Model reduction for the chemical master equation

Model Reduction for the Chemical Master Equation: an Information-Theoretic Approach

Kaan Öcal[1,2], Guido Sanguinetti[3], and Ramon Grima[2,†]

[1]School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK
[2]School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JH, UK
[3]Scuola Internazionale Superiore di Studi Avanzati, 34136 Trieste, Italy
[†]Corresponding author: ramon.grima@ed.ac.uk

Abstract

The complexity of mathematical models in biology has rendered model reduction an essential tool in the quantitative biologist's toolkit. For stochastic reaction networks described using the Chemical Master Equation, commonly used methods include time-scale separation, the Linear Mapping Approximation and state-space lumping. Despite the success of these techniques, they appear to be rather disparate and at present no general-purpose approach to model reduction for stochastic reaction networks is known. In this paper we show that most common model reduction approaches for the Chemical Master Equation can be seen as minimising a well-known information-theoretic quantity between the full model and its reduction, the Kullback-Leibler divergence defined on the space of trajectories. This allows us to recast the task of model reduction as a variational problem that can be tackled using standard numerical optimisation approaches. In addition we derive general expressions for the propensities of a reduced system that generalise those found using classical methods. We show that the Kullback-Leibler divergence is a useful metric to assess model discrepancy and to compare different model reduction techniques using three examples from the literature: an autoregulatory feedback loop, the Michaelis-Menten enzyme system and a genetic oscillator.

Keywords: Chemical Master Equation · Model Reduction · Systems Biology

1 Introduction

Stochastic biochemical reaction networks such as those involved in gene expression, immune response or cellular signalling [1–4] are often described using the Chemical Master Equation (CME). The CME describes the dynamics of biochemical processes on a mesoscopic level, viewing them as a discrete collection of molecules interacting and undergoing reactions stochastically; as such it is generally considered more accurate than continuum approximations such as rate equations and the Chemical Langevin Equation [4]. Despite its explanatory power, the CME poses significant analytical and computational difficulties to modellers that have limited its use in practice. Closed-form solutions to the CME are difficult to obtain and are only known for a small number of biologically relevant systems, and solving the CME numerically requires using approximations such as the Finite State Projection (FSP) [5]. Numerical approaches tend to scale poorly with the number of species and reactions present in a system, and as a result there is significant interest in finding ways to simplify a description of a stochastic reaction network that make it easier to analyse and study - this is the goal of model reduction.

Model reduction for deterministic and continuum-limit models in biology is an active research topic [6, 7], but very few existing methods can be applied to the discrete, stochastic setting of the CME. The Quasi-Steady State Approximation (QSSA) is perhaps the best known technique, first considered in the stochastic case in [8]. Here the system is partitioned into 'slow' and 'fast' species such that the fast species evolve very quickly on the timescale of the slow species. On the slow timescale the states of the fast species can therefore be approximated by their steady-state value (conditioned on the slow species), effectively allowing a description of the system in terms of the

slow species only. By its nature the QSSA is only applicable to systems with a clear separation of timescales between species, the existence of which cannot always be established. The QSSA for stochastic systems is generally believed to require more stringent conditions than in the deterministic case, but the exact validity conditions are not well-understood [9–14]. Recent work [15] has analysed a modification of the QSSA, the total QSSA (tQSSA), in systems involving reversible binding of molecules and shown that it has a wide range of applicability in the stochastic case.

Similar to the QSSA is the Quasiequilibrium Approximation (QEA), which was first considered in [16, 17] for stochastic reaction networks. Here the reaction network is decomposed into 'slow' and 'fast' *reactions*, and the fast reactions are assumed to equilibrate rapidly on the timescale of the slow reactions. Similar to the QSSA, the QEA can be used to reduce the number of species and reactions in a system, but it relies on the existence of a clear timescale separation between reactions, which is not always present for large systems with many distinct reactions. Much like the QSSA, the validity of the QEA for systems without the appropriate timescale separation has not been generally established, and from the asymptotic nature of the descriptions it is not usually possible to quantify the approximation error. Despite this, both the QSSA and the QEA are by far the most commonly used model-reduction technique for chemical reaction networks owing to their physical interpretability and analytical tractability, most famously in the Michaelis-Menten model of enzyme kinetics.

A distinct approach for model reduction with the Chemical Master Equation is state-space lumping, which originates from the theory of finite Markov chains, see e.g. [18]. Here different states in a system are lumped together such that the coarse-grained system is still Markovian and can be modelled using the CME. For a typical biochemical reaction network it may not be possible to perform any nontrivial lumping while preserving Markovianity, whence approximate lumping methods have been considered e.g. in [19–21]. Here the coarse-grained system is approximated by a Markov process, and the approximation error quantified using the KL divergence between the original model and a lift of the approximation to the state space of the original model. State-space lumping for the CME often occurs in the context of the QEA, as states related by fast reactions are effectively lumped together, or averaged [22–24]. For this reason we will not consider this approach separately, although many of our considerations, such as the optimal form of the lumped propensity functions, extend to state-space lumping.

Finally, a more recent model reduction technique specifically for gene expression systems is the Linear Mapping Approximation (LMA) [25]. The LMA replaces bimolecular reactions of the form $G + P \to (\ldots)$, where G is a binary species such as a gene, by a linear approximation where P is replaced by its mean conditioned on $[G] = 1$. While the LMA does not reduce the number of species or reactions, reaction networks with linear reactions are generally easier to analyse: their moments can be computed exactly, and closed-form solutions are known for many cases [26–29].

At a first glance these approaches - timescale separation, state space lumping and the Linear Mapping Approximation - seem rather disparate, and it is unclear what, if any, relationships exist between them. In this paper we show that all of these methods can be viewed as minimising a natural information theoretic quantity, the Kullback-Leibler (KL) divergence, between the full and reduced models. In particular they can be seen as maximal likelihood approximations of the full model, and one can assess the quality of the approximation in terms of the resulting KL divergence. Based on these results we show how the KL divergence can be estimated and minimised numerically, therefore providing an automated method to choose between different candidate reductions of a given model in situations where none of the above model reduction techniques are classically applicable. In its full generality, the Chemical Master Equation describing a reduced model can be obtained by marginalisation of the original CME, and hence our approach recovers the method based on condi-

tional expectations presented in [30].

The KL divergences we consider in this paper are computed on the space of trajectories, and as such include both static information and dynamical information, in contrast to purely distribution-matching approaches. The KL divergence and similar information-theoretic measures between continuous-time Markov chains have previously been considered in [31, 32] in the context of variational inference (with the true model and the approximation reversed compared to our approach), in [33, 34] to obtain approximate non-Markovian reductions, in [35] to analyse information flow for stochastic reaction networks and in [36] in quantifying model discrepancy for Markovian agent-based models.

In Section 2 we introduce the mathematical framework in which we consider model reduction for the Chemical Master Equation, based on KL divergences between continuous-time Markov chains. We show how the KL divergence can be minimised analytically in some important cases, recovering standard results in the literature and providing a mathematical justification for commonly used mean-field arguments as in the QSSA, the QEA and the LMA. We furthermore provide numerical algorithms for estimating as well as minimising the KL divergence in cases where analytical solutions are not available. In Section 3 we illustrate the use of KL divergences as a metric for approximation quality using three biologically relevant examples: an autoregulatory feedback loop exhibiting critical behaviour, Michaelis-Menten enzyme kinematics, where we reanalyse the QSSA and the QEA, and an oscillatory genetic feedback system taken from [11], for which we compare different reductions using our approach. Finally in Section 4 we discuss our observations and how our approach could be used as a stepping-stone towards automated reduction of complex biochemical reaction pathways.

2 Methods

2.1 Stochastic Reaction Networks

The Chemical Master Equation describes a biochemical reaction network as a stochastic process on a discrete state space $\mathcal{X}$. We will use the letter q to denote such a stochastic process, which for the purposes of this paper can be seen as a probability distribution over *trajectories* on $\mathcal{X}$. For a biochemical reaction network the state space will be $\mathcal{X} = \mathbb{N}^s$, where s is the number of species in the system: every state consist of a tuple $\boldsymbol{n} = (n_1, \ldots, n_s)$ of s integers describing the abundances of each species.

Since the model q can consist of many species interacting in complicated ways, we often want to find a reduction that is more tractable, yet approximates q as closely as possible. The reduced model, which we will call p, should be of the same form as q, i.e. described by the CME, but will typically involve fewer species and simpler reactions. In particular p can be defined on a lower-dimensional state space $\tilde{\mathcal{X}}$. A state $\boldsymbol{n}$ in the original model can then be described by its projection $\tilde{\boldsymbol{n}}$ onto this lower-dimensional space, together with some unobserved components, which we will denote $\boldsymbol{z}$. See Fig. 1**a** and **b** for an illustration.

In this paper we assume that the basic structure of p is known *a priori*, i.e. the species and reactions we wish to retain are fixed. Our approach to model reduction therefore consists in finding the optimal propensity functions for the reduced model, and we shall see how this can give rise to various known approximations such as the QSSA, the QEA or the LMA depending on what reductions are performed. We will return to the related problem of choosing the structure of the reduced model p in the discussion.

The original model q defines a probability distribution over trajectories in $\mathcal{X}$, and projecting each trajectory onto the chosen reduced state space we get the (exact) projection of q onto this space, which we denote $\tilde{q}$ (Fig. 1**a**). This is a stochastic process on $\tilde{\mathcal{X}}$ that is generally not Markovian and

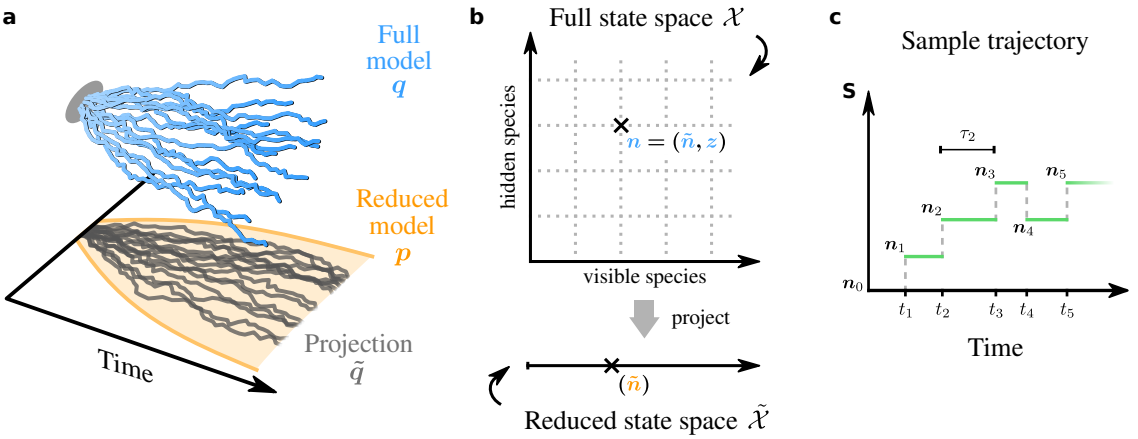

Figure 1: Model reduction for the Chemical Master Equation. **(a)** Model reduction approximates a high-dimensional model q by a lower-dimensional version. Since the direct projection $\tilde{q}$ of the full model is not easy to describe, we approximate it using a family of tractable candidate models: in this paper, the approximation p is described by the CME. **(b)** Comparison of the full state space of a system, consisting of two species S_1 and S_2, and a reduced state space containing S_1 only. Species which are not deemed essential can be discarded in the reduction and become unobserved variables. The dynamics of the original system involves all species, whereas the reduced model aims at an effective description only in terms of the reduced species. **(c)** Sample trajectory for a one-dimensional system, defined by the sequence $\boldsymbol{n}_0, \boldsymbol{n}_1, \ldots$ of states visited and the corresponding jump times $t_1, t_2, \ldots$ (or alternatively the waiting times $\tau_0, \tau_1, \ldots$).

thus cannot be modelled using the CME. We aim to find a tractable approximation p to $\tilde{q}$ that can be described using the CME, and we will do this by minimising the KL divergence $\mathrm{KL}(\tilde{q} \,\|\, p)$ between the two models on the space of trajectories. Several well-known examples of model reduction for the CME are illustrated in Fig. 2.

Jumps in q come in two kinds: those that affect the observed species $\tilde{\boldsymbol{n}}$, which we will call visible jumps, and those that only change z, which we call hidden. The jumps in $\tilde{q}$ correspond to visible jumps in q. In the context of the CME, jumps are typically grouped into reactions with fixed stoichiometry, often also called reaction channels, and we can similarly distinguish visible and hidden reactions in q. We will always assume that different reactions have different stoichiometries, so that every jump in q and p corresponds to a unique reaction. Reactions with the same stoichiometry can always be combined by summing their propensities.

We introduce some more notation at this point, which is summarised in Table 1 and illustrated in Fig. 1**c**. A single realisation, or trajectory, of q is defined by the sequence of states $\boldsymbol{n}_0, \boldsymbol{n}_1, \boldsymbol{n}_2, \ldots \in \mathcal{X}$ visited and jump times $0 < t_1 < t_2 < \ldots$. We will write $\boldsymbol{n}_{[0,T]} = \{\boldsymbol{n}(t)\}_{0 \le t \le T}$ for a trajectory, where $\boldsymbol{n}(0) = \boldsymbol{n}_0$ and $\boldsymbol{n}(t) = \boldsymbol{n}_i$ for $t_i \le t < t_{i+1}$, and denote by $\tau_i = t_{i+1} - t_i$ the waiting times between jumps.

For a continuous-time Markov process p, e.g. one defined using the CME, we denote the transition rate from state $\boldsymbol{n}$ to $\boldsymbol{m} \neq \boldsymbol{n}$ by $p_{\boldsymbol{m}\leftarrow\boldsymbol{n}}$. We let $p_{\leftarrow\boldsymbol{n}} = \sum_{\boldsymbol{m}\neq\boldsymbol{n}} p_{\boldsymbol{m}\leftarrow\boldsymbol{n}}$ be the total transition rate out of $\boldsymbol{n}$. The transition probabilities at $\boldsymbol{n}$ are then given by $p_{\boldsymbol{m}|\boldsymbol{n}} = p_{\boldsymbol{m}\leftarrow\boldsymbol{n}} / p_{\leftarrow\boldsymbol{n}}$. For completeness we define $p_{\boldsymbol{n}|\boldsymbol{n}} := 0$.

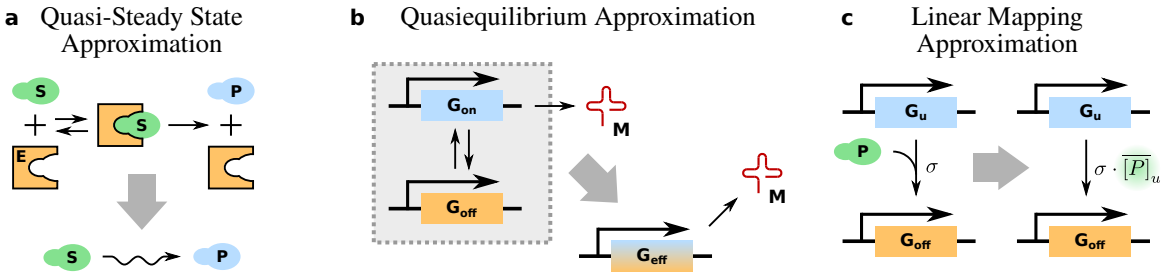

Figure 2: Common model reduction techniques for the Chemical Master Equation. **(a)** The QSSA eliminates intermediate species which evolve on a faster timescale than others, replacing them by their steady-state values. In the case of the pictured Michaelis-Menten enzyme system this is often applied to the enzyme E and the substrate-enzyme complex ES. **(b)** The QEA is an analogue of the QSSA that can be applied when a reaction and its reverse equilibrate rapidly on timescales of interest. This can occur e.g. when a gene switches rapidly between states. **(c)** The LMA replaces protein-gene binding reactions by effective unimolecular reactions, where the concentration of proteins is approximated by its mean conditioned on the gene state. While this does not remove any species from the system, it considerably simplifies the propensities of the reactions.

