Hybrid Stochastic–Deterministic Approaches for Simulation and Analysis of Biochemical Reaction Networks

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von

Stephan Menz

Berlin 2013

Erstgutachter: Prof. Dr. Christof Schütte
Zweitgutachter: Prof. Dr. Tobias Jahnke

Tag der Disputation: 13. Dezember 2012

Contents

List of Figures

List of Tables

1

Introduction

Mathematical modeling has become an indispensable tool for the description and systematic analysis of biological phenomena [1–4]. Biologists have revealed fundamental intelligence about living systems by experimentally studying their structure and physiology in increasing detail. But as the scope and depth of information virtually explodes, understanding the intricate dynamical properties of (micro-)biological processes becomes more complicated. Mathematical modeling offers a formal language to express such complex biological behavior, to test assumptions and hypothesis about the functioning of biological systems with the precision of mathematics [5]. There exist different classes of mathematical modeling approaches to describe the dynamics of biochemical reaction networks [4]: qualitative (e.g., by Bayesian or Boolean networks), quantitative (based on deterministic or stochastic reaction kinetics), and combined qualitative–quantitative approaches. Herein we deal with quantitative approaches, being routinely used, for instance, to model and analyze metabolic pathways [6, 7], protein interactions [8], or gene expression and regulatory networks [9].

Deterministic Approach: Traditionally, a quantitative approach is based on the macroscopic view of chemical kinetics, where reaction processes are modeled by *deterministic* rate equations. It assumes that the quantity of a species S_i $(i = 1, \ldots, N)$ evolves as a *continuous* variable $\boldsymbol{x}_i$ in time t according to a set of coupled ordinary differential equations (ODEs) of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_i = f_i(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \qquad (i = 1, \ldots, N).$$

Usually, such *continuous–deterministic* models are expressed in terms of molar concentrations $\boldsymbol{x}_i = \boldsymbol{X}_i/\Omega$, where $\boldsymbol{X}_i$ is the number of molecules of species S_i and Ω denotes the system volume times the Avogadro constant. The functions f_i are inferred from the stoichiometries and assumed rates of the reactions in a given network [10]. The fundamental rate model in deterministic reaction kinetics is the law of mass action, which has further led to a number of important rate models in enzyme kinetics, such as the Michaelis–Menten equation or the Hill equation, cf. [4, 7]. A deterministic formulation can easily be extended to account

for the spatial organization of biological systems and possible inhomogeneities in species concentrations, e.g., by partitioning the system into spatial compartments or by combining the reaction kinetics model with diffusion equations [4]. In this way, deterministic models have been successfully used to describe and analyze reaction processes not only in chemistry, but also in biochemistry and systems biology [4, 7, 11].

On a microscopic level, however, chemical reactions can be regarded as *discrete, random* events resulting from collisions between individual molecules [4, 12, 13]. Hence, the assumption of *continuously* changing concentrations in deterministic reaction kinetics becomes clearly inaccurate if the number of molecules of some species are low. Moreover, variation in the temporal order of reaction events leads to *fluctuations* in the number of molecules around the average species levels. These inherent stochastic fluctuations (also called intrinsic noise) arise from the lack of total predictability in molecular dynamics and quantum indeterminacy [10, 14], and are different from extrinsic noise caused by environmental changes or interference with other processes [15, 16]. For low species levels, stochastic fluctuations become significant and can affect the qualitative behavior of the system [17, 18]. In a small number of molecules regime, a continuous–deterministic model thus constitutes a substantial simplification and might not accurately describe the true system dynamics [10].

Stochastic Approach: There is considerable experimental evidence that stochastic effects play a crucial role in many cellular processes like gene expression and regulation [19–24], where constituents are typically present in small numbers. For this reason, the last decade has witnessed an increasing interest in stochastic descriptions of biochemical systems. Stochastic reaction kinetics accounts for the inherent fluctuations in the discrete number of molecules by modeling the state of the system $\boldsymbol{X} = (\boldsymbol{X}_1(t), \ldots, \boldsymbol{X}_N(t))^{\mathrm{T}}$ as a continuous-time, discrete-state Markov jump process. By introducing a system compartmentalization, such *discrete–stochastic* models can also be extended to account for spatial inhomogeneities [4]. The fundamental equation of stochastic reaction kinetics is the chemical master equation (CME) that defines the temporal evolution of the probability density function (PDF) of the system state. Unfortunately, only few approaches exist to *directly* solve the CME for a general system [25–29]. The main problem is that the state space grows exponentially with the number of species, which renders most direct approaches computationally infeasible for larger reaction networks.

Indirect Methods: There exist a number of *approximate solution* techniques to the CME, e.g., [30–35]. Instead of solving for the PDF directly, however, most of these methods are

based on Monte Carlo (MC)-simulations of the Markov jump process underlying the CME [30–32]. Henceforth, we call MC-based methods *indirect*. The main advantage of indirect methods is that they are easy to apply. With Gillespie's stochastic simulation algorithm (SSA) [31], for instance, one can compute a statistically exact realization by iteratively generating two random numbers in order to determine when and which will be the next reaction that occurs in the system. However, indirect methods also inherit the common disadvantages of MC-based approaches: As the PDF is *approximated* by a statistical ensemble of realizations, there is always a sampling error introduced. Depending on the quantities of interest (expectations, higher moments or the PDF itself), the number of realizations n required to meet a certain accuracy might be very large. Convergence is rather slow with the sampling error decaying like $C \cdot n^{-1/2}$ [36, 37]. In addition, the constant C can be exponentially large if the system exhibits switching behavior [38].[1]

Apart from the problem of judging the required number of realizations to build-up a sufficient statistics, the computational costs of an *exact* realization is dictated by the total number of reaction events that have to be simulated [10, 31]. This renders exact indirect methods like the SSA numerically infeasible when applied to systems that include many rapidly firing reactions or species present in large numbers. For this reason, it is often necessary to sacrifice some of the accuracy of a discrete–stochastic system description to obtain computationally faster indirect methods.

A common approach to accelerate stochastic simulations is based on a τ-leap condition [33, 34, 39–43] that allows to approximate multiple reaction events as independent Poisson or binomial random variables. If the τ-leap condition is satisfied and all reactions are sufficiently fast, the discrete Markov jump process might also be approximated by a continuous Markov process, whose evolution is described by the chemical Langevin equation (CLE) [33]. Although such continuous–stochastic models capture the effects of stochastic fluctuations [18, 44], their simulation is usually performed indirect and still computationally expensive. Furthermore, the underlying approximation might result in a misdescription of the true system dynamics [18, 45]. Another approach is based on a kind of quasi-steady state approximation (QSSA). QSSA-based methods [46–52] rely on the assumption that variables associated with fast reactions approach a quasi-stationary PDF, cf. [10]. The state changes resulting from fast reactions are then approximated via their quasi-stationary PDFs and the system is propagated by effectively simulating only the slow, unproblematic reactions.

[1] The example in [38] is a Hamiltonian system, where switching times increase exponentially with decreasing system temperature.

Hybrid Approaches: A different strategy rests upon the idea of employing a *hybrid* system representation, i.e., approximating fast reactions associated with large number of molecules as continuous processes, either in a stochastic [50, 53–56] or deterministic [50, 53, 57–62] context, while simulating reactions that are not suited for a continuous approximation as discrete stochastic processes. Hybrid approaches circumvent the problems related to a purely discrete or purely continuous system representation by leveraging the advantages of both approaches: They capture relevant stochastic effects arising from fluctuations in small number of molecules, while reducing the computational complexity substantially. In contrast to QSSA-based approaches, hybrid approaches do not depend on the existence of quasi-stationary distributions of the fast modes. Furthermore, they can easily be generalized in application.

In this thesis, we are concerned with hybrid approaches that couple the *discrete–stochastic* formulation of stochastic reaction kinetics with the *continuous–deterministic* formulation of classical reaction kinetics. Such hybrid stochastic–deterministic approaches are motivated by the limit behavior of the stochastic process that underlies the CME. As shown by T. G. Kurtz [63], on compact time intervals, the deterministic process of classical reaction kinetics approximates the stochastic process in the thermodynamic limit, i.e., the number of molecules of all species and the volume of the system approach infinity ($\boldsymbol{X}_i \to \infty$, $\Omega \to \infty$), while the species concentrations converge to some finite value $\boldsymbol{x}_i = \boldsymbol{X}_i/\Omega$. It further turns out that the reaction intensities grow linearly as the system approaches the thermodynamic limit [14, 64]. Hence, if species are present in large numbers and reactions are fast, deterministic reaction kinetics is a good approximation to the CME. This well-known property is one of the key facts exploited in hybrid stochastic–deterministic approaches [57–62].

Objective: In this thesis, we demonstrate that hybrid approaches are well-suited for the modeling of large reaction networks that exhibit fluctuations due to some species present in small numbers. Current hybrid methods [50, 53–62], however, almost exclusively rely on MC-simulations of the Markov process associated with the discretely modeled subsystem. They are hence *indirect* and suffer from the aforementioned disadvantages of indirect methods.

The main objective of this thesis is to exploit the concept of hybrid approaches in order to derive a *direct* solution of the CME. The idea of our direct hybrid approach is to apply only a partial limit to such species that are present in large quantities. Since a partial volume limit is hard to justify for obvious reasons, we pursue a multi-scale expansion approach with respect to a scaling parameter $\varepsilon \ll 1$. We link this 'artificial' parameter ε to large species levels and fast reactions, such that it plays a similar role as Ω^{-1} in the thermodynamic limit.

Based on a WKB-ansatz and Laplace's method [65], we are then able to derive a system of coupled evolution equations for the PDF of species present in small numbers and the related expectations of species present in large numbers. As this effectively reduces the CME to the small copy number subspace, its direct numerical solution (without MC-simulations) becomes feasible whenever there are few species in low quantities present in the system. In contrast to indirect hybrid methods, the impact of changes of the discrete *distribution* on the dynamics of the continuous variables then has to be taken into account explicitly.

A first approach to directly solve the CME based on a partitioning of the state space was proposed by Henzinger et al. [66]. In their approach a coupling is realized in two separate steps: (i) propagation of the discrete distribution and the continuous variables for a small time step; (ii) distribution of the continuous variables according to the changes in the discrete distribution. In contrast, we present a closed hybrid approach that implicitly integrates these propagation and distribution steps continuously. A similar approach has recently been proposed by T. Jahnke [67]. Based on an error analysis of two model reduction approaches that employ a product ansatz, identifying the direct product approximation of the CME solution as the main source of error, a model reduction by conditional expectations (MRCE) is proposed. The derivation of the MRCE-model is based on certain assumptions on the continuous variables, such as zero covariance. These assumptions become in our multi-scale formalism theoretically justifiable. Our approach thus also contributes to a deeper understanding of the MRCE-model. Interestingly, we further find that the direct contribution of slow reactions on the dynamics of the continuous variables is negligible compared to their indirect contribution that results from changes in the discrete distribution.

Even though a hybrid formulation reduces the computational complexity, it does not reduce the model complexity *per se*. Parameters of detailed biological models are often not available and have thus to be estimated by fitting the corresponding mathematical model to experimental data. But parameter estimation can only provide reliable results if the model complexity is in balance with the amount and quality of data [68]. If this is not the case, the model has to be simplified by exploiting reduction concepts like time-scale separation [69–71], sensitivity analysis [72, 73] or balanced truncation [74–76]. Often, a reduced model also provides further insight into features of the dynamic behavior of a reaction network [4].

We further study model reduction in the context of a detailed model of gene expression. As classical reaction kinetics does generally not allow to model gene expression on the level of single molecules (in particular a single gene), the underlying subprocesses of transcription and translation are usually not explicitly incorporated in a deterministic model, but implicitly

aggregated into effective protein synthesis rates [4, 77–79]. We propose a QSSA-like reduction approach that allows to directly link such effective rates to the transcriptional and translational processes given in a discrete–stochastic formulation of gene expression. As it turns out, the derived functional relations are of particular value when analyzing the sensitivity of a reduced deterministic model with respect to effective synthesis rates.

Outline: We begin by setting the framework and providing the background on stochastic and deterministic reaction kinetics in Chapter 2. The general concepts and relevant properties of both approaches are all reviewed, and, additionally, indirect methods for the simulation of discrete–stochastic models are recalled and discussed. We further study the relationship between deterministic and stochastic reaction kinetics in detail, including the important result by T. G. Kurtz for the thermodynamic limit. We also introduce an asymptotic approximation technique for the CME, providing an alternative approach to derive the ODE-model of deterministic reaction kinetics as the large-size system limit of stochastic reaction kinetics.

In Chapter 3, we outline and discuss indirect approaches for the hybrid simulation of biochemical reaction networks. We present a detailed derivation of a hybrid simulation algorithm, where reactions are dynamically partitioned into stochastic and deterministic processes. We apply this hybrid method to a mathematical model of the replication dynamics of the human immunodeficiency virus (HIV). Based on the hybrid simulation results, we are able to design and validate *in silico* a novel treatment strategy for HIV-infected patients that can lead to significant improvements compared to conventional treatment strategies.

In Chapter 4, we use the asymptotic techniques introduced in Chapter 2 to derive a novel hybrid stochastic–deterministic approach to solve the CME directly. Our direct hybrid approach is capable of reducing the number of degrees of freedom of the CME-description substantially. It does not suffer from the disadvantages of indirect methods studied in the previous chapters and is more efficient if the reaction system includes only a few species associated with low number of molecules. We illustrate the performance of our direct hybrid approach by applying it to model systems of biological interest.

In Chapter 5, we use a detailed stochastic model of gene expression to derive effective protein synthesis rates as typically incorporated in deterministic models of biochemical systems. The model reduction approach we use effectively eliminates the transcriptional and translational subprocesses and discloses their functional relationship to the effective synthesis rates. We demonstrate and validate our approach on a protein interactions model of the flagellar gene regulation cascade in *Escherichia coli* (*E. coli*). We further perform a sensitivity

analysis with respect to the parameters in the resulting deterministic formulation of this system. Based on the derived functional relationships, we find that sensitivity with respect to effective rates does not directly carry over to the aggregated detailed parameters, but breaks down into more diverse relations.

In Chapter 6, we conclude by summarizing the presented results and outlining possible directions for future work.

Acknowledgements: Many people supported me during the last years of my research. First of all, I would like to thank Christof Schütte and Wilhelm Huisinga not only for introducing me to this fascinating field of research and offering me the possibility of writing this thesis, but also for their continuous support. My gratitude also goes to Max von Kleist for sharing his passion on virus research, Juan Cristobal Latorre for his important input and ideas on the asymptotics of the CME, as well as Regine Hengge for sharing her deep knowledge about the molecular interactions in *E. coli*.

I had the chance to visit several research groups and would like to thank all the people there who always made me feel welcome and enjoy science in a stimulating atmosphere, namely, the members of the Biocomputing Group at the Freie Universität Berlin, the Computational Physiology Group at the University of Potsdam, all people from the Hamilton Institute at the National University of Ireland, Maynooth, the participants and assistants of the GRT-program PharMetrX: Pharmacometrics & Computational Disease Modelling, as well as Charlotte Kloft and her members at the Department of Clinical Pharmacy and Biochemistry at the Freie Universität Berlin.

I am also very grateful to Peter Pelikan, Sathej Gopalakrishnan and Surya Shravan Sajja Kumar for proofreading parts of this thesis and providing valuable comments and suggestions.

And finally, I have to thank Anne for all the support and love she gave me during the last years.

This work was financially supported by DFG funding, provided through the Dahlem Research School of the Freie Universität Berlin.

2

Modeling of Biochemical Reaction Networks

In this chapter, we provide the necessary background for the reader unfamiliar with the stochastic and deterministic modeling approaches. We first review the general description of biochemical reaction networks as chemical systems, see Section 2.1. In Sections 2.2 and 2.3, we study both formulations of reaction kinetics, the stochastic and the deterministic, their underlying assumptions and properties. In case of the stochastic formulation, we additionally review basic numerical simulation methods. We conclude the chapter by discussing the relationship between stochastic and deterministic reaction kinetics in detail in Section 2.4.

2.1 The Chemical System

Consider a chemical system of constant volume V that is homogeneous, especially well-stirred, and in thermal equilibrium at a constant absolute temperature T. The system includes molecules of N chemical species S_i, $i = 1, \ldots, N$, which interact through M elementary reactions R_μ, $\mu = 1, \ldots, M$.[1] Given its stoichiometry, we can write an elementary reaction generally as

$$R_\mu\colon \quad s^r_{\mu 1}\, S_1 + \cdots + s^r_{\mu N}\, S_N \;\to\; s^p_{\mu 1}\, S_1 + \cdots + s^p_{\mu N}\, S_N \quad (\mu = 1, \ldots, M),$$

where $s^r_{\mu i}$ and $s^p_{\mu i} \in \mathbb{N}_0$ denote the *stoichiometric coefficients*, specifying how many molecules of S_i react and how many molecules of S_i get produced with a single firing of R_μ, respectively. Hence, we associate a vector $\xi_\mu := (\xi_{\mu 1}, \ldots, \xi_{\mu N})^{\mathrm{T}} \in \mathbb{Z}^N$ to each reaction R_μ with

$$\xi_{\mu i} := (s^p_{\mu i} - s^r_{\mu i}) \quad (i = 1, \ldots, N),$$

that gives the *net changes* in the number of molecules of the species caused by a single firing of R_μ. We denote the possible number of molecules of a species S_i by $X_i \in \mathbb{N}_0$, $i = 1, \ldots, N$, such that $X = (X_1, \ldots, X_N)^{\mathrm{T}} \in \mathbb{N}_0^N$ gives a possible state of the system. Now, if R_μ is of the

[1] An elementary reaction is assumed to occur in a single reaction step with no reaction intermediates and to pass through a single transition state [80, 81].

form $2\,S_1 \longrightarrow S_2$, for instance, then $\xi_\mu = (-2, +1, 0, \ldots, 0)^{\mathrm{T}}$ and a firing of R_μ would change the state of the system from $X = (X_1, X_2, X_3, \ldots, X_N)^{\mathrm{T}}$ to $X = (X_1 - 2, X_2 + 1, X_3, \ldots, X_N)^{\mathrm{T}}$. Typically, the net changes ξ_μ of a chemical system are sparse vectors, as only a few species participate in a particular reaction R_μ.

In the following, we describe the system's dynamics as a continuous-time process[2] $\boldsymbol{X}(t) := (\boldsymbol{X}_1(t), \ldots, \boldsymbol{X}_N(t))^{\mathrm{T}}$ taking values in $\mathbb{N}_0^N$, where

$$\boldsymbol{X}_i(t) := \text{number of molecules of species } S_i \text{ at time } t \quad (i = 1, \ldots, N).$$

Assume the system is in state X_0 at time $t = 0$, i.e., $\boldsymbol{X}(0) = X_0$, and denote by $\boldsymbol{K}_\mu(t)$ the number of firings of a reaction R_μ up to time $t \geq 0$, with $\boldsymbol{K}_\mu(0) = 0$ for every $\mu = 1, \ldots, M$. According to the net changes ξ_μ, the state of the system at time $t \geq 0$ is then given as

$$\boldsymbol{X}(t) = \boldsymbol{X}(0) + \sum_{\mu=1}^{M} \xi_\mu \boldsymbol{K}_\mu(t),$$

with $\boldsymbol{X}(0) = X_0$. In principle, we are thus left with predicting the reaction counts $\boldsymbol{K}_\mu(t)$ for every $\mu = 1, \ldots, M$ and $t > 0$.

In the stochastic framework this is realized by modeling every $\boldsymbol{K}_\mu$ as a Poisson process that increases in discrete steps with intensity a_μ, called reaction propensity, see Section 2.2. Hence, $\boldsymbol{X}(t)$ is modeled as a continuous-time, discrete-state Markov jump process. In the deterministic setting, the reaction count $\boldsymbol{K}_\mu$ gets approximated by an average count that continuously increases with rate v_μ, see Section 2.3. Consequently, in the deterministic framework the system dynamics is approximated by a continuous (in time and state) process $\boldsymbol{x}(t)$ that takes values in $\mathbb{R}_{\geq 0}^N$. However, as one would probably expect, both rate functions, the reaction propensity a_μ and the reaction rate v_μ are found to have a similar form, which is determined by the specific type of a reaction R_μ, described below.

Elementary reactions R_μ can be classified regarding their *molecularity*

$$|s_\mu^r| := \sum_{i=1}^{N} s_{\mu i}^r,$$

that gives the number of molecules that react in a single firing of R_μ, see [80, 81]. In general, one distinguishes between: (1) unimolecular reactions with $|s_\mu^r| = 1$, (2) bimolecular reactions with $|s_\mu^r| = 2$ and (3) termolecular reactions with $|s_\mu^r| = 3$. Termolecular reactions (and reactions

[2] Henceforth, we denote every process or time-dependent variable by a bold symbol, while elements of the state space are set in normal font.

of higher molecularity), however, are usually not considered, as they rely on the physically improbable event of a simultaneous collision of three (or more) molecules [15, 82].[3] If the system is open, we additionally consider a type of reaction that models the possible influx of new molecules from an external source, i.e., $\varnothing \longrightarrow \cdots$.[4] We call such elementary reactions with $|s_\mu^r| = 0$ *source-like* reactions.

2.2 Stochastic Reaction Kinetics

In the stochastic formulation of reaction kinetics, the initial assumptions on the system are interpreted in a probabilistic manner [82]:

(a) *Well-stirred*: The position of a molecule is a uniformly distributed random variable, such that the probability of finding the center of a randomly selected molecule inside a subregion of volume δV is equal to $\delta V/V$.

(b) *Thermal equilibrium*: The velocity of a molecule is a random variable that follows the Maxwell-Boltzmann distribution, such that the probability of finding the velocity of a randomly selected molecule in the infinitesimal region $\mathrm{d}^3\vec{v}$ about velocity $\vec{v}$ is equal to $f_{MB}(\vec{v})\mathrm{d}^3\vec{v}$, where

$$f_{MB}(\vec{v}) = \left(\frac{m}{2\pi k_B T}\right)^{3/2} \exp\left\{-\frac{mv^2}{2k_B T}\right\} \tag{2.1}$$

is the Maxwell-Boltzmann distribution for the velocity $\vec{v} = (v_x, v_y, v_z)^{\mathrm{T}}$ of a particle with mass m. Here, $k_B \approx 1.38 \times 10^{-23}\,\mathrm{J/K}$ denotes the Boltzmann constant, and $\mathrm{d}^3\vec{v} = \mathrm{d}v_x\mathrm{d}v_y\mathrm{d}v_z$ and $v^2 = {v_x}^2 + {v_y}^2 + {v_z}^2$.

2.2.1 Reaction Propensity

Under the above premises, the fundamental hypothesis of stochastic reaction kinetics states that [82]: Given the system is in state X at time t, then (i) the probability of a single firing of reaction R_μ inside V in an infinitesimal time interval $[t; t + \delta t)$ is given by $a_\mu(X)\delta t + o(\delta t)$;

[3] Ter- and higher-molecular reactions are typically stepwise reactions that involve at least two consecutive elementary reactions with at least one reaction intermediate, see [15, 80, 81]. Therefore, they are not considered separately in the following.

[4] Here and in the following, the symbol '$\varnothing$' denotes the empty set.

and (ii) the probability of more than one reaction event inside V in $[t; t + \delta t)$ is equal to $o(\delta t)$, where $o(\delta t)$ denotes terms $f(\delta t)$ satisfying $f(\delta t)/\delta t \to 0$ as $\delta t \to 0$.[5] The function a_μ associated with each reaction R_μ, with

$$\begin{aligned} a_\mu(X)\delta t := {} & \text{probability, to first order in } \delta t, \text{ that reaction } R_\mu \text{ will occur once} \\ & \text{inside } V \text{ in the next infinitesimal time interval } [t; t + \delta t) \\ & \text{given that the system is in state } X \text{ at time } t, \end{aligned}$$

is called *reaction propensity* in the chemical literature [30, 82]. It follows that the unit of a reaction propensity a_μ is always equal to the inverse unit of time.

As shown by D. T. Gillespie in 1976 [30], the propensity a_μ of an elementary reaction R_μ is of the general form

$$a_\mu(X) = c_\mu h_\mu(X), \tag{2.2}$$

where c_μ denotes the *specific probability rate constant* of R_μ, and $h_\mu(X)$ gives the number of distinct combinations of R_μ reactant molecules available in state X.

Distinct Reactant Combinations

The number of distinct combinations of R_μ reactant molecules in a state X is given by the product of binomial coefficients as

$$h_\mu(X) := \prod_{i=1}^{N} \binom{X_i}{s_{\mu i}^r} = \begin{cases} \displaystyle\prod_{i=1}^{N} \frac{X_i!}{s_{\mu i}^r!(X_i - s_{\mu i}^r)!} & \text{if } X_i \geq s_{\mu i}^r \text{ for all } i = 1, \ldots, N, \\ 0 & \text{otherwise,} \end{cases} \tag{2.3}$$

where $n! := \prod_{k=1}^{n} k$ denotes the factorial of a non-negative integer n.[6] Hence, the propensity a_μ of an elementary reaction R_μ is a simple multivariate polynomial function of the number of molecules X_i, $i = 1, \ldots, N$.

Example 2.2.1. The bimolecular reaction R_{2a}: $S_i + S_j \longrightarrow \cdots$ (with $i \neq j$) has a propensity of the form $a_{2a}(X) = c_{2a} X_i X_j$, whereas the bimolecular reaction R_{2b}: $2\,S_i \longrightarrow \cdots$ has a propensity of the form $a_{2b}(X) = c_{2b} X_i(X_i - 1)/2$.

[5] Henceforth, we use the following Bachmann–Landau notations: Suppose f and g are two functions defined on some subset of the real numbers. We write $f(x) = o\big(g(x)\big)$ as $x \to x_0$ if and only if $\lim_{x \to x_0} f(x)/g(x) = 0$, i.e., f grows much slower than g as $x \to x_0$. Similarly, we use the notation $f(x) = \mathcal{O}\big(g(x)\big)$ as $x \to x_0$ if and only if there exists a constant C such that $\lim_{x \to x_0} |f(x)|/|g(x)| \leq C$, indicating that the rate of growth (i.e., the order) of f is not faster than that of g, cf. [65].

[6] Following the convention for empty products, we have $0! = 1$.

Specific Probability Rate Constant

The specific probability rate constant c_μ of a reaction R_μ is defined such that

$$\begin{aligned} c_\mu \delta t := {} & \text{probability that a particular combination} \\ & \text{of } R_\mu \text{ reactant molecules at time } t \text{ will react accordingly} \\ & \text{inside } V \text{ in the next infinitesimal time interval } [t; t + \delta t), \end{aligned}$$

see [82]. Thus, the unit of c_μ is always equal to the inverse unit of time.

Example 2.2.2 (see [82])**.** For bimolecular reactions, the probability $c_\mu \delta t$ is given by the product of (i) the probability that a randomly chosen combination of R_μ reactant molecules at time t will collide during $[t; t + \delta t)$, and (ii) the conditioned probability that a given collision is effective and the molecules will react according to R_μ. An expression for the first probability derives from premises (a) and (b) by considering the relative velocity of a selected combination of R_μ reactant molecules (see eq. (2.1)) and the sub-volume that is critical for their collision. This probability thus depends on the absolute temperature T and the volume V of the system. The second probability usually relates to an energy barrier E_a, called *activation energy*, that has to be overcome in order for R_μ to occur. The fraction of molecules with a kinetic energy greater than E_a depends on the absolute temperature T, and turns out to be equal to the familiar Arrhenius exponential form $\exp\{-E_a/(k_B T)\}$, with E_a given in molecular units.

For instance, if we assume the idealized case where the molecules of a species S_i can be regarded as hard spheres with mass m_i and radius r_i. Then, the rate constant c_{2a} of the bimolecular reaction R_{2a}: $S_i + S_j \longrightarrow \cdots$ is of the form

$$c_{2a} = \sqrt{\frac{8 k_B T}{\pi m_r}} \frac{\pi (r_i + r_j)^2}{V} \exp\left\{ -\frac{E_a}{k_B T} \right\},$$

where $m_r := m_i m_j / (m_i + m_j)$ denotes the reduced mass of S_i and S_j ($i \neq j$).

2.2.2 Representation with Poisson Processes

Below, we briefly review some basic properties of Poisson processes. It follows that the fundamental hypothesis of stochastic reaction kinetics essentially states that the number of firings $\boldsymbol{K}_\mu$ of a reaction R_μ can be represented by a Poisson process whose intensity is given by the reaction propensity a_μ.

Review of Poisson Processes

Remark 2.2.1 (cf. [83–85]). A Poisson process $\boldsymbol{\mathcal{P}}$ with intensity λ is a continuous-time counting process[7] that satisfies:

(i) $\boldsymbol{\mathcal{P}}(0) = 0$ almost surely, i.e., $\mathbb{P}\big[\boldsymbol{\mathcal{P}}(0) = 0\big] = 1$.

(ii) Its increments are independent and stationary.[8]

(iii) $\mathbb{P}\big[\boldsymbol{\mathcal{P}}(t+\delta t) - \boldsymbol{\mathcal{P}}(t) = 1\big] = \lambda\delta t + o(\delta t)$ as $\delta t \to 0$.

(iv) $\mathbb{P}\big[\boldsymbol{\mathcal{P}}(t+\delta t) - \boldsymbol{\mathcal{P}}(t) > 1\big] = o(\delta t)$ as $\delta t \to 0$.

Since $\sum_{k=0}^{\infty} \mathbb{P}[\boldsymbol{\mathcal{P}}(t+\delta t) = k] \equiv 1$, it immediately follows from properties (iii) and (iv) that further $\mathbb{P}[\boldsymbol{\mathcal{P}}(t+\delta t) - \boldsymbol{\mathcal{P}}(t) = 0] = 1 - \lambda\delta t + o(\delta t)$ as $\delta t \to 0$. The expected number of events in an infinitesimal time interval $[t; t+\delta t)$ is given by

$$\mathbb{E}\big[\boldsymbol{\mathcal{P}}(t+\delta t) - \boldsymbol{\mathcal{P}}(t)\big] := \sum_{n=0}^{\infty} n\mathbb{P}\big[\boldsymbol{\mathcal{P}}(t+\delta t) - \boldsymbol{\mathcal{P}}(t) = n\big] = \lambda\delta t + o(\delta t), \tag{2.4}$$

as $\delta t \to 0$. This shows that the intensity λ of a Poisson process $\boldsymbol{\mathcal{P}}$ can be understood as the expected number of events per unit time, i.e., we can intuitively think of λ as a rate.

If $\lambda \equiv 1$, then $\boldsymbol{\mathcal{P}}$ is called a unit Poisson process. If $\boldsymbol{\mathcal{P}}$ is a unit Poisson process, then $\boldsymbol{\mathcal{P}}_\lambda(t) := \boldsymbol{\mathcal{P}}(\lambda t)$ is a Poisson process with intensity λ. So far, we considered λ to be constant in time, in which case $\boldsymbol{\mathcal{P}}_\lambda$ is called a *homogeneous* Poisson process. If the intensity of a Poisson process is not constant, e.g.,

$$\boldsymbol{\mathcal{P}}_\lambda(t) := \boldsymbol{\mathcal{P}}\left(\int_0^t \lambda(s)\,\mathrm{d}s\right), \tag{2.5}$$

then $\boldsymbol{\mathcal{P}}_\lambda$ is a *non-homogeneous* Poisson process with intensity $\lambda(t)$.[9] The cumulative intensity $\Lambda(t) := \int_0^t \lambda(s)\,\mathrm{d}s$ in eq. (2.5) is a dimensionless quantity that can be considered as the *internal time* of $\boldsymbol{\mathcal{P}}$, as it links the absolute time t to the amount of time that has passed for $\boldsymbol{\mathcal{P}}$, see [86]. Thus, problems concerning non-homogeneous Poisson processes can be reduced to the homogeneous case by utilizing time changes like eq. (2.5), cf. [87].

[7] A counting process is an integer-valued stochastic process $\{\boldsymbol{N}(t), t \geq 0\}$, satisfying: (a) $\boldsymbol{N}(t) \geq 0$ for all $t \geq 0$; (b) $\boldsymbol{N}(t)$ is non-decreasing with t, i.e., if $t \geq s$ then $\boldsymbol{N}(t) \geq \boldsymbol{N}(s)$; and (c) $\boldsymbol{N}(t)$ is right-continuous, such that $\boldsymbol{N}(t) - \boldsymbol{N}(s)$ represents the number of events in $(s; t]$, where $t > s \geq 0$.

[8] A counting process $\boldsymbol{N}$ has *independent increments* if the number of events that occur in disjoint time intervals are independent, i.e., if $t_0 < t_1 < t_2 < \cdots$ then $\boldsymbol{N}(t_k) - \boldsymbol{N}(t_{k-1})$, $k = 1, 2, \ldots$, are independent random variables. $\boldsymbol{N}$ has *stationary increments* if $\boldsymbol{N}(t+s) - \boldsymbol{N}(t)$ does not depend on t but only on $s > 0$.

[9] The increments of a non-homogeneous Poisson process are in general not stationary.

By property (ii), the increment $\boldsymbol{\mathcal{P}}(t+s) - \boldsymbol{\mathcal{P}}(t)$ for any $s > 0$ after time t is independent of the past $\boldsymbol{\mathcal{P}}(u)$, with $0 \le u \le t$. Thus, the future $\boldsymbol{\mathcal{P}}(t+s)$ depends upon the past only through the present value $\boldsymbol{\mathcal{P}}(t)$, which shows that a Poisson processes is Markov jump processes.

Definition 2.2.1. A continuous-time stochastic process $\boldsymbol{X}(t)$ is a *Markov process* if for all $0 < s < t$ the probability distribution of $\boldsymbol{X}(t)$ given $\boldsymbol{X}(u)$, with $0 \le u \le s$, depends only upon $\boldsymbol{X}(s)$, i.e., for all $0 \le t_0 < t_1 < \cdots < t_n < s$ with arbitrary $n \in \mathbb{N}$ and every Borel measurable event X,

$$\mathbb{P}[\boldsymbol{X}(t) = X \,|\, \boldsymbol{X}(t_0), \boldsymbol{X}(t_1), \ldots, \boldsymbol{X}(t_n), \boldsymbol{X}(s)] = \mathbb{P}[\boldsymbol{X}(t) = X \,|\, \boldsymbol{X}(s)].$$

A Markov process is called *homogeneous*, if $\mathbb{P}[\boldsymbol{X}(t) = X \,|\, \boldsymbol{X}(s)]$ depends upon s and t only through the difference $t - s > 0$. If the state space of the Markov process is discrete (countable), it is also called a *Markov jump process*.

The following results on Poisson processes can be found in most elementary books on probability and stochastic processes, e.g. [83–85]. For convenience, we consider only the homogenous case.

Theorem 2.2.1. *Let $\boldsymbol{\mathcal{P}}$ be a Poisson process with intensity λ, then $\boldsymbol{\mathcal{P}}(t) \sim Pois(\lambda t)$, i.e.,*

$$\mathbb{P}[\boldsymbol{\mathcal{P}}(t) = n] = \mathrm{e}^{-\lambda t} \frac{(\lambda t)^n}{n!},$$

for all $t \ge 0$ and $n \in \mathbb{N}_0$.[10]

Theorem 2.2.2. *Let $\boldsymbol{\mathcal{P}}$ be a Poisson process with intensity $\lambda > 0$ and denote the successive occurrence times of events by $t_0 = 0 < t_1 < t_2 < \cdots$. Then, the inter-arrival times $\boldsymbol{T}_i := t_i - t_{i-1}$, $i = 1, 2, \ldots$, are independent and identically distributed (iid) random variables with $\boldsymbol{T}_i \sim Exp(\lambda)$, i.e.,*

$$\mathbb{P}[\boldsymbol{T}_i > \tau] = \mathrm{e}^{-\lambda\tau},$$

for every $\tau \ge 0$.

Note. From Theorem 2.2.1 it follows that the expectation and the variance in the number of events during $(t; t+\tau]$ are both equal to $\lambda\tau$, i.e., $\mathbb{E}[\boldsymbol{\mathcal{P}}(\tau)] = \lambda\tau$ and $\mathrm{Var}(\boldsymbol{\mathcal{P}}(\tau)) = \lambda\tau$, respectively. By Theorem 2.2.2, the inter-arrival time $\boldsymbol{T}_i$ of a Poisson process with intensity λ has expectation $\mathbb{E}[\boldsymbol{T}_i] = 1/\lambda$ and variance $\mathrm{Var}(\boldsymbol{T}_i) = 1/\lambda^2$.

[10] Henceforth, we define $0^0 := 1$.

Representation of Reaction Counts

In terms of changes in the reaction counts $\boldsymbol{K}_\mu$ during an infinitesimal time interval $[t; t+\delta t)$, the fundamental hypothesis of stochastic reaction kinetics states that for every $\mu = 1, \dots, M$

$$\mathbb{P}\big[\boldsymbol{K}_\mu(t+\delta t) - \boldsymbol{K}_\mu(t) = 1\big] = a_\mu\big(\boldsymbol{X}(t)\big)\delta t + o(\delta t),$$

and further

$$\begin{aligned}\mathbb{P}\big[\boldsymbol{K}_\mu(t+\delta t) - \boldsymbol{K}_\mu(t) > 1\big] &= o(\delta t),\\ \mathbb{P}\big[\boldsymbol{K}_\mu(t+\delta t) - \boldsymbol{K}_\mu(t) > 0, \boldsymbol{K}_j(t+\delta t) - \boldsymbol{K}_j(t) > 0\big] &= o(\delta t), \quad \text{for} \quad j \neq \mu,\end{aligned}$$

as $\delta t \to 0$. Now, let $\boldsymbol{\mathcal{P}}_\mu$, $\mu = 1, \dots, M$, be independent, unit Poisson processes. Since every propensity $a_\mu(\boldsymbol{X}(t))$ is constant until the next reaction takes place and $\boldsymbol{X}(t)$ changes, we thus have that for given $\boldsymbol{X}(s)$, $s \leq t$, and sufficiently small δt

$$\mathbb{P}\left[\boldsymbol{\mathcal{P}}_\mu\left(\int_0^{t+\delta t} a_\mu\big(\boldsymbol{X}(s)\big)\,\mathrm{d}s\right) - \boldsymbol{\mathcal{P}}_\mu\left(\int_0^{t} a_\mu\big(\boldsymbol{X}(s)\big)\,\mathrm{d}s\right) = 1\right] = a_\mu\big(\boldsymbol{X}(t)\big)\delta t + o(\delta t),$$

and

$$\begin{aligned}&\mathbb{P}\left[\boldsymbol{\mathcal{P}}_\mu\left(\int_0^{t+\delta t} a_\mu\big(\boldsymbol{X}(s)\big)\,\mathrm{d}s\right) - \boldsymbol{\mathcal{P}}_\mu\left(\int_0^{t} a_\mu\big(\boldsymbol{X}(s)\big)\,\mathrm{d}s\right) > 1\right] = o(\delta t),\\ &\mathbb{P}\left[\boldsymbol{\mathcal{P}}_\mu\left(\int_0^{t+\delta t} a_\mu\big(\boldsymbol{X}(s)\big)\,\mathrm{d}s\right) - \boldsymbol{\mathcal{P}}_\mu\left(\int_0^{t} a_\mu\big(\boldsymbol{X}(s)\big)\,\mathrm{d}s\right) > 0,\right.\\ &\qquad\left.\boldsymbol{\mathcal{P}}_j\left(\int_0^{t+\delta t} a_j\big(\boldsymbol{X}(s)\big)\,\mathrm{d}s\right) - \boldsymbol{\mathcal{P}}_j\left(\int_0^{t} a_j\big(\boldsymbol{X}(s)\big)\,\mathrm{d}s\right) > 0\right] = o(\delta t), \quad \text{for} \quad j \neq \mu,\end{aligned}$$

where the first and second equalities directly follow from property (iii) and (iv) of the Poisson processes $\boldsymbol{\mathcal{P}}_\mu$, respectively, and the last equality follows from their independence. In fact, this suggests that we can represent the reaction counts $\boldsymbol{K}_\mu$ as non-homogeneous Poisson processes with intensities a_μ, i.e.,

$$\boldsymbol{K}_\mu(t) = \boldsymbol{\mathcal{P}}_\mu\left(\int_0^t a_\mu\big(\boldsymbol{X}(s)\big)\,\mathrm{d}s\right) \quad (\mu = 1, \dots, M),$$

see [86, 88] for further details. The above equation is an example of a *random time change* representation of stochastic processes, cf. [87].

From the above considerations, we thus find that the state of the system at time t can be represented in stochastic reaction kinetics as

$$\boldsymbol{X}(t) = \boldsymbol{X}(0) + \sum_{\mu=1}^{M} \xi_\mu \boldsymbol{\mathcal{P}}_\mu\left(\int_0^t a_\mu\big(\boldsymbol{X}(s)\big)\,\mathrm{d}s\right), \tag{2.6}$$

where $\mathcal{P}_\mu$, $\mu = 1, \dots, M$, are independent, unit Poisson processes and $\boldsymbol{X}(0)$ denotes the initial state of the system. Given the net changes ξ_μ and propensities a_μ of each reaction R_μ, as well as an initial state $\boldsymbol{X}(0)$, the stochastic equation (2.6) uniquely determines $\boldsymbol{X}(t)$.[11] There are other ways of relating the reaction propensities a_μ to the stochastic process $\boldsymbol{X}(t)$, cf. Subsection 2.2.5. However, eq. (2.6) has the advantage of being intuitive and easily generalized to incorporate additional effects, such as external noise or time delays [88]. As we will see in the following, eq. (2.6) also suggests methods to simulate the stochastic process $\boldsymbol{X}(t)$.

2.2.3 The Stochastic Simulation Algorithm

In the sequel, we introduce the most prominent procedure for generating an exact numerical realization of the stochastic process $\boldsymbol{X}(t)$. This method, called stochastic simulation algorithm (SSA) or Gillespie's Direct Method, was presented by D. T. Gillespie in [30, 31] and is based on the probability function $p(\tau, j \,|\, X; t)$, defined such that

$$p(\tau, j \,|\, X; t)\delta t := \text{probability that the next reaction will occur in the infinitesimal time interval } [t+\tau; t+\tau+\delta t) \text{ and that this reaction will be } R_j, \text{ given that } \boldsymbol{X}(t) = X.$$

Knowledge of $p(\tau, j \,|\, X; t)$ allows to sample when and what will be the next reaction that fires and hence to propagate the system forward in time. By running many of such stochastic simulations, one can compute an ensemble of realizations to *approximate* the statistics of $\boldsymbol{X}(t)$, such as its mean, percentiles, correlations or its full PDF. The SSA and its derivation are given below, mainly following [10, 31]. We also briefly discuss other exact simulation methods.

Derivation of the Stochastic Simulation Algorithm

We notice that the joint probability $p(\tau, j \,|\, X; t)\delta t$ is given as the product of the probability (a) that no reaction fires in $[t; t+\tau)$, given $\boldsymbol{X}(t) = X$; and the conditional probability (b) that reaction R_j fires once in $[t+\tau; t+\tau+\delta t)$, given $\boldsymbol{X}(t+\tau) = X$. Clearly, the latter event has probability $a_j(X)\delta t + o(\delta t)$ as $\delta t \to 0$, such that

$$p(\tau, j \,|\, X; t)\delta t = p_0(\tau \,|\, X; t)\big(a_j(X)\delta t + o(\delta t)\big), \tag{2.7}$$

[11] It should be noted that without any additional assumptions, however, the solution of eq. (2.6) might only exists up to some finite time, i.e., $\boldsymbol{X}(t)$ blows up and reaches infinity in finite time. For instance, consider $\boldsymbol{X}(t) = \boldsymbol{X}(0) + \mathcal{P}_1\big(\int_0^t c_1 \boldsymbol{X}(s)(\boldsymbol{X}(s)-1)/2 \,\mathrm{d}s\big)$ for $\boldsymbol{X}(0) > 1$ and compare with Example 2.3.1.

where $p_0(\tau \mid X; t)$ denotes the probability (a). Since reactions are modeled as independent Poisson processes whose intensities are constant if no event occurs in $[t; t+\tau)$, it follows from their independence and Theorem 2.2.2 that

$$p_0(\tau \mid X; t) = \exp\{-a_0(X)\tau\}, \tag{2.8}$$

where $a_0(X) := \sum_{\mu=1}^{M} a_\mu(X)$. Substituting eq. (2.8) into eq. (2.7), dividing both sides by δt and taking the limit $\delta t \to 0$ results in

$$p(\tau, j \mid X; t) = \exp\{-a_0(X)\tau\} a_j(X). \tag{2.9}$$

Next, we demonstrate how the system can be stochastically simulated based on $p(\tau, j \mid X; t)$. Consider $\boldsymbol{X}(t) = X$, let $p(\tau \mid X; t)$ denote the probability that the next reaction fires at time $t+\tau$, and let $p(j \mid \tau, X; t)$ denote the conditional probability that this reaction will be R_j. From eq. (2.9) it follows that

$$p(\tau \mid X; t) = \sum_{j=1}^{M} p(\tau, j \mid X; t) = \exp\{-a_0(X)\tau\} a_0(X), \tag{2.10}$$

which shows that the time until the next reaction fires is an exponential random variable with parameter $a_0(X)$. Since $p(\tau, j \mid X; t) = p(j \mid \tau, X; t) p(\tau \mid X; t)$, we find from eqs. (2.9) and (2.10) that

$$p(j \mid \tau, X; t) = \frac{p(\tau, j \mid X; t)}{p(\tau \mid X; t)} = \frac{a_j(X)}{a_0(X)}, \tag{2.11}$$

provided that $p(\tau \mid X; t) > 0$.[12] The index of the next reaction that fires is thus an integer random variable with point probabilities $a_j(X)/a_0(X)$, $j = 1, \ldots, M$. There exist several Monte Carlo (MC)-methods for generating samples τ and j according to the distributions implied by eqs. (2.10) and (2.11), respectively. One of the most simplest is the *direct method*, which follows from the standard inversion generating method of MC-theory [10, 89]: Generate two random numbers r_1 and r_2 from the standard uniform distribution $U(0,1)$ and take

$$\tau = \frac{1}{a_0(X)} \ln\left(\frac{1}{r_1}\right), \tag{2.12a}$$

$$j = \text{the smallest integer such that } \sum_{\mu=1}^{j} a_\mu(X) \geq r_2 a_0(X). \tag{2.12b}$$

Given an initial state $\boldsymbol{X}(t_0) = X_0$ at some time t_0, the SSA utilizes this generating method to compute a stochastic realization of $\boldsymbol{X}(t)$. The main steps of the SSA are given in Table 2.1.

[12] If $p(\tau \mid X; t) = 0$ or, equivalently, $a_0(X) = 0$, no further reaction can occur in the system (e.g., due to extinction of all reactant species) and the realization will stay in the absorbing state X, i.e., $\boldsymbol{X}(s) = X$ for all $s \geq t$.

Table 2.1: *The stochastic simulation algorithm, cf. [30, 31].*

1:	Initialize: Set $X \leftarrow X_0$ and $t \leftarrow t_0$.
2:	Evaluate: Compute all $a_\mu(X)$ and $a_0(X) = \sum_{\mu=1}^{M} a_\mu(X)$; stop if $a_0(X) = 0$.
3:	MC-Step: Draw values for τ and $j$according to eqs. (2.12a) and (2.12b), respectively.
4:	Update: Set $X \leftarrow X + \nu_j$ and $t \leftarrow t + \tau$.
5:	Iterate: Go to step **2** or stop simulation.

Other Exact Simulation Algorithms

We note that two random numbers are used in each iteration of Gillespie's Direct Method. In [30], D. T. Gillespie proposed a second algorithm, the *First Reaction Method*, where the putative times to the next R_μ firings, $\mu = 1, \ldots, M$, are calculated in every iteration step and only the earliest of those is accepted. This method also generates values for τ and j in exact agreement with $p(\tau, j \,|\, X; t)$ as given in eq. (2.9), but requires M random numbers per iteration and is thus less efficient than the Direct Method.

