

RESEARCH ARTICLE

Simulation and inference methods for non-Markovian stochastic reaction networks

Thomas P. Steele¹, David J. Warne^{1,2,3*}

1 School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia, **2** Centre for Data Science, Queensland University of Technology, Brisbane, Australia, **3** ARC Centre of Excellence for Mathematical Analysis of Cellular Systems (MACSYS), Queensland University of Technology, Brisbane, Australia

* david.warne@qut.edu.au



Abstract

Stochastic models of reaction networks are widely used to capture intrinsic noise in complex systems in the life sciences. Typical formulations of these models are based on Markov processes for which there is extensive research on efficient simulation and inference. However, there are complex processes in biology, such as gene transcription and translation, that introduce history dependent dynamics requiring non-Markovian processes to accurately capture the stochastic dynamics of the system. This greater realism comes with additional computational challenges for simulation and parameter inference. We develop efficient stochastic simulation algorithms for well-mixed non-Markovian stochastic reaction networks with stochastic delays that depend on system state and time. Our methods generalize the next reaction method and τ -leaping method to support arbitrary inter-event time distributions while preserving computational scalability. We also introduce a coupling scheme to generate exact non-Markovian sample paths that are positively correlated to an approximate non-Markovian τ -leaping sample path. This enables substantial computational gains for simulation and Bayesian inference through multilevel Monte Carlo and multifidelity schemes. We demonstrate the effectiveness of our approach using several non-Markovian examples, showing substantial gains in both simulation accuracy and inference efficiency. These results extend the practical applicability of non-Markovian models in systems biology and beyond.

OPEN ACCESS

Citation: Steele TP, Warne DJ (2026) Simulation and inference methods for non-Markovian stochastic reaction networks. PLoS Comput Biol 22(8): e1014687. <https://doi.org/10.1371/journal.pcbi.1014687>

Editor: Amar Singh, KU: The University of Kansas, UNITED STATES OF AMERICA

Received: January 21, 2026

Accepted: August 6, 2026

Published: August 26, 2026

Copyright: © 2026 Steele, Warne. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: Julia code and simulation data related this project is available on GitHub (<https://github.com/davidwarne/NonMarkovMFABC>).

Funding: The project is supported by the Australian Research Council (ARC). DJW is supported by an ARC Discovery Early Career Researcher Award (DE250100396). DJW

Author summary

To understand and predict how cells within living organisms behave we need mathematical models of biochemical processes that govern cell function. Accurate models of these biochemical processes often require inherent internal randomness to be explicitly described. Unfortunately this leads to substantial computational challenges when simulating these models and analysing them

acknowledges support from the ARC Centre of Excellence for the Mathematical Analysis of Cellular Systems (MACSYS; CE230100001). DJW 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.

statistically. These challenges are especially true when the process is non-Markovian: that is, the history of the process is important to determine the future. We develop new and efficient methods for simulation and statistical inference of such non-Markovian models and demonstrate that our approach achieves a substantial reduction in computation cost without sacrificing accuracy.

Introduction

Stochastic modelling of system dynamics is an essential technique in many fields of science to describe processes driven by intrinsic noise [1–3]. Many real systems are well described by Markovian stochastic processes, where future evolution depends only on the current state and is conditionally independent of the past [4–6]. However, in many systems, history plays an essential role in shaping future dynamics. In such cases, non-Markovian stochastic processes are necessary for accurate simulation, prediction, and parameter inference [7–12].

A particularly important application domain for stochastic models is the study of intracellular processes. Processes such as gene regulation are inherently stochastic, often due to the low copy numbers of key bio-molecules involved [5,13,14]. The dynamics of such systems are governed by the chemical master equation (CME), which is analytically intractable in all but the simplest cases, necessitating computational approaches [15]. Two primary problems for these systems are: the forward problem, that is, the stochastic simulation of a given system; and the inverse problem, that is, inferring model parameters given experimental data [16–18].

For the forward problem, a variety of stochastic simulation methods are available. These include exact methods, such as the Gillespie direct method [4] and the next reaction method (NRM) [19,20], and approximate methods, such as the τ -leaping method [6]. All of these methods are based on models of chemically reacting systems that are continuous-time Markov jump processes where state transitions occur instantaneously at reaction events times [5,15]. However, the assumption of instantaneous reaction events is not always appropriate, particularly for complex reactions that occur in gene transcription and translation. In these cases, it may be more appropriate to model complex reactions using delayed reactions, due to the non-negligible duration between initiation and completion of a reaction [21–24]. The presence of delay reactions in a system presents a computational challenge as it becomes a non-Markovian stochastic process [7,9,10,25].

A popular approach to the inverse problem is Bayesian inference that updates knowledge of model parameters, $\theta \in \Theta$, through combining prior knowledge of these parameters with observed data, y_{obs} . This proceeds through the application of Bayes' theorem,

$$\pi(\theta | y_{\text{obs}}) = \frac{\pi(y_{\text{obs}} | \theta) \pi(\theta)}{\pi(y_{\text{obs}})}, \quad (1)$$

where $\pi(\theta)$ is the prior probability density that represents current knowledge, $\pi(y_{\text{obs}} | \theta)$ is the likelihood of an observation given model parameters, and $\pi(y_{\text{obs}})$ is the evidence or marginal likelihood that acts as a normalisation constant. Through Eq (1) we obtain the posterior probability density, $\pi(\theta | y_{\text{obs}})$, that represents updated knowledge informed by the prior and the data. Since the evidence term is usually intractable, sampling schemes such as Markov chain Monte Carlo (MCMC) [26,27] and sequential Monte Carlo (SMC) [28] are used to sample to posterior distribution. However, these standard tools are not directly applicable in the context of partially observed stochastic processes, since the likelihood will almost always be intractable due to dependence on the CME solution [18,29–31]. This is only rendered more challenging in the non-Markovian setting as the equivalent CME is non-trivial [32].

Due to the likelihood intractability, we rely on so-called likelihood-free inference methods, also known as simulation-based inference. These methods rely on large numbers of model simulations as a substitute for direct likelihood evaluation [33–35]. These include traditional approaches, such as approximate Bayesian computation (ABC) [36–40] and Bayesian synthetic likelihood (BSL) [41,42], and modern machine learning approaches, such as neural likelihood estimation [43] and neural posterior estimation [44,45]. While machine learning approaches have been shown to require fewer stochastic simulations [46], challenges still remain in terms of accuracy guarantees [47,48]. Regardless of the specific approach, all likelihood-free and simulation-based inference approaches are computationally intensive, especially for complex or high-dimensional models, as they require large numbers of model simulations to obtain accurate posterior estimates.

Recent work has focused on acceleration strategies that leverage model approximations. Two notable approaches are multilevel Monte Carlo (MLMC) [49–52] and multifidelity schemes [53, 54, 55]. These techniques reduce the computational cost by combining large numbers of computationally inexpensive low-fidelity approximate simulations with relatively fewer computationally expensive high-fidelity simulations. Accuracy is maintained through low variance bias corrections that rely on variance reduction due to coupling simulation pairs [49,54,56]. As a result, MLMC and multifidelity approaches can achieve substantial improvements in computational efficiency, often by orders of magnitude, without sacrificing inference accuracy [50–54,57,58].

Despite these advances, such multifidelity methods have been almost exclusively applied to Markovian systems [54,57,58]. Utilising multifidelity methods for non-Markovian stochastic processes introduces major challenges. In particular, many simulation schemes in the literature cannot simulate non-Markovian systems accurately [59], and those that do are often restricted to delay time distributions that depend on time only [7, 15]. Moreover, the generation of coupled sample paths in a non-Markovian setting, as required for MLMC and multifidelity approaches, requires careful handling of delay structures to maintain consistency and avoid statistical inaccuracies [7,32]. In addition, the focus of the literature has been on exact simulation schemes [7, 15, 59]. As a result, the necessary approximate simulation and coupling schemes required for MLMC and multifidelity schemes have not been explored in the non-Markovian setting.

In this work, we address this challenge and enable MLMC and multifidelity simulation and inference approaches for non-Markovian reaction networks. We present a novel exact stochastic simulation that leads to a natural approximate stochastic simulation and coupling scheme. We begin by presenting the general framework for simulating jump processes with non-exponential inter-event times, then introduce non-Markovian extensions to both the NRM and the τ -leaping method. Finally, we develop a novel coupling scheme for generating positively correlated approximate and exact non-Markovian sample paths. This enables the application of MLMC and multifidelity methods for the forwards and inverse problems for non-Markovian systems. We demonstrate the efficacy of our approach through the implementation of multifidelity ABC [54], however, our simulation algorithms are applicable to any of the current MLMC or multifidelity methods that rely on coupling for accelerating simulation [49,56] and inference [50,57,58,60]. Together, these contributions provide a foundation for efficient simulation and inference for non-Markovian systems, addressing a critical barrier to the practical implementation of these models.

Methods

In this section, we establish some fundamental background and mathematical foundations that form a basis of our methods and algorithms. We begin by reviewing the standard formulation of a reaction network [17,18,61] and then examine how delay reactions alter the structure and dynamics of these systems [7,25]. From this, we derive the inter-event time distribution in non-Markovian systems and present a generalised representation of a reaction network that accommodates non-exponential waiting times. We then review the ABC approach to likelihood-free inference [34,36,37] and acceleration using multifidelity schemes [54, 57]. Finally, we introduce our algorithmic contributions to enable the application of multifidelity approaches in the non-Markovian setting. Specifically, we develop a non-Markovian extension to the NRM, a non-Markovian τ -leaping scheme, and a novel coupling scheme for generating correlated approximate and exact non-Markovian simulation pairs.

Modelling stochastic reaction networks

A reaction network consists of $\mathcal{N}$ chemical species, $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_{\mathcal{N}}$, with system state, $\mathbf{X}(t) = [X_1(t), \dots, X_{\mathcal{N}}(t)]^T$, where $X_i(t) \in \mathbb{N}$ is the copy number of the i th chemical species. Species interact via a network of $\mathcal{M}$ reaction channels,



where $k_j > 0$ is the kinetic rate parameter for reaction j , and $\nu^- \in \mathbb{N}^{\mathcal{N} \times \mathcal{M}}$ and $\nu^+ \in \mathbb{N}^{\mathcal{N} \times \mathcal{M}}$ denote the reactant and product stoichiometric matrices, respectively. In this paper, for a matrix A , we adopt the notation a_{ij} for element in row i and column j , and a_{*j} as the j th column of A . Given a system state $\mathbf{X}(t)$ the probability that reaction j occurs in the interval $[t, t + \Delta t)$ is $\lambda_j(\mathbf{X}(t))\Delta t + \mathcal{O}(\Delta t^2)$ where $\lambda_j(\mathbf{X}(t))$ is the propensity function for the j th reaction channel. These propensity functions typically have the standard combinatorial form based on the law of mass action,

$$\lambda_j(\mathbf{X}(t)) = k_j \prod_{i=1}^{\mathcal{N}} \frac{X_i(t)!}{(X_i(t) - \nu_{ij}^-)!}. \quad (3)$$

However, nonlinear propensities, such as those based on Hill functions are also routinely used. In both cases, the propensity is a function of the system state only, this means that propensities are constant between reaction events. When a reaction event from channel j occurs, the system state is instantaneously updated $\mathbf{X}(t + \tau) = \mathbf{X}(t) + \nu_{*j}$ where $\nu_{*j} = \nu_{*j}^+ - \nu_{*j}^-$, and τ is the inter-event time. In this setting, the number of reaction events over time for channel j can be described by an inhomogeneous Poisson process, this leads to the so-called Kurtz random time-change representation of a reaction network [62],

$$\mathbf{X}(t + \tau) = \mathbf{X}(t) + \sum_{j=1}^{\mathcal{M}} Y_j \left(\int_0^\tau \lambda_j(\mathbf{X}(t + s)) ds \right) \nu_{*j}, \quad (4)$$

where $Y_j(\cdot)$ is a unit-time homogeneous Poisson process. The state transition probability function related to this process, $p(\mathbf{x}_t, \mathbf{x}_s) = \Pr(\mathbf{X}(t) = \mathbf{x}_t \mid \mathbf{X}(s) = \mathbf{x}_s)$ for $t > s$, can be obtained through the solution to the forwards Kolmogorov equation, which is known as the CME in the biochemistry literature [17,29,63],

$$\frac{dp(\mathbf{x}_t, \mathbf{x}_s)}{dt} = \sum_{j=1}^{\mathcal{M}} \lambda_j(\mathbf{x}_t - \nu_j) p(\mathbf{x}_t - \nu_j, \mathbf{x}_s) - p(\mathbf{x}_t, \mathbf{x}_s) \sum_{j=1}^{\mathcal{M}} \lambda_j(\mathbf{x}_t) \quad (5)$$

Unfortunately, the CME (Eq (5)) is intractable for most realistic networks [18,61,64].

Inclusion of delay reactions

In complex biochemical processes, it may not always be appropriate to treat all reaction events as instantaneous. For example, the transcription process from DNA to mRNA and subsequent translation into proteins are complex biochemical processes that occur over a non-negligible period of time [7,21,23,25,65]. This type of reaction cannot be adequately described by Equations (2)–(5). Instead these reactions must be described as *delay reactions* where the interval between reactant consumption and product formation is explicitly modeled.

Suppose that of our $\mathcal{M}$ reactions $\mathcal{R} = \{1, 2, \dots, \mathcal{M}\}$ there is a subset $\mathcal{R}_D \subseteq \mathcal{R}$ that are delayed reactions and the remaining set $\mathcal{R}_I = \mathcal{R} \setminus \mathcal{R}_D$ are instantaneous reaction set so that $\mathcal{R} = \mathcal{R}_D \cup \mathcal{R}_I$ and $\mathcal{R}_D \cap \mathcal{R}_I = \emptyset$. This leads to the system

$$\begin{aligned} \sum_{i=1}^{\mathcal{N}} \nu_{ij}^- \mathcal{X}_i &\xrightarrow{k_j} \sum_{i=1}^{\mathcal{N}} \nu_{ij}^+ \mathcal{X}_i, \quad j \in \mathcal{R}_I \\ \sum_{i=1}^{\mathcal{N}} \nu_{ij}^- \mathcal{X}_i &\xrightarrow[k_j^*]{k_j} \sum_{i=1}^{\mathcal{N}} \nu_{ij}^+ \mathcal{X}_i, \quad j \in \mathcal{R}_D. \end{aligned} \quad (6)$$

For instantaneous reactions $j \in \mathcal{R}_I$ that are denoted by $\rightarrow$, the standard formulation for the propensities $\lambda_j(\mathbf{X}(t))$ holds with $k_j > 0$ as the kinetic rate parameter (Eq (3)). If the j th reaction is in $\mathcal{R}_D$ then it is a delayed reaction, which is denoted by $\Rightarrow$. This delay reaction has an initiation and a completion event, the initiation event occurs according to a Poisson process with propensity $\lambda_j(\mathbf{X}(t))$ with kinetic rate parameter $k_j > 0$ just as in the instantaneous case. However the completion event, conditional on initiation time, t_0^* , occurs with propensity $\lambda_j^*(\mathbf{X}(t), t^*)$ with rate kinetic parameter $k_j^* > 0$ and internal time $t^* = t - t_0^*$ that leads to a non-exponential waiting-time distribution.

