

Neural network aided approximation and parameter inference of stochastic models of gene expression

Qingchao Jiang^{*1}, Xiaoming Fu^{*1,2}, Shifu Yan^{*1}, Runlai Li⁴, Wenli Du¹, Zhixing Cao^{†1,3}, Feng Qian¹, and Ramon Grima^{†2}

¹*Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China*

²*School of Biological Sciences, the University of Edinburgh, Max Born Crescent, Edinburgh, EH9 3BE, Scotland, United Kingdom*

³*State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai 200237, China*

⁴*Department of Chemistry, National University of Singapore, Science Drive 3, Singapore 117543, Singapore*

December 15, 2020

Abstract

Non-Markov models of stochastic biochemical kinetics often incorporate explicit time delays to effectively model large numbers of intermediate biochemical processes. Analysis and simulation of these models, as well as the inference of their parameters from data, are fraught with difficulties because the dynamics depends on the system's history. Here we use an artificial neural network to approximate the time-dependent distributions of non-Markov models by the solutions of much simpler time-inhomogeneous Markov models; the approximation does not increase the dimensionality of the model and simultaneously leads to inference of the kinetic parameters. The training of the neural network uses a relatively small set of noisy measurements generated by experimental data or stochastic simulations of the non-Markov model. We show using a variety of models, where the delays stem from transcriptional processes and feedback control, that the Markov models learnt by the neural network accurately reflect the stochastic dynamics across parameter space.

^{*}These authors contributed equally.

[†]Correspondence and requests for materials should be addressed to Z.C. or R.G. (email: zcao@ecust.edu.cn, ramon.grima@ed.ac.uk).

Introduction

Over the past two decades, stochastic modelling has provided insight into how cellular dynamics is influenced by noise in gene expression [1,2]. The complexity of cellular biochemistry prevents a full stochastic description of all reaction events and rather these models are effective, in the sense that each reaction provides an effective description of a group of processes. A major assumption behind the majority of stochastic models of biochemical kinetics and gene expression is the memoryless hypothesis, i.e., the stochastic dynamics of the reactants is only influenced by the current state of the system, which implies that the waiting times for reaction events obey exponential distributions. While this Markovian assumption considerably simplifies model analysis [3], it is dubious for modelling certain non-elementary reaction events that encapsulate multiple intermediate reaction steps. For example consider a model of transcription that predicts the distribution of RNA polymerase numbers attached to the gene but which does not explicitly model the microscopic processes behind elongation [4]. In this case, given that RNA synthesis proceeds with approximately constant elongation speed, the reaction modelling termination should occur a fixed time after the reaction modelling initiation, which implies the system has memory and is not Markovian. Of course in this instance, the model could be made Markovian by extending it so that it includes the explicit microscopic description of the movement of the RNA polymerase along the gene [5] but this implies a large increase in the effective number of species which makes explicit solution of the Markovian model impossible. Hence in many cases, a low dimensional stochastic model can be only achieved by a suitable non-Markovian description. Given their practical importance, these systems have been the subject of increased research interest, leading to an exact analytical solution for a few simple cases and the development of exact Monte Carlo algorithms for the study of those with complex dynamics [6–11].

Nevertheless, presently the understanding of non-Markovian models lags much behind that of Markovian models where a wide range of approximation methods are available. Hence there is ample scope for the development of methods to tackle the difficulties inherent in stochastic systems possessing memory. Given the success of artificial neural networks (ANNs) in solving scientific problems where traditional methods have made little progress, it is of interest to consider whether such an approach could be useful for solving the aforementioned stochastic problems. Neural networks being universal function approximators have recently been used to solve partial differential equations commonly used in physics, biology and chemistry. In particular these techniques have been used to approximately solve Burgers equation [12–14], the Schrodinger equation [13,15] and partial differential equations describing collective cell migration in scratch assays [16]; the ANN based methods behind the solution of these problems are a subclass of the universal differential equation framework that has recently been proposed [17].

Inspired by the success of ANNs in other fields, in this article we develop a novel ANN based methodology based to study non-Markov models of gene expression and transcriptional feedback. We propose to use an ANN to approximate non-Markov models by much simpler stochastic models. Specifically the key idea is to approximate the chemical master equation (CME) of non-Markov models (which we refer to as the delay CME) that is in terms of the two-time probability distribution by a CME whose terms are only a function of the current time, i.e. by a time-inhomogeneous Markov process (see Fig. 1a for an illustration). Notably this mapping is achieved without increasing the number of fluctuating species. We refer to the learnt CME describing the time-inhomogeneous Markov process as the neural-network chemical master equation (NN-CME). The latter, because of its simplified form, can then either be studied analytically using standard methods or else straightforwardly simulated using the finite state projection (FSP) method. In what follows, we introduce the ANN based approximation method by means of a simple example and then verify its accuracy in predicting time-dependent molecule number distributions of various realistic models of gene expression and its superior computational efficiency when compared to direct stochastic simulation. We finish by showing the orthogonal use of the method to infer the parameters of bursty gene expression from synthetic data.

Results

Illustration of ANN-aided stochastic model approximation using a simple model of transcription

We consider a simple non-Markovian system where molecules are produced at a rate ρ and are removed from the system (degraded) after a fixed time delay τ :



In other words, each molecule has an internal clock that starts ticking when it is “born”, and when this clock registers a time τ , the molecule “dies”. Note that as a convention in this paper, we use an arrow to denote a reaction in which the products are formed after an exponentially distributed time and an arrow with 2 horizontal lines to denote a reaction which occurs after a fixed elapsed time. The above model, which we denote Model I, describes the fluctuations of nascent RNA (N) numbers due to constitutive expression. Specifically the production reaction models the process of initiation whereby an RNAP (RNA polymerase) molecule binds the promoter; the delayed degradation reaction models, in a combined manner, the processes of elongation and termination whereby an RNAP molecule travels at a constant velocity along the gene and finally detaches from the gene, respectively. Note that the number of RNAPs bound to the gene is equal to the number of nascent RNA since each RNAP molecule has attached to it a nascent RNA tail (for an illustration see Fig. 2a Model I). It can be shown (see SI Note 1) from first principles that the delay CME describing the stochastic dynamics of this process is given by:

$$\frac{dP(n, t)}{dt} = \rho[P(n-1, t) - P(n, t)] + \rho[P(n, t|0, t-\tau) - P(n-1, t|0, t-\tau)], \quad (2)$$

where $P(n, t)$ is the probability that at time t there are n nascent RNA molecules bound to the gene. Similarly $P(n, t|n', t')$ is the conditional probability distribution that at time t there are n molecules given that at a previous time $t' < t$, there were n' molecules. The right hand side of the master equation is a function of the present time t as well as of the previous time $t - \tau$. Master equations of this type are typically much more difficult to solve, analytically or numerically, than conventional master equations where the right hand side is only a function of the present time t . Hence the key idea of our method is to map the master equation (1) to the new master equation Eq. (2):

$$\frac{d}{dt}P(n, t) = \rho[P(n-1, t) - P(n, t)] + NN_{\theta}(n+1, t)P(n+1, t) - NN_{\theta}(n, t)P(n, t), \quad (3)$$

where the function $NN_{\theta}(n, t)$ is an effective time-dependent propensity describing the removal of nascent RNA molecules which is to be learnt by the ANN. This is the NN-CME for reaction scheme (1). Note that this master equation with $NN_{\theta}(n, t) = kn$ is the conventional chemical master equation describing the birth-death process $\emptyset \xrightarrow{\rho} N, N \xrightarrow{k} \emptyset$ where k is the degradation rate. Given the physical intuition that $P(n, t) \approx 0$ for a large enough n , we truncate the probability vector by letting $P(N+1, t) = 0$ for some N , which is indeed the main idea behind the finite state projection algorithm (FSP) [18]. By doing so, Eq. (3) can be compactly written in the form