Another exact stochastic simulation algorithm, the *Next Reaction Method*, was introduced by Gibson & Bruck in [32, 90]. The Next Reaction Method is a modified version of the First Reaction Method. However, it has the potential of being more efficient than the SSA if the system includes many species and reactions, large N and M, respectively. It requires per iteration only one random number and a computational time that is proportional to $\mathcal{O}(\ln M)$. The reduction in numerical costs is realized by storing the putative next firing times of all reactions in an indexed priority cue. This priority cue includes a binary tree structure, dynamically constructed such that each parent node has an earlier next firing time than either of its children nodes. The costs of updating this tree structure are $\mathcal{O}(\ln M)$. As further proven in [32], re-use of the next firing times of all reactions that do not fire becomes statistically correct under some special transformation. By employing this transformation, the Next Reaction Method consumes only one random number per iteration. For further details we refer to [32, 90]; further exact simulation methods have been reviewed in [10, 91, 92].

2.2.4 Approximate Stochastic Simulation Approaches

It is important to notice that the lower limit of computational costs of an exact stochastic simulation is dictated by the number of single reaction events that have to be realized. This renders

exact indirect methods like the SSA numerically impractical when applied to larger systems that include many rapidly firing reactions or species present in large numbers. Therefore, exactness in the simulation of the stochastic process $\boldsymbol{X}(t)$ is usually sacrificed in order to reduce the computational costs [10], as done, for instance, in *tau-leaping methods*, e.g., [33, 34, 39–43], or approaches that exploit a kind of quasi-steady state approximation (QSSA), e.g., [46–52, 93]. Below, we outline the assumptions underlying these approaches and briefly discuss resulting simulation strategies, mainly following [10, 33].

The Tau-Leaping Approach

In tau-leaping methods it is assumed that given $\boldsymbol{X}(t) = X$ then there exists a time increment $\tau > 0$ that satisfies the following *Leap Condition* [33]: No reaction propensity appreciable changes its value during $[t; t + \tau)$, i.e., for every $\mu = 1, \ldots, M$ it holds

$$a_\mu\big(\boldsymbol{X}(s)\big) \approx a_\mu(X) \quad \forall s \in [t; t + \tau). \tag{2.13}$$

Of course, condition (2.13) will be exactly satisfied if τ is chosen so small that no reaction fires in $[t; t + \tau)$, as done in an exact stochastic simulation. But then one would also obtain no computational benefit. The idea of tau-leaping methods is to choose larger values for τ and to simulate all events that are likely to occur in $[t; t + \tau)$ simultaneously. The error resulting from such 'leap' will then depend on the quality of the approximation made in eq. (2.13).

To outline the tau-leaping approach, we assume for the moment that condition (2.13) holds. Then, a possible state of the system at time $t + \tau$ can be approximated by

$$\boldsymbol{X}(t + \tau) \overset{\text{eq. (2.6)}}{=} \boldsymbol{X}(t) + \sum_{\mu=1}^{M} \xi_\mu \boldsymbol{\mathcal{P}}_\mu\left(\int_t^{t+\tau} a_\mu\big(\boldsymbol{X}(s)\big)\, \mathrm{d}s\right) \approx \boldsymbol{X}(t) + \sum_{\mu=1}^{M} \xi_\mu \boldsymbol{\mathcal{P}}_\mu\big(a_\mu(\boldsymbol{X}(t))\tau\big). \tag{2.14}$$

This equation is the basic tau-leaping formula [33]. As pointed out in [88], it is equivalent to defining an (explicit) Euler-type approximation to eq. (2.6), suggesting an obvious strategy for approximately doing a stochastic simulation [10]: (i) given $\boldsymbol{X}(t) = X$, choose a value for τ satisfying condition (2.13), (ii) generate samples K_μ, $\mu = 1, \ldots, M$, of Poisson random variables with mean $a_\mu(X)\tau$, (iii) update the state $X \leftarrow X + \sum_{\mu=1}^{M} \xi_\mu K_\mu$ and time $t \leftarrow t + \tau$, and (iv) continue with (i) or stop simulation. For sufficiently large values of the samples K_μ generated in each iteration, such approximate simulation will then in principle be faster than a corresponding exact stochastic simulation. However, since both, the approximation quality and the gain in computational speed, will depend on τ, the practical question arrises how to efficiently determine appropriate values for the leap-size τ?

As the propensity of a source-like reaction is a constant function, condition (2.13) is always satisfied in this case. But in general the propensity function will depend on the number of molecules $\boldsymbol{X}$. However, $\boldsymbol{X}$ changes only by ξ_μ with any R_μ firing, which is typically in the range of one or two molecules. The Leap Condition (2.13) can thus always be satisfied if species are present in sufficiently large numbers [15, 33], and the approximation error can be kept small by choosing τ in such way that the relative change in the level of all reactants in $[t; t+\tau)$ is negligible. Hence, given $\boldsymbol{X}(t) = X$ it has to hold

$$|\Delta_\tau \boldsymbol{X}_i(t)| \ll X_i \quad \forall\, i \in I_r,$$

where $|\Delta_\tau \boldsymbol{X}_i(t)| := |\boldsymbol{X}_i(t+\tau) - X_i| \approx \left|\sum_{\mu=1}^{M} \xi_{\mu i} \boldsymbol{\mathcal{P}}_\mu(a_\mu(X)\tau)\right|$ is the absolute change in $\boldsymbol{X}_i$ during $[t; t+\tau)$, and I_r is the index set of all reactants, i.e., $i \in I_r$ if and only if $s^r_{\mu i} > 0$ for some $\mu = 1, \ldots, M$. By enforcing bounds on the mean and variance of $|\Delta_\tau \boldsymbol{X}_i(t)|$ with respect to $\boldsymbol{X}_i(t)$, an efficient and currently widely used strategy for an appropriate τ-selection has been proposed in [94]. There also exist approaches to avoid the possibility of driving some number of molecules negative, which can occur in a leap-step if the random samples K_μ associated with reactions that decrease a common species are too large. These approaches are usually based on strategies where such critical reactions are monitored and simulation is switched to the exact SSA if necessary [95], or where the unbounded Poisson random numbers K_μ are approximated by bounded binomial [39, 40, 96] or multinomial [97] random numbers.

If the chemical system is stiff due to the presence of well-separated fast and slow reactions, simulations based on the tau-leaping formula (2.14) or any other explicit leaping scheme [33, 39, 40] will proceed slowly. Because then the leap-size τ has to be restricted to the time scale of the system's fastest mode in order to keep the approximation error small [10]. Following well-know strategies to numerically solve stiff ODEs, partially implicit tau-leaping methods have been developed [34, 42] that are able to produce significantly faster simulations than explicit tau-leaping methods for stiff systems. However, these methods usually suffer from a damping effect: The variance generated from the underlying implicit tau-leaping formula is much smaller than the exact value [10]. This damping effect can be reduced (a) by combining implicit tau-leap steps with a sequence of much shorter explicit tau-leap or SSA steps, a strategy called down-shifting [34], or (b) by employing higher-order approximation formulas like the trapezoidal rule [41, 98] or implicit second-order weak Taylor schemes [43]. Other methods to stochastically simulate stiff systems are either based on the idea of eliminating the fast modes through a kind of QSSA, discussed below, or follow a hybrid approach where fast reactions are approximated as continuous processes, see Chapter 3.

The QSSA-Based Approach

One strategy to stochastically simulate stiff systems relies on the idea underlying the probably best-known model of enzyme kinetics, the Michaelis–Menten approximation [12, 13, 99], where the subset of enzyme-related species (free enzyme and enzyme–substrate complex) is assumed to be asymptotically at steady state on the time scale of interest. Such type of approximation is known as the quasi-steady state approximation (QSSA), cf. [100, 101]. By eliminating the fast modes in stiff systems, the QSSA generally reduces the model complexity and, consequently, also the computational complexity. The QSSA has been applied to stochastic reaction kinetics in several different ways [46–52, 93]. In Chapter 5, we use a QSSA-like approach to eliminate a stochastic model of gene expression in order to derive effective protein synthesis rates. In the following, we outline the main steps of QSSA-based approaches to accelerate the stochastic simulation of chemical systems.

To employ a QSSA for stochastic simulation, the first step is to identify a (provisional) partitioning of the reactions into fast and slow subsets, $\mathcal{R}^f$ and $\mathcal{R}^s$, where reactions those propensity functions tend to have large values are assigned to $\mathcal{R}^f$ and all other reactions are assigned to $\mathcal{R}^s$. Such partitioning can be based on some (biological) insight into the system [46, 49], or by recording and comparing the number of firings of each reaction channel during a few SSA runs [48]. It can be further verified by testing stochastic stiffness conditions of the fast modes, for details see [47], or adaptively changed during simulation based on current propensity values [51, 52]. If a partitioning into fast and slow reactions is not obvious in a given model, then this might indicate that the system is not really stiff.

Usually, one is interested mostly in the effective dynamics on the slow time scale associated with $\mathcal{R}^s$ and hence wants to eliminate the fast modes associated with $\mathcal{R}^f$, which can be done by applying the QSSA. Henceforth, we adopt the notation used in [47], and continue by defining a partitioning of the species into 'fast' and 'slow' species, $\mathcal{S}^f$ and $\mathcal{S}^s$, respectively, which is induced by the reaction partitioning as follows: A species those number of molecules gets changed by a fast reaction is assigned to $\mathcal{S}^f$; all other species that exclusively get changed by slow reactions (but not by fast ones) are assigned to $\mathcal{S}^s$. The process $\boldsymbol{X}(t)$ can be rearranged correspondingly as $\boldsymbol{X}(t) = (\boldsymbol{X}^f(t), \boldsymbol{X}^s(t))^{\mathrm{T}}$, where $\boldsymbol{X}^f(t)$ and $\boldsymbol{X}^s(t)$ denote the process composed by the fast and slow species state variables, respectively. It is important to note that the processes $\boldsymbol{X}^f(t)$ and $\boldsymbol{X}^s(t)$ are generally not independent of each other. Following [47], we thus define the virtual fast process $\widehat{\boldsymbol{X}}^f(t)$ as the process composed by the same fast species state variables as $\boldsymbol{X}^f(t)$ but evolving under only the fast reactions $\mathcal{R}^f$.

With all slow reactions turned off, the QSSA can be applied to $\widehat{\boldsymbol{X}}^f$, i.e., we assume that $\widehat{\boldsymbol{X}}^f$ approaches a well-defined quasi-stationary PDF $u(\widehat{X}^f \,|\, X^f, X^s; t)$ for given $\boldsymbol{X}(t) = (X^f, X^s)^{\mathrm{T}}$ before the next slow reaction is likely to fire. Then, the propensity function $a_\mu^s(X^f, X^s)$ of a reaction $\mathcal{R}^s$ can be approximated on the slow time scale by

$$\bar{a}_\mu^s(X^f, X^s) := \sum_{\widehat{X}^f} a_\mu^s\big(\widehat{X}^f, X^s\big) u\big(\widehat{X}^f \,|\, X^f, X^s; t\big), \tag{2.15}$$

which is the average over the fast variables, treated as these were distributed according to the asymptotic form to which the virtual fast process $\widehat{\boldsymbol{X}}^f$ relaxes, see further [46, 47, 52]. Given $u(\widehat{X}^f \,|\, X^f, X^s; t)$, the system can be effectively propagated on the slow time scale by simulating only the slow reactions, i.e., using a method such as the SSA but replacing the propensity $a_\mu^s(X^f, X^s)$ of each slow reaction $\mathcal{R}^s$ by its average $\bar{a}_\mu^s(X^f, X^s)$ as given in eq. (2.15).

Indirect methods utilizing a QSSA mainly differ in the way the averaged propensities $\bar{a}_\mu^s(X^f, X^s)$ of the slow reactions are computed. In [46–49] this is realized by considering the first moments of $u(\widehat{X}^f \,|\, X^f, X^s; t)$, where moments are computed from either exact (if available) or approximate analytic expressions of $u(\widehat{X}^f \,|\, X^f, X^s; t)$. Such approach benefits from the fact that once algebraic expressions for $\bar{a}_\mu^s(X^f, X^s)$ have been derived, these can be directly used for a simulation on the slow time scale. However, generalizing this approach to complex reaction networks is not straightforward. In [51, 52], the averaged propensities $\bar{a}_\mu^s(X^f, X^s)$ are approximated by a small sequence of relatively short SSA-runs of $\widehat{\boldsymbol{X}}^f$. Of course, such approach requires the successive approximation of $\bar{a}_\mu^s(X^f, X^s)$ after each slow reaction step and is hence more expensive than the previous mentioned approach. Nevertheless, it offers a more general and adaptable way of employing the QSSA that is still capable of substantial computational savings compared to exact SSA-simulations of stiff system, see [51, 52].

2.2.5 The Chemical Master Equation

In the previous subsections, we studied methods to simulate the stochastic process $\boldsymbol{X}(t)$. Such indirect methods suffer from the disadvantages of a MC-based approach (see Chapter 1), as the statistics of $\boldsymbol{X}(t)$ have to be approximated by an ensemble of stochastic realizations. This can be avoided if we *directly* solve for the PDF $P(X; t \,|\, X_0; t_0)$ of $\boldsymbol{X}(t)$ for a given initial state $X_0 \in \mathbb{N}_0^N$ at time t_0, defined as

$$P(X; t \,|\, X_0; t_0) := \mathbb{P}\big[\boldsymbol{X}(t) = X \,|\, \boldsymbol{X}(t_0) = X_0\big],$$

for every $X \in \mathbb{N}_0^N$ and $t \geq t_0$. The function $P(X; t \,|\, X_0; t_0)$ can be physically interpreted as follows: If an ensemble of identical chemical systems started in the same initial state X_0 at time t_0, then $P(X; t \,|\, X_0; t_0)$ gives the fraction of systems that are in state X at time $t \geq t_0$. The time-evolution equation of $P(X; t \,|\, X_0; t_0)$ is the chemical master equation (CME) that stands at the basis of the stochastic framework. As shown by D. T. Gillespie [82], the CME can be directly derived from the fundamental hypothesis of stochastic reaction kinetics. Below, we outline Gillespie's derivation of the CME, referring to [82] for further details.

Derivation of the Chemical Master Equation

Consider an infinitesimal time interval $[t; t + \delta t)$ and the probability $P(X; t + \delta t)$.[13] We can specify $P(X; t + \delta t)$ by considering all events that can lead the system to state X during $[t; t + \delta t)$. These events are: (i) The system was in X at time t and no reaction fired in $[t; t + \delta t)$. The probability of this event is given by $P(X; t)$ times the probability $\mathbb{P}_0[X; t]$ that no reaction fired in $[t; t + \delta t)$ given $\boldsymbol{X}(t) = X$. (ii) For $\mu = 1, \ldots, M$, the system was in state $(X - \xi_\mu) \in \mathbb{N}_0^N$ at time t and reaction R_μ fired once in $[t; t + \delta t)$. The probability of each of these events is given by $P(X - \xi_\mu; t)$ times the probability $\mathbb{P}_\mu[X - \xi_\mu; t]$ that R_μ fired once in $[t; t + \delta t)$ given $\boldsymbol{X}(t) = X - \xi_\mu$. (iii) The system reached X by multiple reaction events in $[t; t + \delta t)$, which has probability $o(\delta t)$ as $\delta t \to 0$.

We observe that the events under (i)–(iii) are mutually exclusive, such that $P(X; t + \delta t)$ is given by the sum of their probabilities, i.e.,

$$P(X; t + \delta t) = P(X; t)\mathbb{P}_0[X; t] + \sum_{\mu=1}^{M} P(X - \xi_\mu; t)\mathbb{P}_\mu[X - \xi_\mu; t] + o(\delta t), \tag{2.16}$$

as $\delta t \to 0$.[14] By the definition of a propensity, we have $\mathbb{P}_\mu[X - \xi_\mu; t] = a_\mu(X - \xi_\mu)\delta t + o(\delta t)$, for every $\mu = 1, \ldots, M$, and similarly $\mathbb{P}_0[X; t] = 1 - \sum_{\mu=1}^{M} a_\mu(X)\delta t + o(\delta t)$. Putting things together, we thus have

$$P(X; t + \delta t) = P(X; t)\left(1 - \sum_{\mu=1}^{M} a_\mu(X)\delta t\right) + \sum_{\mu=1}^{M} P(X - \xi_\mu; t)a_\mu(X - \xi_\mu)\delta t + o(\delta t). \tag{2.17}$$

[13] Henceforth, we omit the dependence on the initial value of $\boldsymbol{X}$ in the notation of its PDF and simply write $P(X; t)$.

[14] All terms in eq. (2.16) that correspond to some $(X - \xi_\mu) \notin \mathbb{N}_0^N$ have to be omitted, since $P(\cdot\,; t)$ is only defined on $\mathbb{N}_0^N$. For convenience, we hence set $P(X; t)$ for all $X \notin \mathbb{N}_0^N$, here and in the following.

Subtracting $P(X;t)$ from both sides of eq. (2.17), dividing by δt and taking the limit $\delta t \to 0$, we obtain the CME

$$\frac{\partial}{\partial t}P(X;t) = \sum_{\mu=1}^{M} a_\mu(X-\xi_\mu)P(X-\xi_\mu;t) - a_\mu(X)P(X;t), \tag{2.18}$$

a special type of the Kolmogorov forward equation that completely characterizes the statistics of $\boldsymbol{X}$.

Solving the Chemical Master Equation

Even though the CME has a relative simple structure, analytical solutions exist only for the special case where the chemical systems solely include unimolecular reactions [102, 103]. Mathematically speaking, the CME can be considered as a system of coupled ODEs, with one equation for every possible state X, or, equivalently, as a discrete partial differential equation (PDE), where continuous derivatives are replace by discrete differences [26]. Now, the main problem arises from the high dimensionality of the CME, which renders most approaches computationally infeasible that aim to solve the CME directly. For example, assume we have 5 species whose number of molecules is each limited to the range 0–99, say, than the chemical system has in total 100^5 possible states. In this example, solving the CME thus means to solve 10^{10} ODEs, which would be a real numerical challenge.

The discrete PDE point of view motivates the use of Galerkin-based approaches to directly solve the CME, e.g., [26–28]. Modifying these methods to tackle the many degrees of freedom that result from the high dimensionality of the CME, in combination with a typical low regularity and multi-modality of its solution, however, is still an ongoing field of research [26, 28]. Another class of approaches is based on the idea to truncate the state space to a finite domain that captures enough information of the PDF and, at the same time, allows to numerically solve the CME, e.g., [25, 29, 104–106]. This method, first suggested by B. Munsky and M. Khammash [25], is known as the finite state projection (FSP) method, discussed below.

The Finite State Projection Algorithm: To illustrate the FSP method in the following, we first notice that the infinitesimal generator $\mathcal{A} := \lim_{t\to 0+}(P(t)-I)/t$ for the process $\boldsymbol{X}$, where $P(t) := P(\,\cdot\,;t)$, is given by eq. (2.18) as

$$\big(\mathcal{A}P(\,\cdot\,;t)\big)(X) = \sum_{\mu=1}^{M} a_\mu(X-\xi_\mu)P(X-\xi_\mu;t) - a_\mu(X)P(X;t), \tag{2.19}$$

such that we can write the CME in operator form as the linear equation[15]

$$\frac{\partial}{\partial t}P(t) = \mathcal{A}P(t). \tag{2.20}$$

The generator $\mathcal{A}$ in eq. (2.20) is a matrix with a usually large, possibly infinite number of rows and columns. As can be seen from eq. (2.19), if the state space is finite, the generator matrix $\mathcal{A}$ obeys a special structure: It is a square matrix with non-negative diagonal elements and non-positive off-diagonal elements such that every column sum is exactly zero. It follows from the Gershgorin Circle Theorem [107] and this special structure that at least one eigenvalue of $\mathcal{A}$ is equal to zero and all others have a non-positive real-part.

The FSP method utilizes the linearity of the CME. Let us denote by

$$\mathcal{X} := \left\{X \in \mathbb{N}_0^N : \underline{X}_1 \leq X_1 \leq \overline{X}_1, \ldots, \underline{X}_N \leq X_N \leq \overline{X}_N\right\}$$

a finite subspace $\mathcal{X} \subset \mathbb{N}_0^N$ for lower and upper truncation vectors $\underline{X} = (\underline{X}_1, \ldots, \underline{X}_N)^{\mathrm{T}} \in \mathbb{N}_0^N$ and $\overline{X} = (\overline{X}_1, \ldots, \overline{X}_N)^{\mathrm{T}} \in \mathbb{N}_0^N$, respectively, satisfying $\underline{X}_i \leq \overline{X}_i$ for all $i = 1, \ldots, N$. Let further $\mathcal{A}_\mathcal{X}$ denote the restriction of $\mathcal{A}$ to the truncated state space $\mathcal{X}$. Then, the solution $P_\mathcal{X}(t)$ to the corresponding CME, i.e., $\frac{\partial}{\partial t}P_\mathcal{X}(t) = \mathcal{A}_\mathcal{X}P_\mathcal{X}(t)$, is given as

$$P_\mathcal{X}(t) = \exp\{\mathcal{A}_\mathcal{X}t\}P_\mathcal{X}(0), \tag{2.21}$$

where $\exp\{A\} := \sum_{n=0}^{\infty}(A)^n/n!$ denotes the matrix exponential of A, and $P_\mathcal{X}(0)$ is the initial PDF $P(0)$ at time $t = 0$ restricted to $\mathcal{X}$. Now, in [25] it is proven that, if $\sum_{X\in\mathcal{X}} P_\mathcal{X}(X;t) \geq 1 - \varepsilon$ for some $\varepsilon \geq 0$ and $t > 0$, then

$$P_\mathcal{X}(X;t) \leq P(X;t) \leq P_\mathcal{X}(X;t) + \varepsilon,$$

for every $X \in \mathcal{X}$. This result leads to the FSP algorithm, where the state space $\mathcal{X}$ is gradually expanded until a sufficient approximation of $P(X;t)$ is computed, see Table 2.2

To apply the FSP algorithm, two practical questions arise: (1) How to solve eq. (2.21) efficiently; and (2) How to construct the truncated state space $\mathcal{X}$ appropriately? The first problem has been addressed in [105], where an improved FSP algorithm is presented that uses a Krylov-based approximation to eq. (2.21). Other suggested approaches to solve eq. (2.21)

[15]So far, we considered to solve the CME for deterministic initial values, i.e., $\boldsymbol{X}(t_0) = X_0$ such that $P(X;t_0) = \delta_{X_0}(X)$ with $\delta_{X_0}(X)$ denoting Kronecker's delta function. However, we note that solving the CME for arbitrary initial probability distributions $p_0(X)$, with $p_0(X) \geq 0$ for all $X \in \mathbb{N}_0^N$ and $\sum_{X\in\mathbb{N}_0^N} p_0(X) = 1$, can always be reduced to deterministic initial value problems by exploiting the linearity of the CME [103].

Table 2.2: *The finite state projection algorithm, cf. [25].*

1:	Initialize: Choose $t > 0$, $P(0)$ and $\varepsilon \geq 0$, and select an initial $\mathcal{X}$.
2:	Evaluate: Solve $P_{\mathcal{X}}(t) = \exp\{\mathcal{A}_{\mathcal{X}} t\} P_{\mathcal{X}}(0)$ for given $\mathcal{X}$.
3:	Iterate: If $\sum_{X \in \mathcal{X}} P_{\mathcal{X}} k(X;t) < 1 - \varepsilon$, expand $\mathcal{X}$ and got to step **2**; otherwise, stop.

include a sparse grid method [35], a dynamical low-rank approach [108] or the exploitation of multiple time scales [104]. Problem (2) has been studied in [25, 29, 104, 109], for instance, and is usually addressed by concepts of reachability or some special insight into the system under consideration. In [93], a QSSA-based approach (see Subsection 2.2.4) has been used to reduce the dimensionality of a stochastic model of gene expression and the resulting CME, describing the evolution on the slow time scale, was numerically solved by a simple Euler method. In Chapter 4, we present a different approach to reduce the dimensionality of the CME. Our idea is to eliminate those species from the discrete state space that are present in large numbers and can hence be appropriately represented by their expected levels, which is in line with the deterministic formulation of reaction kinetics, see Section 2.3.

2.2.6 The Chemical Fokker–Planck Equation and the Chemical Langevin Equation

Under a series of assumptions, the CME can be approximated by a less refined model, the chemical Fokker–Planck equation (CFPE), e.g., see [16, 110]. A simple but non-rigorous derivation of the CFPE was given by H. A. Kramers [111] and J. E. Moyal [112]. It relies on the assumption that at any time t the number of molecules of all species are very large, i.e. $\boldsymbol{X}_i(t) \gg 1$ for all $i = 1, \ldots, N$, such that the discrete-state Markov process $\boldsymbol{X}(t) = (\boldsymbol{X}_1(t), \ldots, \boldsymbol{X}_N(t))^{\mathrm{T}}$ can effectively be approximated by a continuous-state Markov process $\boldsymbol{Y}(t) := (\boldsymbol{Y}_1(t), \ldots, \boldsymbol{Y}_N(t))^{\mathrm{T}}$ taking values in $\mathbb{R}^N_{\geq 0}$, where $\boldsymbol{Y}_i(t)$, $i = 1, \ldots, N$, gives the amount of species S_i at time t, cf. [15, 110].[16] The time-evolution equation of the PDF $P(Y;t)$ of $\boldsymbol{Y}(t)$ is given by the CME

$$\frac{\partial}{\partial t} P(Y;t) = \sum_{\mu=1}^{M} a_\mu(Y - \xi_\mu) P(Y - \xi_\mu; t) - a_\mu(Y) P(Y;t). \tag{2.22}$$

[16] Such continuous approximation is usually motivated by introducing a scaling parameter Ω, related to the volume of the system, as we discuss it in the context of deterministic reaction kinetics in the following sections.

A second assumption is that the functions $f_\mu(Y) := a_\mu(Y)P(Y;t)$ in eq. (2.22) are analytic with respect to Y, such that $f_\mu(Y - \xi_\mu)$ can be represented by it Taylor series expansion about Y, i.e.,

$$f_\mu(Y - \xi_\mu) = f_\mu(Y) + \sum_{|\gamma|\geq 1}^{\infty} \prod_{i=1}^{N} \frac{(-\xi_{\mu i})^{\gamma_i}}{\gamma_i!} \left(\frac{\partial}{\partial Y_i}\right)^{\gamma_i} f_\mu(Y), \tag{2.23}$$

with multi-index $\gamma = (\gamma_1, \ldots, \gamma_N) \in \mathbb{N}_0^N$ and $|\gamma| = \gamma_1 + \ldots + \gamma_N$. Substituting eq. (2.23) into eq. (2.22) gives the Kramers–Moyal expansion of the CME

$$\frac{\partial}{\partial t} P(Y;t) = \sum_{\mu=1}^{M} \sum_{|\gamma|\geq 1}^{\infty} \prod_{i=1}^{N} \frac{(-\xi_{\mu i})^{\gamma_i}}{\gamma_i!} \left(\frac{\partial}{\partial Y_i}\right)^{\gamma_i} \big[a_\mu(Y)P(Y;t)\big]. \tag{2.24}$$

By terminating the Taylor series in eq. (2.24) after the second term, we obtain the CFPE, a second-order PDE of the form

$$\frac{\partial}{\partial t} P(Y;t) = \sum_{\mu=1}^{M} \left(-\prod_{i=1}^{N} \xi_{\mu i} \frac{\partial}{\partial Y_i} + \frac{1}{2} \prod_{i,j=1}^{N} \xi_{\mu i}\xi_{\mu j} \frac{\partial^2}{\partial Y_i Y_j} \right) a_\mu(Y)P(Y;t), \tag{2.25}$$

which replaces the massive ODE-system form of the CME. However, as the CFPE is obtained by an uncontrolled, perfunctory truncation of the Taylor series in the Kramers–Moyal expansion (2.24), its validity had been questioned by several authors, see further [16, 110, 113].

In [15], D. T. Gillespie gave a more rigorous derivation of the CFPE, reviving the question of its validity. More precisely, he derived the chemical Langevin equation (CLE), a stochastic differential equation (SDE) of the form

$$\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{Y}(t) = \sum_{\mu=1}^{M} \xi_\mu a_\mu\big(\boldsymbol{Y}(t)\big) + \sum_{\mu=1}^{M} \xi_\mu \sqrt{a_\mu\big(\boldsymbol{Y}(t)\big)}\boldsymbol{\Gamma}_\mu(t), \tag{2.26}$$

with $\boldsymbol{\Gamma}_\mu(t)$, $\mu = 1, \ldots, M$, denoting temporally uncorrelated, independent Gaussian white noises, satisfying $\mathbb{E}[\boldsymbol{\Gamma}_\mu(t)] = 0$ and $\mathbb{E}[\boldsymbol{\Gamma}_\mu(t)\boldsymbol{\Gamma}_j(s)] = \delta_{\mu j}\delta(t-s)$, where the first delta function is Kronecker's and the second is Dirac's, see [89, 114]. The CLE (2.26) is equivalent to the CFPE (2.25), i.e., its solution generates exact sample paths of the CFPE, cf. [16].

Gillespie's derivation of the CLE relies on the assumption that two dynamical conditions are satisfied [15]: If the system is in state X at time t, then there must exist a time increment $\tau > 0$ such that (a) none of the reaction propensities changes significantly during the time interval $[t; t+\tau)$, and (b) the expected number of firings of each reaction R_μ in $[t; t+\tau)$ must be much larger than 1.

We observe that condition (a) is the Leap Condition (2.13) assumed in the derivation of the basic tau-leaping formula

$$\boldsymbol{X}(t+\tau) = \boldsymbol{X}(t) + \sum_{\mu=1}^{M} \xi_\mu \boldsymbol{\mathcal{P}}_\mu\big(a_\mu(\boldsymbol{X}(t))\tau\big). \tag{2.27}$$

As discussed in Subsection 2.2.4, condition (a) can always be satisfied if the number of molecules are sufficiently large. Condition (b) further stipulates that the mean of each Poisson random variable in eq. (2.27) is much larger than 1, i.e.,

$$\mathbb{E}\Big[\boldsymbol{\mathcal{P}}_\mu\big(a_\mu(\boldsymbol{X}(t))\tau\big)\Big] = a_\mu(\boldsymbol{X}(t))\tau \gg 1. \tag{2.28}$$

Since a Poisson random variable with intensity λ, assuming $\lambda \in \mathbb{N}$, can be thought of as the sum of λ independent Poisson random variables each with intensity 1, see Subsection 2.2.2, for large λ such random variable can be well approximated by a normal random variable with the same mean and variance, as stated by the central limit theorem (CLT):

Theorem 2.2.3 (CLT [85]). *Let $X_1, \ldots, X_n$ be a sequence of iid random variables with common mean μ and variance σ^2. Then, the distribution of*

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

converges to the standard normal distribution $\mathcal{N}(0,1)$ as $n \to \infty$.

Thus, condition (b), or more precisely inequality (2.28), allows to approximate the Poisson random variables in eq. (2.27) by statistically independent normal random variables as follows

$$\boldsymbol{\mathcal{P}}_\mu\big(a_\mu(\boldsymbol{X}(t))\tau\big) \approx a_\mu(\boldsymbol{X}(t))\tau + \sqrt{a_\mu(\boldsymbol{X}(t))}\boldsymbol{\mathcal{N}}_\mu(0,1)\sqrt{\tau} \quad (\mu = 1, \ldots, M).$$

It should be noticed that by employing this approximation to eq. (2.27), i.e., replacing *integer* Poisson random variables by *real* normal random variables, one effectively approximates the *discretely* changing, integer-valued process $\boldsymbol{X}(t)$ by a *continuously* changing, real-valued process $\boldsymbol{Y}(t)$ with

$$\boldsymbol{Y}(t+\tau) = \boldsymbol{Y}(t) + \sum_{\mu=1}^{M} \xi_\mu a_\mu\big(\boldsymbol{Y}(t)\big)\tau + \sum_{\mu=1}^{M} \xi_\mu \sqrt{a_\mu\big(\boldsymbol{Y}(t)\big)}\boldsymbol{\mathcal{N}}_\mu(0,1)\sqrt{\tau}.$$

If we make some notational changes: Denote τ by $\mathrm{d}t$, substitute $\mathrm{d}\boldsymbol{Y}(t) := \boldsymbol{Y}(t+\mathrm{d}t) - \boldsymbol{Y}(t)$ and introduce temporally uncorrelated, independent Wiener processes (Brownian motion)

$\boldsymbol{W}_\mu(t)$, $\mu = 1, \dots, M$, satisfying $\mathrm{d}\boldsymbol{W}_\mu(t) := \boldsymbol{W}_\mu(t + \mathrm{d}t) - \boldsymbol{W}_\mu(t) = \boldsymbol{\mathcal{N}}_\mu(0,1)\sqrt{\mathrm{d}t}$; we can write the above equation as

$$\mathrm{d}\boldsymbol{Y}(t) = \sum_{\mu=1}^{M} \xi_\mu a_\mu\big(\boldsymbol{Y}(t)\big)\mathrm{d}t + \sum_{\mu=1}^{M} \xi_\mu \sqrt{a_\mu\big(\boldsymbol{Y}(t)\big)}\mathrm{d}\boldsymbol{W}_\mu(t),$$

which is the Itô differential form of the CLE, mathematically equivalent to the white-noise form (2.26), cf. [15, 89].

Even though, questions regarding the general validity and accuracy of the CFPE/CLE approximation are not answered by Gillespie's derivation and still subject of ongoing research, see [113, 115–117], it shows that a continuous approximation of $\boldsymbol{X}$ is justified if the above conditions (a) and (b) are satisfied. A large copy number of every species suffices to ensure that condition (a) is satisfied. Condition (b) essentially requires that all reactions have sufficiently large propensity values. Since a propensity function generally scales with the number of molecules, we thus expect that a large number regime also suffices to ensure that condition (b) is satisfied. In Section 2.4, we study this relation more closely with respect to the continuous approximation made in deterministic reaction kinetics.

2.3 Deterministic Reaction Kinetics

In deterministic reaction kinetics, the state of the system is described by a continuous variable $x := (x_1, \dots, x_N)^{\mathrm{T}} \in \mathbb{R}^N_{\geq 0}$ that is typically related to the number of molecules by

$$x_i := X_i/\Omega \quad (i = 1, \dots, N), \tag{2.29}$$

where $\Omega \in \mathbb{R}_{>0}$ denotes some scaling factor. According to eq. (2.29), the unit of x is equal to the inverse unit of Ω. Typically, this scaling factor Ω is related to the volume of the system. For instance, $\Omega = N_A \cdot V$, where $N_A \approx 6.02 \times 10^{23}\,\mathrm{mol}^{-1}$ denotes the Avogadro constant, such that x describes the molar concentrations of the species. As outlined above, such continuous approximation is justified if $X_i \gg 1$, for all $i = 1, \dots, N$. Obviously, if further $\Omega \gg 1$, then the scaled net changes ξ_μ/Ω are small compared to the intensive variable $x = X/\Omega$, such that fluctuations in the species levels caused by the reactions become negligible. Deterministic reaction kinetics can hence be understood as an approximation under such large-size system assumption, cf. Section 2.4.

2.3.1 Reaction Rate

In deterministic reaction kinetics, the *rate* v_μ of a reaction R_μ is customarily defined with respect to the (measurable) change in the amount of a species caused by the firings of R_μ in an infinitesimal time interval, cf. [12, 13, 81]. Hence, a reaction rate is regarded in the deterministic framework as

$$\begin{aligned} v_\mu(x) = {} & \text{average number of } R_\mu \text{ firings per } \Omega \text{ and per unit time} \\ & \text{given that the system is in state } x, \end{aligned} \tag{2.30}$$

see [30]. The unit of v_μ is thus always equal to the product of the inverse units of Ω and time.

Law of Mass Action

The general form of the rates of elementary reactions is given by the law of mass action, first formulated by C. M. Guldberg and P. Waage in 1864 [118]:

> The rate of an elementary reaction R_μ is proportional to the product of molar concentrations of all R_μ reactants raised to the power of their stoichiometric coefficients, where the factor of proportionality is called *reaction rate constant*.

Let k_μ denote the rate constant of an elementary reaction R_μ, then the law of mass action predicts the rate equation

$$v_\mu(x) = k_\mu \prod_{i=1}^{N} (x_i)^{s^r_{\mu i}}. \tag{2.31}$$

The rate v_μ of an elementary reaction R_μ is thus a simple multivariate polynomial function of the species levels x_i, $i = 1, \ldots, N$. The degree of v_μ, called *reaction order*, is determined by the stoichiometric coefficients $s^r_{\mu i}$ of the R_μ reactant species. More precisely, the order of an elementary reaction R_μ is equal to its molecularity $|s^r_\mu| = \sum_{i=1}^{N} s^r_{\mu i}$, where each stoichiometric coefficient $s^r_{\mu i}$ gives the partial order of R_μ with respect to S_i.[17]

Reaction Rate Constant

From eqs. (2.30) and (2.31), we find that the unit of the rate constant k_μ of an elementary reaction R_μ depends on the reaction order: (0) For a source-like reaction R_0 the unit of k_0 is

[17] It should be noted that the partial orders, molecularities and stoichiometric coefficients are only identical in the case of elementary reactions. The partial order of a stepwise reaction is usually a real-valued, possibly negative number, and the concept of molecularity is not defined in that case, see further [80, 81].

equal to the product of the inverse units of Ω and time, (1) for a unimolecular reaction R_1 the unit of k_1 is equal to the product of the inverse unit of time, and (2) for every bimolecular reaction R_2 the unit of k_2 is equal to the unit of Ω times the inverse unit of time.

In general, the rate constant k_μ of a reaction R_μ also depends on the absolute temperature T and activation energy E_a as described by the (empirical) Arrhenius equation [13, 80, 81]

$$k_\mu = A \exp\left\{-\frac{E_a}{k_B T}\right\}, \tag{2.32}$$

where k_B denotes the Boltzmann constant and E_a is the activation energy in molecular units. The pre-exponential factor A in eq. (2.32) is related to the frequency of collisions of R_μ reactant molecules and has the same unit as k_μ. As predicted by collision theory and transition state theory of chemical kinetics, the factor A also weakly depends on the absolute temperature T, see [12]. In contrast to the specific probability rate constant c_μ, however, we find the rate constant k_μ to be independent from the system's volume V.

2.3.2 Representation with ODEs

Let $\boldsymbol{\eta}_\mu(t)$ denote the average number of firings of a reaction R_μ per Ω until time t, with $\boldsymbol{\eta}_\mu(0) = 0$ for all $\mu = 1, \ldots, M$. According to the definition of a reaction rate v_μ, see eq. (2.30), in classical reaction kinetics $\boldsymbol{\eta}_\mu(t)$ is considered to be the solution of the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\eta}_\mu(t) = v_\mu\big(\boldsymbol{x}(t)\big), \quad \text{with} \quad \boldsymbol{\eta}_\mu(0) = 0,$$

or in integral form

$$\boldsymbol{\eta}_\mu(t) = \int_0^t v_\mu\big(\boldsymbol{x}(s)\big)\,\mathrm{d}s.$$

The species levels $\boldsymbol{x}(t)$, necessary for evaluating are given by the net changes ξ_μ via

$$\boldsymbol{x}(t) = \boldsymbol{x}(0) + \sum_{\mu=1}^{M} \xi_\mu \boldsymbol{\eta}_\mu(t) = \boldsymbol{x}(0) + \sum_{\mu=1}^{M} \xi_\mu \int_0^t v_\mu\big(\boldsymbol{x}(s)\big)\,\mathrm{d}s, \tag{2.33}$$

for some initial levels $\boldsymbol{x}(0) = x_0$ at time $t = 0$. Differentiating eq. (2.33) with respect to time t shows that $\boldsymbol{x}(t)$ is given as the solution of the initial value problem

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}(t) = \sum_{\mu=1}^{M} \xi_\mu v_\mu\big(\boldsymbol{x}(t)\big) =: f\big(\boldsymbol{x}(t)\big), \quad \text{with} \quad \boldsymbol{x}(0) = x_0. \tag{2.34}$$

Eqs. (2.34) build the usual ODE-model of deterministic reaction kinetics, a coupled system of M autonomous ODEs of first order. If the considered chemical system does not include a source-like reaction, then this ODE-system is homogeneous; otherwise it is non-homogeneous. If all reactions R_μ are at most of first order, then all rates $v_\mu$will be constant or linear functions such that eqs. (2.34) form a system of linear ODEsand an analytical solution might be available. Given the large variety of well-developed numerical methods for the integration of ODEs, cf. [119–122], finding a numerical solution of the ODE-model of deterministic reaction kinetics is in general not problematic, even for large and more complex, possibly stiff systems. The following well-known result in the study of differential equations guarantees local existence and uniqueness of a solution $\boldsymbol{x}(t)$ of eqs. (2.34).

Theorem 2.3.1 (Picard–Lindelöf Theorem, cf. [119]). *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a locally Lipschitz continuous function, i.e., there exists a constant $L > 0$ such that*

$$\big\|f(x) - f(y)\big\| \leq L\big\|x - y\big\|, \quad \textit{for all } x \in \mathbb{R}^n \textit{ and } y \in \overline{B}_r(x),$$

where $\overline{B}_r(x) := \{y \in \mathbb{R}^n : \|x - y\| \leq r\}$ is some closed r-neighborhood around x with $r > 0$.[18] *Then, the initial value problem*

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}(t) = f\big(\boldsymbol{x}(t)\big), \quad \textit{with} \quad \boldsymbol{x}(0) = x_0.$$

has a unique solution defined on an interval $[0; T]$, with $T > 0$.

In our case, every component of the right hand side f is a polynomial function, because it is a linear combination of the reaction rates v_μ that are polynomial functions by definition. Since every polynomial function is Lipschitz continuous on any bounded interval of its domain of definition, it follows that f is also locally Lipschitz continuous. Thus, from Theorem 2.3.1 we know that the ODE-system (2.34) uniquely determines $\boldsymbol{x}(t)$ on some finite time interval. However, it should be noted that without any additional assumptions this result applies only locally, i.e., the solution of eqs. (2.34) might blow up and reach infinity in finite time, cf. [119].

Example 2.3.1. Let $M = N = 1$, consider the reaction R_1: $2\,S_1 \longrightarrow 3\,S_1$, such that $\xi_1 = 1$ and $v_1(x) = k_1 x^2$, and assume that $x_0 > 0$. Then, the solution of the ODE-model (2.34) is given by $\boldsymbol{x}(t) = x_0/(1 - k_1 x_0 t)$, and we have $\boldsymbol{x}(t) \to \infty$ as $t \to 1/(k_1 x_0) < \infty$.

[18] Here and in the following, $\|\cdot\|$ denotes the Euclidian norm on an n-dimensional vector space $\mathbb{R}^n$, i.e., $\|x\| = \sqrt{\sum_{i=1}^n x_i{}^2}$ for $x \in \mathbb{R}^n$.

2.4 The Relationship between Stochastic and Deterministic Reaction Kinetics

In the sequel, we study the relationship between the stochastic and the deterministic formulation of reaction kinetics. We start by examining the relation between the propensity a_μ and the rate v_μ of a reaction R_μ, see Subsection 2.4.1. That followed, we identify the continuous–deterministic process $\boldsymbol{x}(t)$ that satisfies eq. (2.33) as the large-size system limit of the discrete–stochastic process $\boldsymbol{X}(t)$ satisfying eq. (2.6) for every finite time t. This well-known limit behavior of the stochastic process $\boldsymbol{X}(t)$, known as its thermodynamic limit, was first pointed out by T. G. Kurtz in [63]. In Subsection 2.4.2, we recall Kurtz's Theorem that follows from his more general results on the limit behavior of Markov jump processes [123, 124].

The deterministic process $\boldsymbol{x}(t)$ is casually interpreted as describing the average system dynamics. By computing the evolution of the CME average, it is shown in Subsection 2.4.3 that this interpretation is in general incorrect. From Kurtz's result, however, one expects such relation to hold for general reaction systems whenever the species are present in large numbers and the system is sufficiently close to the thermodynamic limit. We study this relation more closely in Subsection 2.4.4, by seeking an approximate solution to the CME using a Wentzel–Kramers–Brillouin (WKB)-ansatz, which provides an alternative approach to derive the ODE-model of deterministic reaction kinetics as the large-size system limit of stochastic reaction kinetics, cf. [18]. This will also serve as a preliminary to Chapter 4, where the same techniques are applied to the large copy number subspace of a system in order to derive a hybrid stochastic–deterministic solution of the CME.

2.4.1 The Relation between Propensity and Rate of a Reaction

In stochastic reaction kinetics the number of firings of a reaction R_μ is modeled as a non-homogeneous Poisson process with intensity a_μ. Hence, $a_\mu(X)$ gives the expected number of R_μ firings per unit time if the system is in state X, see eq. (2.4). In deterministic reaction kinetics, the rate $v_\mu(x)$ gives the average number of R_μ firings per Ω and per unit time, given the system is in state $x = X/\Omega$. Thus, from the large-size system assumption underlying the continuous approximation of deterministic reaction kinetics (see Section 2.3), we infer that the reaction rate $v_\mu(x)$ approximates the Ω-scaled propensity $\alpha_\mu(x) := a_\mu(X = x\Omega)/\Omega$ for

$\Omega \gg 1$ and $X_i \gg 1$, $i = 1, \ldots, N$. Below, we study this relation more closely for the case of elementary reactions.

The rate v_μ of an elementary reaction R_μ is given by the law of mass action as

$$v_\mu(x) = k_\mu \prod_{i=1}^{N} (x_i)^{s^r_{\mu i}}, \tag{2.35}$$

where the rate constant k_μ, is independent of Ω, see Subsection 2.3.1. By eqs. (2.2) and (2.3), the propensity a_μ of an elementary reaction R_μ is given as

$$a_\mu(X) = c_\mu h_\mu(X) = \begin{cases} c_\mu \prod_{i=1}^{N} \dfrac{X_i!}{s^r_{\mu i}!(X_i - s^r_{\mu i})!} & \text{if } X_i \geq s^r_{\mu i} \text{ for all } i = 1, \ldots, N, \\ 0 & \text{otherwise,} \end{cases} \tag{2.36}$$

where we can rearrange $h_\mu(X)$ as follows

$$h_\mu(X) = \prod_{i=1}^{N} \frac{X_i!}{s^r_{\mu i}!(X_i - s^r_{\mu i})!} = \frac{1}{\prod_{j=1}^{N}\left(s^r_{\mu j}!\right)} \prod_{i=1}^{N} \prod_{s=0}^{s^r_{\mu i}-1} \left(X_i - s\right), \tag{2.37}$$

if $X_i \geq s^r_{\mu i}$ for all $i = 1, \ldots, N$. In order to compare $a_\mu(X)$ and $v_\mu(x)$, we substitute $X_i = x_i\Omega$ in eq. (2.36) and divide by Ω, which gives the Ω-scaled propensity $\alpha_\mu(x)$ as

$$\alpha_\mu(x) = \frac{a_\mu(x\Omega)}{\Omega} = \begin{cases} \dfrac{c_\mu \Omega^{|s^r_\mu|-1}}{\prod_{j=1}^{N}\left(s^r_{\mu j}!\right)} \prod_{i=1}^{N} \prod_{s=0}^{s^r_{\mu i}-1} \left(x_i - \dfrac{s}{\Omega}\right) & \text{if } x_i \geq \frac{s^r_{\mu i}}{\Omega} \text{ for all } i = 1, \ldots, N, \\ 0 & \text{otherwise,} \end{cases} \tag{2.38}$$

where $|s^r_\mu| = \sum_{i=1}^{N} s^r_{\mu i}$ denotes the molecularity of R_μ. The Ω-scaled propensity α_μ gives the expected number of R_μ firings per Ω and per unit time in stochastic reaction kinetics.