2.2 Minimising the KL Divergence

The Kullback-Leibler Divergence between two distributions $\tilde{q}$ and p is defined as

$$\mathrm{KL}(\tilde{q} \,\|\, p) = \int \tilde{q}(x) \log \frac{\tilde{q}(x)}{p(x)} \mathrm{d}x = \mathbb{E}_{\tilde{q}} \left[\log \tilde{q}(x)\right] - \mathbb{E}_{\tilde{q}} \left[\log p(x)\right] \tag{1}$$

$$= \mathbb{E}_{\tilde{q}} \left[\log \tilde{q}\right] + H(\tilde{q}; p) \tag{2}$$

where the quantity $H(\tilde{q}; p)$ is known as the cross-entropy:

$$H(\tilde{q}; p) = -\mathbb{E}_{\tilde{q}} \left[\log p(x)\right] \tag{3}$$

Minimising the KL divergence with respect to p is therefore equivalent to minimising the cross-entropy. Alternatively, it is equivalent to maximising the average log-likelihood under p of samples drawn from $\tilde{q}$, which is a statistical inference problem.

In this paper samples from $\tilde{q}$ and p are trajectories. We therefore need to compute the log-likelihood of an arbitrary trajectory $\boldsymbol{n}_{[0,T]}$ under p, which by assumption is a stochastic reaction network described by the CME. In Appendix A we show that the log-likelihood for a trajectory

Symbol	Explanation	Symbol	Explanation
q	Full model	$q_{\boldsymbol{m}\leftarrow\boldsymbol{n}}$	Transition rate
$\tilde{q}$	Projected model	$q_{\leftarrow\boldsymbol{n}}$	Total transition rate
p	Reduced model	$q_{\boldsymbol{m}\,|\,\boldsymbol{n}}$	Transition probability
$\boldsymbol{n}$	Full state vector	$q_0(\boldsymbol{n})$	Initial distribution
$\tilde{\boldsymbol{n}}$	Reduced state vector	$q_t(\boldsymbol{n})$	Single-time marginal
z	Unobserved state vector		

Table 1: Notation used in this paper.

visiting the states $\boldsymbol{n}_0, \boldsymbol{n}_1, \ldots, \boldsymbol{n}_k$ with waiting times $\tau_0, \tau_1, \ldots, \tau_k$ is given by

$$\log p(\boldsymbol{n}_{[0,T]}) = \log p_0(\boldsymbol{n}_0) - \sum_{i=0}^{k} \tau_i \, p_{\leftarrow \boldsymbol{n}_i} + \sum_{i=1}^{k} \log p_{\boldsymbol{n}_i \leftarrow \boldsymbol{n}_{i-1}}. \tag{4}$$

Note that computing the likelihood of a fully observed trajectory is easier than computing the likelihood of a trajectory sampled at discrete time points, which is the form of data typically observed in biological experiments. The difference is that computing the likelihood in the latter case requires integrating over all possible courses of the trajectory between observations; this integral is computed implicitly by the Chemical Master Equation, which is generally hard to solve. For a fully observed trajectory there are no hidden variables to integrate out, which considerably simplifies the task of computing likelihoods.

Returning to the problem of minimizing the KL divergence (1), or equivalently the cross entropy, we need to compute expectations over $\tilde{q}$ that are not in general available in closed form. It is easier to approximate these expectations by simulating N trajectories $\tilde{\boldsymbol{n}}^{(i)}_{[0,T]}$, $i = 1, \ldots, N$, from $\tilde{q}$ and computing the estimate

$$\widehat{H}(\tilde{q}; p_\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log p(\tilde{\boldsymbol{n}}^{(i)}_{[0,T]}). \tag{5}$$

We can then minimise this with respect to the reduced model parameters by gradient descent, noting that the cross-entropy is a convex function of p. Formulæ for the gradients can be computed by differentiating Eq. (4) by hand, or by using automatic differentiation software. Minimising the estimated cross-entropy via gradient descent yields a probabilistic estimate for the optimal parameters, which will generally converge to the true solution as $N \to \infty$. Summarising the contents of this section we arrive at Algorithm 1 for automatically fitting the optimally reduced model.

Algorithm 1 Model reduction via gradient descent.

Inputs: number of simulations N, simulation length T, full model q, model family $\{p_\theta\}$, learning rate η
Output: reduced model parameters $\widehat{\theta}$

for all $i = 1, \ldots, N$ **do**
  sample $\boldsymbol{n}^{(i)}_{[0,T]}$ **from** q
  project $\boldsymbol{n}^{(i)}_{[0,T]}$ **to** $\tilde{\boldsymbol{n}}^{(i)}_{[0,T]}$
end
initialize parameters $\hat{\theta}$
while not converged **do**
  compute $\widehat{H}(\tilde{q}; p_{\hat{\theta}})$ **using** (4), (5)
  compute $\nabla_\theta \widehat{H}(\tilde{q}; p_{\hat{\theta}})$
  $\hat{\theta} \mathrel{-}= \eta \nabla_\theta \widehat{H}(\tilde{q}; p_{\hat{\theta}})$
return $\hat{\theta}$

This algorithm has the advantage of being completely general, but unlike standard model reduction algorithms it involves numerical optimisation. In the next section we will analyse our loss function more closely to obtain alternative expressions for the optimum, allowing us to bypass Alg. 1 for a wide range of problems.

Instead of minimising $\mathrm{KL}(\tilde{q} \,\|\, p)$, an alternative approach would minimise the opposite KL divergence $\mathrm{KL}(p \,\|\, \tilde{q})$. The latter has been analysed e.g. in [31, 32] for continuous-time Markov processes. Despite the almost symmetric definition, it is well-known that these two quantities behave rather differently in an optimisation context and lead to very different results, cf. [37]. From a pragmatic point of view, minimising $\mathrm{KL}(\tilde{q} \,\|\, p)$ involves computing likelihoods under p, while minimising $\mathrm{KL}(p \,\|\, q)$ involves computing probabilities of trajectories under $\tilde{q}$, which is significantly more difficult.

Example: Telegraph Model

We will illustrate the above by means of a well-known example, the telegraph model of gene transcription [38], simplified to neglect degradation and depicted in Fig. 3. The full model q consists of three species: a gene in the on state (G_{on}) and the off state (G_{off}), together with mRNA (M). The gene switches between both states with constant activation rate σ_{on} and inactivation rate σ_{off}, and in the active state produces mRNA with rate ρ_{on}. The reduced model p consists only of mRNA (M), which is produced at constant rate ρ_{eff}. This instructive example has previously been treated in [33] in a similar context.

A state in the full model can be represented as $\boldsymbol{n} = (g, m)$, where g is the gene state (on or off) and the m is the number of mRNA present, while a reduced state $\tilde{\boldsymbol{n}} = (m)$ is given by the number of mRNA molecules only, with the unobserved component $\boldsymbol{z} = (g)$ being the gene state. The only reaction in q that descends to $\tilde{q}$ is mRNA production, whereas the two gene switching reactions are not observed. Note that the projection $\tilde{q}$ of the telegraph model is not Markovian, as the instantaneous mRNA production rate depends on the current gene state g, but we can sample from $\tilde{q}$ by simulating q and discarding the information about the gene state.

The reduced model p is a Poisson process whose only parameter is the production rate ρ_{eff}. The only trajectories possible under p are those starting at $m = 0$ and increasing by one at every jump. The log-likelihood of such a trajectory $\tilde{\boldsymbol{n}}_{[0,T]}$ under p is given by

$$\log p(\tilde{\boldsymbol{n}}_{[0,T]}) = \tilde{\boldsymbol{n}}(T) \log \rho_{\mathrm{eff}} - \rho_{\mathrm{eff}} T, \tag{6}$$

where $\tilde{\boldsymbol{n}}(T)$ is the total number of mRNA produced up to time T, as can be verified using Eq. (4). The log-likelihood of any other trajectory is $-\infty$.

Note that Eq. (6) describes the likelihood for an entire trajectory, as opposed to the likelihood of observing $\boldsymbol{n}_T$ molecules at time T. The latter follows a Poisson distribution with rate $\rho_{\mathrm{eff}} T$ and can be obtained by integrating Eq. (4) over all trajectories that end at the same $\boldsymbol{n}_T$. Indeed, conditioned on observing $\boldsymbol{n}_T$ mRNA molecules at time T their individual production times are uniformly and independently distributed on $[0, T]$ by the properties of the Poisson process, and integrating Eq. (6) over all possible combinations of production times we recover the usual Poisson likelihood, keeping in mind that any permutation of the production times yields the same trajectory.

The reduced model p can model every trajectory obtained from $\tilde{q}$, so the log-likelihood of any sample from $\tilde{q}$ is finite. The mRNA transcription reactions in both models correspond. If q were to include e.g. mRNA degradation as is usual in the literature, some trajectories from the full model would feature decreases in $\tilde{\boldsymbol{n}}$ and be impossible under p; in this case the KL divergence would infinite, which signals that the reduced model p is not appropriate and needs to be extended to include degradation.

For this simple example the likelihood of a reduced trajectory under p only depends on the total mRNA produced. We can therefore compute the cross-entropy explicitly:

$$H(\tilde{q}; p) = -\mathbb{E}_{\tilde{q}} \left[\tilde{\boldsymbol{n}}(T)\right] \log \rho_{\mathrm{eff}} + \rho_{\mathrm{eff}} T. \tag{7}$$

Minimising this with respect to ρ_{eff} yields the optimum

$$\rho_{\text{eff}}^{*} = \frac{1}{T}\,\mathbb{E}_{\tilde{q}}\left[\tilde{\boldsymbol{n}}(T)\right], \tag{8}$$

which is the average mRNA production rate on the interval $[0, T]$. Thus the optimal approximation to the telegraph model is obtained by setting the mRNA production rate to its mean value. This is the result we obtain by applying the QEA to this system (see 2.5.2), which suggests that the approximation will be better if σ_{on} and σ_{off} are large compared to ρ_{on}. We will see how to evaluate the approximation quality in 2.4.

Even if the analytical minimisation above were not possible, the above minimisation can be performed using Alg. 1. To do so we start by sampling a fixed number N of mRNA trajectories from the telegraph model; the more trajectories we use, the better our estimate of ρ_{eff} will be. The next step is constructing the cross-entropy loss function for the reduced model, which only depends on one parameter ρ_{eff} and is given by Eq. (7). We then perform gradient descent on the cross-entropy, averaged over all trajectories, to arrive at our estimate of ρ_{eff}. Here the learning rate η in Alg. 1 has to be determined by experimentation. The code accompanying this paper contains an example implementation of this procedure in Julia.

2.3 Analysing the KL Divergence

In the last section we have derived a numerical procedure to minimise the KL divergence (1), or equivalently the cross-entropy (3). While this yields a direct computational procedure to fit reduced models, we are equally interested in the theoretical insights to be gained from considering model reduction as a variational problem. Understanding how our approach works in more detail will help us establish links with existing model reduction techniques and derive general principles that can help us generalise these to new systems.

As shown in Appendix B, for a Markovian reduced model p the cross-entropy can be written as

$$H(\tilde{q}; p)_{[0,T]} = H(\tilde{q}_0; p_0) + \int_0^T \sum_{\tilde{\boldsymbol{n}}} \tilde{q}_t(\tilde{\boldsymbol{n}}) \left(p_{\leftarrow\tilde{\boldsymbol{n}}}(t) - \sum_{\tilde{\boldsymbol{m}}\neq\tilde{\boldsymbol{n}}} \tilde{q}_{\tilde{\boldsymbol{m}}\leftarrow\tilde{\boldsymbol{n}}}(t) \log p_{\tilde{\boldsymbol{m}}\leftarrow\tilde{\boldsymbol{n}}}(t) \right) \mathrm{d}t \tag{9}$$

$$= H(\tilde{q}_0; p_0) + \int_0^T H(\tilde{q}; p)_t \,\mathrm{d}t, \tag{10}$$

with the instantaneous cross-entropy rate at time t defined as

$$H(\tilde{q}; p)_t = -\sum_{\tilde{\boldsymbol{n}}} \tilde{q}_t(\tilde{\boldsymbol{n}}) \left(\sum_{\tilde{\boldsymbol{m}}\neq\tilde{\boldsymbol{n}}} \tilde{q}_{\tilde{\boldsymbol{m}}\leftarrow\tilde{\boldsymbol{n}}}(t) \log p_{\tilde{\boldsymbol{m}}\leftarrow\tilde{\boldsymbol{n}}}(t) - p_{\leftarrow\tilde{\boldsymbol{n}}}(t) \right). \tag{11}$$

Here we use the effective transition rates

$$\tilde{q}_{\tilde{\boldsymbol{m}}\leftarrow\tilde{\boldsymbol{n}}}(t) = \lim_{\delta t\to 0} \frac{1}{\delta t} P(\tilde{q}(t+\delta t) = \tilde{\boldsymbol{m}} \,|\, \tilde{q}(t) = \tilde{\boldsymbol{n}}, q_0) \qquad (\tilde{\boldsymbol{m}} \neq \tilde{\boldsymbol{n}}). \tag{12}$$

If $\tilde{q}$ is not Markovian, the transition probability from a state $\tilde{\boldsymbol{n}}$ at time t will be affected by the history of the process, and hence on the initial distribution q_0. If $\tilde{q}$ is Markovian, Eq. (12) reduces to the classical transition rate which is independent of q_0.

Looking at Eq. (9) we make two important observations. As intimated in the discussion of the telegraph example of the previous section, if any trajectory under $\tilde{q}$ has jumps which are not allowed under p, i.e. if $\tilde{q}_{\tilde{\boldsymbol{m}}\leftarrow\tilde{\boldsymbol{n}}} \neq 0$ while $p_{\tilde{\boldsymbol{m}}\leftarrow\tilde{\boldsymbol{n}}} = 0$, then the cross-entropy (9) will be ∞: the reduced model p

is not flexible enough to model $\tilde{q}$. On the other hand, if p contains transitions which are impossible under $\tilde{q}$, i.e. if $p_{\tilde{\boldsymbol{m}}\leftarrow\tilde{\boldsymbol{n}}} \neq 0$ while $\tilde{q}_{\tilde{\boldsymbol{m}}\leftarrow\tilde{\boldsymbol{n}}} = 0$, then these can only increase the cross-entropy (9). This means that the optimal reduced model p does not contain transitions beyond those in $\tilde{q}$, and in the context of the CME we can therefore assume that p and $\tilde{q}$ have the same reactions, with different propensities (Markovian for p, not necessarily for $\tilde{q}$).

If the i-th reaction in p has propensity function $\rho_i(\tilde{\boldsymbol{n}}; t)$ and net stoichiometry $\boldsymbol{s}_i$ we obtain the following decomposition of the cross-entropy rate at time t:

$$H(\tilde{q}; p)_t = -\sum_i \sum_{\tilde{\boldsymbol{n}}} \tilde{q}_t(\tilde{\boldsymbol{n}}) \left[\tilde{q}_{\tilde{\boldsymbol{n}}+\tilde{\boldsymbol{s}}_i\leftarrow\tilde{\boldsymbol{n}}} \log \rho_i(\tilde{\boldsymbol{n}}; t) - \rho_i(\tilde{\boldsymbol{n}}; t)\right], \tag{13}$$

where the first sum is over all reactions in p, or equivalently all visible reactions in q. The total cross-entropy is obtained by integrating Eq. (13) over $[0, T]$, and we can find the optimal p by optimising the cross-entropy for each reaction separately.