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{A}_{\theta}(t)\mathbf{P}(t), \quad (4)$$

with $\mathbf{P}(t) = [P(0,t), \dots, P(N,t)]^\top$. The transition matrix is defined as $\mathbf{A}_\theta = \mathbf{D} + \mathbf{N}_\theta(t)$, where the two components are given by

$$\mathbf{D} = \begin{bmatrix} -\rho & 0 & \cdots & 0 & 0 \\ \rho & -\rho & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -\rho & 0 \\ 0 & 0 & \cdots & \rho & 0 \end{bmatrix},$$

and

$$\mathbf{N}_\theta(t) = \begin{bmatrix} 0 & \text{NN}_\theta(1,t) & \cdots & 0 & 0 \\ 0 & -\text{NN}_\theta(1,t) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -\text{NN}_\theta(N-1,t) & \text{NN}_\theta(N,t) \\ 0 & 0 & \cdots & 0 & -\text{NN}_\theta(N,t) \end{bmatrix}.$$

The output $\text{NN}_\theta(0,t)$ is set to 0 to reflect the fact that nascent RNA cannot be further removed when there is none attached to the gene.

Next we describe how we train the ANN to infer the effective transition matrix $\mathbf{A}_\theta(t)$. We use a multi-layer perceptron with a single hidden layer; this is a simple feedforward ANN consisting of three layers – an input layer with $N + 1$ inputs, a hidden layer with an arbitrary number of neurons and an output layer with N outputs. The simplicity of the ANN here used is motivated by the universal approximation theorem which states that a single hidden-layer feedforward ANN is able to approximate a wide class of functions on compact subsets [19, 20]. The activation functions used in the hidden layer and output layer are \tanh and relu for all examples. In the output layer, we impose an increasing set of fixed biases which we specify later. For more details of the ANN, including the choice of hyperparameters, please see SI Table 1. The training procedure consists of three main steps:

1. Given the probability vector $\mathbf{P}(t)$ at some time t , the $N + 1$ elements of the vector constitute the inputs to the ANN. Given some set of weights and biases θ , the ANN's N outputs are then taken as the elements of the matrix $\mathbf{N}_\theta(t)$, i.e. the n^{th} output of the ANN is $\text{NN}_\theta(n,t)$. By a numerical discretization of Eq. (4), given the inputs and the outputs of the ANN, we obtain $\mathbf{P}(t + \Delta t)$ where Δt is the finite time step. Hence it follows that given the initial conditions specified by $\mathbf{P}(0)$, by means of the ANN, for a given set of weights and biases, we can recursively predict the time-evolution of the probability vector $\mathbf{P}(t)$ over the time points $t \in [t_1, t_2, \dots, t_{N_{\text{shots}}}]$ where N_{shots} is the total number of snapshots.
2. Next we generate approximate probability distributions at the same time points by means of stochastic simulations of the birth delayed-degradation reaction (1). Note that by stochastic simulations in this paper we always mean an exact stochastic simulation algorithm (SSA) modified to incorporate delays (specifically Algorithm 2 in [6]; for proof of its exactness see [7]). Let these distributions be denoted as $\mathbf{H}(t)$. Following this we calculate a distance between the latter and $\mathbf{P}(t)$ calculated in the previous step.
3. If the distance is larger than a threshold then update the set of weights and biases by means of back propagation and gradient descent, and repeat steps (1) and (2). If the distance is smaller than the threshold then the procedure ends and the transition matrix $\mathbf{A}_\theta(t)$ has been learnt.

Note that since the output of the ANN are the propensities $\text{NN}_\theta(n,t)$, these must be positive. We choose the set of biases in the output layer (r_n in Fig. 1a for $n = 1, \dots, N$) to be given by $r_n = \frac{n}{\tau}$. This form is inspired by the fact that for the conventional CME with first-order degradation, $\text{NN}_\theta(n,t)$ is proportional to n where the proportionality constant is the effective rate of decay which has units of inverse time. Hence intuitively, the effective removal propensity of the NN-CME is equal to the propensity assuming first-order degradation plus a correction which is what the ANN effectively learns. This choice of

biases is also similar to that of well-known residual network (ResNet) [21,22] which helps to accelerate the convergence of training and reduce computational cost.

Note also that the objective function is chosen as $J(\theta) = \sum_{j=1}^{N_{\text{shot}}} \|\mathbf{H}(t_j) - \mathbf{P}(t_j)\|_2^2$. While there are more accurate distance measures (such as the Wasserstein distance) we use the mean-squared-error form for two reasons: (i) it is commonly used in neural network training, and (ii) its simple form allows efficient calculation of derivatives through the chain rule (the back propagation method). Steps (1) and (2) are illustrated in Fig. 1b and the whole training procedure is summarized in Fig. 1c. Note that while the gradient descent in Fig. 1c is illustrated using an Euler method, for our training we used the standard adaptive moment estimation algorithm (ADAM).

Once the matrix $\mathbf{A}_\theta(t)$ is learnt, Eq. (4) can be integrated numerically to obtain the time-dependent probability vector at all times in the future. In Fig. 2b (first row), we show that the solution of the learnt NN-CME is practically indistinguishable from distributions estimated from stochastic simulations of Model I (1) – hence this implies that the ANN training protocol is effective as a means to map a master equation with terms having a non-local temporal dependence to a master equation with terms having a purely local temporal dependence.

Testing the accuracy and computational efficiency of ANN-aided stochastic model approximation on more complex models of transcription

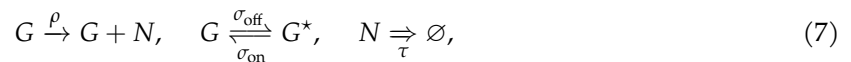
To verify that the accurate mapping characteristics are not specific to Model I, we next consider the application of the procedure to learn the NN-CME of two more complex transcription models incorporating delay (see Fig. 2a). We consider Model II which is the same as Model I except that the binding of RNAPs to the promoter occurs in bursts whose size i are distributed according to the geometric distribution $b^i / (1+b)^{i+1}$; this can be described by the reaction scheme:



where α stands for the burst frequency and b is the mean burst size. This is a minimal delay model to describe the phenomenon of transcriptional bursting [23]. The delay CME describing the nascent RNA dynamics is given by (see SI Note 2):

$$\begin{aligned} \partial_t P(n, t) &= \sum_{m=1}^{\infty} \frac{\alpha b^m}{(1+b)^{m+1}} [P(n-m, t) - P(n, t)] \\ &+ \sum_{m=1}^{\infty} \frac{\alpha b^m}{(1+b)^{m+1}} \sum_{n'} P(n', t) [P(n+m, t|n'+m, t-\tau) - P(n, t|n'+m, t-\tau)]. \end{aligned} \quad (6)$$

We also consider Model III wherein the promoter switches between an active and inactive state, RNAP binding occurs only in the active state which is followed by delayed degradation modelling the RNAP movement along the gene and its detachment; this can be described by the reaction scheme:



where G and G^* stand for the active and inactive gene state respectively, and σ_{on} and σ_{off} are the activation and inactivation rates, respectively. It can be shown that in the limit of large σ_{off} (compared to σ_{on}), Model III reduces to Model II, whereas in the opposite limit of small σ_{off} it reduces to Model I. Hence Model III can describe both constitutive and bursty transcription, as well as regimes in between. The delay CME

