PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Quantifying biochemical reaction rates from static population variability within incompletely observed complex networks

Timon Wittenstein[1,2☯], **Nava Leibovich**[1☯], **Andreas Hilfinger**[1,3,4,5*]

1 Department of Physics, University of Toronto, Ontario, Canada, **2** Department of Physics, Johannes Gutenberg University Mainz, Mainz, Germany, **3** Department of Mathematics, University of Toronto, Toronto, Ontario, Canada, **4** Department of Cell & Systems Biology, University of Toronto, Toronto, Ontario, Canada, **5** Department of Chemical & Physical Sciences, University of Toronto Mississauga, Mississauga, Ontario, Canada

☯ These authors contributed equally to this work.
* andreas.hilfinger@utoronto.ca

OPEN ACCESS

Citation: Wittenstein T, Leibovich N, Hilfinger A (2022) Quantifying biochemical reaction rates from static population variability within incompletely observed complex networks. PLoS Comput Biol 18(6): e1010183. https://doi.org/10.1371/journal.pcbi.1010183

Editor: Attila Csikász-Nagy, Pázmány Péter Catholic University: Pazmany Peter Katolikus Egyetem, HUNGARY

Received: September 21, 2021

Accepted: May 7, 2022

Published: June 22, 2022

Data Availability Statement: All relevant data are within the paper and its Supporting information files. Simulation code is available through the following public repository (also referenced in the manuscript) https://github.com/t-wittenstein/quantify-biochemical-rates.

Funding: This work was supported by a Discovery Grant (AH) of the Natural Sciences and Engineering Research Council of Canada and a New Researcher Award (AH) from the University of Toronto

Abstract

Quantifying biochemical reaction rates within complex cellular processes remains a key challenge of systems biology even as high-throughput single-cell data have become available to characterize snapshots of population variability. That is because complex systems with stochastic and non-linear interactions are difficult to analyze when not all components can be observed simultaneously and systems cannot be followed over time. Instead of using descriptive statistical models, we show that incompletely specified mechanistic models can be used to translate qualitative knowledge of interactions into reaction rate functions from covariability data between pairs of components. This promises to turn a globally intractable problem into a sequence of solvable inference problems to quantify complex interaction networks from incomplete snapshots of their stochastic fluctuations.

Author summary

Statistical models are the dominant tool to interpret co-variability of molecular components in cellular processes because they can be formulated for a subset of components while leaving the full complexity of the processes unspecified. Their drawback lies in the difficulty of translating statistical associations into causal interactions. In contrast, complete mechanistic models of biochemical reaction networks are a powerful tool to describe physical interactions, but often necessitate making a large number of assumptions such that each individual assumption is only marginally tested in global model comparisons with data. We introduce a novel inference method that combines the power of both approaches by exploiting testable predictions for only partially specified mechanistic models. We present numerical proof-of-principle examples in which we reconstruct biochemical reaction rates from partial observations of variability within simulated biochemical reaction networks in the absence of cell division. In contrast to existing approaches, our algorithm does not require perturbations, temporal information, observing all

Connaught Fund. NL was supported by research awards from the Israel Council for Higher Education (VATAT) and a Ben-Gurion University Postdoctoral Fellowship. TW gratefully acknowledges financial support through a Deutschlandstipendium. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

components within a complex network, or complete model knowledge. Its key ingredients are partial knowledge of qualitative network interactions paired with high precision probability distributions to quantify stochastic fluctuations within biochemical reaction networks.

Introduction

Quantifying interactions between components within a complex network from snapshots of their activity is a challenge common to many areas of science. For example, understanding cellular processes requires quantifying biochemical reaction rates between molecules while typical high-throughput methods such as single-cell sequencing [1, 2], flow cytometry [3–5], or a combination thereof [6] generate static population snapshots of a subset of cellular components.

Covariability of components within cellular processes is typically analyzed using statistical associations [7–13] because accurate mechanistic modelling of biochemical reactions is challenging for complex systems due to the large number of unknown parameters and interactions [14]. However, molecular abundances are set by underlying physical interactions that affect each component's rate of production and degradation rather than its instantaneous concentration. How molecular components affect each other is then difficult to infer from statistical associations [15], especially in the absence of perturbation experiments [16, 17]. For example, even perfectly linear *rate* dependencies will lead to non-linear statistical relations between observed *values* of cellular components.

Mechanistic models that quantify causal interactions between components would thus be preferable to describe biochemical reaction networks. However, constructing complete mechanistic models requires describing every interaction in a system. For complex cellular processes, often only some of the mechanistic details of molecular interactions will be known in detail, some aspects will have to be estimated from comparable systems reported in the literature, and some will have to be postulated because of a lack of direct experimental evidence. Here, we introduce a novel data analysis approach to deduce mechanistic rate dependencies *one interaction at a time* using incompletely specified mechanistic models, see Fig 1. Its key advantage is that only the interactions of the local network need to be specified and the mechanistic details of regulation or degradation of all other components within the network do not need to be modelled. This approach exploits a local qualitative understanding of network interactions through probability balance equations [18] that must be satisfied as long as we know how one component is made and degraded. In contrast to existing work [13, 19, 20], our approach does not require temporal information, experimental perturbations, or complete observation of all components within an interaction network.

We present numerical evidence for four distinct network topologies in which we successfully infer how one component affects the production rate of another from their observed joint probability distribution without without making any assumptions about the dynamics of the non-observed components. This proof-of-principle across four different network dynamics illustrates how qualitative knowledge of local network interactions can be translated into quantitative rate functions using only static snapshots of naturally occurring population variability.

Background theory

Describing the dynamics of some components within a complex interaction network in which the interactions between many components are unknown may seem impossible. However, we

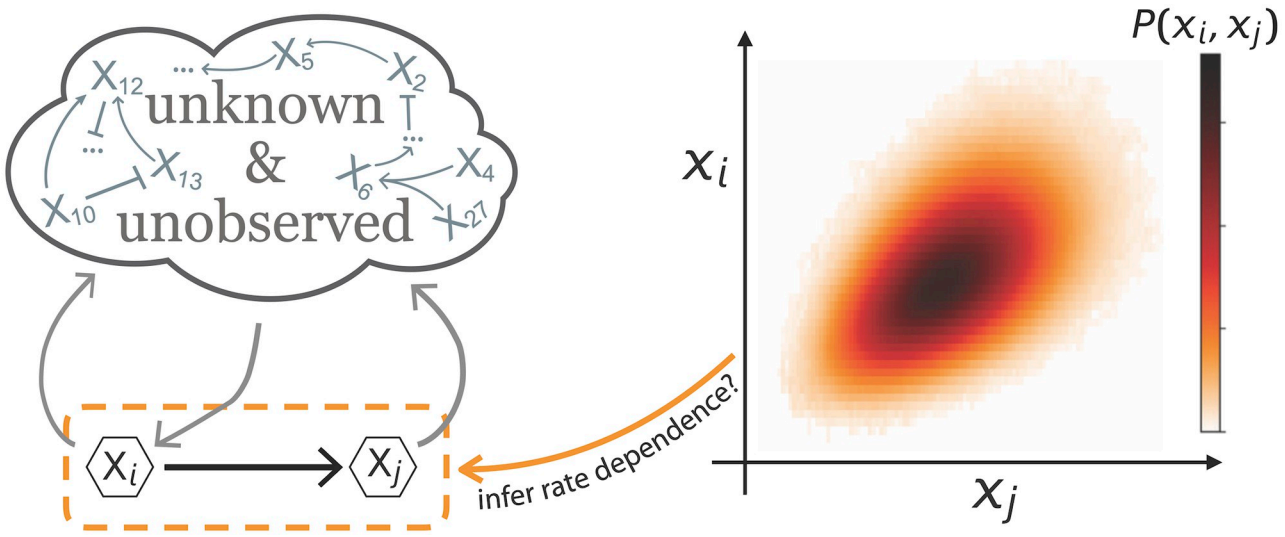

Fig 1. Our goal: Determining how one component affects the production rate of another from observing their joint probability distribution in a sea of unobserved components with unknown dynamics. The joint probability distribution between two components $P(x_i, x_j)$ varies enormously between systems even for identical regulation of X_i through X_j. However, previous work has shown that all systems with a given interaction between the two components satisfy invariant probability flux balance relations [18]. Here, we demonstrate that such relations can translate qualitative local network knowledge into quantitative rate functions using only the observed joint probability distributions between X_i and X_j even when the dynamics within the rest of the network is unobserved and completely unknown.

https://doi.org/10.1371/journal.pcbi.1010183.g001

can trivially do so as long as we are content with describing one component's dynamics in terms of components directly affecting it. The actual dynamics are fundamentally indeterminable for incomplete models but if components of interest are experimentally measurable their empirically observed covariability can be used to close the problem and constrain interaction rates as described below.

We follow the previously established approach [18] to characterize "local" system dynamics within a completely general complex reaction network with probabilistic events

$$\mathbf{x} \xrightarrow{r_k(\mathbf{x})} \mathbf{x} + \mathbf{d}_k \qquad k = 1, 2, 3, \ldots$$

where the state vector $\mathbf{x} = (x_1, x_2, x_3, \ldots)$ of abundances can be arbitrarily high-dimensional, the k^{th} reaction changes levels of component X_i by d_{ki}, and the reaction rates $r_k(\mathbf{x})$ are arbitrarily non-linear functions of the state vector. This notation is motivated by biochemical reaction networks but many areas of science encounter stochastic systems whose dynamics are determined by the corresponding general chemical master equation

$$\frac{\mathrm{d}P(\mathbf{x}, t)}{\mathrm{d}t} = \sum_k \left[r_k(\mathbf{x} - \mathbf{d}_k) P(\mathbf{x} - \mathbf{d}_k, t) - r_k(\mathbf{x}) P(\mathbf{x}, t) \right] \tag{1}$$

where $P(\mathbf{x}, t)$ denotes the probability of the system to be in state $\mathbf{x}$ at some time t. Directly solving Eq (1) is impractical for many complex cellular processes of interest because, generally, not all reactions rates are known, and because non-linear rates in complex systems generally render them analytically intractable due to moment closure problems [21, 22], although exceptions exist [23].

