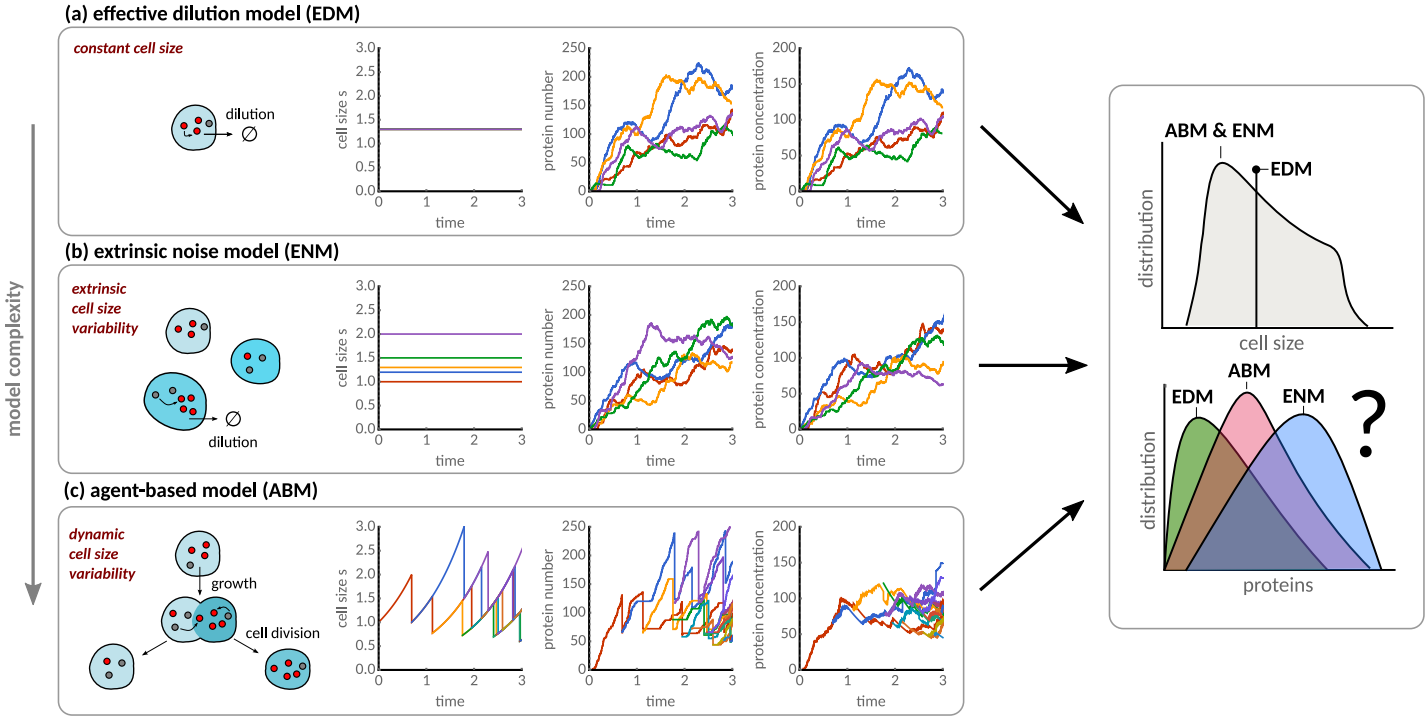

FIG. 1. **Modelling approaches for cell size dependence of gene expression. (a)** The *effective dilution model* describes cells at constant size with intracellular reactions coupled to effective dilution reactions. **(b)** The *extrinsic noise model* incorporates static cell size variability as a source 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.

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 $X = x/s$,

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

and analogous expressions for the molecule number distributions.

4. Analytical solutions and noise decomposition

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 this approximation, the mean concentrations are approximated by the solution $\bar{X}$ of the rate equations (2) and the probability distribution $\Pi_{\mathrm{EDM}}(x|s)$ is approximated by a Gaussian. In the same limit, the covariance matrix Σ_Y can be decomposed into intrinsic and extrinsic components, Σ_Y^{int} and Σ_Y^{ext}, using the law of total variance[18,19]

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

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 $\Sigma_Y^{\mathrm{int}} = E_\Pi[\mathrm{Cov}_{\Pi_Y}[x|s]]$ and $\Sigma_Y^{\mathrm{ext}} = \mathrm{Cov}_\Pi[E_{\Pi_Y}[x|s]]$, where E_Π denotes the expectation value with respect to the distribution Π, and analogously for concentrations. The intrinsic components Σ_Y^{int} satisfy a Lyapunov equation called the linear noise approximation:

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

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 the rate equations (2) and $\mathcal{D}_d$ denotes the diffusion matrix obeying

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

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

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

, (12)

where the last cell follows from $\Sigma_{\text{ENM}}^{\text{ext}} = \text{Cov}_\Pi[E_\Pi[x|s]]$ with $E_\Pi[x|s] = s\bar{X}$.

As a concrete example, we consider transcription of mRNAs with a size-dependent transcription rate that are translated into stable proteins:

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

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]

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

where we also account for size-variability via Ω_Y and Σ_Y^{ext} given by Eqs. (10) and (12), respectively, and the parameters

$$\delta = 1 + \frac{k_{\text{dm}}}{\alpha}, \quad b = \frac{k_{\text{tl}}}{k_{\text{dm}} + \alpha}, \tag{16}$$

correspond to the ratio of mRNA and protein degradation/dilution rates and the translational burst size, respectively. From Eq. (15), (10) and (12), it is clear that size variation acts on the intrinsic noise component of molecule concentrations (via $E_\Pi[s^{-1}] \approx E_\Pi[s]^{-1}(1 + \text{CV}_\Pi^2[s])$) but the extrinsic noise component of molecule numbers (via Σ_Y^{ext} (12)).

B. Agent-based modelling

Little is known about the accuracy of EDMs and ENMs in predicting cellular noise in growing populations. In the following, we introduce an agent-based modelling approach that serves as a gold standard to assess the validity of these effective models. The ABM represents 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.

The ABM simulation algorithm is given in Box 1, which combines the First-Division algorithm, previously introduced for agent-based cell populations[38], with the Extrande method adapted to simulate reaction networks embedded in a growing cell[45]. In the following, we describe the exact analytical framework with which we characterise the snapshot distributions that underlie such a population of agents.

Master equation for agent-based populations

We consider the number of cells $n(\tau, s, x, t)$ with age τ (time since the last division), cell size s and molecule counts x in a snapshot at time t, which evolves as

$$\underbrace{\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial \tau} + \frac{\partial}{\partial s}\alpha s + \bar{\gamma}(s,\tau)\right)}_{\text{growth}} n(\tau,s,x,t) = \underbrace{\mathbb{Q}(s)n(\tau,s,x,t)}_{\text{stochastic reactions}},$$

$$\underbrace{n(0,s,x,t)}_{\#\text{ newborn cells}} = 2\int_0^\infty \mathrm{d}\tau' \int_0^\infty \mathrm{d}s' \underbrace{B(s|s')}_{\text{division error}} \times \sum_{x'} \underbrace{B(x|x',s/s')}_{\text{partitioning of molecules}} \underbrace{\bar{\gamma}(s',\tau')n(\tau',s',x',t)}_{\#\text{ dividing cells}}, \tag{17}$$

and describes cell growth, stochastic reaction kinetics and a boundary condition for cell division that ensures that the number of newborn cells is twice the number of dividing cells after partitioning their size and molecular contents. These evolution equations have been derived in[38,39] for age-dependent snapshots but here we extend such agent-based models to include also cell size dynamics and size-dependent reaction dynamics. We allow for the following generalisations: (i) size increases exponentially in single cells, (ii) cells divide with rate $\bar{\gamma}(s,\tau)$ that is both size- and age-dependent, (iii) the transition matrix $\mathbb{Q}(s)$ of the molecular reactions depends on cell size s via the propensities (see definition after Eq. (4)), and (iv) the molecular partitioning kernel $B(x|x', s/s')$ depends on the inherited size fraction s/s' of the daughter cell. We now describe in detail how we model the individual noise sources associated with cell size control, division errors, and molecule partitioning.

a. Cell size control fluctuations. Recent studies[46,47] have shown that the distribution of sizes with which cells divide does not explicitly depend on cell age but on the birth size s_0. Assuming that $\bar{\gamma}(s,\tau) = \alpha s\gamma(s, s^{-\alpha\tau})$, where $\gamma(s, s_0)$ is the division rate per unit size (see also[48,49]), the division-size distribution is given by

$$\varphi(s_d|s_0) = \gamma(s_d, s_0)e^{-\int_{s_0}^{s_d} \mathrm{d}s\gamma(s,s_0)}. \tag{18}$$

As a concrete example of (18) we consider a model where the division size is linearly related to birth size[49–51]

$$s_d = as_0 + \Delta. \tag{19}$$

The division rate can be calculated from the distribution $\tilde{\varphi}(\Delta)$ of the noise term Δ in (19) via $\tilde{\gamma}(\Delta) = \tilde{\varphi}(\Delta)/(\int_\Delta^\infty \mathrm{d}u\tilde{\varphi}(u))$ and setting $\gamma(s,s_0) = \tilde{\gamma}(s-as_0)$, which gives the correct division-size distribution $\varphi(s_d|s_0) = \tilde{\varphi}(s_d - as_0)$ as expected. 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[46,47,52]. In the following, we will refer to $\mathrm{CV}_\varphi[\Delta]$ as the size-control noise.