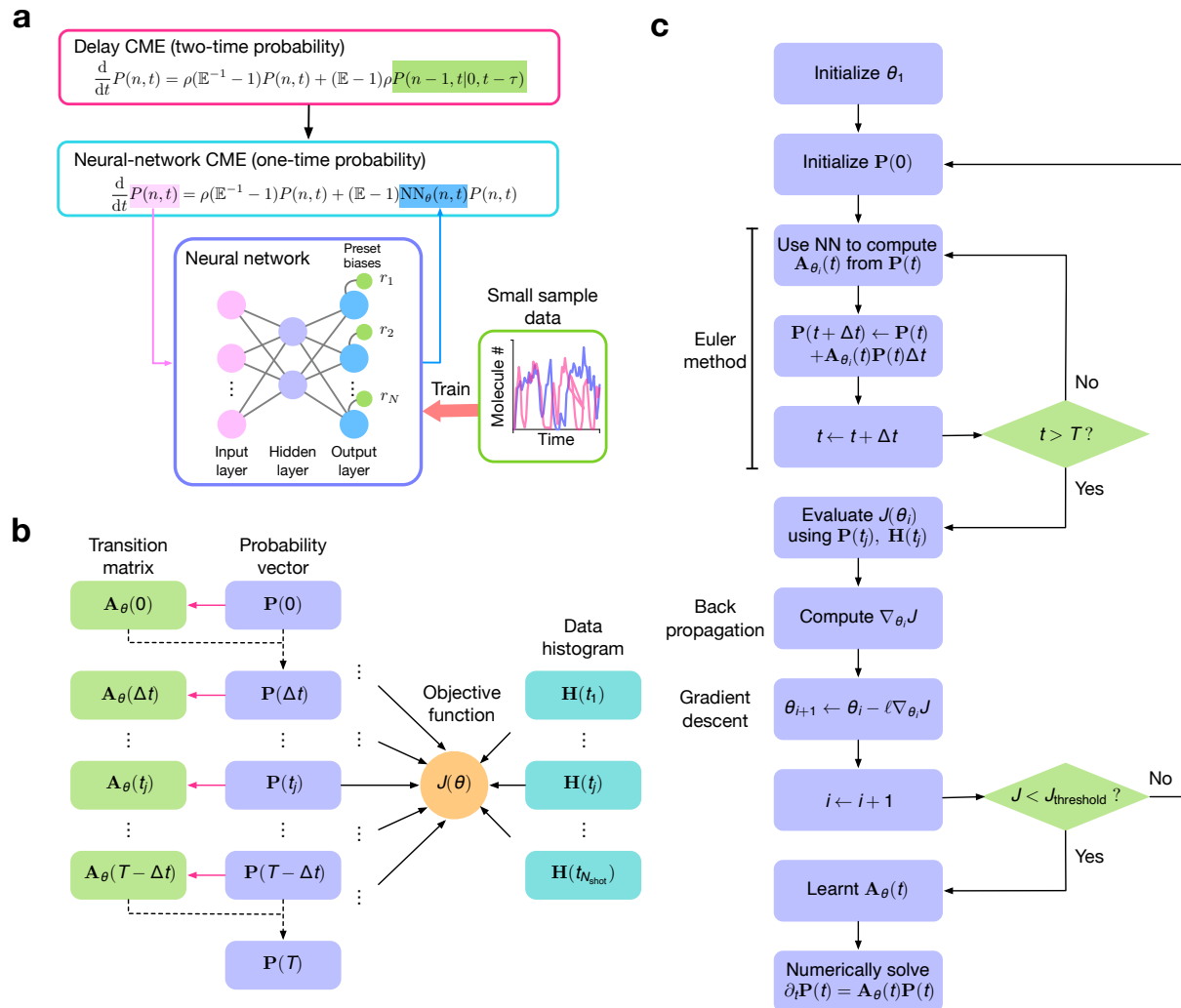


Figure 1: The ANN-aided stochastic model approximation. (a) Illustration of the key idea behind the method, namely the ANN-aided mapping of a delay master equation that is in terms of the two-time probability distribution by the simpler Neural-Network Chemical Master Equation (NN-CME) whose terms are only a function of the current time. (b) Illustration of the procedure behind the calculation of the transition matrix and the objective function. For a given set of weights and biases of the ANN (denoted by θ), taking as input $P(t)$, the ANN's output gives the transition matrix elements $A_{\theta}(t)$ which then by means of the Euler method (or more advanced differential equation solvers) is used to predict the distribution at the next time step $P(t + \Delta t)$. Note that magenta arrows show the ANN computation while the black dashed arrows show the use of the Euler method. Stochastic simulations that sample the solution of the delay master equation are used to produce histograms at several time points $H(t)$; finally the distance $J(\theta)$ is calculated between the latter and $P(t)$ (evaluated at the same time points). (c) Flowchart illustrating all the steps in ANN training. If the objective function calculated as shown in (b) is above a threshold then the weights and biases of the ANN are updated using back propagation followed by gradient descent; this is repeated until the objective function is below a threshold.

describing nascent RNA dynamics is given by (see SI Note 3):

$$\begin{aligned}\frac{dP_0(n,t)}{dt} &= -\sigma_{\text{on}}P_0(n,t) + \sigma_{\text{off}}P_1(n,t) + \sum_{n'}\rho P_1(n',t-\tau)[P_{01}(n,t|0,t-\tau) - P_{01}(n-1,t|0,t-\tau)], \\ \frac{dP_1(n,t)}{dt} &= \sigma_{\text{on}}P_0(n,t) - \sigma_{\text{off}}P_1(n,t) + \rho[P_1(n-1,t) - P_1(n,t)] + \sum_{n'}\rho P_1(n',t-\tau)[P_{11}(n,t|0,t-\tau) - P_{11}(n-1,t|0,t-\tau)],\end{aligned}\tag{8}$$

where $P_i(n,t)$ is the probability that the gene is in state i at time t and the number of nascent RNA is n ; note that $i = 0, 1$ where 0 is the inactive state and 1 is the active state. Similarly $P_{ij}(n,t|n',t')$ denotes the conditional probability distribution that at time t the gene is in state i and the number of molecules is n , given that at a previous time t' the gene was in state j and the number of molecules was n' .

Note that as for the delay CME describing Model I, the delay CMEs describing Models II and III also have terms on the right hand side which are a function of the two-time probability distribution. These terms which stem from delayed degradation, make analytical and numerical solution of the delay master equations non-trivial. However the ANN-aided procedure to solve Models II and III is as easy to implement as for Model I. By replacing the two-time probability distribution terms on the left hand sides of Eqs. (6) and (8) by terms of the type $\text{NN}_{\theta}(n,t)$ (see SI Note 4 for details), one can map the delay master equations into the form $\frac{d}{dt}\mathbf{P}(t) = \mathbf{A}_{\theta}(t)\mathbf{P}(t)$, where the transition matrix $\mathbf{A}_{\theta}(t)$ is learnt by the same training procedure as before.

In Fig. 2b rows 2, 3 and 4 we show the comparison between the time-dependent distribution of nascent RNA predicted by the NN-CME and stochastic simulations of the reaction schemes corresponding to Models II and III. The agreement is excellent at all times and for all models, independent of the modality and skewness of the distribution. This reinforces the result that the ANN-aided procedure enables an accurate mapping of master equations with terms having a non-local temporal dependence (via the two-time probability distribution) to master equations with terms having a purely local temporal dependence.