Since each delay reaction has its own internal time, a delay reaction proceeds in two phases. First the reaction is initiated: suppose the k th event initiation event that occurs at time $t + \tau$ where τ is the inter-event time, the state vector updated according to $\mathbf{X}(t + \tau) = \mathbf{X}(t) - \nu_{*j}^-$, and a new *delayed chemical species* is created, $\mathcal{X}_j^{*,k}$, for the k th initiation of the j th reaction (which is a delay reaction since $j \in \mathcal{R}_D$). The k th initiation time is set, $t_0^{*,k} = t + \tau$, is set and a completion reaction is introduced,



The associated reaction network proceeds with this new reaction included in the network. Suppose at time t_1 the completion event for this k th initiation of reaction j occurs at time τ_1 , the k th delayed chemical species $\mathcal{X}_j^{*,k}$ is removed along with the associated reaction, and the state is updated $\mathbf{X}(t_1 + \tau_1) = \mathbf{X}(t_1) + \nu_{*j}^+$. As formulated here, the virtual delayed chemical species are not considered part of the system state, but simply represent the ongoing progression of the complex reaction from initiation to completion. However, this can also be modified by defining a copy number $X_j^{*,k}(t) = 1$, at the k th initiation event, and including this as part of the system state $\mathbf{X}(t)$. This would enable the presence of ongoing reactions to affect any propensity values in the system for any reaction $j \in \mathcal{R}$. We only consider the total number of ongoing reactions as part of the state vector, that is,

$$X_j^*(t) = \sum_{k=1}^{\infty} X_j^{*,k}(t). \quad (7)$$

This leads to the extended state vector including the copy numbers for the $\mathcal{N}$ chemical species and counts of ongoing delayed reaction, that is, $\mathbf{X}(t) = [X_1(t), X_2(t), \dots, X_{\mathcal{N}}(t), X_{j_1}^*(t), X_{j_2}^*(t), \dots, X_{j_D}^*(t)]$, where $\mathcal{D}$ is the number of delayed reaction and $\mathcal{R}_D = \{j_1, j_2, \dots, j_D\}$. Thus, if we consider a system with $\mathcal{N}$ chemical species and $\mathcal{M}$ reactions, of which $\mathcal{D} \leq \mathcal{M}$ are delayed reactions, then we will have an effective state vector of dimension $\mathcal{N} + \mathcal{D}$.

This leads to the following extension to the Kurtz representation,

$$\begin{aligned} \mathbf{X}(t + \tau) = & \mathbf{X}(t) + \sum_{j \in \mathcal{R}_I} Y_j \left(\underbrace{\int_0^\tau \lambda_j(\mathbf{X}(t+s)) ds}_{\text{Instantaneous reactions}} \right) \nu_{*,j} \\ & + \sum_{j \in \mathcal{R}_D} Y_j \left(\underbrace{\int_0^\tau \lambda_j(\mathbf{X}(t+s)) ds}_{\text{Delay initiation reaction}} \right) (\eta_j - \nu_{*,j}^-) \\ & + \sum_{j \in \mathcal{R}_D} \sum_{k=1}^{\infty} \min \left\{ 1, Y_j^k \left(\underbrace{\int_0^\tau \lambda_j^*(\mathbf{X}(t+s), t_j^k + s) ds}_{\text{Completion of } k\text{th ongoing delay reaction}} \right) \right\} (\nu_{*,j}^+ - \eta_j). \end{aligned} \quad (8)$$

Here, $Y_j(\cdot)$ for $j \in \mathcal{R}$, and $Y_j^k(\cdot)$ for $j \in \mathcal{R}_D$ are $k \in \mathbb{Z}^+$ and unit rate homogeneous Poisson processes. We have $t_j^k = t - t_{j,0}^{*,k}$ as the internal time for the k delay in the j th channel, where sequence, $t_{j,0}^{*,1} < t_{j,0}^{*,2} < \dots < t_{j,0}^{*,k} < \dots$, corresponds to the initiation event times for delay as generated by the initiation processes $Y_j(\cdot)$ for $j \in \mathcal{R}_D$. The vector η_j is an elementary vector that is zero everywhere except for the dimension of the state $X_j^*(t)$ (Eq (7)), that is $\eta_j^T \mathbf{X}(t) = X_j^*(t)$. This representation is essential to the construction of our stochastic simulation schemes presented later. We note here that there are connections to the delay reaction setting presented by Anderson [15] and Boguñá et al., [7]. The distinct properties of our approach by comparison to these previous works are discussed along with the derivation of our methods in subsequent sections.

Simulation-based inference

Suppose that we obtain time series data, $y_{\text{obs}} = [\mathbf{Y}(t_1), \mathbf{Y}(t_2), \dots, \mathbf{Y}(t_n)]$, that is assumed to be noisy observations at n discrete times, $t_1 < t_2 < \dots < t_n$, from a realisation of the stochastic process $\{\mathbf{X}(t)\}_{t \geq t_0}$ (Eq (4) or Eq (8)). The aim is to infer the unknown rate parameters of the model $\theta = [\{k_j\}_{j \in \mathcal{R}}, \{k_j^*\}_{j \in \mathcal{R}_D}]$ via Bayes' theorem (Eq (1)). Typically inference will proceed according to the expectation,

$$\mathbb{E}[f(\theta) | y_{\text{obs}}] = \int_{\Theta} f(\theta) \pi(\theta | y_{\text{obs}}) d\theta, \quad (9)$$

for any $f(\theta)$ that is integrable with respect to the posterior measure with density $\pi(\theta | y_{\text{obs}})$. The form of Eq (9) is very general and includes estimation of means, variances, higher-order moments, probabilities, or probability densities of the posterior distribution through the specification of the function $f(\cdot)$. Assuming one can draw independent, identically distributed (i.i.d.) samples from the posterior, then we estimate using Monte Carlo (MC) integration,

$$\mathbb{E}[f(\theta) | y_{\text{obs}}] \approx \hat{f} = \frac{1}{N} \sum_{i=1}^N f(\theta^{(i)}),$$

where $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N)} \stackrel{\text{iid}}{\sim} \pi(\cdot | y_{\text{obs}})$. Unfortunately, standard Bayesian sampling approaches that are based on MCMC or SMC are not viable since the likelihood is intractable due to noisy partial observations and dependence on the master equation. This remains true even in the Markovian case where CME is substantially simpler solve (Eq (5)).

To handle this intractable likelihood setting, various simulation-based techniques can be used [33,34]. Regardless of the specific method applied, the dominating computational cost is associated with stochastic simulations from the model. A clear example of this are ABC-based methods that approximate the posterior according to,

$$\pi(\theta | y_{\text{obs}}) \approx \pi_{\text{ABC}}(\theta | \rho(y_{\text{obs}}, y_{\text{sim}}) < \varepsilon) \propto \Pr(\rho(y_{\text{obs}}, y_{\text{sim}}) < \varepsilon | \theta) \pi(\theta),$$

where

$$\Pr(\rho(y_{\text{obs}}, y_{\text{sim}}) < \varepsilon | \theta) = \int \mathbb{1}_{[0, \varepsilon]}(\rho(y_{\text{obs}}, y_{\text{sim}})) \pi(y_{\text{sim}} | \theta) dy_{\text{sim}}$$

with $y_{\text{sim}} \sim \pi(\cdot | \theta)$ being simulated data through stochastic simulation, $\rho(y_{\text{obs}}, y_{\text{sim}})$ is a discrepancy metric, and $\varepsilon > 0$ is the acceptance threshold [18,34,35].

Inference proceeds via expectations with respect to this approximation,

$$\mathbb{E}[f(\theta_\varepsilon) | y_{\text{obs}}] = \int_{\Theta} f(\theta_\varepsilon) \pi_{\text{ABC}}(\theta_\varepsilon | \rho(y_{\text{obs}}, y_{\text{sim}}) < \varepsilon) d\theta_\varepsilon,$$

where $\theta_\varepsilon \sim \pi_{\text{ABC}}(\theta | \rho(y_{\text{obs}}, y_{\text{sim}}) < \varepsilon)$. In an accept-reject scheme [54] (Alg 1), i.i.d. samples from the joint prior-data distribution, $\pi(y_{\text{sim}}, \theta) = \pi(y_{\text{sim}} | \theta) \pi(\theta)$, are used to enable the MC estimator,

$$\mathbb{E}[f(\theta_\varepsilon) | y_{\text{obs}}] \approx \hat{f}_\varepsilon = \frac{\sum_{i=1}^N w(\theta^{(i)}) f(\theta^{(i)})}{\sum_{i=1}^N w(\theta^{(i)})}, \quad (10)$$

where $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N)} \stackrel{\text{iid}}{\sim} \pi(\cdot)$, and $w(\theta) = \mathbb{1}_{[0, \varepsilon]}(\rho(y_{\text{obs}}, y_{\text{sim}}))$. Here $\mathbb{1}_A(x)$ is the indicator function, $\mathbb{1}_A(x) = 1$ if $x \in A$ and $\mathbb{1}_A(x) = 0$ otherwise. Due to the acceptance threshold $\varepsilon > 0$ and discrepancy based approximation, any expectations taken with respect to the ABC posterior are not exact as $\mathbb{E}[f(\theta_\varepsilon) | y_{\text{obs}}] \neq \mathbb{E}[f(\theta) | y_{\text{obs}}]$ in general. As a result, we ideally take $\varepsilon \rightarrow 0$, however, this leads to extremely low acceptance probabilities of $\mathcal{O}(\varepsilon^d)$ where $d > 1$ is the dimensionality of the data, y_{obs} [66]. This curse of dimensionality renders Alg 1 inefficient when implemented directly, however, extensions based on MCMC [67], SMC [68,69], MLMC [51,52] and multifidelity schemes [54] are widely applied in practice.

Algorithm 1 ABC accept-reject scheme (See Prescott & Baker [54])

Require: A stochastic model $\pi(\cdot | \theta)$; a prior distribution $\pi(\theta)$; a discrepancy metric $\rho(y_{\text{obs}}, \cdot)$; a discrepancy threshold $\varepsilon > 0$; observed data y_{obs} ; and sample size N .

```

for  $i = 1, \dots, N$  do
    Sample prior  $\theta^{(i)} \sim \pi(\cdot)$ ;
    Generate simulated data  $y_{\text{sim}} \sim \pi(\cdot | \theta^{(i)})$ ;
     $w^{(i)} \leftarrow \mathbb{1}_{[0, \varepsilon]}(\rho(y_{\text{obs}}, y_{\text{sim}}))$ ;
end for
return  $(\theta^{(1)}, w^{(1)}), (\theta^{(2)}, w^{(2)}), \dots, (\theta^{(N)}, w^{(N)})$ .

```

Multifidelity approximate Bayesian computation

Prescott and Baker [54] consider an approach to accelerate the computation of $\hat{f}_\varepsilon$ (Eq (10)) that is based upon the probabilistic telescoping summation idea of Rhee and Glynn [55]. Here, we denote $\tilde{\pi}(\cdot | \theta)$ as an approximate simulator that is computationally inexpensive to generate a sample $\tilde{y}_{\text{sim}}$ from $\tilde{\pi}(\cdot | \theta)$. We refer to this approximation as a *low-fidelity* simulator, and the exact simulator, $\pi(\cdot | \theta)$, as a *high-fidelity* simulator that is computationally expensive to generate a sample y_{sim} from $\pi(\cdot | \theta)$. Direct substitution of the low-fidelity simulator into Alg 1 will be unreliable due to propagation

of simulation error into posterior estimation error [57]. This problem motivates the multifidelity approach that exploits the efficiency of the low-fidelity simulator while maintaining the accuracy of the high-fidelity simulator.

In the context of ABC, a so called multifidelity weighting function is introduced,

$$w_{mf}(\theta) = \mathbb{1}_{[0,\varepsilon)}(\rho(y_{obs}, \tilde{y}_{sim})) + \frac{m}{\mu(\tilde{y}_{sim})} [\mathbb{1}_{[0,\varepsilon)}(\rho(y_{obs}, y_{sim})) - \mathbb{1}_{[0,\varepsilon)}(\rho(y_{obs}, \tilde{y}_{sim}))], \quad (11)$$

where m is the outcome of a Bernoulli random variable, M , such that $\Pr(M = 1 | \tilde{y}_{sim}) = \mu(\tilde{y}_{sim})$. The function, $\mu(\tilde{y}_{sim})$, represents the probability of computing a correction of $\pm 1/\mu(y_{sim})$ when $\mathbb{1}_{[0,\varepsilon)}(\rho(y_{obs}, y_{sim})) \neq \mathbb{1}_{[0,\varepsilon)}(\rho(y_{obs}, \tilde{y}_{sim}))$. Note that if $m=0$, then $w_{mf}(\theta) = \mathbb{1}_{[0,\varepsilon)}(\rho(y_{obs}, \tilde{y}_{sim}))$ and the high-fidelity simulator is not needed. Using this weighting scheme, we can construct the multifidelity estimator,

$$\hat{f}_{mf,\varepsilon} = \frac{\sum_{i=1}^N w_{mf}(\theta^{(i)}) f(\theta^{(i)})}{\sum_{i=1}^N w_{mf}(\theta^{(i)})}, \quad (12)$$

where $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N)} \stackrel{iid}{\sim} \pi(\cdot)$. It can be shown that $\hat{f}_{mf,\varepsilon}$ is a consistent estimator of $\mathbb{E}[f(\theta_\varepsilon) | y_{obs}]$ [54,57]. Furthermore, given a fixed computational budget, one can choose a mean function, $\mu(\tilde{y}_{sim})$, such that $\text{Var}[\hat{f}_{mf,\varepsilon}] < \text{Var}[\hat{f}_\varepsilon]$, provided certain conditions hold [57] that relate to the accuracy of the low-fidelity simulator relative to the computational gains. Broadly speaking, Eq (12) utilises the high-fidelity model to correct bias introduced by the low-fidelity model through the weighting scheme (Eq (11)).

For simplicity we will restrict our description to the special case where the decision to simulate $y_{sim} \sim \pi(\cdot | \theta)$ is informed only by the acceptance decision of a given low-fidelity simulation $\tilde{y}_{sim} \sim \tilde{\pi}(\cdot | \theta)$. That is,

$$\mu(\tilde{y}_{sim}) = \mu_a \mathbb{1}_{[0,\varepsilon)}(\rho(y_{obs}, \tilde{y}_{sim})) + \mu_r \mathbb{1}_{[\varepsilon,\infty)}(\rho(y_{obs}, \tilde{y}_{sim})), \quad (13)$$

where $\mu_a \in [0, 1]$ and $\mu_r \in [0, 1]$ are the continuation probabilities when $\tilde{y}_{sim}$ is accepted or, respectively, rejected. Assuming μ_a and μ_r are given, Eq (11)–(13) can be used to implement a multifidelity ABC (MF-ABC) accept-reject scheme as shown in Alg 2. To achieve optimal performance, Prescott and Baker [54] perform a training step for a set of N_w warm-up simulation pairs $(\tilde{y}_{sim}^{(1)}, y_{sim}^{(1)}), (\tilde{y}_{sim}^{(2)}, y_{sim}^{(2)}), \dots, (\tilde{y}_{sim}^{(N_w)}, y_{sim}^{(N_w)})$. This is extended to an adaptive tuning of μ_a and μ_r in Warne et al., [58], then in Prescott et al., [57] a θ dependent version of the continuation probability function is developed.

In general, positively correlated simulation pairs are essential to obtain substantial performance gains (S1 Appendix). In the Markovian setting, MF-ABC inference and various extensions can be directly implemented using the exact stochastic simulation schemes of Gillespie [5] and Anderson [15], the approximate stochastic simulation scheme of Gillespie [6] and the coupling scheme of Anderson and Higham [56]. However, to-date the literature has primarily focused on exact stochastic simulation schemes for a non-Markovian reaction networks [7,9,10,15,59]. Two notable exceptions are the neural network aided approximation scheme of Jiang et al. [70] and the recent scalable approach of Pélissier et al., [65]. While both approaches are highly effective, neither approach leads to an exact coupling scheme that is required to implement multifidelity methods. This leads to our main contribution, that is, exact and approximate stochastic simulation algorithms with appropriate coupling schemes for non-Markovian chemical reaction networks. These contributions are presented in the follow subsections.