We can minimize Eq. (13) analytically if the full model q is Markovian. Assume there is precisely one reaction in q with net stoichiometry $\tilde{\boldsymbol{s}}_i$ in the projection, and let σ_i be its propensity function. Differentiating the above equation with respect to $\rho_i(\tilde{\boldsymbol{n}}; t)$ and setting the derivative to zero we obtain

$$\rho_i^*(\tilde{\boldsymbol{n}}; t) = \sum_z q_t(z \,|\, \tilde{\boldsymbol{n}})\, \sigma_i(\tilde{\boldsymbol{n}}, z; t) = \mathbb{E}_z[\sigma_i(\tilde{\boldsymbol{n}}, z; t) \,|\, \tilde{\boldsymbol{n}}; t]\,. \tag{14}$$

The optimal propensity for a reaction under p is the expected propensity under the original model conditioned on the observed state $\tilde{\boldsymbol{n}}$. In particular, if the propensity of the original reaction does not depend on unobserved species it can be taken over directly. We explore the implications of these results on the moments of the reduced system in Section 2.5.4.

In practice we often place constraints on the reduced propensities $\rho_i(\tilde{\boldsymbol{n}}; t)$ such as time-homogeneity and mass-action kinetics, which result in constrained optima. For example, if ρ_i is taken to be independent of time we have to integrate Eq. (13) over $[0, T]$, and as $T \to \infty$ it can be verified that the total cross-entropy is minimised for

$$\rho_i^*(\tilde{\boldsymbol{n}}) = \sum_z q_\infty(z \,|\, \tilde{\boldsymbol{n}})\, \sigma_i(\tilde{\boldsymbol{n}}, z) = \mathbb{E}_z[\sigma_i(\tilde{\boldsymbol{n}}, z) \,|\, \tilde{\boldsymbol{n}}; t = \infty]\,, \tag{15}$$

which is the steady-state version of Eq. (14).

For other forms of propensity functions the optimum can generally be obtained by a similar averaging procedure. In all cases the main difficulties lie in estimating the relevant conditional expectations, which can be done numerically using the Monte-Carlo approach presented in the previous section. Eq. (14) is perhaps the central result of this paper, and we will see that it coincides with the propensities obtained using the QSSA, the QEA and the LMA in when these are applied. These methods can therefore all be seen as special cases of our variational approach based on minimising the KL divergence. As an aside, our results provide a new derivation of the marginal CME proposed in [30, 34] and the nontrivial result that Eq. (14) and Alg. 1 target equivalent objectives.

In particular Alg. 1 can be bypassed by computing certain conditional moments of a system, which provides an alternative way to perform model reduction. For many systems these conditional moments can be efficiently computed using moment equations, possibly involving moment closure approximations, which have attracted a large amount of literature [4]. The LMA, for example, involves a bootstrapping approach to estimate the relevant moments. Alternatively the conditional expectations could be computed by numerically integrating the CME, although we do not expect this to scale well to large systems. In consequence Alg. 1, while a convenient way to implement our approach in full generality and empirically verify its performance, can often be bypassed.

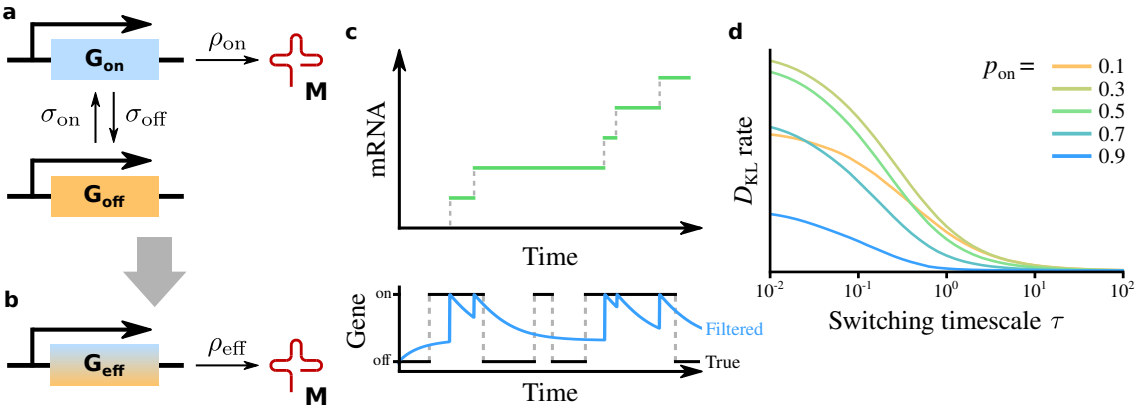

Figure 3: Model reduction illustrated using the telegraph model without degradation. **(a)** Schematic of the model. **(b)** The reduced model is equivalent to a Poisson process with rate ρ_{eff}. **(c)** Example trajectory from the telegraph model, showing mRNA numbers (top) and the gene state (bottom) in time. If we only observe mRNA numbers we can infer the current gene state by filtering, obtaining the probability for the gene being on at a given time (blue line). When mRNA numbers increase the on probability jumps to 1 since transcription only happens in that state. **(d)** KL divergence rate between the telegraph model and its Poisson reduction, for various choices of the switching timescale $\tau = \sigma_{\text{on}} + \sigma_{\text{off}}$ and on probability $p_{\text{on}} = \sigma_{\text{on}}/(\sigma_{\text{on}} + \sigma_{\text{off}})$, assuming a fixed transcription rate $\rho_{\text{on}} = 1$. The Poisson approximation becomes more accurate as the switching timescale increases compared to the transcription timescale. KL divergences were estimated numerically using Alg. 2.

Example: Telegraph Model, cont.

Consider the telegraph model from the previous section. If we make the reduced model p more flexible by allowing a nonlinear and time-dependent propensity $\rho(m;t) = \rho_{\text{eff}}(m;t)$, then the cross-entropy between $\tilde{q}$ and p can be computed as

$$H(\tilde{q};p) = \int_0^T \sum_m \tilde{q}_t(m)\left(\tilde{q}_{m+1\leftarrow m}(t)\log\rho_{\text{eff}}(m;t) - \rho_{\text{eff}}(m;t)\right)\mathrm{d}t. \tag{16}$$

Note that both $\tilde{q}$ and p have the same initial conditions. The effective propensities of $\tilde{q}$ can be computed as

$$\tilde{q}_{m+1\leftarrow m}(t) = \rho_{\text{on}}\, q_t(g = \text{on}\,|\,m), \tag{17}$$

which is ρ_{on} weighed by the probability that the gene is on at time t given that there are m mRNA molecules present. This is the optimal propensity for the reduced model p in accordance with Eq. (14), featuring a nonlinear dependence on m, and this can be shown to be the optimal Markovian approximation to $\tilde{q}$. We will see in Section 2.5.4 that this approximation exactly matches the mRNA distribution of the full model at all times.

If we require mass-action kinetics for the reduced model, i.e. if we let $\rho(m;t) = \rho_{\text{eff}}(t)$ for an effective rate constant $\rho_{\text{eff}}(t)$, minimising the cross-entropy (16) yields

$$\rho_{\text{eff}}(t) = \rho_{\text{on}}\, q_t(g = \text{on}). \tag{18}$$

The effective transcription rate learned by p is the expected transcription rate of q at time t, this time averaged over all m.

Finally if we require $\rho(m;t) = \rho_{\text{eff}}$ to be time-independent we obtain by a similar procedure

$$\rho_{\text{eff}} = \rho_{\text{on}} \cdot \frac{1}{T} \int_0^T q_t(g = \text{on})\, \mathrm{d}t, \tag{19}$$

which is just the previous result averaged over t. This is the special case considered in the previous section, where p was restricted to be a Poisson process with constant intensity. As $T \to \infty$, the optimum (19) converges to the expected transcription rate of q at steady state.

It can be checked that the propensity functions (17) and (18) exactly preserve the mean mRNA numbers of the full system. Similarly, the time-independent propensity function (19) preserves the mean mRNA numbers at steady state. For mass-action propensities, however, the variances will be underestimated. This is because the reduced models do not model gene switching, which increases the noise in the system. This is a common occurrence with model reduction and is well-known for the telegraph model, since mRNA numbers are always distributed following a Poisson mixture distribution, which will always have a higher variance than a single Poisson distribution with the same mean [39]. We will further explore how well reduced models capture marginal distributions of the original in Section 2.5.4.

2.4 Computing KL Divergences

To this point we have been occupied with minimising the KL divergence (1), or rather the cross-entropy (3), with respect to p. Perhaps counter-intuitively, minimising the KL divergence is easier than computing it, since the entropy

$$H(\tilde{q})_{[0,T]} = -\mathbb{E}_{\tilde{\boldsymbol{n}}_{[0,T]}}[\log q(\tilde{\boldsymbol{n}}_{[0,T]})]\,, \tag{20}$$

can be ignored during optimisation. In order to assess the performance of the reduced model obtained this way we want to compute the full KL divergence, and hence the entropy, explicitly. As for the cross-entropy the integral in Eq. (20) is generally intractable, and we will approximate the entropy by simulating N samples $\tilde{\boldsymbol{n}}^{(i)}_{[0,T]}$ from $\tilde{q}$ and computing

$$\widehat{H}(\tilde{q}) = -\frac{1}{N} \sum_{i=1}^{N} \log \tilde{q}(\tilde{\boldsymbol{n}}^{(i)}_{[0,T]}). \tag{21}$$

Here we face another difficulty, for if $\tilde{q}$ is not Markovian we cannot use Eq. (4) to compute log-likelihood of a trajectory $\tilde{\boldsymbol{n}}_{[0,T]}$. Since by assumption q is a Markov process, the projection $\tilde{q}$ is just a partially observed Markov process and we can use the so-called forward algorithm for Hidden Markov Models [40] to compute the log-likelihood of a trajectory.

The forward algorithm computes the joint probability $q(\tilde{\boldsymbol{n}}_{[0,t]}, z_t)$ for the observed trajectory up to time t and the current state of the hidden species, z_t. As shown in Appendix C, the forward algorithm in this case yields the following set of jump ODEs:

$$q(\tilde{\boldsymbol{n}}_0, z_0) = q_0(\tilde{\boldsymbol{n}}_0, z_0), \tag{22}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} q(\tilde{\boldsymbol{n}}_{[0,t]}, z_t) = &\sum_{z' \neq z_t} \left(q_{(\tilde{\boldsymbol{n}}_t, z_t) \leftarrow (\tilde{\boldsymbol{n}}_t, z')}\, q(\tilde{\boldsymbol{n}}_{[0,t]}, z') - q_{(\tilde{\boldsymbol{n}}_t, z') \leftarrow (\tilde{\boldsymbol{n}}_t, z_t)}\, q(\tilde{\boldsymbol{n}}_{[0,t]}, z_t) \right) \\ &- \sum_i \sigma_i(\tilde{\boldsymbol{n}}_t, z_t)\, q(\tilde{\boldsymbol{n}}_{[0,t]}, z_t),\end{aligned} \tag{23}$$

$$\lim_{t \searrow t_k} q(\tilde{\boldsymbol{n}}_{[0,t]}, z_t = z) = \lim_{t \nearrow t_k} \left(\sigma_{j_k}(\tilde{\boldsymbol{n}}_k, \tilde{z}_t = z)\, q(\tilde{\boldsymbol{n}}_{[0,t]}, z') \right), \tag{24}$$

with jumps at the jump times of the observed trajectory $\tilde{\boldsymbol{n}}_{[0,T]}$. The first sum in Eq. (23) represents hidden reactions (those that do not affect $\tilde{\boldsymbol{n}}$), and the second sum is over all visible reactions in q, with propensity functions given by the σ_i. The marginal likelihood of the observed trajectory can then be computed by summing over the unobserved variables z

$$\tilde{q}(\tilde{\boldsymbol{n}}_{[0,T]}) = \sum_z q(\tilde{\boldsymbol{n}}_{[0,T]}, z). \tag{25}$$

The above equations take the form of a modified CME where the observed variables are fixed. They can be solved using an adaptation of the Finite State Projection algorithm [5], and in cases where the unobserved state space is finite, exact solutions can be computed numerically. We summarise the procedure to estimate the KL divergence defined in Eq. (1) in Algorithm 2.

Algorithm 2 Computing the projected KL divergence between a full and a reduced model.

Inputs: number of simulations N, simulation length T, full model q, reduced model p
Output: KL divergence estimate $\widehat{\mathrm{KL}}(\tilde{q} \,\|\, p)$

$\hat{L} := 0$
for all $i = 1, \ldots, N$ **do**
  sample $\boldsymbol{n}_{[0,T]}$ **from** q
  project $\boldsymbol{n}_{[0,T]}$ **to** $\tilde{\boldsymbol{n}}_{[0,T]}$
  compute $p(\tilde{\boldsymbol{n}}_{[0,T]})$ **using** (4)
  compute $\tilde{q}(\tilde{\boldsymbol{n}}_{[0,T]})$ **using** (22)–(24), (25).
  $\hat{L}$ += $\tilde{q}(\tilde{\boldsymbol{n}}_{[0,T]}) - \tilde{p}(\tilde{\boldsymbol{n}}_{[0,T]})$
return $\hat{L}/N$

As further shown in Appendix C, a corollary of the above equations is the following jump ODE for the marginal distribution itself:

$$\log \tilde{q}(\tilde{\boldsymbol{n}}_0) = \log \tilde{q}_0(\tilde{\boldsymbol{n}}(0)), \tag{26}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} \log \tilde{q}(\tilde{\boldsymbol{n}}_{[0,t]}) = -\sum_i \mathbb{E}_{z_t}[\sigma_i(\tilde{\boldsymbol{n}}(t), z_t) \,|\, \tilde{\boldsymbol{n}}_{[0,t]}, q_0; t], \tag{27}$$

$$\lim_{t \searrow t_k} \log \tilde{q}(\tilde{\boldsymbol{n}}_{[0,t]}) = \lim_{t \nearrow t_k} \left(\log \tilde{q}(\tilde{\boldsymbol{n}}_{[0,t]}) + \log \mathbb{E}_{z_t}\left[\sigma_{j_k}(\tilde{\boldsymbol{n}}_k, \tilde{z}_t) \,|\, \tilde{\boldsymbol{n}}_{[0,t]}, q_0; t\right]\right). \tag{28}$$

These equations generalise Eqs. (4) to the presence of unobserved species z. These equations can be found in various places in the literature, e.g. [33, 35, 41–44] although we are not aware of a consistent nomenclature for this modification of the CME.

The equations for the log-likelihood are linear and can be split into a sum of contributions for each reaction in $\tilde{q}$. Integrating over all reduced trajectories $\tilde{\boldsymbol{n}}_{[0,T]}$ shows that the entropy $H(\tilde{q})$ can thus be expressed as a sum of contributions from each reaction in $\tilde{q}$. Eq. (13) shows that the same decomposition holds for the cross-entropy $H(\tilde{q}; p)$, so entire KL divergence $\mathrm{KL}(\tilde{q} \,\|\, p)$ can be decomposed into contributions coming from the individual reactions, which allows us to differentially analyse the accuracy of the reduced model p for each reaction separately. We note that all reactions together determine the distribution q on trajectories over which we integrate, so this decomposition is not strict in practice.