We observe that if $x_i = 0$, for some $i = 1, \ldots, N$, then the related factors in $\alpha_\mu(x)$ and $v_\mu(x)$ are the same, i.e., for $s^r_{\mu i}/\Omega = x_i = 0$, the corresponding factor in $\alpha_\mu(x)$ as well as in $v_\mu(x)$ is equal to one; otherwise, if $s^r_{\mu i}/\Omega > x_i = 0$, then both functions $\alpha_\mu(x)$ and $v_\mu(x)$ are equal to zero. Hence, we consider $x_i > 0$ for every $i = 1, \ldots, N$ and assume that $x_i = X_i/\Omega$ remains constant as $\Omega \to \infty$ and $X_i \to \infty$ (which is the usual definition of the thermodynamic limit, see [10, 64]). As the stoichiometric coefficients $s^r_{\mu i}$ are some non-negative integers (usually

$s^r_{\mu i} \leq 2$), we find that $x_i \geq s^r_{\mu i}/\Omega$ is always satisfied for sufficiently large Ω. The Ω-scaled propensity $\alpha_\mu(x)$ is thus given by eq. (2.38) as

$$\begin{aligned}\alpha_\mu(x) &= \widehat{c}_\mu \prod_{i=1}^{N} \prod_{s=0}^{s^r_{\mu i}-1} \left(x_i - \frac{s}{\Omega}\right) = \widehat{c}_\mu \prod_{i=1}^{N} \left[(x_i)^{s^r_{\mu i}} \prod_{s=0}^{s^r_{\mu i}-1} \left(1 - \frac{s}{x_i \Omega}\right)\right] \\ &= \widehat{c}_\mu \prod_{i=1}^{N} (x_i)^{s^r_{\mu i}} \left(1 + \mathcal{O}(\Omega^{-1})\right), \end{aligned} \tag{2.39}$$

as $\Omega \to \infty$ and $X_i \to \infty$ for every $i = 1, \ldots, N$, where we substituted

$$\widehat{c}_\mu := c_\mu \frac{\Omega^{|s^r_\mu|-1}}{\prod_{j=1}^{N} \left(s^r_{\mu j}!\right)}. \tag{2.40}$$

Therefore, the leading order terms of $\alpha_\mu(x)$ are of the same form as the rate function $v_\mu(x)$, compare eqs. (2.39) and (2.35), with the factor of proportionality given by $k_\mu = \widehat{c}_\mu$.[19] Furthermore, we observe that if $s^r_{\mu i} \leq 1$ for all $i = 1, \ldots, N$, then eq. (2.39) holds without $\mathcal{O}(\Omega^{-1})$.

In summary, the above results show that the rate v_μ of an elementary reaction R_μ is the leading order approximation of the related Ω-scaled propensity α_μ in the thermodynamic limit. If the partial order of all reactant species is less than two, i.e., $s^r_{\mu i} \leq 1$ for all $i = 1, \ldots, N$, then v_μ and α_μ are identical.

The Relation between Specific Probability Rate Constant and Reaction Rate Constant

From eqs. (2.39) and (2.40) we infer that the specific probability rate constant c_μ of an elementary reaction R_μ is related to the reaction rate constant k_μ of R_μ as follows

$$c_\mu = k_\mu \frac{\prod_{i=1}^{N} s^r_{\mu i}!}{\Omega^{|s^r_\mu|-1}}. \tag{2.41}$$

Since k_μ is independent of Ω, this relation points out the Ω-dependence of c_μ.

Only for unimolecular reactions, i.e., $|s^r_\mu| = 1$, the specific probability rate constant is independent of Ω. In a unimolecular reaction, the molecules of a particular species undergo some prescribed change. It is an empirical fact that for any such reaction encountered in practice, the rate constant will depend on structural properties of the molecule, and possibly also on the system temperature, however, it will not depend on the volume of the system, as comprised by Ω, see [64].

[19] Since k_μ is Ω-independent, the relation $k_\mu = \widehat{c}_\mu$ implies that $\widehat{c}_\mu$ is also independent of Ω.

For bimolecular reactions R_2, i.e., $|s_\mu^r| = 2$, the specific probability rate constant is proportional to the inverse of Ω. This inverse Ω-dependence arises from the collision probability comprised by $c_\mu \delta t$, because the probability that two molecules will collide in the next δt decreases linearly in a well-stirred system if the confining volume increases, cf. Subsection 2.2.1 and Example 2.2.2.

For a source-like reaction, i.e., $|s_\mu^r| = 0$, $c_\mu \delta t$ gives the probability that new molecules will be introduced at some uniformly random location inside the system in the next δt. Eq. (2.41) states that the specific probability rate constant *must* have the form $c_\mu = k_\mu \Omega$ in order to behave reasonable in the thermodynamic limit, cf. [64].

The Dependence of a Reaction Propensity on the System Size

The reaction propensity a_μ of an elementary reaction R_μ with respect to the rate constant k_μ is given by eqs. (2.36), (2.37) and (2.41) as

$$a_\mu(X) = \begin{cases} \dfrac{k_\mu}{\Omega^{|s_\mu^r|-1}} \displaystyle\prod_{i=1}^{N} \prod_{s=0}^{s_{\mu i}^r - 1} \left(X_i - s\right) & \text{if } X_i \geq s_{\mu i}^r \text{ for all } i = 1, \ldots, N, \\ 0 & \text{otherwise.} \end{cases} \tag{2.42}$$

In Table 2.3, the propensities of all relevant elementary reactions are given in terms of k_μ and compared with the related rate functions. As can bee nicely seen from Table 2.3 and eq. (2.42), a propensity behaves like $a_\mu(X) = \mathcal{O}(\Omega)$ as $\Omega \to \infty$ and $X_i \to \infty$ while $x_i = X_i/\Omega$ remains constant for all $i = 1, \ldots, N$, i.e., a propensity diverges linearly to leading order as the system approaches the thermodynamic limit. In the following, we denote this Ω-dependence of a propensity by writing a_μ^Ω.

Equivalently, the Ω-scaled propensity α_μ of an elementary reaction R_μ, given in terms of k_μ as

$$\alpha_\mu(x) = \frac{a_\mu^\Omega(x\Omega)}{\Omega} = \begin{cases} k_\mu \displaystyle\prod_{i=1}^{N} \prod_{s=0}^{s_{\mu i}^r - 1} \left(x_i - \frac{s}{\Omega}\right) & \text{if } x_i \geq \frac{s_{\mu i}^r}{\Omega} \text{ for all } i = 1, \ldots, N, \\ 0 & \text{otherwise,} \end{cases}$$

either (i) remains constant, if $s_{\mu i}^r \leq 1$ for all $i = 1, \ldots, N$, or (ii) converges to a constant value, if some $s_{\mu i}^r > 1$, in the thermodynamic limit. Namely, for all $x \in \mathbb{R}_{\geq 0}^N$, we have $\alpha_\mu(x) = \nu_\mu(x)$ in case (i), or $\alpha_\mu(x) \to \nu_\mu(x)$ in case (ii), since then $\alpha_\mu(x) = \nu_\mu(x) + \mathcal{O}(\Omega^{-1})$ by eq. (2.39).

Table 2.3: *Propensity and rate functions of all relevant elementary reactions with respect to rate constants k_μ and the conversion factor Ω, e.g., the system volume V times the Avogadro constant $N_A \approx 6.02 \times 10^{23}\ \mathrm{mol}^{-1}$.*

Order	Reaction	Propensity	Rate
0^{th}	$\varnothing \xrightarrow{k_0} \cdots$	$a_0(X) = k_0\Omega$	$v_0(x) = k_0$
1^{st}	$S_i \xrightarrow{k_1} \cdots$	$a_1(X) = k_1 X_i$	$v_1(x) = k_1 x_i$
2^{nd}	$S_i + S_j \xrightarrow{k_{2a}} \cdots$	$a_{2a}(X) = \frac{k_{2a}}{\Omega} X_i X_j$, with $i \neq j$	$v_{2a}(x) = k_{2a} x_i x_j$
	$2\,S_i \xrightarrow{k_{2b}} \cdots$	$a_{2b}(X) = \begin{cases} \frac{k_{2b}}{\Omega} X_i(X_i - 1) & \text{if } X_i \geq 1, \\ 0 & \text{otherwise} \end{cases}$	$v_{2b}(x) = k_{2b} {x_i}^2$

2.4.2 The Thermodynamic Limit

In the following, we explicitly account for the Ω-dependence of a propensity a_μ^Ω and study the limit behavior of the Ω-dependent Markov jump process $\boldsymbol{X}^\Omega$ described by

$$\boldsymbol{X}^\Omega(t) = \boldsymbol{X}^\Omega(0) + \sum_{\mu=1}^{M} \xi_\mu \boldsymbol{\mathcal{P}}_\mu\left(\int_0^t a_\mu^\Omega\big(\boldsymbol{X}^\Omega(s)\big)\,\mathrm{d}s\right), \tag{2.43}$$

where the propensity a_μ^Ω of a reaction R_μ is given by eq. (2.42), ξ_μ denotes the corresponding net change and $\boldsymbol{\mathcal{P}}_\mu$, $\mu = 1, \ldots, M$, are independent, unit Poisson processes. As before, $\boldsymbol{X}^\Omega(t)$ represents the state of the chemical system at time t in terms of number of molecules $X = (X_1, \ldots, X_N)^{\mathrm{T}} \in \mathbb{N}_0^N$ of the species.

In the previous subsection, we have seen that in the thermodynamic limit, defined as the limit where the number of molecules of all species and the system volume approach infinity, i.e., $X_i \to \infty$ for all $i = 1, \ldots, N$ and $\Omega \to \infty$, while the species concentrations $x = X/\Omega$ remain constant, every propensity $a_\mu^\Omega(X = x\Omega)$ diverges linearly to leading order. The Ω-scaled propensity $\alpha_\mu(x) = a_\mu^\Omega(x\Omega)/\Omega$ hence converges in the thermodynamic limit to a constant value, which is equal to the related rate $\nu_\mu(x)$. Based on this observation, one probably expects a similar limit behavior for the sequence of normalized process $\boldsymbol{x}^\Omega := \boldsymbol{X}^\Omega/\Omega$. Indeed, as pointed out by T. G. Kurtz [63], his results in [123, 124] show that in the thermodynamic limit $\boldsymbol{x}^\Omega$ approaches the related process of deterministic reaction kinetics, given by

$$\boldsymbol{x}(t) = \boldsymbol{x}(0) + \sum_{\mu=1}^{M} \xi_\mu \int_0^t \nu_\mu\big(\boldsymbol{x}\big)\,\mathrm{d}s, \tag{2.44}$$

where the rates ν_μ are given by eq. (2.31). This important result on the relationship between the stochastic and the deterministic modeling approach is stated in the following theorem; for the sake of its importance and clarity, we also recall its proof.

Theorem 2.4.1 (T. G. Kurtz [123]). *Suppose $\boldsymbol{X}^\Omega(t)$ satisfies eq. (2.43) and $\boldsymbol{x}(t)$ satisfies eq. (2.44) for every $t \geq 0$. If*

$$\lim_{\Omega\to\infty} \frac{\boldsymbol{X}^\Omega(0)}{\Omega} = \boldsymbol{x}(0), \tag{2.45}$$

then

$$\mathbb{P}\left[\lim_{\Omega\to\infty} \sup_{s\leq t} \left\| \frac{\boldsymbol{X}^\Omega(t)}{\Omega} - \boldsymbol{x}(t) \right\| = 0\right] = 1,$$

for all finite times $t \geq 0$.

Note. In Theorem (2.4.1), global existence of the solutions of eqs. (2.43) and (2.44) is assumed. However, if these solutions blow up and reach infinity in some finite time T, say, almost-sure convergence of $\boldsymbol{x}^{\Omega}(t) := \boldsymbol{X}^{\Omega}(t)/\Omega$ to $\boldsymbol{x}(t)$ still holds for every $t < T$. The proof of Kurtz's Theorem (2.4.1), as given below, mainly follows the results presented in [87, 88, 125]. It is based on an application of the strong law of large numbers (SLLN) to Poisson processes and a consequential proposition, as well as a variant of the integral form of the Grönwall–Bellmann inequality:

Theorem 2.4.2 (SLLN for Poisson Processes [126]). *Let $\boldsymbol{\mathcal{P}}$ be a Poisson process with intensity λ, then*

$$\mathbb{P}\left[\lim_{t\to\infty}\frac{\boldsymbol{\mathcal{P}}(t)}{t} = \lambda\right] = 1.$$

Proposition 2.4.3 (D. F. Anderson & T. G. Kurtz [88]). *Let $\boldsymbol{\mathcal{P}}$ be a unit Poisson process, then*

$$\mathbb{P}\left[\lim_{\Omega\to\infty}\sup_{\lambda\le\overline{\lambda}}\left|\frac{1}{\Omega}\boldsymbol{\mathcal{P}}(\Omega\lambda) - \lambda\right| = 0\right] = 1,$$

for any $\overline{\lambda} \geq 0$.

Lemma 2.4.4 (An Integral Form of the Grönwall–Bellmann Inequality [87]). *Let g be a Borel-measurable function that is bounded on bounded intervals, satisfying $g(t) \leq A + B\int_0^t g(s)\,\mathrm{d}s$ for all $t \geq 0$ and some constants $A, B \in \mathbb{R}_{\geq 0}$. Then,*

$$g(t) \leq A\mathrm{e}^{Bt},$$

for all $t \geq 0$.

Proof of Theorem 2.4.1. Dividing both sides of eq. (2.43) by $\Omega > 0$, we find that for all $t \geq 0$ the Ω-scaled process $\boldsymbol{x}^{\Omega}(t) = \boldsymbol{X}^{\Omega}(t)/\Omega$ satisfies

$$\begin{aligned}\boldsymbol{x}^{\Omega}(t) &= \boldsymbol{x}^{\Omega}(0) + \sum_{\mu=1}^{M}\frac{\xi_\mu}{\Omega}\boldsymbol{\mathcal{P}}_\mu\left(\int_0^t a_\mu^{\Omega}\big(\boldsymbol{x}^{\Omega}(s)\Omega\big)\,\mathrm{d}s\right)\\ &= \boldsymbol{x}^{\Omega}(0) + \sum_{\mu=1}^{M}\frac{\xi_\mu}{\Omega}\boldsymbol{\mathcal{P}}_\mu\left(\Omega\int_0^t \alpha_\mu\big(\boldsymbol{x}^{\Omega}(s)\big)\,\mathrm{d}s\right),\end{aligned}$$

where we substituted $a_\mu^{\Omega}(x\Omega) = \Omega\alpha_\mu(x)$ in the second equality. Now, consider the centered processes $\widetilde{\boldsymbol{\mathcal{P}}}_\mu(\Omega\Lambda_\mu(t)) := \boldsymbol{\mathcal{P}}_\mu(\Omega\Lambda_\mu(t)) - \Omega\Lambda_\mu(t)$ with $\Lambda_\mu(t) := \int_0^t \alpha_\mu(\boldsymbol{x}^{\Omega}(s))\,\mathrm{d}s$ for every $\mu = 1, \ldots, M$, such that

$$\boldsymbol{x}^{\Omega}(t) = \boldsymbol{x}^{\Omega}(0) + \sum_{\mu=1}^{M}\frac{\xi_\mu}{\Omega}\widetilde{\boldsymbol{\mathcal{P}}}_\mu\big(\Omega\Lambda_\mu(t)\big) + \sum_{\mu=1}^{M}\xi_\mu\int_0^t \alpha_\mu\big(\boldsymbol{x}^{\Omega}(s)\big)\,\mathrm{d}s,$$

for all $t \geq 0$. Since the reaction rates $v_\mu(x)$ are Ω-independent and further $\alpha_\mu(x) \to v_\mu(x)$ for every $x \in \mathbb{R}^N_{\geq 0}$ in the thermodynamic limit, it follows that for all $t \geq 0$

$$\left\|\boldsymbol{x}^\Omega(t) - \boldsymbol{x}(t)\right\| \leq \left\|\boldsymbol{x}^\Omega(0) - \boldsymbol{x}(0)\right\| + \left\|\sum_{\mu=1}^{M} \frac{\xi_\mu}{\Omega} \widetilde{\boldsymbol{\mathcal{P}}}_\mu\big(\Omega \Lambda_\mu(t)\big)\right\| + \varepsilon^1_\Omega + \left\|\int_0^t f\big(\boldsymbol{x}^\Omega(s)\big) - f\big(\boldsymbol{x}(s)\big)\,\mathrm{d}s\right\|, \quad (2.46)$$

where $f(x) := \sum_{\mu=1}^M \xi_\mu v_\mu(x)$ and $\varepsilon^1_\Omega := \left\|\sum_{\mu=1}^M \xi_\mu \int_0^t \alpha_\mu(\boldsymbol{x}^\Omega(s)) - v_\mu(\boldsymbol{x}^\Omega(s))\,\mathrm{d}s\right\|$, satisfying $\varepsilon^1_\Omega \to 0$ as $\Omega \to \infty$. For fixed $t \geq 0$, we know that $\boldsymbol{x}^\Omega(t)$ will be contained in some compact set $K^N_\Omega \subset E^N_\Omega$, where $E^N_\Omega := \mathbb{R}^N_{\geq 0} \cap \{n/\Omega \colon n \in \mathbb{N}^N_0\}$, such that

$$\overline{\alpha}_\mu := \sup_{x \in K^N_\Omega} \alpha_\mu(x) < \infty,$$

for all $\mu = 1, \ldots, M$. Hence,

$$\varepsilon^2_\Omega := \left\|\sum_{\mu=1}^{M} \frac{\xi_\mu}{\Omega} \widetilde{\boldsymbol{\mathcal{P}}}_\mu\big(\Omega \Lambda_\mu(t)\big)\right\| \leq \sum_{\mu=1}^{M} \frac{\|\xi_\mu\|}{\Omega} \sup_{s \leq t} \left|\widetilde{\boldsymbol{\mathcal{P}}}_\mu\big(\Omega \Lambda_\mu(s)\big)\right| \quad (2.47a)$$

$$\leq \sum_{\mu=1}^{M} \frac{\|\xi_\mu\|}{\Omega} \sup_{s \leq t} \left|\boldsymbol{\mathcal{P}}_\mu\big(\Omega \overline{\alpha}_\mu s\big) + \Omega \overline{\alpha}_\mu s\right| \quad (2.47b)$$

$$\leq \sum_{\mu=1}^{M} \|\xi_\mu\| \left(\frac{1}{\Omega} \boldsymbol{\mathcal{P}}_\mu\big(\Omega \overline{\alpha}_\mu t\big) + \overline{\alpha}_\mu t\right). \quad (2.47c)$$

We recall that the superposition of M independent Poisson processes $\boldsymbol{\mathcal{P}}_1(t), \ldots, \boldsymbol{\mathcal{P}}_M(t)$ with intensities $\lambda_1, \ldots, \lambda_M$, as in the right sides of inequalities (2.47b) and (2.47c), for instance, is a Poisson process $\boldsymbol{\mathcal{P}}(t) := \sum_{\mu=1}^M \boldsymbol{\mathcal{P}}_\mu(t)$ with intensity $\lambda = \sum_{\mu=1}^M \lambda_\mu$, cf. [83]. Applying the SLLN for Poisson processes (Theorem 2.4.2), we thus obtain

$$\lim_{\Omega \to \infty} \sum_{\mu=1}^{M} \|\xi_\mu\| \left(\frac{1}{\Omega} \boldsymbol{\mathcal{P}}_\mu\big(\Omega \overline{\alpha}_\mu t\big) + \overline{\alpha}_\mu t\right) = \sum_{\mu=1}^{M} \|\xi_\mu\| 2\overline{\alpha}_\mu t = \sum_{\mu=1}^{M} \|\xi_\mu\| \lim_{\Omega \to \infty} \left(\frac{1}{\Omega} \boldsymbol{\mathcal{P}}_\mu\big(\Omega \overline{\alpha}_\mu t\big) + \overline{\alpha}_\mu t\right),$$

with probability one, which shows that we can interchange limit and summation on the right side of inequality (2.47c). Since inequalities (2.47b) and (2.47c) further imply dominated convergence, we can also interchange limit and summation on the right of inequality (2.47a) and apply Proposition 2.4.3, i.e.,

$$\lim_{\Omega \to \infty} \sum_{\mu=1}^{M} \frac{\|\xi_\mu\|}{\Omega} \sup_{\lambda_\mu \leq \overline{\lambda}_\mu} \left|\widetilde{\boldsymbol{\mathcal{P}}}_\mu\big(\Omega \lambda_\mu\big)\right| = \sum_{\mu=1}^{M} \|\xi_\mu\| \lim_{\Omega \to \infty} \sup_{\lambda_\mu \leq \overline{\lambda}_\mu} \left|\frac{1}{\Omega} \boldsymbol{\mathcal{P}}_\mu\big(\Omega \lambda_\mu\big) - \lambda_\mu\right| = 0,$$

with probability one, where $\overline{\lambda}_\mu := \overline{\alpha}_\mu t$ for each $\mu = 1, \dots, M$. Therefore, we have

$$\mathbb{P}\Big[\lim_{\Omega\to\infty} \varepsilon_\Omega^2 = 0\Big] = 1,$$

for all $t \geq 0$.

Clearly, the right hand side $f(\,\cdot\,) = \sum_{\mu=1}^{M} \xi_\mu v_\mu(\,\cdot\,)$ in eq. (2.46) is Lipschitz continuous on any bounded set $K^N \subset \mathbb{R}^N_{\geq 0}$, i.e., there exists a constant $L_K > 0$ such that

$$\big\|f(x) - f(y)\big\| \leq L_{K^N}\big\|x - y\big\|,$$

for all $x, y \in K^N$. Putting things together, we thus have

$$\big\|\boldsymbol{x}^\Omega(t) - \boldsymbol{x}(t)\big\| \leq \big\|\boldsymbol{x}^\Omega(0) - \boldsymbol{x}(0)\big\| + \varepsilon_\Omega + L_{K^N}\int_0^t \big\|\boldsymbol{x}^\Omega(s) - \boldsymbol{x}(s)\big\|\,\mathrm{d}s,$$

for all $t \geq 0$, where $\varepsilon_\Omega := \varepsilon_\Omega^1 + \varepsilon_\Omega^2 \to 0$ as $\Omega \to \infty$. Applying the Grönwall–Bellmann inequality (Lemma 2.4.4) with $g(t) := \|\boldsymbol{x}^\Omega(t) - \boldsymbol{x}(t)\|$, $A := \|\boldsymbol{x}^\Omega(0) - \boldsymbol{x}(0)\| + \varepsilon_\Omega$ and $B := L_{K^N}$, we get

$$\big\|\boldsymbol{x}^\Omega(t) - \boldsymbol{x}(t)\big\| \leq \Big(\big\|\boldsymbol{x}^\Omega(0) - \boldsymbol{x}(0)\big\| + \varepsilon_\Omega\Big)\mathrm{e}^{L_{K^N}t},$$

for all $t \geq 0$, and the statement of the theorem follows from condition (2.45) and the fact that $\lim_{\Omega\to\infty} \varepsilon_\Omega = 0$ with probability one. □

2.4.3 Averaging of the CME

The deterministic process $\boldsymbol{x}(t)$ of classical reaction kinetics is customarily interpreted as describing the average system dynamics. In the following, we study the origin of such interpretation by deriving evolution equations of the first moments of the discrete–stochastic process $\boldsymbol{X}^\Omega(t)$, whose PDF $P(X;t)$ evolves in time as described by the CME

$$\frac{\partial}{\partial t}P(X;t) = \sum_{\mu=1}^{M} a_\mu^\Omega(X - \xi_\mu)P(X - \xi_\mu;t) - a_\mu^\Omega(X)P(X;t). \tag{2.48}$$

We recall that the expected value of any function f of $\boldsymbol{X}^\Omega \in \mathbb{N}_0^N$ at time t is defined as

$$\mathbb{E}\big[f(\boldsymbol{X}^\Omega)\big] := \sum_{X_1=0}^{\infty}\cdots\sum_{X_N=0}^{\infty} f(X)P(X;t) = \sum_{X\geq 0} f(X)P(X;t). \tag{2.49}$$

Multiplication of the CME with $f(X)$ and subsequent summation over X gives

$$\sum_{X\geq 0} f(X)\frac{\partial}{\partial t}P(X;t) = \sum_{X\geq 0} f(X)\left[\sum_{\mu=1}^{M} a_\mu^\Omega(X - \xi_\mu)P(X - \xi_\mu;t) - a_\mu^\Omega(X)P(X;t)\right]. \tag{2.50}$$

We note that further manipulation on the infinite series in the right hand side of eq. (2.50) is only justified under absolute convergence of the series $\sum_{X\geq 0} f(X)a_\mu^\Omega(X-\xi_\mu)P(X-\xi_\mu;t)$ and $\sum_{X\geq 0} f(X)a_\mu^\Omega(X)P(X;t)$, where the propensities a_μ^Ω are polynomial functions of X. Therefore, in the following[20], we let f be a polynomial function and assume that sufficiently many moments[21] of $\boldsymbol{X}^\Omega$ exist at time t, cf. [127], such that eq. (2.50) can be written as

$$\sum_{X\geq 0} f(X)\frac{\partial}{\partial t}P(X;t) = \sum_{\mu=1}^{M}\left[\sum_{X\geq 0} f(X)a_\mu^\Omega(X-\xi_\mu)P(X-\xi_\mu;t) - \sum_{X\geq 0} f(X)a_\mu^\Omega(X)P(X;t)\right]. \tag{2.51}$$

Next, we show that the lower bound of the first sum in the right hand side of eq. (2.51) can be changed to ξ_μ, i.e.,

$$\sum_{X\geq 0} f(X)a_\mu^\Omega(X-\xi_\mu)P(X-\xi_\mu;t) = \sum_{X\geq \xi_\mu} f(X)a_\mu^\Omega(X-\xi_\mu)P(X-\xi_\mu;t), \tag{2.52}$$

for every $\mu = 1,\ldots,M$. For that purpose, we consider the variables X_i, $i = 1,\ldots,N$, separately and distinguish between the three possible cases: (i) $\xi_{\mu i} = 0$, (ii) $\xi_{\mu i} > 0$, or (iii) $\xi_{\mu i} < 0$.

(i) If $\xi_{\mu i} = 0$, the lower bound of summation in eq. (2.52) remains unchanged.

(ii) If $\xi_{\mu i} > 0$, the terms for $X_i \in I := \{0,\ldots,\xi_{\mu i}-1\}$ are omitted in the right hand side of eq. (2.52). But these are equal to zero, since for every $X_i \in I$ we have

$$X_i < \xi_{\mu i} \;\Rightarrow\; X_i - \xi_{\mu i} < 0 \leq s_{\mu i}^r \overset{\text{eq. (2.3)}}{\Rightarrow}\; a_\mu^\Omega(X-\xi_\mu) = 0.$$

(iii) If $\xi_{\mu i} < 0$, the terms for $X_i \in I := \{\xi_{\mu i},\ldots,-1\}$ are added in the right hand side of eq. (2.52). But these are equal to zero, since by definition of the net changes we have

$$\xi_{\mu i} = s_{\mu i}^p - s_{\mu i}^r \;\Leftrightarrow\; -\xi_{\mu i} = s_{\mu i}^r - s_{\mu i}^p \leq s_{\mu i}^r,$$

such that for every $X_i \in I$

$$X_i < 0 \;\Rightarrow\; X_i - \xi_{\mu i} < -\xi_{\mu i} \leq s_{\mu i}^r \overset{\text{eq. (2.3)}}{\Rightarrow}\; a_\mu^\Omega(X-\xi_\mu) = 0.$$

[20] Alternatively, we could make the physically plausible assumption that $\boldsymbol{X}^\Omega$ can only take values on some finite subset $K^N \subset \mathbb{N}_0^N$; and set $a_\mu^\Omega(X) = 0$ and $P(X;t) = 0$ for all $X \notin K^N$, and $a_\mu^\Omega(X) = 0$, if $(X+\xi_\mu) \notin K^N$, to further ensure that $\boldsymbol{X}^\Omega$ can not leave K^N. Then, the series in eq. (2.50) would be effectively finite sums over all $X \in K^N$.

[21] Given a vector $m = (m_1,\ldots,m_N)^{\mathrm{T}} \in \mathbb{N}_0^N$, then the (uncentered) moment of some $\boldsymbol{X} \in \mathbb{N}_0^N$ associated with m is given by $\mathbb{E}[\boldsymbol{X}^m]$, where $\boldsymbol{X}^m := (\boldsymbol{X}_1)^{m_1}\cdot\ldots\cdot(\boldsymbol{X}_N)^{m_N}$. We call $|m| := \sum_{i=1}^N m_i$ the *order* of the moment. If $N = 2$, for instance, then the moments of first order are $\mathbb{E}[\boldsymbol{X}_1]$ and $\mathbb{E}[\boldsymbol{X}_2]$, and the second order moments are $\mathbb{E}[\boldsymbol{X}_1\boldsymbol{X}_1]$, $\mathbb{E}[\boldsymbol{X}_1\boldsymbol{X}_2]$ and $\mathbb{E}[\boldsymbol{X}_2\boldsymbol{X}_2]$.

Therefore, equality (2.52) holds and since further

$$\sum_{X\geq\xi_\mu} f(X)a_\mu^\Omega(X-\xi_\mu)P(X-\xi_\mu;t) = \sum_{X\geq 0} f(X+\xi_\mu)a_\mu^\Omega(X)P(X;t),$$

for every $\mu = 1,\dots,M$, we find that eq. (2.51) can be written as

$$\sum_{X\geq 0} f(X)\frac{\partial}{\partial t}P(X;t) = \sum_{\mu=1}^{M}\sum_{X\geq 0}\Big[\big(f(X+\xi_\mu)-f(X)\big)a_\mu^\Omega(X)P(X;t)\Big].$$

The time-evolution of the expectation of $f(\boldsymbol{X}^\Omega)$ is hence given by

$$\frac{\partial}{\partial t}\mathbb{E}\big[f(\boldsymbol{X}^\Omega)\big] = \sum_{\mu=1}^{M}\mathbb{E}\Big[\big(f(\boldsymbol{X}^\Omega+\xi_\mu)-f(\boldsymbol{X}^\Omega)\big)a_\mu^\Omega(\boldsymbol{X}^\Omega)\Big]. \tag{2.53}$$

Remark 2.4.1. The above result shows that the total probability mass $\sum_{X\geq 0}P(X;t)\equiv 1$ is conserved by the CME, i.e., by eqs. (2.49) and (2.53) with $f(\,\cdot\,)\equiv 1$ we have

$$\frac{\partial}{\partial t}\sum_{X\geq 0}P(X;t) = 0.$$

By taking $f(X) = X$ in eq. (2.53), the time-evolution of the expected value of $\boldsymbol{X}^\Omega$ is given as

$$\frac{\partial}{\partial t}\mathbb{E}\big[\boldsymbol{X}^\Omega\big] = \sum_{\mu=1}^{M}\xi_\mu\mathbb{E}\big[a_\mu^\Omega(\boldsymbol{X}^\Omega)\big]. \tag{2.54}$$

It should be noted that this is an exact result, valid as long as the underlying reaction model is properly formulated such that the corresponding moments of $\boldsymbol{X}^\Omega$ in eq. (2.54) exist, cf. [127]. Furthermore, observe the similarity of eq. (2.54) to the ODE-model of deterministic reaction kinetics. This becomes more obvious if we take $f(X) = X/\Omega$, such that eq. (2.53) gives the time-evolution of the expected value of the species concentrations $\boldsymbol{x}^\Omega = \boldsymbol{X}^\Omega/\Omega$ as predicted by the CME, i.e.,

$$\frac{\partial}{\partial t}\mathbb{E}\big[\boldsymbol{x}^\Omega\big] = \sum_{\mu=1}^{M}\frac{\xi_\mu}{\Omega}\mathbb{E}\big[a_\mu^\Omega(\Omega\boldsymbol{x}^\Omega)\big] = \sum_{\mu=1}^{M}\xi_\mu\mathbb{E}\big[\alpha_\mu(\boldsymbol{x}^\Omega)\big], \tag{2.55}$$

where we employed linearity of the expectation and substituted the Ω-scaled propensities $\alpha_\mu(x) = a_\mu^\Omega(x\Omega)/\Omega$ in the last equality. Since further $\alpha_\mu(x) = \nu_\mu(x) + \mathcal{O}(\Omega^{-1})$, one might infer from eq. (2.55) that the deterministic process $\boldsymbol{x}(t)$ of classical reaction kinetics approximates the CME average $\mathbb{E}[\boldsymbol{x}^\Omega]$.

However, we note that eq. (2.55) does in general not yield a closed system of equations for the average $\mathbb{E}[\boldsymbol{x}^\Omega]$ due to the unknown expectation $\mathbb{E}[\alpha_\mu(\boldsymbol{x}^\Omega)]$. Only if all propensities $\alpha_\mu(\boldsymbol{x}^\Omega)$ were constant or linear functions of $\boldsymbol{x}^\Omega$ (as it is the case for zero and first order reactions, respectively), then we could put $\mathbb{E}[\alpha_\mu(\boldsymbol{x}^\Omega)] = \alpha_\mu(\mathbb{E}[\boldsymbol{x}^\Omega])$ and completely express the right hand side of eq. (2.55) in terms of $\mathbb{E}[\boldsymbol{x}^\Omega]$. In this case, eq. (2.55) would thus constitute a closed system of equations for $\mathbb{E}[\boldsymbol{x}^\Omega]$, i.e.,

$$\frac{\partial}{\partial t}\mathbb{E}[\boldsymbol{x}^\Omega] = \sum_{\mu=1}^{M} \xi_\mu \alpha_\mu\big(\mathbb{E}[\boldsymbol{x}^\Omega]\big),$$

which is equaivalent to the ODE-model of classical reaction kinetics.

As soon as any reaction R_μ is of second or higher order, however, this will generally not be the case. Because then the corresponding propensity $\alpha_\mu(\boldsymbol{x}^\Omega)$ is a nonlinear function of $\boldsymbol{x}^\Omega$ for which $\mathbb{E}[\alpha_\mu(\boldsymbol{x}^\Omega)] \neq \alpha_\mu(\mathbb{E}[\boldsymbol{x}^\Omega])$ in general. For example, consider a bimolecular reaction R_2, whose propensity is of the form $\alpha_2(\boldsymbol{x}^\Omega) = k_2 \boldsymbol{x}_i^\Omega \boldsymbol{x}_j^\Omega$. Since $\mathbb{E}[\boldsymbol{x}_i^\Omega \boldsymbol{x}_j^\Omega] \neq \mathbb{E}[\boldsymbol{x}_i^\Omega]\mathbb{E}[\boldsymbol{x}_j^\Omega]$, eq. (2.55) would then not be closed, but would include the second moment $\mathbb{E}[\boldsymbol{x}_i^\Omega \boldsymbol{x}_j^\Omega]$ on the right hand side. Of course, the evolution equation of $\mathbb{E}[\boldsymbol{x}_i^\Omega \boldsymbol{x}_j^\Omega]$ could be derived from eq. (2.53) by putting $f(X) = X_i X_j / \Omega^2 = x_i x_j$. But the right hand side of the resulting equation would then introduce moments of order three. In general, eq. (2.55) has thus to be regarded as an open-ended hierarchy of moment equations [128].

The idea of *moment closure methods* is to approximate the exact, open system of evolution-equations of all moments up to some order n by a closed system, where the unknown higher moments are estimated from a small number of stochastic simulations, e.g., [129], or replaced by nonlinear functions of the first n moments, e.g., [127, 130, 131]. The common problems of moment closure methods are, however, to *a priori* predict an adequate truncation order n and that the produced ODE-system can become very stiff as moments of higher order are considered, cf. [127]. Usually, the first issue is tackled by comparing the results obtained for increasing n. Such an *a posteriori* procedure, however, can fail if higher order truncations result in a reduced accuracy of the lower order moments, which depends on the chemical system under consideration and is yet not studied on theoretical grounds [127, 129].

Obviously, in deterministic reaction kinetics no distinction is made between the expectation of products and the product of expectations, i.e., independence of the random variables is automatically assumed in a deterministic context, such that $\mathbb{E}[\boldsymbol{x}_i^\Omega \boldsymbol{x}_j^\Omega] = \mathbb{E}[\boldsymbol{x}_i^\Omega]\mathbb{E}[\boldsymbol{x}_j^\Omega]$, for instance, in the case of bimolecular reactions. This assumption of zero covariance nullifies the effects of correlations ($i \neq j$) and the effects of random fluctuations (if $i = j$). However, from

Kurtz's Theorem, studied in the previous subsection, we infer that $\mathbb{E}[\boldsymbol{x}_i^\Omega \boldsymbol{x}_j^\Omega] \approx \mathbb{E}[\boldsymbol{x}_i^\Omega]\mathbb{E}[\boldsymbol{x}_j^\Omega]$ or more generally $\mathbb{E}[\alpha_\mu(\boldsymbol{x}^\Omega)] \approx \alpha_\mu(\mathbb{E}[\boldsymbol{x}^\Omega])$ can be expected to hold for general reaction networks if the species are present in large numbers (close to the thermodynamic limit). In this case, the CME average should thus be well-approximated by the corresponding deterministic process of classical reaction kinetics.

2.4.4 Leading-Order WKB-Approximation for the CME

In the following, we study the aforementioned implications more closely. Starting point is a system-size related scaling of the CME with $\varepsilon := \Omega^{-1}$ as a parameter of smallness, motivated by the classical scaling in deterministic reaction kinetics. By applying a WKB-approximation [18, 65] to the resulting ε-scaled CME and further use of Laplace's method of integral approximation (see Appendix B), we are then able to demonstrate that the ODE-model of deterministic reaction kinetics approximates the evolution of the CME average to leading order as $\varepsilon \to 0$.

System-Size Scaled CME

Analogous to classical reaction kinetics (see Section 2.3), we assume that the number of molecules X_i, $i = 1, \ldots, N$, are large such that the state space can be represented by real numbers $x = (x_1, \ldots, x_N)^{\mathrm{T}} \in \mathbb{R}^N_{\geq 0}$, where

$$x_i := \varepsilon \cdot X_i \quad (i = 1, \ldots, N), \tag{2.56}$$

with $\varepsilon \ll 1$ denoting some scaling factor. By taking $\varepsilon = \Omega^{-1}$, eq. (2.56) becomes identical to the large-size system assumption underlying the continuous approximation in the deterministic framework. The ε-scaled propensity of an elementary reaction is hence given by $\alpha_\mu(x) = \varepsilon \cdot a_\mu^\varepsilon(X = x/\varepsilon)$, see Subsection 2.4.1. We further assume that the propensity functions α_μ are sufficiently continuously differentiable with respect to x. To keep the probability invariant under the change of variables performed in eq. (2.56), the PDF $P_\varepsilon(x;t)$ of the ε-scaled population levels x is given by $P_\varepsilon(x;t) = \varepsilon^{-N} \cdot P(X;t)$. Hence, under the above system-size scaling, the CME can be written as

$$\varepsilon \frac{\partial}{\partial t} P_\varepsilon(x;t) = \sum_{\mu=1}^{M} \alpha_\mu(x - \varepsilon\xi_\mu) P_\varepsilon(x - \varepsilon\xi_\mu; t) - \alpha_\mu(x) P_\varepsilon(x;t). \tag{2.57}$$

WKB-Approximation of the PDF

In the following, we seek an approximate solution of the ε-scaled CME (2.57) in the form of an asymptotic expansion as $\varepsilon \to 0$. We assume that the PDF $P_\varepsilon(x;t)$ can be represented in a WKB-like series expansion with respect to the spatial coordinate x, i.e.,

$$P_\varepsilon(x;t) = C_\delta \exp\left\{\frac{1}{\delta} s_0(x;t)\right\} \sum_{n=0}^{\infty} \delta^n U_n(x;t) \quad \text{as} \quad \delta \to 0, \tag{2.58}$$

where the factor C_δ is related to the normalization of $P_\varepsilon(x;t)$, cf. [65, 132, 133]. The unknown functions s_0 and U_n, $n = 1, 2, \ldots$, in eq. (2.58) are assumed to be sufficiently continuously differentiable with respect to the arguments x and t.

In general, the WKB-technique is a powerful method for approximating the solution of a linear differential equation whose highest derivative is multiplied by a small parameter ε, well-known from the calculation of semi-classical approximations in quantum mechanics, see [65] for further information. In the context of the CME, the leading order WKB-approximation s_0, also called eikonal function, is known to describe the mode of the PDF in the basin of an attractor [132–134], which, as we will see shortly, naturally leads to the corresponding deterministic formulation of the chemical system [18].

Differentiating eq. (2.58) with respect to time t gives

$$\frac{\partial}{\partial t} P_\varepsilon(x;t) = C_\delta \exp\left\{\frac{1}{\delta} s_0(x;t)\right\} \sum_{n=0}^{\infty} \delta^{n-1} \left(U_n(x;t) \frac{\partial}{\partial t} s_0(x;t) + \delta \frac{\partial}{\partial t} U_n(x;t) \right),$$

which we will use to express the left hand side of the ε-scaled CME (2.57) under the above WKB-ansatz. In the right hand side of eq. (2.57), we first substitute the WKB-expansions of $P_\varepsilon(x;t)$ and $P_\varepsilon(x - \varepsilon\xi_\mu;t)$ as given by eq. (2.58). Then, we Taylor expand the propensities $\alpha_\mu(x - \varepsilon\xi_\mu)$ and the eikonal $s_0(x - \varepsilon\xi_\mu;t)$ and the functions $U_n(y - \varepsilon v_\mu \,|\, Z;t)$, $n = 1, 2, \ldots$, in the expression for $P_\varepsilon(x - \varepsilon\xi_\mu;t)$ around the state x, i.e.,

$$\alpha_\mu(x - \varepsilon\xi_\mu) = \sum_{|\gamma| \geq 0} \frac{(-\varepsilon\xi_\mu)^\gamma}{\gamma!} \partial_x^\gamma \alpha_\mu(x)$$

and

$$P_\varepsilon(x - \varepsilon\xi_\mu;t) = C_\delta \exp\left\{\frac{1}{\delta} \sum_{|\gamma| \geq 0} \frac{(-\varepsilon\xi_\mu)^\gamma}{\gamma!} \partial_x^\gamma s_0(x;t)\right\} \sum_{n=0}^{\infty} \delta^n \left(\sum_{|\gamma| \geq 0} \frac{(-\varepsilon\xi_\mu)^\gamma}{\gamma!} \partial_x^\gamma U_n(x;t) \right),$$

with multi-index $\gamma = (\gamma_1, \ldots, \gamma_N) \in \mathbb{N}_0^N$ and notations $|\gamma| = \gamma_1 + \ldots + \gamma_N$, $\gamma! = \gamma_1! \cdot \ldots \cdot \gamma_N!$, $x^\gamma = (x_1)^{\gamma_1} \cdot \ldots \cdot (x_N)^{\gamma_N}$ and $\partial_x^\gamma = \partial_{x_1}^{\gamma_1} \ldots \partial_{x_N}^{\gamma_N}$, where $\partial_{x_i}^{\gamma_i} := \partial^{\gamma_i} / \partial {x_i}^{\gamma_i}$ for $i = 1, \ldots, N$.

Hence, to this step the CME (2.57) can be written as

$$\begin{aligned}
&\frac{\varepsilon}{\delta} C_\delta \exp\left\{\frac{1}{\delta} s_0(x;t)\right\} \sum_{n=0}^{\infty} \delta^n \left(U_n(x;t) \frac{\partial}{\partial t} s_0(x;t) + \delta \frac{\partial}{\partial t} U_n(x;t) \right) \\
&\quad = C_\delta \sum_{\mu=1}^{M} \Bigg[\left(\sum_{|\gamma|\geq 0} \frac{(-\varepsilon\xi_\mu)^\gamma}{\gamma!} \partial_x^\gamma \alpha_\mu(x) \right) \exp\left\{ \frac{1}{\delta} \sum_{|\gamma|\geq 0} \frac{(-\varepsilon\xi_\mu)^\gamma}{\gamma!} \partial_x^\gamma s_0(x;t) \right\} \\
&\qquad\qquad \times \sum_{n=0}^{\infty} \delta^n \left(\sum_{|\gamma|\geq 0} \frac{(-\varepsilon\xi_\mu)^\gamma}{\gamma!} \partial_x^\gamma U_n(x;t) \right) - \alpha_\mu(x) \exp\left\{ \frac{1}{\delta} s_0(x;t) \right\} \sum_{n=0}^{\infty} \delta^n U_n(x;t) \Bigg].
\end{aligned} \tag{2.59}$$

Comparing the terms of leading order on both sides of eq. (2.59) gives

$$\begin{aligned}
&\frac{\varepsilon}{\delta} C_\delta \exp\left\{\frac{1}{\delta} s_0(x;t)\right\} U_0(x;t) \frac{\partial}{\partial t} s_0(x;t) \\
&\qquad = C_\delta \exp\left\{\frac{1}{\delta} s_0(x;t)\right\} U_0(x;t) \sum_{\mu=1}^{M} \alpha_\mu(x) \left[\exp\left\{ -\frac{\varepsilon}{\delta} \xi_\mu^{\mathrm{T}} \nabla s_0(x;t) \right\} - 1 \right],
\end{aligned} \tag{2.60}$$

where $\nabla s_0(x;t) := \left(\partial_{x_1} s_0(x;t), \ldots, \partial_{x_N} s_0(x;t)\right)^{\mathrm{T}}$ denotes the gradient of s_0 with respect to x. By dominant balance, the order on both sides of eq. (2.60) has to be the same. Thus, δ has to be proportional to ε, i.e., $\delta \sim \varepsilon$ as $\varepsilon \to 0$,[22] and for simplicity we choose $\delta = \varepsilon$. Then, eq. (2.60) is satisfied if we set

$$\frac{\partial}{\partial t} s_0(x;t) + \sum_{\mu=1}^{M} \alpha_\mu(x) \left[1 - \exp\left\{ -\xi_\mu^{\mathrm{T}} \nabla s_0(x;t) \right\} \right] = 0. \tag{2.61}$$

This is a linear PDE of first order for the eikonal function s_0, which can be solved by the methods of characteristics, see [135] and below.

Substituting eq. (2.61) back into the full expression of the CME (2.59), we find the evolution equations of the functions U_n, $n = 1, 2, \ldots$, to be determined by

$$\begin{aligned}
&\sum_{n=0}^{\infty} \varepsilon^{n+1} \frac{\partial}{\partial t} U_n(x;t) \\
&\quad = \sum_{\mu=1}^{M} \exp\left\{ -\xi_\mu^{\mathrm{T}} \nabla s_0(x;t) \right\} \Bigg[\left(\sum_{|\gamma|\geq 0} \frac{(-\varepsilon\xi_\mu)^\gamma}{\gamma!} \partial_x^\gamma \alpha_\mu(x) \right) \exp\left\{ \frac{1}{\varepsilon} \sum_{|\gamma|\geq 2} \frac{(-\varepsilon\xi_\mu)^\gamma}{\gamma!} \partial_x^\gamma s_0(x;t) \right\} \\
&\qquad\qquad \times \sum_{n=0}^{\infty} \varepsilon^n \left(\sum_{|\gamma|\geq 0} \frac{(-\varepsilon\xi_\mu)^\gamma}{\gamma!} \partial_x^\gamma U_n(x;t) \right) - \alpha_\mu(x) \sum_{n=0}^{\infty} \varepsilon^n U_n(x;t) \Bigg].
\end{aligned}$$

[22] We use the asymptotic notation $f(x) \sim g(x)$ as $x \to x_0$ if $\lim_{x\to x_0} f(x)/g(x) = 1$, cf. [65].

Comparing the terms of order $\mathcal{O}(\varepsilon)$ on both sides in the above equation, for instance, gives a linear first order PDE for U_0, which is a sort of transport equation (cf. [133]):

$$\frac{\partial}{\partial t}U_0(x;t) = \sum_{\mu=1}^{M} \exp\Big\{-\xi_\mu^{\mathrm{T}}\nabla s_0(x;t)\Big\}\Big[\alpha_\mu(x)\frac{1}{2}\xi_\mu^{\mathrm{T}}\nabla^2 s_0(x;t)\xi_\mu U_0(x;t) - \xi_\mu^{\mathrm{T}}\nabla\alpha_\mu(x)U_0(x;t) - \alpha_\mu(x)\xi_\mu^{\mathrm{T}}\nabla U_0(x;t)\Big],$$

where $\nabla^2 s_0(x;t)$ denotes the Hessian matrix of s_0 with respect to x.