However, even when many details of a system are unknown, any given molecular component X_i that reaches a time-independent stationary state with probability distribution $P_{ss}(x_i)$

must satisfy

$$0 = \sum_k [\langle r_k(\mathbf{x}) | x_i = m - d_{ki} \rangle P_{ss}(x_i = m - d_{ki}) - \langle r_k(\mathbf{x}) | x_i = m \rangle P_{ss}(x_i = m)] \quad \forall m \in \mathbb{N}_0 \tag{2}$$

which follows from simple summation of Eq (1) over all other variables and has been derived and discussed previously [18, 24]. Here, and throughout the article, angular brackets denote averages over the stationary state distribution.

In this paper, we demonstrate that Eq (2) can be exploited to infer rate functions even when we know nothing about the dynamics of all other components X_j for $j \neq i$ such that conditional rates cannot be predicted from incompletely specified models. Note, Eq (2) is not an approximate coarse-graining but corresponds to an exact balance relation for any variable in a larger complex system at stationarity. This stationarity assumption allows for a broad class of biological systems to be analyzed: Eq (2) requires that the joint probability distribution of interest does not change over time, which can be verified experimentally from population snapshots taken at different time-points. Thus, the only dynamics excluded from our analysis is transient behaviour such that stationary probability distributions are not accessible from experimental data. Even explicitly time-varying systems that technically never reach a stationary state, such as deterministic oscillations, satisfy Eq (2) when considering their time-averaged probability distributions and rates [18]. Although care has to be taken when comparing such time-averages with population averages in growing populations [25].

Next, we present results to show that stationary state probability distributions contain enough information to reconstruct an entirely unknown rate function in an otherwise unknown network of interactions. This is in contrast to an analysis of deterministic steady-states which only give one balance equation for each variable of interest and do not contain enough information to reconstruct a general rate function that is not parameterized by a single parameter.

Results

When the rates of all reactions directly changing X_i-levels are known, Eq (2) represents a self-consistency check that must be satisfied by the observed joint probability distribution between X_i and all variables directly affecting those rates. Next, we show how this relation can be "inverted" to determine rate functions from observed probability distributions.

While Eq (2) must provably hold for all stationary states, any empirically observed distribution will exhibit sampling errors which can have significant effects. For example, any real experiment will have some maximum value $m_{\max}$ for which X_i is observed and thus Eq (2) will clearly be violated for unbounded systems because Eq (2) cannot balance when $m = m_{\max}$ due to the lack of sampling of the rarest states. Inverting the equation system Eq (2) to identify the functional dependencies of $r_k(\mathbf{x})$ thus requires minimizing deviations from the predicted relations. In general, minimizing the sum of squared differences remains an underdetermined problem if we treat each value of the rate function as an independent unknown. However, under the assumption that biochemical rate functions are sufficiently smooth the problem can be solved by limiting the variability of the rate functions across neighbouring states (Materials & methods).

To demonstrate how rate functions can be successfully inferred from partial observations of some components within a larger network we consider four different example networks that exhibited oscillations, bistability, fluctuation control, and noise enhancing feedback. Such markedly different global dynamics was already achievable with non-linear three-component feedback networks while conserving the reaction rates for one of the components, see Fig 2

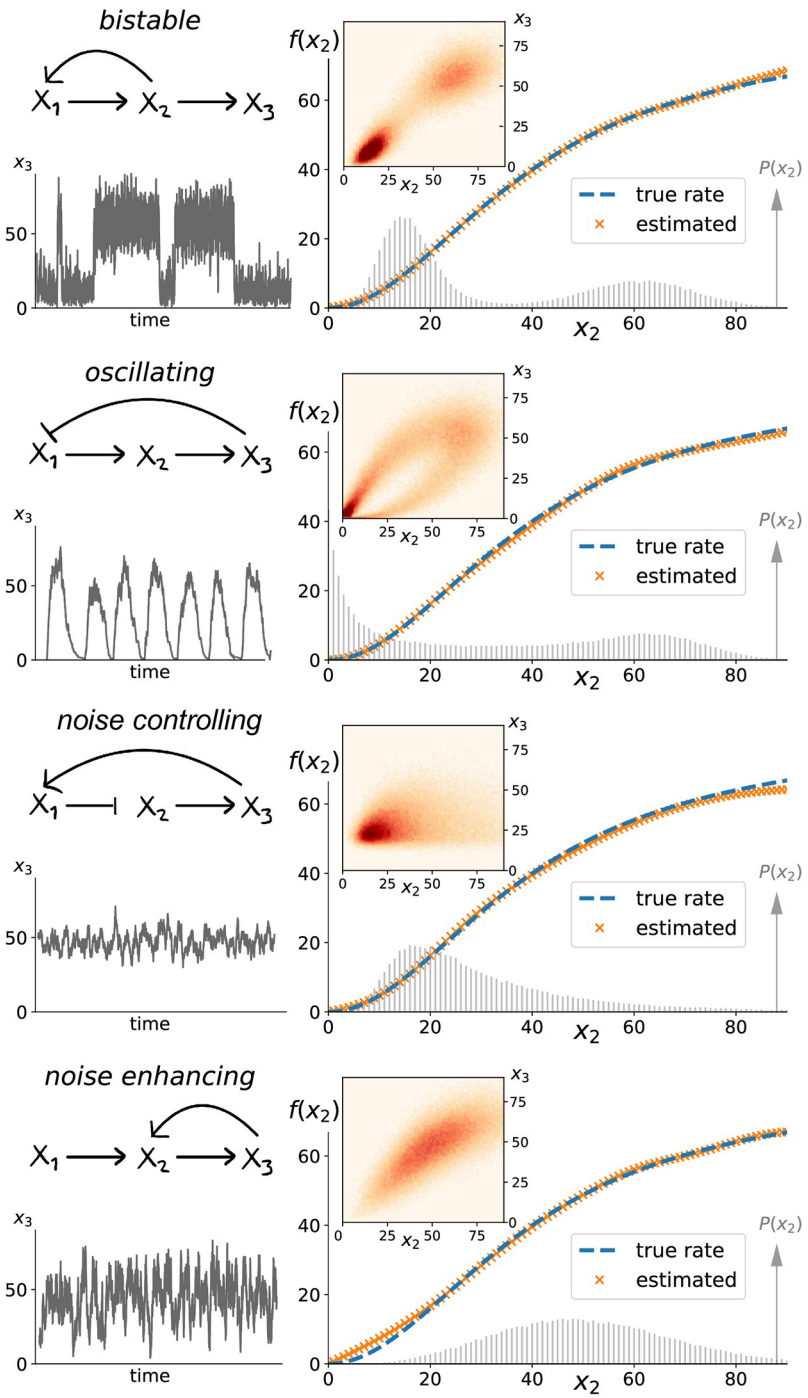

Fig 2. Probability flux balances can determine biochemical rates regardless of global network dynamics. Fixing how the production rate $f(x_2)$ of X_3 depends on X_2-levels, we considered four different global network topologies within the class defined by Eq (3), that exhibit diverse system dynamics and variability in X_3 (left column). The insets of the right column depict numerically observed joint distributions $P(x_2, x_3)$ corresponding to 100,000 independent snapshots. Although probability distributions differed greatly between the four systems, Eq (4) could identify the functional dependence of the production rate of X_3 based on the numerical convex optimization algorithm detailed in the Materials & Methods. We find near perfect agreement between the inferred rate (orange crosses) and the true rate function (dashed blue line) regardless of a system's global dynamics. This inference of $f(x_2)$ does not utilize any temporal information, its only input is the stationary joint probability distribution between the two components of interest. It relies on observing fluctuations across a wide range of X_2-states as illustrated by the shaded probability distribution $P(x_2)$ with deviations occurring where X_2 was rarely or never observed. While the degradation rate of X_3 was assumed to be known, no information about how its production rate depends on X_2, or the dynamics of X_1, X_2 was used.

https://doi.org/10.1371/journal.pcbi.1010183.g002

(left panels). We thus present the performance of our algorithm when applied to simulation data from those simple systems. But the algorithm can equally be applied to much larger systems with hundreds of variables as long as the reactions directly affecting X_i are qualitatively known.

Numerical proof-of-principle examples

The conserved part of our test systems that we want to reconstruct from simulation data corresponds to a stereotypical biochemical reaction rate in cells. In particular, we specified that the production of component X_3 is affected by X_2 through a Hill-type function $f(x_2)$ and X_3 molecules are degraded independently leading to the following class of reaction systems

$$\underbrace{\begin{array}{l} x_3 \xrightarrow{f(x_2)} x_3 + 1 \\ x_3 \xrightarrow{x_3/\tau_3} x_3 - 1 \end{array}}_{\text{conserved part}} + \underbrace{\left[\begin{array}{c} \mathrm{X}_1, \mathrm{X}_2 \\ \text{production \& degradation} \end{array}\right]}_{\text{various feedback dynamics}}, \tag{3}$$

where τ_3 denotes the average life-time of component X_3 and $f(x_2) = \lambda x_2^n/(K^n + x_2^n)$. For the test examples presented in Fig 2 (dashed blue lines in right panels) we chose $\tau_3 = 1$ and $\lambda = 80$, $n = 2$, $K = 40$.

The reaction dynamics of the other variables, i.e., how X_1, X_2 affect each other and how they are affected by X_3 were chosen to achieve diverse system dynamics and are specified in the Materials & Methods. Regardless of the X_1,X_2-dynamics, the probability balance equation Eq (2) applied to the specified reaction in Eq (3) imply that the above systems must satisfy the following probability balance relations at stationarity

$$\frac{P(x_3 = m+1)}{P(x_3 = m)} = \frac{\langle f(x_2)|x_3 = m\rangle}{(m+1)/\tau_3} \quad \forall m \in \mathbb{N}_0. \tag{4}$$

Although Eq (4) is reminiscent of detailed balance, individual backwards and forward reaction fluxes do not need to balance in general for systems that do not operate at thermodynamic equilibrium such as cellular processes. The condition we exploit here is that the marginal probability distribution does not change at stationarity, and thus that each state must on average balance incoming and outgoing probability fluxes. The contrast with detailed balance is directly apparent in systems with dimeric degradation of X_3 as discussed in a later section.