Example: Telegraph Model (cont. 2)

We can solve the filtering problem for the telegraph model exactly as the unobserved state space is 2-dimensional. Let $F(i; t) = q(g(t) = i, m_{[0,t]})$. The telegraph model consists of three reactions, the

hidden switching reactions with parameters σ_{on} and σ_{off}, and the visible transcription reaction ρ_{on}. Specialising Eq. (23) to this system we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\text{on}; t) = -(\rho_{\text{on}} + \sigma_{\text{off}})\, F(\text{on}; t) + \sigma_{\text{on}}\, F(\text{off}; t), \tag{29}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\text{off}; t) = \sigma_{\text{off}}\, F(\text{on}; t) - \sigma_{\text{on}}\, F(\text{off}; t), \tag{30}$$

where $m_{[0,t]}$ denotes the observed mRNA trajectory. At each mRNA production event we update the probabilities according to Eq. (24), which in this case reads:

$$\lim_{t \searrow t_k} F(\text{on}; t) = \rho_{\text{on}} \cdot \lim_{t \nearrow t_k} F(\text{on}; t), \tag{31}$$

$$\lim_{t \searrow t_k} F(\text{off}; t) = 0. \tag{32}$$

The marginal likelihood of the observed mRNA trajectory is then given by

$$q(m_{[0,t]}) = F(\text{on}; t) + F(\text{off}; t). \tag{33}$$

These derivations can also be found in [33].

As a byproduct the above equations yield the conditional distribution over the gene state at each time t given the trajectory prior to that time point. These are the filtered distributions and illustrated in Fig. 3**c**, but we only need the marginal likelihood for our purposes.

Using the above equations to compute the log-likelihood for a large number of trajectories sampled from the telegraph model q yields a numerical approximation of the entropy $H(\tilde{q})$, and together with Eq. (6) we obtain an estimate of the KL divergence $\mathrm{KL}(\tilde{q} \,\|\, p)$. For time-homogeneous propensities we empirically found the KL divergence to be asymptotically proportional to T, and we call the proportionality factor the (steady-state) KL divergence rate. Note that the cross-entropy rate can be defined directly using Eq. (11).

In Fig. 3**d** we show how this KL divergence rate between the telegraph model and its Poisson approximation changes for various choices of σ_{on} and σ_{off}. We observe that the KL divergence rate tends to 0 as the switching rates increase; indeed the Poisson approximation can be obtained from the QEA for this example. We further note that the KL divergence is maximal when the gene spends substantial amounts of time in each state; for p_{on} close to 1 the gene is almost always active and the system approaches constitutive expression with rate ρ, while for p_{on} close to 0 the gene only activates sporadically and resembles a constitutive gene with a very low expression rate $\rho \cdot p_{\text{on}}$. Visually inspecting mRNA trajectories generated in both regimes show that the results become harder to distinguish from the Poisson process.

Our results on the telegraph model without degradation apply almost verbatim when degradation is included: adding the reaction $M \longrightarrow \emptyset$ to the full and reduced models does not affect the KL divergences considered. *A priori* adding degradation changes the probability distribution over trajectories, which can now exhibit decreasing mRNA numbers, but the log-likelihood contributed by the degradation reaction is the same between the telegraph model and its reduction, and therefore the difference cancels in Eqs. (26)–(28) and Eq. (13). Since the log-likelihood contributed by the gene switching and transcription reactions does not depend on mRNA numbers, which do change in the presence of degradation, the total KL divergence is unaffected. In the presence of feedback, degradation would indirectly affect the KL divergence via its effect on mRNA numbers; we refer the interested reader to the recent paper [35] for an analysis of this type of information flow in stochastic biochemical reaction networks. We can readily verify that our observations on the mean and variance of mRNA numbers under the telegraph model and its reduction remain valid in the presence of degradation.

2.5 Relationship with Known Approaches

2.5.1 The Quasi-Steady State Approximation

In deterministic chemical kinetics the Quasi-Steady State Approximation (QSSA) is a model reduction technique that can be applied when a system can be partitioned into so-called slow species $\boldsymbol{n}^s$ and fast species $\boldsymbol{n}^f$ such that the fast species $\boldsymbol{n}^f$ evolve on a much faster time scale than the slow species. On the timescale on which the slow species evolve, the fast species can therefore be assumed to reach their steady state almost instantaneously. Thus one assumes $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{n}^f = 0$, which allows for simplification of the remaining equations for the slow species. The QSSA has famously been applied to Michaelis-Menten enzyme kinetics with the enzyme-substrate complex ES as the fast species, resulting in the classical Michaelis-Menten propensity for product formation (see 3.2 for more information).

The QSSA was extended to the stochastic case in [8]. Here one assumes that conditioned on $\boldsymbol{n}^s$, the fast species reach steady state nearly instantaneously compared to the timescale of interest. In our formulation the reduced state space consists of the slow species, $\tilde{\boldsymbol{n}} = \boldsymbol{n}^s$, and the unobserved species are the fast species, $z = \boldsymbol{n}^f$. As shown in [8] the QSSA yields the following reduced CME for the approximation p (Eqs. (10), (11) in [8]):

$$\frac{\mathrm{d}}{\mathrm{d}t}p_t(\boldsymbol{n}^s) = \sum_{i=1}^{r}\left[\tilde{\rho}_i(\boldsymbol{n}^s - \boldsymbol{S}_i^s; t)\, p_t(\boldsymbol{n}^s - \boldsymbol{S}_i^s) - \tilde{\rho}_i(\boldsymbol{n}^s; t)\, p_t(\boldsymbol{n}^s)\right], \tag{34}$$

where the reduced propensities $\tilde{\rho}_i$ are defined as

$$\tilde{\rho}_i(\boldsymbol{n}^s; t) = \mathbb{E}_{\boldsymbol{n}^f}\left[\rho_i(\boldsymbol{n}^s, \boldsymbol{n}^f) \,|\, \boldsymbol{n}^s; t\right]. \tag{35}$$

This agrees precisely with Eq. (14), which illustrates how the QSSA can be seen as minimising the KL divergence (B4).We remark that Eq. (34) is a special case of the phenomenon discussed in Section 2.5.4. We can compare (35) with the true propensities of the projection $\tilde{q}$, which depend on the entire history of the observed trajectory and the initial distribution q_0:

$$\rho_i^{\text{exact}}(\boldsymbol{n}; t \,|\, q_0, \boldsymbol{n}^s_{[0,t]}) = \mathbb{E}_{\boldsymbol{n}^f}[\rho_i(\boldsymbol{n}^s, \boldsymbol{n}^f) \,|\, \boldsymbol{n}^s_{[0,t]}, q_0; t]. \tag{36}$$

This equation, which describes the instantaneous reaction rate of the non-Markovian process $\tilde{q}$, has previously been considered e.g. in [8, 33].

Some intuition for the relationship between timescale separation and our approach can be gained from equations Eqs. (26)–(28), which describe the probability of a reduced trajectory $\tilde{\boldsymbol{n}}_{[0,T]}$ under $\tilde{q}$. Reaction propensities under q in general depend on the unobserved species, whose distribution is correlated with the history of the current trajectory. If we assume that the timescale of the unobserved species $\boldsymbol{n}^f$ is very fast, then this correlation will decay very quickly and the time-dependence of the conditional distributions in Eqs. (26)–(28) will be negligible, so we can replace these equations by

$$\log \tilde{q}(\tilde{\boldsymbol{n}}_0) = \log \tilde{q}_0(\tilde{\boldsymbol{n}}(0)), \tag{37}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\log \tilde{q}(\tilde{\boldsymbol{n}}_{[0,t]}) = -\sum_i \mathbb{E}_{z_t}[\sigma_i(\tilde{\boldsymbol{n}}(t), z_t) \,|\, \tilde{\boldsymbol{n}}_t], \tag{38}$$

$$\lim_{t \searrow t_k} \log \tilde{q}(\tilde{\boldsymbol{n}}_{[0,t]}) = \lim_{t \nearrow t_k}\left(\log \tilde{q}(\tilde{\boldsymbol{n}}_{[0,t]}) + \log \mathbb{E}_{z_t}\left[\sigma_{j_k}(\tilde{\boldsymbol{n}}_k, \tilde{z}_t) \,|\, \tilde{\boldsymbol{n}}_t\right]\right). \tag{39}$$

which is another way of showing that $\tilde{q}$ has propensities given by (35).

2.5.2 The Quasiequilibrium Approximation

While the QSSA relies on a reaction network being divisible into slow and fast species that evolve on two different timescales, in practice it is more frequently the case that some *reactions* in a system will be fast and that some will be slow. The Quasiequilibrium Approximation (QEA) developed in [17, 45, 46] modifies the QSSA to model this scenario by reformulating the system using extents $\boldsymbol{a} = (a_1, \ldots, a_r)$, where the extent a_i is defined the number of times reaction i has occurred.

The extents themselves form a Markov chain, and the state of a system can be obtained from its extents as $\boldsymbol{n}(t) = \boldsymbol{n}_0 + \boldsymbol{S}\boldsymbol{a}(t)$, where $\boldsymbol{S}$ is the stoichiometry matrix. As derived e.g. in [17], the CME of this system is

$$\frac{\mathrm{d}}{\mathrm{d}t} q_t(\boldsymbol{a}) = \sum_i \left[\rho_i(\boldsymbol{S}\boldsymbol{a} - \boldsymbol{S}_i)\, q_t(\boldsymbol{a} - \boldsymbol{e}_i) + \rho_i(\boldsymbol{S}\boldsymbol{a})\, q_t(\boldsymbol{a})\right]. \tag{40}$$

Here the sum is over reactions and $\boldsymbol{e}_i$ is the vector with a 1 in the i-th position and 0 elsewhere. The QEA assumes that the extent vector can be divided into slow and fast components $\boldsymbol{a}^s$ and $\boldsymbol{a}^f$, respectively, corresponding to slow and fast reactions, and we define our reduced system to consist only of the slow reactions. From here we can proceed analogously to the QSSA discussed above and obtain that the reduced propensities are given by

$$\tilde{\rho}_i(\boldsymbol{a}^s; t) = \mathbb{E}_{\boldsymbol{a}^f}\left[\rho_i(\boldsymbol{S}\boldsymbol{a}) \,|\, \boldsymbol{a}^s; t\right], \tag{41}$$

if the fast reactions can be assumed to equilibrate instantaneously.

A subtlety of this argument is the fact that the fast extents $\boldsymbol{a}^f$ can only increase in time and therefore will not admit a steady-state distribution in general. The conditional means, however, will converge in many cases, e.g. if $\boldsymbol{a}^f$ consists of both directions of a reversible reaction. If this is the case the QEA yields a well-defined reduction, which moreover agrees with (14) and therefore minimises the KL divergence to the full model on the space of extents.

2.5.3 The Linear Mapping Approximation

The Linear Mapping Approximation (LMA) [25] replaces a bimolecular reaction of the form $G + X \xrightarrow{\sigma} G^*$, where G is a binary species representing a gene state, by a reaction $G \xrightarrow{\overline{\sigma}} G^*$ with effective propensity $\overline{\sigma}$. Assuming mass action kinetics, the propensity function of the bimolecular reaction is $\rho(\boldsymbol{n}) = gx\sigma$, where g is the state of G and x the abundance of the species X, and the propensity of the linearised version is $\tilde{\rho}(\boldsymbol{n}) = \overline{\sigma} g$ for an appropriate choice of $\overline{\sigma}$. Taking the derivative with respect to $\overline{\sigma}$ of the cross-entropy rate (13) at time t yields

$$\frac{\partial H(\tilde{q}; p)_t}{\partial \overline{\sigma}} = \sum_{\boldsymbol{n}} q_t(\boldsymbol{n}) \left(g - \frac{\sigma g x}{\overline{\sigma}}\right), \tag{42}$$

which vanishes if and only if

$$\overline{\sigma} = \sigma \cdot \mathbb{E}\left[x \,|\, g = 1; t\right]. \tag{43}$$

This provides a mathematical justification for the mean-field assumption underlying [25]. The LMA approximates this conditional expectation by imposing a self-consistency condition on the linearised system, thereby deriving an effective approximation to (43).

As an aside we can compute the KL divergence rate between q and the optimal reduction p given by (43) analytically to obtain

$$\mathrm{KL}(q \,\|\, p)_t = \sigma \cdot P(g = 1) \cdot \left(\mathbb{E}\left[x \log x \,|\, g = 1\right] - \mathbb{E}\left[x \,|\, g = 1\right] \mathbb{E}\left[\log x \,|\, g = 1\right]\right). \tag{44}$$

Here the dependence of the expectations on the time t are suppressed. This expression for the discrepancy incurred by the LMA, which resembles the definition of the variance, quantifies the intuition that the LMA should be accurate if fluctuations of X in the unbound gene state are small.

The above derivation is not entirely rigorous as the linearisation has a different net stoichiometry than the original reaction and KL divergence between q and p is infinite. To remedy this we can either neglect fluctuations in X due to binding in the original model or consider the KL divergence on a reaction-by-reaction basis as discussed in 2.4, since the two reactions correspond despite their different net stoichiometries.

If fluctuations in X are absent or can be neglected, the optimal reduction using Eq. (43) preserves first-order moments for all species, assuming that all other reactions are linear. Indeed, replacing the reaction $G \xrightarrow{\sigma[X]} G^*$ by its linearisation only changes the moment equation for $\mathbb{E}[g]$. For the full model, the latter is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[g;t] = \sigma\,\mathbb{E}[gx;t] + \dots, \tag{45}$$

while the reduction has

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[g;t] = \sigma\,\mathbb{E}[x\,|\,g=1;t]\,\mathbb{E}[g;t] + \dots, \tag{46}$$

which is equivalent for a binary species g (the remaining terms only depend on species means by our linearity assumption). Second and higher-order moments will in generally differ between the two as we will see. This sheds light on a more subtle aspect of the LMA: the self-consistent approach in [25] involves computing the second-order moment $\mathbb{E}[gx;t]$, which can differ in the reduction. For this reason the original approach, while generally quite accurate, can lead to slightly suboptimal propensities as we shall see in Section 3.1.

2.5.4 Moment-Matching

In Appendix D we show that the optimal reduced model p, defined using the propensity functions (14), exactly preserves the marginal distributions of the full model. In particular, if we place no constraints on the reduced propensity functions the resulting system p will have the same means, variances and higher-order moments for the observed species as the full system q. Thus our approach is also related to moment-matching, in particular to the proposed moment-based methods in [11, 47].

The fact that the reduced model p can exactly capture the probability distribution of the projected model $\tilde{q}$ at any time t does not mean that the two are equivalent. The difference can be seen on the level of individual trajectories. The reduced model p is always memoryless, whereas this is not usually the case for $\tilde{q}$, resulting in different dynamical dynamical properties such as autocorrelations. This ability to discriminate models on a trajectory-level illustrates an advantage of our path-based approach over purely moment-based model reduction.

We emphasise that exact preservation of moments only holds for the globally optimal propensity functions defined by Eq. (14), which can depend on $\tilde{n}$ and t in a complicated way. In practice one often use parametric propensity functions, such as mass-action propensities or Hill functions, and as a consequence this feature is lost. As with the LMA or the telegraph model it may still be possible to preserve some moments (such as means) exactly with the right choice of propensity function. Our results imply that discrepancies in the moments between the full and reduced models result from the fact that Eq. (14) is only computed approximately for methods like the QSSA or the QEA.