Next we test the computational efficiency of the ANN-aided procedure compared to stochastic simulations. Fig. 3a shows the Hellinger distance between the probability distribution of nascent RNA numbers according to the NN-CME and the exact analytical solution of Model I (see SI Note 1) as a function of the number of snapshots N_{shots} and of the number of stochastic simulations used to train the ANN. As expected, increasing the number of snapshots and the number of stochastic simulations in the ANN's training enhances the accuracy of the NN-CME's distribution (manifested as a reduction in the Hellinger distance). More interestingly, we found that the NN-CME obtained from training the ANN with just a thousand stochastic simulations outputs a distribution that has the same Hellinger distance from the exact distribution as the one obtained from a histogram generated using 30,000 stochastic simulations (direct simulation). Moreover in this case, the time to acquire samples plus the time for ANN training takes 1/6 of the computation time if we only use simulations. Another way of distinguishing our method and stochastic simulations is to compare the distributions predicted by both methods, given the same number of stochastic simulations; as shown in Fig. 3b, while at short times the two are comparable, at long times the NN-CME's prediction is far more accurate than that of the SSA. Note that training can also be done in steady-state, i.e. solving the algebraic equations $\mathbf{A}_{\theta}(t)\mathbf{P}(t) = 0$; the precision and efficiency of this alternative mode of training the ANN is illustrated and discussed in SI Fig. S2.

Rapid exploration of parameter space and the prediction of a novel type of zero-inflation phenomenon

Given the computational efficiency of the ANN-aided model approximation, one would expect it to be useful as a means to rapidly explore the phases of a system's behaviour across large swathes of parameter space. This endeavour is only possible if the NN-CME's predictions are accurate across parameter space which is yet to be seen since we have only shown its accuracy for few parameter sets in Figs. 2 and 3. In what follows, we explicitly verify that the NN-CME can correctly capture all the phases of Model III's

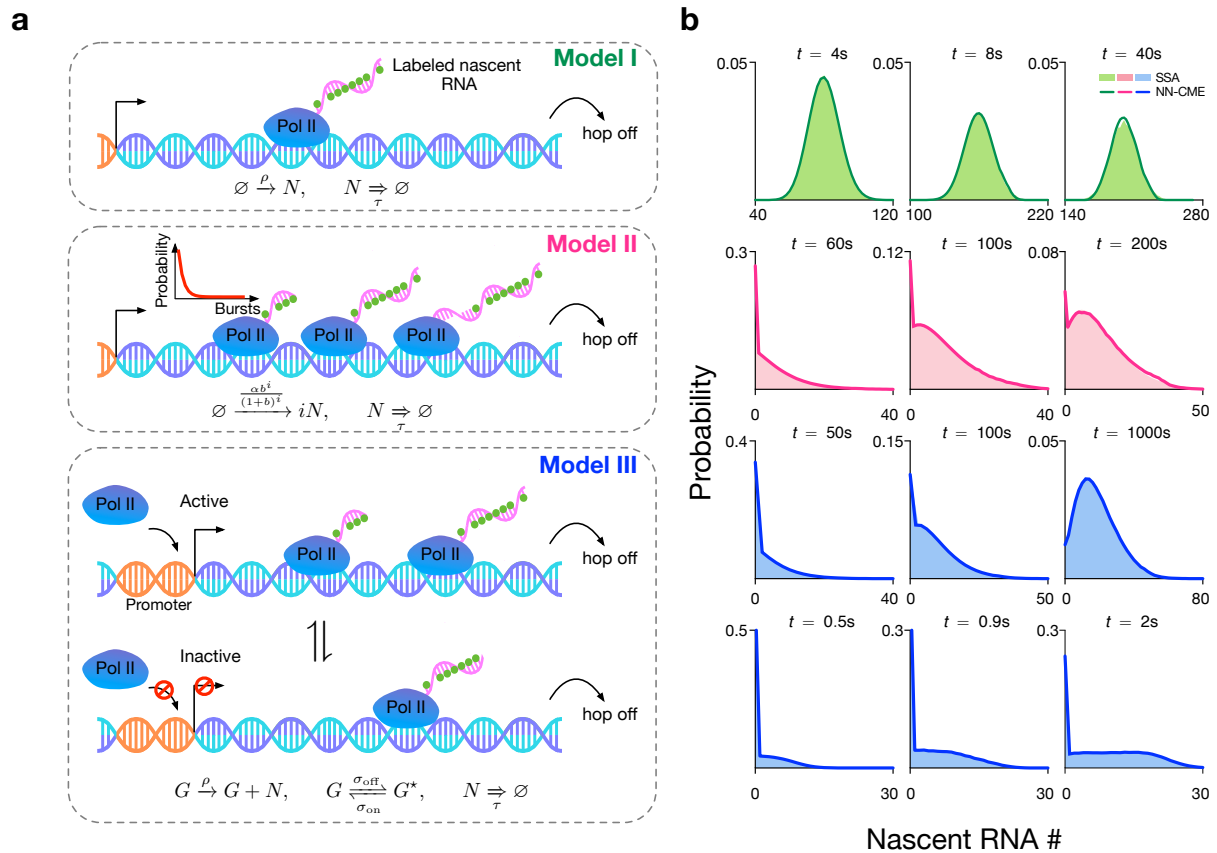


Figure 2: ANN-aided stochastic model approximation of various models of transcription. (a) Illustration of three models of transcription. The models describe initiation, elongation and termination and specifically predict the numbers of nascent RNAs (equivalently the number of RNAP polymerases, Pol IIs) at the gene locus. In all models, a nascent RNA molecule detaches after a constant time has elapsed from its binding to the promoter. The models differ in how they model Pol II binding: in Model I, the binding is modeled as a Poisson process, hence one at a time; in Model II, binding occurs in bursts, whose size conforms to a geometric distribution; in Model III, the gene switches between active and inactive states, and only the active state permits Pol II binding. (b) For all models, the FSP solution of the NN-CME derived by the ANN-aided procedure is in excellent agreement with the SSA of the delay CME. The accuracy is independent of the modality and skewness of the distribution. The rate constants and other parameters related to the ANN's training are specified in SI Table 1.

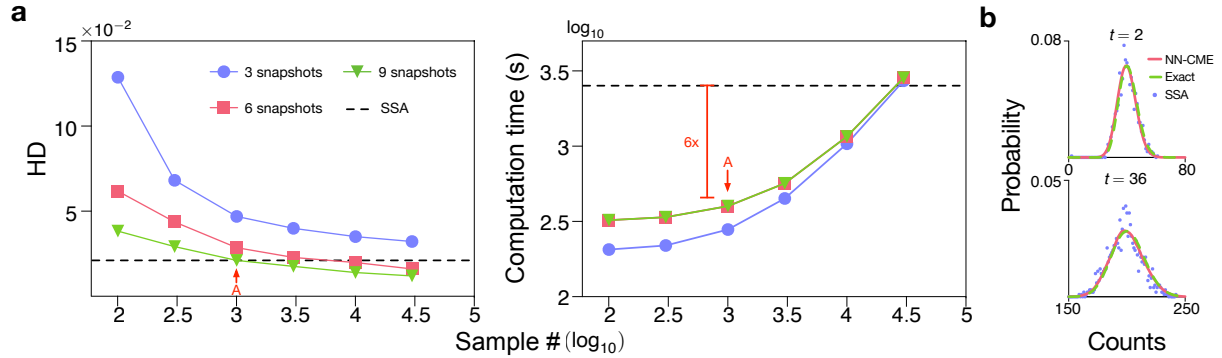


Figure 3: (a) Precision and computational efficiency of the ANN-aided model approximation as a function of sample size and number of snapshots. The method is benchmarked on Model I since the time-dependent solution of the delay CME is exactly known (see SI Note 1) and hence the accuracy of our method can be precisely quantified. A measure of the accuracy is the average Hellinger Distance (HD) between the NN-CME and exact distributions at four different time points. The computation time is equal to the time to acquire samples plus time for training. Each data point in the graphs is averaged on 3 independent trainings. Note that the NN-CME obtained from training with 10^3 samples produces a distribution that is as precise as that from 3×10^4 samples using the SSA of the delay CME (shown as a black dashed line); in this case the computation time of the NN-CME is also just 1/6 of the SSA. (b) Comparison of the NN-CME distributions, exact analytical distributions and histograms from stochastic SSA simulations of the delay CME at two different time points; the sampling for both training and the SSA is 10^3 . Note that the NN-CME leads to much more accurate distributions than the SSA for the same number of samples. The rate constants and other parameters related to the ANN’s training are specified in SI Table 1.

behaviour.