Algorithm 2 Multifidelity ABC accept-reject scheme

Require: A low-fidelity simulator $\tilde{\pi}(\cdot | \theta)$; a high-fidelity simulator $\pi(\cdot | \theta)$; a prior distribution $\pi(\theta)$; a discrepancy function $\rho(y_{obs}, \cdot)$; a discrepancy threshold ε ; observed data y_{obs} ; a sample size N ; and continuation probability function, $\mu(\cdot)$.

for $i = 1, \dots, N$ **do**

```

Generate a proposal  $\theta^{(i)} \sim \pi(\theta)$ ;
Generate low-fidelity simulation,  $\tilde{y}_{\text{sim}} \sim \tilde{\pi}(\cdot | \theta^{(i)})$ ;
Set  $\tilde{w}^{(i)} \leftarrow \mathbb{1}_{[0, \varepsilon)}(\rho(y_{\text{obs}}, \tilde{y}_{\text{sim}}))$ ;
Sample  $M \sim \text{Bernoulli}(\mu(\tilde{y}_{\text{sim}}))$ ;
if  $M=1$  then
    Generate high-fidelity simulation,  $y_{\text{sim}} \sim \pi(\cdot | \theta)$ ;
    Set  $w^{(i)} \leftarrow \mathbb{1}_{[0, \varepsilon)}(\rho(y_{\text{obs}}, y_{\text{sim}}))$ ;
    Set  $w_{\text{mf}}^{(i)} \leftarrow \tilde{w}^{(i)} + (w^{(i)} - \tilde{w}^{(i)}) / \mu(\tilde{y}_{\text{sim}})$ ;
else
     $w_{\text{mf}}^{(i)} \leftarrow \tilde{w}^{(i)}$ ;
end if
end for
return  $(\theta^{(1)}, w_{\text{mf}}^{(1)}), (\theta^{(2)}, w_{\text{mf}}^{(2)}), \dots, (\theta^{(N)}, w_{\text{mf}}^{(N)})$ .

```

Stochastic simulation algorithms for non-Markovian processes

Stochastic simulation schemes for Markovian reaction networks have been well studied [16–18,71] with many exact and approximate stochastic simulation algorithms available. In addition, exact schemes for handling certain time-dependent propensity functions and certain types of delays are also available. For example, Anderson's modified NRM [15] handles time dependent propensities via numerical integration, and the Extrande method of Voliotis et al., [59] provides an efficient alternative based on thinning techniques. Anderson also considers delayed reactions with either fixed completion times, or random completion times that can depend on time [15]. However, Boguñá et al., [7] note that the Anderson approach cannot directly be extended to the setting where delays are random and depend on both time and system state. To deal with this Boguñá et al., [7] extend the Gillespie method through numerical inversion of the next reaction time in this general setting. By contrast, there have been relatively few works to date on approximate stochastic simulation schemes of non-Markovian systems [65,70]. For the existing approximations, there is no natural exact coupling scheme that arises. As result, application of MLMC and multifidelity schemes is not achievable without a new approximation scheme.

In this section, we present our main contributions: i) An exact stochastic simulation algorithms for non-Markovian reaction networks of the general form with state and time dependent delays. This is achieved through a novel generalisation to the Anderson modified NRM [15] for fixed delays; ii) an approximate stochastic simulation algorithm that is based on a τ -leaping discretisation of our non-Markovian modified NRM; then iii) we extend the exact coupling scheme of similar form to Anderson and Higham [56], Lester et al. [72], and Prescott et al. [57], to generate positively correlated exact and approximate simulation pairs in the non-Markovian case.

Exact stochastic simulation algorithms

The NRM [19,20] is an alternative exact stochastic simulation algorithm to the Gillespie direct method [5]. Both methods generate sample paths that exactly follow the probability law of a discrete-state continuous-time Markov process given in Eq (4). The Gillespie direct method operates through sampling the global waiting time to the next reaction and then sampling the next event conditional on this waiting time. In contrast, the NRM is derived directly from the Kurtz random time-change representation and proceeds through maintaining independent unit rate Poisson processes for each reaction channel and determining the next even via the minimum of the waiting times. This typically offers a computational advantage as only one random variable is required per event. For non-Markovian processes, such as those presented in Section, the equivalent approach to the Gillespie direct method, that is the non-Markovian Gillespie Algorithm (nM-GA) of Boguñá et al., [7], requires additional approximations in practice since the cumulative hazard function for the process rarely admits a closed-form quantile function required to sample the global waiting time.

We extend the NRM to the non-Markovian setting by considering a race of Poisson processes with time and history dependent cumulative hazard functions. That is the next reaction time of each reaction channel is computed

independently assuming the state vector $\mathbf{X}(t)$ remains unchanged. To proceed, we first define the cumulative hazard function for instantaneous reactions, $j \in \mathcal{R}_I$, and delay initiations, $j \in \mathcal{R}_D$, as

$$\Lambda_j(\tau; \mathbf{X}(t), t) = \int_0^\tau \lambda_j(\mathbf{X}(t+s), t+s) ds. \quad (14)$$

Similarly, we define the cumulative hazard function for delay completions, $j \in \mathcal{R}_D$,

$$\Lambda_j^*(\tau; \mathbf{X}(t), t, a) = \int_0^\tau \lambda_j^*(\mathbf{X}(t+s), t-a+s) ds, \quad (15)$$

where $a > 0$ is some time-shift based on the history of the reaction network. Given the definitions in Eq (14)–(15), we observe that the next event time for a instantaneous reaction $j \in \mathcal{R}_I$ or the initiation of a delayed reaction $j \in \mathcal{R}_D$, is given by solving

$$\tau_{j,u} = \Lambda_j(\tau_j; \mathbf{X}(t), t), \quad (16)$$

for τ_j , where $\tau_{j,u}$ is the next event time for the unit rate Poisson process $Y_j(\cdot)$, which is an exponentially distributed random variable with unit rate, $\tau_{j,u} \sim \text{Exp}(1)$ as determined under the Kurtz random time-change formulation (Eq (8)). Similarly, the completion of the k th delayed reaction event, $j \in \mathcal{R}_D$ with initiation times $t_{j,0}^{*,k}$, the completion time is obtained by solving

$$\tau_{j,u}^k = \Lambda_j^*(\tau_j^k; \mathbf{X}(t), t, t_{j,0}^{*,k}) \quad (17)$$

for τ_j^k , where $\tau_{j,u}^k$ is the next event time for the unit rate Poisson process $Y_j^k(\cdot)$. Note that for $t - t_{j,0}^{*,k} + s < 0$ then we define $\lambda_j^*(\mathbf{X}(t+s), t - t_{j,0}^{*,k} + s) = 0$, that is, a completion event cannot occur before it is initiated. Given the set of next event times for each channel independently, we can obtain the overall next event time for the system via $\tau_{\min} = \min(\{\tau_j\}_{j \in \mathcal{R}} \cup \{\tau_j^k\}_{j \in \mathcal{R}_D, k \in \mathcal{D}_j})$ where $\mathcal{D}_j$ is the set of delayed reactions yet to be completed, with $|\mathcal{D}_j| = X_j^*(t)$. The specific value of j and k that is associate with $\tau_{\min}$ will be the next event that occurs.

Equations (16) and (17) provide the *first reaction* time only, given the time and state dependence. Thus for all reactions channels other than the one that fired we need to solve for the remaining time left, $r_{j,u} > 0$, in the time interval of the unit rate Poisson process following the state update. Focusing on instantaneous and delay initiation channels, at time $t > t_0$ we denote the internal time, that is the time since the last reaction, as t_j . We have at time $t > t_0$, for all reactions,

$$\tau_{j,u} = r_{j,u} + \Lambda_j(t_j + \tau_{\min}; \mathbf{X}(t), t) - \Lambda_j(t_j; \mathbf{X}(t), t). \quad (18)$$

Note that the firing reaction will, according to Eq (18), result in $r_{j,u} = 0$. In this case, a new remaining time interval for the unit rate Poisson process is generated, that is, $r_{j,u} = \tau_{j,u} \sim \text{Exp}(1)$.

Eq (18) leads to an update rule for $r_{j,u}$ following any instantaneous reaction event,

$$r_{j,u} \leftarrow r_{j,u} - [\Lambda_j(t_j + \tau_{\min}; \mathbf{X}(t), t) - \Lambda_j(t_j; \mathbf{X}(t), t)]. \quad (19)$$

For the completion of the k th delay event, the equivalent update would similarly be,

$$r_{j,u}^k \leftarrow r_{j,u}^k - [\Lambda_j^*(t_j^k + \tau_{\min}; \mathbf{X}(t), t, t_{j,0}^{*,k}) - \Lambda_j^*(t_j^k; \mathbf{X}(t), t, t_{j,0}^{*,k})]. \quad (20)$$

Given remaining intervals and internal times enables the update of next reaction times. This is given for the instantaneous and initiation reactions by solving,

$$\Lambda_j(\tau_j + t_j; \mathbf{X}(t), t) = r_{j,u} + \Lambda_j(t_j; \mathbf{X}(t), t), \quad (21)$$

for τ_j , and for the delay completions by solving

$$\Lambda_j^*(\tau_j^k + t_j^k; \mathbf{X}(t), t, t_{j,0}^{*,k}) = r_{j,u}^k + \Lambda_j^*(t_j^k; \mathbf{X}(t), t, t_{j,0}^{*,k}). \quad (22)$$

for τ_j^k .

Equations (16)–(22) provide a mechanism for updating next reaction times and remaining interval lengths in the general setting of time and state dependent propensities and including time delays. This enables direct simulation of the inhomogeneous Poisson processes in Eq (8). Unlike the Markovian case, the remaining interval lengths Eq (19)–(20) must be updated for every reaction channel following every reaction event. In settings with large numbers of concurrent incomplete delay channels, a global bounding scheme on the instantaneous rates, following Pélissier et al., [65], could be implemented at the cost of some accuracy.

To track ongoing delays within reaction channels $j \in \mathcal{R}_D$ we maintain a set $\mathcal{D}_j \subset \mathbb{N}$ that contains the k indices for initiated delays that are yet to be completed. For any $k \in \mathbb{N}$, the only completion processes that need to be simulated are the $Y_j^k(\cdot)$ such that $k \in \mathcal{D}_j$. For such a process, we only need to simulate to the first event time of $Y_j^k(\cdot)$ as we take $\min(1, Y_j^k(\cdot))$, therefore after the completion event and $Y_j^k(\cdot) = 1$, we remove k from $\mathcal{D}_j$. Similarly, at the completion of a delayed reaction cannot occur before the initiation event, that is, $\lambda_j^*(\mathbf{X}(t), t - t_{j,0}^{*,k}) = 0$ when $t < t_{j,0}^{*,k}$. Thus, when the k th initiation event occurs, the index k is added to $\mathcal{D}_j$. As a result of this, the number of elements in the set $\mathcal{D}_j$ is always the same as the copy number for the ongoing reactions $X_j^*(t)$ (Eq (7)) at any time $t > t_0$.

Combining the direct implementation of the Kurtz random time-change representation with the dynamic tracking of progressing delayed reactions via the sets $\mathcal{D}_j$ for $j \in \mathcal{R}_D$ leads to a non-Markovian next reaction method (nM-NRM). The method, presented in Alg 3, is an extension of Anderson's NRM that dealt with fixed delays [15] to enable random delays that depend on both time and system state. This means we achieve the same generality as the non-Markovian Gillespie method of Boguñá et al., [7]. However, our approach leads to convenient approximation and coupling schemes that are essential for efficient sampling based on MLMC or multifidelity methods such as the multifidelity ABC method (Alg 2). In particular, the random time-change representation allows all necessary interval lengths $r_{j,u}$ and $r_{j,u}^k$ for the unit rate Poisson processes to be generated prior to the start of a simulation. This provides direct analogues of the τ -leaping schemes [6] and coupling based on common Poisson clocks [54,56].

Algorithm 3 Non-Markovian next reaction method (nM-NRM)

Require: A non-Markovian reaction network with $\mathcal{N}$ species and $\mathcal{M}$ reaction channels, of which $\mathcal{D}$ are delay reactions (Eq (6)); an initial time $t_0 \geq 0$; a final time $T > t_0$; an initial system state $\mathbf{x}_0 = \mathbf{X}(t_0)$;

1: Initialise state $\mathbf{X} \leftarrow \mathbf{x}_0$, time $t \leftarrow t_0$, internal times $t_j \leftarrow 0$ for $j \in \mathcal{R}$, next reaction intervals, $r_{j,u} \sim \text{Exp}(1)$ for $j \in \mathcal{R}$, and delay sets $\mathcal{D}_j \leftarrow \emptyset$, $j \in \mathcal{R}_D$;

while $t < T$ **do**

 Update reaction times τ_j for $j \in \mathcal{R}_I$, and τ_j^k for $k \in \mathcal{D}_j$, $j \in \mathcal{R}_D$ (Eq (21)–(22));

 Find $\tau_{\mu_1} \leftarrow \min_{j \in \mathcal{R}} (\tau_j)$, and $\tau_{\mu_2}^k \leftarrow \min_{j \in \mathcal{R}_D, k \in \mathcal{D}_j} (\tau_j^k)$;

 Set next event time $\tau_{\min} \leftarrow \min(\tau_{\mu_1}, \tau_{\mu_2}^k)$;

 Update intervals $r_{j,u}$ for $j \in \mathcal{R}$, and $r_{j,u}^k$ for $k \in \mathcal{D}_j$, $j \in \mathcal{R}_D$ (Eq (19)–(20));

 Update internal times $t_j \leftarrow t_j + \tau_{\min}$ for $j \in \mathcal{R}$, and $t_j^k \leftarrow t_j^k + \tau_{\min}$ for $k \in \mathcal{D}_j$, $j \in \mathcal{R}_D$;

if $\tau_{\min} = \tau_{\mu_1}$ **then**

 Set $r_{\mu_1,u} \sim \text{Exp}(1)$;

```

if  $\mu_1 \in \mathcal{R}_I$  then
    Set  $\nu \leftarrow \nu_{*,\mu_1}$ ;
else
    Set  $\nu \leftarrow \eta_{\mu_1} - \nu_{*,\mu_1}^-$ ;
    Set  $k' \leftarrow X_{\mu_1}^* + 1$ ;
    Set  $t_{\mu_1,0}^{*,k'} \leftarrow t + \tau_{\mu_1}$ ,  $t_{\mu_1}^{k'} \leftarrow 0$ , and  $t_{\mu_1,u}^{k'} \sim \text{Exp}(1)$ ;
    Initiate delayed reaction  $\mathcal{D}_{\mu_1} \leftarrow \mathcal{D}_{\mu_1} \cup \{k'\}$ ;
end if
else
    Set  $\nu \leftarrow \nu_{*,\mu_2}^+ - \eta_{\mu_2}$ ;
    Complete delayed reaction  $\mathcal{D}_{\mu_2} \leftarrow \mathcal{D}_{\mu_2} \setminus \{k\}$ ;
end if
Update the system state  $\mathbf{X} \leftarrow \mathbf{X} + \nu$ , and time  $t \leftarrow t + \tau_{\min}$ ;
end while

```

Approximate stochastic simulation algorithms

The τ -leaping method introduced by Gillespie [6] is an approximate stochastic simulation algorithm that efficiently simulates reaction networks by discretising time and allowing for multiple reaction events per time step. The τ -leaping method is widely used in the simulation of reaction networks where exact methods like the Gillespie algorithm are prohibitively expensive [16,18]. While many τ -leaping variants exist, to our knowledge, none are suitable for non-Markovian reaction networks where reaction propensities may depend on internal clocks or history. Leveraging the general framework for non-Markovian reaction network with delays, given by Eq (8), we develop here an extension to the τ -leaping algorithm to approximate sample paths.