In the following, we show that the leading order WKB-approximation of the mode of $P_\varepsilon(x;t)$, i.e., the absolute maximum point of s_0, follows the deterministic path of classical reaction kinetics.

Leading Order Approximation of the PDF

A convenient approach for the analysis of the eikonal function s_0 is to consider the PDE (2.61) as the Hamilton–Jacobi equation for the action of an auxiliary system with coordinates $\boldsymbol{x}(t)$ and momenta $\boldsymbol{p}(t) := \nabla s_0(x;t)$, cf. [132, 136]. The Hamiltonian $\mathcal{H}$ of this system is defined as

$$\mathcal{H}(\boldsymbol{x},\boldsymbol{p}) := -\frac{\partial}{\partial t}s_0(x;t) \overset{\text{eq. (2.61)}}{=} \sum_{\mu=1}^{M}\alpha_\mu(\boldsymbol{x})\Big[1-\exp\big\{-\xi_\mu^{\mathrm{T}}\boldsymbol{p}\big\}\Big],$$

and Hamilton's equations of motion of the system are thus given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}(t) = \frac{\partial}{\partial \boldsymbol{p}}\mathcal{H}(\boldsymbol{x},\boldsymbol{p}) = \sum_{\mu=1}^{M}\xi_\mu\alpha_\mu\big(\boldsymbol{x}(t)\big)\exp\big\{-\xi_\mu^{\mathrm{T}}\boldsymbol{p}(t)\big\}, \tag{2.62a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{p}(t) = -\frac{\partial}{\partial \boldsymbol{x}}\mathcal{H}(\boldsymbol{x},\boldsymbol{p}) = \sum_{\mu=1}^{M}\nabla\alpha_\mu\big(\boldsymbol{x}(t)\big)\Big[\exp\big\{-\xi_\mu^{\mathrm{T}}\boldsymbol{p}(t)\big\}-1\Big]. \tag{2.62b}$$

These equations define the characteristics of the PDE (2.61) (where $\mathcal{H} \equiv \text{const}$) and their solution determines $s_0(x;t)$, cf. [135].

At this point, it is essential for us to point out how maxima of s_0 are propagated by eqs. (2.62) in time. Observe that if the initial function $s_0(x;t=0)$ has a *unique* absolute maximum at $x = \overline{\boldsymbol{x}}_0$, say, where

$$\overline{\boldsymbol{p}}_0 \equiv \nabla s_0(x=\overline{\boldsymbol{x}}_0;t=0) = 0,$$

then $\mathcal{H}(\overline{\boldsymbol{x}}_0,\overline{\boldsymbol{p}}_0) \equiv 0$ and the equations of motion that propagate the maximum point $\overline{\boldsymbol{x}}(t)$ and the corresponding $\overline{\boldsymbol{p}}(t)$ in time are given by eqs. (2.62a) and (2.62b) as

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{\boldsymbol{x}}(t) = \sum_{\mu=1}^{M}\xi_\mu\alpha_\mu\big(\overline{\boldsymbol{x}}(t)\big) \quad \text{with} \quad \overline{\boldsymbol{x}}(0) = \overline{\boldsymbol{x}}_0, \tag{2.63a}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{\boldsymbol{p}}(t) = 0 \quad \text{with} \quad \overline{\boldsymbol{p}}(0) = \overline{\boldsymbol{p}}_0 = 0, \tag{2.63b}$$

respectively. This shows that $\overline{\boldsymbol{x}}(t)$ *remains* a unique maximum point of s_0 for all times $t \geq 0$, because of the uniqueness of paths under given initial conditions.[23] The solution of eq. (2.63a) hence gives the most probable values $\overline{\boldsymbol{x}}(t)$ of the species levels as $\varepsilon \to 0$. Moreover, we note that eq. (2.63a) is identical to the ODE-model of deterministic reaction kinetics. Since further $\alpha_\mu(x) = \nu_\mu(x) + \mathcal{O}(\varepsilon)$ as $\varepsilon \to 0$, see Subsection 2.4.1, the above result shows that if the system is close to the thermodynamic limit, then the classical deterministic solution more closely describes the mode trajectory of the corresponding CME probability distribution rather than its average [18].

Laplace's Integral Approximation of the PDF

We linked the ODE-model of deterministic reaction kinetics to the evolution equation (2.63a) for the most probable value of the species levels x (as given by the maximum of s_0) and showed that this maximum remains unique at $\overline{\boldsymbol{x}}(t)$ for all times $t \geq 0$. In the following, we demonstrate that this result has an important implication for computing expectations of the underlying process $\boldsymbol{x}^\varepsilon(t)$. Because of the exponential form of the WKB-ansatz (2.58), we can use Laplace's method of integral approximation (see Appendix B) to compute integrals of the form

$$\mathbb{E}\big[f(\boldsymbol{x}^\varepsilon)\big] := \int_0^\infty f(x)P_\varepsilon(x;t)\,\mathrm{d}x = C_\varepsilon \int_0^\infty f(x)\exp\left\{\frac{1}{\varepsilon}s_0(x;t)\right\}\sum_{n=0}^{\infty}\varepsilon^n U_n(x;t)\,\mathrm{d}x,$$

as $\varepsilon \to 0$. If the eikonal function $s_0(x;t)$ has a unique absolute maximum at the point $x = \overline{\boldsymbol{x}}(t)$, then by Laplace's method it follows that

$$\int_0^\infty f(x)\exp\left\{\frac{1}{\varepsilon}s_0(x;t)\right\}\mathrm{d}x = \exp\left\{\frac{1}{\varepsilon}s_0\big(\overline{\boldsymbol{x}}(t);t\big)\right\}\sqrt{\frac{(2\pi\varepsilon)^N}{\big|\nabla^2 s_0\big(\overline{\boldsymbol{x}}(t);t\big)\big|}}\Big(f\big(\overline{\boldsymbol{x}}(t)\big) + \mathcal{O}(\varepsilon)\Big),$$

as $\varepsilon \to 0$. Here and in the following, we consider the generic case that $\overline{\boldsymbol{x}}$ is a non-degenerate maximum of s_0 and further use that the Hessian matrix $\nabla^2 s_0(x = \overline{\boldsymbol{x}}(\,\cdot\,);\,\cdot\,)$ is hence negative definite, such that the determinant $|\nabla^2 s_0(x = \overline{\boldsymbol{x}}(\,\cdot\,);\,\cdot\,)| \neq 0$.

[23] Assume there exists a solution $(\overline{\boldsymbol{y}}(t), \overline{\boldsymbol{p}}(t))$ of eqs. (2.62) with initial conditions $\overline{\boldsymbol{y}}_0 \neq \overline{\boldsymbol{x}}_0$ and $\overline{\boldsymbol{p}}_0 \neq 0$, but such that $\overline{\boldsymbol{p}}(t = T) = 0$ and hence $\overline{\boldsymbol{y}}(t = T)$ is an extremum of $s_0(x; t = T)$, for some $T > 0$. Then, one could reverse time in eqs. (2.63a) and (2.63b) and conclude that necessarily $\overline{\boldsymbol{p}}_0 = 0$, which yields a contradiction.

The above result has an immediate consequence: Since the total probability *has to* integrate to one for *any* value of ε, we can compute $s_0(x = \overline{\boldsymbol{x}}(\,\cdot\,);\,\cdot\,)$ and $U_0(x = \overline{\boldsymbol{x}}(\,\cdot\,);\,\cdot\,)$ directly, i.e.,

$$\begin{aligned} 1 &\equiv \int_0^\infty P_\varepsilon(x;t)\,\mathrm{d}x \\ &= C_\varepsilon \int_0^\infty \exp\left\{\frac{1}{\varepsilon}s_0(x;t)\right\}\big(U_0(x;t)+\mathcal{O}(\varepsilon)\big)\,\mathrm{d}x \\ &= \frac{\exp\left\{\frac{1}{\varepsilon}s_0\big(\overline{\boldsymbol{x}}(t);t\big)\right\}}{\sqrt{\left|\nabla^2 s_0\big(\overline{\boldsymbol{x}}(t);t\big)\right|}}\Big(U_0\big(\overline{\boldsymbol{x}}(t);t\big)+\mathcal{O}(\varepsilon)\Big), \end{aligned} \tag{2.64}$$

where we substituted $C_\varepsilon = (2\pi\varepsilon)^{-N/2}$ in the last equality, as the above equation holds for all ε and we have assumed s_0 and U_0 to be independent of ε. For the same reason, it further follows from eq. (2.64) that

$$s_0\big(\overline{\boldsymbol{x}}(t);t\big) = 0 \quad \text{and} \quad U_0\big(\overline{\boldsymbol{x}}(t);t\big) = \sqrt{\left|\nabla^2 s_0\big(\overline{\boldsymbol{x}}(t);t\big)\right|}, \tag{2.65}$$

for all $t \geq 0$. Moreover, we find that the $\mathcal{O}(\varepsilon)$ corrections in the last equality of eq. (2.64) must be identical to zero. In particular, this shows that the expectation of any function f of $\boldsymbol{x}^\varepsilon$ can be approximated as

$$\begin{aligned} \mathbb{E}\big[f(\boldsymbol{x}^\varepsilon)\big] &= C_\varepsilon \int_0^\infty f(x)\exp\left\{\frac{1}{\varepsilon}s_0(x;t)\right\}\big(U_0(x;t)+\mathcal{O}(\varepsilon)\big)\,\mathrm{d}x \\ &= \frac{\exp\left\{\frac{1}{\varepsilon}s_0\big(\overline{\boldsymbol{x}}(t);t\big)\right\}}{\sqrt{\left|\nabla^2 s_0\big(\overline{\boldsymbol{x}}(t);t\big)\right|}}\Big(f\big(\overline{\boldsymbol{x}}(t)\big)U_0\big(\overline{\boldsymbol{x}}(t);t\big)+\mathcal{O}(\varepsilon)\Big) \\ &= f\big(\overline{\boldsymbol{x}}(t)\big)+\mathcal{O}(\varepsilon), \end{aligned} \tag{2.66}$$

as $\varepsilon \to 0$, where we applied Laplace's method and the results for s_0 and U_0 given in eqs. (2.65). It should be noticed that, according to the conservation relation (2.64), the last equalities in eq. (2.66) hold without the $\mathcal{O}(\varepsilon)$ corrections if f is a constant function.

The result in eq. (2.66) specifically shows that $\mathbb{E}[\boldsymbol{x}^\varepsilon] = \overline{\boldsymbol{x}}(t) + \mathcal{O}(\varepsilon)$ and $\mathrm{Var}(\boldsymbol{x}^\varepsilon) = \mathcal{O}(\varepsilon)$ as $\varepsilon \to 0$. In the large-size system regime, i.e., $\varepsilon = \Omega^{-1} \ll 1$, the PDF $P_\varepsilon(x;t)$ of $\boldsymbol{x}^\varepsilon(t)$ will thus be tightly concentrated about its distribution mode, such that fluctuations can be neglected. Hence, the stochastic process $\boldsymbol{x}^\varepsilon(t)$ is well-approximated by its distribution mode that, in agreement with Kurtz's Theorem, follows the deterministic path of classical reaction kinetics.

We can further use the above results in view of our studies in Subsection 2.4.3 on the evolution of the expected value $\mathbb{E}[\boldsymbol{x}^\varepsilon]$. First notice that by assuming existence of sufficiently many

moments of $\boldsymbol{x}^\varepsilon$, an averaging of the ε-scaled CME (2.57) can be performed as demonstrated in Subsection 2.4.3 for the unscaled CME (2.48), i.e.,

$$\begin{aligned}
\frac{\partial}{\partial t}\mathbb{E}[\boldsymbol{x}^\varepsilon] &= \int_0^\infty x \frac{\partial}{\partial t} P_\varepsilon(x;t)\,\mathrm{d}x \\
&= \frac{1}{\varepsilon}\int_0^\infty x \sum_{\mu=1}^{M} \alpha_\mu(x-\varepsilon\xi_\mu)P_\varepsilon(x-\varepsilon\xi_\mu;t) - \alpha_\mu(x)P_\varepsilon(x;t)\,\mathrm{d}x \\
&= \frac{1}{\varepsilon}\sum_{\mu=1}^{M}\left[\int_0^\infty (x+\varepsilon\xi_\mu)\alpha_\mu(x)P_\varepsilon(x;t)\,\mathrm{d}x - \int_0^\infty x\alpha_\mu(x)P_\varepsilon(x;t)\,\mathrm{d}x\right] \\
&= \sum_{\mu=1}^{M}\left[\xi_\mu \int_0^\infty \alpha_\mu(x)P_\varepsilon(x;t)\,\mathrm{d}x\right]. \qquad (2.67)
\end{aligned}$$

By employing the WKB-ansatz (2.58) for $P_\varepsilon(x;t)$ in eq. (2.67) and applying Laplace's integral approximation in combination with the results for s_0 and U_0 given in eqs. (2.65), we find

$$\begin{aligned}
\frac{\partial}{\partial t}\mathbb{E}[\boldsymbol{x}^\varepsilon] &= \sum_{\mu=1}^{M}\left[\xi_\mu C_\varepsilon \int_0^\infty \alpha_\mu(x)\exp\left\{\frac{1}{\varepsilon}s_0(x;t)\right\}\big(U_0(x;t)+\mathcal{O}(\varepsilon)\big)\,\mathrm{d}x\right] \\
&= \sum_{\mu=1}^{M}\xi_\mu\alpha_\mu\big(\overline{\boldsymbol{x}}(t)\big) + \mathcal{O}(\varepsilon), \qquad (2.68)
\end{aligned}$$

as $\varepsilon \to 0$. Again, this shows that in the large-size system regime the expectation $\mathbb{E}[\boldsymbol{x}^\varepsilon]$ is well-approximated by the distribution mode $\overline{\boldsymbol{x}}$. It can be nicely seen that the error made by neglecting the effects of correlations and fluctuations in deterministic reaction kinetics (e.g., assuming $\mathbb{E}[\boldsymbol{x}_i^\varepsilon \boldsymbol{x}_j^\varepsilon] = \mathbb{E}[\boldsymbol{x}_i^\varepsilon]\mathbb{E}[\boldsymbol{x}_j^\varepsilon]$) is of order $\mathcal{O}(\varepsilon)$ as $\varepsilon \to 0$. As we have seen in Subsection 2.4.3, for systems including only reactions up to order one, the evolution of $\mathbb{E}[\boldsymbol{x}^\varepsilon]$ and the ODE-model of the deterministic process $\boldsymbol{x}(t)$ are identical, such that $\mathbb{E}[\boldsymbol{x}^\varepsilon] = \boldsymbol{x}(t)$. This implies that in such case also eqs. (2.66) and (2.68) hold without $\mathcal{O}(\varepsilon)$.

In Chapter 4, we use the above methods to derive a direct hybrid formulation of biochemical reaction networks, where the levels of species limited to small numbers are modeled discretely, while the levels of species present in large numbers are approximated by (partial) expectations. The derivation is based on a decomposition of the joint PDF into the marginal PDF of the 'discrete' species and the conditional PDF of the 'continuous' species. In contrast to above, we then apply the WKB-approximation and Laplace's method only on the conditional PDF of the continuous species.

3

Hybrid Stochastic–Deterministic Simulation of Biochemical Reaction Networks

The costs of computing an exact stochastic simulation generally scale with the number of reaction events that have to be simulated. This renders exact indirect methods, such as the SSA, numerically infeasible whenever the system includes many rapidly firing reactions and species with a large number of molecules. In hybrid simulation methods, fast reactions associated with large species levels are approximated as continuous processes, either in a stochastic or deterministic context, whereas all other reactions are still realized as discrete stochastic processes. By exploiting such partial continuous approximations, hybrid methods are capable of decreasing the computational cost substantially, while capturing relevant stochastic effects arising from small numbers or fluctuations in slow reaction events.

We outline and discuss hybrid simulation approaches in Section 3.1, providing a detailed derivation of a hybrid stochastic–deterministic variant of the SSA, mainly following [56, 61, 137]. In Section 3.2, we study an application of this indirect hybrid method to a mathematical model of the replication dynamics of the human immunodeficiency virus (HIV) [138]. Here, the main objective is the *in silico*-based design and validation of a drug treatment strategy for HIV-infected, treatment-naïve patients. Based on hybrid simulations, we are able to utilize and evaluate a novel mathematical concept that prevents the emergence of drug-resistance by applying a single, pro-active treatment switch that can lead to significant improvements compared to conventional treatment strategies [139].

3.1 Hybrid Simulation Approaches

We recall that in stochastic reaction kinetics the state of the system after elapse of some small increment in time $\mathrm{d}t > 0$ can be represented as

$$\boldsymbol{X}(t+\mathrm{d}t) = \boldsymbol{X}(t) + \sum_{\mu=1}^{M} \xi_\mu \boldsymbol{\mathcal{P}}_\mu\left(\int_t^{t+\mathrm{d}t} a_\mu\big(\boldsymbol{X}(s)\big)\,\mathrm{d}s\right).$$

For convenience, we adopt the notation of differential equations in the following and write the above equation as

$$\mathrm{d}\boldsymbol{X}(t) = \sum_{\mu=1}^{M} \xi_\mu \mathrm{d}\boldsymbol{\mathcal{P}}_\mu(t),$$

where $\mathrm{d}\boldsymbol{X}(t) := \boldsymbol{X}(t+\mathrm{d}t) - \boldsymbol{X}(t)$ and $\mathrm{d}\boldsymbol{\mathcal{P}}_\mu(t) := \boldsymbol{\mathcal{P}}_\mu\big(\int_t^{t+\mathrm{d}t} a_\mu(\boldsymbol{X}(s))\,\mathrm{d}s\big)$. Now, the idea of hybrid simulation methods is to partition the reactions in the above equation into a subset $\mathcal{M}^d \subseteq \{1,\ldots,M\}$ of discretely treated reactions and a subset $\mathcal{M}^c := \{1,\ldots,M\} \setminus \mathcal{M}^d$ of continuously approximated reactions. The discrete reactions are then simulated by a Monte Carlo (MC)-method, e.g., the direct method as in the SSA, while the continuous reactions are approximated either by employing a CLE approximation, e.g., [50, 53–56], or deterministically using the ODE-formulation of classical reaction kinetics, e.g., [50, 53, 57–62]. The key principle of such indirect hybrid methods is to efficiently monitor and simulate the occurrences of the discrete reaction events while simultaneously simulating the dynamics of the continuous reaction processes. A review of different hybrid simulation methods is also given in [92].

In the following, we successively discuss appropriate partition strategies, the resulting hybrid system representation, as well as the realization of the discrete reactions in a stochastic–deterministic simulation.

3.1.1 Partitioning of the Reactions

Suggested hybrid simulation approaches vary in their underlying partitioning strategies. In general, the partitioning of the reactions into discrete and continuous processes can either be static [50, 53, 57, 59] or dynamic [54–56, 58, 60–62], i.e., the subsets $\mathcal{M}^c$ and $\mathcal{M}^d$ are constant or change during simulation, respectively. The partitioning might be based on some heuristics [50, 53], e.g., obtained by a few exact stochastic simulations or some insight into the system under consideration [57, 59]. For example, it is often reasonable to approximate metabolic reactions as continuous processes, while gene regulatory processes usually necessitate a discrete formulation.

An automated repartitioning of the reactions during simulation can be realized according to some predefined criteria with respect to the levels of species involved in the reactions [58], the values of the reaction propensities [50, 53], or, most commonly, some combination of the two [54–57, 59–62]. Clearly, a dynamic partitioning strategy introduces some additional computational costs compared to a static partitioning. At the same time, however, it offers a more general and reliable way of employing hybrid simulations, e.g., when the state of the

system varies considerably over time. Therefore, we consider a dynamic partitioning of the reactions into time-dependent subsets $\mathcal{M}^c(t)$ and $\mathcal{M}^d(t)$, based on criteria defined below.

In the context of the CLE (see Subsection 2.2.6), we already encountered sufficient conditions for the continuous approximation of a reaction R_μ during a time interval $[t; t + \mathrm{d}t)$: (a) The expected number of R_μ firings in $[t; t + \mathrm{d}t)$ has to be much larger than 1, i.e.,

$$\mathbb{E}\big[\mathrm{d}\boldsymbol{\mathcal{P}}_\mu(t)\big] = \int_t^{t+\mathrm{d}t} a_\mu(\boldsymbol{X}(s))\,\mathrm{d}s \gg 1, \tag{3.1}$$

and (b) the value of its propensity a_μ has to change insignificantly during $[t; t + \mathrm{d}t)$, i.e.,

$$a_\mu\big(\boldsymbol{X}(s)\big) \approx a_\mu(X) \quad \forall s \in [t; t + \mathrm{d}t), \tag{3.2}$$

where $\boldsymbol{X}(t) = X$, such that eq. (3.1) can be approximated as $\mathbb{E}[\mathrm{d}\boldsymbol{\mathcal{P}}_\mu(t)] \approx a_\mu(\boldsymbol{X}(t))\mathrm{d}t \gg 1$. As further discussed in Subsection 2.2.4, the above Leap Condition (3.2) can always be satisfied if the chemical species are present in sufficiently large numbers. In particular, the relative change in the value of a_μ during $[t; t + \mathrm{d}t)$ will be negligible if the level $\boldsymbol{X}_i(t)$ of every R_μ reactant is much larger than the absolute change $|\mathrm{d}\boldsymbol{X}_i(t)|$ it undergoes by *all* reactions in $[t; t + \mathrm{d}t)$. By choosing $\mathrm{d}t$ such that no discrete reaction fires during $[t; t + \mathrm{d}t)$, it is sufficient in a hybrid setting to require that $\boldsymbol{X}_i(t)$ is much larger than the absolute change caused by the *continuous* reactions only.

Given $\boldsymbol{X}(t) = X$, it is thus typically assumed that a reaction R_μ can be modeled continuously if the following two criteria are satisfied [56, 61, 62]:

$$a_\mu\big(\boldsymbol{X}(t)\big) > C_a, \tag{3.3a}$$

$$\boldsymbol{X}_i(t) > C_X \cdot |\xi_{\mu i}| \quad (i = 1, \ldots, N), \tag{3.3b}$$

for some pre-defined values of the parameters $C_a \gg 1$ and $C_X \gg 1$. If one of the above criteria is not satisfied, then the reaction is modeled as a discrete stochastic process.

The parameter C_a in criterium (3.3a) defines how many reaction events have to occur on average per unit time in order for the process to appear continuously. In other words, a reaction R_μ is continuously approximated if its expected next reaction time, given by $1/a_\mu(\boldsymbol{X}(t))$, is shorter than $1/C_a \ll 1$. This is exactly the property that renders an exact stochastic simulation of R_μ slow. By criterium (3.3b) it is ensured that *every* species which gets affected by a continuous reaction R_μ and contributes to a_μ or the propensity of any other continuous reaction is present in large enough numbers, such that the relative change $|\xi_{\mu i}|/\boldsymbol{X}_i(t)$ caused by an R_μ firing is smaller than or equal to $1/C_X \ll 1$. Of course, the exact

relation of criterium (3.3b) to condition (3.2) more specifically depends on the actual system under consideration, and some examples exists where criterium (3.3b) can be relaxed while still obtaining good approximations [61]. However, more general studies on its practical implementation are yet not available, and the handling of fast reactions involving small number of molecules is sill an open question in hybrid approaches [92].

It should be noticed that as the values of both parameters C_a and C_X increase towards infinity, we approach the conditions of the thermodynamic limit where a deterministic approximation becomes accurate. Usually, values in the range of ten to hundred are used for C_a and C_X, see [56, 61]; the resulting approximation error, however, will depend on the specific problem. For the hybrid stochastic–deterministic simulations of the HIV-dynamics model presented in Section 3.2, we set both, C_a and C_X, equal to 20. The reaction partitioning was then dynamically updated as follows: Initially and after each discrete reaction event, *every* reaction was (re-)classified as a continuous or discrete process by (re-)evaluating the partitioning criteria (3.3). Between discrete reaction events, however, we only monitored criteria (3.3) with respect to the *continuous* reactions, i.e., R_μ, with $\mu \in \mathcal{M}^c(t)$, was reclassified as a discrete reaction at time t if $a_\mu(\boldsymbol{X}(t))$ reached C_a or any $\boldsymbol{X}_i(t)$ reached $C_X \cdot |\xi_{\mu i}|$ during numerical integration of the corresponding ODE-system, see next subsection. This updating strategy was chosen in order to avoid potential instabilities in the reaction partitioning, due to some propensities or species level hovering around the critical values C_a or C_X, respectively. Furthermore, it reduced the computational costs that are introduced by a dynamic partitioning.

3.1.2 Hybrid System Representation

Given $\boldsymbol{X}(t) = X$, we define a partitioning of the reactions according to criteria (3.3) into the subsets $\mathcal{M}^d(t)$ and $\mathcal{M}^c(t)$ of discrete and continuous reactions, respectively, such that

$$\mathrm{d}\boldsymbol{X}(t) = \sum_{\mu \in \mathcal{M}^d(t)} \xi_\mu \mathrm{d}\boldsymbol{\mathcal{P}}_\mu(t) \;+ \sum_{\mu \in \mathcal{M}^c(t)} \xi_\mu \mathrm{d}\boldsymbol{\mathcal{P}}_\mu(t), \tag{3.4}$$

We recall that in the CLE approximation (see Subsection 2.2.6), the discrete-valued Poisson random variable $\mathrm{d}\boldsymbol{\mathcal{P}}_\mu$ is approximated by a corresponding real-valued normal random variable $\boldsymbol{\mathcal{N}}_\mu$, i.e., for $\boldsymbol{X}(t) = X$

$$\mathrm{d}\boldsymbol{\mathcal{P}}_\mu(t) \approx \boldsymbol{\mathcal{P}}_\mu\big(a_\mu(X)\mathrm{d}t\big) \approx \boldsymbol{\mathcal{N}}_\mu\big(a_\mu(X)\mathrm{d}t,\, a_\mu(X)\mathrm{d}t\big),$$

where $\boldsymbol{\mathcal{N}}_\mu(m, \sigma^2)$ denotes a normal random variable with mean m and variance σ^2. The first step in this approximation is justified by condition (3.2) and the second by condition (3.1). Thus, applying the CLE approximation to all continuous reactions in eq. (3.4) results in

$$\mathrm{d}\boldsymbol{X}(t) = \sum_{\mu\in\mathcal{M}^d(t)} \xi_\mu \mathrm{d}\boldsymbol{\mathcal{P}}_\mu(t) \;+ \sum_{\mu\in\mathcal{M}^c(t)} \xi_\mu \left[a_\mu\big(\boldsymbol{X}(t)\big)\mathrm{d}t + \sqrt{a_\mu\big(\boldsymbol{X}(t)\big)}\boldsymbol{\mathcal{N}}_\mu(0,1)\sqrt{\mathrm{d}t}\right], \tag{3.5}$$

which is the hybrid formula underlying discrete stochastic–continuous stochastic simulation methods, e.g., [50, 53–56].

It should be noticed that the relative fluctuations in the increment $\mathrm{d}\boldsymbol{\mathcal{P}}_\mu(t)$ of a continuous reaction in eq. (3.5) scale like

$$\frac{\sqrt{\mathrm{Var}\big(\mathrm{d}\boldsymbol{\mathcal{P}}_\mu(t)\big)}}{\mathbb{E}\big[\mathrm{d}\boldsymbol{\mathcal{P}}_\mu(t)\big]} = \frac{1}{\sqrt{\mathbb{E}\big[\mathrm{d}\boldsymbol{\mathcal{P}}_\mu(t)\big]}} \approx \frac{1}{\sqrt{a_\mu\big(\boldsymbol{X}(t)\big)\mathrm{d}t}},$$

which will be much smaller than 1 for any $\mu \in \mathcal{M}^c$ and small time increment $\mathrm{d}t$ as guaranteed by criterium (3.3a). Hence, we assume that it is reasonable to neglect fluctuations related to the continuous reactions in eq. (3.5), i.e., the random components $\xi_\mu\sqrt{a_\mu(\boldsymbol{X}(t))}\boldsymbol{\mathcal{N}}_\mu(0,1)\sqrt{\mathrm{d}t}$, and that the contribution of all continuous reactions is well-represented by the deterministic components $\xi_\mu a_\mu(\boldsymbol{X}(t))\mathrm{d}t$, such that eq. (3.5) further simplifies to

$$\mathrm{d}\boldsymbol{X}(t) = \sum_{\mu\in\mathcal{M}^d(t)} \xi_\mu \mathrm{d}\boldsymbol{\mathcal{P}}_\mu(t) \;+ \sum_{\mu\in\mathcal{M}^c(t)} \xi_\mu a_\mu\big(\boldsymbol{X}(s)\big)\mathrm{d}t. \tag{3.6}$$

This is the general hybrid system representation of stochastic–deterministic simulation approaches, e.g., [50, 53, 57, 59–62].

In line with our studies in Section 2.4, the partial deterministic approximation exploited in eq. (3.6) is justified by large values of the propensities and species levels being associated with the continuous reactions, which can both be guaranteed by the partitioning criteria (3.3). Although eq. (3.6) is less accurate than the CLE approximation (3.5), it is acceptable if one is primarily interested in the fluctuations of species present in small numbers [53, 62], as it is typically the case. A major benefit of the deterministic approximation is that as long as no discrete reaction fires, the system dynamics is simply given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{X}(t) = \sum_{\mu\in\mathcal{M}^c(t)} \xi_\mu a_\mu\big(\boldsymbol{X}(s)\big), \tag{3.7}$$

which can usually be efficiently solved with high accuracy using one of the well-documented numerical integration methods for ODEs, cf. [119–122].

3.1.3 Simulation of the Discrete Reactions

We proceed by deriving a stochastic simulation procedure for the discrete reactions in eq. (3.6), analogously to the derivation of the SSA, see Subsection 2.2.3. In the hybrid setting, however, we now have to account for changes in the propensity value of a discrete reaction, resulting from changes in the number of molecules by the continuous reactions. Given $\boldsymbol{X}(t) = X$, the probability $p^d(\tau, j \,|\, X; t)\delta t$ that the next *discrete* reaction fires during $[t + \tau; t + \tau + \delta t)$ and that this will be the discrete reaction R_j is hence given by

$$p^d(\tau, j \,|\, X; t)\delta t = p_0^d(\tau \,|\, X; t)\Big(a_j\big(\boldsymbol{X}(t + \tau)\big)\delta t + o(\delta t)\Big), \tag{3.8}$$

with $j \in \mathcal{M}^d(t + \tau)$ and $p_0^d(\tau \,|\, X; t)$ denoting the probability that no *discrete* reaction fires in $[t; t + \tau)$, given $\boldsymbol{X}(t) = X$.

To determine $p_0^d(\tau \,|\, X; t)$ in eq. (3.8), we consider the change in this probability during an infinitesimal time increment $\delta\tau$. As the increments of Poisson processes are independent, we have

$$p_0^d(\tau + \delta\tau \,|\, X; t) = p_0^d(\tau \,|\, X; t)\Bigg(1 - \sum_{\mu \in \mathcal{M}^d(t+\tau)} a_\mu\big(\boldsymbol{X}(t + \tau)\big)\delta\tau + o(\delta\tau)\Bigg), \tag{3.9}$$

where the second factor gives the probability that no discrete reaction fires in $[t + \tau; t + \tau + \delta\tau)$. Subtracting $p_0^d(\tau \,|\, X; t)$ from both sides of eq. (3.9), dividing by $\delta\tau$ and taking the limit $\delta\tau \to 0$ results in

$$\frac{\mathrm{d}p_0^d(\tau \,|\, X; t)}{\mathrm{d}\tau} = -a_0^d\big(\boldsymbol{X}(t + \tau), t + \tau\big)p_0^d(\tau \,|\, X; t),$$

where $a_0^d(\boldsymbol{X}(t), t) := \sum_{\mu \in \mathcal{M}^d(t)} a_\mu(\boldsymbol{X}(t))$ is the collective propensity of all discrete reactions at time t. Since $p_0^d(0 \,|\, X; t) = 1$, the solution of the above ODE is given by

$$p_0^d(\tau \,|\, X; t) = \exp\Bigg\{-\int_t^{t+\tau} a_0^d\big(\boldsymbol{X}(s), s\big)\,\mathrm{d}s\Bigg\}.$$

Substituting this result into eq. (3.8), dividing both sides by δt and taking the limit $\delta t \to 0$ finally results in

$$p^d(\tau, j \,|\, X; t) = \exp\Bigg\{-\int_t^{t+\tau} a_0^d\big(\boldsymbol{X}(s), s\big)\,\mathrm{d}s\Bigg\}a_j\big(\boldsymbol{X}(t + \tau)\big), \tag{3.10}$$

with $j \in \mathcal{M}^d(t + \tau)$.

We continue by exploiting the relation $p^d(\tau, j \,|\, X; t) = p^d(j \,|\, \tau, X; t)p^d(\tau \,|\, X; t)$ in order to derive the marginal probability $p^d(\tau \,|\, X; t)$ that the next *discrete* reaction fires at time

$t + \tau$ and the conditional probability $p^d(j \mid \tau, X; t)$ that this will be the discrete reaction R_j. Summation of eq. (3.10) over the subset of discrete reactions at time $t + \tau$ gives

$$p^d(\tau \mid X; t) = \sum_{j \in \mathcal{M}^d(t+\tau)} p^d(\tau, j \mid X; t) = \exp\left\{ - \int_t^{t+\tau} a_0^d\big(\boldsymbol{X}(s), s\big)\, \mathrm{d}s \right\} a_0^d\big(\boldsymbol{X}(t+\tau), t+\tau\big), \quad (3.11)$$

which shows that the time until the next discrete reaction event is an exponential random variable with time-dependent parameter $a_0^d(\boldsymbol{X}(t), t)$. From eqs. (3.10) and (3.11) it further follows that

$$p^d(j \mid \tau, X; t) = \frac{p^d(\tau, j \mid X; t)}{p^d(\tau \mid X; t)} = \frac{a_j\big(\boldsymbol{X}(t+\tau)\big)}{a_0^d\big(\boldsymbol{X}(t+\tau), t+\tau\big)}, \quad (3.12)$$

with $j \in \mathcal{M}^d(t+\tau)$ and $p^d(\tau \mid X; t) > 0$.[1] Thus, the index of the next discrete reaction that fires is an integer random variable with point probabilities $a_j(\boldsymbol{X}(t+\tau)) / a_0^d(\boldsymbol{X}(t+\tau), t+\tau)$, for every $j \in \mathcal{M}^d(t+\tau)$.

We employ the direct method to generate random samples τ and j according to the distributions implied by eqs. (3.11) and (3.12), respectively. Given τ and $\boldsymbol{X}(t+\tau)$ as the solution of the ODE-system (3.7), we can generate a random sample $j \in \mathcal{M}^d(t+\tau)$ as in the SSA but restricted to the subset of discrete reactions: Draw a random number r_2 from the standard uniform distribution $U(0,1)$ and take

$$j = \text{the smallest integer such that} \sum_{\mu=1}^{j} \mathbb{1}_{\mathcal{M}^d(t+\tau)}(\mu)\, a_\mu\big(\boldsymbol{X}(t+\tau)\big) \geq r_2 a_0^d\big(\boldsymbol{X}(t+\tau), t+\tau\big), \quad (3.13)$$

where $\mathbb{1}_{\mathcal{M}^d(t)}$ denotes the indicator function of the subset $\mathcal{M}^d(t)$ of discrete reactions at time t

$$\mathbb{1}_{\mathcal{M}^d(t)}(\mu) = \begin{cases} 1 & \text{if } \mu \in \mathcal{M}^d(t), \\ 0 & \text{otherwise.} \end{cases}$$

The cumulative distribution function (CDF) of the time until the next discrete reaction event is given by eq. (3.11) as

$$F(\tau \mid X; t) = \int_0^\tau p^d(s \mid X; t)\, \mathrm{d}s = 1 - \exp\left\{ - \int_t^{t+\tau} a_0^d\big(\boldsymbol{X}(s), s\big)\, \mathrm{d}s \right\}. \quad (3.14)$$

By the inversion generating method [89], we can generate a random sample from this distribution by drawing a random number r_1 from the standard uniform distribution $U(0,1)$ and

[1] If $p^d(\tau \mid X; t) = 0$, no discrete reaction fires during $[t; t+\tau]$, such that we do not have to evaluate $p^d(j \mid \tau, X; t)$, but only simulate the continuous reactions.

solving the equation $r_1 = F(\tau \,|\, X; t)$ for τ. Setting eq. (3.14) equal to $(1 - r_1)$, which like r_1 is a random number from $U(0, 1)$, yields

$$g_0(\tau \,|\, X; t) := \int_t^{t+\tau} a_0^d\big(\boldsymbol{X}(s), s\big)\, \mathrm{d}s + \ln(r_1) = 0, \tag{3.15}$$

which is the generating formula we have to solve in order to compute the next reaction time, cf. [56, 61]. To illustrate the similarity of the above equation to the corresponding generating formula (2.12a) in the SSA, we observe that if a_0^d is a constant function, i.e., $a_0^d(\boldsymbol{X}(s), s) \equiv a_0^d(X)$ for all $s \in [t; t + \tau]$, then by eq. (3.15)

$$g_0(\tau \,|\, X; t) = a_0^d(X)\tau + \ln(r_1) = 0 \quad \Rightarrow \quad \tau = \frac{1}{a_0^d(X)} \ln\!\left(\frac{1}{r_1}\right).$$

It should be noticed that the function $g_0(\tau \,|\, X; t)$ defined in eq. (3.15) is monotonically non-decreasing for $\tau \geq 0$, since propensities are non-negative functions by definition. Furthermore, since $r_1 \in (0; 1)$, we have $g_0(0 \,|\, X; t) = \ln(r_1) < 0$ and differentiation of eq. (3.15) with respect to τ shows that the value of $g_0(\tau \,|\, X; t)$ increases according to the ODE

$$\frac{\mathrm{d}}{\mathrm{d}\tau} g_0(\tau \,|\, X; t) = a_0^d\big(\boldsymbol{X}(t + \tau), t + \tau\big) \tag{3.16}$$

This suggests a relative simple procedure to determine when the next discrete reaction fires: We integrate eq.(3.16) together with the ODE-system (3.7) for the initial values $g_0(0 \,|\, X; t) = \ln(r_1)$ and $\boldsymbol{X}(t) = X$, respectively, until time $\tau = s$ such that $g_0(s \,|\, X; t) = 0$. Then, we use eq. (3.13) to decide which discrete reaction has to be simulated and update $\boldsymbol{X}(t + \tau)$ accordingly.

As mentioned before, the above algorithmic approach is equivalent to Gillespie's Direct Method implemented in the SSA, (see Subsection 2.2.3), but it additionally accounts for time-varying propensities of the discrete reactions. Gillespie's Direct Method has been utilized in the hybrid methods proposed in [50, 53, 54, 57, 58, 60–62]; time-varying propensities, however, have been explicitly incorporated only by Alfonsi et al. [61] and are otherwise approximated during simulation (e.g., by artificially introducing a 'probability of no reaction' that adjusts the discrete time steps and allows to gradually update the propensity values [53]). It is also possible to simulate the discrete reactions based on variants of Gillespie's Next Reaction Method [55, 56, 59, 61] or Gibson & Bruck's First Reaction Method [61]. However, this requires to solve a system of differential equations of the type of eq. (3.16), one for each discrete reaction, which generally increases the computationally complexity.

3.1.4 Algorithmic Implementation

In summary, the derived hybrid stochastic–deterministic simulation method comprises the following algorithmic workflow:[2]

(1) Initialize $t \leftarrow t_0$ and $\boldsymbol{X}(t_0) \leftarrow X_0$, set $\mathcal{M}^d(t_0) \leftarrow \{1, \ldots, M\}$ and $\mathcal{M}^c(t_0) \leftarrow \varnothing$, and choose partition thresholds C_a and C_X and a final simulation time t_{end}.

(2) Draw a random number r_1 from $U(0,1)$ and set $g_0(t \,|\, t) = \ln(r_1)$.

(3) For every reaction R_μ ($\mu = 1, \ldots, M$):

  (i) If $a_\mu(\boldsymbol{X}(t)) > C_a$ and $\boldsymbol{X}_i(t) > C_X \cdot |\xi_{\mu i}|$ for all $i = 1, \ldots, N$, set $\mathcal{M}^c(t) \leftarrow \mathcal{M}^c(t) \cup \mu$ and $\mathcal{M}^d(t) \leftarrow \mathcal{M}^d(t) \setminus \mu$.

  (ii) Otherwise, set $\mathcal{M}^c(t) \leftarrow \mathcal{M}^c(t) \setminus \mu$ and $\mathcal{M}^d(t) \leftarrow \mathcal{M}^d(t) \cup \mu$.

(4) Solve the ODE-system for the subset $\mathcal{M}^c(t)$ starting at time $\tau = t$

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{X}(\tau) = \sum_{\mu \in \mathcal{M}^c(\tau)} \xi_\mu a_\mu\big(\boldsymbol{X}(\tau)\big) \quad \text{together with} \quad \frac{\mathrm{d}}{\mathrm{d}\tau} g_0(\tau \,|\, t) = \sum_{\mu \in \mathcal{M}^d(\tau)} a_\mu\big(\boldsymbol{X}(\tau)\big), \tag{3.17}$$

until the first time $\tau = s$, where:

  (a) $g_0(s \,|\, t) = 0$ and a discrete reaction R_j with $j \in \mathcal{M}^d(s)$ has to be performed:

    (i) Draw a random number r_2 from $U(0,1)$.

    (ii) Take j as the smallest integer such that

$$\sum_{\mu=1}^{j} \mathbb{1}_{\mathcal{M}^d(s)}(\mu) a_\mu\big(\boldsymbol{X}(s)\big) \geq r_2 \sum_{\mu=1}^{M} \mathbb{1}_{\mathcal{M}^d(s)}(\mu) a_\mu\big(\boldsymbol{X}(s)\big).$$

    (iii) Set $t \leftarrow s$ and $\boldsymbol{X}(s) \leftarrow \boldsymbol{X}(s) + \xi_j$, and go to Step (2).

  (b) $a_\mu(\boldsymbol{X}(s)) = C_a$ or $\boldsymbol{X}_i(s) = C_X \cdot |\xi_{\mu i}|$ for some $\mu \in \mathcal{M}^c(s)$ and some $i = 1, \ldots, N$:

    (i) Set $\mathcal{M}^c(s) \leftarrow \mathcal{M}^c(s) \setminus \mu$ and $\mathcal{M}^d(s) \leftarrow \mathcal{M}^d(s) \cup \mu$.

    (ii) Set $t \leftarrow s$ and go to Step (4).

  (c) $s = t_{\text{end}}$, then stop simulation.

[2] It should be noticed that as long as either the subset $\mathcal{M}^c(t)$ or $\mathcal{M}^d(t)$ is empty, automatically a full discrete–stochastic (SSA) or a full continuous–deterministic simulation is performed, respectively.

In Step (4) of the above algorithmic scheme, events related to (a) the occurrence of a discrete reaction or (b) the repartitioning of a continuous reaction have to be detected during numerical integration of the ODEs (3.17). To this end, event functions $F_i(\tau)$ can be introduced, e.g., $F_0(\tau) := g_0(\tau \mid t)$ or $F_\mu(\tau) := C_a - a_\mu(\boldsymbol{X}(\tau))$, such that $F_i(s) = 0$ at those points in time $\tau = s$ where a corresponding event is located. Then, the problem of event detection during an successful integration step from time t_{n-1} to t_n becomes: (i) Detecting the possible occurrence of zero-crossings of any F_i between t_{n-1} and t_n, and (ii) finding (within a predefined numeric tolerance) the smallest root of $F_i(s)$ with $s \in (t_{n-1}; t_n]$.

Problem (i) is usually tackled by comparing the signs of $F_i(t_{n-1})$ and $F_i(t_n)$, i.e., if $F_i(t_{n-1}) \times F_i(t_n) \leq 0$, then an event occurred during $(t_{n-1}; t_n]$. This strategy works correct as long as the number of roots of F_i in $(t_{n-1}; t_n]$ is odd, which will always be the case for $F_0(\tau) = g_0(\tau \mid t)$, for instance, since $g_0(\tau \mid t)$ is monotonic in $(t_{n-1}; t_n]$. A multiple (even) number of zero-crossings may occur if the dynamics of F_i is faster than the dynamics of the system state [140]. The possibility of an even number of roots can be eliminated by explicitly including the dynamics of F_i in the ODE-system (3.17) or in the step-size selection of the numerical integration method, cf. [141, 142]. If a zero-crossing is detected, an algorithm such as the *regula falsi* method or Brent's method [143] can be used to find the smallest root of F_i in $(t_{n-1}; t_n]$, where a successive numerical solution of the ODEs (3.17) can be avoided by use of a numerical integration method that supplies a dense output interpolation, cf. [119–122]. Typically, common computational software programs, e.g., MATLAB® or *Mathematica*®, already include ODE-integration methods that comprise event detection and dense output.

3.2 An Application: HIV Quasi-Species Dynamics during Pro-Active Treatment Switching[†]

In the sequel, we study an application of the above stochastic–deterministic simulation algorithm to an HIV-dynamics model that we previously proposed in [138]. Using hybrid simulations, we are able to predict and evaluate (in terms of the mathematical model) a simple treatment strategy, where a single, pro-active switch from an induction to a maintenance drug regimen is applied. The proposed induction–maintenance (IM) strategy demonstrates

[†]The contents of this section are based on: M. von Kleist, S. Menz, H. Stocker, K. Arasteh, C. Schütte, and W. Huisinga. HIV Quasispecies Dynamics during Pro-Active Treatment Switching: Impact on Multi-Drug Resistance and Resistance Archiving in Latent Reservoirs. *PLoS ONE*, 6(3): e18204, 2011.

significant improvements in terms of resistance archiving and virological response compared to conventional treatment strategies. While continuous pro-active treatment alternation improved the clinical outcome in a randomized trial [144, 145], our results indicate that a similar improvement might be reached by a single pro-active treatment switch. The clinical validity of this finding, however, remains to be shown in a corresponding trial.

3.2.1 Background

Since 1996, HIV infection is treated with a combination therapy, known as highly active anti-retroviral therapy (HAART) [146, 147]. The initial clinical success of HAART had led many researchers to believe that eradication of HIV could be feasible. However, it was soon realized that inducible pro-virus persists in latently infected cells despite ongoing therapy, and that this latent reservoir prevents HIV eradication within the patients lifetime [148–153]. Latent infection is established when $CD4^+$ T-lymphoblasts with integrated pro-virus escape both immune effector mechanisms and cytopathic effects of the virus [148, 154], and revert to a resting memory state [155]. Besides preventing eradication, the latent reservoir also serves as a memory for any quasi-species replicating during infection [156, 157], including drug-resistant variants. The contents of this archive are strong predictors of future treatment failure [156, 158].

Despite the impressive improvement of antiviral therapy, many patients still experience virological failure caused by selection of drug-resistant virus populations. Current guidelines recommend changing treatment after confirmation of virological failure. However, in face of the rapid viral turnover this approach could be sub-optimal [159], as treatment change after appearance of virological failure allows for expansion and potential archiving of drug-resistant strains. An optimal treatment strategy should thus prevent viral relapse with drug-resistant strains and, more importantly, prevent these mutants from establishing latent infection.

IM approaches are used for the treatment of a growing number of infectious and neoplastic diseases [160–162]. Treatment begins with an intensified induction regimen (composed of a number of potent and potentially toxic drugs), which is subsequently replaced by a maintenance regimen (composed of a smaller number of less toxic drugs) [163]. However, patients treated with a large number of drugs are particularly vulnerable to drug interactions [164] and adverse side effects that complicate therapy and seriously undermine the success of clinical management [165]. Another approach to overcome the development of resistance is to

alternate anti-retroviral therapy [166]. This strategy has been shown to significantly delay virological failure [144, 145], yet it is flawed by its high psychological and physical burden [167].