b. Division errors. After 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 $\bar{\pi}(\theta)$ (see Box 1). This can be modelled using the division kernel

$$B(s_0|s') = \int_0^1 \mathrm{d}\theta\, \pi(\theta)\delta(\theta - s_0/s'),$$

where $\pi(\theta) = \frac{1}{2}\bar{\pi}(\theta) + \frac{1}{2}\bar{\pi}(1-\theta)$ including the case of asymmetric division. We will refer to $\mathrm{CV}_\pi[\theta]$ as the division error about the centre $E_\pi[\theta] = \frac{1}{2}$.

c. Molecule partitioning at cell division. The partitioning kernel $B(x|x',\theta)$ denotes the probability that a cell inherits x molecules from a total of x' molecules from its mother and this probability depends on the daughter's inherited size fraction θ. We assume that cells are sufficiently well mixed and each molecule is partitioned independently with probability θ such that the division kernel is binomial

$$B(x|x',\theta) = \prod_{i=1}^{N} \binom{x_i'}{x_i} \theta^{x_i}(1-\theta)^{x_i'-x_i}. \tag{20}$$

To make analytical progress, we assume that the population 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. To this end, we let $n(\tau,s,x,t) \propto e^{\alpha t}\Pi(s,\tau,x)$ and change variables from cell age τ to birth size s_0 such that $\Pi(s,s_0,x) = (\alpha s)^{-1}\Pi(s,\tau = \ln(s/s_0)/\alpha, x)$. We find that this transformation reduces the PDE (17) to an integro-ODE:

$$\left(\alpha + \frac{\partial}{\partial s}\alpha s + \alpha s\gamma(s,s_0)\right)\Pi(s,s_0,x) = \mathbb{Q}(s)\Pi(s,s_0,x) \tag{21a}$$

$$s_0\Pi(s_0,s_0,x) = 2\sum_{x'}\int_0^\infty \mathrm{d}s'\int_0^{s'}\mathrm{d}s_0'\, B(x|x',s_0/s')B(s_0|s')s'\gamma(s',s_0')\Pi(s',s_0',x') \tag{21b}$$

We finally characterise the marginal cell size distribution $\Pi(s,s_0)$ and the conditional molecule number distribution $\Pi(x|s,s_0)$ via Bayes' formula

$$\Pi(s,s_0) = \textstyle\sum_x \Pi(s,s_0,x), \quad \Pi(x|s,s_0) = \dfrac{\Pi(s,s_0,x)}{\Pi(s,s_0)}, \tag{22}$$

which together provide the full information about the population process.

Cell size distribution

The evolution of the size distribution $\Pi(s,s_0)$ is obtained by summing Eqs. (21) over all possible x, which yields:

$$\left(\alpha + \frac{\partial}{\partial s}\alpha s + \alpha s\gamma(s,s_0)\right)\Pi(s,s_0) = 0 \tag{23a}$$

$$s_0\Pi(s_0,s_0) = 2\int_0^\infty \mathrm{d}s'\int_0^{s'}\mathrm{d}s_0'\, B(s_0|s')s'\gamma(s',s_0')\Pi(s',s_0'). \tag{23b}$$

Eqs. (23) can be solved analytically

$$\Pi(s,s_0) = \frac{2}{Z}\psi_{\mathrm{bw}}(s_0)\Phi(s|s_0)\frac{1}{s^2}, \tag{24}$$

where $\psi_{\mathrm{bw}}(s_0)$ is the birth size distribution in a backward lineage (see[48] 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 for cells of a certain size

The conditional molecule number distribution $\Pi(x|s,s_0)$ gives the probability to find the molecule numbers x in a cell of size s that was born at size s_0 and satisfies

$$\alpha s\frac{\partial}{\partial s}\Pi(x|s,s_0) = \mathbb{Q}(s)\Pi(x|s,s_0), \tag{25a}$$

$$\Pi(x|s_0,s_0) = \sum_{x'}\int_0^\infty \mathrm{d}s'\int_0^{s'}\mathrm{d}s_0'\, B(x|x',s_0/s')\rho(s',s_0'|s_0)\Pi(x'|s',s_0'). \tag{25b}$$

Eqs. (25) follow directly from substituting Eq. (22) into (21) and using (18) and (23). The solution of these equations depends implicitly on the ancestral cell size distribution ρ,

$$\rho(s',s_0'|s_0) = \frac{1}{\psi_{\mathrm{bw}}(s_0)}\frac{s_0}{s'}B(s_0|s')\varphi(s'|s_0')\psi_{\mathrm{bw}}(s_0'), \tag{25c}$$

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

We here introduce the concept of *stochastic concentration homeostasis* (SCH) as a generalisation of concentration homeostasis in deterministic systems (see Sec. II A 1)

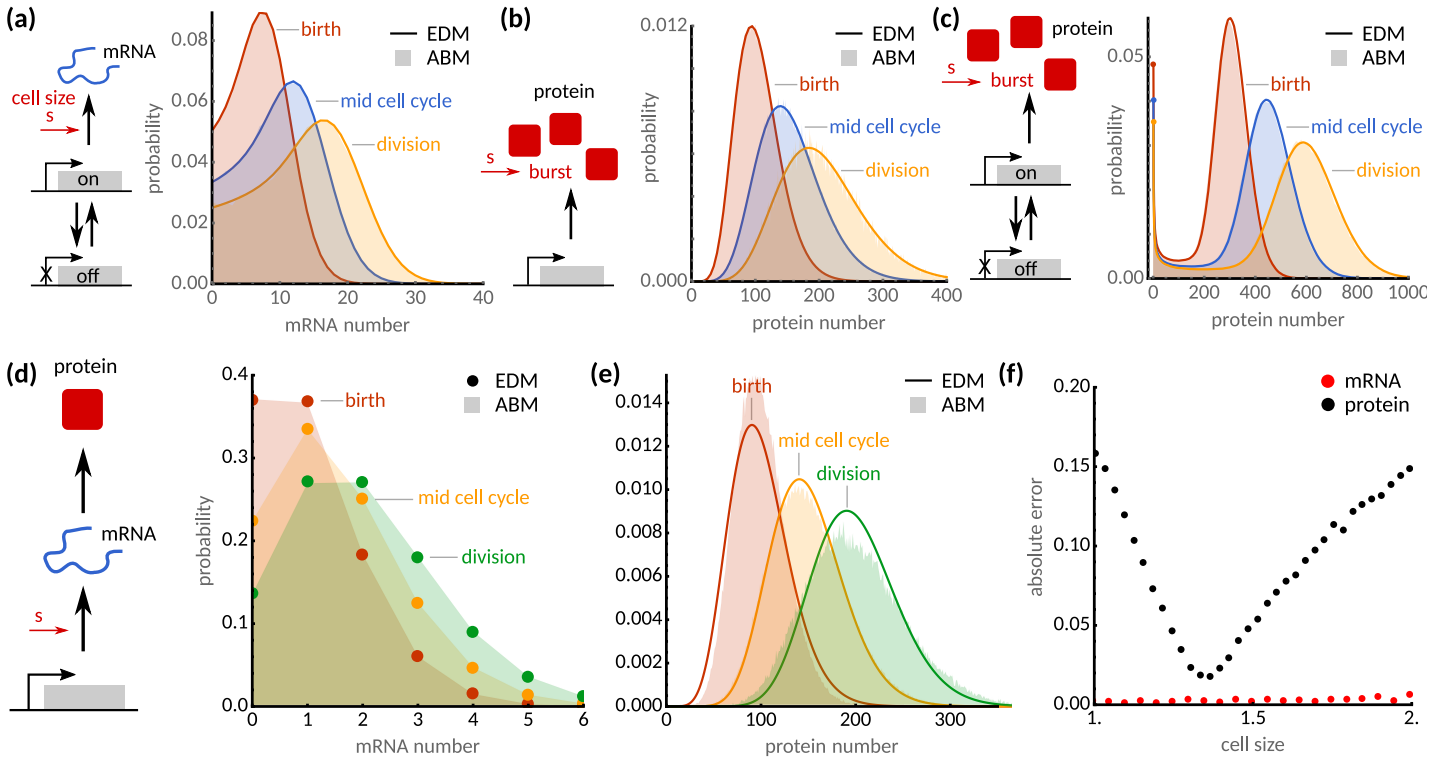

FIG. 2. **Distributions of CME and agent-based models agree for reaction networks with stochastic concentration homeostasis. (a-c)** The EDM (solid lines, analytical solution[53,54]) agrees with agent-based simulations (shaded areas) for a range of gene expression models. Panels show (a) bursty transcription[53] with transcription rate proportional to cell size, (b) bursty translation[54] , and (c) bursty transcription and translation[54] with geometrically distributed bursts m_s whose average is proportional to cell size s (see main text for details). **(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.[55]). **(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 $k_0 = 10$, $k_{\rm dm} = 9$, $k_{\rm tl} = 100$, $\alpha = 1$. Cell cycle noise assumes gamma distribution $\tilde{\varphi}(\Delta)$ with unit mean and $\mathrm{CV}_\varphi[\Delta] = 0.1$, while division noise assumes symmetric beta distribution with $\mathrm{CV}_\pi[\theta] = 0.01$.

to higher moments and distributions in stochastic reaction networks. SCH is a homeostatic condition for the distribution $p(x|s)$ of a size-dependent stochastic process to be expressed as a mixture of Poisson random variables drawn from an underlying continuous stochastic concentration vector $\kappa = (\kappa_1, \kappa_2, \ldots, \kappa_N)^T$ that is statistically independent of s:

$$p(x|s) = \int_{\mathcal{K}} \mathrm{d}\kappa\, \chi(\kappa) \prod_{i=1}^{N} \frac{(\kappa_i s)^{x_i}}{x_i!} e^{-s\kappa_i}. \tag{26}$$

The fact that κ and its density $\chi(\kappa)$ are independent of s ensures concentration homeostasis in the stochastic sense.