The nM-NRM (Alg 3) proceeds by explicitly simulating every reaction event that occurs. As a result, the computational cost depends heavily on the global average reaction rate. The τ -leaping method mitigates this by invoking a simplifying assumption that the system state changes very little over the interval $[t, t + \tau)$, that is $\mathbf{X}(t + s) \approx \mathbf{X}(t)$ for $s \in [0, \tau)$. We denote $\mathbf{Z}(t) = [Z_1(t), Z_2(t), \dots, Z_N(t), Z_{j_1}^*(t), Z_{j_2}^*(t), \dots, Z_{j_D}^*(t)]$, as the approximation to $\mathbf{X}(t)$ that arises by assuming $\mathbf{Z}(t + s) = \mathbf{Z}(t)$ for $s \in [0, \tau)$. This leads to an approximation of the cumulative hazard functions,

$$\begin{aligned}\Lambda_j(\tau; \mathbf{X}(t), t) &\approx \tilde{\Lambda}_j(\tau; \mathbf{Z}(t), t) = \int_0^\tau \lambda_j(\mathbf{Z}(t), t + s) ds, \\ \Lambda_j^*(\tau; \mathbf{X}(t), t, a) &\approx \tilde{\Lambda}_j^*(\tau; \mathbf{Z}(t), t, a) = \int_0^\tau \lambda_j^*(\mathbf{Z}(t), t - a + s) ds.\end{aligned}\quad (23)$$

By substitution of Eq (23) into Eq (8) we obtain

$$\mathbf{Z}(t + \tau) = \mathbf{Z}(t) + \sum_{j \in \mathcal{R}_I} P_j \nu_{*,j} + \sum_{j \in \mathcal{R}_D} P_j (\eta_j - \nu_{*,j}^-) + \sum_{j \in \mathcal{R}_D} \sum_{\kappa=1}^{\infty} \min \{ Z_j^{*,\kappa}(t), P_j^\kappa \} (\nu_{*,j}^+ - \eta_j), \quad (24)$$

where $P_j \sim \text{Poisson}(\tilde{\Lambda}_j(\tau; \mathbf{Z}(t), t))$ and $P_j^\kappa \sim \text{Poisson}(\tilde{\Lambda}_j^*(\tau; \mathbf{Z}(t), t, t_{j,0}^{*,\kappa}))$ are independent Poisson random variates. Note that in most cases the integrals in Eq (23) are analytically tractable. Unlike Eq (8) where every delayed reaction initiation has a unique initiation time $t_{j,0}^{*,k}$, the initiation process described by Eq (24), for $j \in \mathcal{R}_D$, will initiate delays in groups of size $P_j > 0$. Within this group, each delay will have the same initiation time of $t_{j,0}^{*,\kappa} = t + \tau$. Note, we adopt the notation $t_{j,0}^{*,\kappa}$ for the initiation time of the κ th group in the time-discretised approximation, this is not to be confused with the notation of $t_{j,0}^{*,k}$ for the exact initiation times in the continuous-time process in Eq (8). The grouping occurs due to the time discretisation. That is, over time interval $[t, t + \tau)$ exactly $P_j \sim \text{Poisson}(\tilde{\Lambda}_j(\tau; \mathbf{Z}(t), t))$ initiation events occur and enter the system in bulk at time $t_{j,0}^{*,\kappa} = t + \tau$ which is the next time-step. As a result, all of these P_j delayed

chemical species are identical and can be treated with the same reaction channel. That is at the κ th initiation group, we set $Z_j^{*,\kappa} = P_j$ for $j \in \mathcal{R}_D$ and include κ in the delay group set $\mathcal{D}_j$. This new reaction channel remains in the system until it has fired a total of $Z_j^{*,\kappa}$ times, after which κ is removed from the delay set $\mathcal{D}_j$. This leads to the non-Markovian τ -leaping method (nM-TLM) in Alg 4.

Algorithm 4 Non-Markovian τ -leaping method (nM-TLM)

Require: A non-Markovian reaction network with $\mathcal{N}$ species and $\mathcal{M}$ reaction channels, of which $\mathcal{D}$ are delay reactions (Eq (6)); an initial time $t_0 \leq 0$; a final time $T > t_0$; an initial approximate system state $\mathbf{z}_0 = \mathbf{Z}(t_0)$; a time discretisation τ ;

Initialise state $\mathbf{Z} \leftarrow \mathbf{z}_0$, time $t \leftarrow t_0$, internal times $t_j \leftarrow 0$ for $j \in \mathcal{R}$, and delay sets $\mathcal{D}_j \leftarrow \emptyset$, $j \in \mathcal{R}_D$;
while $t < T$ **do**
 Evaluate $\tilde{\Lambda}_j(\tau; \mathbf{Z}(t), t)$ for $j \in \mathcal{R}$, and $\tilde{\Lambda}_j^*(\tau; \mathbf{Z}(t), t, t_{j,0}^{*,\kappa})$ for $\kappa \in \mathcal{D}_j$ and $j \in \mathcal{R}_D$;
 Sample $P_j \sim \text{Poisson}(\tilde{\Lambda}_j(\tau; \mathbf{Z}(t), t))$ for $j \in \mathcal{R}$;
 Sample $P_j^\kappa \sim \text{Poisson}(\tilde{\Lambda}_j^*(\tau; \mathbf{Z}(t), t_{j,0}^{*,\kappa}))$ for $\kappa \in \mathcal{D}_j$ and $j \in \mathcal{R}_D$;
 Update system state using Eq (24);
 Update delayed species $Z_j^{*,\kappa} \leftarrow \max(0, Z_j^{*,\kappa} - P_j^\kappa)$ for $\kappa \in \mathcal{D}_j$ and $j \in \mathcal{R}_D$;
 Remove completed delay channels $\mathcal{D}_j \leftarrow \mathcal{D}_j \setminus \{\kappa : Z_j^{*,\kappa} = 0\}$ for $j \in \mathcal{R}_D$;
 for $j \in \mathcal{R}_D$ such that $P_j > 0$ **do**
 Set $\kappa' \leftarrow Z_j^* + 1$
 Set $t_{j,0}^{*,\kappa'} \leftarrow t + \tau$ and $Z_j^{*,\kappa'} \leftarrow P_j$ for $j \in \mathcal{R}_D$;
 Update $\mathcal{D}_j \leftarrow \mathcal{D}_j \cup \{\kappa'\}$
 end for
 Update time $t \leftarrow t + \tau$;
end while

While we do not analytically consider the rate of convergence, it is straightforward to show using the same reasoning as Gillespie [6] that Eq (24) converges to Eq (8) as $\tau \rightarrow 0$. We report numerical results for both weak and strong convergence rates in the Results Section.

Unbiased multilevel Monte Carlo

Given the exact and approximate stochastic simulation schemes (Algs 3 and 4), we can develop a coupling scheme that generates an exact simulation that is driven by the same inhomogeneous Poisson processes underlying a given approximate simulation (See S2 Appendix for details). Using this coupling scheme we can implement an unbiased MLMC estimator following the approaches of Anderson and Higham [56] and Lester et al., [72]. In this setting, we aim to estimate the expectation,

$$\mathbb{E}[f(\mathbf{X}(T))] = \sum_{\mathbf{x}_T \in \mathbb{N}^{\mathcal{N}}} f(\mathbf{x}_T) \Pr(\mathbf{X}(T) = \mathbf{x}_T \mid \mathbf{X}(t_0) = \mathbf{x}_0),$$

where $f(\cdot)$ is some function of the system state, and $\Pr(\mathbf{X}(T) = \mathbf{x}_T \mid \mathbf{X}(t_0) = \mathbf{x}_0)$ is the solution to the forward problem of the non-Markovian network at time $t = T$, given an initial state $\mathbf{X}(t_0) = \mathbf{x}_0$. We expand the expectation using the telescoping summation,

$$\mathbb{E}[f(\mathbf{X}(T))] = \mathbb{E}[f(\mathbf{Z}(T))] + \mathbb{E}[f(\mathbf{X}(T)) - f(\mathbf{Z}(T))],$$

where $\mathbf{Z}(T)$ is the approximate stochastic process as determined by Eq (24) for a given discretisation τ . This leads to the MLMC estimator

$$\hat{f}_{\mathbf{X}} = \frac{1}{N_0} \sum_{i=1}^{N_0} f(\mathbf{Z}^{(i)}(T)) + \frac{1}{N_1} \sum_{i=1}^{N_1} f(\mathbf{X}^{(i)}(T)) - f(\mathbf{Z}^{(i)}(T)). \quad (25)$$

Here, $\mathbf{Z}^{(1)}(T), \mathbf{Z}^{(2)}(T), \dots, \mathbf{Z}^{(N_0)}(T)$ are i.i.d. samples from the nM-TLM (Alg 4) and $(\mathbf{X}^{(1)}(T), f(\mathbf{Z}^{(1)}(T))), (\mathbf{X}^{(2)}(T), f(\mathbf{Z}^{(2)}(T))), (\mathbf{X}^{(N_1)}(T), f(\mathbf{Z}^{(N_1)}(T)))$ i.i.d. coupled nM-NRM and nM-TLM pairs (Alg 4 and [S2 Appendix](#)). For a target mean square error (MSE), h^2 , the optimal choice of sample sizes, N_ℓ for $\ell = 0, 1$, is given by

$$N_\ell \propto \frac{1}{h^2} \sqrt{\frac{V_\ell}{C_\ell}} [\sqrt{V_0 C_0} + \sqrt{V_1 C_1}]$$

where $V_0 = \text{Var}[f(\mathbf{Z}(T))]$ and $V_1 = \text{Var}[f(\mathbf{X}(T)) - f(\mathbf{Z}(T))]$, and C_0 and C_1 are the expected computational cost (i.e., total simulation runtime) of a draw from $\mathbf{Z}(T)$ and $(\mathbf{X}(T), \mathbf{Z}(T))$, respectively. Importantly, since we can generate exact couplings between $\mathbf{X}(T)$ and $\mathbf{Z}(T)$, the resulting MLMC estimator ([Eq \(25\)](#)) is unbiased regardless of the choice of discretisation τ .

Results

In this section, we demonstrate the performance and applicability of the algorithms introduced in Section 2. We begin by specifying two representative examples of non-Markovian reaction networks that include both delays and history-dependent dynamics. Using these models, we demonstrate sample paths generated by our algorithms and evaluate the accuracy and efficiency of the nM-TLM approximation. We then apply our exact coupling scheme to demonstrate accelerated simulation and inference using unbiased MLMC simulation [\[49,56\]](#) and multifidelity ABC for inference [\[54\]](#). Our results show substantial estimator efficiency improvements of an order of magnitude are feasible using our methods compared with direct MC sampling.

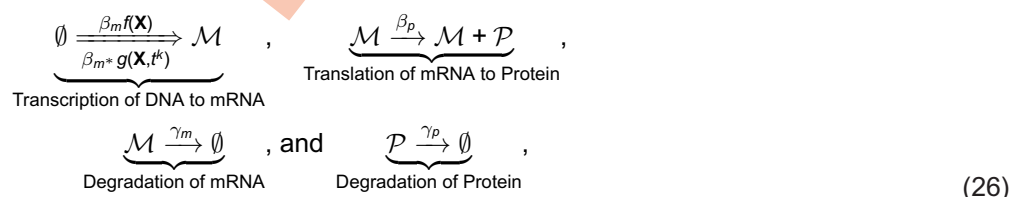
Example non-Markovian reaction networks

We consider two examples of non-Markovian biochemical reaction networks to demonstrate the efficacy of our methods. The first is a genetic regulation network that includes auto-inhibition with oscillatory dynamics arising from a delay in messenger RNA transcription [\[73,74\]](#). The second is a non-Markovian extension to the susceptible-infectious-susceptible (SIS) model from epidemiology, in which both the infection and recovery processes have Weibull waiting time distributions.

Example 1: Gene regulation with delayed auto-inhibition

The model describes the transcription of a gene coding DNA sequence into an messenger RNA (mRNA) molecule and subsequent translation of mRNA to proteins that, in-turn, auto-inhibit the transcription process. The process of transcription and subsequent transport of mRNA from cell nucleus to cytoplasm is non-negligible [\[22,73\]](#), thus we define the transcription process as a delay reaction. We also consider the resources required for transcription to be finite [\[7\]](#), therefore the processing time slows as the total number of mRNA in the increases, requiring a state-dependent completion function. Such delay models are of interest in the study of gene regulation as they are capable of oscillatory behaviour arising from auto-inhibitory mechanism, such as is observed in the Hes1 system associated with cell differentiation [\[73,75\]](#).

The delayed auto-inhibition model is a non-Markovian reaction network consisting of one delayed reaction, and three instantaneous reactions. The network is given by,



where chemical species $\mathcal{M}$ and $\mathcal{P}$, respectively, represent the mRNA and protein molecules. For the delay reaction for mRNA transcription, $\beta_m > 0$ is the maximum initiation rate and $\beta_{m^*} > 0$ is the maximum completion rate. For the remaining instantaneous reactions, β_p is the translation rate, γ_m is the mRNA degradation rate, and γ_p is the protein degradation rate. Based on the formalism from sections and we have reaction sets $\mathcal{R}_D = \{1\}$ and $\mathcal{R}_I = \{2, 3, 4\}$, the system state at time $t > t_0$ is $\mathbf{X}(t) = [M(t), P(t), M^*(t)]^T$ with the copy numbers given by $M(t) \geq 0$ for mRNA, $P(t) \geq 0$ for proteins, and $M^*(t) \geq 0$ for incomplete mRNA transcriptions. The stoichiometric matrices and elementary vectors described for Eq (8) are

$$\nu^- = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \nu^+ = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \nu = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \text{ and } \eta_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

and the propensity functions are

$$\lambda_1(\mathbf{X}(t)) = \beta_m f(\mathbf{X}(t)), \quad (27)$$

$$\begin{aligned} \lambda_1^*(\mathbf{X}(t), t, t_{0,1}^{*,k}) &= \beta_{m^*} g(\mathbf{X}(t), t - t_{0,1}^{*,k}), \\ \lambda_2(\mathbf{X}(t)) &= \beta_p M(t), \\ \lambda_3(\mathbf{X}(t)) &= \gamma_m M(t), \\ \lambda_4(\mathbf{X}(t)) &= \gamma_p P(t). \end{aligned} \quad (28)$$

The mRNA transcription propensities for initiation (Eq (27)) and completion (Eq (28)) are both regulated by Hill functions. For the transcription initiation, the Hill function depends on the protein copy numbers $P(t)$ and capture the auto-inhibition mechanism,

$$f(\mathbf{X}(t)) = \frac{P_a^h}{P_a^h + P(t)^h}, \quad (29)$$

where $h > 0$ is the Hill coefficient and P_a is a constant such that if $P(t) = P_a$ then $f(\mathbf{X}(t)) = 1/2$. We note that $f(\mathbf{X}(t)) \in (0, 1)$, since $f(\mathbf{X}(t)) \rightarrow 0$ as $P(t) \rightarrow \infty$, and $f(\mathbf{X}(t)) \rightarrow 1$ as $P(t) \rightarrow 0$. In the case of transcription completion, the Hill function depends on the mRNA copy numbers, $M(t)$ and $M^*(t)$, and the internal time $t_1^k = t - t_{0,1}^{*,k}$ to capture the delay effect and finite transcription resources,

$$g(\mathbf{X}(t), t^k) = \frac{\alpha M_a^\alpha (\beta_{m^*} t^k)^{\alpha-1}}{M_a^\alpha + v(M(t) + M^*(t))^\alpha}. \quad (30)$$

Here, M_a and v are constants such that if $M(t) + M^*(t) = M_a$ then $g(\mathbf{X}(t), t_1^k) = \alpha (\beta_{m^*} t^k)^{\alpha-1} / (1 + v)$ and the shape parameter, $\alpha > 0$, is the only parameter that affects the existence of the non-Markovian delay as it leads to a Weibull distribution for the waiting time. For fixed internal time t_1^k we have that $g(\mathbf{X}(t), t_1^k) \in (0, \alpha (\beta_{m^*} t^k)^{\alpha-1})$ using similar arguments to those used for Eq (29). Note that if $\alpha = 1$, the dependence on the internal time is removed and the system becomes Markovian. However, when $\alpha > 1$, the completion event becomes more likely to occur as the internal time increases. Conversely, when $\alpha < 1$, the completion event becomes less likely to occur as the internal time increases and can lead to stalling of the transcription process as $g(\mathbf{X}(t), t^k) \rightarrow 0$ as the slowing completion processes build up (e.g., $M^* \rightarrow \infty$).