We consider the case when the gene spends most of its time in the OFF state (the bursty regime of gene expression). In this case, Model III is well approximated by Model II (see SI Note 3), and by means of the exact analytical solution of the latter we identify four regions (I-IV; see Fig 4a) according to the type of steady-state distribution (see Fig. 4b and its caption for their description). Specifically phase IV is the only region of space where bimodal distributions (peak at zero and at a non-zero value) are found. Theory shows that the conditions (see SI Note 2) that need to be satisfied for this bimodality to manifest are:

$$\frac{2 + \frac{2}{b}}{\alpha} < \tau < \frac{b + \frac{1}{b} + 2}{\alpha}, \quad b > 1. \quad (9)$$

By using the NN-CME to randomly sample points in parameter space, we find the same as the theoretical prediction: the distributions are unimodal (dots) except in Region IV where they are bimodal (crosses). Hence this verifies the accuracy of our method across parameter space.

We also note that bimodality in the bursty regime is unexpected because the standard model of gene expression (Model II/III with delayed-degradation replaced by first-order degradation [2, 24, 25]) predicts a unimodal steady-state distribution which is well approximated by a negative binomial distribution. Note that Model III is more appropriate to model nascent RNA dynamics than the standard model of gene expression because unlike mature RNA, nascent RNA typically does not get degraded while the RNAP is traversing the gene; rather nascent “degradation” occurs after a finite elapsed time when it detaches from the gene and becomes mature RNA. Hence Region IV can be understood as delay-induced bimodality or a delay-induced zero-inflation phenomenon. Since there is evidence that the delay time is stochastic rather than fixed [26, 27] we also used the NN-CME to investigate how the nascent RNA distributions change with variance in the delay time when the mean is kept constant: as shown in Fig. 4c we find that a large increase in the variability of the delay time tends to destroy the peak at zero. Note that for systems with stochastic delay, the training of the ANN-aided approximation remains the same

as for those with fixed delay; the advantage of our method is that it can just as easily solve non-Markov models with stochastic delay as those with deterministic delay whereas analytically only the latter are amenable to exact solutions (see SI Note 1).

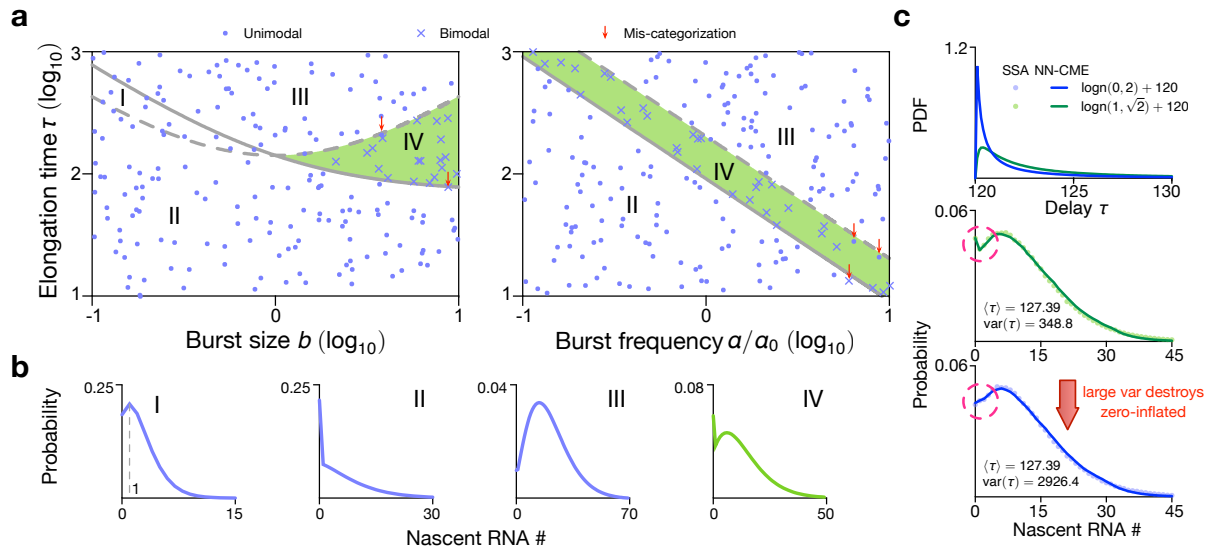


Figure 4: Stochastic bifurcation diagram for Model III in the bursty regime ($\sigma_{\text{off}} \gg \sigma_{\text{on}}$) using the NN-CME and comparison with theory. (a) From an analytical approximation of Model III in the bursty regime, the space is divided into four regions according to the type of distributions (shown in b): type I, a unimodal distribution with mode = 1; type II, a unimodal distribution with mode = 0; type III, a unimodal distribution with mode > 1; type IV, a bimodal distribution with two modes at zero and a non-zero value. Region IV is highlighted in green since it is a phase that does not exist in the bursty regime of the standard model of gene expression (Model III with delayed degradation replaced by first-order degradation) – this is hence delay-induced bimodality. The lines defining the division of space are: solid line is $(2 + \frac{2}{b})/\alpha$ and the dashed line is $(b + \frac{1}{b} + 2)/\alpha$ which respectively are the lower and upper bounds on τ given by Eq. (9). To check the accuracy of the ANN-aided model approximation for Model III, we used it to compute the NN-CME and then solved using FSP to obtain nascent number distributions for 200 points in parameter space. These are randomly sampled from the space $\{\rho = 2.11, \sigma_{\text{off}} \in [2.11 \times [10^{-1}, 10], \sigma_{\text{on}} = 0.0282, \tau \in [10, 10^3]\}$ (left) and $\{\rho = 2.11, \sigma_{\text{off}} = 0.609, \sigma_{\text{on}} = 0.0282 \times [10^{-1}, 10], \tau \in [10, 10^3]\}$ (right). Dots denote parameter sets for which the NN-CME distributions are unimodal and crosses show those for which the distributions are bimodal. The fact that the vast majority of crosses fall in region IV and the dots outside of it shows that the NN-CME agrees with the analytical approximation of Model III (parameter sets which mismatch between the NN-CME and theory are highlighted with red arrows and are very few in number). Note in the left figure of (a), the burst frequency is fixed to $\alpha = 0.0282$ (left) while in the right figure, we use $\alpha_0 = 0.0282$ and the burst size is fixed to $b = 3.46$. (c) The NN-CME is learnt from stochastic simulations of the delay model of Model III with the added feature that the elongation time τ is a random variable sampled from two different lognormal distributions (see top figure). In the middle and bottom figures we show that the delay-induced bimodality (phase IV) disappears as the variance on the elongation time τ increases at constant mean. The rate constants and other parameters related to the ANN’s training are specified in SI Table 1.

In summary, we find that delayed degradation induces an extra mode peaked at $n = 0$, a type of zero-inflation phenomenon. This phenomenon is commonly seen in single-cell RNA-seq data, and it is usually attributed to the expression drop-off caused by technical noises or sequencing sensitivity [28]. It has also been shown [29, 30] that it may arise from an extra number of gene states. However our results suggest that delay due to elongation (when the variability in elongation times is small) is another important source contributing to the zero-inflated distributions evident in RNA-seq data.