A. The effective dilution model is valid for reaction networks with stochastic concentration homeostasis

Theorem 1 (Appendix A) is a central result of our analysis and it states that if the EDM (4) satisfies SCH, i.e., Eq. (26) holds for $\Pi_{\rm EDM}(x|s)$, then its stationary solution is also a solution of the ABM (25):

$$\Pi(x|s, s_0) = \Pi_{\rm EDM}(x|s), \tag{27}$$

and the solution is independent of the birth size s_0. Equivalently (Appendix A), we can say that the EDM/ABM satisfies SCH if the factorial-moment generating function is of the form

$$G_{\rm EDM}(z|s) = \sum_x \prod_{i=1}^{N} z_i^{x_i} \Pi_{\rm EDM}(x|s) = F(s(z-1)), \tag{28}$$

where $F(t) = E_\chi[e^{\sum_{i=1}^N t_i \kappa_i}]$, the moment-generating function of the concentration vector, is cell-size (s) indepen-

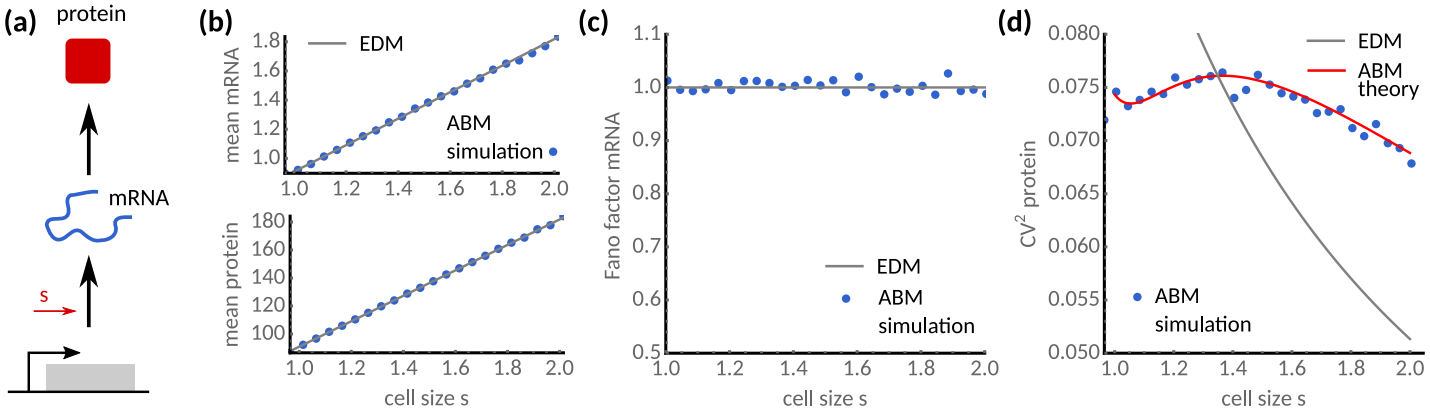

FIG. 3. **Comparing the statistics of the effective dilution and agent-based models. (a)** Simple model of mRNA transcription and protein translation transcriptional size-scaling (13). **(b)** Mean mRNA (top) and protein levels (bottom) agree with the EDM (solid grey lines) and ABM simulations (blue 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 $k_0 = 10$, $k_{\rm dm} = 10$, $k_{\rm tl} = 100$, $\alpha = 1$. Cell size control parameters are as in Fig. 2.

dent. An interesting observation is that SCH implies that the mean numbers and coefficients of variation for cells of the same size s are given by

$$E_\Pi[x|s] = sE_\chi[\kappa], \quad \mathrm{CV}^2_\Pi(x|s) = \frac{1}{sE_\chi[\kappa]} + \mathrm{CV}^2_\chi(\kappa). \quad (29)$$

Since κ is independent of s, SCH implies homeostasis of the mean concentrations in (29) but concentration homeostasis on average does not necessarily imply SCH. The coefficients of variation coincide both for concentrations and molecule numbers and have size-dependent and size-independent components. In the following, we provide examples of reaction networks for which SCH holds for all values of the rate constants and demonstrate the validity of the EDM by comparing its distribution solutions to ABM simulations.

It can be seen from (26) and (29) that when the EDM's stationary distribution is Poissonian with deterministic concentration vector κ, this distribution satisfies SCH and hence is also a solution of the ABM. More generally, SCH can be checked without solving for $G_{\rm EDM}(z|s)$ (or $\Pi_{\rm EDM}(x|s)$). Assuming mass-action kinetics (6), for example, a sufficient condition for SCH is that the network consists entirely of mono-molecular reactions (see Appendix A) of the form:

$$\varnothing \xrightarrow{D(t)s} S_1, \text{ or } \varnothing \xrightarrow{D(t)} m_s \times S_1,$$
$$\text{or } \varnothing \xrightarrow{s} S_1, \text{ or } S_1 \to \varnothing, \text{ or } S_1 \to S_2, \quad (30)$$

where S_1 and S_2 denote any pair of species that are partitioned at cell division and $D(t)$ is a exogenous stationary stochastic process modelling a genetic state which is not partitioned at cell division and does not scale with cell size. The propensities of zero-order reactions in SCH networks must either be proportional to cell size or include size-dependent random bursts m_s whose burst distribution satisfies SCH itself. We illustrate the predictive power of this result by demonstrating SCH for common gene expression models involving reactions of the form (30) and show that the analytical solution of the chemical master equations agrees exactly with the agent-based models (Fig. 2a-c).

mRNA expression involving a two-state promoter[53] (Fig. 2a),

$$D_{\rm off} \underset{k_{\rm off}}{\overset{k_{\rm on}}{\rightleftharpoons}} D_{\rm on} \xrightarrow{sk_0} D_{\rm on} + M, \quad M \xrightarrow{k_{\rm dm}} \varnothing$$

satisfies SCH for all parameter values since the network is of the form (30) whenever the transcription rate is proportional to cell size. The stochastic concentration variable is distributed as $\kappa \sim \frac{k_0}{(k_{\rm dm}+\alpha)}\mathrm{Beta}\left(\frac{k_{\rm on}}{(k_{\rm dm}+\alpha)}, \frac{k_{\rm off}}{(k_{\rm dm}+\alpha)}\right)$.

Bursty protein expression (Fig. 2b) of a stable (non-degrading) protein arising from a two-stage model of gene expression can be modelled using stochastic bursts:

$$\varnothing \xrightarrow{k_0} m_s \times P.$$

According to (30) the model satisfies SCH for all parameter values when the burst distribution obeys SCH. This is the case for geometrically distributed bursts m_s whose mean is proportional to cell size, $E[m_s|s] = bs$. It can be shown that the stochastic concentration variable follows $\kappa \sim \mathrm{Gamma}(k_0/\alpha, b)$.

Similarly, bursty protein expression from a two-state promoter arising from a three-stage gene expression model[54,56] (Fig. 2c),

$$D_{\rm off} \underset{k_{\rm off}}{\overset{k_{\rm on}}{\rightleftharpoons}} D_{\rm on} \xrightarrow{k_0} D_{\rm on} + m_s \times P,$$

also satisfies SCH for geometrically distributed bursts (with mean bs) but the concentration variable is doubly stochastic[57] $\kappa \sim \text{Gamma}(\frac{a_-}{\alpha}, br)$, $r \sim \text{Beta}(\frac{a_-}{\alpha}, \frac{a_+ + k_{\text{on}} + k_{\text{off}}}{\alpha})$ where $a_\pm = k_{\text{off}} + k_{\text{on}} + k_0 \pm \sqrt{(k_{\text{off}} + k_{\text{on}} + k_0)^2 - 4k_0 k_{\text{on}}}$. We observe excellent numerical agreement between the ABM simulations and analytical EDM solutions in all these cases validating our theoretical predictions (Fig. 2a-c).

A complex example of the reactions (30) that obeys SCH for all parameter values but yet defies analytical solution is

$$D_i \xrightarrow{g_{ij}} D_j \quad \forall i,j = 1,\dots,N_D$$
$$D_i \xrightarrow{st_i} D_i + M_1, \forall i = 1,\dots,N_D$$
$$M_1 \xrightarrow{k_1} M_2 \xrightarrow{k_2} \dots \xrightarrow{k_S} M_S, \quad M_i \xrightarrow{\delta_i} \varnothing \quad \forall i = 1,\dots,N_M$$

where the exogenous genetic states D_i undergo switching with rates g_{ij} but are not partitioned at cell division, transcription rates st_i are assumed to be proportional to cell size s, and processing of transcripts M_i follows a multi-step process with rates k_i and degradation with rates δ_i. For example, it can be checked that for $N_M = 1$ we recover the 2^m-multistate model[58] as a special case whose EDM has a factorial-moment generating function (compare Eq. (7) in[58] with (28)) satisfies SCH precisely when the transcription rates are proportional to cell size.