We propose an approach that combines the advantages of conventional IM and treatment alternation strategies, but minimizes their inherent disadvantages. We suggest a single, pro-active treatment switch from an inducer to a maintenance drug combination. Initially, the inducer combination should rapidly lower the viral population size and eliminate resistant mutants. Before drug-resistant strains are likely to be archived, the inducer combination will be replaced by a maintenance regimen with a completely different resistance profile. Based on hybrid stochastic–deterministic simulations of a HIV-dynamics model that integrates the mode of action of all approved and some developmental drugs [138], we are able to study a novel mathematical concept, which prevents the emergence of drug-resistance in each individual realization (virtual patient) by switching between therapies. Utilizing this concept, we deduce a distribution of switching times from the hybrid simulations, used to determine a single fixed duration for the induction therapy that increases the probability of treatment success in the whole virtual patient population and minimizes the risk of resistance archiving in the latent reservoir. Finally, we evaluate the performance of this novel IM-strategy against conventional HAART by comparing the statistics obtained from corresponding hybrid

3.2.2 Results

Mathematical HIV-Dynamics Model of Viral Replication, Mutation and Drug Interference

We have previously introduced a novel model of HIV-dynamics [138] that allows the mechanistic integration of all novel and some developmental HIV drugs. We extended this model by the compartment of very long lived, latently infected T-cells T_L, which are believed to prevent eradication of HIV [168] and to lead to archiving of drug-resistance [156, 157]. The resulting model is depict in Figure 3.1. It comprises uninfected T-cells and macrophages, T_U and M_U, respectively, non-infectious virus V_{NI}, infectious virus $V_I(i)$ of mutant strain $i = 1, 2, \ldots$, and infected cells belonging to each mutant strain i: infected T-cells and macrophages *prior* to proviral genomic integration, $T_1(i)$ and $M_1(i)$, and infected T-cells and macrophages *after* proviral genomic integration $T_2(i)$, $T_L(i)$ and $M_2(i)$. Latently infected cells $T_L(i)$ do not express viral genes but can become activated, transforming $T_L(i)$ into virus producing cells $T_2(i)$. A detailed description of the model and its parameterization is given in Appendix A.

The overall dynamics in the model comprises several viral strains, each with species whose copy numbers can vary over several orders of magnitude (from 0 to $\approx 10^{11}$). For this reason, we employed a hybrid approach in order to (i) correctly account for stochastic fluctuations in slow reactions and species present in small numbers, and (ii) reduce the computational costs

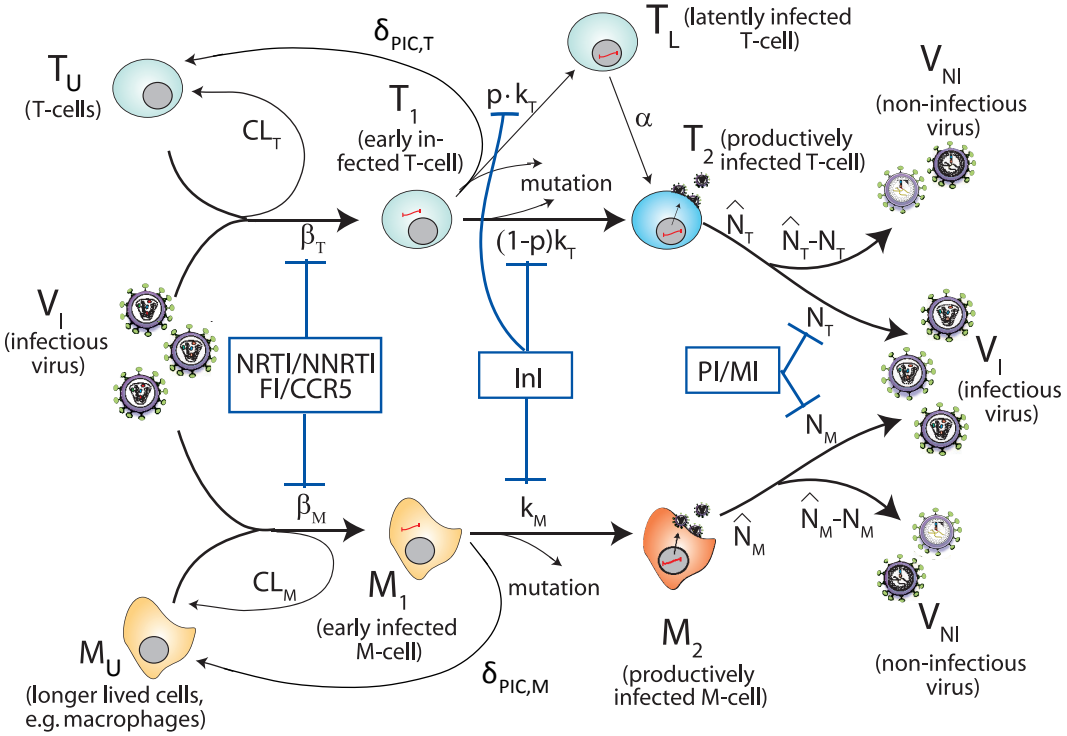

Figure 3.1: *HIV-dynamics model of viral replication, mutation and drug interference. Target cells T_U and M_U can become infected by infective virus V_I (with rate constants β_T and β_M), producing early infected cells T_1 and M_1. Infection can also be unsuccessful after viral fusion (CL_T and CL_M), eliminating the virus and rendering the target cell uninfected. Early infected cells T_1 and M_1 can also destroy essential viral proteins or DNA prior to integration ($\delta_{PIC,T}$ and $\delta_{PIC,M}$), returning the cells to an uninfected stage. The genomic viral DNA can become integrated (k_T and k_M) creating post-integration, infected cells T_2, T_L and M_2. The latently infected cell type T_L (created with probability p) does not express viral genes, but can become activated (with rate constant α), transforming T_L into a productively infected T-cell T_2. Virus producing cells T_2 and M_2 release new infectious and non-infectious virus V_I and V_{NI} (N_T, $\widehat{N}_T - N_T$ and N_M, $\widehat{N}_M - N_M$, respectively). Phenotypic mutation occurs at the stage of viral genomic integration. T_U and M_U are produced by the immune system with constant rates λ_T and λ_M, all cell types x get destroyed with respective rate constants δ_x, and free virus (infectious and non-infectious) gets cleared with rate constant CL (not shown in the illustration). The sites of drug interference are indicated by blue bars for the respective drug classes (NRTIs, NNRTIs, FIs, CCR5-inhibitors, INIs, PIs, and MIs).*

substantially by approximating the fast, large-number system dynamics deterministically. We used the hybrid method derived in Section 3.1, where a reaction R_μ was deterministically approximated during simulation if both, the value of its propensity and the levels of all species affected by R_μ, were above a threshold value of 20. Otherwise, the reaction was treated as a discrete stochastic process.

Realization and Implementation of the Model: We implemented the stochastic–deterministic simulation algorithm from Subsection 3.1.4 for the above HIV-dynamics model in C++, utilizing an ODE integrator that is based on numerical differentiation formulas [169] and uses strategies for event detection and error- and step size-control comparable to `ode15s` in MATLAB® [170].[3] Numerical integration and event detection was performed with a relative error tolerance of 10^{-6} and an absolute error tolerance of 10^{-9}. Simulations were run on 112 Intel® Xeon™ CPU E5440 cores with 2.83 GHz at the high performance computing cluster of the National University of Ireland, Maynooth (NUIM), utilizing the Intel® C++ and FORTRAN compilers and the Intel® math kernel library (MKL). To evaluate the proposed IM-strategy against conventional HAART (data shown in Figure 3.4), for instance, we performed 12,000 hybrid realizations. With realization start ($t_0 = 0$) the effects of drug treatment were simulated until $t = 730$ days was reached. The average runtime on a single core was ≈ 1.5 h per realization, resulting in an overall computation time of $\approx 18,000$ h.

Treatment Change Before Virological Failure

Currently, changes of anti-retroviral treatment regimes are largely triggered by virological failure or toxicity. In Figure 3.2A, we show the viral load during a hybrid simulation in the case of first-line treatment failure. The corresponding population dynamics of HIV is shown in Figure 3.2B. During first-line treatment failure, resistant mutants (green and cyan lines) are selected from the quasi-species population and quickly become the dominant population, leading to viral rebound (observe that fluctuations are captured for low levels in the hybrid simulation). While the total virus population is temporarily shrinking, mutants that confer resistance against a potential follow-up treatment (red line, dark grey shaded area) are depleted (possibly eradicated, by correctly accounting for discretely changing low levels in the hybrid setting). However, during viral rebound the total population re-expands and consequently erroneous reverse transcription generates novel mutants that can confer resistance against a

[3] A MATLAB®-implementation of our simulation code is provided in the Supporting Information to [139].

second-line therapy. Once the total population size has been restored, a second-line therapy, although composed of entirely different drugs, is as likely to fail as before initiation of the first-line therapy. Furthermore, drug-resistant viral strains are likely to become archived while they dominate the population (light grey shaded area).

In Figure 3.2C, we show the viral load during the proposed IM-therapy. The corresponding population dynamics of HIV is shown in Figure 3.2D. The inducer combination reduces the viral load (see Figure 3.2C). However, treatment is changed (vertical dashed lines) to the maintenance combination, before resistant strains (green and cyan lines) can become more

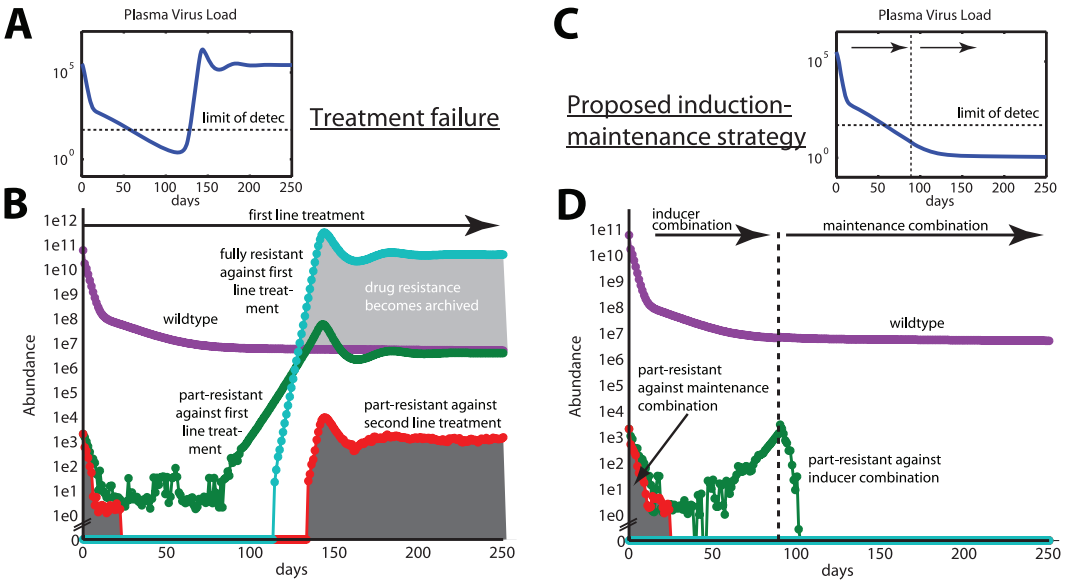

Figure 3.2: *Abundance of viral mutants during first-line treatment failure and proposed IM-strategy. A and B: Plasma virus load and abundance of distinct viral mutants during first-line treatment failure, respectively. C and D: Plasma virus load and abundance of distinct viral mutants during IM-strategy, respectively, switching from induction to maintenance therapy after* 80 *days (vertical dashed lines). Magenta lines show the abundance of wild type; green and cyan lines show the abundance of mutants that are part-resistant (against two out of three drugs) and full-resistant against the first-line regimen, respectively; red lines show the abundance of mutants that are part-resistant against the second-line treatment, where the area under the curve is dark grey shaded to stress the negative impact on a success of the second-line regimen. The light shaded area in panel B highlights that full-resistant mutants are more abundant than wild type virus, indicating when drug-resistance archiving in latently infected cells takes place. Simulations were performed for drug efficacy* $\eta(wt, d) = 70$ *% and fitness loss* $s = 20$ *% per drug-resistance mutation. It was further assumed that a single point mutation can confer absolute resistance to a single drug.*

abundant than the wild type (magenta line). Therefore, at the time of treatment change, total virus has been decreased and mutants that confer resistance to the maintenance therapy (red line, dark grey shaded area) are likely to be eradicated, which improves the probability to achieve durable virological suppression. With this strategy, the abundance of the wild type is larger than the abundance of drug-resistant mutants, which further lowers the probability that drug-resistance enters the latent reservoir (light grey shaded area is absent in Figure 3.2D).

Determination of Treatment Changing Time

To determine an optimal time point for switching from inducer to maintenance combinations, we first determined relevant sets of parameters for (i) the efficacy $\eta(wt, d)$ of drug d against the wild type '*wt*' and (ii) the fitness loss $s(d)$ that is associated with resistance development, since the corresponding *in vivo* parameters are known to vary substantially between different patients [171]. For simulation purposes, we assumed that one point mutation is sufficient to create high-level resistance (99 %) to a single drug. This is somewhat a worst-case assumption, but is justified for a number of drugs, see, e.g., [172, 173].

Then, we use an algorithm that automatically switches from inducer to maintenance combination, minimizing virological failure in each realization (virtual patient) by utilizing the 'reproductive capacity' $R_{\text{cap}}(d)$. We introduced the reproductive capacity in [138] and provide its derivation for the extended HIV-model 3.1 in Appendix A. The reproductive capacity $R_{\text{cap}}(d)$ can be interpreted as the expected total number of infectious offspring that the whole viral population produces under treatment d during one round of replication. It can be calculated from any simulation and enables to evaluate the state of infection from the

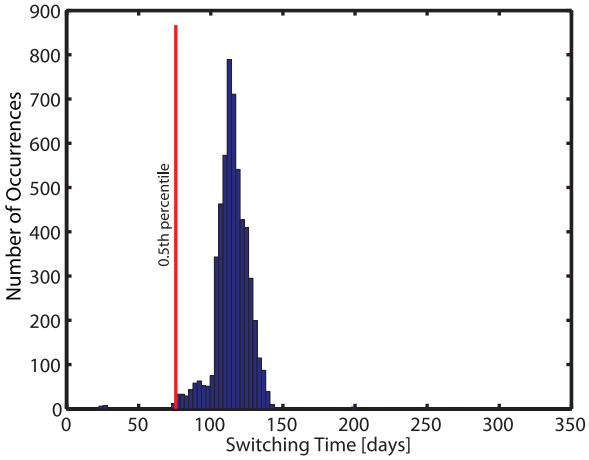

Figure 3.3: *Histogram of optimal, individual treatment switching times. Times for switching from induction to maintenance therapy were automatically determined and carried out in the simulations, using eq.* (3.18). *The* 0.5*th percentile* t_{switch} = 80 *days was used as a fixed switching time in the suggested IM-strategy. Hybrid simulations were performed at clinically relevant parameter sets, cf. [139]. Drug switches occurred in a total of* 5, 478 *out of* 6, 000 *simulations.*

perspective of any potential treatment d. As the viral population adapts to some currently applied treatment, $R_{\text{cap}}(d)$ changes accordingly: $R_{\text{cap}}(d)$ is large initially and decreases subsequently until drug-resistant strains develop and begin to render treatment d inefficient. We want to assess the point in time when an inducer combination stops to provide benefits (in terms of the viral population) for the next (maintenance) drug combination. Thus, we evaluate $R_{\text{cap}}(d)$ for d = *maintenance combination* while the inducer combination is applied and change from induction to maintenance therapy when $R_{\text{cap}}(d)$ reaches its minimum, i.e.,

$$\text{switch if:} \qquad \frac{\mathrm{d}}{\mathrm{d}t} R_{\text{cap}}(d) = 0. \tag{3.18}$$

The derived individual switching times from a total of 6,000 hybrid simulations are shown in Figure 3.3. We chose the 0.5th percentile t_{switch} = 80 days as a fixed time for treatment change in the proposed IM-strategy. In the sequel, we evaluate, if the proposed IM-strategy with the chosen value for t_{switch} leads to a general improvement compared to conventional HAART, in terms of treatment success and drug-resistance archiving.

Implementation of Conventional vs. Proposed IM-Strategy

In order to reflect the clinical practice of HIV care, we implemented the following routine for assessing the efficacy of the applied treatment combinations. Our virtual patients are monitored every month for efficacy assessment until total virus level falls below the detection limit (50 HIV RNA/mL plasma). Thereafter, they are monitored every other month. Virological failure has been defined according to current treatment guidelines [168]: At the first efficacy assessment (one month after treatment initiation), viral load should have fallen by at least 2 logs HIV RNA/mL plasma. Each consecutive measurement should be below the previous assessment. By month 4, viral load should be below the level of detection (50 HIV RNA/mL plasma). After that, detectable virus is defined as virological failure.

We implemented HAART as follows: Initially, the virtual patient is treated with a combination of two nucleoside reverse transcriptase inhibitors (NRTIs) and one non-nucleoside reverse transcriptase inhibitor (NNRTI) (e.g., tenofovir (TDF) + emtricitabine (FTC) + efavirenz (EFV)), until virological failure is detected. Then, treatment is changed to a second-line regimen consisting of a protease inhibitor (PI), an integrase inhibitor (INI) and an entry inhibitor (EI) (e.g., ritonavir (RTV) boosted PI + raltegravir (RLV) + maraviroc (MVR)).

In the proposed IM-strategy, patients are initially treated with a combination consisting of a PI, an InI and an EI, until t_{switch} = 80 days. After that, a treatment consisting of two NRTIs

and one NNRTI is applied. If virological failure is detected at any efficacy assessment time point, treatment change is applied.

We performed 1,000 hybrid simulations of both strategies, conventional HAART and the proposed IM-therapy, for each beforehand identified, clinically relevant parameter set (see first column in Table 3.1 and further [139]), and counted the number of realizations in which virological failure occurred. Furthermore, we assessed if the number of drug-resistant mutants in the very long-lived, latently infected T-cells T_L was higher at simulation end than upon treatment initiation. In such case, we recorded 'archiving' of drug-resistance.

Table 3.1: *Probability of virological failure and archiving of multi-drug resistant virus during proposed IM-strategy vs. conventional HAART. Columns* 2–5 *show the distinct treatment outcome of* 1,000 *hybrid stochastic–deterministic simulations for the IM-strategy (left entry) and HAART (right entry), respectively, for different parameter sets* P_1–P_{12} *in terms of drug efficacy* $(1-\eta)$ *and mutation-associated reproductive fitness loss* s*. Column* 2*: Percentage of virological failure after* 2 *years of therapy according to current HIV treatment guidelines [168]. Columns* 3–5*: Probability of the archiving of multi-drug resistance during treatment. Further, the results of cross-tab* χ^2*-tests of independence between treatment strategy (IM vs. HAART) and outcome are stated (§:* $p = 0.001$*,* $*$*:* $p = 0.05$*). A small p-value indicates that different outcomes are due to the treatment strategy and not to random effects.*

Parameter Set	**Failure Rate**	**Probability of Multi-Drug Resistance Archiving**		
ID $(1-\eta; s)$	IM, HAART	≥ 2 mutations	≥ 3 mutations	≥ 5 mutations
P_1 (0.70; 0.30)	1.7 %, 4.8 %§	1.8 %, 4.8 %§	1.7 %, 4.8 %§	0.0 %, 0.1 %
P_2 (0.70; 0.25)	4.2 %, 14.2 %§	4.8 %, 14.2 %§	4.2 %, 13.9 %§	0.1 %, 0.2 %
P_3 (0.70; 0.20)	6.6 %, 41.8 %§	18.5 %, 42.2 %§	9.6 %, 41.6 %§	0.1 %, 2.9 %§
P_4 (0.75; 0.25)	0.9 %, 2.8 %*	0.9 %, 2.9 %*	0.9 %, 2.8 %*	0.0 %, 0.0 %
P_5 (0.75; 0.20)	1.8 %, 12.5 %§	2.2 %, 12.6 %§	1.8 %, 12.5 %§	0.0 %, 0.4 %
P_6 (0.80; 0.20)	0.7 %, 2.2 %*	0.8 %, 2.3 %*	0.7 %, 2.2 %*	0.0 %, 0.2 %
P_7 (0.80; 0.15)	3.1 %, 21.9 %§	2.8 %, 22.1 %§	3.1 %, 21.9 %§	0.2 %, 0.9 %*
P_8 (0.80; 0.10)	7.9 %, 44.0 %§	9.3 %, 44.0 %§	8.3 %, 44.0 %§	0.7 %, 14.6 %§
P_9 (0.85; 0.15)	0.6 %, 0.6 %	0.9 %, 1.3 %	0.6 %, 0.6 %	0.0 %, 0.0 %
P_{10} (0.85; 0.10)	2.4 %, 7.1 %§	2.7 %, 8.1 %§	2.4 %, 7.2 %§	0.3 %, 0.4 %
P_{11} (0.85; 0.05)	33.7 %, 59.1 %§	34.7 %, 59.5 %§	34.0 %, 59.3 %§	3.4 %, 17.2 %§
P_{12} (0.90; 0.05)	1.2 %, 1.8 %	2.3 %, 2.5 %	1.3 %, 1.8 %	0.1 %, 0.1 %

Proposed IM-Strategy Improves Success Rate and Minimizes Archiving of Drug-Resistance

Table 3.1 (third–fifth column) shows the number of cases in which archiving of multi-drug resistant viral strains (with ≥ 2, ≥ 3 and ≥ 5 drug-resistance mutations) occurred in the latent reservoir under the proposed IM-strategy and conventional HAART, respectively. It can be seen that the proposed treatment strategy leads to a significant reduction in multi-drug resistance archiving for the majority of evaluated parameters. This indicates, that although two treatment lines have been used for the novel IM therapy, more therapeutic options are on average available in the follow-up period, compared to conventional therapy.

Figure 3.4A shows that the proposed IM-strategy (blue line) with a fixed treatment switching time of t_{switch} = 80 days leads to a significant reduction in the probability to experience virological failure compared to the conventional treatment strategy (red line). This observation holds true for a wide range of parameters (see Table 3.1, second column). In only two cases, where virological failure rarely occurs during conventional therapy, we do not get significant differences at the $p = 0.05$ level. As further depict in Figure 3.4B, virological failure and the average number of archived drug-resistance mutations are strongly correlated (Spearman's

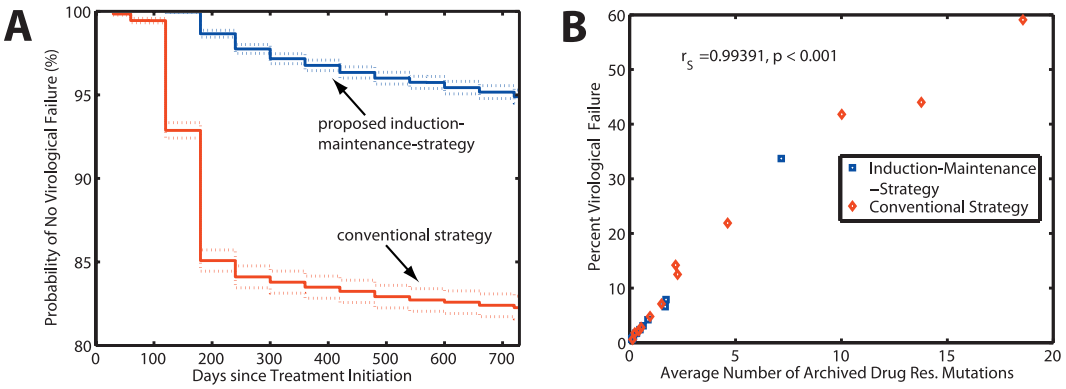

Figure 3.4: *Kaplan–Meier estimates of treatment success, and correlation between virological failure and drug-resistance archiving. The plots summarize the results trough the whole simulated parameter space from Table 3.1 (*12,000 *hybrid simulations in total). A: Probability of no virological failure for the IM-strategy (blue line) and conventional HAART (red line), respectively. Dashed lines are the* 95 % *confidence ranges, calculated using Greenwood's formula. Virological failure was defined according to current treatment guidelines [168]. B: Probability of virological failure vs. average number of drug-resistance archiving in the latent reservoir. A strong positive correlation ($p < 0.001$) between virological failure and drug-resistance archiving exists, as indicated by Spearman's non-parametric rank correlation coefficient $r_S > 0.99$.*

correlation coefficient $r_S > 0.99$, $p < 0.001$). This indicates that virological failure is a strong predictor for drug-resistance archiving.

3.2.3 Discussion

We studied an application of a hybrid method to a comprehensive HIV-dynamics model that well-illustrated the advantages obtained by a stochastic–deterministic system representation: Neither a purely stochastic (due to numerical complexity) nor a purely deterministic approach (due to lack of stochastic effects) would have allowed to perform our *in silico* study. Using hybrid simulations, however, we were able to derive and test (in terms of the mathematical model) a very simple treatment strategy that can lead to significant reductions in virological failure in comparison to conventional HAART.

A unique inducer combination is used for a short time (80 days) and pro-actively switched to a maintenance combination. The purpose of the inducer combination is to decrease the viral population size and thereby increase the likelihood that the subsequent (maintenance) therapy will achieve durable suppression. Clinical implementation of this novel treatment strategy requires only one additional clinical visit after 80 days in comparison to conventional HAART. The important finding of our study is that although two drug combinations are always utilized during the proposed IM-strategy, less archiving of drug-resistance occurs compared to conventional HAART, where a second treatment line would be applied only in the case of virological failure or toxicity. Less drug-resistance archiving implies that more treatment options will be available for the follow-up and long-term management of HIV-infected patients when the proposed IM-strategy is used (see Table 3.1, third–fifth column).

Only a few archiving events (≥ 40 fully resistant mutants) are sufficient to eliminate treatment options permanently, cf. [139]. The number of circulating latently infected cells is small [152, 154, 174, 175]. Detecting a small subset of mutants within the circulating latently infected cells is experimentally not feasible, because standard sequencing technology will detect the major strains [176], while novel, second generation methods require large samples [177]. Hence, mathematical modeling is a reasonable tool to investigate drug-resistance archiving following treatment application.

The time for switching between combinations t_{switch} (= 80 days) is the most critical parameter for the success of the proposed IM-strategy. The following two considerations have to be taken into account: (i) The inducer combination should be applied only for a short time, to prevent the selection and archiving of mutants, which are resistant to the *current* drug

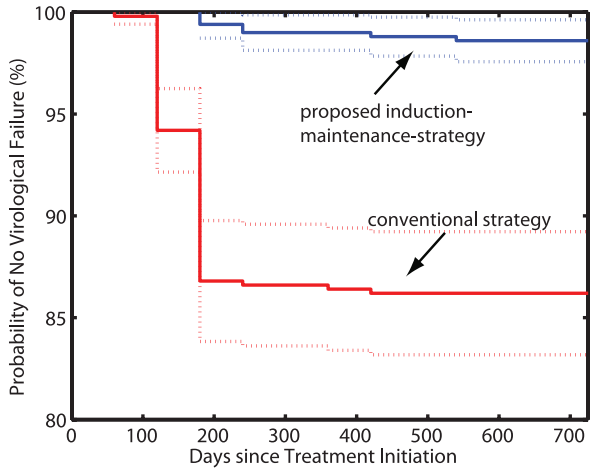

Figure 3.5: *Kaplan–Meier estimates of treatment success (probability of no virological failure) for very high initial abundance of drug-resistant mutants. The outcome of* 500 *hybrid simulation for the proposed IM-strategy (blue line) and for the conventional HAART (red line), respectively, are shown. Dashed lines indicate the* 95 % *confidence ranges, calculated using Greenwood's formula. The initial abundance of drug-resistant mutants was set to* 1 % *of the population. Other parameter values:* $(1 - \eta) = 0.75$, $s = 0.8$.

combination and would limit its further use (risk of the strategy), (ii) while at the same time, it has to be applied long enough to possibly eradicate viral mutants, which are resistant to the *next* drug combination (benefit of the strategy).

The time required for resistant mutants to emerge during therapy depends on their abundance before therapy initiation (if they pre-exist and are selected from the quasi-species population) and on their genetic distance to the wild type (if resistance is *de novo* developed). As discussed above, we determine the abundance of mutants at the time of therapy initiation by utilizing the deterministic fix-point of the model as a starting condition for the hybrid simulations. In Figure 3.5, we demonstrate the non-inferiority of our approach if drug-resistant mutants are more abundant than expected (1 % of the wild type level, i.e., the detection limit of second generation sequencing technologies [177–179]). We assumed the shortest genetic distance possible between wild type and fully drug-resistant mutants (one mutation creates full resistance against a single drug, such that three distinct mutations create full resistance against a triple drug combination). For some drugs, however, only subsequent accumulation of mutations creates fully drug-resistant mutants [180]. In the hybrid simulations, drug-resistance thus develops more rapidly than *in vivo* for drugs with a large genetic barrier [181]. This implies that the inducer combination could possibly be applied for a longer time than the predicted 80 days. However, our results demonstrate that even this very short time can improve the clinical outcome significantly (see Figures 3.4 and 3.5, and Table 3.1).

Eradication of viral mutants also depends critically on their abundance before therapy initiation and on the rate at which viral compartments (and hence resistant mutants) are

cleared. The *in vivo* elimination rate of viral compartments has been quantified and validated in a number of clinical studies [182–184]. We used the expected abundance of viral mutants (the deterministic fix-point of the model) to estimate the abundance of different viral mutants at the time of treatment initiation. The results in Figure 3.5 also show non-inferiority of our approach for the case where an unexpectedly high abundance of drug-resistant mutants is present, which would require longer time for eradication.

One limitation of the proposed IM-strategy is the potential inability to eliminate viral strains that carry resistance to the maintenance therapy. This is particularly the case, if viral mutants, carrying resistance against all (or the majority of) drugs in the maintenance combination, are already archived in the latent reservoir before treatment initiation. In patients infected by wild type virus, however, the likelihood that part-resistant mutants (against two out of three drugs in the maintenance regimen) enter the latent reservoir prior to treatment is so small that their expected number archived in the latent reservoir is uncritical (much less than one), cf. [139]. Furthermore, part-resistant mutants are still susceptible to one out of the three drugs in the maintenance combination, and for triple-drug (full-)resistant strains, the likelihood of archival copies is even smaller.

Infection with drug-resistant strains against established drug classes is a major, growing health concern [185]. During infection with drug-resistant viral strains, archiving in the latent reservoir is likely, as it is established early in the infection [186]. If the circulating viral population reverses to a drug-susceptible type, archived drug-resistant mutants from the time of infection might remain undetected and can complicate subsequent treatment, cf. [139]. However, this particular circumstance applies equally to the proposed IM-therapy and conventional HAART.

For our strategy, we have chosen drugs from novel classes (e.g., INI, EI) for the inducer combination, while we selected drugs from well-established classes for the maintenance combination (NNRTI, NRTI). This has the following rationale: The inducer combination will only be applied for a short time (80 days), while the maintenance combination could possibly be applied for much longer periods of time (until it fails, or toxicological events occur). Second or third generation drugs within the established drug classes are often more convenient to apply and less toxic, which has important implications for the long-term management of HIV [187]. Secondly, drugs from the novel drug classes (i.e., INI, EI) are currently not available as generic formulations, whereas low-cost alternative drugs exist for established drug classes. Therefore, in order to reduce treatment costs, it is of advantage to select a strategy, in which

inexpensive drugs can be used for the majority of time, while cost-intensive ones are only applied for short treatment periods.

Intuitively, it might be more advantageous to use drug-resistance tests to guide treatment switches, instead of using a fixed time for a pro-active switch from inducer to maintenance combination [166]. However, under the considerations discussed above, such switch should be applied before any resistant strains become abundant. This implies that the most frequent viral strain at the time of switch should be the wild type. Standard assays fail to detect minority species [176]. Ultra-deep/pyro-sequencing might provide a more holistic picture of the quasi-species composition and can pick up viral mutants that are abundant in ≈ 1 % of the quasi-species population if the sample is large enough [177–179]. However, even in this case, viral mutants are likely to dominate once the results are available (> 1 week), owing to the rapid viral kinetics [188].

In our *in silico* study, we considered time-invariant, as well as anatomically homogeneous average drug efficacy $(1 - \eta)$ for ease of modeling. It is also possible to consider drug- and patient-specific time-varying pharmacokinetics and to study the impact of compliance on drug-resistance development. According to the performed hybrid simulations, however, the qualitative difference between the outcome of conventional HAART versus the proposed IM-strategy is not expected to change if compliance is identical between the two study arms. As shown in Table 3.1, the proposed IM-therapy performs better than conventional HAART for a wide range of parameter values for $(1 - \eta)$. Furthermore, it was shown in a clinical study [144, 145] that treatment alternation leads to significantly less virological failure than conventional HAART, when compliance is imperfect but identical between the two study arms. However, since the study in [144, 145] is not identical to our IM-strategy, a clinical study should be performed to fully investigate its potential. Ideally, this prospective randomized trial could evaluate the time to virological failure in patients taking a single unchanged regimen and patients on IM-regimens. Importantly, the trial should be designed to evaluate whether the IM-strategy affects the durability of second- and third-line regimens. The presence and relative frequency of viral minority populations as well as their mutational patterns could be monitored by analyzing proviral DNA from circulating T-cells using, e.g., next-generation sequencing. This data could serve to validate the improvements expected from our *in silico* predicted IM-treatment strategy.

4

Hybrid Stochastic–Deterministic Solution of the CME[‡]

In this chapter, we present and theoretically justify a hybrid stochastic–deterministic approach to solve the CME *directly*. Starting point is a partitioning of the system into discrete and continuous species and reactions, see Subsection 4.1.1. Accordingly, we decompose the joint PDF $P(Y, Z; t)$ into the marginal PDF $P(Z; t)$ of the discrete species and the conditional PDF $P(Y \mid Z; t)$ of the continuous species. In light of the thermodynamic limit, we assume that the PDF $P(Y \mid Z; t)$ of the continuous species will be tightly concentrated about a deterministic path for given levels Z of the discrete species. Based on a scaling parameter ε, motivated by large population levels and fast reactions, and a corresponding multi-scale expansion of the ε-scaled CME that uses the WKB-approximation, we demonstrate that fluctuations in the levels of the continuous species can indeed be neglected and $P(Y \mid Z; t)$ is well-approximated by its distribution mode. Applying Laplace's method of integral approximation, we are then able to derive hybrid evolution equations that comprise a CME for the PDF $P(Z; t)$ of the discrete species coupled to evolution equations for the related expected levels of the continuous species for each discrete state Z. This approach can be interpreted as taking 'partial' expectations over the continuous species, see Subsection 4.1.2.

Our hybrid method does not suffer from the disadvantages of indirect methods, see Chapter 1. In contrast to indirect hybrid methods, the impact of changes in the *distribution* of the discrete species on the dynamics of the continuous species is taken into account explicitly, see Subsection 4.1.7. Our direct hybrid approach is more efficient than indirect approaches if the reaction system comprises a few species in low quantities and the remaining species in larger levels or associated with rapidly firing reaction channels. This is typically the case for systems that integrate gene expression, regulation and metabolic pathways. We illustrate the performance of our hybrid stochastic–deterministic approach in applications to different model systems of biological interest.

[‡]The contents of this chapter are based on: S. Menz, J. C. Latorre, C. Schütte, and W. Huisinga. Hybrid Stochastic–Deterministic Solution of the Chemical Master Equation. *Multiscale Modeling and Simulation* (under review), 2011.

4.1 Derivation of the Hybrid CME–ODE Method

In the following, we derive a general hybrid description of the system dynamics where the time-evolution of the PDF $P(Z;t)$ of species present in small numbers Z is coupled to the time-evolution of the partial expectations

$$\mathbb{E}_Z[\boldsymbol{Y}] := \sum_Y YP(Y,Z;t) = P(Z;t)\sum_Y YP(Y\,|\,Z;t)$$

of species with population levels Y adequate for a continuous deterministic approximation. The derivation is based on the WKB-ansatz used in Subsection 2.4.4 to link the ODE-model of classical reaction kinetics to the leading order WKB-approximation of the full CME. In order to maintain the effects of fluctuations on the system dynamics, we apply this ansatz in the following only partially, i.e., we seek a WKB-approximation for the conditional PDF $P(Y\,|\,Z;t)$ of the continuous species. This approach will allow us to derive the dynamics of both species on the different time scales. The resulting equations for the short-time scale can be understood in the sense of the central limit theorem (CLT): Species with high population levels and fast reaction rates will be sharply distributed around the unique most probable value. The time-evolution of this point will be formally derived and its uniqueness, together with the resulting asymptotic form of the conditional probability $P(Y\,|\,Z;t)$, allow us to apply Laplace's method in order to obtain evolution equations that live on the longer-time scale.

4.1.1 Partitioning of the System

We partition the system with respect to the species and their expected number of molecules. Assume that for a given reaction network it can be distinguished between:

(i) 'Continuous' species S_i^c, $i \in \mathcal{N}^c \neq \varnothing$, whose changes in number of molecules are approximated by continuous deterministic processes.

(ii) 'Discrete' species S_i^d, $i \in \mathcal{N}^d = \{1,\ldots,N\}\setminus\mathcal{N}^c$, whose changes in number of molecules retain a discrete stochastic description.

Intuitively, we aim at a partitioning of the set of species into those that are present in small numbers and where stochastic fluctuations might be important (discrete species), and those that are present in larger numbers such that a representation by continuous variables

introduces only a negligible error and where fluctuations are expected to be less important. This partitioning is disjoint, i.e., $\mathcal{N}^c \cup \mathcal{N}^d = \{1, \ldots, N\}$, and we rearrange the species-related variables accordingly:

$$X = (Y, Z)^{\mathrm{T}}, \quad \text{and} \quad \xi_\mu = (\upsilon_\mu, \zeta_\mu)^{\mathrm{T}} \quad (\mu = 1, \ldots, M),$$

where Y and υ_μ denote the number of molecules and net changes of all continuous species S_i^c, with $i \in \mathcal{N}^c$, and Z and ζ_μ denote the number of molecules and net changes of all discrete species S_i^d, with $i \in \mathcal{N}^d$, respectively. Finally, we represent the joint probability function $P(Y, Z; t)$ using conditional probabilities as

$$P(Y, Z; t) = P(Y \mid Z; t) P(Z; t).$$

To correctly account for stochastic fluctuations in the discrete variable Z, we assume that all reactions that act on a discrete species S_i^d are modeled as discrete stochastic processes (in line with the partitioning criteria in indirect hybrid approaches, see Subsection 3.1.1). All other reactions influence only the continuous species. Based on the above assumption, these reactions are approximated as continuous deterministic processes. Hence, the discrete–continuous partitioning of the species induces a corresponding partition of the reactions:

(i) 'Continuous' reactions do not change the number of molecules of any discrete species S_i^d, i.e.,

$$\zeta_\mu = 0 \quad \forall \mu \in \mathcal{M}^c, \tag{4.1a}$$

where $\mathcal{M}^c \neq \varnothing$ denotes the set of all continuous reactions.

(ii) 'Discrete' reactions change the number of molecules of at least one discrete species S_i^d, i.e.,

$$\zeta_\mu \neq 0 \quad \forall \mu \in \mathcal{M}^d, \tag{4.1b}$$

where $\mathcal{M}^d = \{1, \ldots, M\} \setminus \mathcal{M}^c$ denotes the set of all discrete reactions.

Notice that this partitioning is again disjoint, i.e., $\mathcal{M}^c \cup \mathcal{M}^d = \{1, \ldots, M\}$.

As a consequence of the above partitioning, the net changes of the reactions can be rearranged as

$$\begin{array}{c} \\ \text{species} \\ \begin{array}{c} \text{cont.} \\ \text{disc.} \end{array} \end{array} \begin{array}{c} \begin{array}{cc} \text{reactions} & \\ \text{cont.} & \text{disc.} \end{array} \\ \begin{pmatrix} \upsilon_{\mathcal{M}^c} & \upsilon_{\mathcal{M}^d} \\ 0 & \zeta_{\mathcal{M}^d} \end{pmatrix} \end{array}.$$

Grouping terms together, we thus obtain the CME of the partitioned system as

$$\begin{aligned}\frac{\partial}{\partial t}\big[P(Y\,|\,Z;t)P(Z;t)\big] = P(Z;t)\sum_{\mu\in\mathcal{M}^c} a_\mu(Y-v_\mu,Z)P(Y-v_\mu\,|\,Z;t) - a_\mu(Y,Z)P(Y\,|\,Z;t)\\ + \sum_{\mu\in\mathcal{M}^d}\big[a_\mu(Y-v_\mu,Z-\zeta_\mu)P(Y-v_\mu\,|\,Z-\zeta_\mu;t)P(Z-\zeta_\mu;t)\\ - a_\mu(Y,Z)P(Y\,|\,Z;t)P(Z;t)\big].\end{aligned} \tag{4.2}$$

4.1.2 Partial Averaging of the CME

Our hybrid solution of the CME can be interpreted as taking partial expectations over the continuous species. In the following, we introduce this approach by studying the evolution of the partial expectation $\mathbb{E}_Z[f(\boldsymbol{Y})] = \sum_Y f(Y)P(Y,Z;t)$ of a suitable test function f of $\boldsymbol{Y}$. First, notice that a partial averaging of the CME (4.2) can be performed along the lines of the total averaging discussed in Subsection 2.4.3: By multiplying the partitioned CME (4.2) with $f(Y)$, summing over Y, assuming existence of sufficiently many partial moments[1] of $\boldsymbol{Y}$ and re-arranging terms, we find the time-evolution of the partial expectation of f to be given as

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}_Z\big[f(\boldsymbol{Y})\big] = \sum_{\mu\in\mathcal{M}^c}\mathbb{E}_Z\big[\big(f(\boldsymbol{Y}+v_\mu)-f(\boldsymbol{Y})\big)a_\mu(\boldsymbol{Y},Z)\big]\\ + \sum_{\mu\in\mathcal{M}^d}\mathbb{E}_{Z-\zeta_\mu}\big[f(\boldsymbol{Y}+v_\mu)a_\mu(\boldsymbol{Y},Z-\zeta_\mu)\big] - \mathbb{E}_Z\big[f(\boldsymbol{Y})a_\mu(\boldsymbol{Y},Z)\big].\end{aligned} \tag{4.3}$$

For $f(\,\cdot\,)\equiv 1$ the above equation gives the evolution of the PDF $P(Z;t)$ of the discrete species, i.e.,

$$\begin{aligned}\frac{\partial}{\partial t}P(Z;t) &= \sum_{\mu\in\mathcal{M}^d}\mathbb{E}_{Z-\zeta_\mu}\big[a_\mu(\boldsymbol{Y},Z-\zeta_\mu)\big] - \mathbb{E}_Z\big[a_\mu(\boldsymbol{Y},Z)\big]\\ &= \sum_{\mu\in\mathcal{M}^d}\mathbb{E}\big[a_\mu(\boldsymbol{Y},Z-\zeta_\mu)\,|\,Z-\zeta_\mu\big]P(Z-\zeta_\mu;t) - \mathbb{E}\big[a_\mu(\boldsymbol{Y},Z)\,|\,Z\big]P(Z;t),\end{aligned} \tag{4.4}$$

where $\mathbb{E}[f(\boldsymbol{Y},\,\cdot\,)\,|\,\cdot\,] \coloneqq \sum_Y f(Y,\,\cdot\,)P(Y\,|\,\cdot\,;t)$ denotes the conditional expectation of a function f of $\boldsymbol{Y}$, and by definition $\mathbb{E}_Z[f(\boldsymbol{Y},Z)] = \mathbb{E}[f(\boldsymbol{Y},Z)\,|\,Z]P(Z;t)$. We observe that if all discrete reactions are at most of first order with respect to the continuous species, i.e.,

[1] In line with the definition of partial expectations, we call $\mathbb{E}_Z[\boldsymbol{Y}^m]$ the *partial moment* of $\boldsymbol{Y}\in\mathbb{N}_0^{N^c}$ associated with the vector $m = (m_1,\dots,m_{N^c})^{\mathrm{T}}\in\mathbb{N}_0^{N^c}$, where N^c denotes the number of continuous species.

$|s^r_\mu|_{\mathcal{N}^c} := \sum_{i\in\mathcal{N}^c} s^r_{\mu i} \leq 1$ for all $\mu \in \mathcal{M}^d$, then we can put $\mathbb{E}[a_\mu(\boldsymbol{Y}, \cdot)\,|\,\cdot] = a_\mu(\mathbb{E}[\boldsymbol{Y}\,|\,\cdot], \cdot)$ in eq. (4.4), such that

$$\frac{\partial}{\partial t}P(Z;t) = \sum_{\mu\in\mathcal{M}^d} a_\mu\big(\mathbb{E}[\boldsymbol{Y}\,|\,Z-\zeta_\mu], Z-\zeta_\mu\big)P(Z-\zeta_\mu;t) - a_\mu\big(\mathbb{E}[\boldsymbol{Y}\,|\,Z], Z\big)P(Z;t). \quad (4.5)$$

Given the related partial expectations $\mathbb{E}_Z[\boldsymbol{Y}]$ of the continuous species,[2] we could hence propagate $P(Z;t)$ in this case by solving eq. (4.5). For general reaction networks, however, eq. (4.5) is not exact.

By taking $f(Y) = Y$ in eq. (4.3), we find the time-evolution of the partial expectations $\mathbb{E}_Z[\boldsymbol{Y}]$ of the continuous species to be given as

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}_Z[\boldsymbol{Y}] = &\sum_{\mu\in\mathcal{M}^c} v_\mu \mathbb{E}_Z\big[a_\mu(\boldsymbol{Y}, Z)\big] \\ &+ \sum_{\mu\in\mathcal{M}^d} \mathbb{E}_{Z-\zeta_\mu}\big[(\boldsymbol{Y}+v_\mu)a_\mu(\boldsymbol{Y}, Z-\zeta_\mu)\big] - \mathbb{E}_Z\big[\boldsymbol{Y}a_\mu(\boldsymbol{Y}, Z)\big]. \quad (4.6)\end{aligned}$$

Similarly as above, taking expectations and propensities can be exchanged in the first sum on the right hand side of eq. (4.6) if all continuos reactions are at most of first order with respect to Y. In the second sum, however, this is only possible if all discrete reactions are of zero order with respect to Y, such that their propensities are independent of Y, i.e., $a_\mu(Y, Z) = a_\mu(Z)$ for all $\mu \in \mathcal{M}^d$. Thus, if (a) $|s^r_\mu|_{\mathcal{N}^c} \leq 1$ for all $\mu \in \mathcal{M}^c$ and (b) $|s^r_\mu|_{\mathcal{N}^c} = 0$ for all $\mu \in \mathcal{M}^d$, we find that

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}_Z[\boldsymbol{Y}] = &\sum_{\mu\in\mathcal{M}^c} v_\mu a_\mu\big(\mathbb{E}[\boldsymbol{Y}\,|\,Z], Z\big)P(Z;t) + \sum_{\mu\in\mathcal{M}^d} v_\mu a_\mu(Z-\zeta_\mu)P(Z-\zeta_\mu;t) \\ &+ \sum_{\mu\in\mathcal{M}^d} \mathbb{E}_{Z-\zeta_\mu}[\boldsymbol{Y}]a_\mu(Z-\zeta_\mu) - \mathbb{E}_Z[\boldsymbol{Y}]a_\mu(Z). \quad (4.7)\end{aligned}$$

We notice that under condition (b) $|s^r_\mu|_{\mathcal{N}^c} = 0$ for all $\mu \in \mathcal{M}^d$, the evolution of $P(Z;t)$ is given by eq. (4.5). Thus, if conditions (a) and (b) are satisfied in a given reaction system, then eqs. (4.5) and (4.7) give a closed system of exact evolution equations for the PDF $P(Z;t)$ of the discrete species and the partial expectations $\mathbb{E}_Z[\boldsymbol{Y}]$ of the continuous species for every discrete

[2] Notice that terms of the form $a_\mu(\mathbb{E}[\boldsymbol{Y}\,|\,Z], Z)P(Z;t)$ can be computed from (i) $P(Z;t)$ or (ii) $\mathbb{E}_Z[\boldsymbol{Y}]$ without solving for $\mathbb{E}[\boldsymbol{Y}\,|\,Z]$ if the corresponding reaction is of (i) zero or (ii) first order with respect to the continuous species, as assumed in eq. (4.5). Because then: (i) the propensity a_μ is independent of $\mathbb{E}[\boldsymbol{Y}\,|\,Z]$, or (ii) we have $a_\mu(\mathbb{E}[\boldsymbol{Y}\,|\,Z], Z)P(Z;t) = a_\mu(\mathbb{E}_Z[\boldsymbol{Y}], Z)$.

state Z.[3] The full discrete state space is then effectively reduced to the subspace associated with the discrete species and all problematic dimensions, associated with species present in large numbers, are canceled out. In the following, we demonstrate that this approach can be generalized to approximate the dynamics of more complex systems.