Example 2: A non-Markovian epidemic model

Our second example is a non-Markovian extension to the classic Markovian SIS epidemic model, typically used to model the spread of an infection that does not induce any form of immunity upon recovery. Our non-Markovian network is given by



where the species S and $\mathcal{I}$, respectively, represent the susceptible and infectious individuals in the population of size P . The rate $\beta_I^* > 0$ is the maximum completion rate for infection process and the maximum initiation rate for the recovery process. Similarly, $\beta_R^* > 0$ is the maximum completion rate for recovery process and the maximum initiation rate for the infection process. Here, both reactions are delay type with $\mathcal{R}_D = \{1, 2\}$ and $\mathcal{R}_I = \emptyset$. The system state at time $t > t_0$ is given by $\mathbf{X}(t) = [S(t), I(t), S^*(t), I^*(t)]$ with the total population given, $S(t) \geq 0$ to susceptibles, $I(t) \geq 0$ for infectious, $S^*(t) \geq 0$ for infectious that are recovering, and $I^*(t)$ for susceptibles becoming infectious. Due to the form of the two reactions in Eq (31), the completion event for the recovery reaction is the initiation event for the infection reaction and vice-versa. As a result, beyond the initial condition, the population is effectively split between the delay states $S^*(t)$ and $I^*(t)$.

The stoichiometric matrices are given by

$$\nu^- = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \nu^+ = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \nu = \begin{bmatrix} -1 & 1 \\ 1 & -1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \eta_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \text{ and } \eta_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

To completely define the propensities, we denote the internal time for the k th recovery and infection completion process, respectively, as $t_R^k = t - t_{0,R}^{*,k}$ and $t_I^k = t - t_{0,I}^{*,k}$. Here, $t_{0,R}^{*,1}, t_{0,R}^{*,2}, \dots$, are initiation times for recovery reactions arising from infection completion events. Similarly, $t_{0,I}^{*,1}, t_{0,I}^{*,2}, \dots$, are initiation times for infection reactions arising from recovery completion events. The propensity functions are,

$$\lambda_1(\mathbf{X}(t), t, t_{0,R}^{*,k}) = \lambda_2(\mathbf{X}(t), t, t_{0,R}^{*,k}) = \beta_R^* g_R^*(\mathbf{X}(t), t - t_{0,R}^{*,k}), \quad (32)$$

$$\lambda_1^*(\mathbf{X}(t), t, t_{0,I}^{*,k}) = \lambda_2^*(\mathbf{X}(t), t, t_{0,I}^{*,k}) = \beta_I^* g_I^*(\mathbf{X}(t), t - t_{0,I}^{*,k}), \quad (33)$$

where $g_R^*(\mathbf{X}(t), t_R^k) = \alpha_R (\beta_R^* t_R^k)^{\alpha_R - 1}$ and $g_I^*(\mathbf{X}(t), t_I^k) = \alpha_I (\beta_I^* t_I^k)^{\alpha_I - 1} (I(t) + S^*(t))/P$ with the total population being $P = S(t) + I(t) + S^*(t) + I^*(t)$, which is a conserved constant quantity. Just as in the delayed gene regulation model, the shape parameters, $\alpha_R > 0$ and $\alpha_I > 0$, correspond to non-Markovian delay behaviour with Weibull completion times, and in the special case that $\alpha_R = \alpha_I = 1$ then the system becomes Markovian.

Simulation results

To evaluate the practical performance of our simulation algorithms, we begin by examining sample paths generated from the delayed gene regulation model (Eqs. (29)–(30)). We generate independent stochastic simulations of this model using both the exact nM-NRM (Alg 3) and the approximate nM-TLM (Alg 4). For all auto-inhibition model simulations in these results we use the parameter values $\beta_m = 10$, $\beta_m^* = 0.175$, $\beta_p = 1$, $\gamma_m = 0.08$, $\gamma_p = 0.05$, $\alpha = 2.5$, $M_a = 10$, $\nu = 0.5$, $h = 1.5$, and $P_a = 5$, and the initial conditions are set to $\mathbf{X}_0 = [0, 0, 0]$ [7].

By inspection of the auto-inhibition model (Eq (26)), we see that in the limit as $\alpha \rightarrow 1$, the system becomes Markovian as the function $g^*(\cdot)$ loses its time dependence. As a result, our nM-NRM approach should reduce precisely to the Markovian NRM [15,19] and give equivalent results to the standard Markovian GDM [4]. To confirm this we generate MC averaged simulation paths, using $n=4,000$ simulations, for several delay shape parameter values $\alpha = 0.5, 1.0, 1.5, 2.5$. We then compare against the same average generated from the Markovian GDM by removing the time dependence on the transcription completion event. As shown in Fig 1, we see that there is excellent agreement between our nM-NRM for $\alpha = 1$ and the Markovian GDM. Furthermore we note the effect of the history on the dynamics decreases as $\alpha \rightarrow 1$.

Fig 2 presents example sample paths from the exact nM-NRM (Alg 3; Fig 2(a)–2(b)) and approximate nM-TLM (Alg 4; Fig 2(c)–2(d)) showing the evolution of mRNA $M(t)$, proteins $P(t)$, and progressing mRNA transcriptions $M^*(t)$. Using a small number of sample paths ($n=5$) we can observe qualitatively similar behaviour between exact realisations (Fig 2(a)) and their approximate counterparts (Fig 2(c)). In both cases, we observe the expected effect of the auto-inhibition (Eq

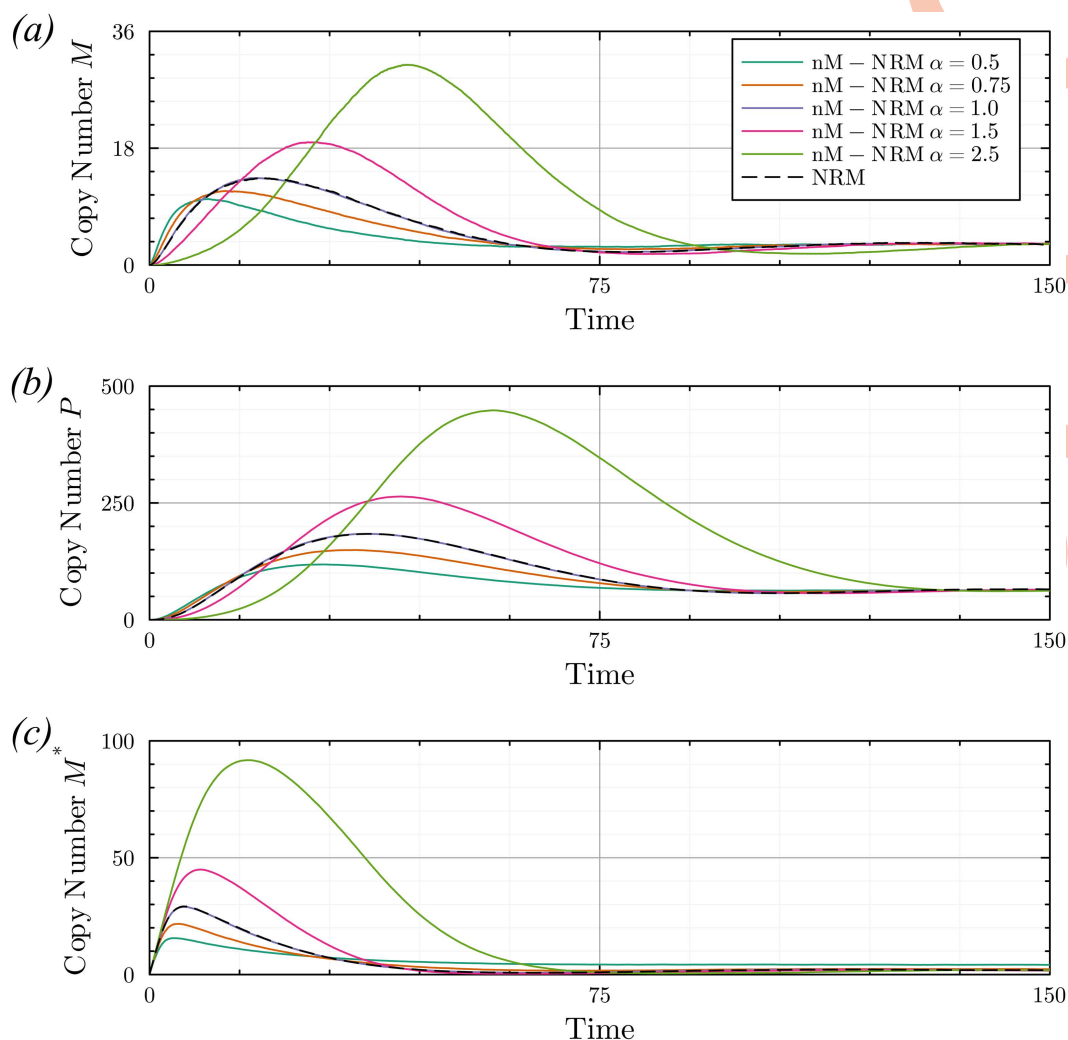


Fig 1. Demonstration of convergence of the non-Markovian setting to the Markovian setting as $\alpha \rightarrow 1$. Expected samples paths, computed using $n=4,000$ simulations, are shown for: (a) mRNA molecules M , (b) Protein molecules P ; and (c) progressing transcriptions M^* . There is excellent agreement between the $\alpha = 1$ case (purple solid line) and the standard NRM (black dashed line).

<https://doi.org/10.1371/journal.pcbi.1014687.g001>

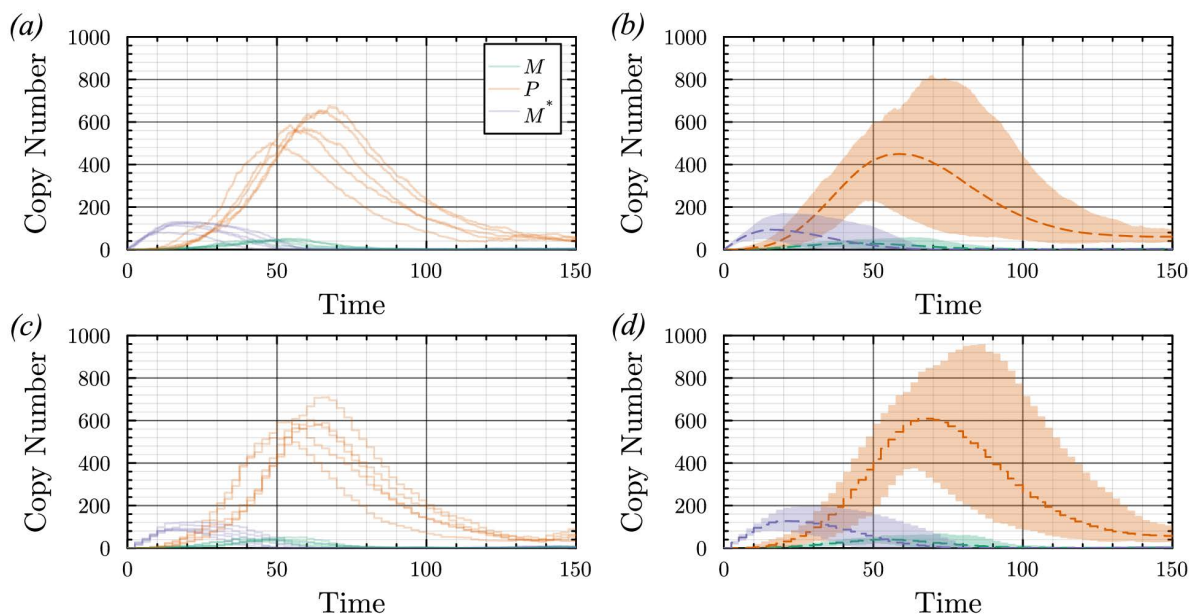


Fig 2. Example realisations of the gene regulation with delayed autoinhibition model (Eqs. (29)–(30)) using the exact nM-NRM (a)–(b) and approximate nM-TLM with time step $\tau = 2.5$ (c)–(d). $n=5$ independent paths are shown for (a) nM-NRM and (c) nM-TLM. Path distributions are shown using $n=100$ realisations for (b) nM-NRM and (d) nM-TLM with solid dashed lines indicating mean copy numbers. Model parameters are $\beta_m = 10$, $\beta_{m^*} = 0.175$, $\beta_p = 1$, $\gamma_m = 0.08$, $\gamma_p = 0.05$, $\alpha = 2.5$, $M_a = 10$, $v = 0.5$, $P_a = 5$, and $h = 1.5$.

<https://doi.org/10.1371/journal.pcbi.1014687.g002>

(29)) with increase then decline in protein copy numbers, eventually leading to smaller oscillations. However, as expected path-wise numerical error can be observed due to the nM-TLM discretisation $\tau = 2.5$ (Fig 2(c)) compared with exact sample paths (Fig 2(a)). This error propagates into the path distributions (Fig 2(b), 2(d)) estimated with a larger number of sample paths ($n = 100$). Here we see that the nM-TLM (Fig 2(d)) is consistently over-estimating the peak of the protein path (mean $P \approx 600$; Fig 2(d)) in comparison to the exact nM-NRM (mean $P \approx 400$; Fig 2(c)). In addition, the peak of the protein in the nM-TLM approximation is occurring at a later time (mean peaks at $t \approx 75$; Fig 2(d)) compared with the exact nM-NRM (mean peaks at $t \approx 55$; Fig 2(b)).

We also demonstrate the correctness of our coupling scheme. To do this we generate nM-TLM realisations using Alg 4 and generate $n = 10$ coupled exact nM-NRM realisations (S2 Appendix). This is then compared against $n = 10$ independent realisations of the exact nM-NRM using Alg 3. The results, shown in Fig 3, give a comparison of the independent (Fig 3(a), 3(c)) and coupled (Fig 3(b), 3(d)) cases for two different step sizes, $\tau = 2.5$ (Fig 3(a)–3(b)) and $\tau = 0.05$ (Fig 3(c)–3(d)). We note a substantial reduction in variance in the exact realisations due to the coupling scheme (compare Fig 3(a) with Fig 3(b) and Fig 3(c) with Fig 3(d)). The coupled simulations also clearly demonstrate the approximation error due to the discretisation $\tau = 2.5$ (Fig 3(b)) and convergence of nM-TLM to the nM-NRM as $\tau \rightarrow 0$ (Fig 3(d)). Note that the $n = 10$ exact simulations in Fig 3(a) are independent simulations to the $n = 5$ samples in Fig 2(a).