On the other hand, discrepancies between the EDM and ABM solutions will be apparent when reactions do not obey SCH. To illustrate this point, we return to the gene expression model with transcriptional size-scaling and explicit protein translation reaction (13). Note that in the EDM extra reactions are being added for the dilution of mRNAs and proteins, while for the ABM proteins are diluted through growth and divisions. Using our condition (28), it is straight-forward to verify that the Poissonian mRNA distributions of the EDM 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 (30). To demonstrate the breakdown of the EDM, we compare the analytical steady state distributions obtained by Bokes et al.[55] 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 pronounced both for newborn and dividing cells. The remainder of this article is dedicated to investigate the sources and consequences of these discrepancies.

B. The EDM approximates the mean concentrations of ABMs lacking SCH

SCH provides a general criterion with which to probe the validity of the EDM probability distributions. In practice, however, approximate agreement of the first few moments, e.g., mean and variances, often suffices. Here, we establish that under the mass-action scaling assumption (6) the mean concentrations of the ABM and EDM agree approximately, and they satisfy concentration homeostasis on average. This can be seen by multiplying Eq. (25) by x and averaging, which yields ODEs for the mean numbers:

$$\alpha s \frac{\partial}{\partial s} E_\Pi[x|s,s_0] = \sum_{r=1}^{R} \nu_r E_\Pi[w_r(x)|s,s_0], \tag{31a}$$

and the boundary conditions

$$E_\Pi[x|s_0,s_0] = E_\rho\left[\frac{s_0}{s'} E_\Pi[x|s',s_0']\middle| s_0\right]. \tag{31b}$$

Unfortunately, Eqs. (31) are not necessarily closed and the equation for the mean may involve higher order moments when $w_r(x)$ depends nonlinearly on x and we need to resort to approximations. Analogously to the linear noise approximation, we set $E[x|s,s_0] = s\bar{X}$ and $E_\Pi[w_r(x)|s,s_0] \approx s f_r(\bar{X})$ and insert the resulting expression into Eqs. (31). It follows that $\bar{X}$ is independent of size and satisfies the rate equations (2). We conclude that, for mass-action kinetics, the EDM agrees exactly with the ABM on average for networks with linear propensities and approximately for large cell size for nonlinear reaction networks.

C. Scaling of fluctuations with size in individual cells manifests the breakdown of the EDM lacking SCH

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

$$\alpha s \frac{\partial}{\partial s}\Sigma(s,s_0) = \mathcal{J}\Sigma(s,s_0) + \Sigma(s,s_0)\mathcal{J}^T + s\mathcal{D}(\bar{X}), \tag{32a}$$

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 ($\text{CV}_\varphi[\Delta] = \text{CV}_\pi[\theta] = 0$), which implies the following boundary condition

$$4\Sigma(s_0,s_0) = 2s_0\text{diag}(\bar{X}) + \Sigma(2s_0,s_0). \tag{32b}$$

The first term is due to binomial partitioning of molecules and the second stems from gene expression noise at division. It is implicit in the deterministic division assumption ($\text{CV}_\varphi[\Delta] = \text{CV}_\pi[\theta] = 0$) that the birth size s_0 across cells is fixed and that the size distribution in Eq. (24) reduces to

$$\Pi(s) = \Pi(s|s_0) = \frac{2s_0}{s^2} \tag{33}$$

for $s_0 \le s \le 2s_0$ and zero otherwise, in agreement with previous results[59,60]. Similarly, the ancestral distribution (25c) reduces to $\rho(s',s_0'|s_0) = \delta(s'-2s_0)\delta(s_0'-s_0)$.

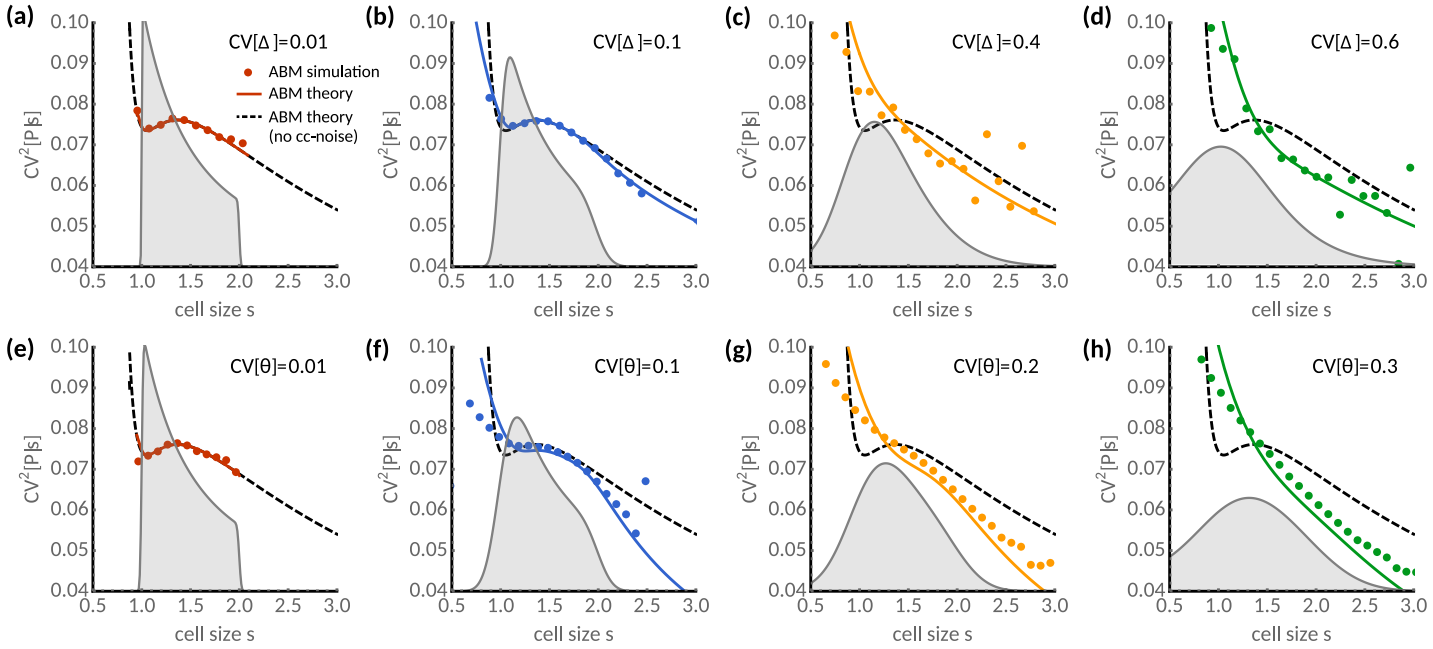

FIG. 4. **Effect of noise in cell size control and division on gene expression noise in single cells. (a-d)** Protein noise as a function of cell size s for various noise levels in added size CV[Δ]. For comparison, the prediction without cell cycle noise (dashed black line, Eq. (36)) and the cell size distributions (shaded grey) are shown. **(e-h)** Same as (a-d) but with division noise affecting the inherited size fraction CV[θ]. Analytical predictions (solid lines, Eq. (38a) with (36) and (38b)) 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 CV[Δ] = 0.01 (e-f) while division errors θ followed a symmetric beta distribution with CV[θ] = 0.01 (a-d).

Eqs. (32) can be solved in closed form using the eigendecomposition of the Jacobian $\mathcal{J}$. The solution to (32a) that respects the boundary condition (32b) is (Appendix D)

$$\Sigma(s,s_0) = \sum_{ij} \frac{s\,\hat{u}_i \hat{u}_j^\dagger}{(\alpha - \lambda_i - \lambda_j^*)} \times \left[\tilde{\mathcal{D}}_{ij} + \frac{\tilde{\mathcal{D}}_{ij} + \tilde{X}_{ij}(\lambda_i + \lambda_j^* - \alpha)}{2^{\frac{\lambda_i+\lambda_j^*}{\alpha}-1} - 2} \left(\frac{s_0}{s}\right)^{1-\frac{\lambda_i+\lambda_j^*}{\alpha}}\right], \quad (34)$$

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 result using the gene expression example with transcriptional size-scaling and explicit translation reaction (13). The mean of mRNA numbers m and protein numbers p are

$$E_\Pi[m|s] = s\frac{k_0}{\alpha\delta}, \quad E_\Pi[p|s] = s\frac{bk_0}{\alpha}, \quad (35)$$

where the constants are defined in Eq. (16). These expressions hold both for the EDM and the ABM, and they exhibit concentration homeostasis on average as shown in the previous section. The exact agreement between EDM and ABM is also confirmed by ABM simulations (Fig. 3b).

Using Eq. (32) we find that the cell size dependent fluctuations satisfy

$$\begin{aligned}
&\mathrm{Var}_\Pi[m|s,s_0] = E_\Pi[m|s],\\
&\mathrm{Cov}_\Pi[m,p|s,s_0] = \frac{sbk_0}{\alpha\delta}\left(1 + \left(\frac{s_0}{s}\right)^\delta \frac{2^\delta}{1-2^{\delta+1}}\right),\\
&\mathrm{Var}_\Pi[p|s,s_0] = E_\Pi[p|s]\times\\
&\qquad\left(1 + 2b - \frac{s_0}{s}\frac{4b\delta}{3(\delta-1)} + \left(\frac{s_0}{s}\right)^\delta \frac{b}{\delta-1}\frac{2^{\delta+1}}{(2^{\delta+1}-1)}\right).
\end{aligned} \quad (36)$$

We note that the mRNA variance of the ABM agrees precisely with the EDM (Fig. 3c). The agreement is a direct consequence of SCH exhibited by the mRNA transcription and degradation reactions (cf. (30)). However, the expressions for the predicted mRNA-protein covariance and protein variance disagree with their EDM counterpart since the reactions involving the protein violate SCH. 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 underestimates 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 coefficient of variation peaks as cells

progress through the cell cycle (Fig. 3d, solid red line), which is in excellent agreement with the ABM simulations (blue dots) but is not seen in the EDM (solid grey). This can be seen directly from Eqs. (36) for which protein fluctuations can be approximated in the limit of fast mRNA degradation ($\delta \gg 1$) as

$$\mathrm{CV}^2_{\mathrm{ABM}}[p|s,s_0] \approx \frac{1}{E_\Pi[p|s]}\left(1+b\left(2-\frac{s_0}{s}\frac{4}{3}\right)\right), \qquad (37)$$

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 b, the peak shifts from $s = s_0$ for $b = 3/2$ to $s = 4/3s_0$ for $b \gg 1$. The qualitative difference between the scalings of gene expression noise with cell size of EDMs and ABMs manifests the breakdown of the EDM, which is observed both for concentrations and molecule numbers since their coefficients of variation coincide when considering cells of the same size.