4.1.3 Scaling of the Continuous Species and Reactions

According to our assumption on the continuous species, we scale their population levels with a factor $\varepsilon \ll 1$, i.e.,

$$y := \varepsilon \cdot Y. \tag{4.8}$$

The parameter ε is related to the abundance of the continuous species S_i^c and used in the following asymptotic approximation to derive a partial limit of reaction kinetics. The exact value of ε may not be required, since the final equations in the scaled state space can be transformed back to the original unscaled state space. However, as the following hybrid approach gives an asymptotic approximation, the resulting error depends on the validity of this partial continuous–deterministic approximation and only vanishes in the limit as $\varepsilon \to 0$.

In order to keep the probability invariant under the change of variables (4.8), the PDF of the scaled population levels is given by

$$P_\varepsilon(y, Z; t) = P_\varepsilon(y \,|\, Z; t) P_\varepsilon(Z; t) := \varepsilon^{-N^c} \cdot P(Y \,|\, Z; t) P(Z; t),$$

where N^c denotes the number of continuous species. Hence, with respect to the scaled levels, the partitioned CME (4.2) reads

$$\begin{aligned}
&\frac{\partial}{\partial t}\big[P_\varepsilon(y \,|\, Z; t) P_\varepsilon(Z; t)\big] \\
&\qquad = P_\varepsilon(Z; t) \sum_{\mu \in \mathcal{M}^c} a_\mu^\varepsilon(y - \varepsilon v_\mu, Z) P_\varepsilon(y - \varepsilon v_\mu \,|\, Z; t) - a_\mu^\varepsilon(y, Z) P_\varepsilon(y \,|\, Z; t) \\
&\qquad\quad + \sum_{\mu \in \mathcal{M}^d} \big[a_\mu^\varepsilon(y - \varepsilon v_\mu, Z - \zeta_\mu) P_\varepsilon(y - \varepsilon v_\mu \,|\, Z - \zeta_\mu; t) P_\varepsilon(Z - \zeta_\mu; t) \\
&\qquad\qquad - a_\mu^\varepsilon(y, Z) P_\varepsilon(y \,|\, Z; t) P_\varepsilon(Z; t)\big],
\end{aligned} \tag{4.9}$$

[3] It should be noticed that even though a discrete reaction does not necessarily has to act on the continuous species, i.e., $v_\mu = 0$ for some $\mu \in \mathcal{M}^d$, it *always* contributes to the evolution of the partial expectations $\mathbb{E}_Z[\boldsymbol{Y}]$ by changing their distribution on the discrete state space through the terms in the last sum on the right hand side of eq. (4.7). This impact of a discrete reaction on the expectations of the continuous species is not immediately present in indirect hybrid methods. It becomes only *implicitly* incorporated when the statistical properties of an ensemble of hybrid realizations are computed, see further Subsection 4.1.7.

where $a_\mu^\varepsilon(y, Z) = a_\mu(Y = y/\varepsilon, Z)$ for all $\varepsilon > 0$. Intuitively it is clear that the intensity of a reaction process does not depend on the scale of the reactant levels, see also [189]. For example, one might interpret the scaling as some transformation of units.

In accordance with the definition of the continuous species, we assume that the scaling of their levels imposes a corresponding scaling of the continuous reactions, analogously to the deterministic formulation of reaction kinetics. If $N^c = N$ and thus all levels are scaled, we require that our hybrid approach coincides with the purely deterministic limit. In this case, the parameter ε can be linked to the Ω-scaling in classical reaction kinetics via $\varepsilon = \Omega^{-1}$ (cf. Section 2.4). Therefore, in line with the results in Subsection 2.4.1, we assume a corresponding ε-scaling of the propensities of all continuous reactions

$$\alpha_\mu(y, Z) := \varepsilon \cdot a_\mu^\varepsilon(y, Z) = \varepsilon \cdot a_\mu(Y = y/\varepsilon, Z) \quad \forall \mu \in \mathcal{M}^c. \tag{4.10a}$$

In our context, however, the above ε-scaling of propensities is only applied to the subset $\mathcal{M}^c$ of the reaction system, since firing of a discrete channel results in changes of the process $\boldsymbol{Z}(t)$, which, by definition, necessitates stochastic reaction kinetics. Hence, the propensities of all discrete reactions are assumed to satisfy

$$\alpha_\mu(y, Z) := a_\mu^\varepsilon(y, Z) = a_\mu(Y = y/\varepsilon, Z) \quad \forall \mu \in \mathcal{M}^d. \tag{4.10b}$$

Intuitively, we require that—due to the larger number of molecules of the continuous species and their appropriate representation by continuous variables—the propensities of reactions that exclusively act on the continuous species are of the order of $1/\varepsilon \gg 1$ larger than the propensities of reactions that act on the discrete species. Under assumptions (4.10), eq. (4.9) becomes

$$\begin{aligned}
&\frac{\partial}{\partial t}\big[P_\varepsilon(y \,|\, Z; t) P_\varepsilon(Z; t)\big] \\
&\qquad = \frac{1}{\varepsilon} P_\varepsilon(Z; t) \sum_{\mu \in \mathcal{M}^c} \alpha_\mu(y - \varepsilon v_\mu, Z) P_\varepsilon(y - \varepsilon v_\mu \,|\, Z; t) - \alpha_\mu(y, Z) P_\varepsilon(y \,|\, Z; t) \\
&\qquad\quad + \sum_{\mu \in \mathcal{M}^d} \big[\alpha_\mu(y - \varepsilon v_\mu, Z - \zeta_\mu) P_\varepsilon(y - \varepsilon v_\mu \,|\, Z - \zeta_\mu; t) P_\varepsilon(Z - \zeta_\mu; t) \\
&\qquad\qquad\qquad - \alpha_\mu(y, Z) P_\varepsilon(y \,|\, Z; t) P_\varepsilon(Z; t)\big].
\end{aligned} \tag{4.11}$$

In the following, we seek an approximate solution of the ε-scaled CME (4.11) in the form of a multi-scale expansion. Similar as in Subsection 2.4.4, we assume that the conditional

probability $P_\varepsilon(y\,|\,Z;t)$ can be represented in a WKB-like series expansion with respect to the spatial coordinate, i.e.,

$$P_\varepsilon(y\,|\,Z;t) = C_\varepsilon \exp\left\{\frac{1}{\varepsilon}s_0(y\,|\,Z;t)\right\}\big(U_0(y\,|\,Z;t) + \varepsilon U_1(y\,|\,Z;t) + \dots\big), \tag{4.12}$$

where the factor C_ε is related to the normalization of $P_\varepsilon(y\,|\,Z;t)$. As before, we assume that the functions s_0 and U_n, $n = 0, 1, \dots$, in the above asymptotic expansion of $P_\varepsilon(y\,|\,Z;t)$ are sufficiently continuously differentiable with respect to the arguments y and t. We represent the PDF $P_\varepsilon(Z;t)$ of the discrete species in an asymptotic series with respect to the spatial coordinate of the form

$$P_\varepsilon(Z;t) = P_0(Z;t) + \varepsilon P_1(Z;t) + \dots, \tag{4.13}$$

where the functions P_n, $n = 0, 1, \dots$, are assumed to be sufficiently continuously differentiable with respect to time t.

4.1.4 Leading Order Approximation of the Conditional PDF

Below, we determine a solution of the conditional PDF $P_\varepsilon(y\,|\,Z;t)$ to its leading order $\mathcal{O}(\varepsilon^{-1})$, following the same steps as in Subsection 2.4.4. First, we study the left hand side of the ε-scaled CME (4.11). Differentiation of eqs. (4.12) and (4.13) with respect to t gives

$$\begin{aligned}\frac{\partial}{\partial t}P_\varepsilon(y\,|\,Z;t) = {}& \frac{1}{\varepsilon}P_\varepsilon(y\,|\,Z;t)\frac{\partial}{\partial t}s_0(y\,|\,Z;t)\\ &+ C_\varepsilon \exp\left\{\frac{1}{\varepsilon}s_0(y\,|\,Z;t)\right\}\left(\frac{\partial}{\partial t}U_0(y\,|\,Z;t) + \varepsilon\frac{\partial}{\partial t}U_1(y\,|\,Z;t) + \dots\right),\end{aligned}$$

and

$$\frac{\partial}{\partial t}P_\varepsilon(Z;t) = \frac{\partial}{\partial t}P_0(Z;t) + \varepsilon\frac{\partial}{\partial t}P_1(Z;t) + \dots.$$

Hence, on the left hand side of eq. (4.11), we find to leading order

$$\begin{aligned}&\frac{\partial}{\partial t}\big[P_\varepsilon(y\,|\,Z;t)P_\varepsilon(Z;t)\big]\\ &\qquad= P_\varepsilon(Z;t)\frac{\partial}{\partial t}P_\varepsilon(y\,|\,Z;t) + P_\varepsilon(y\,|\,Z;t)\frac{\partial}{\partial t}P_\varepsilon(Z;t)\\ &\qquad= C_\varepsilon\left(\frac{1}{\varepsilon}P_0(Z;t)\exp\left\{\frac{1}{\varepsilon}s_0(y\,|\,Z;t)\right\}U_0(y\,|\,Z;t)\frac{\partial}{\partial t}s_0(y\,|\,Z;t) + \mathcal{O}(1)\right).\end{aligned} \tag{4.14}$$

In the right hand side of eq. (4.11), we Taylor-expand the eikonal function $s_0(y - \varepsilon v_\mu \,|\, Z; t)$ and $U_0(y - \varepsilon v_\mu \,|\, Z; t)$ around the state y, i.e.,

$$s_0(y - \varepsilon v_\mu \,|\, Z; t) = s_0(y \,|\, Z; t) - \varepsilon v_\mu^{\mathrm{T}} \nabla s_0(y \,|\, Z; t) + \mathcal{O}(\varepsilon^2)$$

and

$$U_0(y - \varepsilon v_\mu \,|\, Z; t) = U_0(y \,|\, Z; t) + \mathcal{O}(\varepsilon),$$

respectively. This implies the following expansion of the conditional PDF

$$P_\varepsilon(y - \varepsilon v_\mu \,|\, Z; t) = C_\varepsilon \exp\left\{\frac{1}{\varepsilon} s_0(y \,|\, Z; t)\right\} \exp\left\{-v_\mu^{\mathrm{T}} \nabla s_0(y \,|\, Z; t)\right\}\big(U_0(y \,|\, Z; t) + \mathcal{O}(\varepsilon)\big), \tag{4.15}$$

Similarly, we Taylor-expand $\alpha_\mu(y - \varepsilon v_\mu, Z)$ in eq. (4.11) as

$$\alpha_\mu(y - \varepsilon v_\mu, Z) = \alpha_\mu(y, Z) + \mathcal{O}(\varepsilon). \tag{4.16}$$

Substituting expansions (4.15) and (4.16) into the right hand side of the ε-scaled CME (4.11), using eq. (4.14) on the left hand side, dividing by C_ε and comparing the terms of order $\mathcal{O}(\varepsilon^{-1})$ on both sides, we find that the leading order approximation of the continuous processes for a given discrete state Z to be determined by

$$\frac{\partial}{\partial t} s_0(y \,|\, Z; t) = \sum_{\mu \in \mathcal{M}^c} \alpha_\mu(y, Z)\Big[\exp\left\{-v_\mu^{\mathrm{T}} \nabla s_0(y \,|\, Z; t)\right\} - 1\Big], \tag{4.17}$$

where we divided both sides by $P_0(Z; t) \exp\{\frac{1}{\varepsilon} s_0(y \,|\, Z; t)\} U_0(y \,|\, Z; t)$.[4]

The PDE (4.17) can be considered as the Hamilton–Jacobi equation for the action of a system with coordinates $\boldsymbol{y}(t \,|\, Z)$ and momenta $\boldsymbol{p}(t \,|\, Z) := \nabla s_0(y \,|\, Z; t)$, see Subsection 2.4.4, where the Hamiltonian $\mathcal{H}_Z$ is defined as

$$\mathcal{H}_Z(\boldsymbol{y}, \boldsymbol{p}) := -\frac{\partial}{\partial t} s_0(y \,|\, Z; t) = -\sum_{\mu \in \mathcal{M}^c} \alpha_\mu(\boldsymbol{y}, Z)\big[\exp\{-v_\mu^{\mathrm{T}} \boldsymbol{p}\} - 1\big].$$

The corresponding Hamilton's equations of motion read

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{y}(t \,|\, Z) &= \frac{\partial}{\partial \boldsymbol{p}} \mathcal{H}_Z(\boldsymbol{y}, \boldsymbol{p}) = \sum_{\mu \in \mathcal{M}^c} v_\mu \alpha_\mu\big(\boldsymbol{y}(t \,|\, Z), Z\big) \exp\left\{-v_\mu^{\mathrm{T}} \boldsymbol{p}(t \,|\, Z)\right\}, \\
\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{p}(t \,|\, Z) &= -\frac{\partial}{\partial \boldsymbol{y}} \mathcal{H}_Z(\boldsymbol{y}, \boldsymbol{p}) = \sum_{\mu \in \mathcal{M}^c} \nabla \alpha_\mu\big(\boldsymbol{y}(t \,|\, Z), Z\big)\Big[\exp\left\{-v_\mu^{\mathrm{T}} \boldsymbol{p}(t \,|\, Z)\right\} - 1\Big].
\end{aligned} \tag{4.18}$$

[4] If $P_0(Z; t) \exp\{\frac{1}{\varepsilon} s_0(y \,|\, Z; t)\} U_0(y \,|\, Z; t) = 0$, eqs. (4.11) and (4.14) result in an algebraic equation for $s_0(y \,|\, Z; t)$. This situation will be jointly dealt with the case where $P_0(Z; t) \ll 1$, see Subsection 4.1.6.

In Subsection 2.4.4, we have seen that maxima of $s_0(y \,|\, Z; t = 0)$ are propagated along the characteristic where $\mathcal{H}_Z \equiv 0$.

We assume that the initial function $s_0(y \,|\, Z; t = 0)$ has a *unique* global maximum at $y = \overline{\boldsymbol{y}}_0(Z)$, where

$$\overline{\boldsymbol{p}}_0(Z) \equiv \nabla s_0\big(y = \overline{\boldsymbol{y}}_0(Z) \,|\, Z; t = 0\big) = 0.$$

The equations of motion for the propagation in time of the maximum point $\overline{\boldsymbol{y}}(t \,|\, Z)$ and corresponding $\overline{\boldsymbol{p}}(t \,|\, Z)$ are

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{\boldsymbol{y}}(t \,|\, Z) = \sum_{\mu \in \mathcal{M}^c} \upsilon_\mu \alpha_\mu\big(\overline{\boldsymbol{y}}(t \,|\, Z), Z\big) \quad \text{with} \quad \overline{\boldsymbol{y}}(t = 0 \,|\, Z) = \overline{\boldsymbol{y}}_0(Z) \tag{4.19}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{\boldsymbol{p}}(t \,|\, Z) = 0 \quad \text{with} \quad \overline{\boldsymbol{p}}(t = 0 \,|\, Z) = \overline{\boldsymbol{p}}_0(Z) = 0,$$

respectively. Because of the uniqueness of paths under given initial conditions, it follows that $\overline{\boldsymbol{y}}(t \,|\, Z)$ remains the unique maximum of $s_0(t \,|\, Z)$ for all $t > 0$. Thus, the solution $\overline{\boldsymbol{y}}(t \,|\, Z)$ of eq. (4.19) gives the most probable values of the continuous species for a given discrete state Z, which is identical to the classical solution of biochemical reaction kinetics if we consider the discrete state Z to be constant.

However, it is important to realize that a solution of eq. (4.19), or more generally eqs. (4.18), is only valid on the $\mathcal{O}(\varepsilon)$ scale, but the continuous processes also depend on the discrete–stochastic dynamics evolving on the $\mathcal{O}(1)$ scale. Hence, we wish to derive evolution equations for the continuous species that live on scales of order $\mathcal{O}(1)$. In *indirect* multi-scale methods, for instance, eq. (4.19) is usually solved up to the predicted time of a next stochastic reaction event where the system is then updated accordingly and propagation of the characteristics continued for the corresponding new initial conditions, see Chapter 3 and Subsection 4.1.7 (see also [190] for an analysis on this type of multi-scale numerical methods).

4.1.5 Laplace's Integral Approximation of the Conditional PDF

In the previous subsection, we derived the evolution equation (4.19) for the most probable value of $\boldsymbol{y}^\varepsilon$ for each discrete state Z and showed that this maximum will remain unique at $\overline{\boldsymbol{y}}(t \,|\, Z)$ for all times $t > 0$. As demonstrated in Subsection 2.4.4, because of the special form

of the WKB-ansatz (4.12) and the above results, we can use Laplace's method to compute expectations of the process $\boldsymbol{y}^\varepsilon$, e.g.,

$$\begin{aligned}\mathbb{E}\big[f(\boldsymbol{y}^\varepsilon)\,|\,Z\big] &= \int_0^\infty f(y)P_\varepsilon(y\,|\,Z;t)\,\mathrm{d}y \\ &= C_\varepsilon\int_0^\infty f(y)\exp\left\{\frac{1}{\varepsilon}s_0(y\,|\,Z;t)\right\}\big(U_0(y\,|\,Z;t)+\mathcal{O}(\varepsilon)\big)\,\mathrm{d}y \\ &= C_\varepsilon\exp\left\{\frac{1}{\varepsilon}s_0\big(\overline{\boldsymbol{y}}(t\,|\,Z)\,|\,Z;t\big)\right\}\sqrt{\frac{(2\pi\varepsilon)^{N^c}}{\big|\nabla^2 s_0\big(\overline{\boldsymbol{y}}(t\,|\,Z)\,|\,Z;t\big)\big|}} \\ &\qquad\times\Big(f\big(\overline{\boldsymbol{y}}(t\,|\,Z)\big)U_0\big(\overline{\boldsymbol{y}}(t\,|\,Z)\,|\,Z;t\big)+\mathcal{O}(\varepsilon)\Big), \quad (4.20)\end{aligned}$$

where we consider the generic case that the maximum of $s_0(y\,|\,Z;t)$ at $y=\overline{\boldsymbol{y}}(t\,|\,Z)$ is non-degenerate, such that $|\nabla^2 s_0(x=\overline{\boldsymbol{x}}(\,\cdot\,);\,\cdot\,)|\neq 0$. By definition, the conditional probability $P_\varepsilon(y\,|\,Z;t)$ *has to* integrate to one for *any* value of ε, i.e.,

$$\begin{aligned}1\equiv\int_0^\infty P_\varepsilon(y\,|\,Z;t)\,\mathrm{d}y &= C_\varepsilon\int_0^\infty\exp\left\{\frac{1}{\varepsilon}s_0(y\,|\,Z;t)\right\}\big(U_0(y\,|\,Z;t)+\mathcal{O}(\varepsilon)\big)\,\mathrm{d}y \\ &= \frac{\exp\big\{\frac{1}{\varepsilon}s_0\big(\overline{\boldsymbol{y}}(t\,|\,Z)\,|\,Z;t\big)\big\}}{\sqrt{\big|\nabla^2 s_0\big(\overline{\boldsymbol{y}}(t\,|\,Z)\,|\,Z;t\big)\big|}}\Big(U_0\big(\overline{\boldsymbol{y}}(t\,|\,Z)\,|\,Z;t\big)+\mathcal{O}(\varepsilon)\Big), \quad (4.21)\end{aligned}$$

where we substituted $C_\varepsilon=(2\pi\varepsilon)^{-N^c/2}$ in the last equality, see Subsection 2.4.4. It thus follows that $s_0(y=\overline{\boldsymbol{y}}(\,\cdot\,|\,Z)\,|\,Z;\,\cdot\,)$ and $U_0(y=\overline{\boldsymbol{y}}(\,\cdot\,|\,Z)\,|\,Z;\,\cdot\,)$ must be given as

$$s_0\big(\overline{\boldsymbol{y}}(t\,|\,Z)\,|\,Z;t\big)=0\quad\text{and}\quad U_0\big(\overline{\boldsymbol{y}}(t\,|\,Z)\,|\,Z;t\big)=\sqrt{\big|\nabla^2 s_0\big(\overline{\boldsymbol{y}}(t\,|\,Z)\,|\,Z;t\big)\big|}, \quad (4.22)$$

for all Z and all $t\geq 0$. At the same time, the $\mathcal{O}(\varepsilon)$ corrections in the last equality of eq. (4.21) must be identical to zero. This shows that the conditional expectation in eq. (4.20) can be approximated by

$$\mathbb{E}\big[f(\boldsymbol{y}^\varepsilon)\,|\,Z\big]=f\big(\overline{\boldsymbol{y}}(t\,|\,Z)\big)+\mathcal{O}(\varepsilon). \quad (4.23)$$

From the conservation relation (4.21), it further follows that the above equality holds without $\mathcal{O}(\varepsilon)$ if f is a constant function.

In the following, we are interested in the partial expectation of a function f of $\boldsymbol{y}^\varepsilon$ with respect to the PDF $P_\varepsilon(\,\cdot\,, Z; t)$, which can be approximated with the above results as

$$\begin{aligned}\mathbb{E}_Z\big[f(\boldsymbol{y}^\varepsilon)\big] &:= \int_0^\infty f(y) P_\varepsilon(y, Z; t)\,\mathrm{d}y = \int_0^\infty f(y) P_\varepsilon(y \,|\, Z; t)\,\mathrm{d}y \cdot P_\varepsilon(Z; t) \\ &= \mathbb{E}\big[f(\boldsymbol{y}^\varepsilon) \,|\, Z\big] \cdot \big(P_0(Z; t) + \mathcal{O}(\varepsilon)\big) \\ &= f\big(\overline{\boldsymbol{y}}(t \,|\, Z)\big) P_0(Z; t) + \mathcal{O}(\varepsilon). \end{aligned} \tag{4.24}$$

Similarly, by applying Laplace's method and the results for s_0 and U_0 as given in eqs. (4.22), we find that all expectations with respect to $P_\varepsilon(\,\cdot\,, Z; t)$ can be computed from $\overline{\boldsymbol{y}}(t \,|\, Z)$ and $P_0(Z; t)$.[5]

Approximation of the PDF of the Discrete Species

We use the previous results to derive the evolution equation of $P_0(Z; t)$. First, we note that according to eq. (4.4) the evolution of $P_\varepsilon(Z; t)$ is given as

$$\frac{\partial}{\partial t} P_\varepsilon(Z; t) = \sum_{\mu \in \mathcal{M}^d} \mathbb{E}\big[\alpha_\mu(\boldsymbol{y}^\varepsilon, Z - \zeta_\mu) \,|\, Z - \zeta_\mu\big] P_\varepsilon(Z - \zeta_\mu; t) - \mathbb{E}\big[\alpha_\mu(\boldsymbol{y}^\varepsilon, Z) \,|\, Z\big] P_\varepsilon(Z; t),$$

where we substituted α_μ for the discrete reactions as given by eq. (4.10b). This equation implies that the evolution of $P_\varepsilon(Z; t)$ is given by scales of order $\mathcal{O}(1)$. Therefore, the dynamics of $P_\varepsilon(Z; t)$ can be expressed as

$$\begin{aligned}\frac{\partial}{\partial t} P_\varepsilon(Z; t) &= \frac{\partial}{\partial t} P_0(Z; t) + \varepsilon \frac{\partial}{\partial t} P_1(Z; t) + \dots \\ &= \sum_{\mu \in \mathcal{M}^d} \mathbb{E}\big[\alpha_\mu(\boldsymbol{y}^\varepsilon, Z - \zeta_\mu) \,|\, Z - \zeta_\mu\big] P_0(Z - \zeta_\mu; t) - \mathbb{E}\big[\alpha_\mu(\boldsymbol{y}^\varepsilon, Z) \,|\, Z\big] P_0(Z; t) + \mathcal{O}(\varepsilon). \end{aligned} \tag{4.25}$$

Employing the Laplace approximation given in eq. (4.23) and comparing the terms of order $\mathcal{O}(1)$ on both sides of eq. (4.25) yields the time-evolution of $P_0(Z; t)$ as

$$\frac{\partial}{\partial t} P_0(Z; t) = \sum_{\mu \in \mathcal{M}^d} \alpha_\mu\big(\overline{\boldsymbol{y}}(t \,|\, Z - \zeta_\mu), Z - \zeta_\mu\big) P_0(Z - \zeta_\mu; t) - \alpha_\mu\big(\overline{\boldsymbol{y}}(t \,|\, Z), Z\big) P_0(Z; t). \tag{4.26}$$

[5] From this result we can infer that the assumption that s_0 has a *unique* maximum at $t = 0$ may be relaxed. If initially s_0 has many local maxima at $\overline{\boldsymbol{y}}_0^i$, $i = 1, 2, \dots$, these maxima will evolve independently in time according to eq. (4.19). The approximation of partial expectations of the form (4.24) will then be a superposition of terms of the form $f(\overline{\boldsymbol{y}}^i(t \,|\, Z))$, as long as the distance between these maxima is $\mathcal{O}(1)$ so that Laplace's method can be applied.

Approximation of the Partial Expectations of the Continuous Species

Next, we derive the evolution equation of the leading order approximation for the partial expectations of the continuous species on the $\mathcal{O}(1)$ scale. As shown above, the partial expectation of any function f of $\boldsymbol{y}^\varepsilon$ with respect to $P_\varepsilon(\,\cdot\,, Z; t)$ can be approximated by $\mathbb{E}_Z[f(\boldsymbol{y}^\varepsilon)] = f(\overline{\boldsymbol{y}}(t \,|\, Z))P_0(Z; t) + \mathcal{O}(\varepsilon)$. This approximation is consistent with the derived evolution of $P_0(Z; t)$ in eq. (4.26). More importantly, it also allows us to derive the evolution of the leading order approximation of $\mathbb{E}_Z[\boldsymbol{y}^\varepsilon]$, i.e.,

$$\mathbb{E}_Z^0[\boldsymbol{y}] \coloneqq \overline{\boldsymbol{y}}(t \,|\, Z)P_0(Z; t). \tag{4.27}$$

First note that according to eq. (4.6), the evolution of $\mathbb{E}_Z[\boldsymbol{y}^\varepsilon]$ is given as

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}_Z[\boldsymbol{y}^\varepsilon] = &\sum_{\mu\in\mathcal{M}^c} v_\mu \mathbb{E}_Z\big[\alpha_\mu(\boldsymbol{y}^\varepsilon, Z)\big] \\ &+ \sum_{\mu\in\mathcal{M}^d} \mathbb{E}_{Z-\zeta_\mu}\big[(\boldsymbol{y}^\varepsilon + \varepsilon v_\mu)\alpha_\mu(\boldsymbol{y}^\varepsilon, Z-\zeta_\mu)\big] - \mathbb{E}_Z\big[\boldsymbol{y}^\varepsilon \alpha_\mu(\boldsymbol{y}^\varepsilon, Z)\big]\end{aligned} \tag{4.28}$$

where we substituted α_μ for the continuous and discrete reactions as given by eqs. (4.10a) and (4.10b), respectively. Once again, we observe that terms of order $\mathcal{O}(\varepsilon^{-1})$ do not appear in the right hand side of eq. (4.28), and therefore the dynamics of $\mathbb{E}_Z[\boldsymbol{y}^\varepsilon]$ can be approximated by

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}_Z[\boldsymbol{y}^\varepsilon] &= \frac{\partial}{\partial t}\mathbb{E}_Z^0[\boldsymbol{y}] + \mathcal{O}(\varepsilon) \\ &= \sum_{\mu\in\mathcal{M}^c} v_\mu \mathbb{E}\big[\alpha_\mu(\boldsymbol{y}^\varepsilon, Z) \,|\, Z\big]P_0(Z; t) + \mathcal{O}(\varepsilon) \\ &\quad + \sum_{\mu\in\mathcal{M}^d} \mathbb{E}\big[\boldsymbol{y}^\varepsilon \alpha_\mu(\boldsymbol{y}^\varepsilon, Z-\zeta_\mu) \,|\, Z-\zeta_\mu\big]P_0(Z-\zeta_\mu; t) - \mathbb{E}\big[\boldsymbol{y}^\varepsilon \alpha_\mu(\boldsymbol{y}^\varepsilon, Z) \,|\, Z\big]P_0(Z; t).\end{aligned}$$

Applying the Laplace approximation given in eq. (4.23) and comparing the terms of order $\mathcal{O}(1)$ on both sides, we find

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}_Z^0[\boldsymbol{y}] = &\sum_{\mu\in\mathcal{M}^c} v_\mu \alpha_\mu\big(\overline{\boldsymbol{y}}(t \,|\, Z), Z\big)P_0(Z; t) \\ &+ \sum_{\mu\in\mathcal{M}^d} \alpha_\mu\big(\overline{\boldsymbol{y}}(t \,|\, Z-\zeta_\mu), Z-\zeta_\mu\big)\mathbb{E}_{Z-\zeta_\mu}^0[\overline{\boldsymbol{y}}] - \alpha_\mu\big(\overline{\boldsymbol{y}}(t \,|\, Z), Z\big)\mathbb{E}_Z^0[\overline{\boldsymbol{y}}].\end{aligned} \tag{4.29}$$

The above eqs. (4.26) and (4.29) describe the evolutions of $P_0(Z;t)$ and $\mathbb{E}_Z^0[\boldsymbol{y}]$. It is worth mentioning that although $\mathbb{E}_Z^0[\boldsymbol{y}]$ is determined explicitly by $\overline{\boldsymbol{y}}(t\,|\,Z)$ and $P_0(Z;t)$, it cannot be recovered by direct integration of eqs. (4.19) and (4.26), since the evolution of $\overline{\boldsymbol{y}}(t\,|\,Z)$ as given by eq. (4.19) is only valid on the order $\mathcal{O}(\varepsilon)$. The evolution equation (4.29) of $\mathbb{E}_Z^0[\boldsymbol{y}]$, on the other hand, is valid on the order $\mathcal{O}(1)$, which is the scale we are interested in. The corresponding conditional levels $\overline{\boldsymbol{y}}(t\,|\,Z)$ of interest (i.e., where $P_0(Z;t) > 0$) can then be recovered by relation (4.27), as we shall see in the next subsections.

4.1.6 Final Equations of the Hybrid CME–ODE Approach

Summarizing, we derived the following hybrid system approximating the coupled dynamics of the stochastic and deterministic processes. The time-evolution of the probability distribution $P_0\big(Z;t\big)$ of the discrete species is given by eq. (4.26), i.e.,

$$\frac{\partial}{\partial t}P_0(Z;t) = \sum_{\mu\in\mathcal{M}^d} a_\mu\big(\overline{\boldsymbol{Y}}(t\,|\,Z-\zeta_\mu), Z-\zeta_\mu\big)P_0(Z-\zeta_\mu;t) - a_\mu\big(\overline{\boldsymbol{Y}}(t\,|\,Z), Z\big)P_0(Z;t), \quad (4.30a)$$

where we re-substituted the functions α_μ by the original propensities a_μ for the discrete reactions using eq. (4.10b). By eq. (4.29), the partial expectations $\mathbb{E}_Z^0[\boldsymbol{Y}] = \varepsilon^{-1}\mathbb{E}_Z^0[\boldsymbol{y}]$ of the continuous species for a discrete state Z are determined by

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}_Z^0[\boldsymbol{Y}] = {} & \underbrace{\sum_{\mu\in\mathcal{M}^c} v_\mu a_\mu\big(\overline{\boldsymbol{Y}}(t\,|\,Z), Z\big)P_0(Z;t)}_{\text{impact of the continuous reactions}} \\ & \underbrace{+ \sum_{\mu\in\mathcal{M}^d} a_\mu\big(\overline{\boldsymbol{Y}}(t\,|\,Z-\zeta_\mu), Z-\zeta_\mu\big)\mathbb{E}_{Z-\zeta_\mu}^0[\boldsymbol{Y}] - a_\mu\big(\overline{\boldsymbol{Y}}(t\,|\,Z), Z\big)\mathbb{E}_Z^0[\boldsymbol{Y}]}_{\text{impact of the discrete reactions}}, \quad (4.30b)\end{aligned}$$

where we replaced all α_μ by the original propensities a_μ as given by eqs. (4.10).

To obtain the solution of the conditional levels $\overline{\boldsymbol{Y}}(t\,|\,Z)$ from the above equations, we use definition (4.27), i.e., divide $\mathbb{E}_Z^0[\boldsymbol{Y}] = \overline{\boldsymbol{Y}}(t\,|\,Z)P_0(Z;t)$ by $P_0(Z;t) > 0$. If $P_0(Z;t) = 0$, the conditional levels $\overline{\boldsymbol{Y}}(t\,|\,Z)$ are not defined but also not of any practical interest. We further note that $P_0(Z;t) = 0$ for a discrete state Z implies that also $\mathbb{E}_Z^0[\boldsymbol{Y}] = 0$, and thus all terms in the right hands sides of the above evolution equations (4.30) that are associated with Z are also equal to zero. Hence, we do not need to specify the levels $\overline{\boldsymbol{Y}}(t\,|\,Z)$ in our hybrid model if $P_0(Z;t) = 0$, but for convenience may set them equal to zero in such case.

As the size of the discrete state space is drastically reduced, a standard numerical integration of eqs. (4.30) becomes applicable even for more complex systems. During numerical integration of the hybrid system the propensities have to be evaluated at every integration step. This requires the explicit computation of the conditional value $\overline{\boldsymbol{Y}}_i(t \,|\, Z)$ of a continuous species S_i^c from $P_0(Z;t)$ and $\mathbb{E}_Z^0[\boldsymbol{Y}_i]$ whenever $P_0(Z;t) > 0$ and: (a) S_i^c is a reactant of a discrete reaction, or (b) S_i^c is a reactant of a second or higher-order continuous reaction. To avoid numerical instabilities for $P_0(Z;t) \approx 0$, we introduce a threshold value $\delta \ll 1$: If $P_0(Z;t) \leq \delta$, we set $P_0(Z;t) = 0$ and $\mathbb{E}_Z^0[\boldsymbol{Y}] = 0$ in the evaluation of eqs. (4.30), such that the actual dynamics of the hybrid system will be constrained to those discrete states Z where $P_0\,(Z;t) > \delta$ during integration. Basically, this is the same strategy as proposed by Henzinger et al. [66], but carried over to our closed hybrid formulation. For appropriate choices of δ, the additional error made with this criterion can be expected to be negligible, cf. [66, 191]. A reasonable choice for δ would be a value not greater than the allowed absolute error used in the numerical integration of the hybrid system.

4.1.7 Related Indirect Hybrid Approach

Let us point out the equations of the processes $\boldsymbol{Y}$ and $\boldsymbol{Z}$ associated with the continuous and the discrete species, respectively, that underly our hybrid model (4.30). As shown in Subsection 2.2.2, the full stochastic process $\boldsymbol{X}(t) = \big(\boldsymbol{Y}(t), \boldsymbol{Z}(t)\big)^{\mathrm{T}}$ can be represented by

$$\boldsymbol{X}(t) = \boldsymbol{X}(0) + \sum_{\mu=1}^{M} \xi_\mu \boldsymbol{\mathcal{P}}_\mu\!\left(\int_0^t a_\mu\big(\boldsymbol{X}(s)\big)\,\mathrm{d}s\right),$$

where $\boldsymbol{\mathcal{P}}_\mu$, $\mu = 1, \ldots, M$, are independent, unit Poisson processes. Based on the system partitioning of our hybrid approach (see Subsection 4.1.1), we can write the above equation separately as

$$\boldsymbol{Y}(t) = \boldsymbol{Y}(0) + \sum_{\mu\in\mathcal{M}^c} \upsilon_\mu \boldsymbol{\mathcal{P}}_\mu\!\left(\int_0^t a_\mu\big(\boldsymbol{Y}(s), \boldsymbol{Z}(s)\big)\,\mathrm{d}s\right) + \sum_{\mu\in\mathcal{M}^d} \zeta_\mu \boldsymbol{\mathcal{P}}_\mu\!\left(\int_0^t a_\mu\big(\boldsymbol{Y}(s), \boldsymbol{Z}(s)\big)\,\mathrm{d}s\right) \tag{4.31a}$$

and

$$\boldsymbol{Z}(t) = \boldsymbol{Z}(0) + \sum_{\mu\in\mathcal{M}^d} \zeta_\mu \boldsymbol{\mathcal{P}}_\mu\!\left(\int_0^t a_\mu\big(\boldsymbol{Y}(s), \boldsymbol{Z}(s)\big)\,\mathrm{d}s\right), \tag{4.31b}$$

where we used that $\xi_\mu = (\upsilon_\mu, \zeta_\mu)^{\mathrm{T}}$ with $\zeta_\mu = 0$ for all $\mu \in \mathcal{M}^c$, see eq. (4.1a). We notice that the processes $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are not independent of each other, as the propensities in eqs. (4.31a) and

(4.31b) will in general depend on both, $\boldsymbol{Y}$ and $\boldsymbol{Z}$, respectively. According to our assumptions on the continuous species and reactions (see Subsection 4.1.3), eqs. (4.31) scale like

$$\boldsymbol{y}(t) = \boldsymbol{y}(0) + \varepsilon \sum_{\mu \in \mathcal{M}^c} v_\mu \boldsymbol{\mathcal{P}}_\mu\left(\frac{1}{\varepsilon}\int_0^t \alpha_\mu\big(\boldsymbol{y}(s), \boldsymbol{Z}(s)\big)\,\mathrm{d}s\right) + \varepsilon \sum_{\mu \in \mathcal{M}^d} \zeta_\mu \boldsymbol{\mathcal{P}}_\mu\left(\int_0^t \alpha_\mu\big(\boldsymbol{y}(s), \boldsymbol{Z}(s)\big)\,\mathrm{d}s\right) \tag{4.32a}$$

and

$$\boldsymbol{Z}(t) = \boldsymbol{Z}(0) + \sum_{\mu \in \mathcal{M}^d} \zeta_\mu \boldsymbol{\mathcal{P}}_\mu\left(\int_0^t \alpha_\mu\big(\boldsymbol{y}(s), \boldsymbol{Z}(s)\big)\,\mathrm{d}s\right), \tag{4.32b}$$

where we substituted $\boldsymbol{y} = \varepsilon \boldsymbol{Y}$ and replaced all propensities a_μ by the scaled propensities α_μ as given by eqs. (4.10). The above set of stochastic equations (4.32) is equivalent to the ε-scaled CME (4.11), which was the starting point of our multi-scale analysis.

Next, we study the limit behavior of $\boldsymbol{y}$ and $\boldsymbol{Z}$ as $\varepsilon \to 0$. Obviously, the stochastic process $\boldsymbol{Z}$ associated with the discrete species remains unchanged as $\varepsilon \to 0$, see eq. (4.32b), which is in complete agreement with our desired partial limit. For the process $\boldsymbol{y}$ associated with the continuous species, we can apply the SLLN for Poisson processes (Theorem 2.4.2) as in the classical thermodynamic limit (see Subsection 2.4.2) and find that

$$\lim_{\varepsilon \to 0} \left\| \sum_{\mu \in \mathcal{M}^c} \varepsilon\, v_\mu \boldsymbol{\mathcal{P}}_\mu\big(\Lambda_\mu(t)/\varepsilon\big) - v_\mu \Lambda_\mu(t) \right\| = 0,$$

almost surely, where $\Lambda_\mu(t) \coloneqq \int_0^t \alpha_\mu(\boldsymbol{y}(s), \boldsymbol{Z}(s))\,\mathrm{d}s$. The contribution of the discrete reactions in eq. (4.32a), however, vanishes in the limit as $\varepsilon \to 0$, i.e.,

$$\lim_{\varepsilon \to 0} \left\| \sum_{\mu \in \mathcal{M}^d} \varepsilon\, \zeta_\mu \boldsymbol{\mathcal{P}}_\mu\big(\Lambda_\mu(t)\big) \right\| = 0,$$

since Poisson processes have a finite activity (i.e., almost all paths have only a finite number of jumps along finite time intervals). Hence, we have

$$\boldsymbol{y}(t) \xrightarrow{\text{a.s.}} \boldsymbol{y}(0) + \sum_{\mu \in \mathcal{M}^c} v_\mu \int_0^t \alpha_\mu\big(\boldsymbol{y}(s), \boldsymbol{Z}(s)\big)\,\mathrm{d}s,$$

as $\varepsilon \to 0$. Putting things together, the processes $\boldsymbol{Y}$ and $\boldsymbol{Z}$ underlying our hybrid model (4.30) are thus given by

$$\boldsymbol{Y}(t) = \boldsymbol{Y}(0) + \sum_{\mu \in \mathcal{M}^c} v_\mu \int_0^t a_\mu\big(\boldsymbol{Y}(s), \boldsymbol{Z}(s)\big)\,\mathrm{d}s \tag{4.33a}$$

and

$$\boldsymbol{Z}(t) = \boldsymbol{Z}(0) + \sum_{\mu \in \mathcal{M}^d} \zeta_\mu \boldsymbol{\mathcal{P}}_\mu\left(\int_0^t a_\mu\big(\boldsymbol{Y}(s), \boldsymbol{Z}(s)\big)\,\mathrm{d}s\right), \tag{4.33b}$$

where we re-substituted $\boldsymbol{Y} = \varepsilon^{-1}\boldsymbol{y}$ and the original propensities a_μ as given by eqs. (4.10).

Eqs. (4.33a) and (4.33b) could be used in a corresponding *indirect* hybrid method to compute realizations of $\boldsymbol{Y}$ and $\boldsymbol{Z}$, see Chapter 3. Suppose the system is initially in the states $\boldsymbol{Y}(\tau_0) = Y_0$ and $\boldsymbol{Z}(\tau_0) = Z_0$ at time $\tau_0 = 0$. Then, eq. (4.33a) or, equivalently, the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{Y}(t) = \sum_{\mu\in\mathcal{M}^c} \upsilon_\mu a_\mu\big(\boldsymbol{Y}(t), \boldsymbol{Z}(t)\big) \tag{4.34}$$

will be solved in a hybrid simulation for $\boldsymbol{Y}(\tau_0) = Y_0$ from time $t = \tau_0$ up to the predicted firing time $t = \tau_1$ of the first discrete reaction R_j with $j \in \mathcal{M}^d$. The discrete states will then be updated according to R_j, i.e., $Z_1 \leftarrow Z_0 + \zeta_j$, and integration of eq. (4.34) will be continued from time $t = \tau_1$ for $\boldsymbol{Z}(\tau_1) = Z_1$ until the predicted firing time $t = \tau_2$ of the next discrete reaction. A realization of the coupled processes $\big\{(\boldsymbol{Y}(t), \boldsymbol{Z}(t)), t \geq 0\big\}$ will hence be of the form $\bigcup_{i=0,1,\ldots}\big\{(\boldsymbol{Y}(t), Z_i), t \in [\tau_i; \tau_{i+1})\big\}$, where $\boldsymbol{Y}(t)$ is the solution of the ODE (4.34) with $\boldsymbol{Z}(t) = Z_i$ for $t \in [\tau_i; \tau_{i+1})$. In an individual hybrid realization the discrete reactions thus affect the dynamics of $\boldsymbol{Y}$ by changing the values of the propensities in eq. (4.34). Most importantly, however, these 'switches' in the dynamics of $\boldsymbol{Y}$ also have an impact on its ensemble average, which appears in our direct hybrid model as *additional* coupling terms in eq. (4.30b), but may be unexpected when only looking at eq. (4.34).

4.1.8 Coarse Graining of the Continuous Processes

In most systems, reactions are typically of zero, first or second order and the propensity functions will thus depend on the level of a few species only, see Table 2.3. If the propensities of the continuous reactions are constant on a subset $\mathcal{Z}_k$ of discrete states, then we may associate a single ODE with the entire subset $\mathcal{Z}_k$ rather than an ODE for each element of the subset. We therefore seek subsets $\mathcal{Z}_k$ of discrete states such that for all $\mu \in \mathcal{M}^c$:

$$a_\mu(\,\cdot\,, Z) = const \quad \text{for all } Z \in \mathcal{Z}_k. \tag{4.35}$$

If criterion (4.35) holds, it follows that the evolution equations of the eikonal functions $s_0(\,\cdot \,|\, Z; t)$ are identical for every $Z \in \mathcal{Z}_k$. Given equal initial conditions $Y = \overline{\boldsymbol{Y}}_0(Z)$ for all $Z \in \mathcal{Z}_k$, they hence propagate in time along the same characteristic $\overline{\boldsymbol{Y}}(t \,|\, \mathcal{Z}_k)$. Consequently, in this case we can assign the same partial expectation $\mathbb{E}_{\mathcal{Z}_k}[\boldsymbol{Y}]$ to each subset $\mathcal{Z}_k$. The time-

evolution of its leading order approximation $\mathbb{E}^0_{\mathcal{Z}_k}[\boldsymbol{Y}] := \overline{\boldsymbol{Y}}(t \,|\, \mathcal{Z}_k) \sum_{Z \in \mathcal{Z}_k} P_0(Z;t)$ is then given by summation of eq. (4.30b) over all discrete states $Z \in \mathcal{Z}_k$, i.e.,

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}^0_{\mathcal{Z}_k}[\boldsymbol{Y}] = &\sum_{\mu \in \mathcal{M}^c} v_\mu a_\mu\big(\overline{\boldsymbol{Y}}(t \,|\, \mathcal{Z}_k), \mathcal{Z}_k\big) \sum_{Z \in \mathcal{Z}_k} P_0(Z;t) \\ &+ \sum_{\mu \in \mathcal{M}^d} \sum_{Z \in \mathcal{Z}_k} \Big[a_\mu\big(\overline{\boldsymbol{Y}}(t \,|\, Z - \zeta_\mu), Z - \zeta_\mu\big)\mathbb{E}^0_{Z-\zeta_\mu}[\boldsymbol{Y}] - a_\mu\big(\overline{\boldsymbol{Y}}(t \,|\, Z), Z\big)\mathbb{E}^0_Z[\boldsymbol{Y}] \Big],\end{aligned} \tag{4.36}$$

where $a_\mu(\overline{\boldsymbol{Y}}(t \,|\, \mathcal{Z}_k), \mathcal{Z}_k)$ in the first sum on the right hand side denotes the propensity of a continuous reaction on $\mathcal{Z}_k$ evaluated for $\overline{\boldsymbol{Y}}(t \,|\, \mathcal{Z}_k)$. Again, the second summand in the right hand side of eq. (4.36) describes the impact of the discrete reactions on the partial expectation of the continuous species. As these are coarse grained, all terms related to an exchange on the same subset $\mathcal{Z}_k$ cancel out, and only those terms related to an in- or outflow of probability to be in $\mathcal{Z}_k$ remain in the second summand of eq. (4.36).

Any conditional value $\overline{\boldsymbol{Y}}(t \,|\, \mathcal{Z}_k)$ necessary for the evaluation of eq. (4.36) can be computed from $\mathbb{E}^0_{\mathcal{Z}_k}[\boldsymbol{Y}] := \overline{\boldsymbol{Y}}(t \,|\, \mathcal{Z}_k) \sum_{Z \in \mathcal{Z}_k} P_0(Z;t)$ and the PDFs $P_0(Z;t)$, for all $Z \in \mathcal{Z}_k$, as described in the previous subsection. Further, we have $\overline{\boldsymbol{Y}}(t \,|\, Z) = \overline{\boldsymbol{Y}}(t \,|\, \mathcal{Z}_k)$ for all $Z \in \mathcal{Z}_k$, which can also be used to compute the leading order approximation $\mathbb{E}^0_Z[\boldsymbol{Y}] = \overline{\boldsymbol{Y}}(t \,|\, Z) P_0(Z;t)$ of the partial expectations at a specific state Z, necessary for evaluating the second summand in the right hand side of eq. (4.36). The evolution equations for the PDF $P_0(Z;t)$ of the discrete species are not affected by the suggested coarse graining.