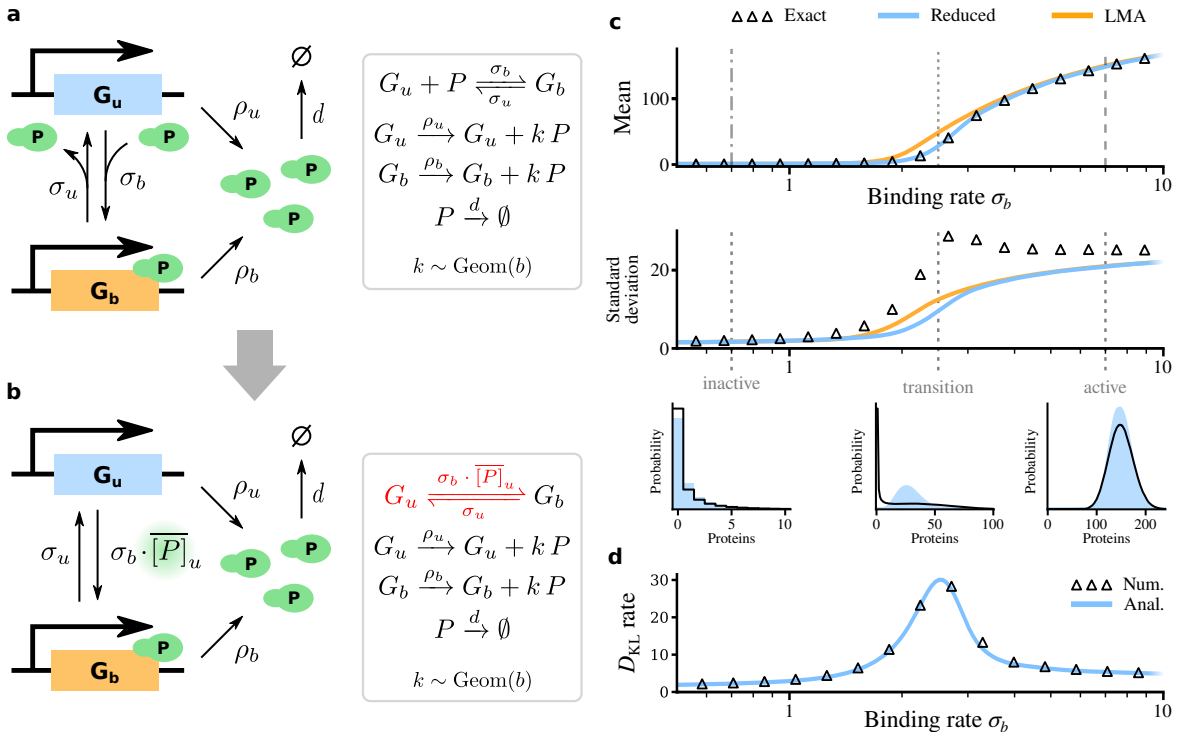

Figure 4: Reduction of an autoregulatory feedback loop using the Linear Mapping Approximation. **(a)** Schematic of the full reaction network. The number $k \geq 0$ of proteins produced at each reaction is geometrically distributed with mean b. **(b)** Reduced form of the same reaction system, where the protein-gene binding reaction is replaced by a mean-field approximation. **(c)** For the chosen parameter values the system exhibits critical behaviour around $\sigma_b = 2.5$, transitioning from a mostly unbound to a mostly bound state. Near the transition protein fluctuations increase and the mean-field assumption in the LMA breaks down. A comparison of steady-state moments and distributions for the full model, the optimal reduction and the LMA shows that the effective reaction rate computed by the latter is generally close to optimal. **(d)** KL divergence rate at steady state between the full model and the reduced version, computed analytically using Eq. (44) and using Monte Carlo simulations. The peak around the transition matches the observed discrepancy between the full and the reduced model. The remaining model parameters are $\sigma_u = 400$, $\rho_u = 0.3$, $\rho_b = 105$, $b = 2$, $d = 1$.

3 Numerical Experiments

3.1 Autoregulatory Feedback Loop

In this section we analyse a simple model of stochastic gene expression featuring positive autoregulation (see Fig. 4**a**). The system in question consists of a single gene found in two states, bound (G_b) and unbound (G_u), as well as the coded protein P which can bind to the gene to increase its own transcription rate. For simplicity we do not consider mRNA dynamics explicitly, instead modelling protein production as occurring in geometrically distributed bursts (see [48] for a derivation). The linearised version of the system, using the optimal propensity derived in Section 2.5.3, is shown in Fig. 4**b**. In Fig. 4**c** we compare the steady-state distributions of the full model with its linearisation, where the effective binding rate is computed numerically using Alg. 1, and the LMA, where the binding rate is approximated using the self-consistent approach in [25].

This system exhibits critical behaviour for some parameter values (see Fig. 4**c**), and it was shown

in [49] that computing the moments of this system is difficult near points of criticality as most moment closure techniques as well as the LMA yield inaccurate results. Our results show that for all parameters the self-consistent equations of the LMA provides results close to the numerically computed optimum. As discussed in Section 2.5.3, the exact reduction very closely reproduces mean protein numbers, with the LMA incurring a small bias near the critical point. In contrast, neither reduction is able to capture the increased fluctuations near the critical point, significantly underestimating the variance in protein numbers. Comparing the protein distributions for the full and the optimally reduced model shows a large discrepancy near the critical point, compared to parameters far from it.

We compute the KL divergence rate at steady state between the full model and its linearisation using (44) and via Monte Carlo estimation (see Fig. 4**d**). The steady-state KL divergence rate exhibits a notable peak near the critical point, coinciding with the parameter regime where the linearisation fails to capture the full model. As we move away in either direction from the critical point the KL divergence decreases in accordance with the better approximation of the system by its linearised version. This shows how the KL divergence can be used to assess how well model reduction works for different parameter regimes.

3.2 Michaelis-Menten Kinetics

We next consider one of the most-studied reaction networks in biology, the Michaelis-Menten model of enzyme kinetics, consisting of an enzyme E and a substrate S that reversibly bind to form an enzyme-substrate complex ES which converts the substrate into the product P (see Fig. 5**a**). This system is often treated as an example application of both the QSSA and the QEA, which remain the standard model reduction techniques applied to this example.

The QSSA classically models the enzyme E and the enzyme-substrate complex ES as the fast species, and replaces the full system by a one-step reaction where the substrate is converted to the product with a Hill-type propensity function (see Fig. 5**b**). This approximation is known to be valid when

$$E_T \ll [S] + K_m, \tag{47}$$

where E_T is the total number of enzymes and K_m is the Michaelis-Menten constant $(k_{-1} + k_2)/k_1$ [50, 51]. Note that the reduced model is invariant under the scaling

$$E_T \mapsto \varepsilon \cdot E_T \qquad\qquad k_i \mapsto \varepsilon^{-1} \cdot k_i, \quad i = -1, 1, 2, \tag{48}$$

where the parameter ε can be seen as controlling the accuracy of the reduction: the rescaled QSSA condition (47) reads

$$\varepsilon E_T \ll [S] + K_m, \tag{49}$$

which suggests that small values of ε should result in the full model more closely resembling the QSSA reduction.

The stochastic QEA for this system is analysed in [17] and also results in a one-step reaction where the propensity function is now piecewise linear (see Fig. 5**b**). As opposed to the QSSA, the QEA is accurate when substrate binding and unbinding is fast, i.e. k_1, k_{-1} are sufficiently large. Note that the QEA reduction is invariant under the scaling

$$k_1 \mapsto c \cdot k_1 \qquad\qquad k_{-1} \mapsto c \cdot k_{-1}, \tag{50}$$

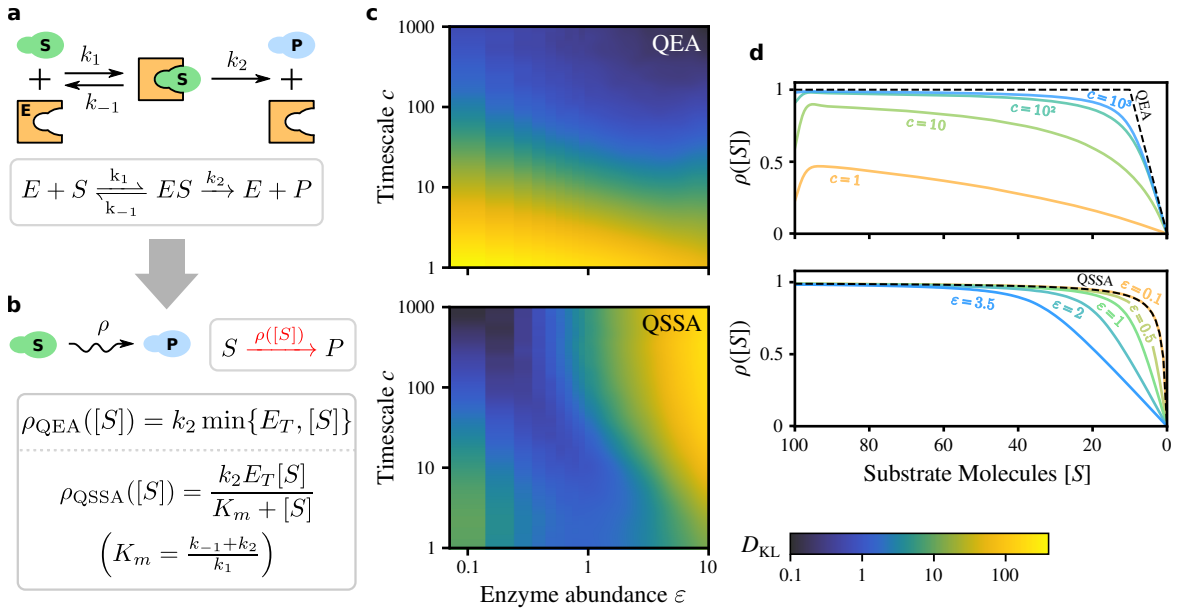

Figure 5: Comparison of the QEA and the QSSA for Michaelis-Menten kinetics. **(a)** Schematic of the Michaelis-Menten system. We assume mass action kinetics. **(b)** Reduction of the Michaelis-Menten system under the QEA and the QSSA yields a one-step process with effective (non-mass action) propensities that depend on the method used. **(c)** Total KL divergence between the full model and the QEA and QSSA for different values of the timescale parameter c in (50) and the enzyme abundance parameter ε in (48). The QEA generally becomes more accurate as c increases, and the QSSA becomes more accurate as ε decreases. Note the small copy number effects that become apparent in the QSSA for low values of ε and c. We used the analytically obtained propensities for the QEA and the QSSA, and numerically estimated the KL divergences using Alg. 2. **(d)** Comparison of the effective reaction propensities computed according to Eq. (14) with those predicted by the QEA and the QSSA. Here substrates include enzyme-bound substrates. The top figure shows the effective propensities for $\varepsilon = 1$: as $c \to \infty$ the effective propensities converge to the function predicted by the QEA. The bottom figure shows the effective reaction propensities for $c = 1000$, and as ε decreases the propensities approach the function predicted by the QSSA. The remaining parameters were fixed to $k_1 = k_{-1} = 0.001$, $k_2 = 0.1$, $E_T = 10$ and $S_T = 100$.

where c controls the timescale at which quasiequilibrium is reached, and therefore the accuracy of the QEA. We stress that (48) and (50) are two independent scalings and can be performed simultaneously, as will be the case in Fig. 5**c**.

In general it may be difficult to predict which of the two approaches is more accurate unless one is clearly in one of the limiting regimes. Based on the KL divergence between the full model and either of the two reactions we can investigate this question for a range of parameters. The reduced model consists of two species, S and P, and since $[S] + [P]$ is conserved we can describe it in terms of either of the two. The correct projection for this example identifies the species S and ES in the full model, ie. we define $[S]_{\text{red}} = [S] + [ES]$ in order to eliminate the binding and unbinding reactions from the projection. This lumping of two rapidly equilibrating species is standard when using the QEA (see e.g. [24]), but it applies equally well to the QSSA in this case.

In Fig. 5**c** we use Alg. 2 to estimate the total KL divergence over $[0, \infty]$ between the full model and both reductions for a fixed number of substrate molecules. The total KL divergence is finite since all trajectories enter the same absorbing state defined by $[S] = 0$ in a finite amount of time. As expected for the QEA its accuracy increases with c, but whereas the QSSA tends to become less accurate for large E_T, in the low enzyme regime we observe a similar decrease in accuracy that is not explained by the deterministic theory.

The observed decrease is a small copy number effect caused by the fact that the waiting time distribution between two productions of P is not exponential: for very small E_T the unbinding of an enzyme immediately after such an event implies that a significant fraction of such productions occur as a two-step process (where the free enzyme binds another substrate and then converts it) [52]. For large enough c the binding step is very fast, and assuming that $k_2 \gg k_{-1}$ so that unbinding is unlikely to occur again, the conversion of substrates to products can be viewed as an effective one-step process. The effect of nonexponential waiting time distributions has previously been analysed in [53]; we see that the KL divergence on trajectories can be sensitive to subtle dynamical effects such as waiting time distribution that are not visible when considering e.g. the moments of a system, which are well predicted by the QSSA for small values of c and ε.

Overall we can see that for the chosen parameter values the QSSA generally performs better than the QEA in the regime of small ε, corresponding to low enzyme numbers, and the QEA is most accurate for c, keeping in mind the scalings in (48) and (50). Neither approximation is satisfactory if the number of total enzymes is larger than the amount of substrates, but the binding and unbinding rates are small. In this scenario other approximations, such as the total QSSA [54], which we do not investigate here, will generally be more accurate.

Figure 5**d** compares the effective propensities for the full system, as a function of the unconverted substrate abundance, with the predictions made by the QEA resp. the QSSA. In the case of the QEA we see that the effective propensities slowly converge to their limit as the timescale parameter c increases. In contrast the QSSA provides a good approximation to the effective propensity as long as the number of substrate molecules is larger than the number of enzymes. While the effective propensities are the optimal choice for the reduced model, the actual quality of the approximation is affected both by the size of the fluctuations of the actual propensities around their mean as well as the degree to which the waiting times of the original system follow an exponential distribution.