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 cell size control and division errors cause variable birth sizes, variable division times and hence noisy expression levels. mRNA fluctuations in the gene expression model with transcriptional size-scaling (13) obey SCH and hence are unaffected by these noise sources. The effect on protein noise remains yet to be elucidated.

To this end, we assume small birth-size variations and approximate 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

$$\Sigma(s) = E_\Pi[\Sigma(s,s_0)|s] \approx \Sigma(s, E_\Pi[s_0|s])). \qquad (38a)$$

This simplification can formally be justified through a saddle-point approximation as the joint distribution $\Pi(s,s_0)$ has a maximum at $\Pi(s, E_\Pi[s_0|s]))$. Generally no analytical expression of $E_\Pi[s_0|s]$ can be derived from Eq. (24) in the presence of cell size control fluctuations, however, and we approximate $E_\Pi[s_0|s]$ by a matched asymptotic expansion (Appendix C):

$$E_\Pi[s_0|s] \approx \bar{s}_0 - \underbrace{\sigma\frac{\sqrt{\frac{2}{\pi}}e^{-\frac{(s-\bar{s}_0)^2}{2\sigma^2}}}{1+\mathrm{erf}\left(\frac{s-\bar{s}_0}{\sqrt{2}\sigma}\right)}}_{\text{small cells}} + \underbrace{2a\sigma^2\gamma(s-a\bar{s}_0)}_{\text{large cells}}, \qquad (38b)$$

which holds for the linear cell size control model (19). 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 interpretation of this conditional expectation is that small cells were born with sizes smaller than average while larger cells were born with sizes above average depending on their size control (Fig. 6 in Appendix C). The parameters in Eq. (38b) 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.[48] for details):

$$\bar{s}_0 = \frac{(2-a)\left(1+\mathrm{CV}^2_\theta\right)}{2-a(1+\mathrm{CV}^2_\theta)},$$
$$\sigma^2 = \bar{s}_0^2\frac{\mathrm{CV}^2_\Delta\left(1+3\mathrm{CV}^2_\theta\right)\left(2-a(1+\mathrm{CV}^2_\theta)\right)^2+4\mathrm{CV}^2_\theta\left(1-\mathrm{CV}^2_\theta\right)}{\left(\mathrm{CV}^2_\theta+1\right)^2\left(4-a^2(1+3\mathrm{CV}^2_\theta)\right)}. \qquad (38c)$$

Eqs. (38) provide a closed form approximation of the cell size dependence of any given moment accurate to order $O(\sigma^3)$.

To verify the accuracy of proposed approximation, we test the theory for various strengths of size control noise and division errors (Fig. 4). We observe that increasing noise results in the monotonic decrease of 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 mean birth size (Fig. 4e-h), presumably since the effect of large variability in birth sizes is not captured in our 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 of networks lacking SCH.

D. ENMs provide surprisingly accurate approximations of total noise in ABMs lacking SCH

We go on to compare the ENM introduced in Sec. III D with the ABM. In contrast to the EDM, the ENM predicts the total noise statistics including the variability introduced by the cell size distribution. It is clear that the ENM agrees exactly with the ABM whenever the network obeys SCH. In particular, the marginal factorial-moment generating function of the ABM's molecule numbers $G(z) = E_\Pi\left[\prod_{i=1}^N z_i^{x_i}\right]$ (irrespective of cell size) follows from Eq. (26) as

$$G(z) = E_\chi\left[\prod_{i=1}^N K(\kappa_i(z_i-1))\right], \qquad (39)$$

when the concentration distribution χ has been identified (as we did for the models in Sec. III A) and the moment-generating function $K(t) = E_\Pi[e^{ts}]$ of the cell size distribution (24) is known. When SCH does not hold, the ABM statistics can in principle be obtained through integrating Eq. (34) against the size distribution $\Pi(s,s_0)$. Specifically, denoting molecule numbers by x and concentrations

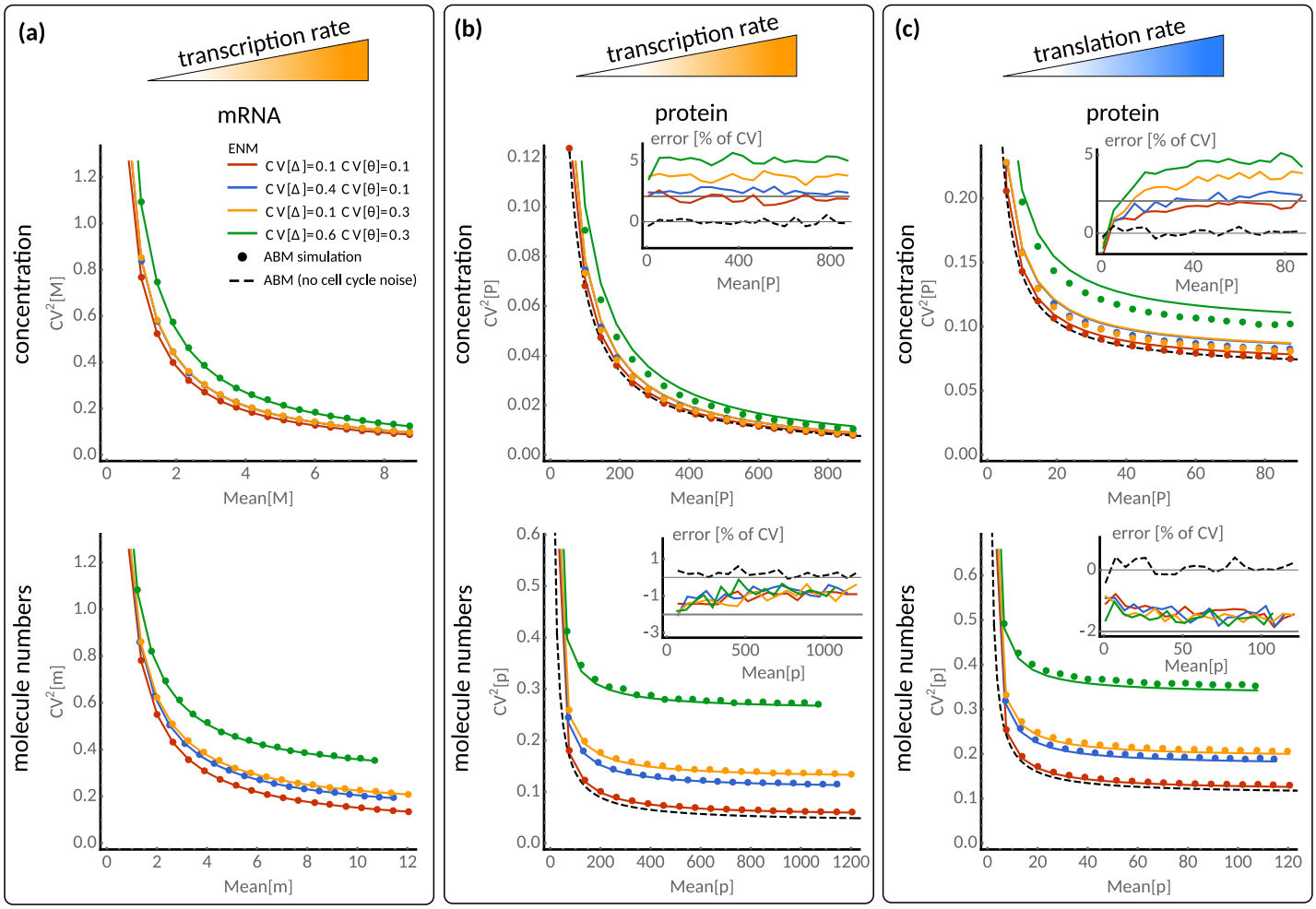

FIG. 5. **The extrinsic noise model approximates gene expression noise with size control and division errors.** **(a)** Scaling of mRNA concentration noise with mean concentrations for various noise levels in added size CV[Δ] and partition noise CV[θ] when the transcription rate k_0 is varied (top). 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 EDM 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_{tl} is varied. **(c)** Same as (b) but varying the transcription rate k_{tl}. For comparison, the ABM predictions without cell cycle noise are shown (dashed black lines, Eq. (41)) and the error bounds of 2% predicted by the theory (solid grey). See caption of Fig. 3 for the remaining parameters.

by $X = \frac{x}{s}$, as before, we have

$$\mathrm{Cov}_\Pi[X] = E_\Pi[\Sigma(s, s_0)/s^2],$$
$$\mathrm{Cov}_\Pi[x] = E_\Pi[\Sigma(s, s_0)] + \mathrm{Var}_\Pi(s)\bar{X}\bar{X}^T, \qquad (40)$$

where $\Sigma(s, s_0)$ is the size-dependent covariance matrix discussed in Sec. III C.