Learning the effective master equation of genetic feedback loops from partial abundance data

Feedback inhibition of gene expression is a common phenomenon in which gene expression is down-regulated by its protein product. Given there is sufficient time delay in the regulation process as well as sufficient nonlinearity in the mass-action law describing the kinetics of certain reaction steps [31], feedback inhibition can lead to oscillatory gene expression such as that observed in circadian clocks [32].

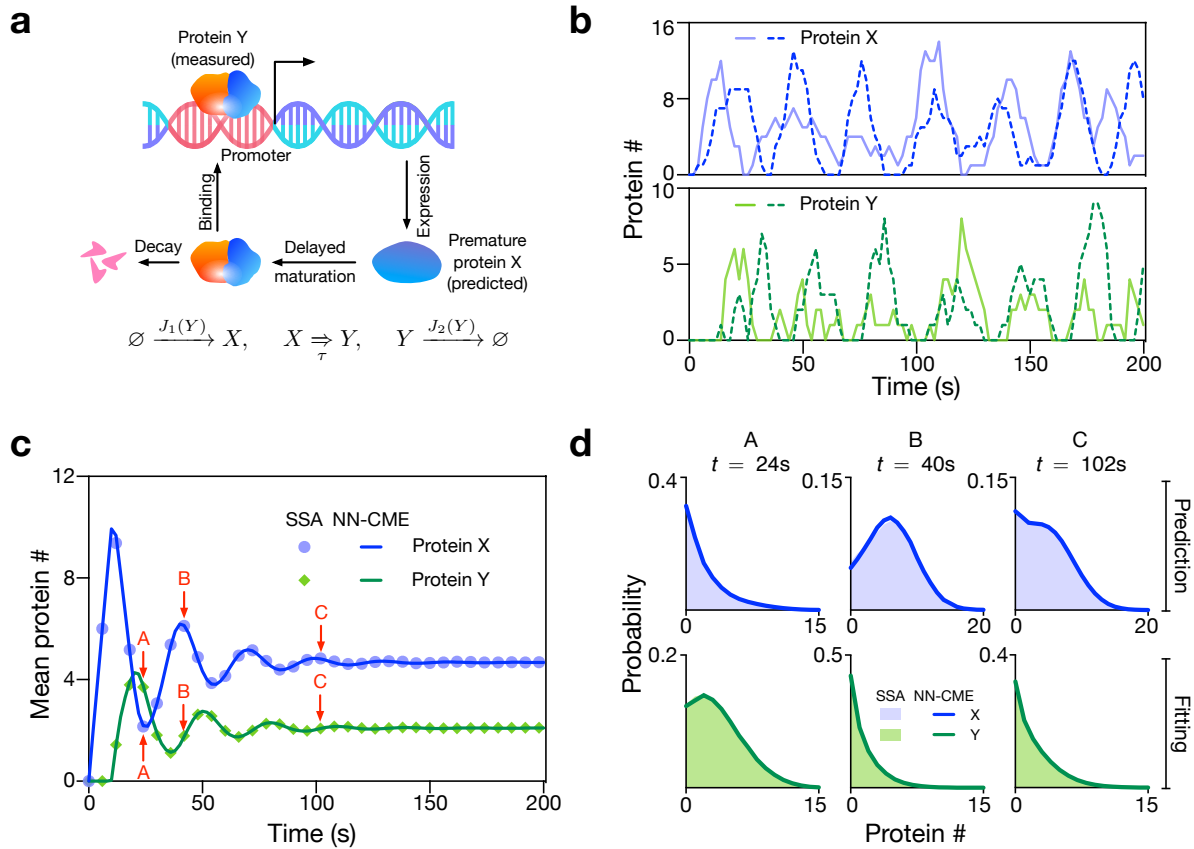


Figure 5: NN-CME accurately predicts the properties of a stochastic auto-regulatory model of oscillatory gene expression when only partial data is used for ANN training. (a) Illustration of a model of auto-regulation whereby a protein X is transcribed by a gene, then it is transformed after a delay time τ into a mature protein Y, which binds the promoter and represses transcription of X. The functions $J_1(Y)$ and $J_2(Y)$ can be found in SI Note 5. (b) Two typical SSA simulations of proteins X and Y, clearly showing that single-cell oscillations while noisy, they are sustained. (c,d) The NN-CME is obtained from training the ANN using only protein Y data from SSA simulations of the delay model of the auto-regulatory model. Surprisingly the NN-CME's solution for the temporal variation of the mean number of *both* proteins X and Y, and for their distributions is in excellent agreement with that of the SSA. Note the distributions in (d) are for the 3 time points labelled A, B and C in (c). The rate constants and other parameters related to the ANN's training are specified in SI Table 1.

Here we consider a simple genetic negative feedback loop (see Fig. 5a) whereby (i) a protein X is transcribed by a promoter, (ii) subsequently after a fixed time delay τ , X turns (via some set of unspecified biochemical processes) into a protein Y and (iii) finally Y binds the promoter and reduces the rate of transcription of X. Simulation trajectories verify oscillatory behaviour of this circuit; see Fig. 5b. We use the simulated trajectories of mature protein Y to train the ANN (the objective function only measures the distance between the ANN predicted distribution of Y and the distribution from stochastic simulations),

in a similar way as previous examples (see SI Note 5). In Fig. 5c and d we show that the time-dependent distributions of both proteins and their means output using the NN-CME are in excellent agreement with the SSA, even clearly capturing the damped oscillatory behaviour; while for Y this is maybe not so surprising, for X it is remarkable because simulated trajectories of X were not used in the training of the ANN. Hence this shows that the ANN-aided model approximation can learn the effective form of master equations from partial trajectory information, a very useful property if the training data is sparse and available only for some molecular species as commonly the case with experimental data.

ANN-aided inference of the parameters of bursty transcription

With a small modification, the ANN-aided model approximation technique besides constructing an approximate NN-CME model, it can also infer the values of kinetic parameters of the data used for training. This is brought about by optimizing not only the weights and biases of the ANN but also simultaneously for the kinetic parameter values. In Fig. 6 we show the results of this method using training data generated by the SSA of Model II with parameters (burst frequency α and size b) that have been measured for 5 mammalian genes [33]. Comparing the latter true parameter values with those obtained from the ANN-aided inference, we conclude that the inference procedure leads to accurate results. Note that the 95% confidence intervals of the estimates are obtained using the profile likelihood method (see SI Note 6 and Fig. 6a,b for an illustration).

We also show that the distribution solution of the NN-CME (which is obtained at one go, together with the inferred parameters) is practically indistinguishable from distributions constructed using the SSA of Model II (the quantile-quantile plots in Fig. 6c are linear with unit slope and zero intercept). In SI Fig. 3 we show the application of the ANN-aided inference to Model III.

Discussion

In this paper, we have shown how the training of a three-layer perceptron with a single hidden layer is enough to approximate the delay CME of a non-Markov model by the NN-CME which is a master equation with time-dependent propensities (time-inhomogeneous Markov process). Notably this mapping has been achieved without increasing the effective number of species in the model. Since the NN-CME has no delay terms, it simplifies analysis and simulation; for e.g. the NN-CME can be accurately approximated by a wide range of standard methods [3] and its solution is straightforward using finite state projection [18]. We note that while neural networks have been recently used to approximate partial differential equations in physics, chemistry and biology, to our knowledge our work represents their first use in approximating equations describing the time-evolution of stochastic processes in continuous time and with a discrete state space e.g., systems describing cellular biochemistry where the discreteness is an important feature of the system due to the low copy number of DNA and mRNA molecules involved [34].