As detailed in the Materials & Methods, we employed a numerical algorithm to approximately solve Eq (4) for $f(x_2)$ and thus infer how the X_3 production rate depends on X_2 from the observed $P(x_2, x_3)$. To do so we generated exact realizations of the above stochastic processes using the standard Doob-Gillespie algorithm [26, 27]. We then sampled X_2, X_3 from the numerically observed stationary distribution to generate an observed joint probability from $N = 100,000$ independent samples. Any information about the dynamics of X_1 was discarded because our method does not utilize any information beyond the pairs of components under consideration. The numerical algorithm is straightforward and detailed in the in the Materials & Methods. In short, a convex optimization algorithm can identify the $f(x_2)$ that minimizes Eq (4) for the numerically observed $P(x_2, x_3)$. Doing so requires finding a large-dimensional but finite solution vector with elements $f_n := f(x_2 = n)$ over the observed states that solves a regular least-squares problem as defined in Eq (9). Standard convex optimization allows to find a solution vector constrained to be non-negative and sufficiently smooth by penalizing large second derivative terms as detailed in the Materials & Methods. The results were a near perfect inference for the production rate of X_3 in all systems as illustrated by the orange crosses in Fig 2.

These numerical proof-of-concept examples thus illustrate how the balance equations Eq (4) can be used to reconstruct the functional form of $f(x_2)$ from pairwise observation of X_2, X_3 in the absence of any temporal information and independent of any information about the vastly different global system dynamics.

Note, in these successful proof-of-principle examples we deliberately made no assumption about $f(x_2)$ beyond its smoothness and non-negativity. Alternatively, one could, e.g., assume that $f(x_2)$ is a Hill-function and identify its characterizing parameters K, n, λ through minimizing violations of Eq (4). However, this necessarily requires performing a non-linear optimization with all the drawbacks that entails compared to a linear optimization. If one knows that $f(x_2)$ is a Hill-function, one can always fit a Hill-function through the $f(x_2)$ obtained from our algorithm and identify K, n, λ that way.

To fully specify a rate function rather than a finite table of values, e.g., through fitting a Hill function or simply continuing $f(x_2) = const.$ above and below the largest and smallest state, requires additional *a priori* information. Because such information cannot be deduced from the experimental observations we refrain from doing so throughout the manuscript.

Also note, that when X_2 varies very slowly $f(x_2)$ can be directly read off from the conditional averages $\langle f(x_2)|x_3\rangle$ as discussed in the next section. The above algorithm solves the problem of determining $f(x_2)$ when the timescales of X_2 and X_3 are not separable.

Sampling requirements

Information cannot be created from nothing and the above inference cannot determine rates for states that were never observed. In practice, making additional assumptions to fill in gaps, such as monotonicity or the functional form of $f(x_2)$, could prove useful (and are easily incorporated into the algorithm), but here we want to illustrate the core of the inference quality based solely on the convex optimization of Eq (4). We thus define an error heuristic E to quantify the quality of our inference by weighting errors in the rate function by the probability of the system to have been observed in that state, relative to the overall average of the rate function $\langle f_{\text{true}}\rangle$:

$$E = \sum_{x_2} \frac{|f_{\text{inferred}}(x_2) - f_{\text{true}}(x_2)|}{\langle f_{\text{true}}\rangle} P(x_2) \quad . \tag{5}$$

To illustrate how the relative time-scale of X_2 and X_3 affect this inference error E we consider the above "noise enhancing" system (Materials & methods) for which changing lifetimes did not introduce different system dynamics. For such systems, inferring $f(x_2)$ is straightforward when the variability of X_2 is slow such that X_3 has enough time to adjust to X_2-levels and the conditional average $\langle x_3|x_2\rangle$ directly identifies the production rate of X_3 (SI). For faster upstream fluctuations the effect of X_2-variability on X_3 decreases and the inference of the production rate becomes more challenging. However, compared to the naive statistical approach of interpreting conditional averages as rates, our inference algorithm based on Eq (4) reliably identifies the correct rate function even when the time-scale of X_2-fluctuations is fast relative to X_3, see Fig 3B. When the upstream variably becomes more than an order of magnitude faster than X_3 our inferred rate function deviates significantly from the true one when inferred from a joint probability distribution constructed from $N = 100,000$ samples. However, even in this unfavourable regime with a 40-fold separation of time-scale between the upstream and downstream variable such that the conditional average $\langle x_3|x_2\rangle$ levels-off, $N = 5 \times 10^6$ were enough sample observations to correctly infer $f(x_2)$, see Fig 3D.

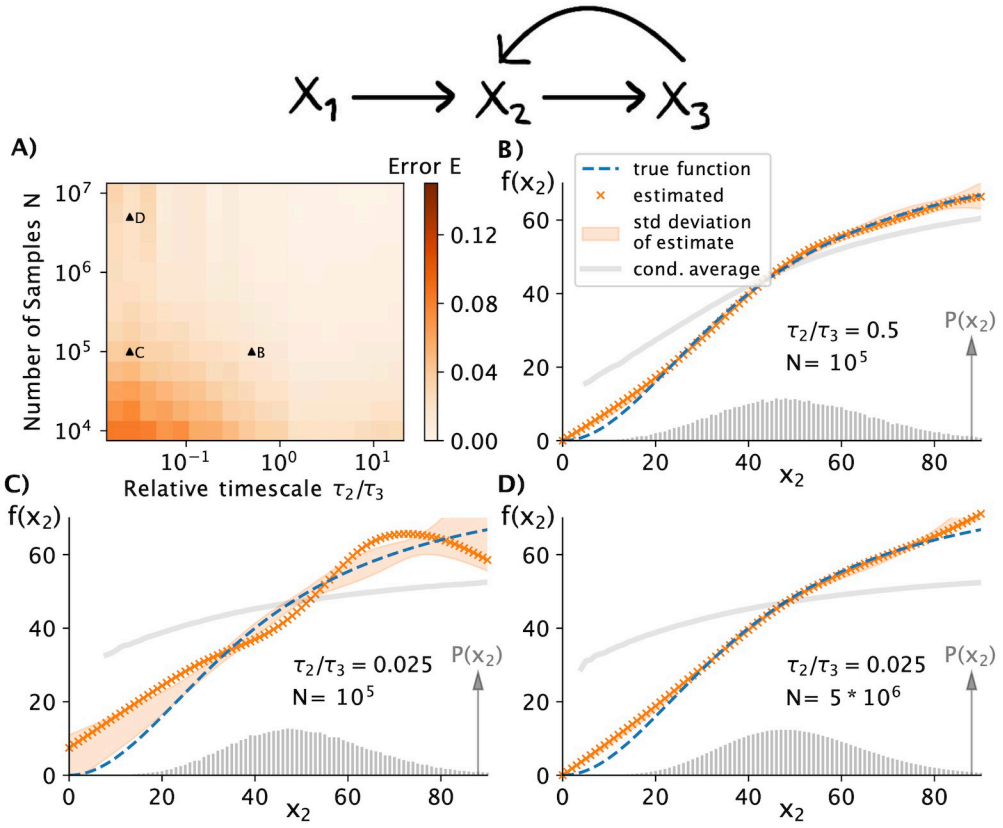

Fig 3. Experimentally achievable sampling leads to accurate inference even when fast upstream variability masks rate dependencies. A) The number of data points required to successfully infer $f(x_2)$ from an empirical $P(x_2, x_3)$ depends on the relative time-scales between the two components of interest. Simulations of the noise enhancing three-component system (Materials & methods) show that several thousand measurement samples can be enough to reliably infer the rate dependence $f(x_2)$ when upstream fluctuations are relatively slow, i.e., $\tau_2 > \tau_3$. B) When upstream fluctuations are fast, the downstream variable does not have time to adjust and the conditional average $\langle x_3|x_2\rangle$ no longer follows $f(x_2)$ as indicated by the grey line. In contrast, our inference method based on Eq (4) accurately estimates the actual rate function from $N = 100,000$ samples. The orange crosses depict an example of one inference while the shaded area displays the standard deviation of individual inferences from different samples of the same process. C) As the upstream fluctuations in X_2 become faster, the inference gets worse when using the same number of sampling points. However, even for systems in which X_2, and X_3 time-scales are separated 40-fold, the production rate $f(x_2)$ can be accurately inferred from $N = 5 \times 10^6$ samples (panel D). Such sampling is experimentally achievable using flow-cytometry approaches to characterize single-cell heterogeneity [28].

https://doi.org/10.1371/journal.pcbi.1010183.g003

Note, that the parameter τ_3 is structurally unidentifiable by our approach. However, if it is unknown, our approach will still determine $\tau_3 \cdot f(x_2)$, i.e., we can identify the shape of the rate function but not its scale.

Upregulated *vs.* downregulated production rates

While the above examples exhibit vastly different global system dynamics the functional form of the production rate of X_3 was conserved across all systems. Next, we demonstrate that our inference method works for arbitrary Hill-type functions for the production rate. We simulated the noise enhancing three-component system (Materials & methods) from Fig 3 with differently shaped Hill-functions $f(x_2) = \lambda x_2^n/(x_2^n + K^n)$ by systematically varying the parameters K, n. The tested range of parameters reflects biologically relevant different shapes, including negative n corresponding to X_2 suppressing X_3, small values of n such that X_3 is barely

affected by X_2, as well as strongly cooperative effects with $n \to \pm 4$. As illustrated in Fig 4A, the inference works satisfactorily for $N = 100,000$ across a broad range Hill-functions with specific examples of successful inference depicted in Fig 4C, 4D and 4E. Note, that those rate functions could not be inferred for states that were never (or extremely rarely) observed.