To illustrate this dependence, we consider the gene expression model with transcriptional size-scaling (13) and integrate Eq. (36) numerically against the size distribution (24). We observe that the mRNA noise-mean relationship of the ABM follows exactly the ENM predictions when the mean varies through the transcription rate (Fig. 5a). This agreement is confirmed for various strengths of cell size control fluctuations and division errors, both for mRNA concentrations and numbers, which validates our theoretical predictions that the mRNA distribution satisfies SCH and hence the ENM is exact.

However, the protein noise-mean relationships of the ABM and ENM differ (Fig. 5b). The discrepancy, albeit small, exists even for deterministic divisions ($\mathrm{CV}_\varphi[\Delta] = \mathrm{CV}_\pi[\theta] = 0$) for which the averages (40) over the size dis-

tribution can be carried out analytically and result in:

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

$$\mathrm{CV}^2_{\mathrm{ABM}}[p] = \mathrm{CV}^2_{\Pi}[s] + \frac{1}{\Omega_p \bar{P}}\left[1 + \frac{2b\left(\frac{2\left(2^{\delta-1}-1\right)}{(2^{\delta+1}-1)(\delta-1)} + \delta(\ln(8)-1) - 1\right)}{3\delta\ln(2)}\right], \quad (41)$$

where $\Omega_P = E_\Pi[s^{-1}]^{-1}$ and $\Omega_p = E_\Pi[s]$ and δ, b are defined as in Eq. (16). It can be verified by optimising (40) over δ that the ENM underestimates ABM noise of protein numbers by at most 2%, while it overestimates noise in protein concentrations by the same amount. The difference between the ENM and ABM predictions increases with cell size control noise for concentration measures but appears to be practically independent of cell size control noise for protein number fluctuations (Fig. 5b, insets). The protein number noise, but not concentration noise, exhibits an noise floor ($\mathrm{CV}^2_\Pi[s]$ in (41)) for large mean numbers due to extrinsic cell size variability across the population and this noise floor increases with noise in size control ($\mathrm{CV}_\varphi[\Delta]$) and division errors ($\mathrm{CV}_\pi[\theta]$) as it is also predicted by the ENM (Fig. 5b).

Similar conclusions hold for the noise-mean relationship when the mean is varied through translation rate (Fig. 5c) but there appears an additional noise floor due to stochastic bursting (Fig. 5c and Eq. (41)) that is present both in the protein concentration and number noise. This phenomenon is in qualitative agreement with previous findings[27,35,40] and similarly predicted by the ENM (15). Presumably, the better quantitative agreement for molecule numbers as compared to concentrations (Fig. 5b and c) is due the fact that the ENM and ABM predictions are generally dominated by extrinsic noise, which has the same effect in both models. Our observations suggest that, surprisingly, the ENM provides much more accurate approximations of the ABM statistics than the EDM.

IV. DISCUSSION

We presented an agent-based framework to study gene expression noise coupled to cell size dynamics across growing and dividing cell populations. The framework consists of an exact algorithm for simulating the stochastic dynamics of dividing cells (Box 1), which generalises previous algorithms for isolated lineages[4,17,32,61–63] towards growing cell populations, and a master equation framework (Sec. II) that exactly characterises the snapshot-distribution of gene expression and cell size across such a agent-based population.

Our theory shows that the newly defined stochastic concentration homeostasis (SCH) (cf. Sec. III A and Theorem 1) provides a necessary and sufficient condition for the stationary distributions of the chemical master equation (EDMs and ENMs) to agree exactly with the snapshot distributions of detailed agent-based models (ABM). A broad class of gene networks (30), involving mono-molecular reactions, multi-state promoters and bursting, satisfy SCH irrespective of network parameters when the reaction rates scale with cell size according to the law of mass-action. SCH is however not restricted to this particular class of network and can generally be checked on a case-by-case basis using the generating function equations, which can be accomplished without solving the chemical master equation analytically (see Appendix A).

SCH guarantees that gene expression distributions for cells of a given size are entirely independent of extrinsic noise sources affecting birth size such as cell size control and division noise. They thus reveal whether a network embedded in a growing cell can be insulated against such noise sources, an important feature that can guide the design of synthetic circuits.

Nevertheless, most gene regulatory networks of interest do not obey SCH. To address this issue, we developed the linear noise approximation for ABMs lacking SCH embedded in growing and dividing cells. The theory provides ODEs for the mean molecule numbers (31) and their covariances (32), which – unlike the conventional linear noise approximation describing EDMs[6,43] – describe their evolution across sizes and features a boundary condition describing the stochastic partitioning of molecules at cell division. We showed that, when the reactions follow mass-action size-scaling, the mean concentrations of EDMs and ABMs agree, because they exhibit concentration homeostasis (Sec. III B). The theory further provides closed-form analytical expressions for the covariance matrix of gene expression fluctuations in the absence of SCH. We note that like the conventional linear noise approximation, the linear noise approximation for ABMs is exact for linear reaction networks but represents an approximation for nonlinear reaction networks (Sec. III C).

While the EDM always predicts birth-size-independent noise, the ABM's covariance matrices generally depend on it (Sec. III C), both for concentrations and molecule numbers. This means that, unlike in SCH conditions, size-control noise and division errors can propagate to gene expression levels, and we unveiled quantitative and qualitative differences for the dependence of expression noise on cell size between EDMs and ABMs. Such differences prevail even for relatively simple gene networks involving protein expression (see (13)) for which our linear noise approximation readily provides exact expressions for mean and noise statistics.

Despite these discrepancies, we found that the ENM of these simple gene expression models provides surprisingly accurate total noise estimates (Sec. III D). In fact, we showed analytically that the ENM (and EDM) of bursty production with translational size-scaling agrees exactly

with the ABM since it obeys SCH. For transcriptional size-scaling, which implies the absence of SCH, the ENM deviates at most a few percent from ABM's total protein noise prediction. To resolve such small differences experimentally, one would need to probe on the order of $10,000$ cells for measuring the squared coefficient of variation accurate to three leading digits (assuming sampling errors inversely proportional to cell number), which is achievable only with high-throughput techniques.

An outstanding question is whether the good agreement we observed is specific to the particular model or parameter values we have chosen or whether the ENM is more generally valid. We have made an initial step in this direction by providing closed-form expressions for the ABM's linear noise approximation of any single-species reaction network with deterministic size-distribution (Appendix D). These results demonstrate that the ENM overestimates the ABM's coefficient of variation of concentrations by at most 8% but underestimates it by at most 2%, and vice versa for molecule numbers, and these bounds hold independently of the choice of parameters. This suggests that ENMs could be surprisingly accurate approximation of ABMs. Other effective models of bursty protein production without any size-scaling as proposed in[64] cannot obey SCH since they ignore cell size and generally produce larger errors than the ENM even for deterministic cell cycles.

A limitation of our study is that we assumed the validity of the linear noise approximation for the noise statistics of networks lacking SCH. Mean and covariance of the linear noise approximation are exact for linear reaction networks, as those we have studied here, but it represents an approximation for networks with nonlinear propensities valid in the limit of large molecule numbers. To improve the estimates of our theory, one could consider higher-order terms in the system size expansion[44], resort to moment-closure approximations[65], or to compute moment bounds[66] for nonlinear reaction networks.

Another limitation is that we neglected growth rate variability, which is a significant source of noise at the single-cell level[47]. It would be interesting to include these features in our ABMs, compare them to the effective models, and investigate whether SCH can be generalised to this case. Previous studies[3] have investigated the dependence of gene expression noise on growth rate dynamics in isolated lineages using small noise approximations similar to the one used here. Nevertheless, it may be expected that selection plays a pronounced role in populations where cells compete for growth unlike in isolated lineages, which in turn may lead to significant deviations of ENMs from ABMs[38,67,68] that we have not studied here.

In summary, we proposed SCH as a general condition for exactness of EDMs. In the absence of SCH, we found that despite qualitative differences in the predictions of EDMs, ENMs closely approximate the total noise statistics of ABMs. Our results reinstate the validity of effective models as approximations of the agent-based dynamics, and thus they 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

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

AUTHOR CONTRIBUTIONS

PT and VS designed the study and interpreted the results. PT developed the theory and analysed the data.

FUNDING

This work has been supported by a UKRI Future Leaders Fellowship (MR/T018429/1) to PT 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] 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).

[31] 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).

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

[33] 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).

[34] 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).

[35] 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).

[36] O. G. Berg, "A model for the statistical fluctuations of protein numbers in a microbial population," J Theor Biol **71**, 587–603 (1978).

[37] N. Brenner, K. Farkash, and E. Braun, "Dynamics of protein distributions in cell populations," Phys Biol **3**, 172 (2006).

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

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

[40] 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).

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

[49] B. M. Martins, A. K. Tooke, P. Thomas, and J. C. Locke, "Cell size control driven by the circadian clock and environment in cyanobacteria," Proc Natl Acad Sci **115**, E11415–E11424 (2018).

[50] 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).

[51] 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).

[52] 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).

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

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

[55] 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).

[56] K. Choudhary and A. Narang, "Urn models for stochastic gene expression yield intuitive insights into the probability distributions of single-cell mRNA and protein counts," Phys Biol **17**, 066001 (2020).