4.1.9 Algorithmic Flow

The main steps of our hybrid approach can be summarized as follows:

Step 1 (Partition the System): Define a partition of the species into continuous species S_i^c, with $i \in \mathcal{N}^c \neq \varnothing$, and discrete species S_i^d, with $i \in \mathcal{N}^d = \{1, \ldots, N\} \setminus \mathcal{N}^c$. Such partitioning can be based on the expected levels and conservation properties of the species, different time-scales of the reactions or other prior knowledge on the system dynamics. Partition the reactions according to conditions (4.1a) and (4.1b), i.e., assign every reaction that changes the level of any discrete species to the subset $\mathcal{M}^d$ of discrete processes, and each reaction that solely affects the level of the continuous species to the subset $\mathcal{M}^c$ of continuous processes. The subset $\mathcal{M}^c$ is assumed to be non-empty in our hybrid approach, i.e., a continuous–deterministic

approximation should not be applied if all reactions act on the discrete species. Conversely, if the subset $\mathcal{M}^d$ is empty, a purely deterministic simulation of the system can be performed.

Step 2 (Formulate Hybrid CME–ODE Equations): If the dynamics of the discrete species is *a priori* known to be restricted to a specific subspace, formulate the corresponding system of (linear) ODEs for the time-evolution of the PDF as described by eq. (4.30a). Otherwise, choose a reasonable initial subset of discrete states, the approximation error can then be bounded by the FSP algorithm, cf. [25] and Subsection 2.2.5. If coarse graining can be applied to the continuous processes (see Subsection 4.1.8), assign a system of ODEs to propagate the partial expectations of the continuous species on every discrete subset $\mathcal{Z}_k$ as given by eq. (4.36). Otherwise, assign an ODE-system as given by eq. (4.30b) to every discrete state Z.

Step 3 (Compute Numerical Solution): Numerical integration of the final hybrid model requires appropriate methods for the efficient solution of differential algebraic equations (DAEs). Since a variety of such methods are available and choosing the optimal one for the case at hand is a problem on its own, we refrain from addressing this topic in depth. Instead, we apply a straight-forward approach that exploits the threshold value δ introduced above: Choose an appropriate δ for the evaluation of the continuous levels from $\mathbb{E}^0_Z[\boldsymbol{Y}]$ or $\mathbb{E}^0_{\mathcal{Z}_k}[\boldsymbol{Y}]$, respectively. Start numerical integration of the system. If necessary, monitor the loss in probability mass and dynamically update the support of the system by expanding the discrete state space based on the FSP algorithm, i.e., update the ODE-system for the PDF of the discrete species and the associated expectations of the continuous species. For state space expansion, strategies elaborated in [25, 29, 104, 109] can be used.

To highlight the benefits of our approach, we compare numerical costs in terms of number of equations that have to be integrated. Let N^d and N^c denote the number of different discrete and continuous species, respectively. Assume that the number of molecules of each discrete species is bounded by the same maximal value m and the level of every continuous species can maximally reach h, with $h \gg m$ by assumption. Hence, the state space of the full system would include $(m+1)^{N^d} \cdot (h+1)^{N^c}$ possible states that have to be considered in the spatial discretization of the CME. In contrast, in our approach the system dynamics would be described by $(m+1)^{N^d} \cdot (1+N^c)$ equations only. These are given by the $(m+1)^{N^d}$ states necessary for the support of the PDF of the discrete species, and the ODEs for the partial expectations of the N^c continuous species associated with each every discrete state. The reduction would be even higher if the continuous processes can be further coarse grained to subsets $\mathcal{Z}_k$ of discrete states. In that case we would only have $(m+1)^{N^d} + K \cdot N^c$ equations, for

instance, where $K \ll (m+1)^{N^d}$ denotes the number of discrete subsets $\mathcal{Z}_k$. As the numerical costs of our hybrid approach basically scale with the number of discrete states, we expect it to be especially efficient for systems that include a few species which have to be modeled discretely, such as gene regulatory or signaling networks.

4.2 Numerical Studies

The SSA and an explicit Runge–Kutta method of order 4 with error and step size control were implemented in C++ . All numerical experiments were performed on an Intel® Core™2 Duo processor with 2 GHz and 2 GB RAM. In each example, the numerical solution of the proposed hybrid approach was compared to the corresponding predictions obtained from ten thousand SSA simulations of the full CME. Numerical integration of the hybrid equations was performed with an absolute tolerance of 10^{-6} and a relative tolerance of 10^{-3}. The above introduced δ-threshold was set to 10^{-6}, in line with the absolute tolerance. For state space truncation, appropriate maximal levels of each discrete species were chosen initially, accounting for higher values than observed in all corresponding SSA simulations. No additional boundary conditions were applied to the truncated systems, such that the resulting truncation error could be monitored by loss of probability mass. At final time points of all experiments, the truncation error was found to be much lower than the tolerated absolute error.

4.2.1 A Simple Switch-Model

Metastability is an important property of biological systems, necessitating in general a stochastic modeling approach for its *in silico* analysis. As a first test example for our hybrid approach, we designed a simple bistable model:

$$\begin{aligned} R_1\colon \quad & S_1 \xrightarrow{k_1} S_2, \qquad & R_2\colon \quad & S_2 \xrightarrow{k_2} S_1, \\ R_3\colon \quad & C + S_1 \xrightarrow{k_3} S_1, \qquad & R_4\colon \quad & S_2 \xrightarrow{k_4} C + S_2, \end{aligned} \tag{4.37}$$

The concept behind this network is that S_1 and S_2 represent two metastable states of a more complex system, derived by some coarse graining of underlying subprocesses. Even though S_1 and S_2 are actually no molecular species but account for different internal states of the system, we can treat these states as discrete species with a possible value of either zero or one within

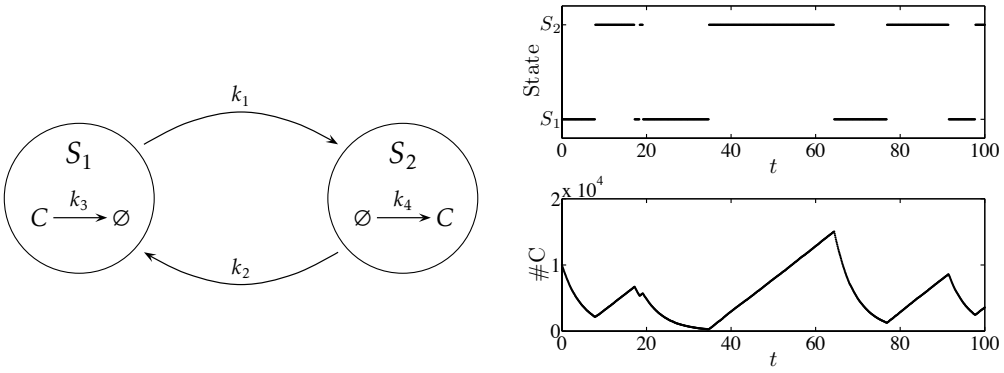

Figure 4.1: *Illustration of the simple bistable system* (4.37)*. Left: The system can switch between two (discrete) states S_1 and S_2. The dynamics of the (continuous) species C depends on the actual state of the system: C is degraded in S_1 and produced in S_2, respectively. Right: Typical results of a single SSA-realization of system* (4.37) *for the parameter values listed in Table 4.1.*

the framework of our approach. Transitions between S_1 and S_2 are modeled by reactions R_1 and R_2 that are treated as discrete processes in the following. Depending on the actual state of the system, the continuously treated species C gets either degraded or produced through reactions R_3 and R_4, respectively, as illustrated in Figure 4.1 (left panel). We included this state-dependence of R_3 and R_4 in their stoichiometries, see eqs. (4.37). The discrete states S_1 and S_2 are not changed by firings of the channels R_3 or R_4, and hence, these reactions are treated as continuous deterministic processes (see also Figure 4.1, right panel, for a typical SSA-realization of the system).

The full set of hybrid equations of network (4.37) is given by

$$\begin{aligned}
\frac{\partial}{\partial t}P_0(S_1;t) &= -k_1 P_0(S_1;t) + k_2 P_0(S_2;t) = -\frac{\partial}{\partial t}P_0(S_2;t),\\
\frac{\partial}{\partial t}\mathbb{E}^0_{S_1}[\boldsymbol{Y}] &= -k_1\mathbb{E}^0_{S_1}[\boldsymbol{Y}] + k_2\mathbb{E}^0_{S_2}[\boldsymbol{Y}] - k_3\mathbb{E}^0_{S_1}[\boldsymbol{Y}],\\
\frac{\partial}{\partial t}\mathbb{E}^0_{S_2}[\boldsymbol{Y}] &= +k_1\mathbb{E}^0_{S_1}[\boldsymbol{Y}] - k_2\mathbb{E}^0_{S_2}[\boldsymbol{Y}] + k_4 P_0(S_2;t),
\end{aligned} \tag{4.38}$$

where $P_0(S_1;t)$ and $P_0(S_2;t)$ denote the probabilities that the system is at time t either in state S_1 or state S_2, respectively. The approximated partial expectations of species C with respect to

k_1	k_2	k_3	k_4
0.06	0.04	0.2	500.0

Table 4.1: *Parameter values used for the bistable system* (4.37) *($\Omega = 1$). Initially, the system is with probability one in state S_1 with ten thousand entities of species C.*

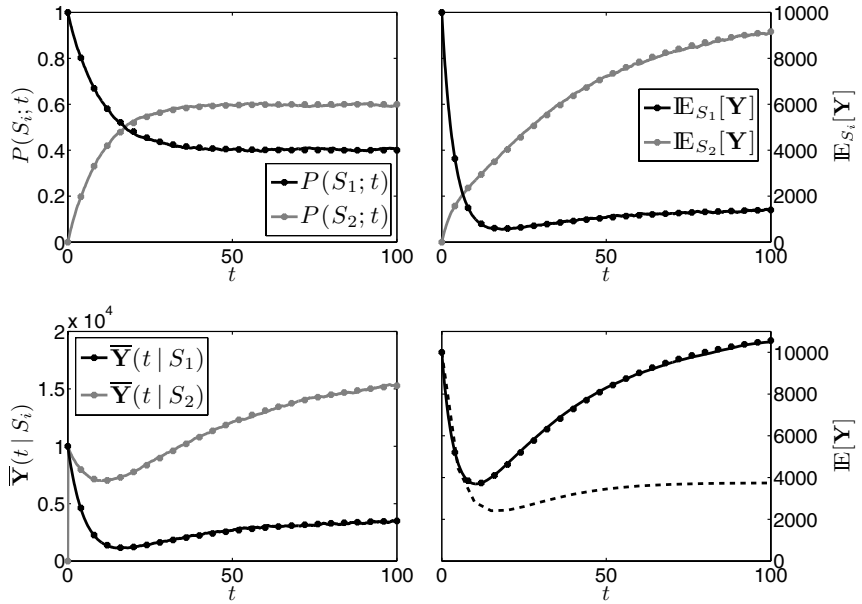

Figure 4.2: *Time evolution of system* (4.37) *as predicted by ten thousand SSA runs (solid lines) and the numerical solution (marked by dots) of the corresponding hybrid equations* (4.38) *for the parameter values listed in Table 4.1. The level of C as predicted by a purely deterministic simulation of system* (4.37) *is shown in the lower right panel (dashed line).*

the states S_1 and S_2 are denoted by $\mathbb{E}^0_{S_1}[\boldsymbol{Y}]$ and $\mathbb{E}^0_{S_2}[\boldsymbol{Y}]$, respectively. It should be noticed that the corresponding conditional levels $\overline{\boldsymbol{Y}}(t \mid S_1)$ and $\overline{\boldsymbol{Y}}(t \mid S_2)$ of C do not have to be computed for the evaluation of the hybrid equations (4.38). This is always the case if the propensities of all continuos reactions are constant or linear functions of Y and the propensities of all discrete reactions are constant with respect to Y.

In Figures 4.2 and 4.3, the numerical solution of eqs. (4.38) is compared to predictions obtained by ten thousand SSA simulations. In the upper panels of Figure 4.2, the time-evolution of the probabilities $P_0(S_1;t)$ and $P_0(S_2;t)$ (upper left), as well as the evolution of the partial expectations $\mathbb{E}^0_{S_1}[\boldsymbol{Y}]$ and $\mathbb{E}^0_{S_2}[\boldsymbol{Y}]$ (upper right) are shown. In the lower panels of Figure 4.2 the corresponding conditional levels $\overline{\boldsymbol{Y}}(t \mid S_1)$ and $\overline{\boldsymbol{Y}}(t \mid S_2)$ (lower left) and the approximation of the total expectation $\mathbb{E}^0[\boldsymbol{Y}] = \mathbb{E}^0_{S_1}[\boldsymbol{Y}] + \mathbb{E}^0_{S_2}[\boldsymbol{Y}]$ of C (lower right) are depicted. Additionally, the level of C as predicted by a purely deterministic simulation of system (4.37) is shown in the lower right panel of Figure 4.2 (dashed line). The relative error between the solution of the hybrid equations (4.38) and the results obtained by the SSA simulations is plotted against time in Figure 4.3.

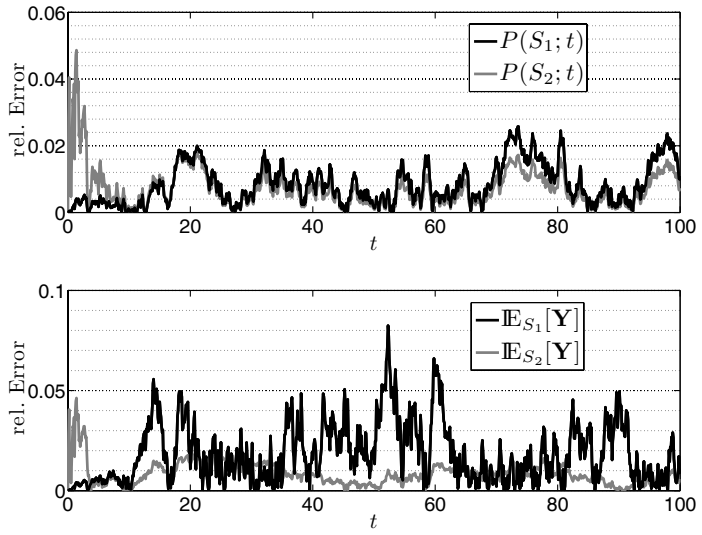

Figure 4.3: *Relative error of the numerical solution of the hybrid equations* (4.38) *with respect to the predictions obtained by ten thousand SSA runs of system* (4.37)*, as shown in Figure 4.2.*

Our hybrid solution is in excellent agreement with the SSA results, whereas a purely deterministic simulation fails to correctly predict the dynamics of the continuous species C. As discussed in Subsection 2.4.3, independence of all random variables is automatically assumed in the deterministic approach, such that the propensity of the continuous, bimolecular reaction R_3 in network (4.37), for instance, gets approximated as $a_3(\mathbb{E}[\boldsymbol{S}_1], \mathbb{E}[\boldsymbol{Y}])$. Obviously, this approximation is inappropriate, since the expected level of the discrete species S_1 is always small ($\mathbb{E}[\boldsymbol{S}_1] \leq 1$). In our hybrid approach, we account for the dependence of the continuous deterministic processes on the discrete stochastic processes—an averaging is only partially applied with respect to the continuous species that are expected to be present in large copy numbers. Furthermore, for the model system (4.37), our hybrid equations (4.38) correspond to the exact evolution equations (4.4) and (4.6) derived in Subsection 4.1.2. This is guaranteed by two properties of the network: Decoupling between the continuous species C and the discrete reactions R_1 and R_2, and linearity of the continuous reactions R_3 and R_4 with respect to the continuous species C. For that reason, the discrepancies to the SSA results, as shown in Figure 4.3, are mainly associated with the sampling error of the MC-method.

4.2.2 Viral Infection Kinetics

As a second example, we consider the infection model of a non-lytic virus proposed by Srivastava et al. [17]. We also use this model to illustrate the suggested coarse graining of the

continuous processes. The model includes three viral components: Viral nucleic acids, classified as either genomic (*gen*) or template (*tem*), and viral structural protein (*struct*), governed by six elementary reactions:

$$\begin{aligned} R_1&: \quad gen \xrightarrow{k_1} tem, & R_2&: \quad tem \xrightarrow{k_2} \varnothing, \\ R_3&: \quad tem \xrightarrow{k_3} tem + gen, & R_4&: \quad gen + struct \xrightarrow{k_4} \varnothing, \\ R_5&: \quad tem \xrightarrow{k_5} tem + struct, & R_6&: \quad struct \xrightarrow{k_6} \varnothing. \end{aligned} \tag{4.39}$$

The infection of the host-cell is initiated with a single molecule of *tem*, where the template *tem* denotes the 'active' form of nucleic acids that is involved in the catalytic synthesis of the viral components *gen* and *struct* (reactions R_3 and R_5, respectively). The *gen* component refers to nucleic acids that transport the viral genetic information, e.g., DNA or RNA, which gets either processed into the active form *tem* (reaction R_1), or used together with structural proteins *struct* to build a new viral cell that gets released from the host (reaction R_4). The number of *tem* and *struct* molecules is further regulated by degradation (reactions R_2 and R_6, respectively). The considered values of all rate constants are given in Table 4.2.

Even though the network (4.39) has a relative simple structure, it is capable of resembling realistic scenarios of a viral infection by rendering two steady states; the first representing a successful infection with approximately 20 *gen*, 200 *tem* and 10,000 *struct* molecules, and the second representing a successful rejection with no molecule of any viral component left. However, a purely deterministic simulation of system (4.39) will never account for these different scenarios, as the first steady state is deterministically stable, whereas the second, absorbing state is deterministically unstable, cf. [17]. To study the system behavior correctly, the discrete–stochastic formulation has to be used instead. Unfortunately, the system dynamics lives on a much too large state space to solve the corresponding CME directly (e.g., there are more than 42 million states that 'directly' connect the two steady states), and usually SSA realizations of the system are considered instead.

Table 4.2: *Parameter values of the viral infection kinetics model* (4.39) *in unit per day* ($\Omega = 1$)*, cf. [17]. Initially, the system is with probability one in the state with one tem, and zero gen and struct molecules.*

k_1	k_2	k_3	k_4	k_5	k_6
0.025	0.25	1.0	7.5×10^{-6}	1000	2.0

As the level of *struct* is at least two orders of magnitude larger in a successful infection than the levels of *tem* and *gen* (except for some transient phase), we treat *struct* as a continuous species; and regard *tem* and *gen* as discrete species. Accordingly, reactions R_1–R_4 are treated as discrete processes; reactions R_5 and R_6 are approximated as continuous processes. Although reactions R_3 and R_5 include *tem* as a reactant species, they are handled differently in our approach: R_3 acts on the discrete species *gen*, whereas R_5 acts on the continuous species *struct* (*tem* is not affected by both reactions). Therefore, R_3 is treated as a discrete process, while R_5 is treated as a continuous process.

Since the propensities of both continuous reactions, R_5 and R_6, have constant values for a fixed number of *tem* molecules, we further coarse grain the number of continuous processes, i.e., we associate a single continuous process to the subset of discrete states where the level of *tem* is the same. We use each of these processes to obtain a deterministic approximation of the expected level of *struct* conditioned on the number of *tem* molecules. This further reduces the number of hybrid equations significantly.

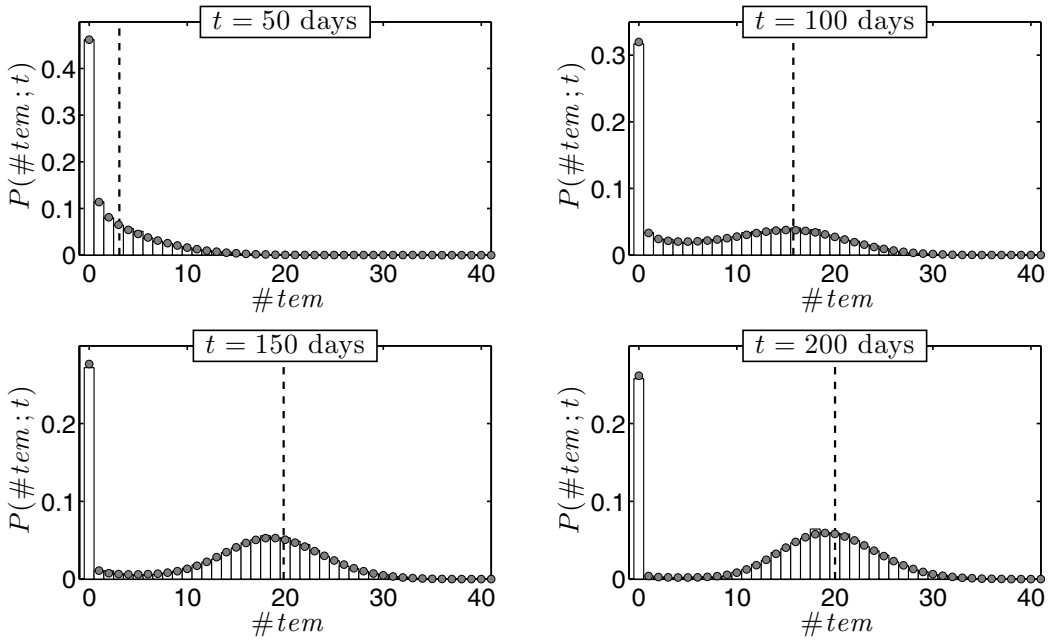

Figure 4.4: *Marginal PDF of tem after* $t = 50, 100, 150$ *and* 200 *days in the viral infection kinetics model* (4.39) *for parameter values listed in Table 4.2. The approximations obtained by ten thousand SSA runs are indicated by bars, the numerical solution of the suggested hybrid model is marked by dots, purely deterministic predictions are highlighted by dashed lines.*

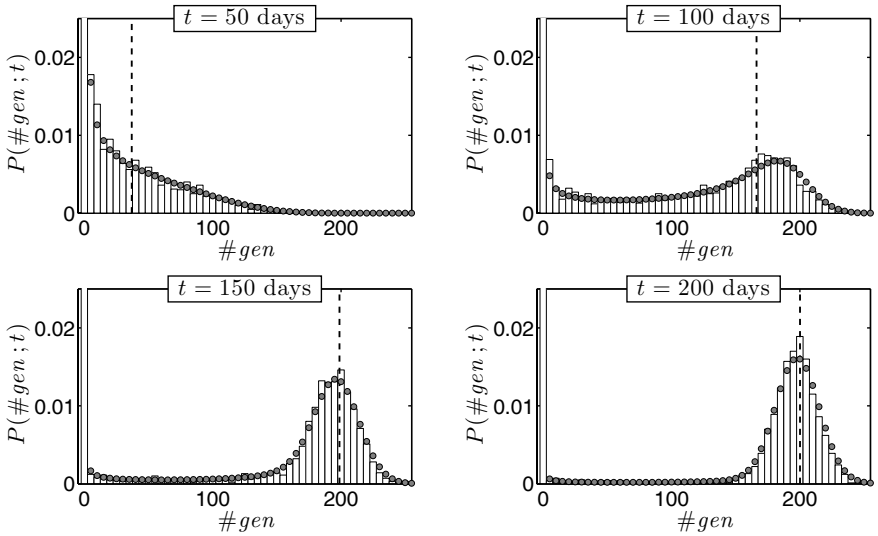

Figure 4.5: *Marginal PDF of gen after* $t = 50, 100, 150$ *and* 200 *days in the viral infection kinetics model* (4.39) *for parameter values listed in Table 4.2 (visualized with a bin-width of five). The approximations obtained by ten thousand SSA runs are indicated by bars, the numerical solution of the suggested hybrid model is marked by dots, purely deterministic predictions are highlighted by dashed lines.*

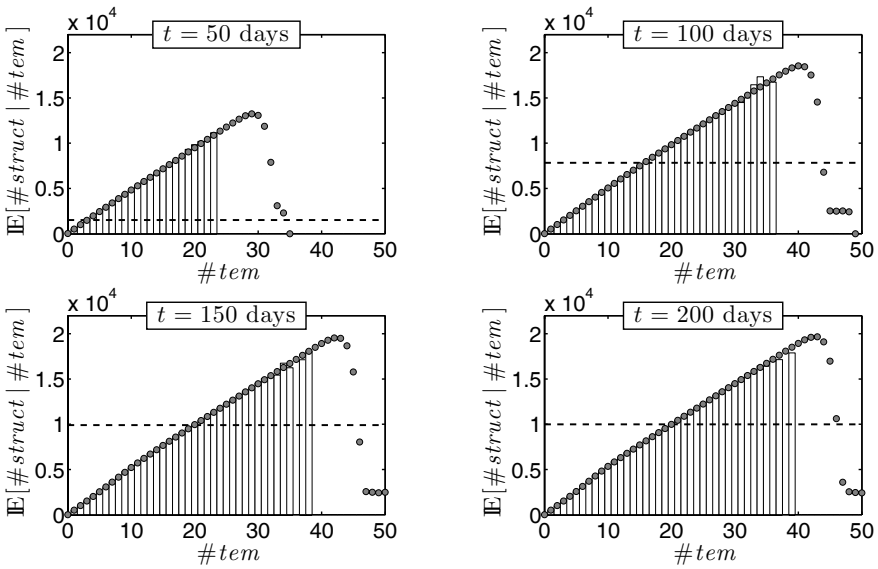

Figure 4.6: *Expected values of struct conditioned on the number of tem molecules after* $t = 50, 100, 150$ *and* 200 *days in the viral infection kinetics model* (4.39) *for parameter values listed in Table 4.2. The approximations obtained by ten thousand SSA runs are indicated by bars, the numerical solution of the suggested hybrid model is marked by dots, purely deterministic predictions are highlighted by dashed lines.*

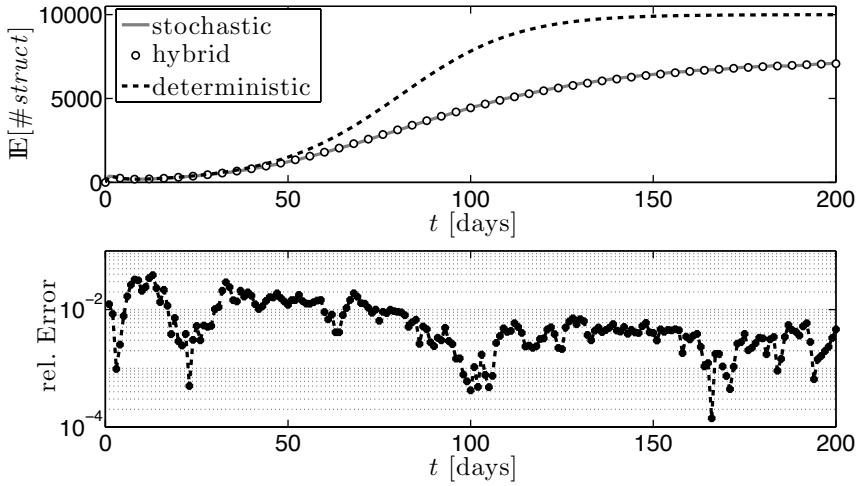

Figure 4.7: *Evolution of the expected level of struct in system* (4.39) *for parameter values listed in Table 4.2. The approximation obtained by ten thousand SSA runs is plotted as a solid line, the predicted expectation in the hybrid approach is marked by dots, the purely deterministic prediction is illustrated by a dashed line. Bottom: The relative error of the hybrid solution with respect to the SSA results against time.*

In Figures 4.4 and 4.5 the predicted PDFs of the discrete species *tem* and *gen*, respectively, are shown at four different time points. The corresponding expected levels of *struct* for a given number of *tem* molecules are plotted in Figure 4.6. The results of the hybrid model are in excellent agreement with the approximations obtained by SSA realizations. Although the initial number of *struct* is low, we observed no problems in treating *struct* as a continuous species in the initial phase. This illustrates that the required large copy numbers of continuously treated species seems to be only a sufficient condition and that our hybrid approach might also be applicable to systems where initial levels of the continuous species are low. Further investigations are thus needed to illuminate under which circumstances the conditions on the continuous species can be weakened.

Furthermore, it can be seen that the conditional expected levels of *struct* remain almost constant during later phase (compare the results for t = 100, 150 and 200 days in Figure 4.6), indicating that the continuously approximated processes are already in equilibrium. Instead it is the probability to be in a specific subset of the discrete state space that changes (see Figures 4.4 and 4.5), and hence the indirect impact of the discrete reactions that results in changes of the partial expectations of *struct* (data not shown). As illustrated in Figure 4.7, this results in a very accurate prediction of the total expectation of *struct* by the hybrid model, whereas a purely deterministic model fails.

4.2.3 Transcriptional Regulation

To further illustrate the proposed hybrid approach on a more complex biological system, we consider the transcriptional regulatory system published in [49]. The system includes six species, interacting through ten elementary reactions:

$$
\begin{aligned}
R_1&: \quad mRNA \xrightarrow{k_1} mRNA + M, & R_2&: \quad M \xrightarrow{k_2} \varnothing,\\
R_3&: \quad DNA{\cdot}D \xrightarrow{k_3} DNA{\cdot}D + mRNA, & R_4&: \quad mRNA \xrightarrow{k_4} \varnothing,\\
R_5&: \quad DNA + D \xrightarrow{k_5} DNA{\cdot}D, & R_6&: \quad DNA{\cdot}D \xrightarrow{k_6} DNA + D,\\
R_7&: \quad DNA{\cdot}D + D \xrightarrow{k_7} DNA{\cdot}2D, & R_8&: \quad DNA{\cdot}2D \xrightarrow{k_8} DNA{\cdot}D + D,\\
R_9&: \quad 2\,M \xrightarrow{k_9} D, & R_{10}&: \quad D \xrightarrow{k_{10}} 2\,M,
\end{aligned}
\tag{4.40}
$$

where M is a protein (monomer) the can reversibly dimerise (R_9 and R_{10}) to form the transcription factor D (dimer). The DNA template has two different binding sites for D, where DNA denotes the state where both sites are free, $DNA{\cdot}D$ denotes the state where D is bound at the first site, and $DNA{\cdot}2D$ denotes the state where the transcription factor D is bound at both sites. In the model it is assumed that D can only bind to the second site (R_7 and R_8) if the first site is already occupied (R_5 and R_6). Transcription only occurs from the state $DNA{\cdot}D$ (reaction R_3), where the transcription factor is exclusively bound to the first binding site of the DNA. The produced messenger RNA, $mRNA$, is translated into protein M (reaction R_1). Both, M and $mRNA$, are subject to degradation (R_2 and R_4).

Table 4.3: *Parameter values of the transcriptional regulatory system* (4.40) *in unit per second, for an average cell volume* $V \approx 1.44 \times 10^{-15}$ *l (i.e.,* $\Omega = N_A V \approx 8.64 \times 10^8$ *l/mol) and a* 10*-fold increased transcription rate* k_3*, cf. [49]. The system is started with probability one in the state with* 2 *M,* 6 *D,* 2 *DNA, and no mRNA, DNA·D and DNA·2D molecules.*

k_1	k_2	k_3	k_4	k_5/Ω
0.043	7.0×10^{-4}	0.72	3.9×10^{-3}	0.014
k_6	k_7/Ω	k_8	k_9/Ω	k_{10}
0.48	1.4×10^{-4}	8.8×10^{-12}	0.029	0.5

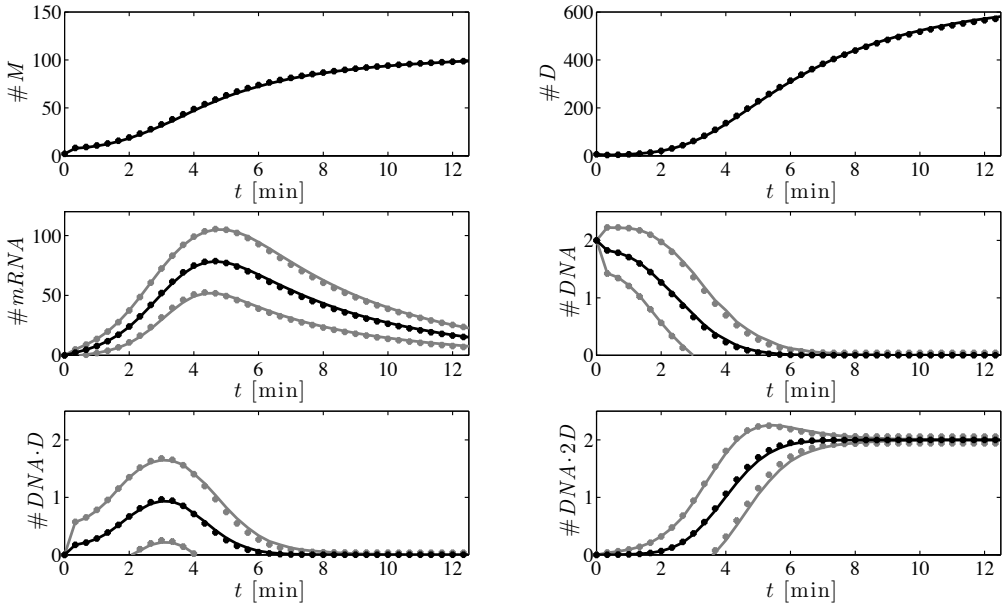

Figure 4.8: *Time evolutions of the expected values (black) plus/minus standard deviations (gray) in the transcriptional regulatory system* (4.40) *for parameter values listed in Table 4.3. The approximations obtained by ten thousand SSA runs are plotted as lines, the numerical solutions of the suggested hybrid model are marked by dots.*

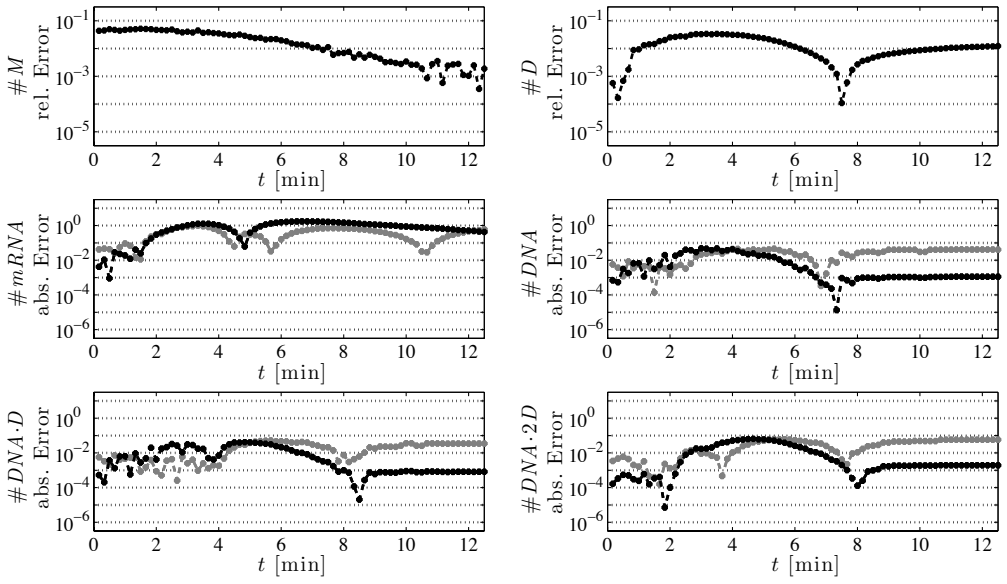

Figure 4.9: *Absolute and relative (with respect to the SSA predictions) errors, respectively, of the results shown in Figure 4.8. Black lines refer to the errors in expected values, gray lines to errors in standard deviations.*

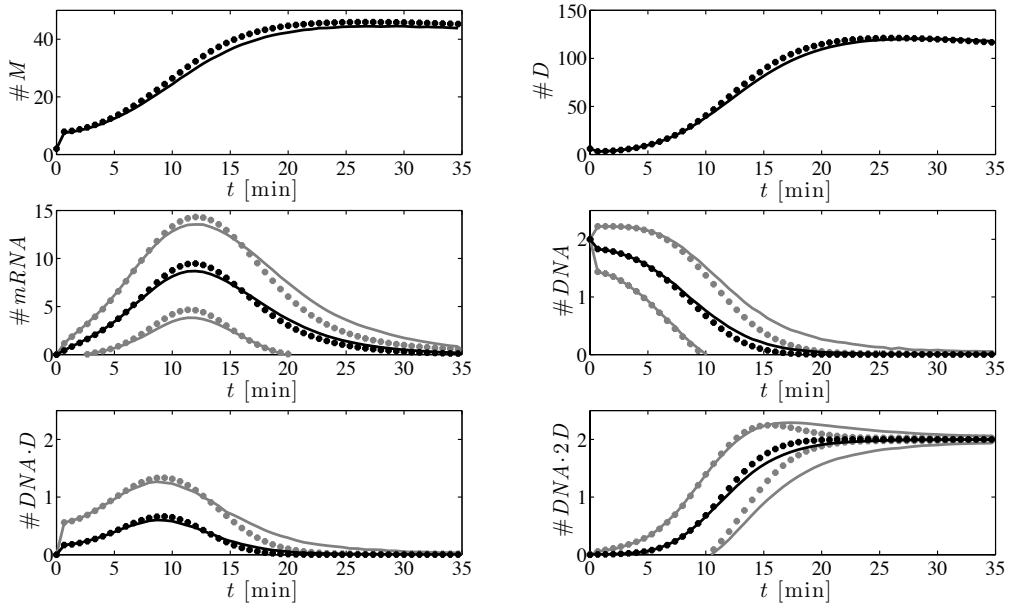

Figure 4.10: *Time evolutions of the expected values (black) plus/minus standard deviations (gray) in the transcriptional regulatory system* (4.40) *for parameter values listed in Table 4.3, but with a ten-fold decreased transcription rate k_3. The approximations obtained by ten thousand SSA runs are plotted as lines, the numerical solutions of the suggested hybrid model are marked by dots.*

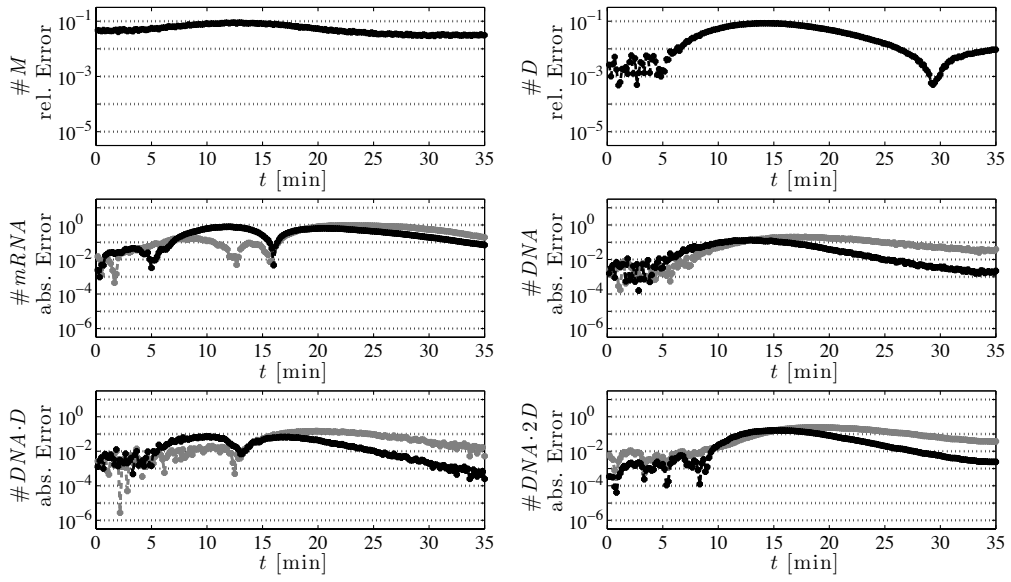

Figure 4.11: *Absolute and relative (with respect to the SSA predictions) errors, respectively, of the results shown in Figure 4.10. Black lines refer to errors in expected values, gray lines to errors in standard deviations.*

Based on common knowledge in cell biology, we expect the level of proteins to be much larger than the levels of mRNA and DNA. Therefore, we regard the monomer and dimer forms of the protein, M and D, as continuous species in our hybrid model, whereas *mRNA* and all DNA forms (*DNA*, $DNA \cdot D$ and $DNA \cdot 2D$) are treated as discrete species. As reactions R_3–R_8 change the levels of the discrete species, we treat these reactions as discrete processes. Reactions R_1, R_2, R_9 and R_{10} only change the levels of the continuous species (M and D) and hence, are approximated as continuous processes. Given this partition of the species and reactions, we can formulate and numerically solve the corresponding hybrid equations of the system.

In Figure 4.8 the evolution of the system as predicted by the hybrid model is compared to the results obtained from ten thousand SSA realizations. The predictions are in excellent agreement, as further illustrated in Figure 4.9. We again emphasize that the results computed with the SSA necessarily include an unknown sampling error. While the error in the numerical integration of the proposed hybrid equations is controlled by standard methods, the error made with an indirect Monte Carlo method is hard to estimate. In the hybrid model, however, an additional error arises from approximating part of the network continuously–deterministically.

We further studied the error related to modeling the two protein forms M and D as continuous species by decreasing the transcription rate k_3, resulting in lower levels of monomer M and, consequently, dimer D. As can be seen from the results shown in Figures 4.10 and 4.11, the error of our hybrid approximation increases in this scenario. Particularly in the later transient phase (after ≈ 10 min), the predicted standard deviations of the discrete species show a noticeable deviation from the SSA results. However, these deviations are not surprising as for lower levels of M and D an error resulting from their continuous–deterministic approximation increases, which is also intuitively clear.

4.3 Concluding Remarks

We used multi-scale analysis techniques to derive a novel hybrid model for approximation of the PDF solution of the CME. To this end, we singled out a subspace associated with species of low copy numbers and assumed that its complement subspace is associated with large copy numbers which can be well approximated by a continuous distribution. We exploited the natural decomposition $P(Y, Z; t) = P(Y \mid Z; t)P(Z; t)$ of the joint PDF into the marginal PDF

$P(Z;t)$ on the low copy number space and the conditional PDF $P(Y \mid Z;t)$ on the large copy number space, where the condition is on the discrete states of the low copy number subspace. The hybrid model resulting from multi-scale asymptotics based on this decomposition couples a CME for $P(Z;t)$ to DAEs for the first moments of $P(Y \mid Z;t)$. Therefore, in first instance, there have to be as many DAEs for every discrete state Z of the low copy number CME as 'continuous' species are present in the system. Although this might first appear very complex, the hybrid model solution is expected to be particularly suitable for networks including a few 'discrete' species only, since the numerical costs directly scale with the size of the CME subspace which is much smaller for the hybrid model than for the original CME. Moreover, as shown in Subsection 4.1.8, under certain conditions the hybrid description can be further coarse grained and the number of DAEs effectively reduced without introducing any additional approximation error. Hence, a direct solution of the proposed hybrid model becomes feasible, which was demonstrated on a viral infection kinetics model and a transcriptional regulatory network by using a straight-forward simulation approach.

The focus of our work was not on a particular numerical integration method for the CME but rather on a general, theoretically justified approach for its direct, hybrid stochastic–deterministic approximation. Similar approaches have been proposed by Henzinger et al. [66] and, recently, by T. Jahnke [67]. In contrast to Henzinger et al. [66], we derived a closed formulation where the distribution of the discrete species and the first moments of the continuous species are coupled continuously in time. Our hybrid model (4.30) is more closely related to Jahnke's MRCE-model [67]. The evolution equations of the MRCE-model are derived by a partial averaging of the CME (see Section 4.1.2) with respect to some variables that are assumed to have zero covariance. This assumption becomes theoretically justified for the continuous species in our hybrid approach under the considered partial scaling of the system. In the light of our multi-scale expansion, we hence expect the MRCE-model to approximate the system dynamics on the $\mathcal{O}(1)$ scale. At the same time, however, our analysis also shows that the direct impact of slow reactions (associated with the discrete species) on the dynamics of the continuous variables lives on a smaller scale and can thus be neglected in a corresponding hybrid approximation up to order $\mathcal{O}(1)$.

It should be emphasized that the resulting DAEs that govern the evolution of the continuous variables are *not* identical with the usually expected equations of classical reaction kinetics, e.g., compare eq. (4.34) with eq. (4.30b). Instead, they include additional coupling terms resulting from changes in the distribution of the population on the discrete subspace as given by the discrete reactions in eq. (4.30b). The numerical experiments indicated that the required large

number of molecules of all continuous species might be only a sufficient condition and that our hybrid approach can also be applied for more general situations. In principle, we just have to require the propensity scalings (4.10) to be valid in the specific subspaces, which can still be satisfied even though the conditional level of a continuously treated species is (temporarily) low. Further investigations are thus needed to illuminate under which circumstances the conditions on the continuous species can be weakened. However, observation of a critically large propensity (order ε^{-1}) for one of the discrete reactions R_μ with $\mu \in \mathcal{M}^d$ may spoil the approximation property of the continuous variables if there exists a direct dependence (i.e., $\zeta_\mu \neq 0$), and thus should be considered in choosing the discrete species and reactions. Such a potential scenario has to be analyzed in further research, e.g., by additionally accounting for different scales of the rate constants, which was out of scope for the present study.

5

Elimination of Discrete–Stochastic Submodels: Effective Protein Synthesis Rates in the FlgM–FliA Regulatory Network[§]

In the previous chapters, we studied hybrid approaches that couple the discrete–stochastic with the continuous–deterministic formulation of biochemical reaction networks. A hybrid system representation allows to reduce the computational costs of simulation while preserving characteristic properties related to fluctuations in small number of molecules. It is important to realize, however, that this does not reduce the complexity of the model with respect to the number of molecular species, reaction channels and, more generally, parameters. Usually, some model parameters can be directly obtained from experiments or literature. Other parameters, such as kinetic constants, can often not be measured directly and have to be estimated indirectly by fitting the model to experimental data. However, parameter estimation can only provide reliable results if the model complexity is in balance with the amount and quality of experimental data [68]. If this is not the case, either additional data is required or the model has to be simplified, e.g., based on time-scale separation [69–71], sensitivity analysis [72, 73] or balanced truncation [74–76].

In this chapter, we employ a QSSA-like reduction approach to a detailed stochastic model of gene expression. Our strategy is to eliminate the processes and parameters associated with this detailed model in order to derive effective protein synthesis rates as typically used in deterministic models of biochemical reaction networks. We demonstrate our approach on a biological model of the interactions of two key proteins in the flagellar gene regulation cascade of *Escherichia coli* (*E. coli*): the flagellar sigma factor FliA and its anti-sigma factor FlgM, see Section 5.2. The interactions between FlgM and FliA serve as a major checkpoint during the assembly process that temporally separates the expression of middle and late genes, see next

[§]The contents of this chapter are based on: S. Menz, R. Hengge, and W. Huisinga. Robustness and Sensitivity of the FlgM–FliA Regulatory Network Controlling Flagellar Gene Expression in Escherichia Coli. In preparation, 2012.

section. Our reduction approach, see Section 5.3, allows to mechanistically incorporate the feedback between middle and late gene expression in a purely deterministic formulation of the model. Sensitivity analysis of the derived model reveals that the checkpoint mechanism is very sensitive to changes in levels of competing sigma factors, allowing the bacterium to rapidly adapt to a changing environment, see Subsection 5.4.3. We also find a high sensitivity to changes in the effective synthesis rates. However, this high sensitivity does not generally carry over to parameters in the detailed model, demonstrating that sensitivity to changes in effective reaction rates has to be interpreted with care.

5.1 Background

E. coli is a non-differentiating bacterium that exhibits very different 'life-styles': The bacteria can occur as single planktonic and motile cells or they can exist as multicellular sessile aggregates, i.e., in biofilms [192–197]. The motile state depends on properly controlled biosynthesis of flagella that are complex rotating organelles anchored in the cell envelope. The flagella comprise three parts: the basal body, the hook and the filament, which are sequentially assembled from the base to the distal end [198]. A remarkable aspect of flagellar assembly in *E. coli* is that gene expression is temporally ordered and coupled to the assembly process [199]. The same has been observed for other bacteria, like *Salmonella typhimurium* [200].