We find that to obtain an accurate NN-CME, training only needs a small sample size (of the order of a thousand SSA trajectories which is computationally cheap), it can be done with partial data (some species data can be missing) and simultaneously one can obtain estimates of the kinetic parameters. The latter is particularly relevant if the training data is collected experimentally, e.g. by measuring nascent RNA numbers using live-cell imaging techniques (such as the MS2 system) at several time points for many cells [35, 36]. Our ANN based inference method rests upon the matching of distributions and hence similar to non-ANN based methods developed in Refs [37, 38], it avoids the pitfalls of moment-based inference [39, 40]. We note that the ability to approximate solutions of delay master equations from simulated data while simultaneously optimizing for the parameters has not been demonstrated before; deep learning frameworks have previously achieved similar feats for deterministic models [13, 16] and more recently for stochastic models described by multi-dimensional Fokker-Planck equations [41, 42].

Recent work has shown that differential equation models describing the time-evolution of average agent density can be learnt (using sparse regression) from agent-based model simulations of spatial

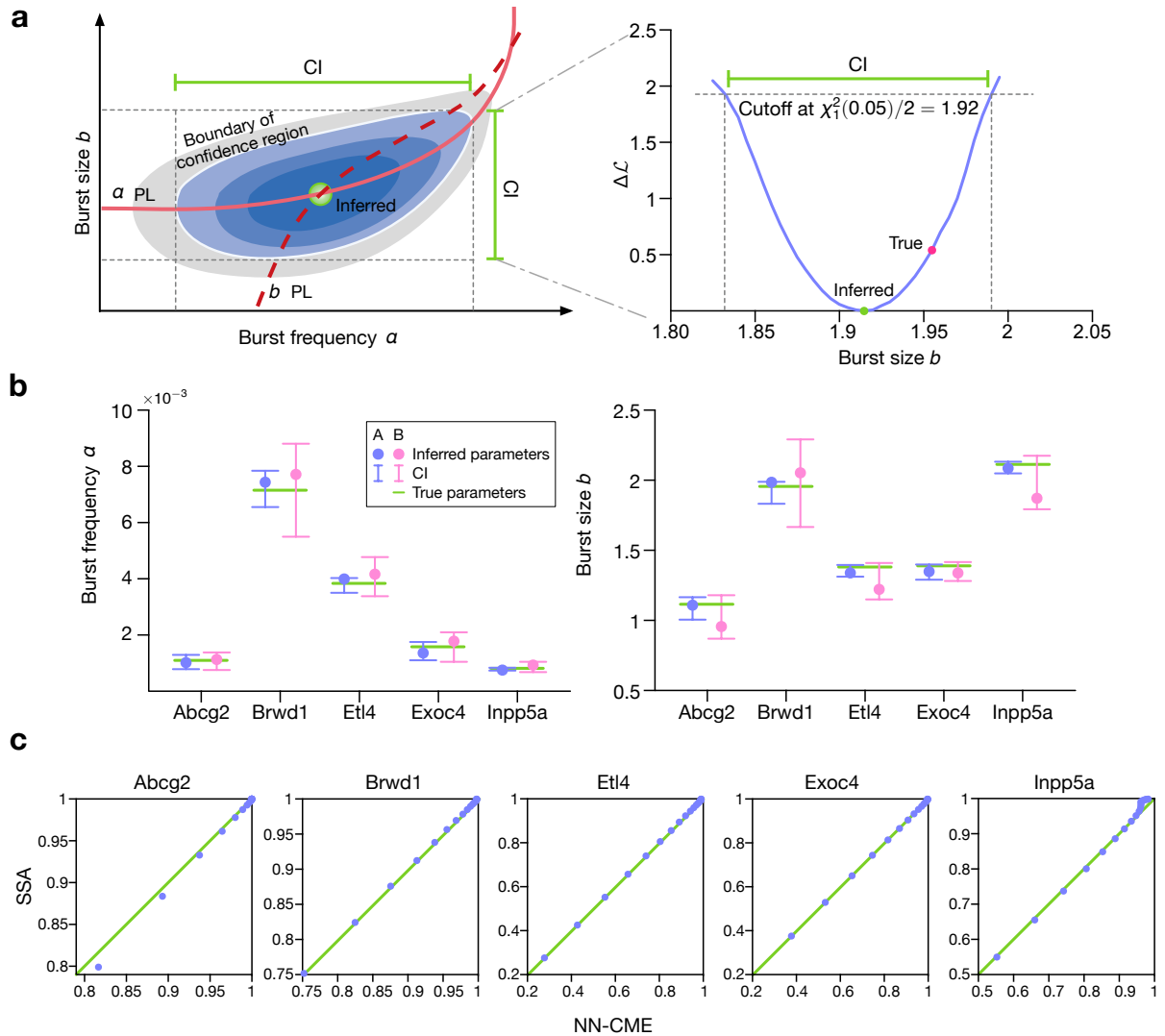


Figure 6: ANN-aided model approximation seamlessly integrates the inference of kinetic parameters and approximation of the delay CME by a NN-CME. The unknown kinetic parameters can be treated in the same way as neural network coefficients (weight and biases) and optimized to minimize the objective function. Application to Model II. (a) Sketch of the computation of the 95% confidence interval (CI) of the inferred kinetic parameters. Blue areas indicate the 95% confidence region, while the gray area shows the non-confidence region. Both solid and dashed red lines show the profile likelihoods of burst frequency α and burst size b , respectively (See SI Note 6 for details). (b) Inferred values of α and burst size b (blue dots), their 95% CIs (blue lines) and the true values (green lines) for five mammalian genes. Inference by using ANN-aided model approximation is robust against size of dataset: Dataset A (100 snapshots and 10^4 cells) and Dataset B (50 snapshots and 10^3 cells) produce similar results. (c) quantile-quantile plots for the steady-state distributions of the NN-CME and those obtained from the SSA; the linearity confirms that the ANN-aided model approximation can accurately approximate the distribution using the NN-CME even when the optimization is over both the kinetic parameters and the neural network coefficients. The rate constants and other parameters related to the ANN's training are specified in SI Table I.

reaction-diffusion processes [43]. Such models can describe for e.g. intracellular biochemical processes in crowded conditions [44] or multi-scale tissue dynamics including cell movement, growth, adhesion, death, mitosis and chemotaxis [45–47]. Some of these models have been shown to display non-Markovian behaviour [48]. While here we showcased the ANN based method using non-Markovian delay CMEs, one could also use for training, data generated by spatially resolved particle-based simulations, as the examples above. The application of our method would provide a master equation that effectively captures stochastic dynamics at the population level of description and avoids the pitfalls of commonly used analytical approximation methods e.g., mean-field approximations, to obtain reduced stochastic descriptions.

Data availability

All relevant data are available from the authors.

Acknowledgement

Z.C., W.D. and F.Q. acknowledge the support from Natural Science Foundation of China (NSFC No. 61988101), Z.C. acknowledges the support from NSFC No. 62073137, Q.J. and S.Y. acknowledge the support from National Key Research and Development Program of China (2020YFA0908303), and R.G. thanks the support from the Leverhulme Trust grant (RPG-2018-423). We thank James Holehouse, Kaan Öcal and Guido Sanguinetti for useful discussions and insightful feedback.

Author contributions

Z.C. and R.G. designed research, supervised research, acquired funding and wrote the manuscript with input from the others. Q.J., X.F. and S.Y. performed research, analyzed the data and wrote the manuscript. W.D. and F.Q. analyzed the data and acquired funding. R.L. was involved in data analysis and graphical illustration.

Competing interests

The authors declare no competing interests.