[57] This follows from the fact that $F(x) = {}_2F_1(a, \lambda; \gamma + \lambda; bx)$ is the moment-generating function of $\kappa \sim \text{Gamma}(a, (1+br)^{-1})$ and $r \sim \text{Beta}(\lambda, \gamma)$. Letting $E_\Pi[z^p|s] = F(s(z-1))$, $a = a_-/\alpha$, $\lambda = a_-\alpha$ and $\lambda + \gamma = \frac{k_{\text{on}}+k_{\text{off}}}{\alpha}$ then yields the factorial-moment generating function solutions in[54,56].

[58]L. Ham, D. Schnoerr, R. D. Brackston, and M. P. Stumpf, "Exactly solvable models of stochastic gene expression," J Chem Phys **152**, 144106 (2020).
[59]F. Maclean and R. Munson, "Some environmental factors affecting the length of Escherichia coli organisms in continuous cultures," Microbiology **25**, 17–27 (1961).
[60]A. Koch and M. Schaechter, "A model for statistics of the cell division process," Microbiology **29**, 435–454 (1962).
[61]T. Lu, D. Volfson, L. Tsimring, and J. Hasty, "Cellular growth and division in the Gillespie algorithm," IET Syst Biol **1**, 121–128 (2004).
[62]J. H. Van Heerden, H. Kempe, A. Doerr, T. Maarleveld, N. Nordholt, and F. J. Bruggeman, "Statistics and simulation of growth of single bacterial cells: illustrations with B. subtilis and E. coli," Sci Rep **7**, 1–11 (2017).
[63]C. Blanco, C. Nieto, C. Vargas, and J. Pedraza, "PyEcoLib: a python library for simulating E. coli stochastic size dynamics," bioRxiv , 319152 (2020).
[64]C. H. Beentjes, R. Perez-Carrasco, and R. Grima, "Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics," Phys Rev E **101**, 032403 (2020).
[65]D. Schnoerr, G. Sanguinetti, and R. Grima, "Approximation and inference methods for stochastic biochemical kinetics—a tutorial review," J Phys A **50**, 093001 (2017).
[66]J. Kuntz, P. Thomas, G.-B. Stan, and M. Barahona, "Bounding the stationary distributions of the chemical master equation via mathematical programming," J Chem Phys **151**, 034109 (2019).
[67]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).
[68]M. Ciechonska, M. Sturrock, A. Grob, G. Larrouy-Maumus, V. Shahrezaei, and M. Isalan, "Ohm's law for increasing fitness gene expression with selection pressure," bioRxiv , 693234 (2020).
[69]C. Gardiner, *Stochastic methods*, Vol. 4 (Springer Berlin, 2009).
[70]J. Kuntz, P. Thomas, G.-B. Stan, and M. Barahona, "Stationary distributions of continuous-time markov chains: a review of theory and truncation-based approximations," SIAM Review **63**, 3–64 (2021).

Appendix A: Stochastic concentration homeostasis and the validity of EDMs and ENMs

Appendices A and B use multi-index notation. In brief, a multi-index is a N-tuple $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$. One defines powers of a vector x via $x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_N^{\alpha_N}$, derivatives $\partial_x^\alpha = \frac{\partial^{\alpha_1}}{\partial x^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial x^{\alpha_2}} \cdots \frac{\partial^{\alpha_N}}{\partial x^{\alpha_N}}$, sum of components $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_N$, and the factorials $\alpha! = \alpha_1! \cdot \alpha_2! \cdots \alpha_N!$ and analogously $\binom{\alpha}{\beta} = \frac{\alpha!}{\beta!(\alpha-\beta)!}$.

Definition 1 (Stochastic concentration homeostasis)**.** *A probability mass function $\Pi(x|s)$ with state space $\mathcal{X}$ obeys SCH if for all $s \in [0,\infty)$ there exists a random variable κ on $\mathcal{K}$ with density $\chi(\kappa)$ satisfying:*

$$\Pi(x|s) = \int_{\mathcal{K}} \mathrm{d}\kappa \chi(\kappa) \frac{(\kappa s)^x}{x!} e^{-s|\kappa|}, \quad \forall x \in \mathcal{X}. \tag{A1}$$

Definition 1 implies that if κ has a moment-generating function $F(t) = E_\chi[e^{t^T\kappa}]$ then using (A1) one finds the factorial-moment generating function

$$E_\Pi[z^x|s] = \sum_{x\in\mathcal{X}} z^x \Pi(x|s) = \int_{\mathcal{K}} \mathrm{d}\kappa \chi(\kappa) e^{-s|\kappa|} \sum_{x\in\mathcal{X}} z^x \frac{(\kappa s)^x}{x!}$$
$$= E_\chi[e^{s(z-1)^T\kappa}] = F(s(z-1)) \tag{A2}$$

and similarly the moments

$$E_\Pi\left[\frac{x!}{(x-n)!}\middle| s\right] = (\partial_z^n E_\Pi[z^x|s])_{z=1} = s^{|n|} E_\chi[\kappa^n]. \tag{A3}$$

Since the factorial-moment generating function uniquely determines the distribution, Eqs. (A2) and (A3) may equivalently serve as definitions of SCH. Furthermore, when $\{x(s)\}_{s\in[0,\infty)}$ is interpreted as a point processes along the size coordinate s, SCH emphasises the fact that it is mixed Poisson process with stationary concentration vector κ.

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. A stationary solution of the EDM (4), if it exists, is also a solution of the ABM (25):*

$$\Pi(x|s, s_0) = \Pi_{EDM}(x|s), \tag{A4}$$

if and only if the EDM's solution (4) obeys SCH.

The utility of the theorem is that SCH 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 $w_r(x,s) = s^{1-|\nu_r^-|} k_r \frac{x!}{(x-\nu_r^-)!}$, whose factorial-moment generating function (see Chapter 7 in[69]) obeys:

$$\alpha s \frac{\partial}{\partial s} G(z|s,s_0)$$
$$= \sum_{r=1}^{R} k_r s^{1-|\nu_r^-|} \left(z^{\nu_r^+} - z^{\nu_r^-}\right) \partial_z^{\nu_r^-} G(z|s,s_0) \tag{A5}$$

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

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

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 comprise only the monomolecular reactions given in (30).

Similarly, we check that adding bursty reactions of the form $\varnothing \xrightarrow{k} m_s \times X$ leads to a generating function equation

$$\alpha s \frac{\partial}{\partial s} G(z|s,s_0) = k[g(z|s) - 1] G(z|s,s_0) + \ldots, \tag{A6}$$

where $g(z|s) = E[z^{m_s}|s]$ is the factorial-moment generating function of the burst distribution. Eq. (A6) transforms to $\alpha x \cdot \nabla F(x) = k[g(x/s+1|s) - 1]F(x)$ after substituting $G(z|s, s_0) = F(s(z-1))$ and $x = s(z-1)$. Hence, bursty reactions obey SCH if and only if the burst distribution obeys SCH, i.e., if there exist a moment-generating function f satisfying $f(x) = g(x/s+1|s)$ as in (A2).

Appendix B: Proof of Theorem 1

The proof of Theorem 1 is divided in three steps. The first step shows that a general condition (B1) guarantees snapshot distributions that are independent of birth size. We then show that (B1) satisfies the effective dilution model and reduces to SCH for binomial partitioning at cell division. The proof also clarifies that the assumption of binomial partitioning cannot be removed under biological constraints conserving the total number of molecule numbers at cell division. General conditions for the existence of the EDM's stationary distributions have been discussed in[70].

Step 1: Distributions invariant of birth size

The conditional distribution $\Pi(x|s, s_0)$ is independent of birth size s_0 if and only if

$$\Pi(x|\theta s, s_0) = \sum_{x'} B(x|x', \theta)\Pi(x'|s, s_0), \tag{B1}$$

where $B(x|x', \theta)$ is the partitioning kernel in Eq. (25b) that depends only the inherited size fraction $\theta = s_0/s'$. This fact can be verified using (B1) in the boundary condition (25b), 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 (B1) $\Pi(x|s)$ coincides with the EDM solution.

Step 2: Transformation into an effective dilution model

Let us denote the factorial-moment 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 (B1) becomes

$$G(z|\theta s) = \sum_{x'} G_B(z|x', \theta)\Pi(x'|s). \tag{B2}$$

Assume that additionally

$$\theta\partial_\theta G_B(z|x', \theta) = \sum_{i=1}^N (z_i - 1)\partial_{z_i} G_B(z|x', \theta), \tag{B3}$$

which holds for the binomial partition kernel $G_B(z|x', \theta) = (1 - \theta + \theta z)^{x'}$. Differentiating Eq. (B2) with respect to θ then gives

$$\begin{aligned}\theta\partial_\theta G(z|\theta s) &= \sum_{x'} \theta\partial_\theta G_B(z|x', \theta)\Pi(x'|s) \\ &= \sum_{i=1}^N (z_i - 1)\partial_{z_i} G(z|\theta s),\end{aligned} \tag{B4}$$

where in the last line we used assumption (B3). Changing variables ($\theta s \to s$) in (B4) 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 EDM

$$\begin{aligned}&\left(\alpha s \frac{\partial}{\partial s}\right)\Pi(x|s) \\ &= -\alpha \sum_{i=1}^N [(x_i + 1)\Pi(x_1, ..., x_i + 1, ..., x_N|s) - x_i\Pi(x|s)] \\ &= -\mathbb{D}\Pi(x|s).\end{aligned} \tag{B5}$$

Using the above relation, we see that (25a) coincides with (4) and (A4) follows.