To determine the cause of unsatisfactory inferences as illustrated by an example in Fig 4F, we utilize the general noise propagation relation [18] to describe all systems within the class of Eq (3)

$$\underbrace{\frac{\mathrm{Var}(x_3)}{\langle x_3\rangle^2}}_{\eta_{x_3,x_3}} = \frac{1}{\langle x_3\rangle} + \underbrace{\frac{\mathrm{Cov}(x_3, f(x_2))}{\langle x_3\rangle\langle f(x_2)\rangle}}_{\eta_{x_3 f}} \quad , \quad I := \frac{|\eta_{x_3 f}|}{\eta_{x_3,x_3}}. \tag{6}$$

Here, I quantifies how much X_2-fluctuations affect X_3-levels. We find that regions of unsatisfactory inference correspond to systems in which the upstream variability has only a small effect on X_3 (compare panels A and B of Fig 4) Inference in the regime where $n \approx 0$ can be significantly improved through a simple cross-validation step as discussed in a later section. Alternatively, enforcing additional constrains such as monotonicity can overcome this problem when applicable (SI).

Non-linear degradation rates

In all of the above systems, X_3-molecules were degraded in a first-order reaction as is commonly the case for cellular components [29, 30] and would be approximately true for all cellular components that are not actively degraded but effectively diluted by cellular growth [31]. Next, we demonstrate that our inference method works equally well for systems in which X_3 undergoes non-linear degradation reactions.

Analogous to Fig 4, we varied the shape of the production rate $f(x_2)$ in a class of noise enhancing (Materials & methods) systems with the following conserved part

$$\begin{aligned} x_3 &\xrightarrow{f(x_2)} x_3 + 1 \\ x_3 &\xrightarrow{\gamma x_3(x_3-1)} x_3 - 2 \end{aligned} \quad + \underbrace{\left[\begin{array}{c} \mathrm{X}_1, \mathrm{X}_2 \\ \\ \text{production \& degradation} \end{array}\right]}_{\text{noise enhancing feedback}} \tag{7}$$

where a non-linear degradation rate corresponding to a dimerization event was added. We again find that our inference method works reliable for most parameters, see Fig 5A, with unsatisfactory results corresponding again to parameter regimes in which the upstream variable has only a marginal effect on the downstream fluctuations.

Additionally, we analyzed how the inference quality of $f(x_2)$ behaves for individual states. As intuitively expected, we find that the inference error initially decreases $\propto 1/\sqrt{N}$ as the number of samples N increases. However, for large N a plateau becomes apparent that is most severe for the most rarely observed states, Fig 5B. This behaviour was generally observed across different classes of systems (SI). Explicitly accounting for the effects of sampling error may lead to lower plateaus for the estimation errors with more advanced statistical methods to invert Eq (4) but are beyond the scope of the current work.

Note, in these systems the component of interest X_3 degrades as a dimer, such that its specified degradation rate in Eq (7) is non-linear, and the reaction eliminates two molecules at a time. The probability balancing Eq (2) therefore no longer involves just neighbouring states and detailed balance is broken. Instead, the above systems must satisfy the following balance

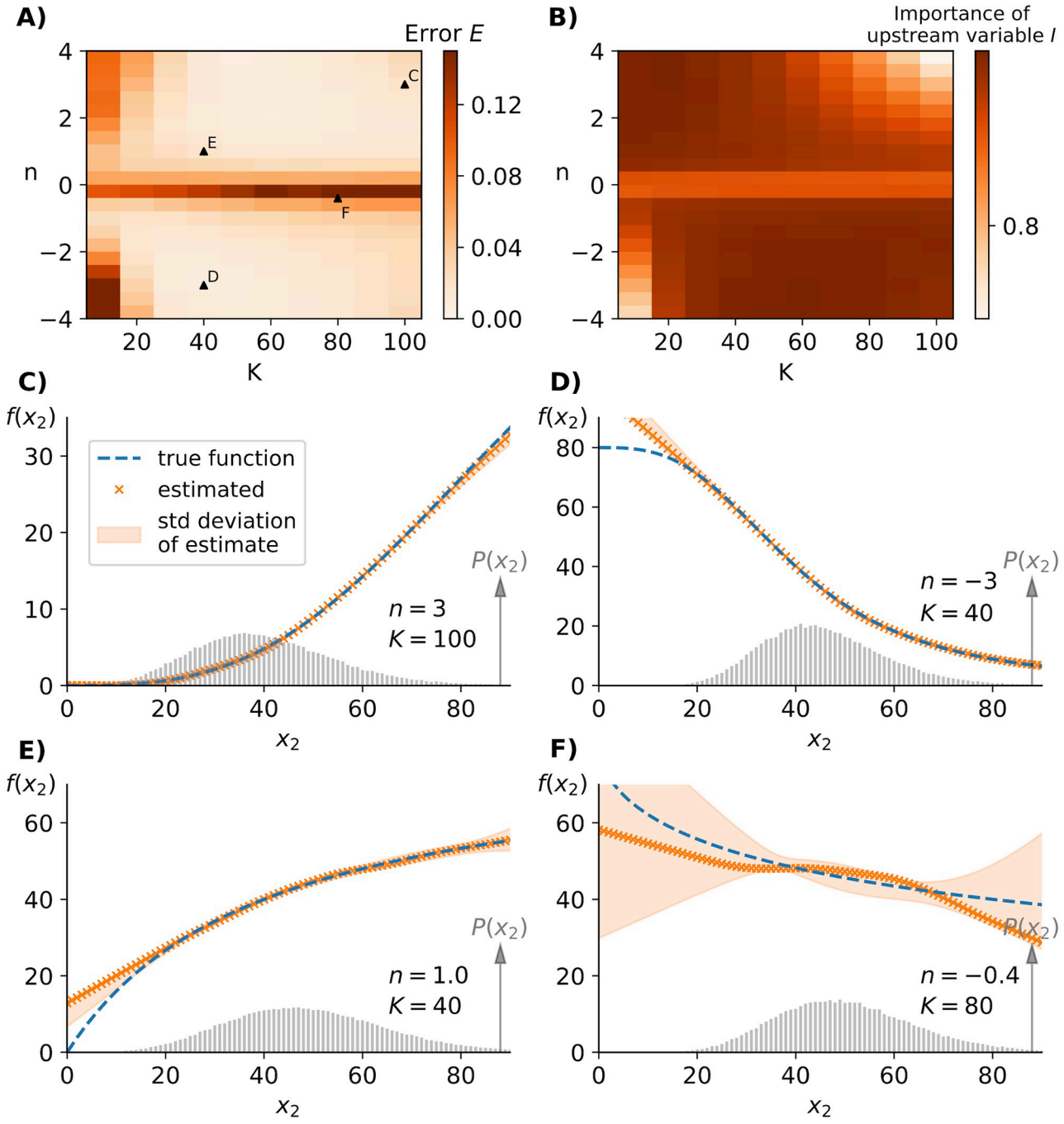

Fig 4. Different shapes of rate functions can be inferred. A) Changing the shape of the production rate $f(x_2) = \lambda x_2^n/(x_2^n + K^n)$ by varying K and n while keeping time-scales fixed and equal, we find that the inferred reaction rates using Eq (4) and $P(x_2, x_3)$ agree well with the true rate for $N = 100,000$ samples across a broad range of the parameter regime. Data shown are for the same noise enhancing three-component system (Materials & methods) as in Fig 3. B) Unsatisfactory inference of $f(x_2)$ corresponds to regimes in which the upstream variable has only little influence on the downstream fluctuations as quantified by the relative importance term I defined in Eq (6). C,D,E) Successful inference examples for different Hill-functions including repressing effects of X_2 on X_3. Any deviations from the true rate function are in the region in which the system is rarely or never observed. F) Example of unsatisfactory inference when the reaction rate varies only little over the majority of observed states resulting in a small effect of X_2-fluctuations on X_3-levels. The poor inference is also highlighted by the extremely broad shaded region indicating the standard deviation of inferred $f(x_2)$ for identical systems subject to different random sampling.

https://doi.org/10.1371/journal.pcbi.1010183.g004

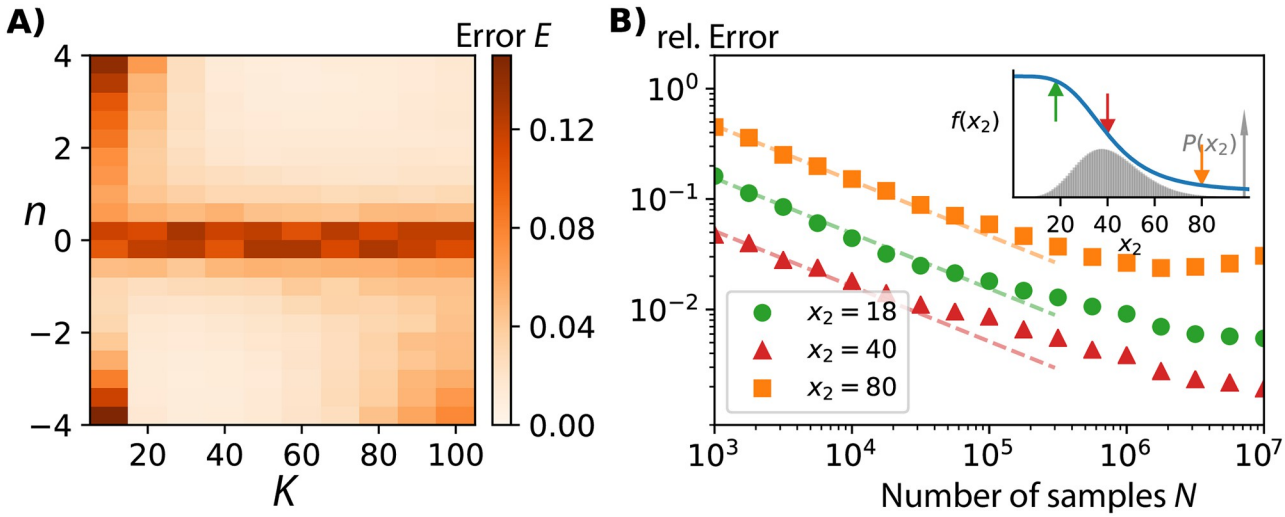

Fig 5. Non-linear degradation does not affect the inference. A) For simulated systems with non-linear degradation of X_3-molecules as defined in Eq (7), the inference quality is satisfactory when using empirically determined joint probability distributions $P(x_2, x_3)$ from $N = 100,000$ samples. Plotted are the inference error E for different production rates $f(x_2) = \lambda x_2^n/(x_2^n + K^n)$. Poor inference corresponds to parameter regimes in which the upstream variable X_2 has only a negligible effect on the downstream variable X_3, i.e., when K or n are small. Data are for the same noise enhancing three-component system as in Fig 3 with the only difference that the degradation of X_3 is now non-linear. B) For a given state, the relative error of the inferred reaction rate $f(x_2)$ initially decreases $\propto 1/\sqrt{N}$ (dashed lines) as the number of sampling points N increases. However, for large N, the relative error levels off, with higher probability states reaching a lower plateau than those only visited rarely. The inset depicts the true function and the specific states considered here, as well as the resulting probability distribution of x_2.