References

- [1] Shahrezaei, V. & Swain, P. S. Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci USA* **105**, 17256–17261 (2008).
- [2] Cao, Z. & Grima, R. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proc Natl Acad Sci USA* **117**, 4682–4692 (2020).
- [3] Schnoerr, D., Sanguinetti, G. & Grima, R. Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. *J Phys A* **50**, 093001 (2017).
- [4] Xu, H., Skinner, S. O., Sokac, A. M. & Golding, I. Stochastic kinetics of nascent RNA. *Phys Rev Lett* **117**, 128101 (2016).
- [5] Filatova, T., Popovic, N. & Grima, R. Statistics of nascent and mature rna fluctuations in a stochastic model of transcriptional initiation, elongation, pausing, and termination. *bioRxiv* (2020).
- [6] Barrio, M., Burrage, K., Leier, A. & Tian, T. Oscillatory regulation of hes1: discrete stochastic delay modelling and simulation. *PLoS Comput Biol* **2**, e117 (2006).
- [7] Cai, X. Exact stochastic simulation of coupled chemical reactions with delays. *J Chem Phys* **126**, 124108 (2007).
- [8] Lafuerza, L. F. & Toral, R. Exact solution of a stochastic protein dynamics model with delayed degradation. *Phys Rev E* **84**, 051121 (2011).
- [9] Park, S. J. *et al.* The chemical fluctuation theorem governing gene expression. *Nat Commun* **9**, 1–12 (2018).
- [10] Zhang, J. & Zhou, T. Markovian approaches to modeling intracellular reaction processes with molecular memory. *Proc Natl Acad Sci USA* **116**, 23542–23550 (2019).
- [11] Leier, A. & Marquez-Lago, T. T. Delay chemical master equation: direct and closed-form solutions. *Proc R Soc A* **471**, 20150049 (2015).
- [12] Sirignano, J. & Spiliopoulos, K. Dgm: A deep learning algorithm for solving partial differential equations. *J Comput Phys* **375**, 1339–1364 (2018).
- [13] Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J Comput Phys* **378**, 686–707 (2019).
- [14] Bar-Sinai, Y., Hoyer, S., Hickey, J. & Brenner, M. P. Learning data-driven discretizations for partial differential equations. *Proc Natl Acad Sci USA* **116**, 15344–15349 (2019).
- [15] Hermann, J., Schätzle, Z. & Noé, F. Deep-neural-network solution of the electronic schrödinger equation. *Nat Chem* **12**, 891–897 (2020).
- [16] Lagergren, J. H., Nardini, J. T., Baker, R. E., Simpson, M. J. & Flores, K. B. Biologically-informed neural networks guide mechanistic modeling from sparse experimental data. *arXiv preprint arXiv:2005.13073* (2020).
- [17] Rackauckas, C. *et al.* Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385* (2020).
- [18] Munsky, B. & Khammash, M. The finite state projection algorithm for the solution of the chemical master equation. *J Chem Phys* **124**, 044104 (2006).

- [19] Hornik, K., Stinchcombe, M., White, H. *et al.* Multilayer feedforward networks are universal approximators. *Neural Netw* **2**, 359–366 (1989).
- [20] Pinkus, A. Approximation theory of the mlp model in neural networks. *Acta Numer* **8**, 143–195 (1999).
- [21] Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning* (MIT Press, 2016).
- [22] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc IEEE Conf on Computer Vision and Pattern Recognition*, 770–778 (2016).
- [23] Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472–474 (2011).
- [24] Paulsson, J. & Ehrenberg, M. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Phys Rev Lett* **84**, 5447 (2000).
- [25] Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* **4**, e309 (2006).
- [26] Hocine, S., Raymond, P., Zenklusen, D., Chao, J. A. & Singer, R. H. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. *Nat Methods* **10**, 119–121 (2013).
- [27] Liu, J. *et al.* Quantitative characterization of the eukaryotic transcription cycle using live imaging and statistical inference. *bioRxiv* (2020).
- [28] Vu, T. N. *et al.* Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32**, 2128–2135 (2016).
- [29] Engl, C., Jovanovic, G., Brackston, R. D., Kotta-Loizou, I. & Buck, M. The route to transcription initiation determines the mode of transcriptional bursting in *E. Coli*. *Nat Commun* **11**, 1–11 (2020).
- [30] Jia, C. Kinetic foundation of the zero-inflated negative binomial model for single-cell RNA sequencing data. *SIAM J Appl Math* **80**, 1336–1355 (2020).
- [31] Novák, B. & Tyson, J. J. Design principles of biochemical oscillators. *Nat Rev Mol Cell Biol* **9**, 981–991 (2008).
- [32] Wenden, B., Toner, D. L., Hodge, S. K., Grima, R. & Millar, A. J. Spontaneous spatiotemporal waves of gene expression from biological clocks in the leaf. *Proc Natl Acad Sci USA* **109**, 6757–6762 (2012).
- [33] Larsson, A. J. *et al.* Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).
- [34] Cai, L., Friedman, N. & Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* **440**, 358–362 (2006).
- [35] Larson, D. R., Zenklusen, D., Wu, B., Chao, J. A. & Singer, R. H. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science* **332**, 475–478 (2011).
- [36] Lenstra, T. L., Coulon, A., Chow, C. C. & Larson, D. R. Single-molecule imaging reveals a switch between spurious and functional ncRNA transcription. *Mol Cell* **60**, 597–610 (2015).
- [37] Öcal, K., Grima, R. & Sanguinetti, G. Parameter estimation for biochemical reaction networks using wasserstein distances. *J Phys A* **53**, 034002 (2019).
- [38] Munskey, B., Li, G., Fox, Z. R., Shepherd, D. P. & Neuert, G. Distribution shapes govern the discovery of predictive models for gene regulation. *Proc Natl Acad Sci USA* **115**, 7533–7538 (2018).

- [39] Cao, Z. & Grima, R. Accuracy of parameter estimation for auto-regulatory transcriptional feedback loops from noisy data. *J R Soc Interface* **16**, 20180967 (2019).
- [40] Zechner, C. *et al.* Moment-based inference predicts bimodality in transient gene expression. *Proc Natl Acad Sci USA* **109**, 8340–8345 (2012).
- [41] Chen, X., Yang, L., Duan, J. & Karniadakis, G. E. Solving inverse stochastic problems from discrete particle observations using the fokker-planck equation and physics-informed neural networks. *arXiv preprint arXiv:2008.10653* (2020).
- [42] Yang, L., Daskalakis, C. & Karniadakis, G. E. Generative ensemble-regression: Learning stochastic dynamics from discrete particle ensemble observations. *arXiv preprint arXiv:2008.01915* (2020).
- [43] Nardini, J. T., Baker, R. E., Simpson, M. J. & Flores, K. B. Learning differential equation models from stochastic agent-based model simulations. *arXiv preprint arXiv:2011.08255* (2020).
- [44] Schöneberg, J. & Noé, F. Readdy-a software for particle-based reaction-diffusion dynamics in crowded cellular environments. *PLoS One* **8**, e74261 (2013).
- [45] Swat, M. H. *et al.* Multi-scale modeling of tissues using compucell3d. In *Methods in cell biology*, vol. 110, 325–366 (Elsevier, 2012).
- [46] Matsiaka, O. M., Penington, C. J., Baker, R. E. & Simpson, M. J. Continuum approximations for lattice-free multi-species models of collective cell migration. *J Theor Biol* **422**, 1–11 (2017).
- [47] Middleton, A. M., Fleck, C. & Grima, R. A continuum approximation to an off-lattice individual-cell based model of cell migration and adhesion. *J Theor Biol* **359**, 220–232 (2014).
- [48] Newman, T. & Grima, R. Many-body theory of chemotactic cell-cell interactions. *Phys Rev E* **70**, 051916 (2004).