Using coupled simulations pairs $(\mathbf{X}(T), \mathbf{Z}(T))$, we can numerically estimate the error rates of our nM-TLM. Specifically we consider the weak error,

$$E_{\text{weak}} = |\mathbb{E}[\mathbf{X}(T)] - \mathbb{E}[\mathbf{Z}(T)]|, \quad (34)$$

and the strong error,

$$E_{\text{strong}} = \mathbb{E} [|\mathbf{X}(T) - \mathbf{Z}(T)|] . \quad (35)$$

Using direct MC with $n = 120,000$ simulations and $T = 300$, we estimate [Eq \(34\)](#) and [Eq \(35\)](#) for a range of values for τ . This enables us to quantify empirical convergence rates in both the weak and strong sense. These empirical rates are shown in [Fig 4](#) along with reference lines for typical weak and strong orders of convergence for a first order scheme such as standard Markovian τ -leaping method and the Euler-Maruyama scheme for stochastic differential equations [6, 16]. That is, order $\mathcal{O}(\tau)$ for the weak convergence rate and order $\mathcal{O}(\tau^{1/2})$ for the strong convergence rate. We observe good agreement between the empirical rates and the expected theoretical rates.

Finally, we demonstrate that this coupling scheme is also effective for driving computational gains using the unbiased MLMC estimator in [Eq \(25\)](#). Here we use the non-Markovian SIS model ([Eq \(31\)](#)), with parameters $\alpha_R = 1.8$, $\beta_R^* = 1.5$, $\alpha_I = 1.3$, $\beta_I^* = 2.0$, and initial conditions $S^*(t_0) = 1,583$, $I^*(t_0) = 417$, $S(t_0) = I(t_0) = 0$ [7]. We apply the unbiased MLMC to estimate $\mathbb{E}[I^*(T)]$ at $T = 3$, for a range of discretisations $\tau \in \{0.01, 0.05, 0.1, 0.2, 0.5, \}$. We take a range of samples sizes (up to $N = 4,000$) for direct MC estimators using the exact nM-NRM (Alg 3), and apply the MLMC estimator ([Eq \(25\)](#)) with optimal N_0 and N_1 for each value of τ for equivalent target mean-squared error (MSE) for the direct MC. For each sample configuration (samples size and τ), we estimate the MSE and total computational cost (that is, the runtime).

Results in [Fig 5](#) show that we can achieve the substantial performance improvements of an order of magnitude (in terms of MSE for equivalent cost to direct MC). An important observation here is that smaller τ can be more computationally efficient than larger τ as the effect of the coupling is stronger in this setting.

Inference results

Given the simulation results ([Figs 2–4](#)), we have the necessary components to implement the multifidelity ABC inference scheme (Alg 2). We generate a synthetic dataset using nM-NRM for the delayed auto-inhibition model (Alg 3) to replicate an observed experimental time series of gene transcription. We then evaluate the efficiency of the multifidelity ABC scheme that is accelerated using the nM-TLM (Alg 4) with exact coupling ([S2 Appendix](#)).

The observed data is denoted as $y_{\text{obs}} = [\mathbf{X}(t_1), \mathbf{X}(t_2), \dots, \mathbf{X}(t_n)]$ where $\mathbf{X}(t_i) = [M(t_i), P(t_i), M^*(t_i)]^T$ with the observations times $t_i = t_0 + 5i$ for $i \leq 20$ and $t_i = t_0 + 20i$ for $i > 20$. The data is generated using a single exact realisation of the delayed auto-inhibition model ([Eq \(26\)](#)) using the nM-NRM (Alg 3) with initial condition $M(t_0) = P(t_0) = M^*(t_0) = 0$ and parameter values $\beta_m = 10$, $\beta_{m^*} = 0.175$, $\beta_p = 1$, $\gamma_m = 0.08$, $\gamma_p = 0.05$, $\alpha = 2.5$, $M_a = 10$, $v = 0.5$, $h = 1.5$, and $P_a = 5$. This synthetic data and the full underlying realisation is shown in [Fig 6](#). Note the observations times start at a higher resolution to capture the initial peak in the gene expression.

For the inference problem we consider the mRNA maximum completion rate, β_{m^*} , the protein production rate parameters, β_p , and the transcription completion time shape parameter, α , to be unknown and to be inferred from the data, y_{obs} . Specifically we consider the problem of estimating the posterior mean,

$$\mathbb{E}[\theta | y_{\text{obs}}] = \int_{\Theta} \theta \pi(\theta | y_{\text{obs}}) d\theta,$$

where $\theta = [\beta_{m^*}, \beta_p, \alpha]^T$. The priors used are independent uniform priors, $\beta_{m^*} \sim \mathcal{U}(0, 0.5)$, $\beta_p \sim \mathcal{U}(0, 5)$ and $\alpha \sim \mathcal{U}(0, 10)$. We choose uniform priors for simplicity and highlight that our methods are not restricted to uniform priors. We use this problem set-up to demonstrate the efficiency gains we obtain through the multifidelity ABC method.

We compare the ABC accept-reject scheme (Alg 1) with exact nM-NRM simulations (Alg 3) against the multifidelity ABC accept-reject of Prescott and Baker [54] that uses the nM-TLM (Alg 4) for low-fidelity simulations and the coupling scheme ([S2 Appendix](#)) for the correction terms ([Eq \(11\)](#)). In both cases the ABC approximation is taken with respect to the Euclidean discrepancy metric on the full data, that is

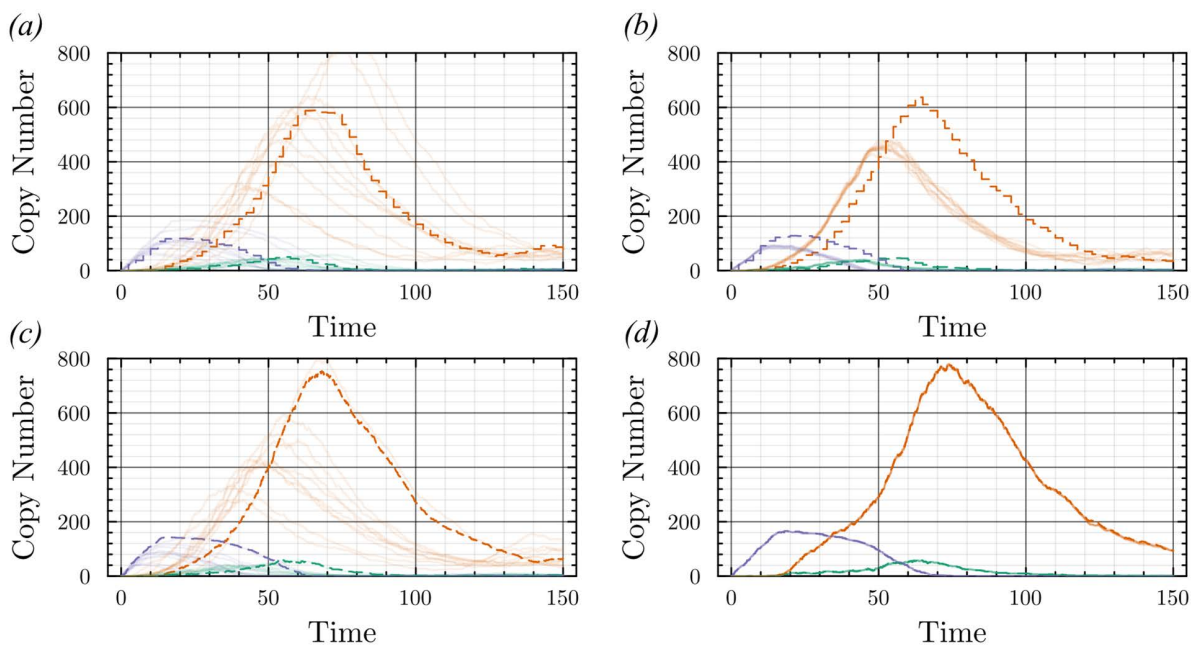


Fig 3. Demonstration of the coupling scheme (S2 Appendix) for nM-NRM realisations (solid lines) that are correlated to a given nM-TLM realisation (dashed lines). Each panel shows $n=10$ independent nM-NRM simulations that are: (a) uncoupled and independent of the nM-TLM simulation with $\tau = 2.5$; (b) coupled and correlated to the nM-TLM simulation with $\tau = 2.5$; (c) uncoupled and independent of the nM-TLM simulation with $\tau = 0.05$; and (d) coupled and correlated to the nM-TLM simulation with $\tau = 0.05$. Model parameters are $\beta_m = 10$, $\beta_{m^*} = 0.175$, $\beta_p = 1$, $\gamma_m = 0.08$, $\gamma_p = 0.05$, $\alpha = 2.5$, $M_a = 10$, $v = 0.5$, $P_a = 5$, and $h = 1.5$.

<https://doi.org/10.1371/journal.pcbi.1014687.g003>

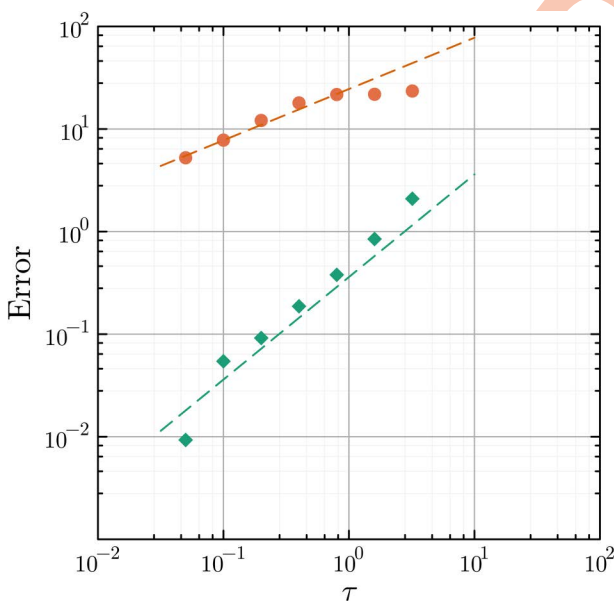


Fig 4. Empirical convergence of the nM-TLM under weak error (green diamonds) and strong error (orange circles). Lines corresponding to a weak convergence rate of $\mathcal{O}(\tau)$ (green dashed line) and a strong convergence rate of $\mathcal{O}(\tau^{1/2})$ (orange dashed line) are shown for reference.

<https://doi.org/10.1371/journal.pcbi.1014687.g004>

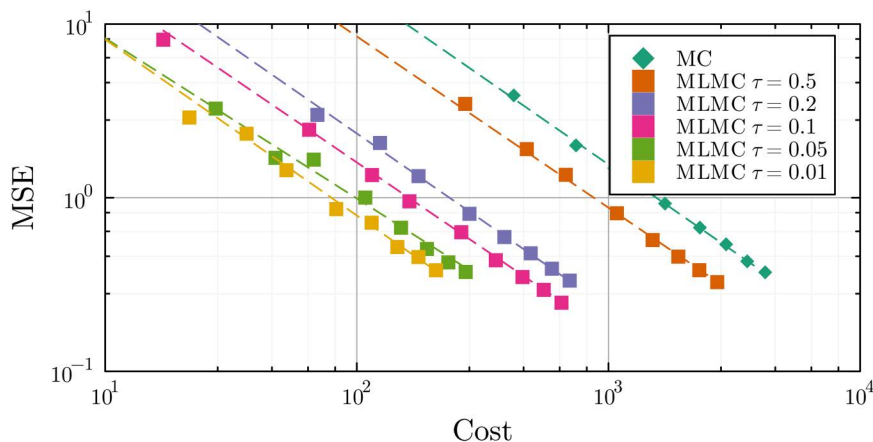


Fig 5. Illustration of the MSE of the estimate of $\mathbb{E}[I^*(T)]$ as a function of computational cost (that is, the total runtime in seconds) for the direct MC estimator compared with the MLMC estimator with different nM-TLM discretisations τ . The target expectation is $\mathbb{E}[I^*(T)]$ under the non-Markovian SIS model with $T=3$, $\alpha_R = 1.8$, $\beta_R^* = 1.5$, $\alpha_I = 1.3$, $\beta_I^* = 2.0$, and initial conditions $S^*(t_0) = 1,583$, $I^*(t_0) = 417$, $S(t_0) = I(t_0) = 0$.

<https://doi.org/10.1371/journal.pcbi.1014687.g005>

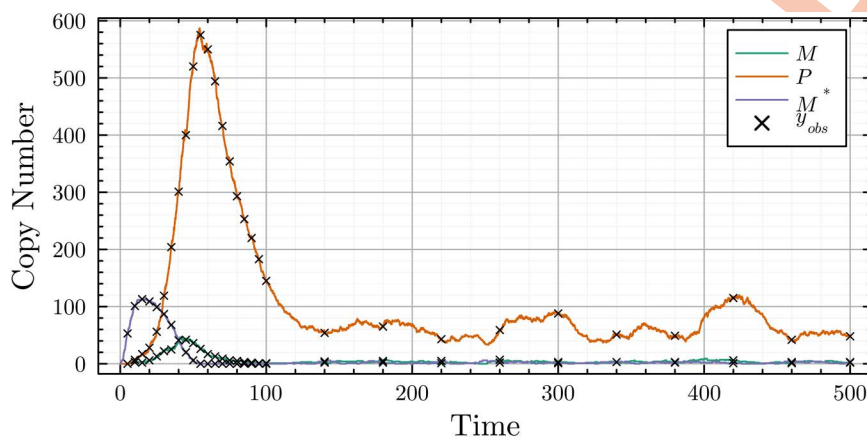


Fig 6. Synthetic data generated from an exact realisation of the gene transcription with delayed auto-inhibition model (Eqs. (29)–(30)). The full sample path for mRNA $M(t)$ (green line), proteins $P(t)$ (orange line), and incomplete transcriptions $M^*(t)$ (purple line) are shown along with the discrete observations contributing to the data y_{obs} (black crosses). Here the model parameters are $\beta_m = 10$, $\beta_{m^*} = 0.175$, $\beta_p = 1$, $\gamma_m = 0.08$, $\gamma_p = 0.05$, $\alpha = 2.5$, $M_a = 10$, $v = 0.5$, $h = 1.5$, and $P_a = 5$. The initial condition is $M(t_0) = P(t_0) = M^*(t_0) = 0$.

<https://doi.org/10.1371/journal.pcbi.1014687.g006>

$$\rho(y_{sim}, y_{obs}) = \left(\sum_{i=1}^n \|\mathbf{x}_{sim}(t_i) - \mathbf{x}_{obs}(t_i)\|_2^2 \right)^{1/2}$$

where $\|\cdot\|_2$ is the vector 2-norm. The acceptance threshold is taken as $\varepsilon = 1200$.

To implement the multifidelity ABC scheme we need to optimise the continuation probability function $\mu(\tilde{y}_{sim})$ given in Eq (13). Following the tuning process outlined by Prescott and Baker [54] we initially set $\mu_a = 1$ and $\mu_r = 1$ and preform a relatively small number of warm-up samples ($N=20,000$) to refine these continuation probabilities according to the optimality condition set out in S1 Appendix. We apply this tuning process for $\tau = 0.68$ and $\tau = 2.31$. We note that our exact coupling scheme (S2 Appendix) is essential for achieving high performance in multifidelity ABC.

We estimate posterior means using the direct ABC rejection sampling and multifidelity ABC sampling for a range of MC sample sizes. Denoting the estimate as $\hat{f}$, in each instance we estimate the MSE of this estimate, $\text{MSE}[\hat{f}]$, and the average computational cost (total runtime) to obtain the estimate, $C(\hat{f})$. Here the MSE is measured against a gold standard ABC estimate with $N = 1,000,000$, and the computational cost is the average runtime to ensure fair comparison as the total number of samples will be smaller for ABC accept-reject than the multifidelity approach of equivalent computational cost. Fig 7 demonstrates the empirical convergence of this variance as a function of computational cost for the mRNA maximum completion rate, β_m^* (Fig 7(a)), the protein production rate β_p (Fig 7(b)), and the transcription completion time shape parameter, α (Fig 7(c)). In each case, the measurements align with a rate of $\text{MSE}[\hat{f}] = \mathcal{O}(C(\hat{f})^{-1})$ as is expected from the theory in Prescott et al., [57] for a fixed acceptance threshold, $\varepsilon > 0$.