https://doi.org/10.1371/journal.pcbi.1010183.g005

equations $\forall m \in \mathbb{N}_0$

$$\gamma(m+1)mP(x_3 = m+1) + \gamma(m+2)(m+1)P(x_3 = m+2) = P(x_3 = m)\langle f(x_2)|x_3 = m\rangle. \quad (8)$$

Experimental measurement noise

Due to finite sampling and unavoidable measurement noise, empirically observed probability distributions will not perfectly reproduce the stationary distributions of the underlying chemical reaction network. Next, we analyze how measurement noise affects our inference method by explicitly accounting for small absolute and relative error terms as well as systemic undercounting of molecules.

To simulate "empirically observed" probability distributions of the above noise enhancing system (Materials & methods) we resampled from the exact stationary distribution $P(x_2, x_3)$ while adding a two-dimensional normally distributed error with zero mean and a standard deviation of $\sigma_{abs} = 1, 3, 8$ molecules respectively (SI). For small absolute errors, we find that the inference method still succeeds to satisfactorily determine the original rate function, see Fig 6B. Note, adding Gaussian noise can lead to negative numbers of molecules. In our analysis of additive noise we discarded those negative "measurements" and re-normalized the resulting distribution.

Furthermore, we analyzed the effect of relative measurement noise by multiplying each sampled data point with a two-dimensional normally distributed error term to simulate a relative error of 1%, 5%, 20% in the observed variables respectively (SI). In its current form, our inference is significantly affected by large multiplicative noise because it causes the "measured" probability distribution to differ significantly from the underlying stationary distribution of

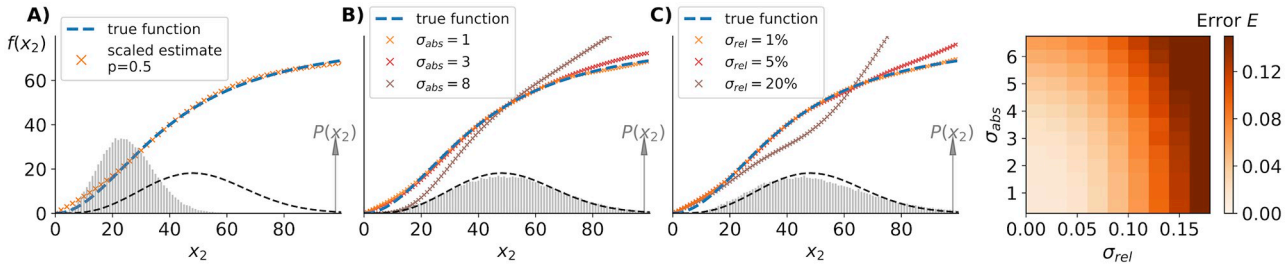

Fig 6. Small measurement noise does not prohibit accurate inference. A) Simulated "empirical" data (histogram) with binomial undercounting in which each molecule is detected with a fixed probability p. Our inference algorithm identifies the correct reaction rate if we simply multiply the measured molecule numbers by $1/p$ again to obtain the input function (dashed blue line). If we do not know p then the algorithm identifies the correct shape $f(x_2)$ but cannot identify the correct scale of X_2 over which it varies. B) Simulated "empirical" data (grey histogram) with added absolute measurement errors modelled as a two-dimensional Gaussian with zero mean and standard deviation of $\sigma_{abs} = 8$. This distribution is not identical to the theoretical one (dashed black line) and leads to significant deviations in the inference (brown crosses). For smaller absolute errors we observe satisfactory inference, as illustrated by the data for $\sigma_{abs} = 1, 3$. C) Simulating relative measurement errors by multiplying "observed" samples from the exact stationary distribution with a random number from a two-dimensional Gaussian with mean one and standard deviation of $\sigma_{rel} = 0.01, 0.1, 0.2$. For relative errors less than 10% we found satisfactory inference but larger errors led to unacceptable estimates for the production rate $f(x_2)$ because of the significant deviation of the "empirical" distribution (grey histogram) from the exact stationary distribution (dashed black line) illustrated for $\sigma_{rel} = 0.2$. Explicit de-convolution steps for known types of measurement noise may significantly improve the inference performance of future algorithms based on Eq (2). D) Quantifying the inference error for absolute and relative measurement errors. Relative measurement errors larger than 10% led to unsatisfactory inference.

https://doi.org/10.1371/journal.pcbi.1010183.g006

the stochastic process, see Fig 6C. Future variants of our inference algorithm may potentially improve on this by performing explicit de-convolution steps [32] to estimate stationary state distributions from experimentally recorded ones before exploiting Eq (2). In fact, measuring error due to probabilistic undercounting can be exactly accounted for by determining the probability p to detect a specific molecule, and applying our inference method to the re-scaled probability distribution, see Fig 6A.

Weakly connected components

Our presented method relies on knowing that one component directly affects the production rate of another. We thus obtained unsatisfactory inferences when the upstream variable barely affects the downstream variable, as illustrated in the regime when $n \to 0$ or $K \ll \langle x_2 \rangle$, see Fig 4A.

Breakdown of satisfactory inference in that regime can be prevented by explicitly considering the possibility that $f(x_2)$ is approximately constant across the observed stochastic fluctuations of X_2. Following a standard cross-validation approach, we can use one half of the observed data as a "training set" [33]. Using this subset of data we infer $f(x_2)$ using our usual unconstrained method but additionally perform a constrained optimization for constant production rates, i.e., we find the best $f(x_2) = \lambda$ for some $\lambda > 0$. This by itself will not pick a constant production rate over the freely optimized $f(x_2)$ because the latter has many more degrees of freedom. However, because the free optimization overfits sampling errors when minimizing deviations of Eq (4) in the regime in which $f(x_2)$ is approximately constant, it will do relatively worse than a constant production rate when applied to the "validation set" of the data. Incorporating this cross-validation approach into our inference methods as detailed in the Materials & Methods, removes the most unsatisfactory regime while leaving the successful inferences unaffected as illustrated in Fig 7A (compare to Fig 4A) with an example detection of constant $f(x_2)$ illustrated in Fig 7B.

Identifying constant production rates is a special case of the general problem of identifying which component affects which in complex biochemical reaction networks [34]. Future

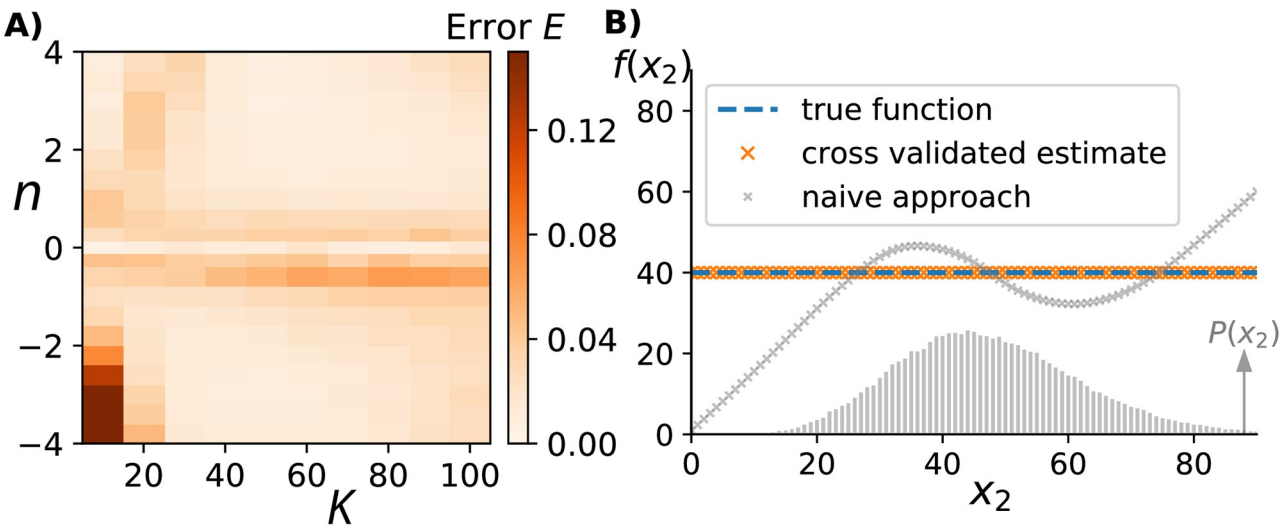

Fig 7. Cross-validation improves inference in regions with constant to near constant rates. A) Analogous to Fig 4 we change the shape of the production rate $f(x_2) = \lambda x_2^n/(x_2^n + K^n)$ by varying K and n. Shown here is the error of the inferred $f(x_2)$ when applying our method with an additional cross-validation step as detailed in the Materials & Methods. The quality of the inference is significantly improved in the regime in which the production rate is essentially independent of the upstream variable around $n \approx 0$. B) Directly comparing the cross-validated and freely inferred $f(x_2)$ when applied to a system in which the true production rate was constant. The cross-validated answer picks out the constant production rate, whereas the naive approach gives a varying estimate due to overfitting of Eq (4).

https://doi.org/10.1371/journal.pcbi.1010183.g007

variants of such a cross-validation approach might thus prove useful in identifying the topology of network interactions based on Eq (4).