3.3 Genetic Oscillator

Our final example is the gene expression system shown in Fig. 6**a**. This system consists of a gene, mRNA and protein, as well as a Michaelis-Menten type protein degradation mechanism. Up to two protein molecules can bind the gene, and in the twice bound state the gene pauses transcription. This model has been considered in [11] as an example system where naive reduction of the CME using

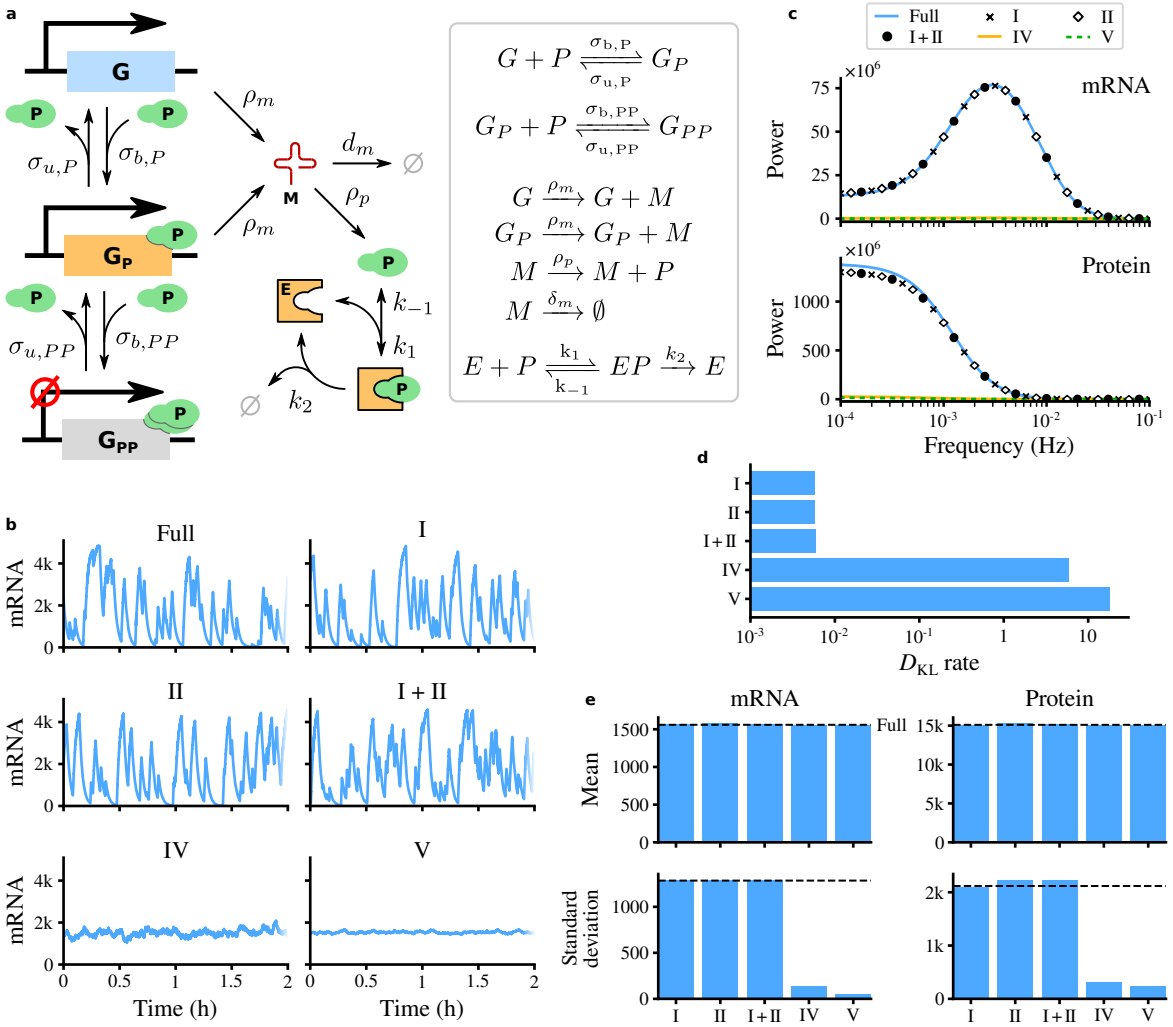

Figure 6: Comparison of different reductions for the oscillatory gene network. **(a)** Schematic of the reaction system. We assume mass action kinetics for all reactions in the full model. **(b)** Example trajectories for the full model and its reductions. Note the oscillatory behaviour in the original model that is not preserved in reductions IV and V. We used Alg. 1 to compute the effective parameters for each reduction (cf. Table 2) from a single long trajectory probing steady-state behaviour. **(c)** Power spectra of mRNA and protein concentrations (defined as copy numbers normalised by the system size Ω). Models I, II and I+II closely reproduce the oscillatory behaviour of the full model while IV and V do not show sustained oscillations. **(d)** KL divergence rates between the full model and all reductions, estimated using Alg. 2. **(e)** Steady-state means and standard deviations for mRNA and protein numbers, estimated numerically using the Gillespie algorithm. While all reductions closely approximate the mean, Models IV and V do not match the variance of the full model. The parameters for the full model, taken from [11], are $\Omega = 1000$, $E_T = 10$, $\rho_m = 50$, $\rho_p = 0.0045$, $d_m = 0.01$, $k_1 = 0.1$, $k_{-1} = 10$, $k_2 = 10$, $\sigma_{b,P} = 0.001$, $\sigma_{u,P} = 100$, $\sigma_{b,PP} = 1000$, $\sigma_{u,PP} = 1$.

Hill-type effective propensities is unable to accurately capture the noise in the original. The true system can exhibit oscillations in mRNA numbers that are caused by the two-step negative feedback and are not present in the reduced version. In this section we want to analyse what reductions can be performed on the full model while still keeping oscillatory behaviour.

We consider five different reductions for this model which are listed in Table 2. Each reduction removes or combines several species such as gene states, thereby reducing the dimensionality of the system. The rightmost column describes the correct projection for each model, derived using an analogous argument as in the Michaelis-Menten example. We fit the unknown (effective) parameters for each model numerically by minimising the KL divergence from the full model using Alg. 1. Typical mRNA trajectories for all models can be seen in Fig. 6**b**. The oscillations in the full model are clearly visible, and they are inherited by reductions I, II and I+II. In contrast, simplifying the first binding step as in IV and V results in a non-oscillatory system (see Fig. 6**c**).

In Fig. 6**d** we compare the KL divergence rates between the full model and reductions I-V (note that the different reductions are defined on different state spaces). While models I, II and I+II have comparatively low KL divergences from the original there is a sharp increase with models IV and V. From this alone one could expect that the latter perform significantly worse, which is indeed the case as shown in Fig. 6**b**. Fig. 6**e** presents a comparison of the mean and standard deviations for mRNA and protein abundances; while all models approximate the means very closely, the predicted standard deviations for model IV and V are very far from their true values consistent with the lack oscillatory behaviour. This resembles the results in Section 3.1 which also show excellent agreement of the full and reduced model on the mean level, independent of the total approximation quality.

This comparison of various reductions shows how variational model reduction can be employed on a more sophisticated scale, where multiple reductions are possible. Given a list of possible simplifications we can automatically find optimal parameters for each reduction and compute the KL divergence from the true model. Reductions which do not significantly affect the output will generally lead to very low KL divergences compared to those which do, as in this example where a large gap between KL divergence rates separates reductions I, II and I+II from IV and V. The latter do not retain much of the moments or oscillatory dynamics of the full model compared to the other reductions.

4 Discussion

In this work we presented an information-theoretic approach to model reduction for the CME based on minimising the KL divergence between the full model and the proposed reduction. Based on this variational principle we determine the optimal reaction propensities for the reduced model in Eq. (14) and show that it underlies some of the most common approaches to model reduction for the CME: the QSSA [8–10], the QEA [16, 17] and the LMA [25]. As a consequence, we establish that these methods can be seen as special cases of our approach. We furthermore obtain a general justification for the mean-field arguments proposed in the literature and connect them with information theory and likelihood-based approaches. We provide a numerical algorithm for automated fitting of a reduced model based on minimising the KL divergence via stochastic gradient descent, and a numerical algorithm for estimating this KL divergence in order to assess model fit. While the KL divergence between Markov chains has been considered before in e.g. [19–21, 31, 32, 34, 35, 55], to our knowledge it has not been studied in connection with standard model reduction methods as done in this paper. Our numerical results show how the KL divergence can be computed in practice and used in the context of model reduction.

Model	Full reactions	Reduced reactions	Projection map
I	$E + P \rightleftharpoons EP \longrightarrow E$	$P \xrightarrow{d_{\text{eff}}} \emptyset$	$[P]_{\text{red}} := [P] + [EP]$
II	$G + P \rightleftharpoons G_P$ $G_P + P \rightleftharpoons G_{PP}$ $G_P \longrightarrow G_P + P$	$G + P \underset{\sigma_{\text{u}}/(\text{K}_{\text{u}} + [\text{P}])}{\overset{\sigma_{\text{b,P}}}{\rightleftharpoons}} G_{P/PP}$ $G_{P/PP} \xrightarrow{\rho_{\text{eff}}} G_{P/PP} + P$	$[P]_{\text{red}} := [P] + [G_{PP}]$ $[G_{P/PP}]_{\text{red}} := [G_P] + [G_{PP}]$
I + II	(I & II above)	(I & II above)	$[P]_{\text{red}} := [P] + [EP] + [G_{PP}]$ $[G_{P/PP}]_{\text{red}} := [G_P] + [G_{PP}]$
IV	$G + P \rightleftharpoons G_P$ $G_P + P \rightleftharpoons G_{PP}$ $G \longrightarrow G + P$ $G_P \longrightarrow G_P + P$	$G_{-/P} + P \underset{\sigma_{\text{u,PP}}}{\overset{\sigma_{\text{b}}[\text{P}]/(\text{K}_{\text{b}} + [\text{P}])}{\rightleftharpoons}} G_{PP}$ $G_{-/P} \xrightarrow{\rho_m} G_{-/P} + P$	$[P]_{\text{red}} := [P] + [G_P] + [G_{PP}]$ $[G_{-/P}]_{\text{red}} := [G] + [G_P]$
V	$G + P \rightleftharpoons G_P$ $G_P + P \rightleftharpoons G_{PP}$ $G \longrightarrow G + P$ $G_P \longrightarrow G_P + P$	$\emptyset \xrightarrow{\frac{\alpha + \beta[P]}{1 + \gamma[P] + \delta[P]^2}} P$	$[P]_{\text{red}} := [P] + [G_P] + 2[G_{PP}]$

Table 2: Five possible reductions for the oscillator model. Model I+II is the combination of the reductions in Models I and II. All propensities are shown without their mass action terms. Species that are lumped or removed in the reduced model are shown in gray, and species that do not exist in the full model in blue. Parameters that are introduced in the reduced model are shown in red and found by minimisation of the KL divergence between the models; all remaining parameters are taken over from the full model. The functional forms of the propensity functions for Models II, IV and V were derived using the QEA.

Using three biologically relevant examples we illustrated how these model reduction techniques can be analysed from this variational perspective. For the autoregulatory feedback loop we showed that the KL divergence provides a useful metric of approximation quality and can detect parameter regimes where the mean-field approximation behind the LMA fails. Using the Michaelis-Menten system we demonstrated how our approach can be used to decide between possible reductions of a model, particularly in the non-asymptotic regime where neither the QSSA nor the QEA are strictly valid. Finally we used the genetic oscillator in [11] as an example to show how different reductions of a given model can be fit automatically, separately and in combination, and assesses which steps in a putative model reduction procedure would impact approximation quality more than others.

In this paper we only focused on discrete stochastic models, mainly those described using the Chemical Master Equation. This approach disregards continuum approximations such as the Chemical Langevin Equation and the System Size Expansion [56], which are frequently used in practice, and it is worth asking whether information-theoretic approaches can be extended to this context. A priori this appears difficult since the KL divergence and related quantities tend to diverge in the continuum limit, such as the divergence between Brownian motions with different diffusion coefficients. This similarly applies to the wide class of hybrid approaches [30, 42, 57, 58]. In particular, abundance separation, as opposed to time-scale separation with the QSSA and the QEA, does not fall under the umbrella approach we considered. In this scenario other perspectives on model reduction are likely needed.

A major drawback of the general approach we presented is that the KL divergences we use cannot usually be optimised analytically. As we have shown, the minimum typically corresponds to computing conditional expectations as in (14), which is rarely possible exactly, e.g. in the case of the LMA, where moment equations are used to approximate conditional means of protein numbers, or the QSSA, where deterministic rate equations are commonly used to derive the approximate propensities. The Monte-Carlo approach we proposed in Alg. 1 has the advantage of being completely generic, but unlike most existing techniques it requires numerical optimisation and will inevitably be slower than those. The optimisation problem in question typically being low-dimensional and convex, most of the computational time is required to generate sample trajectories and compute gradients, with variable computational complexity that heavily depends on the system.

One important aspect of model reduction we have not addressed is that of choosing an appropriate architecture for a reduced model. As we investigated in the case of the genetic oscillator, while we can always optimise the parameters given an architecture, the quality of the approximation can greatly depend on which reductions are performed. Although methods for automatically choosing from a predefined set of approximations of a system exist in special cases [15], a fully general algorithm that computes a maximally reduced version of a given model within a given threshold is still missing from the literature. In combination with other machine learning approaches to model reduction such as [59], however, we hope that our approach may be one of several steps towards such a procedure.

The KL divergence is a well-studied quantity in statistics but it lacks some desirable properties in the context of model reduction. While we were able to establish that optimising the KL divergence in theory leads to a reduced system with the same marginal distributions, establishing a relationship between the KL divergence and dynamical information such as the autocorrelation and power spectra is more difficult. Our investigations suggest that the KL divergence on the space of trajectories is a useful quantity that relates to both the marginal distributions and dynamical properties of the system, and can be a useful metric of model fit. We are optimistic that further investigations in this area will demarcate more clearly which properties of a model are well captured by the KL divergence and which are not.

Code availability

Code implementing the methods introduced in this paper can be found at https://github.com/kaandocal/modred.

Acknowledgments

This work was supported by the EPSRC Centre for Doctoral Training in Data Science (EPSRC grant EP/L016427/1) and the University of Edinburgh for K. Ö. and a Leverhulme Trust grant (Grant No. RPG-2020-327) for R. G. The authors are grateful to the anonymous reviewers for their useful feedback and suggestions.

References

[1] M. Kærn, T. C. Elston, W. J. Blake, and J. J. Collins, "Stochasticity in gene expression: from theories to phenotypes," *Nat. Rev. Genet.*, **6**(6): 451–464, 2005.

[2] R. Satija and A. K. Shalek, "Heterogeneity in immune responses: from populations to single cells," *Trends Immunol.*, **35**(5): 219–229, 2014.

[3] T. Lipniacki, B. Hat, J. R. Faeder, and W. S. Hlavacek, "Stochastic effects and bistability in T cell receptor signaling," *J. Theor. Biol.*, **254**(1): 110–122, 2008.

[4] D. Schnoerr, G. Sanguinetti, and R. Grima, "Approximation and inference methods for stochastic biochemical kinetics - a tutorial review," *J. Phys. A*, **50**(9): 093001, 2017.

[5] B. Munsky and M. Khammash, "The Finite State Projection algorithm for the solution of the Chemical Master Equation," *J. Chem. Phys.*, **124**(4): 044104, 2006.

[6] N. Ali Eshtewy and L. Scholz, "Model reduction for kinetic models of biological systems," *Symmetry*, **12**(5): 863, 2020.

[7] T. J. Snowden, P. H. van der Graaf, and M. J. Tindall, "Methods of model reduction for large-scale biological systems: a survey of current methods and trends," *Bull. Math. Biol.*, **79**(7): 1449–1486, 2017.

[8] C. V. Rao and A. P. Arkin, "Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm," *J. Chem. Phys.*, **118**(11): 4999–5010, 2003.

[9] J. K. Kim, K. Josić, and M. R. Bennett, "The relationship between stochastic and deterministic quasi-steady state approximations," *BMC Syst. Biol.*, **9**(1): 87, 2015.

[10] J. K. Kim, K. Josić, and M. R. Bennett, "The validity of quasi-steady-state approximations in discrete stochastic simulations," *Biophys. J.*, **107**(3): 783–793, 2014.

[11] P. Thomas, A. V. Straube, and R. Grima, "The slow-scale Linear Noise Approximation: an accurate, reduced stochastic description of biochemical networks under timescale separation conditions," *BMC Syst. Biol.*, **6**(1): 39, 2012.

[12] P. Thomas, A. V. Straube, and R. Grima, "Limitations of the stochastic quasi-steady-state approximation in open biochemical reaction networks," *J. Chem. Phys.*, **135**(18): 181103, 2011.

[13] H.-W. Kang, W. R. KhudaBukhsh, H. Koeppl, and G. A. Rempała, "Quasi-steady-state approximations derived from the stochastic model of enzyme kinetics," *Bull. Math. Biol.*, **81**(5): 1303–1336, 2019.

[14] J. Eilertsen, K. Srivastava, and S. Schnell, "Stochastic enzyme kinetics and the quasi-steady-state reductions: Application of the slow-scale Linear Noise Approximation à la Fenichel," *J. Math. Biol.*, **85**(1): 3, 2022.