Across all parameters we observe a reduction in estimator MSE for equivalent levels of computation cost (that is, total runtime). The level of improvement depends on the value of τ , just as in the forward problem case. While $\tau = 0.68$ achieves a small improvement over $\tau = 2.31$, in both cases the speed-up factor is approximately $5\times$ over direct implementation of ABC. These results demonstrate the efficacy of our approximate stochastic simulation and exact coupling schemes can be exploited in to substantially accelerate simulation-based inference approaches though variance reduction of bias correction terms [52,54]. Beyond this demonstration of the posterior mean estimator, we can also demonstrate accelerated posterior probability density estimation. Fig 8 shows excellent agreement between the posterior marginal probability densities obtained using the MF-ABC samples used in Fig 7 (each with total cost $C(\hat{f}) \approx 10^5$ seconds) and expensive ABC gold standard (with total cost $C(\hat{f}) \approx 10^6$ seconds). This accuracy holds across all three unknown parameters, β_m^* , β_p , and α . While such results have been known and demonstrated before in the context of Markov processes [52,57,58], our methods extend the applicability of multifidelity methods to applications involving non-Markovian systems for the first time.

Discussion

In this work, we present a novel extension to the next reaction method to exactly simulate stochastic networks that are non-Markovian due to time dependent propensities, stochastic delays that depend on both time and system state [7,15]. This extension leads to a natural first order approximation of the Kurtz random time-change representation to result in an

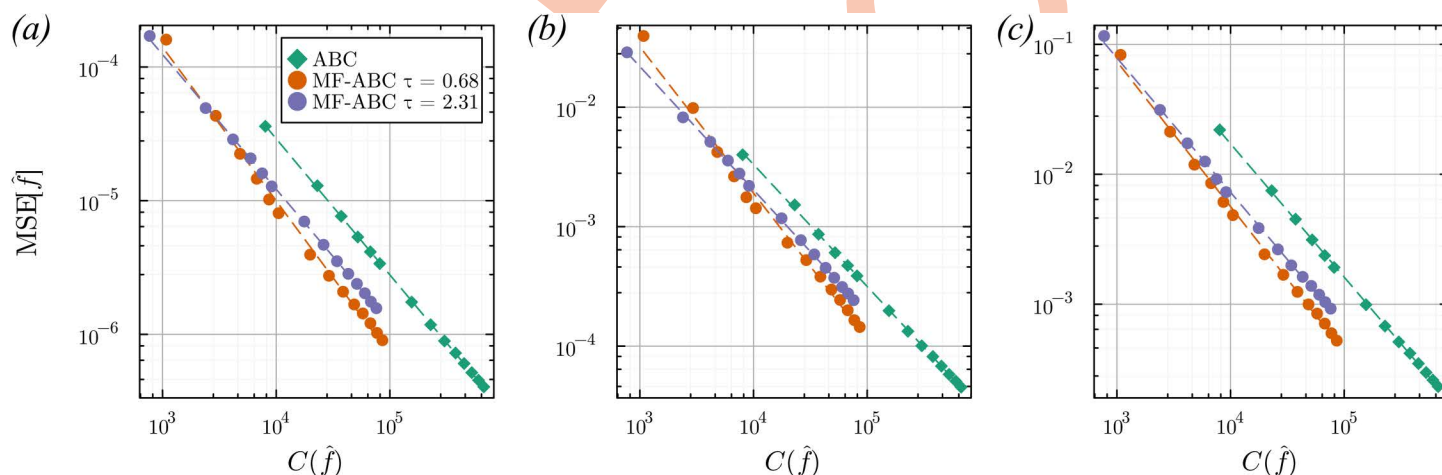


Fig 7. Comparison of convergence rate of posterior mean estimator MSE as a function of computational cost (that is, total runtime in seconds) for: (a) mRNA transcription initiation rate β_m^* ; (b) protein translation rate β_p ; (c) transcription completion time shape parameter α . Results are shown for standard ABC methods (green diamonds), MF-ABC with $\tau = 0.68$ (orange circles), and MF-ABC with $\tau = 2.31$ (purple squares). Regression lines are also shown (dashed lines). The ABC discrepancy threshold used is $\varepsilon = 1200$.

<https://doi.org/10.1371/journal.pcbi.1014687.g007>

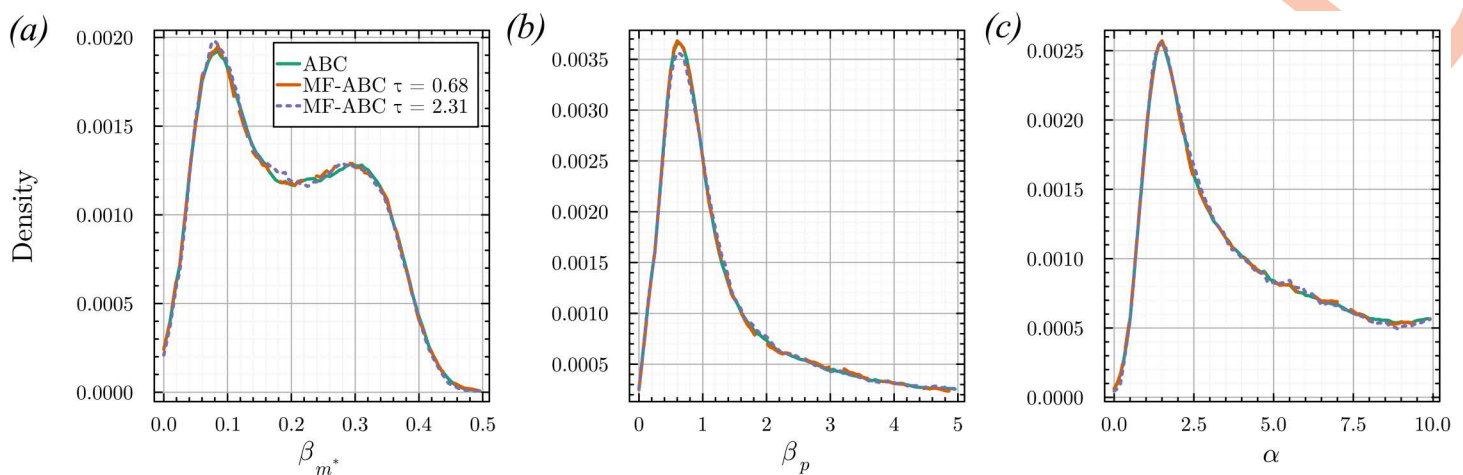


Fig 8. Comparison of posterior marginal probability density estimates obtained for: (a) mRNA transcription initiation rate β_m^* ; (b) protein translation rate β_p ; (c) transcription completion time shape parameter α . Results are shown for the gold standard ABC samples (green solid line), MF-ABC with $\tau = 0.68$ (orange dashed line), and MF-ABC with $\tau = 2.31$ (purple dotted line). The MF-ABC densities are obtained with a factor $10\times$ Computational speed-up. The ABC discrepancy threshold used is $\varepsilon = 1200$.

<https://doi.org/10.1371/journal.pcbi.1014687.g008>

equivalent non-Markovian τ -leaping method [6]. Finally, we develop an exact coupling scheme based on common Poisson clocks and demonstrate its applicability to variance reduction methods such as MLMC [49,56] and multifidelity methods [54,58].

While we focus on the original multifidelity ABC method of Prescott and Baker [54] and achieve orders of magnitude improvement in efficiency, it is important to highlight the generality of our approach to alternative methods in the simulation-based inference literature. In particular, any variant of multifidelity scheme that fundamentally relies on coupling of exact and approximation simulation pairs can be applied in our setting. This includes MLMC for ABC [51,52,58], multifidelity SMC for ABC [60] and adaptive extension [57], and state-of-the-art approaches for multifidelity neural posterior estimation [50,53]. In this work, our main contribution is the necessary stochastic simulation tools to enable MLMC and multifidelity approaches to simulation and inference in the non-Markovian setting.

In addition to parameter inference, our methods are also conducive to acceleration of summary statistic estimators for the forward problem following the Markovian work in networks [18,56,72] and stochastic differential equations [49, 55]. The potential for orders of magnitude improvement in simulation time could be of tremendous significance in the setting of whole cell modelling in quantitative biology that are typically extremely computationally expensive and often rely on stochastic simulation for key biochemical processes [76–79].

We primarily demonstrate our non-Markovian simulation schemes in the context of biochemical processes. This is largely due to the substantial amount of stochastic modelling and simulation literature that has arisen from the area of chemical physics [4–6,15,19,20,63]. However, our methods are not limited to this setting and are general to discrete-state continuous-time non-Markovian processes of the form given in Eq (8). As demonstrated by our epidemiology examples using MLMC simulation, our approach is widely applicable to alternative contexts involving non-Markovian stochastic models, such as ecology [80,81], physics [7,82] or queues [83,84].

Another area of applicability for this work is stochastic simulation following Markovian projection [85,86]. In the setting of reaction networks, Markovian projection can substantially reduce the dimensionality of the state vector while preserving the marginal path distribution for a selected subset of state dimensions [87]. This can be extremely beneficial in partially observed systems as it enables simulation to be performed directly on the observable chemical species without

considering any latent variables. The cost of the Markovian projection is that the reduced system can become non-Markovian [88]. Our methods will enable both exact and approximate simulation in this promising area of dimensionality reduction.

In the setting of optimal configuration of multifidelity simulation-based inference, whether it be based on ABC or alternatives, one open challenge is the automatic tuning of the step size parameter, τ . Importantly, the choice of τ does not impact bias due to the exact coupling scheme [18, 56], however, there is a trade-off between the computational gains of larger τ and the receiver operator characteristic of the low-fidelity simulator as a predictor for the high-fidelity simulator (S1 Appendix). While the coupling scheme enables optimal continuation probability tuning for a fixed τ [54,57], it does not help with adapting τ . One possible way forward could be to exploit the estimated strong convergence rate (Fig 4) to obtain an optimal scaling τ within the MLMC telescoping summation [58]. Other possible solutions could consider a randomised τ to generate a family of low-fidelity models and extending Eq (11) accordingly [57]. In practice a reasonable choice of τ can be obtained using a small numerical experiment using the warm-up step of MF-ABC that optimises the continuation probabilities for a range of candidate values for τ (e.g., see Warne et al., [58]), with smaller optimal continuation probabilities indicating more efficient sampling. This approach also provides some simple diagnostics that indicate when τ is too large to be effective (See Prescott and Baker [54] and Prescott et al., [57]). In particular, if one or both of the optimal continuation probabilities is close to unity for all candidate values for τ , then a finer approximation is required.

Conclusion

We provide a versatile extension of approximate stochastic simulation schemes with exact coupling mechanisms in the context of discrete-state continuous-time non-Markovian stochastic processes. We demonstrate the computational performance of our methods for both the generation of sample paths from the forward problem, and accelerated simulation-based inference for the inverse problem. We obtain an order of magnitude improvements in computational efficiency which could lead to new possibilities for the practical application of complex non-Markovian processes in quantitative biology and many other fields of computational science.

Supporting information

S1 Appendix. Coupling and continuation probabilities.
(PDF)

S2 Appendix. Coupling exact and approximate paths.
(PDF)

S3 Appendix. Inter-event time distribution.
(PDF)

Acknowledgments

Computational resources and services used in this work were provided by the eResearch Office, Queensland University of Technology, Brisbane, Australia.

Author contributions

Conceptualization: David J Warne.

Data curation: Thomas P Steele.

Formal analysis: Thomas P Steele, David J Warne.

Funding acquisition: David J Warne.

Investigation: Thomas P Steele, David J Warne.

Methodology: Thomas P Steele, David J Warne.

Project administration: David J Warne.

Software: Thomas P Steele, David J Warne.

Supervision: David J Warne.

Validation: Thomas P Steele, David J Warne.

Visualization: Thomas P Steele, David J Warne.

Writing – original draft: Thomas P Steele, David J Warne.

Writing – review & editing: Thomas P Steele, David J Warne.

References

1. Abkowitz JL, Catlin SN, Guttrop P. Evidence that hematopoiesis may be a stochastic process in vivo. *Nat Med.* 1996;2(2):190–7. <https://doi.org/10.1038/nm0296-190> PMID: 8574964
2. Cox DR, Miller HD. The theory of stochastic processes. Chapman & Hall/ CRC. 1965.
3. Kaern M, Elston TC, Blake WJ, Collins JJ. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet.* 2005;6(6):451–64. <https://doi.org/10.1038/nrg1615> PMID: 15883588
4. Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics.* 1976;22(4):403–34. [https://doi.org/10.1016/0021-9991\(76\)90041-3](https://doi.org/10.1016/0021-9991(76)90041-3)
5. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem.* 1977;81(25):2340–61. <https://doi.org/10.1021/j100540a008>
6. Gillespie DT. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics.* 2001;115(4):1716–33. <https://doi.org/10.1063/1.1378322>
7. Boguñá M, Lafuerza LF, Toral R, Serrano MÁ. Simulating non-Markovian stochastic processes. *Phys Rev E Stat Nonlin Soft Matter Phys.* 2014;90(4):042108. <https://doi.org/10.1103/PhysRevE.90.042108> PMID: 25375439
8. Browning R, Sulem D, Mengersen K, Rivoirard V, Rousseau J. Simple discrete-time self-exciting models can describe complex dynamic processes: A case study of COVID-19. *PLoS One.* 2021;16(4):e0250015. <https://doi.org/10.1371/journal.pone.0250015> PMID: 33836020
9. Cai X. Exact stochastic simulation of coupled chemical reactions with delays. *J Chem Phys.* 2007;126(12):124108. <https://doi.org/10.1063/1.2710253> PMID: 17411109
10. Fu X, Zhou X, Gu D, Cao Z, Grima R. DelaySSAToolkit.jl: stochastic simulation of reaction systems with time delays in Julia. *Bioinformatics.* 2022;38(17):4243–5. <https://doi.org/10.1093/bioinformatics/btac472>
11. Whitaker SA, Golightly A, Gillespie CS, Kypraios T. Sequential Bayesian Inference for Stochastic Epidemic Models of Cumulative Incidence. *Bayesian Anal.* 2025;1(1). <https://doi.org/10.1214/25-ba1547>
12. Yang Q, Reutens DC, Vegh V. Generalisation of continuous time random walk to anomalous diffusion MRI models with an age-related evaluation of human corpus callosum. *Neuroimage.* 2022;250:118903. <https://doi.org/10.1016/j.neuroimage.2022.118903> PMID: 35033674
13. Arkin A, Ross J, McAdams HH. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics.* 1998;149(4):1633–48. <https://doi.org/10.1093/genetics/149.4.1633> PMID: 9691025
14. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. *Science.* 2002;297(5584):1183–6. <https://doi.org/10.1126/science.1070919> PMID: 12183631
15. Anderson DF. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *J Chem Phys.* 2007;127(21):214107. <https://doi.org/10.1063/1.2799998> PMID: 18067349
16. Higham DJ. Modeling and Simulating Chemical Reactions. *SIAM Rev.* 2008;50(2):347–68. <https://doi.org/10.1137/060666457>
17. Schnoerr D, Sanguinetti G, Grima R. Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. *J Phys A: Math Theor.* 2017;50(9):093001. <https://doi.org/10.1088/1751-8121/aa54d9>
18. Warne DJ, Baker RE, Simpson MJ. Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art. *J R Soc Interface.* 2019;16(151):20180943. <https://doi.org/10.1098/rsif.2018.0943> PMID: 30958205
19. Gibson MA, Bruck J. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *The Journal of Physical Chemistry A.* 2000;104(9):1876–89. <https://doi.org/10.1021/jp993732q>