Limitations of the presented evidence

Multi-variable dependent rate function. In principle, the basic idea of matching probability fluxes can be applied to components with more than one birth and death reaction as illustrated by the complete generality of Eq (2) which makes no assumption about the number of production and degradation reactions for the species of interest. However, in our proof-of-principle examples we successfully specifically inferred univariable rate functions $f = f(x_2)$. Whether our algorithm can also successfully identify more complex rate functions such as $f = f(x_2, x_1)$ will require future work.

Importantly, the value of our approach lies not necessarily in identifying arbitrarily complex rate functions but in identifying simple rate functions in arbitrarily complex systems. The benefit of our approach lies in the fact that it effectively "eliminates" network complexity when reconstructing simple rate laws for local interactions between variables. Note, many complex biological control systems indeed involve many simple interactions which in turn lead to complex systems through feedback loops and an overall large number of interacting components [35].

Experimental accuracy. As presented, our method requires highly accurate and discrete probability distributions which is motivated by measuring small numbers of molecules in cells where such discreteness is fundamentally built into the problem and single-molecule accuracy is possible [36–40].

This type of data is at the forefront of what is technologically possible [41, 42] and many current real-life experimental methods will be much less accurate than our simulated input distributions. We have partially addressed this issue by analyzing the effect of experimental errors in the form of Gaussian noise and probabilistic undercounting, see Fig 6. However in

real-world experiments, the discreteness of measurements may become completely washed out and sampling might not be as high as in our input distributions obtained from 100,000 data points over around 100 discrete states. Generalizing our approach to continuous distributions and more sparsely sampled distributions with significant sampling error is left for future work.

Growing and dividing cells. As is, our method cannot be directly applied to experimental data from growing and dividing cells. Future work is required to demonstrate that our approach could potentially be exploited to infer production rates of components from single-cell data.

In the simplest case, biochemical reactions in growing and dividing cells can be modelled as cyclo-stationary processes of periodically changing probability distributions [43]. In principle, Eq (2) can be generalized to time-averages of time-dependent distributions (see supplement of [18]) but all our proof-of-principle examples considered a stationary component of interest undergoing first or second order degradation in the absence of cell divison. In contrast, much recent progress has been made to analyze stochastic biochemical kinetics in cellular models that include cell-division, gene replication, growth dependent effects, and variability in cell-division times [44–49]. Future work should bring together the approach developed in this manuscript with the existing work on modelling cell-cycle effects.

Additionally, in growing populations time-averaged ensemble averages do not agree with population averages because the former corresponds to a uniform cell-age distribution whereas in a growing population we must have twice as many newborn cells as dividing cells which leads to a non-uniform age-structure [50]. As presented, our evidence corresponds to cellular populations of constant size. Future work is required to show the same approach works in populations with growing number of cells, which in principle can be related to the time-averages in the cyclo-stationary version of Eq (2) as long as the age-structure of cells in the population is known [25].

Formulating stochastic models in terms of concentrations rather than abundances might partially resolve these issues because on average concentrations are the same at the beginning and the end of the cell-cycle unlike abundances which must double on average.

Discussion

Early aeronautical engineering faced a challenge analogous to that of current synthetic biology. While the qualitative requirements for motored flight were well known, designing a reliable flyer required a breakthrough in *quantitatively* understanding each subproblem through painstaking experimental measurements [51].

Similarly, designing reliable synthetic circuits in biology requires a quantitative description of biochemical reaction dynamics rather than qualitative network interaction models. Here, we present a method that promises to iteratively turn a qualitative model of biochemical interactions into a network of quantitative reaction rates. Given one arrow within a network of interacting components, our method can identify the functional dependence of the actual reaction rate, one interaction at a time, from fluctuations of a subset of components without having to perturb the system.

Existing mechanistic modelling approaches often rely on temporal data [52–59] and generally have to be identified from data in one fell swoop [60–65] with all the reliability issues that come with optimizing over many degrees of freedom at once. Prior work has established approaches that can infer a small number of parameters from perturbation observations of a subset of components in otherwise completely specified multivariable models that were assumed to be a complete and correct representation of the process of interest [19, 66–68].

Some work specifically determines instantaneous production rates from static snapshots of partially observed networks [69], but these previous mechanistic approaches estimate rates and parameters by assuming (near) complete knowledge of the entire network dynamics whereas our approach does not rely on such information. Instead we leave all dynamics unspecified that do not directly affect the component of interest.

While approaches that combine mechanistic models with Bayesian inference [70–73] have been used to account for significant noise in experimental data, they too rely on inferring parameters for fully defined mechanistic models all at once. Statistical approaches that rely on perturbations [7–11] can be straightforwardly applied to static snapshots of incompletely observed complex systems but fail to account for the dynamic ways in which one component affects another, and thus do generally not quantitatively describe physical interactions between components.

Our results show that we can reliably identify reaction rates independent of the larger network structures, in contrast to previous approaches in which small models of gene expression were inverted under the assumption they were correct as a whole and only a handful of parameter values needed to be determined [36, 74, 75].

The presented numerical proof-of-principle results establish that our presented algorithm works across a broad range of tested systems and that the number of required observations for algorithmic success is comparable to those accessible by modern experimental single-cell techniques in biology such as flow-cytometry that routinely measure millions of isogenic cells at a time.

As presented, our algorithm requires highly accurate input distributions, and we cannot exclude the possibility that our method works only in theory but not in practice. However, here we presented only the most basic inference algorithm to extract information from the probability flux balance relations. We thereby established that the information contained in Eq (2) is enough to reconstruct the shape of biochemical rate functions in principle. Hybrid approaches that, e.g., combine statistical tools and neural networks with first-principle mathematical modelling have been successfully used to identify parameters in biochemical reaction network [19, 76]. We thus expect future work that combines advanced statistical methods with our first-principle approach of extracting information from Eq (2) to outperform our naive algorithm in the face of significant sampling or measurement error. Additionally, more general approaches of utilizing Eq (2) may in the future be able to identify the network topology within interaction networks from static snapshots of joint probability distributions. This is a problem of immense current interest in systems biology, e.g., for detecting causal interactions from static snapshots of single-cell RNA-seq data.

Materials & methods

Basic algorithm

To utilize probability distributions that contain sampling errors we write the balance Eq (4) for our example systems as $G\mathbf{f} = \mathbf{h}$, where $G_{ij} = P(x_2 = j, x_3 = i)$, $f_i = f(x_2 = i)$ and $h_i = (i + 1)P(x_3 = i + 1)/\tau_3$. To avoid overfitting the solution $\mathbf{f}$ to sampling noise we add a regularization term that effectively penalizes the discrete "second derivatives" $f_{i+2} - 2f_{i+1} + f_i$. The effect of this regularization is to smoothen the resulting reaction rate function. Motivated by complex cellular processes where $f(x_2)$ represent biochemical reaction rates, we furthermore constrain the solution vector $\mathbf{f}$ to be non-negative. We thus solve the following optimization problem

$$\min_{\mathbf{f}} \quad \{||G\mathbf{f} - \mathbf{h}||^2 + \epsilon||\Gamma\mathbf{f}||^2\} \quad \text{s.t.} \quad \mathbf{f} \geq 0 \tag{9}$$

where $\Gamma_{ij} = \delta_{i,j} - 2\delta_{i,j+1} + \delta_{i,j+2}$ is the regularization matrix and ϵ corresponds to the strength of the smoothening. The strength of this regularization parameter affects the quality of the inference. We found $\epsilon = 1/\sqrt{N}$ to lead to satisfactory results because it appropriately decreases in strength as the number of sampling points N increases. This regularization was used throughout the paper. In other applications, alternative heuristics to choose ϵ may prove useful to achieve satisfactory results (SI). Note, that if the lifetime τ_3 is unknown, the inference method will still correctly identify $\tau_3 \cdot f(x_2)$ meaning that we get the correct shape of the reaction rate function up to an unknown scale-factor.

We solved Eq (9) using a standard convex optimization approach which is guaranteed to converge to the optimal solution as a linear program [77]. The code is provided at https://github.com/t-wittenstein/quantify-biochemical-rates. While it is straightforward to add further constraints, such as monotonicity, about the solution function $f(x_2)$ we here deliberately only present results without making any additional assumptions beyond Eq (9). Alternatively one could utilize additional information about $f(x_2)$ and, say, fit a Hill-function directly by minimizing Eq (9). However, this necessarily involves a non-linear optimization routine to find the best-fit parameters K, n, λ which is in general less robust than the linear optimization problem we solve above.

Cross-validation algorithm

In order to avoid overfitting systems when production rates are approximately constant we compared our inferred production rate against a constant one as follows: We divided the sampling data into two equally sized sets, and applied our inference method first to the "training" set and then checked against a constant production rate using the second "validation" set [33]. To compare the two rates we calculated the violation of Eq (4) as the sum of all squared errors. If the constant production rate's error was smaller or within a 5% margin of the freely fitted production rate from the training set, we determined a constant production rate to be the best inference and optimized it over the whole data. Otherwise we applied our regular inference method to the full data set instead.

Definition of example systems

We considered simple three-component systems in which the production and degradation rates of X_1,X_2,X_2 took the following form

$$\begin{array}{ccc} x_i & \xrightarrow{f_i(\mathbf{x})} & x_i + 1 \\ & & \\ x_i & \xrightarrow{x_i/\tau_i} & x_i - 1 \end{array} \quad . \tag{10}$$

In particular, we simulated the following example systems depicted in Fig 2 where the production and degradation rate for X_3 was kept the same as specified in the main text while the other components were subject to the following dynamics.