Step 3: SCH and the necessity of binomial partitioning

Finally, we show that condition (B3) required for the validity of the EDM implies independent binomial partitioning of molecules. (B3) 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 of molecules in (20). It then follows that (B2) (and (B1)) are equivalent to

$$G(z|\theta s) = G((1 - \theta) + \theta z|s). \tag{B6}$$

Finally, we show that (B6) is equivalent to $G(z|s) = F(s(z-1))$ for binomial partitioning. Expanding (B6) 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 factorial-moment generating function is

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

with $F(x) = \sum_n x^n\mu_n(1)/n!$. It remains to be shown that F is indeed a moment-generating function. To this end, we note that $\frac{\partial^k G(z|s)}{\partial z^k} = s^{|k|}F^{(k)}(s(z-1)) \geq 0$ for $z \in (1,\infty)$ and hence $F(-x)$ is a completely monotone function on $x \in [0,\infty)$, which implies that there exists a distribution χ for which $F(-x) = E_\chi[e^{-sx}]$ is a Laplace transform, which concludes the proof of Theorem 1.

Appendix C: Approximation of birth size moments

Here derive an analytical approximation (38b) 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. (24):

$$\begin{aligned} E_\Pi[s_0|s] &= \int_0^s \mathrm{d}s_0\, s_0\Pi(s_0|s) \\ &= \frac{\int_0^s \mathrm{d}s_0\, s_0\Pi(s_0,s)}{\int_0^s \mathrm{d}s_0\,\Pi(s_0,s)} = \frac{E_\psi[s_0\Phi(s|s_0)1_{s_0\leq s}]}{E_\psi[\Phi(s|s_0)1_{s_0\leq s}]}. \end{aligned} \quad \text{(C1)}$$

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. (C1) 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. (C1), 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 (19), and $\bar{s}_0$ and σ are the mean and standard deviation of the backward lineage distribution ψ_{bw} given by Eqs. (38c).

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 (C1) can be approximated by

$$E_\Pi[s_0|s] \approx \frac{E_\psi[s_0 1_{s_0\leq s}]}{E_\psi[1_{s_0\leq s}]}, \quad \text{(C2)}$$

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

$$\begin{aligned} E_\psi[s_0 1_{s_0\leq s}] &\approx \frac{1}{2}\bar{s}_0\left(1+\mathrm{erf}\left(\frac{s-\bar{s}_0}{\sqrt{2}\sigma}\right)\right) - \frac{\sigma e^{-\frac{(s-\bar{s}_0)^2}{2\sigma^2}}}{\sqrt{2\pi}} \\ E_\psi[1_{\bar{s}_0\leq s}] &\approx \frac{1}{2}\left(1+\mathrm{erf}\left(\frac{s-\bar{s}_0}{\sqrt{2}\sigma}\right)\right) \end{aligned} \quad \text{(C3)}$$

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

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

the uniformly valid matched asymptotic expansion is

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

which gives Eq. (38b).

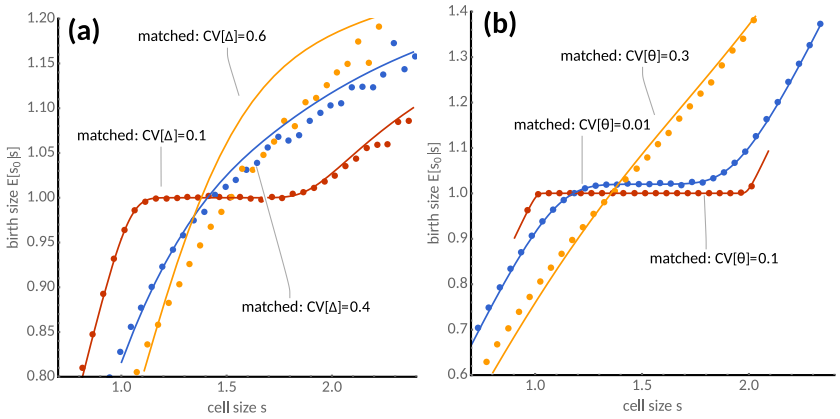

FIG. 6. **Retrospective averages of birth size.** Matched asymptotic expansions of $E_\Pi[s_0|s]$ (Eq. (38b), lines) for the adder size control ($a = 1$, $s_0 = 1$) agree well with agent-based simulations (dots) for **(a)** varying size control noise CV[Δ] and **(b)** division errors CV[θ].

Appendix D: Analytical solutions and error bounds using the linear noise approximation for deterministic cell division

We begin by outlining the solution of (32). Defining $\tilde{\Sigma}(s, s_0) = U^{-1}\Sigma(s, s_0)U^{-\dagger}$, Eq. (32) of the main text becomes

$$\alpha s \partial_s \tilde{\Sigma}_{ij} = (\lambda_i + \lambda_j^*)\tilde{\Sigma}_{ij} + s\tilde{\mathcal{D}}_{ij} \tag{D1}$$

$$4\tilde{\Sigma}_{ij}(s_0, s_0) = \tilde{\Sigma}_{ij}(2s_0, s_0) + 2s_0\tilde{X}_{ij}. \tag{D2}$$

Eq. (D1) has the solution

$$\tilde{\Sigma}_{ij}(s, s_0) = c_{ij} s^{\frac{\lambda_i+\lambda_j^*}{\alpha}} + \frac{\tilde{\mathcal{D}}_{ij} s}{\alpha\left(1 - \frac{\lambda_i+\lambda_j^*}{\alpha}\right)}, \tag{D3}$$

where the constants c_{ij} are fixed using the boundary condition (D2) which gives the solution of the EDM, Eq. (34) of the main text.

Using Eq. (40) and averaging (D3) over the deterministic size distribution (33), we find the covariance matrix of concentrations X,

$$\mathrm{Cov}_\Pi[X] = \frac{1}{\Omega}\sum_{i,j=1}^{N} \hat{u}_i\hat{u}_j^\dagger \frac{\beta_{ij}\left(\coth\left(\frac{1}{2}\beta_{ij}\ln 2\right) + 3\right)}{3(\beta_{ij}+1)}\frac{\alpha\tilde{X}_{ij}}{\xi_{ij}} + \frac{\beta_{ij}\left(3\beta_{ij} - \coth\left(\frac{1}{2}\beta_{ij}\ln 2\right)\right)}{3\left(\beta_{ij}^2 - 1\right)}\frac{\tilde{\mathcal{D}}_{ij}}{\xi_{ij}}, \tag{D4}$$

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. (34). Similarly, considering molecule numbers x, we have $\mathrm{Cov}_\Pi[x] = \Sigma^{\mathrm{int}}_{\mathrm{ABM}} + \mathrm{Var}_\Pi(s)\bar{X}\bar{X}^T$ where the intrinsic noise contribution is given by

$$\Sigma^{\mathrm{int}}_{\mathrm{ABM}} = \Omega\sum_{i,j=1}^{N}\hat{u}_i\hat{u}_j^\dagger\left[\frac{\left(2^{\beta_{ij}}-2\right)\beta_{ij}}{\left(2^{\beta_{ij}}-1\right)(\beta_{ij}-1)\ln 4}\frac{\alpha\tilde{X}_{ij}}{\xi_{ij}} + \frac{\beta_{ij}\left(\left(2^{\beta_{ij}}-1\right)\beta_{ij}\ln 4 - 2^{\beta_{ij}}(1+\ln 4) + 2 + \ln 4\right)}{\left(2^{\beta_{ij}}-1\right)(\beta_{ij}-1)^2\ln 4}\frac{\tilde{\mathcal{D}}_{ij}}{\xi_{ij}}\right], \tag{D5}$$

with $\Omega = E_\Pi[s] = s_0 \ln 4$.

The expressions greatly simplify for a single species since $\hat{u} = \hat{u}_i$, $\beta = \beta_{ij}$ and $\xi = \xi_{ij}$. We note that in this case (D4) increases monotonically with β while (D5) decreases monotonically with β. Using the limits $\beta \to 0$ and $\beta \to \infty$, we find that the ABM's coefficients of variation can be bounded by the EDM's coefficients:

$$\frac{6}{7} \le \frac{\mathrm{CV}^2_{\mathrm{ENM}}[X]}{\mathrm{CV}^2_{\mathrm{ABM}}[X]} \le \frac{3}{2}\ln 2, \tag{D6}$$

$$2\ln^2(2) \le \frac{\mathrm{CV}^2_{\mathrm{ENM}}[x]}{\mathrm{CV}^2_{\mathrm{ABM}}[x]} \le \frac{4\ln 2}{1+\ln 4}, \tag{D7}$$

where we have used the fact that $\mathcal{D}(\bar{X}) \ge \alpha\bar{X}$ and the ENM solution of (9) is $\Sigma^{\mathrm{int}}_{\mathrm{ENM}} = \frac{\alpha\bar{X}}{\Omega\xi} + \frac{\mathcal{D}}{\Omega\xi}$. The result implies that the ENM overestimates the ABM's coefficient of variation of concentrations by at most 8% but underestimates it by at most 2%, and vice versa for molecule numbers.

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[46,50,52]. The algorithm combines the Extrande method[45] for simulating reaction networks embedded in a growing cell and the First-Division algorithm[38] 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 $\{t_{0,i}, s_{0,i}, s_i, x_i\}_{i=1,..,M}$ for population of M cells.
Output: Cell states at time $T > 0$.
Require: Growth rate α, cell size control model (19) 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 (19);
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 $\theta \sim \bar{\pi}$ 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