[15] Y. M. Song, H. Hong, and J. K. Kim, "Universally valid reduction of multiscale stochastic biochemical systems using simple non-elementary propensities," *PLoS Comput Biol*, **17**(10): e1008952, 2021.

[16] E. L. Haseltine and J. B. Rawlings, "Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics," *J. Chem. Phys.*, **117**(15): 6959–6969, 2002.

[17] J. Goutsias, "Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems," *The Journal of Chemical Physics*, **122**(18): 184102, 2005.

[18] L. Cardelli, I. Pérez Verona, M. Tribastone, *et al.*, "Exact maximal reduction of stochastic reaction networks by species lumping," *Bioinform.*, **37**: 2021.

[19] R. A. Amjad, C. Blöchl, and B. C. Geiger, "A generalized framework for Kullback-Leibler Markov aggregation," *IEEE Trans. Automat. Contr.*, **65**(7): 3068–3075, 2020.

[20] K. Deng, P. G. Mehta, and S. P. Meyn, "Optimal Kullback-Leibler aggregation via spectral theory of Markov chains," *IEEE Trans. Automat. Contr.*, **56**(12): 2793–2808, 2011.

[21] B. C. Geiger, T. Petrov, G. Kubin, and H. Koeppl, "Optimal Kullback-Leibler aggregation via information bottleneck," *IEEE Trans. Automat. Contr.*, **60**(4): 1010–1022, 2015.

[22] S. Bo and A. Celani, "Multiple-scale stochastic processes: decimation, averaging and beyond," *Phys. Rep.*, **670**: 1–59, 2017.

[23] J. Holehouse, A. Sukys, and R. Grima, "Stochastic time-dependent enzyme kinetics: closed-form solution and transient bimodality," *J. Chem. Phys.*, **153**(16): 164113, 2020.

[24] C. Jia and R. Grima, "Dynamical phase diagram of an auto-regulating gene in fast switching conditions," *J. Chem. Phys.*, **152**(17): 174110, 2020.

[25] Z. Cao and R. Grima, "Linear Mapping Approximation of gene regulatory networks with stochastic dynamics," *Nat. Commun.*, **9**(1): 3305, 2018.

[26] T. Jahnke and W. Huisinga, "Solving the Chemical Master Equation for monomolecular reaction systems analytically," *J. Math. Biol.*, **54**(1): 1–26, 2007.

[27] C. Gadgil, C. H. Lee, and H. G. Othmer, "A stochastic analysis of first-order reaction networks," *Bull. Mathem. Biol.*, **67**(5): 901–946, 2005.

[28] T. Zhou and J. Zhang, "Analytical results for a multistate gene model," *SIAM J. Appl. Math.*, **72**(3): 789–818, 2012.

[29] Y. Li, D.-Q. Jiang, and C. Jia, "Steady-state joint distribution for first-order stochastic reaction kinetics," *Phys. Rev. E*, **104**(2): 024408, 2021.

[30] T. Jahnke, "On Reduced Models for the Chemical Master Equation," *Multiscale Model. Simul.*, **9**(4): 1646–1676, 2011.

[31] M. Opper and G. Sanguinetti, "Variational inference for Markov jump processes," in *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[32] C. Wildner and H. Koeppl, "Moment-based variational inference for Markov jump processes," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, 6766–6775.

[33] C. Zechner and H. Koeppl, "Uncoupled analysis of stochastic reaction networks in fluctuating environments," *PLOS Computational Biology*, **10**(12): e1003942, 2014.

[34] L. Bronstein and H. Koeppl, "Marginal process framework: a model reduction tool for Markov jump processes," *Phys. Rev. E*, **97**(6): 062147, 2018.

[35] A.-L. Moor and C. Zechner, *Dynamic information transfer in stochastic biochemical networks*, 2022. arXiv: 2208.04162 [q-bio.MN].

[36] W. R. KhudaBukhsh, A. Auddy, Y. Disser, and H. Koeppl, "Approximate lumpability for Markovian agent-based models using local symmetries," *J. Appl. Probab.*, **56**(3): 647–671, 2019.

[37] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. The MIT Press, 2023.

[38] J. Peccoud and B. Ycart, "Markovian modeling of gene-product synthesis," *Theor. Popul. Biol.*, **48**(2): 222–234, 1995.

[39] J. K. Kim and J. C. Marioni, "Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data," *Genome Biol*, **14**(1): R7, 2013.

[40] L. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proc. IEEE*, **77**(2): 257–286, 1989.

[41] F. Confortola and M. Fuhrman, "Filtering of continuous-time Markov chains with noise-free observation and applications," *Stochastics*, **85**: 216–251, 2010.

[42] J. Hasenauer, V. Wolf, A. Kazeroonian, and F. J. Theis, "Method of conditional moments (MCM) for the Chemical Master Equation: a unified framework for the method of moments and hybrid stochastic-deterministic models," *J Math Biol*, **69**(3): 687–735, 2014.

[43] L. Duso and C. Zechner, "Selected-node stochastic simulation algorithm," *The Journal of Chemical Physics*, **148**(16): 164108, 2018.

[44] M. Rathinam and M. Yu, "State and parameter estimation from exact partial state observation in stochastic reaction networks," *The Journal of Chemical Physics*, **154**(3): 034103, 2021.

[45] Y. Cao, D. T. Gillespie, and L. R. Petzold, "The slow-scale Stochastic Simulation Algorithm," *J. Chem. Phys.*, **122**(1): 014116, 2005.

[46] E. L. Haseltine and J. B. Rawlings, "On the origins of approximations for stochastic chemical kinetics," *J. Chem. Phys.*, **123**(16): 164115, 2005.

[47] D. W. Kim, H. Hong, and J. K. Kim, "Systematic inference identifies a major source of heterogeneity in cell signaling dynamics: The rate-limiting step number," *Science Advances*, **8**(11): eabl4598, 2022.

[48] C. Jia and R. Grima, "Small protein number effects in stochastic models of autoregulated bursty gene expression," *J. Chem. Phys.*, **152**(8): 084115, 2020.

[49] K. Öcal, R. Grima, and G. Sanguinetti, "Parameter estimation for biochemical reaction networks using Wasserstein distances," *J. Phys. A*, **53**(3): 034002, 2019.

[50] L. A. Segel, "On the validity of the steady state assumption of enzyme kinetics," *Bull. Math. Biol.*, **50**(6): 579–593, 1988.

[51] D. Gillespie, K. Sanft, and L. Petzold, "Legitimacy of the stochastic Michaelis–Menten approximation," *IET Syst. Biol.*, **5**(1): 58–69, 2011.

[52] R. Grima and A. Leier, "Exact product formation rates for stochastic enzyme kinetics," *J. Phys. Chem. B*, **121**(1): 13–23, 2017.

[53] D. T. Gillespie, Y. Cao, K. R. Sanft, and L. R. Petzold, "The subtle business of model reduction for stochastic chemical kinetics," *J. Chem. Phys.*, **130**(6): 064103, 2009.

[54] S. MacNamara, A. M. Bersani, K. Burrage, and R. B. Sidje, "Stochastic chemical kinetics and the total quasi-steady-state assumption: application to the Stochastic Simulation Algorithm and Chemical Master Equation," *J. Chem. Phys.*, **129**(9): 095105, 2008.

[55] Z. Rached, F. Alajaji, and L. Campbell, "The Kullback-Leibler divergence rate between Markov sources," *IEEE Trans. Automat. Contr.*, **50**(5): 917–921, 2004.

[56] N. van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd. Elsevier, 2007.

[57] A. Hellander and P. Lötstedt, "Hybrid method for the Chemical Master Equation," *Journal of Computational Physics*, **227**(1): 100–122, 2007.

[58] S. Smith, C. Cianci, and R. Grima, "Model reduction for stochastic chemical systems with abundant species," *J. Chem. Phys.*, **143**(21): 214105, 2015.

[59] G. Caravagna, L. Bortolussi, and G. Sanguinetti, "Matching models across abstraction levels with Gaussian processes," in *Computational Methods in Systems Biology*, E. Bartocci, P. Lio, and N. Paoletti, Eds., Cham: Springer International Publishing, 2016, 49–66.

[60] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *J. Comput. Phys.*, **22**(4): 403–434, 1976.

A Likelihoods of Fully Observed Trajectories

Let p be a continuous-time Markov chain defined on the space $\mathcal{X}$ with initial distribution p_0. If p has time-independent transition rates the Stochastic Simulation Algorithm [60], reproduced in Algorithm 3, returns exact samples from p. We can inspect the SSA to obtain the probability of drawing any given trajectory as follows:

- The first state $\boldsymbol{n}(0)$ is drawn from the initial distribution p_0 and has probability $p_0(\boldsymbol{n}(0))$
- The time τ_1 until the next jump is drawn from an exponential distribution with rate $p_{\leftarrow \boldsymbol{n}_0}$ and has probability distribution function $p_{\leftarrow \boldsymbol{n}_0} \exp(-\tau_1 p_{\leftarrow \boldsymbol{n}_0})$.
- The next state $\boldsymbol{n}_1$ is drawn from the transition distribution $p_{\boldsymbol{n}_1 \,|\, \boldsymbol{n}_0}$ and has probability $p_{\boldsymbol{n}_1 \leftarrow \boldsymbol{n}_0} / p_{\leftarrow \boldsymbol{n}_0}$
- The time τ_2 until the next jump is drawn from an exponential distribution with rate $p_{\leftarrow \boldsymbol{n}_1}$ and has probability distribution function $p_{\leftarrow \boldsymbol{n}_1} \exp(-\tau_2 p_{\leftarrow \boldsymbol{n}_1})$, etc.

We thus arrive at the probability

$$p(\boldsymbol{n}_{[0,T]}) = p_0(\boldsymbol{n}(0)) \cdot p_{\leftarrow \boldsymbol{n}_0} \cdot \exp(-\tau_1 p_{\leftarrow \boldsymbol{n}_0}) \cdot \frac{p_{\boldsymbol{n}_1 \leftarrow \boldsymbol{n}_0}}{p_{\leftarrow \boldsymbol{n}_0}} \cdot \exp(-\tau_2 p_{\leftarrow \boldsymbol{n}_1}) \cdot \ldots \cdot \exp(-\tau_k p_{\leftarrow \boldsymbol{n}_{k-1}}) \tag{A1}$$

where the last term corresponds to the probability that $t_{k+1} > T$, ie. that the next jump occurs after time T (note the absence of the prefactor $p_{\leftarrow \boldsymbol{n}_k}$). Each term $p_{\leftarrow \boldsymbol{n}_i}$ in the above expression cancels with the denominator in the next term, and upon taking logarithms we arrive at Eq. (4).

Algorithm 3 The Stochastic Simulation Algorithm to simulate samples from a Markov chain p.

Input: simulation length T, Markov chain p, initial distribution p_0
Output: $\boldsymbol{n}_{[0,T]}$ - sampled trajectory

sample $\boldsymbol{n}_0 \sim p_0$
$t \leftarrow 0$, $i \leftarrow 0$
while $t < T$ **do**
  sample $\tau_{i+1} \sim \mathrm{Exp}(1/p_{\leftarrow \boldsymbol{n}_i})$
  sample $\boldsymbol{n}_{i+1} = \boldsymbol{m}$ **with probability** $p_{\boldsymbol{m} \leftarrow \boldsymbol{n}_i} / p_{\leftarrow \boldsymbol{n}_i}$
  $t \mathrel{+}= \tau_{i+1}$
return states $(\boldsymbol{n}_0, \boldsymbol{n}_1, \ldots)$, jump times $(t_1, t_2, \ldots)$

B Kullback-Leibler Divergence between Markov Chains

In [31] the authors derive an expression for the Kullback-Leibler divergence between two continuous-time Markov chains q and p defined on a common state space $\mathcal{X}$. For any integer $N > 1$ let $q^{(N)}$ resp. $p^{(N)}$ be the time discretisations of q and p with time step $\delta t = T/N$. These define probability distributions on $\mathcal{X}^{N+1}$ with KL divergence

$$\mathrm{KL}(q^{(N)} \,\|\, p^{(N)}) = \sum_{\boldsymbol{n}} q_0(\boldsymbol{n}) \log \frac{q_0(\boldsymbol{n})}{p_0(\boldsymbol{n})} + \sum_{i=1}^{N} \sum_{\boldsymbol{n},\boldsymbol{m}} q_{i-1}^{(N)}(\boldsymbol{n})\, q_i^{(N)}(\boldsymbol{m} \,|\, \boldsymbol{n}) \log \frac{q_i^{(N)}(\boldsymbol{m} \,|\, \boldsymbol{n})}{p_i^{(N)}(\boldsymbol{m} \,|\, \boldsymbol{n})} \tag{B1}$$

The transition rates are related to the discrete-time transition probabilities as follows:

$$q_i^{(N)}(\boldsymbol{n} \,|\, \boldsymbol{n}) = 1 - (\delta t)\, q_{\leftarrow \boldsymbol{n}}(t^{(i)}) + o(\delta t) \tag{B2}$$

$$q_i^{(N)}(\boldsymbol{m} \,|\, \boldsymbol{n}) = (\delta t)\, q_{\boldsymbol{m} \leftarrow \boldsymbol{n}}(t^{(i)}) + o(\delta t) \qquad (\boldsymbol{m} \neq \boldsymbol{n}) \tag{B3}$$

where $t^{(i)} = iT/N$ is the i-th discretisation time point. Using these identities for q and p it can be verified that the KL divergence (B1) converges to the following as $N \to \infty$:

$$\begin{aligned} \mathrm{KL}(q \,\|\, p) = \mathrm{KL}(q_0 \,\|\, p_0) &- \int_0^T \mathrm{d}t \sum_{\boldsymbol{n}} q_t(\boldsymbol{n})\, (q_{\leftarrow \boldsymbol{n}}(t) - p_{\leftarrow \boldsymbol{n}}(t)) \\ &+ \int_0^T \mathrm{d}t \sum_{\boldsymbol{m} \neq \boldsymbol{n}} q_t(\boldsymbol{n})\, q_{\boldsymbol{m} \leftarrow \boldsymbol{n}}(t) (\log q_{\boldsymbol{m} \leftarrow \boldsymbol{n}}(t) - \log p_{\boldsymbol{m} \leftarrow \boldsymbol{n}}(t)) \end{aligned} \tag{B4}$$

In general the KL divergence can be written as the difference of the cross-entropy $H(q; p)$ and the entropy $H(q; p)$, and we define the cross-entropy of two continuous-time Markov chains as

$$H(q; p) = H(q_0; p_0) + \int_0^T \mathrm{d}t \sum_{\boldsymbol{n}} q_t(\boldsymbol{n})\, p_{\leftarrow \boldsymbol{n}}(t) - \int_0^T \mathrm{d}t \sum_{\boldsymbol{m} \neq \boldsymbol{n}} q_t(\boldsymbol{n})\, q_{\boldsymbol{m} \leftarrow \boldsymbol{n}}(t) \log p_{\boldsymbol{m} \leftarrow \boldsymbol{n}}(t). \tag{B5}$$

The above derivation is valid only if $q_{\boldsymbol{m} \leftarrow \boldsymbol{n}} \neq 0$ implies $p_{\boldsymbol{m} \leftarrow \boldsymbol{n}} \neq 0$; otherwise the cross-entropy and KL divergence are both infinite.