- Bistable system: $f_1(x_2) = \lambda x_2^n/(K^n + x_2^n) + c, f_2(x_1) = x_1$ with $\lambda = 50$, $n = 6$, $K = 37$, $c = 15$ and life-times $\tau_1 = \tau_2 = 1$.
- Oscillating system: $f_1(x_3) = \lambda_1 K_1^{n_1}/(K_1^{n_1} + x_3^{n_1}), f_2(x_1) = \lambda_2 x_1^{n_2}/(K_2^{n_2} + x_1^{n_2})$ with $\lambda_1 = 50000$, $n_1 = 10$, $K_1 = 0.1$, $\lambda_2 = 80$, $n_2 = 1$, $K_2 = 100$ and life-times $\tau_1 = \tau_2 = 1$.
- Noise controlling system: $f_1(x_3) = \lambda_1 x_3, f_2(x_1) = \lambda_2 K_2^{n_2}/(K_2^{n_2} + x_1^{n_2})$ with $\lambda_1 = 50$, $\lambda_2 = 3000$, $n_2 = 10$, $K_2 = 10$ and life-times $\tau_1 = 50$, $\tau_2 = 1$.

- Noise enhancing system: $f_1 = 5, f_2(x_1, x_3) = \lambda x_3^n/(K^n + x_3^n) + cx_1$ with $\lambda = 25$, $n = 4$, $K = 50$, $c = 8$ and life-times $\tau_1 = \tau_2 = 1$.

Data in Figs 3–7 correspond to the above noise enhancing system with modifications as specified in the main text. The systems with non-linear degradation rate $\gamma x_3(x_3 - 1)$ were simulated with $\gamma = 2$.

Supporting information

S1 Text. Supplementary Information of the paper: "Quantifying biochemical reaction rates from static population variability within incompletely observed complex networks". Detailed description of simulation process, models, and supplementary data.
(PDF)

Acknowledgments

We thank M. Assaf for the suggestion of the regularization term and N. Lord and C. Zechner for valuable feedback on the manuscript. We thank Raymond Fan, Brayden Kell, Seshu Iyengar, Euan Joly-Smith for many helpful discussions and suggestions to improve the inference method.

Author Contributions

Conceptualization: Timon Wittenstein, Nava Leibovich, Andreas Hilfinger.

Formal analysis: Timon Wittenstein, Nava Leibovich.

Funding acquisition: Nava Leibovich, Andreas Hilfinger.

Investigation: Timon Wittenstein, Nava Leibovich, Andreas Hilfinger.

Methodology: Timon Wittenstein, Nava Leibovich.

Project administration: Andreas Hilfinger.

Validation: Timon Wittenstein, Nava Leibovich.

Writing – original draft: Timon Wittenstein, Nava Leibovich, Andreas Hilfinger.

Writing – review & editing: Timon Wittenstein, Nava Leibovich, Andreas Hilfinger.

References

1. Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome research. 2011; 21(7):1160–1167. https://doi.org/10.1101/gr.110882.110 PMID: 21543516
2. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols. 2018; 13(4):599–604. https://doi.org/10.1038/nprot.2017.149 PMID: 29494575
3. Robinson JP, Roederer M. Flow cytometry strikes gold. Science. 2015; 350(6262):739–740.
4. Zlokarnik G, Negulescu PA, Knapp TE, Mere L, Burres N, Feng L, et al. Quantitation of transcription and clonal selection of single living cells with β-lactamase as reporter. Science. 1998; 279(5347):84–88. https://doi.org/10.1126/science.279.5347.84 PMID: 9417030
5. Porichis F, Hart MG, Griesbeck M, Everett HL, Hassan M, Baxter AE, et al. High-throughput detection of miRNAs and gene-specific mRNA at the single-cell level by flow cytometry. Nature Communications. 2014; 5(1):1–12. https://doi.org/10.1038/ncomms6641 PMID: 25472703
6. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014; 343(6172):776–779. https://doi.org/10.1126/science.1247651 PMID: 24531970

7. Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, et al. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. Nature Biotechnology. 2017; 35(6):551. https://doi.org/10.1038/nbt.3854 PMID: 28459448
8. Chen R, Wu X, Jiang L, Zhang Y. Single-cell RNA-seq reveals hypothalamic cell diversity. Cell Reports. 2017; 18(13):3227–3241. https://doi.org/10.1016/j.celrep.2017.03.004 PMID: 28355573
9. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016; 539(7628):309–313. https://doi.org/10.1038/nature20123 PMID: 27806376
10. Filbin MG, Tirosh I, Hovestadt V, Shaw ML, Escalante LE, Mathewson ND, et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science. 2018; 360 (6386):331–335. https://doi.org/10.1126/science.aao4750 PMID: 29674595
11. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert JP. A general and flexible method for signal extraction from single-cell RNA-seq data. Nature Communications. 2018; 9(1):1–17. https://doi.org/10.1038/s41467-017-02554-5 PMID: 29348443
12. Liu F, Zhang SW, Guo WF, Wei ZG, Chen L. Inference of gene regulatory network based on local Bayesian networks. PLoS computational biology. 2016; 12(8):e1005024. https://doi.org/10.1371/journal.pcbi.1005024 PMID: 27479082
13. Weistuch C, Agozzino L, Mujica-Parodi LR, Dill KA. Inferring a network from dynamical signals at its nodes. PLoS computational biology. 2020; 16(11):e1008435. https://doi.org/10.1371/journal.pcbi.1008435 PMID: 33253160
14. Mayer J, Khairy K, Howard J. Drawing an elephant with four complex parameters. American Journal of Physics. 2010; 78(6):648–649. https://doi.org/10.1119/1.3254017
15. Pearl J, et al. Causal inference in statistics: An overview. Statistics surveys. 2009; 3:96–146. https://doi.org/10.1214/09-SS057
16. Millstein J, Zhang B, Zhu J, Schadt EE. Disentangling molecular relationships with a causal inference test. BMC Genetics. 2009; 10(1):23. https://doi.org/10.1186/1471-2156-10-23 PMID: 19473544
17. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics. 2005; 37(7):710–717. https://doi.org/10.1038/ng1589 PMID: 15965475
18. Hilfinger A, Norman TM, Vinnicombe G, Paulsson J. Constraints on Fluctuations in Sparsely Characterized Biological Systems. Phys Rev Lett. 2016; 116:058101. https://doi.org/10.1103/PhysRevLett.116.058101 PMID: 26894735
19. Lee D, Jayaraman A, Kwon JS. Development of a hybrid model for a partially known intracellular signaling pathway through correction term estimation and neural network modeling. PLoS Computational Biology. 2020; 16(12):e1008472. https://doi.org/10.1371/journal.pcbi.1008472 PMID: 33315899
20. Haggerty RA, Purvis JE. Inferring the structures of signaling motifs from paired dynamic traces of single cells. PLoS computational biology. 2021; 17(2):e1008657. https://doi.org/10.1371/journal.pcbi.1008657 PMID: 33539338
21. Grima R. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. The Journal of chemical physics. 2012; 136(15):04B616. https://doi.org/10.1063/1.3702848 PMID: 22519313
22. Lakatos E, Ale A, Kirk PD, Stumpf MP. Multivariate moment closure techniques for stochastic kinetic models. The Journal of chemical physics. 2015; 143(9):094107. https://doi.org/10.1063/1.4929837 PMID: 26342359
23. Mazo RM. On the discrepancy between results of Nicolis and Saito concerning fluctuations in chemical reactions. The Journal of Chemical Physics. 1975; 62(10):4244–4244. https://doi.org/10.1063/1.430277
24. Kelly FP. Reversibility and stochastic networks. Cambridge University Press; 2011.
25. Thomas P. Making sense of snapshot data: ergodic principle for clonal cell populations. Journal of The Royal Society Interface. 2017; 14(136):20170467. https://doi.org/10.1098/rsif.2017.0467 PMID: 29187636
26. Doob JL. Markoff Chains–Denumerable Case. Transactions of the American Mathematical Society. 1945; 58(3):455–473.
27. Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics. 1976; 22(4):403–434. https://doi.org/10.1016/0021-9991(76)90041-3
28. Arrigucci R, Bushkin Y, Radford F, Lakehal K, Vir P, Pine R, et al. FISH-Flow, a protocol for the concurrent detection of mRNA and protein in single cells using fluorescence in situ hybridization and flow

cytometry. Nature Protocols. 2017; 12(6):1245–1260. https://doi.org/10.1038/nprot.2017.039 PMID: 28518171

29. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO. Precision and functional specificity in mRNA decay. Proceedings of the National Academy of Sciences. 2002; 99(9):5860–5865. https://doi.org/10.1073/pnas.092538799 PMID: 11972065

30. Belle A, Tanay A, Bitincka L, Shamir R, O'Shea EK. Quantification of protein half-lives in the budding yeast proteome. Proceedings of the National Academy of Sciences. 2006; 103(35):13004–13009. https://doi.org/10.1073/pnas.0605420103 PMID: 16916930

31. Huh D, Paulsson J. Non-genetic heterogeneity from stochastic partitioning at cell division. Nature Genetics. 2011; 43(2):95–100. https://doi.org/10.1038/ng.729 PMID: 21186354

32. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nature Communications. 2019; 10(1):1–14. https://doi.org/10.1038/s41467-018-07931-2 PMID: 30674886

33. Shao J. Linear model selection by cross-validation. Journal of the American statistical Association. 1993; 88(422):486–494. https://doi.org/10.1080/01621459.1993.10476299

34. Mercatelli D, Scalambra L, Triboli L, Ray F, Giorgi FM. Gene regulatory network inference resources: a practical overview. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms. 2020; 1863 (6):194430. https://doi.org/10.1016/j.bbagrm.2019.194430 PMID: 31678629

35. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science. 2002; 298(5594):824–827. https://doi.org/10.1126/science.298.5594.824 PMID: 12399590

36. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, Hearn J, et al. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science. 2010; 329(5991):533–538. https://doi.org/10.1126/science.1188308 PMID: 20671182

37. Okumus B, Landgraf D, Lai GC, Bakshi S, Arias-Castro JC, Yildiz S, et al. Mechanical slowing-down of cytoplasmic diffusion allows in vivo counting of proteins in individual cells. Nature communications. 2016; 7(1):1–11. https://doi.org/10.1038/ncomms11641

38. Uphoff S, Lord ND, Okumus B, Potvin-Trottier L, Sherratt DJ, Paulsson J. Stochastic activation of a DNA damage response causes cell-to-cell mutation rate variation. Science. 2016; 351(6277):1094–1097. https://doi.org/10.1126/science.aac9786 PMID: 26941321

39. Sepúlveda LA, Xu H, Zhang J, Wang M, Golding I. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. Science. 2016; 351(6278):1218–1222. https://doi.org/10.1126/science.aad0635 PMID: 26965629

40. Lepore A, Taylor H, Landgraf D, Okumus B, Jaramillo-Riveri S, McLaren L, et al. Quantification of very low-abundant proteins in bacteria using the HaloTag and epi-fluorescence microscopy. Scientific reports. 2019; 9(1):1–9. https://doi.org/10.1038/s41598-019-44278-0 PMID: 31133640

41. Elf J, Barkefors I. Single-molecule kinetics in living cells. Annual review of biochemistry. 2019; 88:635–659. https://doi.org/10.1146/annurev-biochem-013118-110801 PMID: 30359080

42. Hardo G, Bakshi S. Challenges of analysing stochastic gene expression in bacteria using single-cell time-lapse experiments. Essays in Biochemistry. 2021; 65(1):67–79. https://doi.org/10.1042/EBC20200015 PMID: 33835126

43. Huh D, Paulsson J. Random partitioning of molecules at cell division. Proc Natl Acad Sci U S A. 2011; 108(36):15004–15009. https://doi.org/10.1073/pnas.1013171108 PMID: 21873252

44. Cao Z, Grima R. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. Proceedings of the National Academy of Sciences. 2020; 117(9):4682–4692. https://doi.org/10.1073/pnas.1910888117 PMID: 32071224

45. Beentjes CH, Perez-Carrasco R, Grima R. Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics. Physical Review E. 2020; 101(3):032403. https://doi.org/10.1103/PhysRevE.101.032403 PMID: 32290003

46. Jędrak J, Kwiatkowski M, Ochab-Marcinek A. Exactly solvable model of gene expression in a proliferating bacterial cell population with stochastic protein bursts and protein partitioning. Physical Review E. 2019; 99(4):042416. https://doi.org/10.1103/PhysRevE.99.042416 PMID: 31108597

47. Thomas P, Shahrezaei V. Coordination of gene expression noise with cell size: analytical results for agent-based models of growing cell populations. Journal of the Royal Society Interface. 2021; 18 (178):20210274. https://doi.org/10.1098/rsif.2021.0274 PMID: 34034535

48. Jia C, Grima R. Frequency domain analysis of fluctuations of mRNA and protein copy numbers within a cell lineage: theory and experimental validation. Physical Review X. 2021; 11(2):021032. https://doi.org/10.1103/PhysRevX.11.021032

49. Jędrak J, Ochab-Marcinek A. Contributions to the 'noise floor' in gene expression in a population of dividing cells. Scientific Reports. 2020; 10(1):1–13.

50. Powell E. Growth rate and generation time of bacteria, with special reference to continuous culture. Microbiology. 1956; 15(3):492–511. PMID: 13385433

51. Jakab PL. Visions of a flying machine: The Wright brothers and the process of invention. Smithsonian Institution; 2014.

52. Reinker S, Altman RM, Timmer J. Parameter estimation in stochastic biochemical reactions. IEE Proceedings-Systems Biology. 2006; 153(4):168–178. https://doi.org/10.1049/ip-syb:20050105 PMID: 16986618

53. Tian T, Xu S, Gao J, Burrage K. Simulated maximum likelihood method for estimating kinetic rates in gene expression. Bioinformatics. 2007; 23(1):84–91. https://doi.org/10.1093/bioinformatics/btl552 PMID: 17068087

54. Fröhlich F, Kaltenbacher B, Theis FJ, Hasenauer J. Scalable parameter estimation for genome-scale biochemical reaction networks. PLoS Computational Biology. 2017; 13(1):e1005331. https://doi.org/10.1371/journal.pcbi.1005331 PMID: 28114351

55. Feng Xj, Rabitz H. Optimal identification of biochemical reaction networks. Biophysical Journal. 2004; 86(3):1270–1281. https://doi.org/10.1016/S0006-3495(04)74201-0 PMID: 14990460

56. Schmidt H, Cho KH, Jacobsen EW. Identification of small scale biochemical networks based on general type system perturbations. The FEBS Journal. 2005; 272(9):2141–2151. https://doi.org/10.1111/j.1742-4658.2005.04605.x PMID: 15853799

57. Ruttor A, Opper M. Efficient statistical inference for stochastic reaction processes. Physical Review Letters. 2009; 103(23):230601. https://doi.org/10.1103/PhysRevLett.103.230601 PMID: 20366136

58. Kim J, Bates DG, Postlethwaite I, Heslop-Harrison P, Cho KH. Least-squares methods for identifying biochemical regulatory networks from noisy measurements. BMC Bioinformatics. 2007; 8(1):1–15. https://doi.org/10.1186/1471-2105-8-8 PMID: 17212835

59. Kitayama T, Kinoshita A, Sugimoto M, Nakayama Y, Tomita M. A simplified method for power-law modelling of metabolic pathways from time-course data and steady-state flux profiles. Theoretical Biology and Medical Modelling. 2006; 3(1):1–9. https://doi.org/10.1186/1742-4682-3-24 PMID: 16846504

60. Ashyraliyev M, Fomekong-Nanfack Y, Kaandorp JA, Blom JG. Systems biology: parameter estimation for biochemical models. The FEBS journal. 2009; 276(4):886–902. https://doi.org/10.1111/j.1742-4658.2008.06844.x PMID: 19215296

61. Mendes P, Kell D. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. Bioinformatics (Oxford, England). 1998; 14(10):869–883. https://doi.org/10.1093/bioinformatics/14.10.869 PMID: 9927716

62. Rodriguez-Fernandez M, Mendes P, Banga JR. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. Biosystems. 2006; 83(2-3):248–265. https://doi.org/10.1016/j.biosystems.2005.06.016 PMID: 16236429

63. Ocone A, Haghverdi L, Mueller NS, Theis FJ. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. Bioinformatics. 2015; 31(12):i89–i96. https://doi.org/10.1093/bioinformatics/btv257 PMID: 26072513

64. Stathopoulos V, Girolami MA. Markov chain Monte Carlo inference for Markov jump processes via the linear noise approximation. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2013; 371(1984):20110541. https://doi.org/10.1098/rsta.2011.0541 PMID: 23277599

65. Dixit PD, Jain A, Stock G, Dill KA. Inferring transition rates of networks from populations in continuous-time Markov processes. Journal of chemical theory and computation. 2015; 11(11):5464–5472. https://doi.org/10.1021/acs.jctc.5b00537 PMID: 26574334

66. Hasenauer J, Waldherr S, Doszczak M, Radde N, Scheurich P, Allgöwer F. Identification of models of heterogeneous cell populations from population snapshot data. BMC bioinformatics. 2011; 12(1):1–15. https://doi.org/10.1186/1471-2105-12-125 PMID: 21527025

67. Hasenauer J, Waldherr S, Wagner K, Allgöwer F. Parameter identification, experimental design and model falsification for biological network models using semidefinite programming. IET systems biology. 2010; 4(2):119–130. https://doi.org/10.1049/iet-syb.2009.0030 PMID: 20232992

68. Lee D, Jayaraman A, Kwon JSI. Identification of a time-varying intracellular signalling model through data clustering and parameter selection: application to NF-κB signalling pathway induced by LPS in the presence of BFA. IET systems biology. 2019; 13(4):169–179. https://doi.org/10.1049/iet-syb.2018.5079 PMID: 31318334

69. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018; 560(7719):494–498. https://doi.org/10.1038/s41586-018-0414-6 PMID: 30089906

70. Liepe J, Kirk P, Filippi S, Toni T, Barnes CP, Stumpf MP. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. Nature Protocols. 2014; 9(2):439–456. https://doi.org/10.1038/nprot.2014.025 PMID: 24457334

71. Boys RJ, Wilkinson DJ, Kirkwood TB. Bayesian inference for a discretely observed stochastic kinetic model. Statistics and Computing. 2008; 18(2):125–135. https://doi.org/10.1007/s11222-007-9043-x

72. Golightly A, Wilkinson DJ. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. Interface Focus. 2011; 1(6):807–820. https://doi.org/10.1098/rsfs.2011.0047 PMID: 23226583

73. Zechner C, Pelet S, Peter M, Koeppl H. Recursive Bayesian estimation of stochastic rate constants from heterogeneous cell populations. In: 2011 50th IEEE Conference on Decision and Control and European Control Conference. IEEE; 2011. p. 5837–5843.

74. Cai L, Friedman N, Xie XS. Stochastic protein expression in individual cells at the single molecule level. Nature. 2006; 440(7082):358–362. https://doi.org/10.1038/nature04599 PMID: 16541077

75. Friedman N, Cai L, Xie XS. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. Physical Review Letters. 2006; 97(16):168302. https://doi.org/10.1103/PhysRevLett.97.168302 PMID: 17155441

76. Lee D, Jayaraman A, Kwon JSI. Identification of cell-to-cell heterogeneity through systems engineering approaches. AIChE Journal. 2020; 66(5):e16925. https://doi.org/10.1002/aic.16925

77. Boyd S, Boyd SP, Vandenberghe L. Convex optimization. Cambridge University Press; 2004.