20. Cao Y, Li H, Petzold L. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *J Chem Phys.* 2004;121(9):4059–67. <https://doi.org/10.1063/1.1778376> PMID: [15332951](#)
21. Bratsun D, Volfson D, Tsimring LS, Hasty J. Delay-induced stochastic oscillations in gene regulation. *Proc Natl Acad Sci U S A.* 2005;102(41):14593–8. <https://doi.org/10.1073/pnas.0503858102> PMID: [16199522](#)
22. Brett T, Galla T. Stochastic processes with distributed delays: chemical Langevin equation and linear-noise approximation. *Phys Rev Lett.* 2013;110(25):250601. <https://doi.org/10.1103/PhysRevLett.110.250601> PMID: [23829723](#)
23. Meeussen JVV, Lenstra TL. Time will tell: comparing timescales to gain insight into transcriptional bursting. *Trends Genet.* 2024;40(2):160–74. <https://doi.org/10.1016/j.tig.2023.11.003> PMID: [38216391](#)
24. Miękisz J, Poleszczuk J, Bodnar M, Foryś U. Stochastic models of gene expression with delayed degradation. *Bull Math Biol.* 2011;73(9):2231–47. <https://doi.org/10.1007/s11538-010-9622-4> PMID: [21210242](#)
25. Bodnar M, Foryś U, Poleszczuk J. Analysis of biochemical reactions models with delays. *Journal of Mathematical Analysis and Applications.* 2011;376(1):74–83. <https://doi.org/10.1016/j.jmaa.2010.10.038>
26. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika.* Oxford University PressOxford. 2001. p. 331–43. <https://doi.org/10.1093/oso/9780198509936.003.0015>
27. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics.* 1953;21(6):1087–92. <https://doi.org/10.1063/1.1699114>
28. Del Moral P, Doucet A, Jasra A. Sequential Monte Carlo Samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 2006;68(3):411–36. <https://doi.org/10.1111/j.1467-9868.2006.00553.x>
29. Erban R, Chapman SJ, Kevrekidis IG, Vejchodský T. Analysis of a Stochastic Chemical System Close to a SNIPER Bifurcation of Its Mean-Field Model. *SIAM J Appl Math.* 2009;70(3):984–1016. <https://doi.org/10.1137/080731360>
30. Golightly A, Wilkinson DJ. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus.* 2011;1(6):807–20. <https://doi.org/10.1098/rsfs.2011.0047> PMID: [23226583](#)
31. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interface.* 2009;6(31):187–202. <https://doi.org/10.1098/rsif.2008.0172> PMID: [19205079](#)
32. Kanazawa K, Sornette D. Standard form of master equations for general non-Markovian jump processes: The Laplace-space embedding framework and asymptotic solution. *Physical Review Research.* 2024 Jun;6(2):023270. <https://doi.org/10.1103/physrevresearch.6.023270>
33. Cranmer K, Brehmer J, Louppe G. The frontier of simulation-based inference. *Proc Natl Acad Sci U S A.* 2020;117(48):30055–62. <https://doi.org/10.1073/pnas.1912789117> PMID: [32471948](#)
34. Sisson SA, Fan Y, Beaumont MA. Handbook of Approximate Bayesian Computation. Sisson SA, Fan Y, Beaumont MA, editors. Chapman and Hall/CRC. 2018. <https://doi.org/10.1201/9781315117195>
35. Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C. Approximate Bayesian computation. *PLoS Comput Biol.* 2013;9(1):e1002803. <https://doi.org/10.1371/journal.pcbi.1002803> PMID: [23341757](#)
36. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. *Genetics.* 2002;162(4):2025–35. <https://doi.org/10.1093/genetics/162.4.2025> PMID: [12524368](#)
37. Blum MGB. Approximate Bayesian Computation: A Nonparametric Perspective. *Journal of the American Statistical Association.* 2010;105(491):1178–87. <https://doi.org/10.1198/jasa.2010.tm09448>
38. Frazier DT, Martin GM, Robert CP, Rousseau J. Asymptotic properties of approximate Bayesian computation. *Biometrika.* 2018;105(3):593–607. <https://doi.org/10.1093/biomet/asv027>
39. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol.* 1999;16(12):1791–8. <https://doi.org/10.1093/oxfordjournals.molbev.a026091> PMID: [10605120](#)
40. Tavaré S, Balding DJ, Griffiths RC, Donnelly P. Inferring coalescence times from DNA sequence data. *Genetics.* 1997;145(2):505–18. <https://doi.org/10.1093/genetics/145.2.505>
41. Frazier DT, Nott DJ, Drovandi C, Kohn R. Bayesian Inference Using Synthetic Likelihood: Asymptotics and Adjustments. *Journal of the American Statistical Association.* 2022;118(544):2821–32. <https://doi.org/10.1080/01621459.2022.2086132>
42. Price LF, Drovandi CC, Lee A, Nott DJ. Bayesian Synthetic Likelihood. *Journal of Computational and Graphical Statistics.* 2017;27(1):1–11. <https://doi.org/10.1080/10618600.2017.1302882>
43. Papamakarios G, Sterratt D, Murray I. In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019. 837–48. <https://proceedings.mlr.press/v89/papamakarios19a.html>
44. Papamakarios G, Murray I. Fast <-free inference of simulation models with Bayesian conditional density estimation. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016. 1036–44.
45. Papamakarios G, Nalisnick E, Rezende DJ, Mohamed S, Lakshminarayanan B. Normalizing flows for probabilistic modeling and inference. *J Mach Learn Res.* 2021;22(1).
46. Lueckmann JM, Boelts J, Greenberg D, Goncalves P, Macke J. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021. 343–51. <https://proceedings.mlr.press/v130/lueckmann21a.html>

47. Wang X, Kelly RP, Warne DJ, Drovandi C. Preconditioned neural posterior estimation for likelihood-free inference. *Trans Mach Learn Res*. 2024.
48. Frazier DT, Kelly R, Drovandi C, Warne DJ. The Statistical Accuracy of Neural Posterior and Likelihood Estimation. In: 2024. <https://doi.org/10.48550/ARXIV.2411.12068>
49. Giles MB. Multilevel Monte Carlo Path Simulation. *Operations Research*. 2008;56(3):607–17. <https://doi.org/10.1287/opre.1070.0496>
50. Hikida Y, Bharti A, Jeffrey N, Briol FX. Multilevel neural simulation-based inference. 2025. <https://doi.org/10.48550/ARXIV.2506.06087>
51. Jasra A, Jo S, Nott D, Shoemaker C, Tempone R. Multilevel Monte Carlo in approximate Bayesian computation. *Stochastic Analysis and Applications*. 2019;37(3):346–60. <https://doi.org/10.1080/07362994.2019.1566006>
52. Warne DJ, Baker RE, Simpson MJ. Multilevel rejection sampling for approximate Bayesian computation. *Computational Statistics & Data Analysis*. 2018;124:71–86. <https://doi.org/10.1016/j.csda.2018.02.009>
53. Krouglova AN, Johnson HR, Confavreux B, Deistler M, Gonçalves PJ. Multifidelity Simulation-based Inference for Computationally Expensive Simulators. 2025. <https://doi.org/10.48550/ARXIV.2502.08416>
54. Prescott TP, Baker RE. Multifidelity Approximate Bayesian Computation. *SIAM/ASA J Uncertainty Quantification*. 2020;8(1):114–38. <https://doi.org/10.1137/18m1229742>
55. Rhee C-H, Glynn PW. Unbiased Estimation with Square Root Convergence for SDE Models. *Operations Research*. 2015;63(5):1026–43. <https://doi.org/10.1287/opre.2015.1404>
56. Anderson DF, Higham DJ. Multilevel Monte Carlo for Continuous Time Markov Chains, with Applications in Biochemical Kinetics. *Multiscale Model Simul*. 2012;10(1):146–79. <https://doi.org/10.1137/110840546>
57. Prescott TP, Warne DJ, Baker RE. Efficient multifidelity likelihood-free Bayesian inference with adaptive computational resource allocation. *J Comput Phys*. 2024;496:112577. <https://doi.org/10.1016/j.jcp.2023.112577> PMID: 40822115
58. Warne DJ, Prescott TP, Baker RE, Simpson MJ. Multifidelity multilevel Monte Carlo to accelerate approximate Bayesian parameter inference for partially observed stochastic processes. *Journal of Computational Physics*. 2022;469:111543. <https://doi.org/10.1016/j.jcp.2022.111543>
59. Voliotis M, Thomas P, Grima R, Bowsher CG. Stochastic Simulation of Biomolecular Networks in Dynamic Environments. *PLoS Comput Biol*. 2016;12(6):e1004923. <https://doi.org/10.1371/journal.pcbi.1004923> PMID: 27248512
60. Prescott TP, Baker RE. Multifidelity Approximate Bayesian Computation with Sequential Monte Carlo Parameter Sampling. *SIAM/ASA J Uncertainty Quantification*. 2021;9(2):788–817. <https://doi.org/10.1137/20m1316160>
61. Erban R, Chapman J, Maini P. A practical guide to stochastic simulations of reaction-diffusion processes. *arXiv*. 2007. <https://doi.org/10.48550/ARXIV.0704.1908>
62. Kurtz TG. The Relationship between Stochastic and Deterministic Models for Chemical Reactions. *The Journal of Chemical Physics*. 1972;57(7):2976–8. <https://doi.org/10.1063/1.1678692>
63. Gillespie DT. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*. 1992;188(1–3):404–25. [https://doi.org/10.1016/0378-4371\(92\)90283-v](https://doi.org/10.1016/0378-4371(92)90283-v)
64. Jahnke T, Huisinga W. Solving the chemical master equation for monomolecular reaction systems analytically. *J Math Biol*. 2007;54(1):1–26. <https://doi.org/10.1007/s00285-006-0034-x> PMID: 16953443
65. Péliissier A, Phan M, Le Bail D, Beerenwinkel N, Rodríguez Martínez M. Unifying non-Markovian dynamics and agent heterogeneity in scalable stochastic networks. *Nat Commun*. 2026;17(1):3345. <https://doi.org/10.1038/s41467-026-69817-y> PMID: 41771859
66. Barber S, Voss J, Webster M. The rate of convergence for approximate Bayesian computation. *Electron J Statist*. 2015;9(1). <https://doi.org/10.1214/15-ejs988>
67. Marjoram P, Molitor J, Plagnol V, Tavaré S. Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci U S A*. 2003;100(26):15324–8. <https://doi.org/10.1073/pnas.0306899100> PMID: 14663152
68. Sisson SA, Fan TMMY. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*. 2007;104(6):1760–5. <https://doi.org/10.1073/pnas.0607208104>
69. Drovandi CC, Pettitt AN. Estimation of parameters for macroparasite population evolution using approximate bayesian computation. *Biometrics*. 2011;67(1):225–33. <https://doi.org/10.1111/j.1541-0420.2010.01410.x> PMID: 20345496
70. Jiang Q, Fu X, Yan S, Li R, Du W, Cao Z, et al. Neural network aided approximation and parameter inference of non-Markovian models of gene expression. *Nat Commun*. 2021;12(1):2618. <https://doi.org/10.1038/s41467-021-22919-1> PMID: 33976195
71. Gillespie DT, Hellander A, Petzold LR. Perspective: Stochastic algorithms for chemical kinetics. *J Chem Phys*. 2013;138(17):170901. <https://doi.org/10.1063/1.4801941> PMID: 23656106
72. Lester C, Baker RE, Giles MB, Yates CA. Extending the Multi-level Method for the Simulation of Stochastic Biological Systems. *Bull Math Biol*. 2016;78(8):1640–77. <https://doi.org/10.1007/s11538-016-0178-9> PMID: 27515935
73. Jensen MH, Sneppen K, Tiana G. Sustained oscillations and time delays in gene expression of protein Hes1. *FEBS Lett*. 2003;541(1–3):176–7. [https://doi.org/10.1016/s0014-5793\(03\)00279-5](https://doi.org/10.1016/s0014-5793(03)00279-5) PMID: 12706840
74. Barrio M, Burrage K, Leier A, Tian T. Oscillatory regulation of Hes1: Discrete stochastic delay modelling and simulation. *PLoS Comput Biol*. 2006;2(9):e117. <https://doi.org/10.1371/journal.pcbi.0020117> PMID: 16965175

75. Hirata H, Yoshiura S, Ohtsuka T, Bessho Y, Harada T, Yoshikawa K, et al. Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop. *Science*. 2002;298(5594):840–3. <https://doi.org/10.1126/science.1074560> PMID: [12399594](https://pubmed.ncbi.nlm.nih.gov/12399594/)
76. Feig M, Sugita Y. Whole-Cell Models and Simulations in Molecular Detail. *Annual Review of Cell and Developmental Biology*. 2019;35(1):191–211. <https://doi.org/10.1146/annurev-cellbio-100617-062542>
77. Roberts E, Magis A, Ortiz JO, Baumeister W, Luthey-Schulten Z. Noise contributions in an inducible genetic switch: a whole-cell simulation study. *PLoS Comput Biol*. 2011;7(3):e1002010. <https://doi.org/10.1371/journal.pcbi.1002010> PMID: [21423716](https://pubmed.ncbi.nlm.nih.gov/21423716/)
78. Stumpf MPH. Statistical and computational challenges for whole cell modelling. *Current Opinion in Systems Biology*. 2021;26:58–63. <https://doi.org/10.1016/j.coisb.2021.04.005>
79. Yeom J-S, Georgouli K, Blake R, Navid A. Towards dynamic simulation of a whole cell model. In: *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2021. 1–10. <https://doi.org/10.1145/3459930.3471161>
80. Fleming CH, Calabrese JM, Mueller T, Olson KA, Leimgruber P, Fagan WF. Non-Markovian maximum likelihood estimation of autocorrelated movement processes. *Methods Ecol Evol*. 2014;5(5):462–72. <https://doi.org/10.1111/2041-210x.12176>
81. Williams BK. Optimal management of non-Markovian biological populations. *Ecological Modelling*. 2007;200(1–2):234–42. <https://doi.org/10.1016/j.ecolmodel.2006.07.039>
82. de Vega I, Alonso D. Dynamics of non-Markovian open quantum systems. *Rev Mod Phys*. 2017;89(1). <https://doi.org/10.1103/revmodphys.89.015001>
83. Karthikeyan K, Maragathasundari S, Kameswari M. Structural and queue characterization of a non-Markovian boundary value problem. *AIP Advances*. 2023;13(3). <https://doi.org/10.1063/5.0138431>
84. Maragathasundari S, Radha S. A study on the investigation of mathematical modelling in non markovian queue. In: *AIP Conference Proceedings*, 2019. 020042. <https://doi.org/10.1063/1.5135217>
85. Gyöngy I. Mimicking the one-dimensional marginal distributions of processes having an ito differential. *Probab Th Rel Fields*. 1986;71(4):501–16. <https://doi.org/10.1007/bf00699039>
86. Bayer C, Häppölä J, Tempone R. Implied stopping rules for American basket options from Markovian projection. *Quantitative Finance*. 2018;19(3):371–90. <https://doi.org/10.1080/14697688.2018.1481290>
87. Ben Hammouda C, Ben Rached N, Tempone R, Wiechert S. Automated importance sampling via optimal control for stochastic reaction networks: A Markovian projection–based approach. *Journal of Computational and Applied Mathematics*. 2024;446:115853. <https://doi.org/10.1016/j.cam.2024.115853>
88. Ben Hammouda C, Chupin M, Munker S, Tempone R. Filtered Markovian Projection: Dimensionality Reduction in Filtering for Stochastic Reaction Networks. 2025. <https://doi.org/10.48550/ARXIV.2502.07918>