This definition of the cross-entropy generalises to the case where q is not Markovian, e.g. when it is the projection of a Markov process. In the discrete-time case we have

$$H(\tilde{q}^{(N)}; p^{(N)}) = H(\tilde{q}_0; p_0) - \sum_{i=1}^{N} \sum_{\tilde{\boldsymbol{n}}, \tilde{\boldsymbol{m}}} \tilde{q}_{i-1}^{(N)}(\tilde{\boldsymbol{n}} \,|\, q_0)\, \tilde{q}_i^{(N)}(\tilde{\boldsymbol{m}} \,|\, \tilde{\boldsymbol{n}}, q_0) \log p_i^{(N)}(\tilde{\boldsymbol{m}} \,|\, \tilde{\boldsymbol{n}}), \tag{B6}$$

where the marginal and conditional distributions of $\tilde{q}$ depend on the initial distribution of q since $\tilde{q}$ is no longer assumed to be memoryless. Expressing the transition probabilities for p in terms of transition rates using Eqs. (B2) and (B3), as well as the identities

$$\tilde{q}_i^{(N)}(\tilde{\boldsymbol{n}} \,|\, \tilde{\boldsymbol{n}}, q_0) = 1 - (\delta t)\, \tilde{q}_{\leftarrow \tilde{\boldsymbol{n}}}(t^{(i)}) + o(\delta t), \tag{B7}$$

$$\tilde{q}_i^{(N)}(\tilde{\boldsymbol{m}} \,|\, \tilde{\boldsymbol{n}}, q_0) = (\delta t)\, q_{\tilde{\boldsymbol{m}} \leftarrow \tilde{\boldsymbol{n}}}(t^{(i)}) + o(\delta t), \qquad (\tilde{\boldsymbol{m}} \neq \tilde{\boldsymbol{n}}) \tag{B8}$$

which follow from the definition of the marginal transition rates in Eq. (12), we obtain

$$\begin{aligned} H(\tilde{q}^{(N)}; p^{(N)}) = H(\tilde{q}_0; p_0) &+ \sum_{i=1}^{N} \sum_{\tilde{\boldsymbol{n}}} \tilde{q}_{i-1}^{(N)}(\tilde{\boldsymbol{n}} \,|\, q_0) \left(1 - (\delta t) \sum_{\tilde{\boldsymbol{m}} \neq \tilde{\boldsymbol{n}}} \tilde{q}_{\tilde{\boldsymbol{m}} \leftarrow \tilde{\boldsymbol{n}}}(t^{(i)})\right) \log \left(1 - (\delta t) \sum_{\tilde{\boldsymbol{m}} \neq \tilde{\boldsymbol{n}}} p_{\tilde{\boldsymbol{m}} \leftarrow \tilde{\boldsymbol{n}}}\right) \\ &- (\delta t) \sum_{i=1}^{N} \sum_{\tilde{\boldsymbol{n}}} \tilde{q}_{i-1}^{(N)}(\tilde{\boldsymbol{n}} \,|\, q_0) \sum_{\tilde{\boldsymbol{m}} \neq \tilde{\boldsymbol{n}}} \tilde{q}_{\tilde{\boldsymbol{m}} \leftarrow \tilde{\boldsymbol{n}}}(t^{(i)}) \log p_{\tilde{\boldsymbol{m}} \,|\, \tilde{\boldsymbol{n}}}(t^{(i)}) - N \log(\delta t) + o(\delta t). \end{aligned} \tag{B9}$$

The term $N \log(\delta t)$ is independent of p and we will renormalise it to 0; this is standard when deriving continuum-limit versions of many information-theoretic quantities and was implicitly used in our definition of the cross-entropy in Eq. (B5). Proceeding with the above equation to get

$$\begin{aligned} H(\tilde{q}^{(N)}; p^{(N)}) \approx H(\tilde{q}_0; p_0) &+ (\delta t) \sum_{i=1}^{N} \sum_{\tilde{\boldsymbol{n}}} \tilde{q}_{i-1}^{(N)}(\tilde{\boldsymbol{n}} \,|\, q_0) \left(\tilde{q}_{\leftarrow \tilde{\boldsymbol{n}}}(t^{(i)}) - p_{\leftarrow \tilde{\boldsymbol{n}}}(t^{(i)})\right) \\ &- (\delta t) \sum_{i=1}^{N} \sum_{\tilde{\boldsymbol{n}}} \tilde{q}_{i-1}^{(N)}(\tilde{\boldsymbol{n}} \,|\, q_0) \sum_{\tilde{\boldsymbol{m}} \neq \tilde{\boldsymbol{n}}} \tilde{q}_{\tilde{\boldsymbol{m}} \leftarrow \tilde{\boldsymbol{n}}}(t^{(i)}) \log p_{\tilde{\boldsymbol{m}} \,|\, \tilde{\boldsymbol{n}}}(t^{(i)}) - N \log(\delta t) + o(\delta t). \end{aligned}$$

which, after dropping the $N \log(\delta t)$ term and rearranging, yields Eq. (9)in the limit.

C Likelihoods of Reduced Trajectories

Let q be a continuous-time Markov chain defined on the discrete state space $\mathcal{X}$ with initial distribution q_0, and let $\tilde{q}$ be its projection onto $\tilde{\mathcal{X}}$. Computing the log probability of a trajectory $\tilde{\boldsymbol{n}}_{[0,T]}$ under $\tilde{q}$ requires integrating over all possible full trajectories $\boldsymbol{n}_{[0,T]} = (\tilde{\boldsymbol{n}}_{[0,T]}, z_{[0,T]})$ that are compatible with $\tilde{\boldsymbol{n}}_{[0,T]}$, that is,

$$\tilde{q}(\tilde{\boldsymbol{n}}_{[0,T]}) = \int_{z_{[0,T]}} q(\tilde{\boldsymbol{n}}_{[0,T]}, z_{[0,T]}) \mathrm{d}z_{[0,T]}. \tag{C1}$$

In order to compute the integral on the right-hand side we consider the time-discretisations of q and $\tilde{q}$. The marginal likelihood in this case can be obtained using the well-known forward algorithm for HMMs, which sequentially computes the joint probabilities $q^{(N)}(\tilde{\boldsymbol{n}}_{0:i}, z_i)$:

$$q^{(N)}(\tilde{\boldsymbol{n}}_0, z_0) = q_0(\tilde{\boldsymbol{n}}_0, z_0), \tag{C2}$$

$$q^{(N)}(\tilde{\boldsymbol{n}}_{0:i+1}, z_{i+1}) = \sum_z q^{(N)}(\tilde{\boldsymbol{n}}_{0:i}, z_i = z)\, q^{(N)}(\tilde{\boldsymbol{n}}_{i+1}, z_{i+1} \,|\, \tilde{\boldsymbol{n}}_i, z_i = z). \tag{C3}$$

Using (B2), (B3) we can write the second equation as

$$\begin{aligned} q^{(N)}(\tilde{\boldsymbol{n}}_{0:i+1}, z_{i+1}) = q^{(N)}(\tilde{\boldsymbol{n}}_{0:i}, z_i = z_{i+1}) \left(1 - (\delta t)\, q_{\leftarrow(\tilde{\boldsymbol{n}}_i, z_{i+1})}(t^{(i)})\right) \\ + (\delta t) \sum_{z \neq z_{i+1}} q^{(N)}(\tilde{\boldsymbol{n}}_{0:i}, z_i = z)\, q_{(\tilde{\boldsymbol{n}}_i, z_{i+1}) \leftarrow (\tilde{\boldsymbol{n}}_i, z)}(t^{(i)}) + o(\delta) \qquad (\boldsymbol{n}_{i+1} = \boldsymbol{n}_i), \end{aligned} \tag{C4}$$

$$q^{(N)}(\tilde{\boldsymbol{n}}_{0:i+1}, z_{i+1}) = (\delta t) \sum_z q^{(N)}(\tilde{\boldsymbol{n}}_{0:i}, z_i = z)\, q_{(\tilde{\boldsymbol{n}}_{i+1}, z_{i+1}) \leftarrow (\tilde{\boldsymbol{n}}_i, z)}(t^{(i)}) + o(\delta) \qquad (\boldsymbol{n}_{i+1} \neq \boldsymbol{n}_i). \tag{C5}$$

Here the two cases correspond to no visible jump and a visible jump occurring at the i-th step, respectively. Eq. (C4) can be reorganised as

$$\begin{aligned} q^{(N)}(\tilde{\boldsymbol{n}}_{0:i+1}, z_{i+1}) = q^{(N)}(\tilde{\boldsymbol{n}}_{0:i}, z_i = z_{i+1}) + (\delta t) \sum_{z \neq z_{i+1}} q^{(N)}(\tilde{\boldsymbol{n}}_{0:i}, z_i = z)\, q_{(\tilde{\boldsymbol{n}}_i, z_{i+1}) \leftarrow (\tilde{\boldsymbol{n}}_i, z)}(t^{(i)}) \\ - (\delta t) \sum_{z \neq z_{i+1}} q^{(N)}(\tilde{\boldsymbol{n}}_{0:i}, z_i = z_{i+1})\, q_{(\tilde{\boldsymbol{n}}_i, z) \leftarrow (\tilde{\boldsymbol{n}}_i, z_{i+1})}(t^{(i)}) \\ - (\delta t) \sum_{\boldsymbol{n} \neq \boldsymbol{n}_{i+1}} \sum_z q^{(N)}(\tilde{\boldsymbol{n}}_{0:i}, z_i = z_{i+1})\, q_{(\tilde{\boldsymbol{n}}, z) \leftarrow (\tilde{\boldsymbol{n}}_i, z_{i+1})}(t^{(i)}) + o(\delta), \end{aligned} \tag{C6}$$

Here the first and second sums correspond to hidden reactions that do not change $\tilde{n}$ and the last double represents visible reactions that affect $\tilde{n}$. Expressing the transition rates for the visible reactions in terms of the reaction propensities σ_j and taking the continuum limit $\delta t \to 0$ now yields Eqs. (22)-(24).

To arrive at Eqs. (26)-(28) we sum Eqs. (22)-(24) over all z. Marginalising Eq. (22) immediately yields Eq. (26), and for the other two we obtain

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \sum_{z_t} q(\tilde{\boldsymbol{n}}_{[0,t]}, z_t) = \sum_{z_t} \sum_{z' \neq z_t} \left(q_{(\tilde{\boldsymbol{n}}_t, z_t) \leftarrow (\tilde{\boldsymbol{n}}_t, z')}\, q(\tilde{\boldsymbol{n}}_{[0,t]}, z') - q_{(\tilde{\boldsymbol{n}}_t, z') \leftarrow (\tilde{\boldsymbol{n}}_t, z_t)}\, q(\tilde{\boldsymbol{n}}_{[0,t]}, z_t)\right) \\ - \sum_{i \text{ vis.}} \sum_{z_t} \sigma_i(\tilde{\boldsymbol{n}}_t, z_t)\, q(\tilde{\boldsymbol{n}}_{[0,t]}, z_t), \end{aligned} \tag{C7}$$

$$\lim_{t \searrow t_k} \sum_{z_t} q(\tilde{\boldsymbol{n}}_{[0,t]}, z_t) = \lim_{t \nearrow t_k} \sum_{z_t} \sigma_{j_k}(\tilde{\boldsymbol{n}}_k, \tilde{z}_t)\, q(\tilde{\boldsymbol{n}}_{[0,t]}, z'). \tag{C8}$$

The first double sum in Eq.(C7), corresponding to all hidden reactions, vanishes. Write the joint distributions in terms of conditional distributions as

$$q_t(\tilde{n}_{[0,t]}, z) = q_t(\tilde{n}_{[0,t]})\, q_t(z \,|\, \tilde{n}_{[0,t]}, q_0) \tag{C9}$$

where we explicitly include the dependence on the initial distribution. We arrive at

$$\frac{\mathrm{d}}{\mathrm{d}t} q(\tilde{\boldsymbol{n}}_{[0,t]}) = -\sum_{i \text{ vis.}} \sum_{z_t} \sigma_i(\tilde{\boldsymbol{n}}_t, z_t)\, q(z_t \,|\, \tilde{\boldsymbol{n}}_{[0,t]}, q_0)\, q(\tilde{\boldsymbol{n}}_{[0,t]}), \tag{C10}$$

$$\lim_{t \searrow t_k} q(\tilde{\boldsymbol{n}}_{[0,t]}) = \lim_{t \nearrow t_k} \sum_{z_t} \sigma_{j_k}(\tilde{\boldsymbol{n}}_k, \tilde{z}_t)\, q(z' \,|\, \tilde{\boldsymbol{n}}_{[0,t]}, q_0)\, q(\tilde{\boldsymbol{n}}_{[0,t]}). \tag{C11}$$

We recognise the sum over z_t as a conditional expectation:

$$\frac{\mathrm{d}}{\mathrm{d}t} q(\tilde{\boldsymbol{n}}_{[0,t]}) = -\sum_{i \text{ vis.}} \mathbb{E}_{z_t}\left[\sigma_i(\tilde{\boldsymbol{n}}(t), z_t) \,|\, \tilde{\boldsymbol{n}}_{[0,t]}, q_0; t\right]\, q(\tilde{\boldsymbol{n}}_{[0,t]}), \tag{C12}$$

$$\lim_{t \searrow t_k} q(\tilde{\boldsymbol{n}}_{[0,t]}) = \lim_{t \nearrow t_k} \mathbb{E}_{z_t}\left[\sigma_i(\tilde{\boldsymbol{n}}(t), z_t) \,|\, \tilde{\boldsymbol{n}}_{[0,t]}, q_0; t\right]\, q(\tilde{\boldsymbol{n}}_{[0,t]}). \tag{C13}$$

The two equations rearrange to yield (27) and (28).

D Marginal distributions after reduction

In this section we compute the marginal distributions of a projected reaction network $\tilde{q}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{q}_t(\tilde{\boldsymbol{n}}) = \frac{\mathrm{d}}{\mathrm{d}t}\sum_z q_t(\tilde{\boldsymbol{n}}, z) = \sum_i \sum_z \left(\sigma_i(\tilde{\boldsymbol{n}} - \tilde{\boldsymbol{S}}_i, z - \boldsymbol{S}_i^z)\, q_t(\tilde{\boldsymbol{n}} - \tilde{\boldsymbol{S}}_i, z - \boldsymbol{S}_i^z) - \sigma_i(\boldsymbol{n})\, q_t(\tilde{\boldsymbol{n}}, z)\right) \tag{D1}$$

$$= \sum_i \sum_z \left(\sigma_i(\tilde{\boldsymbol{n}} - \tilde{\boldsymbol{S}}_i, z - \boldsymbol{S}_i^z)\, q_t(z - \boldsymbol{S}_i^z \,|\, \tilde{\boldsymbol{n}} - \tilde{\boldsymbol{S}}_i)\, \tilde{q}_t(\tilde{\boldsymbol{n}} - \tilde{\boldsymbol{S}}_i) - \sigma_i(\boldsymbol{n})\, q_t(z \,|\, \tilde{\boldsymbol{n}})\, \tilde{q}_t(\tilde{\boldsymbol{n}})\right) \tag{D2}$$

$$= \sum_i \left(\mathbb{E}\left[\sigma_i(\tilde{\boldsymbol{n}} - \tilde{\boldsymbol{S}}_i, z - \boldsymbol{S}_i^z) \,|\, \tilde{\boldsymbol{n}} - \tilde{\boldsymbol{S}}_i; t\right]\, \tilde{q}_t(\tilde{\boldsymbol{n}} - \tilde{\boldsymbol{S}}_i) - \mathbb{E}\left[\sigma_i(\boldsymbol{n}) \,|\, \tilde{\boldsymbol{n}}; t\right]\, \tilde{q}_t(\tilde{\boldsymbol{n}})\right). \tag{D3}$$

This corresponds exactly to the CME for the reduced system with propensities given by Eq. (14). As a result, the marginal distributions of the optimal reduction p and $\tilde{q}$ agree at all times.