The flagellar gene regulation cascade of *E. coli* consists of more than sixty genes that are organized in three hierarchically and temporally regulated transcriptional classes [201–203]. Global regulators feed into a single class 1 promoter which leads to the initiation of flagellar synthesis. The class 1 (early) genes code for the subunits of the transcription factor FlhDC, the flagellar master regulator that activates class 2 promoters [203, 204]. The protein products of the class 2 (middle) genes are structural components of the flagellar hook basal body, as well as the transcriptional regulators FliA and FlgM. FliA is an alternative sigma factor (σ^F) that enables transcription of the class 3 (late) genes which encode the proteins for the flagellar filament and the control of motility and chemotaxis [205, 206]. In the middle phase of flagellar assembly, FliA is actively inhibited by FlgM, its anti-sigma factor, that tightly binds to FliA.

With the hook basal body, a type III secretion system is formed that is necessary for the secretion of the flagellar filament subunits [198, 207]. The type III secretion system exports FlgM to the outside of the cell with FliA acting as a type III chaperone that delivers FlgM to the export machinery. [208]. The FliA-mediated export of FlgM results in the release of FliA from

the FlgM:FliA complex, an increase in free FliA levels and eventually in activation of class 3 transcription [209]. In this way, class 3 gene expression of filament proteins is coupled to the assembly process of the hook basal body [210]. While the sequential nature of middle and late gene expression has been studied using real-time monitoring of transcriptional activation based on β-galactosidase [195] and green fluorescent protein [199, 211] fusion measurements, the dynamics of the FlgM–FliA checkpoint mechanism and the switch from middle to late gene expression are only poorly understood.

The objective of our *in silico* study was to analyze the timing and robustness of the FlgM–FliA core regulatory mechanism in terms of a mathematical model. Since regulation based on protein-protein interaction can not be studied by means of gene transcription data, quantitative molecule data has to be used. In [210], a semi-quantitative model (i.e., in terms of dimensionless concentration units) of the FlgM–FliA checkpoint in *Salmonella enterica* has been constructed and analyzed with respect to gene transcription data. We used for the first time quantitative measurements of FliA and FlgM protein numbers over time [195] to develop and analyze a detailed deterministic reaction kinetics model of the FlgM–FliA regulatory network in *E. coli*. To incorporate the feedback between middle and late gene expression in the deterministic model, we derived effective protein synthesis rates based on a detailed stochastic model of gene expression that explicitly accounts for initiation of late gene expression by FliA. The resulting model extends the current verbal description of the FlgM–FliA interactions by accounting for sigma factor competition for RNAP and the relation of effective protein synthesis rates to key transcriptional and translational processes, necessary for thorough analysis of their robustness and sensitivity.

Robustness is one of the fundamental characteristics of biological systems, defined as the ability to maintain a persistent response in the presence of perturbations or alterations in the involved molecular processes. Equally important, however, is the ability of a biological system to rapidly adapt to a changing environment [212, 213]. *In silico*, the underlying question of robustness: ''How sensitive is the model to perturbations in the molecular processes?'' is typically addressed by a sensitivity analysis of the predicted output in terms of model parameters (e.g., rates, rate constants or concentrations of molecular species that were not explicitly modeled). Our results provide detailed insight into the timing and robustness of the FlgM–FliA checkpoint mechanism. Since flagella are a common and conserved motive among motile bacteria [214], our findings are expected to have implications beyond the present study in *E. coli*.

5.2 Mathematical Model of the FlgM–FliA Regulatory Network

Description of the Model

The proposed model of the FlgM–FliA regulatory network is illustrated in Figure 5.1. In the model, the two key regulatory proteins FlgM and FliA form the FlgM:FliA complex with association rate constant k_{on} and dissociation rate constant k_{off} [215]. FliA binds to the RNA

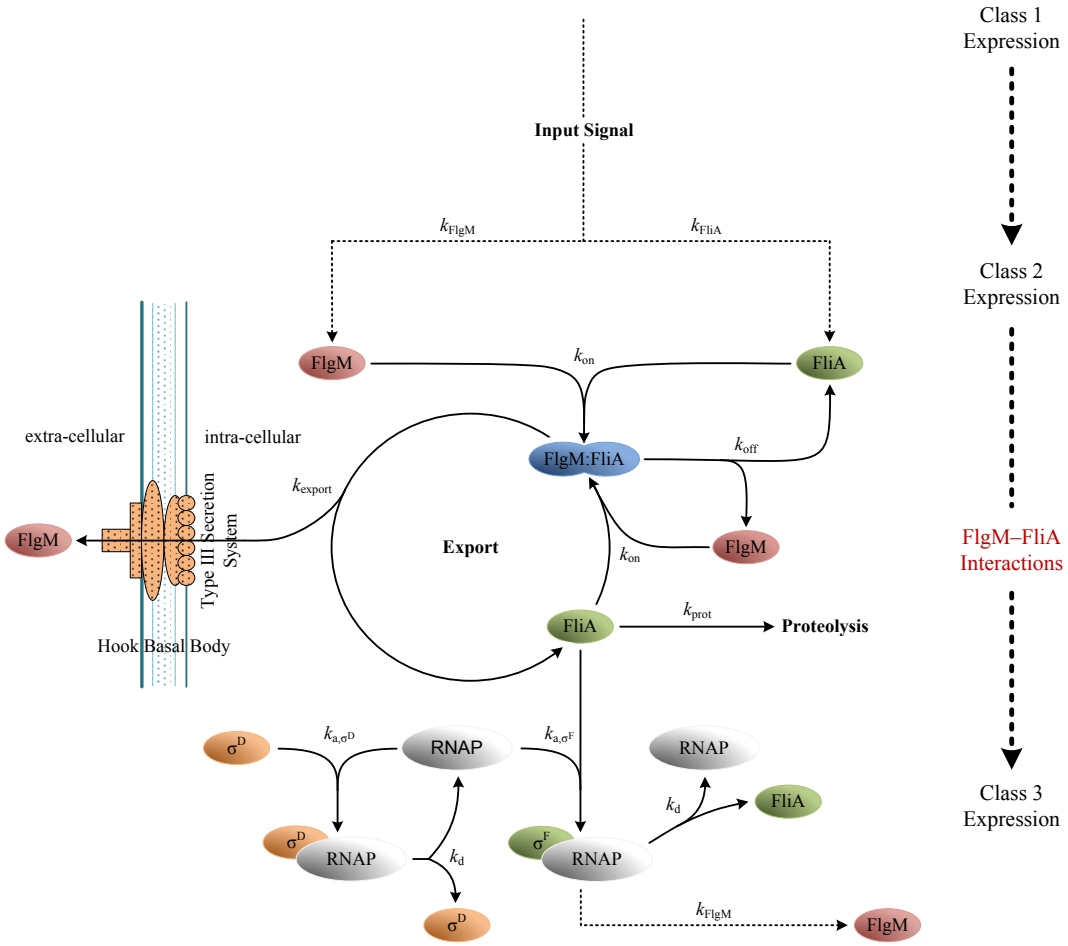

Figure 5.1: *Model of the central flagellar checkpoint mechanism in E. coli. Interactions between FliA and FlgM are at the core, i.e., FlgM:FliA complex formation, FliA-mediated export of FlgM through the type III secretion system, proteolysis of free FliA, FliA binding to RNAP, and σ^F:RNAP enabled class 3 gene expression of FlgM. FliA is competing for binding to RNAP with the sigma factor σ^D.*

polymerase core enzyme RNAP with rate constant k_{a,σ^F} and dissociates from the σ^F:RNAP holoenzyme with rate constant k_d [215]. FliA is competing for binding to RNAP with the sigma factor σ^D, that binds the RNA polymerase core enzyme with rate constant k_{a,σ^D} and dissociates from the σ^D:RNAP holoenzyme with rate constant k_d [215, 216].

FlgM and FliA are produced with rates $k_{\mathrm{FlgM}}(t)$ and $k_{\mathrm{FliA}}(t)$, respectively. These rates comprise a low basal expression rate and a class 2 expression rate that is induced at time t_{class2} by the master regulator FlhDC [203, 204] (only implicitly considered in our model, see Appendix C). In addition, FlgM is subject to class 3 expression that is induced by the σ^F:RNAP holoenzyme. Class 3 expression of FliA was experimentally found to be negligible in comparison to its class 2 expression [195].

As part of the hook basal body, a type III secretion system is established [198]. Upon completion at time t_{export}, FliA acts as a type III secretion chaperon [208] and intra-cellular FlgM is exported with rate constant k_{export} from the FlgM:FliA complex into the extra-cellular space, resulting in extra-cellular $\mathrm{FlgM}_{\mathrm{extern}}$ and free intra-cellular FliA.

The molecular species FliA, FlgM, FlgM:FliA, $\mathrm{FlgM}_{\mathrm{extern}}$ and FliA as part of the σ^F:RNAP complex are subject to dilution with rate constant k_{dil}, accounting for cell growth and division. In addition, like some other sigma factors [217, 218], FliA is subject to proteolysis with rate constant k_{prot} [195]. The total number of sigma factor σ^D, given by σ^D plus σ^D:RNAP, and the total number of RNA polymerases, i.e., RNAP plus σ^F:RNAP plus σ^D:RNAP, are assumed to remain constant due to regulatory mechanisms that were not explicitly included in the model.

In summary, the proposed model of the FlgM–FliA regulatory network model involves the following reactions:

$$
\begin{aligned}
\varnothing &\underset{k_{\mathrm{dil}}}{\overset{k_{\mathrm{FlgM}}(t)}{\rightleftarrows}} \mathrm{FlgM}, & \varnothing &\underset{k_{\mathrm{dil}}}{\overset{k_{\mathrm{FliA}}(t)}{\rightleftarrows}} \mathrm{FliA},\\
\mathrm{FlgM} + \mathrm{FliA} &\underset{k_{\mathrm{off}}}{\overset{k_{\mathrm{on}}}{\rightleftarrows}} \mathrm{FlgM{:}FliA}, & \mathrm{FliA} &\xrightarrow{k_{\mathrm{prot}}} \varnothing,\\
\mathrm{FlgM{:}FliA} &\xrightarrow{k_{\mathrm{export}}(t)} \mathrm{FlgM}_{\mathrm{extern}} + \mathrm{FliA}, & \mathrm{FlgM}_{\mathrm{extern}} &\xrightarrow{k_{\mathrm{dil}}} \varnothing,\\
\mathrm{FlgM{:}FliA} &\xrightarrow{k_{\mathrm{dil}}} \varnothing, & \mathrm{FliA} + \mathrm{RNAP} &\underset{k_d}{\overset{k_{a,\sigma^F}}{\rightleftarrows}} \sigma^F\mathrm{{:}RNAP},\\
\sigma^F\mathrm{{:}RNAP} &\xrightarrow{k_{\mathrm{dil}}} \mathrm{RNAP}, & \sigma^D + \mathrm{RNAP} &\underset{k_d}{\overset{k_{a,\sigma^D}}{\rightleftarrows}} \sigma^D\mathrm{{:}RNAP}. \qquad (5.1)
\end{aligned}
$$

As the number of molecules of the species are expected to be sufficiently large according to experimental measurements (see Appendix C), we choose the deterministic formulation of reaction kinetics to transform the above biological model into a system of ODEs:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{FlgM} &= +k_{\mathrm{FlgM}}(t) - k_{\mathrm{on}}\cdot\mathrm{FliA}\cdot\mathrm{FlgM} + k_{\mathrm{off}}\cdot\mathrm{FlgM{:}FliA} - k_{\mathrm{dil}}\cdot\mathrm{FlgM},\\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{FliA} &= +k_{\mathrm{FliA}}(t) - k_{\mathrm{on}}\cdot\mathrm{FliA}\cdot\mathrm{FlgM} + \left(k_{\mathrm{off}} + k_{\mathrm{export}}(t)\right)\cdot\mathrm{FlgM{:}FliA}\\
&\quad - k_{\mathrm{a},\sigma^{\mathrm{F}}}\cdot\mathrm{FliA}\cdot\mathrm{RNAP} + k_{\mathrm{d}}\cdot\sigma^{\mathrm{F}}{:}\mathrm{RNAP} - \left(k_{\mathrm{dil}} + k_{\mathrm{prot}}\right)\cdot\mathrm{FliA},\\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{FlgM{:}FliA} &= +k_{\mathrm{on}}\cdot\mathrm{FliA}\cdot\mathrm{FlgM} - \left(k_{\mathrm{off}} + k_{\mathrm{export}}(t) + k_{\mathrm{dil}}\right)\cdot\mathrm{FlgM{:}FliA},\\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{FlgM}_{\mathrm{extern}} &= +k_{\mathrm{export}}(t)\cdot\mathrm{FlgM{:}FliA} - k_{\mathrm{dil}}\cdot\mathrm{FlgM}_{\mathrm{extern}},\\
\frac{\mathrm{d}}{\mathrm{d}t}\sigma^{\mathrm{F}}{:}\mathrm{RNAP} &= +k_{\mathrm{a},\sigma^{\mathrm{F}}}\cdot\mathrm{FliA}\cdot\mathrm{RNAP} - \left(k_{\mathrm{d}} + k_{\mathrm{dil}}\right)\cdot\sigma^{\mathrm{F}}{:}\mathrm{RNAP},\\
\frac{\mathrm{d}}{\mathrm{d}t}\sigma^{\mathrm{D}}{:}\mathrm{RNAP} &= +k_{\mathrm{a},\sigma^{\mathrm{D}}}\cdot\sigma^{\mathrm{D}}\cdot\mathrm{RNAP} - k_{\mathrm{d}}\cdot\sigma^{\mathrm{D}}{:}\mathrm{RNAP},\\
\frac{\mathrm{d}}{\mathrm{d}t}\sigma^{\mathrm{D}} &= -k_{\mathrm{a},\sigma^{\mathrm{D}}}\cdot\sigma^{\mathrm{D}}\cdot\mathrm{RNAP} + k_{\mathrm{d}}\cdot\sigma^{\mathrm{D}}{:}\mathrm{RNAP},\\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{RNAP} &= -k_{\mathrm{a},\sigma^{\mathrm{F}}}\cdot\mathrm{FliA}\cdot\mathrm{RNAP} + \left(k_{\mathrm{d}} + k_{\mathrm{dil}}\right)\cdot\sigma^{\mathrm{F}}{:}\mathrm{RNAP} - k_{\mathrm{a},\sigma^{\mathrm{D}}}\cdot\sigma^{\mathrm{D}}\cdot\mathrm{RNAP},\\
&\quad + k_{\mathrm{d}}\cdot\sigma^{\mathrm{D}}{:}\mathrm{RNAP},
\end{aligned}
\tag{5.2}
$$

where zero-order rates are in units nM/min, first-order rate constants are in units min^{-1}, and second-order rate constants are in units $(\mathrm{nM}\cdot\mathrm{min})^{-1}$.

Class 2 expression is regulated by the master regulator FlhDC and the sigma factor σ^{D}, while class 3 expression is regulated by the sigma factor FliA. Since the deterministic model of biochemical reaction kinetics does not allow modeling of gene expression on the level of single molecules (in particular a single gene), we used the stochastic formulation of reaction kinetics to derive effective class 2 and class 3 synthesis rates $k_{\mathrm{FlgM}}(t)$ and $k_{\mathrm{FliA}}(t)$. As a consequence, we retained a mechanistic interpretation of the effective synthesis rates while keeping the number of model parameters low.

5.3 Derivation of Effective Protein Synthesis Rates

We derived effective rates of protein synthesis based on the detailed mechanistic model of gene transcription and translation illustrated in Figure 5.2. In this detailed model, we

explicitly included the impact of transcription and sigma factors on the intermediate steps of mRNA synthesis. This allowed us to derive protein synthesis rates as a function of the master regulator FlhDC and the σ^D:RNAP complex for class 2 gene expression, and as a function of the σ^F:RNAP complex for class 3 gene expression. Our reduction approach uses a QSSA-like

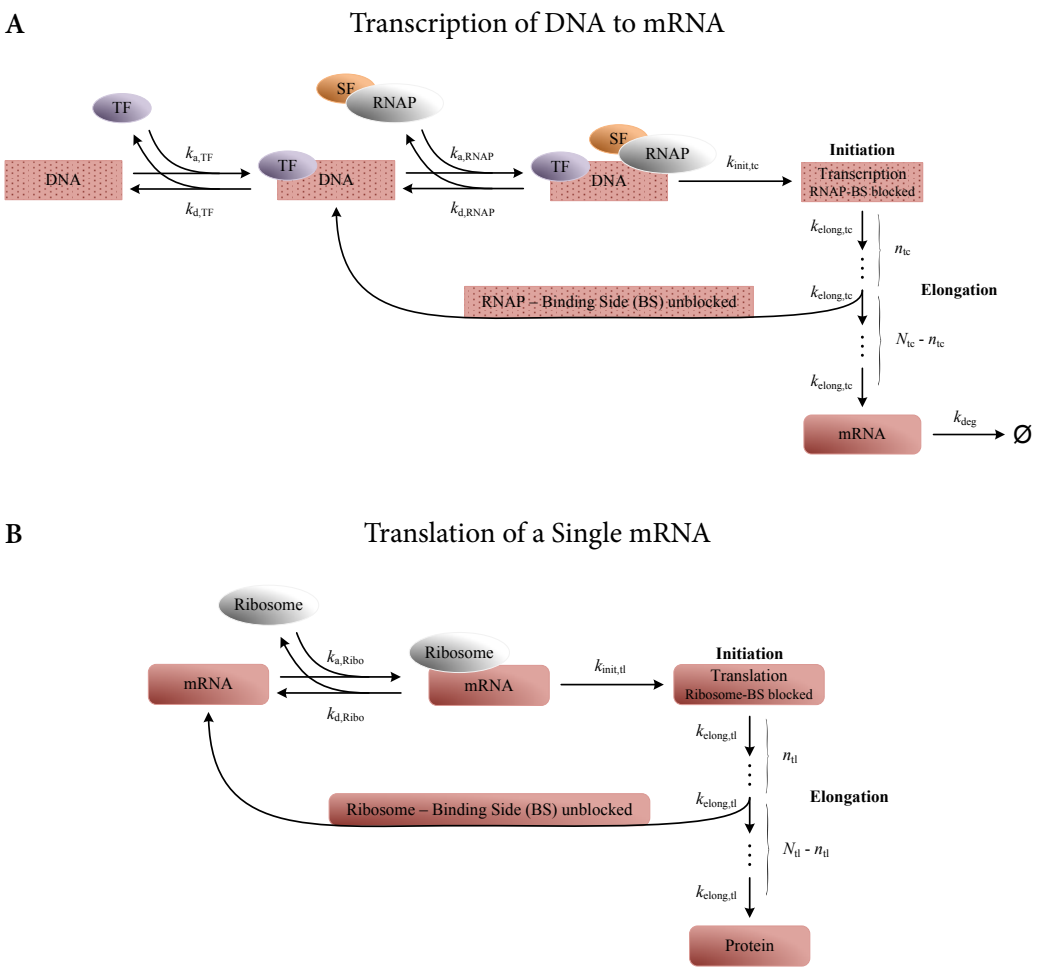

Figure 5.2: *Detailed mechanistic model of gene expression. (A) The transcription of DNA to mRNA is separated into four states: free DNA, transcription factor (TF)-bound DNA, TF/SF:RNAP-bound DNA, and initiated transcription with blocked RNAP-binding side (BS). From the last state, the RNAP-binding side gets unblocked after a distance of n_{tc} nucleotides is reached (promoter clearance). The exit rate of this state determines the effective synthesis rate of mRNA. (B) The translation of a single mRNA is separated into three states: free mRNA, ribosome-bound mRNA and initiated translation with blocked ribosome-BS. Another ribosome can bind to the mRNA after n_{tl} amino acids are synthesized.*

argument: If transcriptional and translational processes are not explicitly incorporated in a model, we assume that in corresponding protein synthesis rates the variables associated with these subprocesses are considered to be effectively in a quasi-steady state distribution for given levels of transcription factors, RNAP holoenzymes and ribosomes.

For the process of gene transcription, we considered free DNA to which a transcription factor (TF) may reversibly bind with association and dissociation rate constants $k_{\mathrm{a,TF}}$ and $k_{\mathrm{d,TF}}$, respectively. The transcription factor (TF)-bound DNA has an increased affinity for binding the RNAP holoenzyme (SF:RNAP) to the promoter of the gene with association and dissociation rate constants $k_{\mathrm{a,RNAP}}$ and $k_{\mathrm{d,RNAP}}$, respectively. Once the RNAP complex is bound, transcription is initiated with rate constant $k_{\mathrm{init,tc}}$ and the RNAP traverses the template strand for the full gene length N_{tc} with elongation rate constant $k_{\mathrm{elong,tc}}$. For a length of n_{tc} nucleotides, the RNAP-binding side (BS) of the DNA is blocked due to steric constraints [219]. Thereafter, it is assumed to be unblocked again, allowing the next SF:RNAP complex to bind and multiple transcription processes to run simultaneously along a single gene or operon. The corresponding biological model is shown in Figure 5.2A.

Based on the stochastic formulation of reaction kinetics, we associated a Markov jump process with the above model of gene transcription. Considering the transcription factor and the RNAP holoenzyme as parametric input, the resulting Markov process is linear on the state space of the four different DNA states ($X_1 :=$ free DNA, $X_2 :=$ TF-bound DNA, $X_3 :=$ TF/SF:RNAP-bound DNA, $X_4 :=$ initiated transcription with blocked RNAP-binding side). For our purpose, it was sufficient to lump the n_{tc} and the $N_{\mathrm{tc}} - n_{\mathrm{tc}}$ elongation steps with rate constant $k_{\mathrm{elong,tc}}$ into single steps with rate constants $k^*_{\mathrm{elong,tc}} := k_{\mathrm{elong,tc}}/n_{\mathrm{tc}}$ and $k_{\mathrm{elong,tc}}/(N_{\mathrm{tc}} - n_{\mathrm{tc}})$, respectively (explicit consideration of the N_{tc} elongation steps would give the same results). The infinitesimal generator A_{tc} of the Markov jump process on the states space (X_1, X_2, X_3, X_4) is given by

$$A_{\mathrm{tc}} = \begin{pmatrix} -k^*_{\mathrm{a,TF}} & k^*_{\mathrm{a,TF}} & 0 & 0 \\ k_{\mathrm{d,TF}} & -(k_{\mathrm{d,TF}} + k^*_{\mathrm{a,RNAP}}) & k^*_{\mathrm{a,RNAP}} & 0 \\ 0 & k_{\mathrm{d,RNAP}} & -(k_{\mathrm{d,RNAP}} + k_{\mathrm{init,tc}}) & k_{\mathrm{init,tc}} \\ 0 & k^*_{\mathrm{elong,tc}} & 0 & -k^*_{\mathrm{elong,tc}} \end{pmatrix},$$

where $k^*_{\mathrm{a,TF}} := k_{\mathrm{a,TF}} \cdot \mathrm{TF}$ and $k^*_{\mathrm{a,RNAP}} := k_{\mathrm{a,RNAP}} \cdot \mathrm{SF{:}RNAP}$ are the binding rate constants parameterized in terms of the transcription factor TF and the RNAP holoenzyme (i.e., SF:RNAP complex), respectively.

We defined the effective synthesis rate k_{mRNA} of mRNA as the steady state rate resulting in the synthesis of new mRNA molecules. This is given, for instance, by the steady state rate through the state X_4, i.e., the product between the stationary distribution $\pi = (\pi_1, \ldots, \pi_4)$ of the Markov jump process associated with state X_4 and the exit rate constant of state X_4

$$k_{\mathrm{mRNA}} := \pi_4 \cdot k^*_{\mathrm{elong,tc}}. \tag{5.3}$$

The stationary distribution satisfies the equation $\pi A_{\mathrm{tc}} = 0$. Solving for π and using eq. (5.3) resulted in

$$k_{\mathrm{mRNA}} = \frac{V_{\mathrm{max,tc}} \cdot \mathrm{SF{:}RNAP}}{K_{\mathrm{tc}} + \mathrm{SF{:}RNAP}}, \tag{5.4}$$

with maximal rate

$$V_{\mathrm{max,tc}} := \frac{k_{\mathrm{init,tc}} \cdot k^*_{\mathrm{elong,tc}}}{k_{\mathrm{init,tc}} + k^*_{\mathrm{elong,tc}}}$$

and a Michaelis–Menten-like constant

$$K_{\mathrm{tc}} = \frac{(k_{\mathrm{d,RNAP}} + k_{\mathrm{init,tc}}) \cdot k^*_{\mathrm{elong,tc}}}{k_{\mathrm{a,RNAP}} \cdot (k_{\mathrm{init,tc}} + k^*_{\mathrm{elong,tc}})} \cdot \left(1 + \frac{K_{\mathrm{D,TF}}}{\mathrm{TF}}\right), \tag{5.5}$$

i.e., the SF:RNAP concentration for which $k_{\mathrm{mRNA}} = V_{\mathrm{max,tc}}/2$, where $K_{\mathrm{D,TF}} = k_{\mathrm{d,TF}}/k_{\mathrm{a,TF}}$. Assuming a linear degradation of the mRNA species, e.g., due to dilution processes, resulted in a total steady state mRNA level of

$$\mathrm{mRNA}^*_{\mathrm{total}} = \frac{k_{\mathrm{mRNA}}}{k_{\mathrm{deg}}}. \tag{5.6}$$

We next determined the protein synthesis rate per single mRNA molecule. In combination with the above result, this finally allowed us to derive the effective protein synthesis rate as a function of both transcription factor and RNAP holoenzyme concentration. The derivation of the effective protein synthesis rate per mRNA is based on the biological model of translation shown in Figure 5.2B. For the process of translation, we considered a ribosome reversibly binding to the mRNA with association and dissociation rate constants $k_{\mathrm{a,Ribo}}$ and $k_{\mathrm{d,Ribo}}$, respectively. From this ribosome-bound state the translation of the mRNA is initiated with rate constant $k_{\mathrm{init,tl}}$, and the ribosome starts to synthesize the decoded amino acid sequence of length N_{tl} with elongation rate constant $k_{\mathrm{elong,tl}}$. Multiple ribosomes can bind to the same mRNA and boost the translation process, with an average distance of n_{tl} between them.

The infinitesimal generator A_{tl} corresponding to the Markov jump process on the states space ($Y_1 :=$ free mRNA, $Y_2 :=$ ribosome-bound mRNA, $Y_3 :=$ initiated translation with blocked ribosome-binding side) is given by

$$A_{\mathrm{tl}} = \begin{pmatrix} -k^*_{\mathrm{a,Ribo}} & k^*_{\mathrm{a,Ribo}} & 0 \\ k_{\mathrm{d,Ribo}} & -(k_{\mathrm{d,Ribo}} + k_{\mathrm{init,tl}}) & k_{\mathrm{init,tl}} \\ k^*_{\mathrm{elong,tl}} & 0 & -k^*_{\mathrm{elong,tl}} \end{pmatrix},$$

where $k^*_{\mathrm{elong,tl}} := k_{\mathrm{elong,tl}}/n_{\mathrm{tl}}$ and $k^*_{\mathrm{a,Ribo}} := k_{\mathrm{a,Ribo}} \cdot \mathrm{Ribo}$ denotes the rate of ribosomal binding to mRNA parameterized by the concentration Ribo of free ribosomes.

We defined the effective protein synthesis rate per mRNA as the steady state rate resulting in synthesis of new protein molecules. This is given, e.g., by the steady state rate through the state Y_3, that is the stationary distribution $\mu = (\mu_1, \mu_2, \mu_3)$ of the Markov jump process associated with state Y_3 times the exit rate constant of state Y_3. The total protein synthesis rate k_{synth} is thus finally given by the synthesis rate per mRNA molecule times the total number of mRNA molecules, i.e.,

$$k_{\mathrm{synth}} := \mu_3 \cdot k^*_{\mathrm{elong,tl}} \cdot \mathrm{mRNA}^*_{\mathrm{total}}. \tag{5.7}$$

The stationary distribution satisfies the equation $\mu A_{\mathrm{tl}} = 0$. Solving for μ and using eq. (5.7) resulted in

$$k_{\mathrm{synth}} = \frac{V_{\mathrm{max,tl}} \cdot \mathrm{Ribo}}{K_{\mathrm{tl}} + \mathrm{Ribo}} \cdot \mathrm{mRNA}^*_{\mathrm{total}}, \tag{5.8}$$

with maximal rate

$$V_{\mathrm{max,tl}} := \frac{k_{\mathrm{init,tl}} \cdot k^*_{\mathrm{elong,tl}}}{k_{\mathrm{init,tl}} + k^*_{\mathrm{elong,tl}}}$$

and a Michaelis–Menten-like constant

$$K_{\mathrm{tl}} := \frac{(k_{\mathrm{d,Ribo}} + k_{\mathrm{init,tl}}) \cdot k^*_{\mathrm{elong,tl}}}{k_{\mathrm{a,Ribo}} \cdot (k_{\mathrm{init,tl}} + k^*_{\mathrm{elong,tl}})}.$$

The equation for k_{synth} can be further simplified if we assume that ribosomes are present in such quantities that ribosomal binding is not the limiting step in the translation process. This implies $\mathrm{Ribo} \gg K_{\mathrm{tl}}$, resulting in

$$k_{\mathrm{synth}} = V_{\mathrm{max,tl}} \cdot \mathrm{mRNA}^*_{\mathrm{total}}, \tag{5.9}$$

or equivalently

$$k_{\mathrm{synth}} = \frac{V_{\mathrm{max}} \cdot \mathrm{SF{:}RNAP}}{K_{\mathrm{M}} + \mathrm{SF{:}RNAP}}, \tag{5.10}$$

where $V_{\max} := V_{\max,\mathrm{tl}} \cdot V_{\max,\mathrm{tc}}/k_{\mathrm{deg}}$ and $K_{\mathrm{M}} := K_{\mathrm{tc}}$. If condition Ribo $\gg K_{\mathrm{tl}}$ is not satisfied, then we would nevertheless obtain the same eq. (5.10) for k_{synth} by replacing eq. (5.9) by eq. (5.8) and assuming that the number of available ribosomes remains roughly constant.

In the flagellar gene regulation cascade, we have SF:RNAP = σ^{D}:RNAP and TF = FlhDC for class 2 gene expression, resulting in a potentially saturable class 2 synthesis rate given by

$$k_{\mathrm{class2}} = \frac{V_{\max} \cdot \sigma^{\mathrm{D}}\text{:RNAP}}{K_{\mathrm{M2}} + \sigma^{\mathrm{D}}\text{:RNAP}}, \tag{5.11}$$

with

$$K_{\mathrm{M2}} = \frac{(k_{\mathrm{d,RNAP}} + k_{\mathrm{init,tc}}) \cdot k^*_{\mathrm{elong,tl}}}{k_{\mathrm{a,RNAP}} \cdot (k_{\mathrm{init,tc}} + k^*_{\mathrm{elong,tl}})} \cdot \left(1 + \frac{K_{\mathrm{D,FlhDC}}}{\mathrm{FlhDC}}\right). \tag{5.12}$$

For class 3 gene expression, it is SF:RNAP = σ^{F}:RNAP but no transcription factor is involved such that the state X_1 is not present. This can be accounted for by replacing the term $(1 + K_{\mathrm{D,TF}}/\mathrm{TF})$ by 1 in eq. (5.5). Hence, we derived a class 3 synthesis rate of the form

$$k_{\mathrm{class3}} = \frac{V_{\max} \cdot \sigma^{\mathrm{F}}\text{:RNAP}}{K_{\mathrm{M3}} + \sigma^{\mathrm{F}}\text{:RNAP}}, \tag{5.13}$$

with

$$K_{\mathrm{M3}} = \frac{(k_{\mathrm{d,RNAP}} + k_{\mathrm{init,tc}}) \cdot k^*_{\mathrm{elong,tl}}}{k_{\mathrm{a,RNAP}} \cdot (k_{\mathrm{init,tc}} + k^*_{\mathrm{elong,tl}})},$$

which was used in the FlgM–FliA interactions model to account for the feedback of FliA on class 3 gene expression (see Table C.2 in Appendix C).

5.4 Results

The model was parameterized based on available *in vivo* data from *E. coli* or related bacteria, which applied to the majority of parameters. If *in vivo* data were not available, data were taken from *in vitro* measurements. By dint of the above reduction process, only four parameters related to the effective synthesis rates of FlgM and FliA had to be estimated based on our experimental measurements [195]. The parameterization of the model is described in detail in Appendix C. In the following, we present and discuss the results obtained from different simulations performed with the model.

5.4.1 Comparison to Experimental Data

Wild type: Figures 5.3A–C show the experimental data (mean ± one standard deviation) of total intra- and extra-cellular FlgM and total FliA. Experimental data are given in number of molecules, as determined from Western blots [195]. The solid line shows predictions of the FlgM–FliA regulatory model based on the parameter values listed in Table C.2. Since the experimental measurements only allow a first rough estimate for the start of class 2 expression

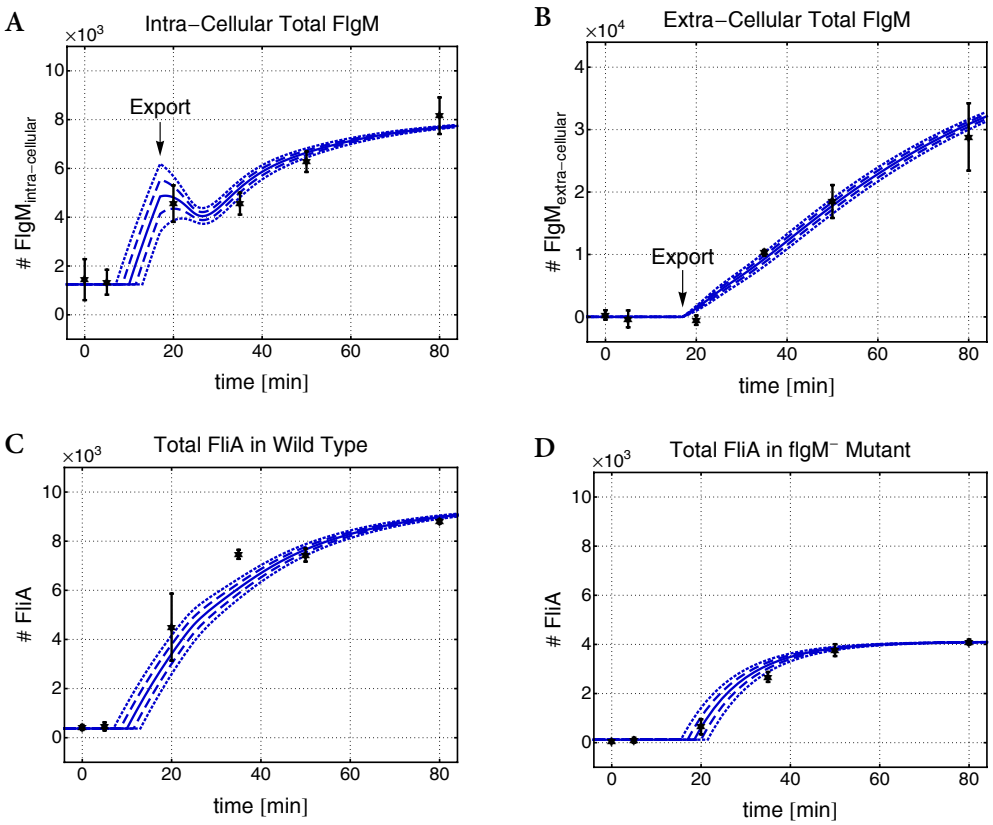

Figure 5.3: *In silico predictions compared to in vivo measurements of FlgM and FliA. Mean of experimental data marked by stars; bars indicate ± one standard deviation. Wild type: (A) intra-cellular total FlgM (free FlgM plus FlgM:FliA), (B) extra-cellular FlgM (relative to baseline level), and (C) total FliA (free FliA plus FlgM:FliA plus σ^F:RNAP) vs. time. flgM⁻ mutant: (D) total FliA (free FliA plus σ^F:RNAP) vs. time. Predictions are based on the parametrization in Table C.2 (solid line), and varied starting time of class 2 expression (otherwise identical parameters): t_{class2} = 7, 8.5, 10, 11.5, 13 min for the wild type and t_{class2} = 15.5, 17, 18.5, 20, 21.5 min for the flgM⁻ mutant.*

t_{class2} (between 5 and 20 min), model predictions are also shown for varying t_{class2} times and otherwise identical parameters: t_{class2} = 7 min (upper dotted line), t_{class2} = 8.5 min (upper dashed line), t_{class2} = 11.5 min (lower dashed line), t_{class2} = 13 min (lower dotted line). We observed that variations of t_{class2} were most pronounced initially and decreased over time.

The model predictions are in very good agreement with our experimental data [195] and other experimental findings [209]. For FliA, we observed a slight underestimation at 35 min. Upon start of class 2 expression at time t_{class2}, the levels of FlgM and FliA rapidly increase. After completion of the type III secretion system, FlgM is exported to the extra-cellular space, resulting in a continuous increase in extra-cellular FlgM (see Figure 5.3B). While FliA continues to increase (see Figure 5.3C), the model predicts a noticeable transient decrease in intra-cellular FlgM, until eventually newly synthesized FlgM molecules resulting from class 3 gene expression increase the level again (see Figure 5.3A).

flgM⁻ mutant: By design, a *flgM⁻* mutant lacks FlgM and results can only be evaluated in terms of FliA. Figure 5.3D shows the experimental data (mean ± one standard deviation) of total FliA, as determined from Western blots [195]. The solid line shows predictions based on the parameter values listed in Table C.2. As for the wild type, we further varied the starting time of class 2 gene expression (otherwise identical parameters): t_{class2} = 15.5 min (upper dotted line), t_{class2} = 17 min (upper dashed line), t_{class2} = 20 min (lower dashed line), t_{class2} = 21.57 min (lower dotted line). Again, the variation is most pronounced initially and decreases over time. The *in silico* predictions are in very good agreement with experimental data.

5.4.2 Validation of the Reduction Process

Having the effective synthesis rates of FlgM and FliA parameterized (as reported in Table C.2), we were also able to parameterize the underlying detailed model of gene expression (depict in Figure 5.2) on the basis of data available from literature and by exploiting the functional relations as given in eqs. (5.3)–(5.10). The parameterization of the detailed gene expression model is described in Appendix C; all parameter values are listed in Table C.3. This allowed to transform the biological model of the FlgM–FliA regulatory network (5.1) into a discrete-stochastic reaction kinetics model, where the class 2 and class 3 related syntheses of FlgM and FliA are now implemented according to the detailed transcriptional and translational processes shown in Figure 5.2.

In Figure 5.4, we compare the average mRNA and protein levels of FlgM and FliA in the wild type as predicted by 100 SSA-runs of this detailed stochastic model to the corresponding levels in the reduced deterministic model (5.2). Even though the transcription and translation related processes and variables are eliminated in the deterministic model, we can use eqs. (5.3)–(5.10) and the parameterization of the detailed model (see Table C.2) to recalculate their

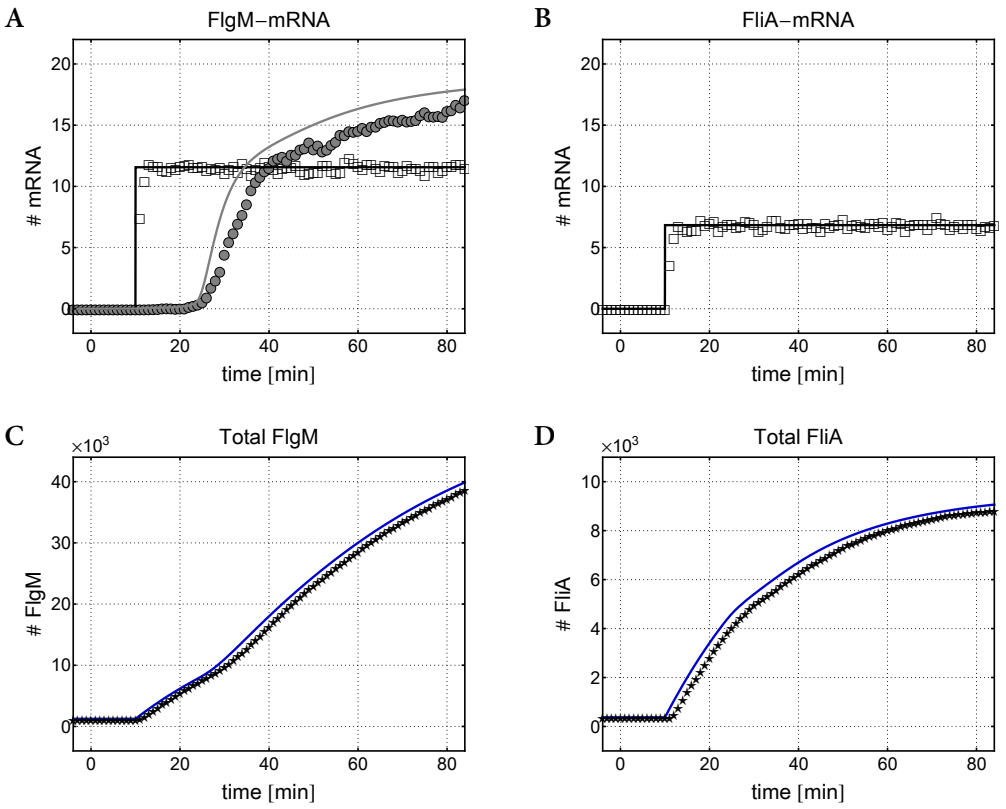

Figure 5.4: *Comparison of mRNA and protein levels in the detailed stochastic model (average of* 100 *SSA-runs, marked by symbols) and in the reduced deterministic model (solid lines) of the FlgM–FliA regulatory network. The panels show the results obtained for the wild type with initiation of class 2 expression at time* $t_{class2} = 10\,\text{min}$*: (A) mRNA of FlgM from class 2 (squares and black line) and class 3 (circles and gray line) transcription, (B) mRNA of FliA from class 2 transcription (squares and black line), (C) total FlgM (intra-cellular free FlgM plus FlgM:FliA plus extra-cellular FlgM), and (D) total FliA (free FliA plus FlgM:FliA plus* σ^F*:RNAP). In the stochastic model, protein synthesis is realized according to the detailed transcriptional and translational processes shown in Figure 5.3 (parameterization given in Table C.3). In the deterministic model, these processes are eliminated and protein synthesis is implemented by the derived effective rates (parameterization given in Table C.2), with mRNA levels given by eq.* (5.6).

values during simulation. The mRNA levels shown in Figure 5.4C+D, for instance, have been computed in this way from eq. (5.6).

The predictions of the reduced deterministic model and the detailed stochastic model are in very good agreement. In the reduced model, the mRNA and protein levels increase a bit earlier and faster than in the detailed model (see Figure 5.4). This effect is related to the time it takes the transcriptional and translational subprocesses to reach quasi-steady state, which, by assumption, is not rendered in the effective synthesis rates. Thus, synthesis is slightly delayed in the detailed but not in the reduced model. This becomes most pronounced in the approximation of class 3 synthesis of FlgM (see Figure 5.4A), as it (a) integrates the class 2 related approximation of FliA and (b) is initially much slower than class 2 synthesis (which is almost immediately saturated after initiation at time $t_{\text{class2}} = 10\,\text{min}$). The resulting deviations in the protein levels, however, are negligible (see Figure 5.4C+D).

5.4.3 Robustness and Timing of the Regulatory Mechanism

Class 3 Gene Expression is Induced only when Pool of Free FlgM is Drastically Reduced

We further studied in detail the checkpoint mechanism, i.e., the feedback of completion of the type III secretion system on class 3 gene expression. In contrast to the experimental

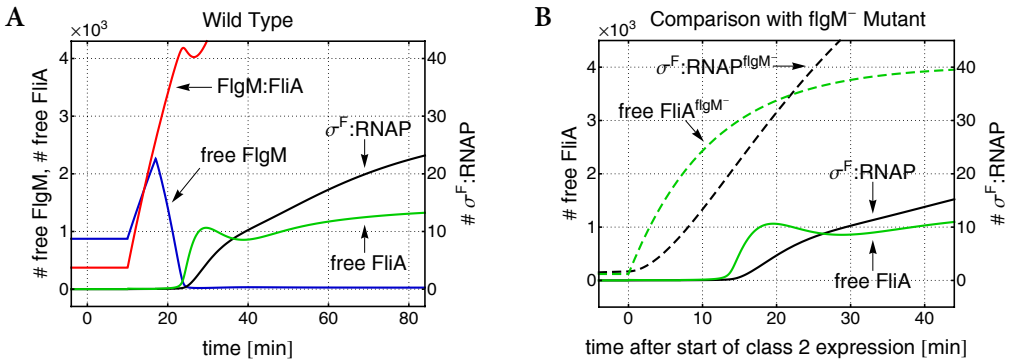

Figure 5.5: *Detailed analysis of the FlgM–FliA interactions: free FlgM (blue), free FliA (green), FlgM:FliA complex (red) and σ^F:RNAP complex (black). For the wild type (A), the most pronounced change in terms of number of molecules is the rapid decrease in free FlgM upon start of export, but not as one might expect in the level of FlgM:FliA. In the flgM⁻ mutant (B), no delay between class 2 and 3 expression is observed (dashed lines), as it is present in the wild type (solid lines).*

measurements, the mathematical model allows to distinguish between free and bound FliA, as well as to monitor the predicted σ^F:RNAP level in order to study the onset of class 3 expression. In Figure 5.5A, the predictions for FlgM, FliA and FlgM:FliA (left axis), and σ^F:RNAP (right axis) in the wild type are shown.

Experimentally, it has been demonstrated that FlgM is exported from the FlgM:FliA complex with FliA acting as a type III secretion chaperone [208]. *In silico*, the most pronounced change in terms of number of molecules is the rapid decrease in free FlgM upon completion of the secretion system around 18 min, but not as one might intuitively expects in the level of FlgM:FliA. A closer look at the key reactions resolves this observation: Since FliA is released from the complex when FlgM is exported, the availability of free FliA significantly increases. However, due to the high affinity of FliA for FlgM, it immediately binds again to free FlgM. Hence, initially the FliA-mediated export of FlgM effectively decreases the level of free FlgM, with FlgM:FliA remaining at high level, but having a very short life span and being 'produced just in time' for the export.

The tight balance of FlgM–FliA association, FlgM export and FliA release results in low levels of free FliA during the first minutes of export (see Figure 5.5B). With continued export of FlgM this balance is changed towards increasing levels of free FliA, which eventually results in the formation of σ^F:RNAP and initiation of class 3 gene expression. As noted before, the number of FlgM transiently decreases upon completion of the type III secretion system. Based on the above analysis, we may now associate this transient decay with the decrease in the pool of *free* FlgM. Only if this pool has been strongly reduced, class 3 gene expression gets initiated.

In Figure 5.5B, the timing of wild type and *flgM*$^-$ mutant is compared. Due to lack of FlgM in the mutant, rising FliA levels immediately initiate class 3 expression. Thus, the delay between class 2 and class 3 expression as observed in the wild type is absent in the *flgM*$^-$ mutant (as also observed experimentally [195]).

Based on the reduced deterministic model we analyzed the robustness and sensitivity of the FlgM–FliA regulatory checkpoint mechanism in its ability to tightly control σ^F:RNAP level, which was considered as the critical marker for initiation of class 3 gene expression. Wild type levels were compared to levels of different *in silico* mutants with 'perturbed' parameter values, in order to access the robustness with respect to such alterations. We were interested in both, the impact on the delay between class 2 and class 3 gene expression, as well as in changes in the intensity of class 3 initiation defined by how fast σ^F:RNAP levels increase in comparison to the wild type.

Checkpoint is Robust to Perturbations in FlgM–FliA Association and Dissociation

Levels of σ^F:RNAP for wild type and *in silico* mutants with altered FlgM:FliA dissociation rate constant (1000-fold decrease, and 100-, 500- and 1000-fold increase) are shown in Figure 5.6A and inset. Changes of up to three orders of magnitude have only marginal influence on σ^F:RNAP levels, thus rendering the regulatory network robust with respect to the binding affinity of FlgM and FliA.

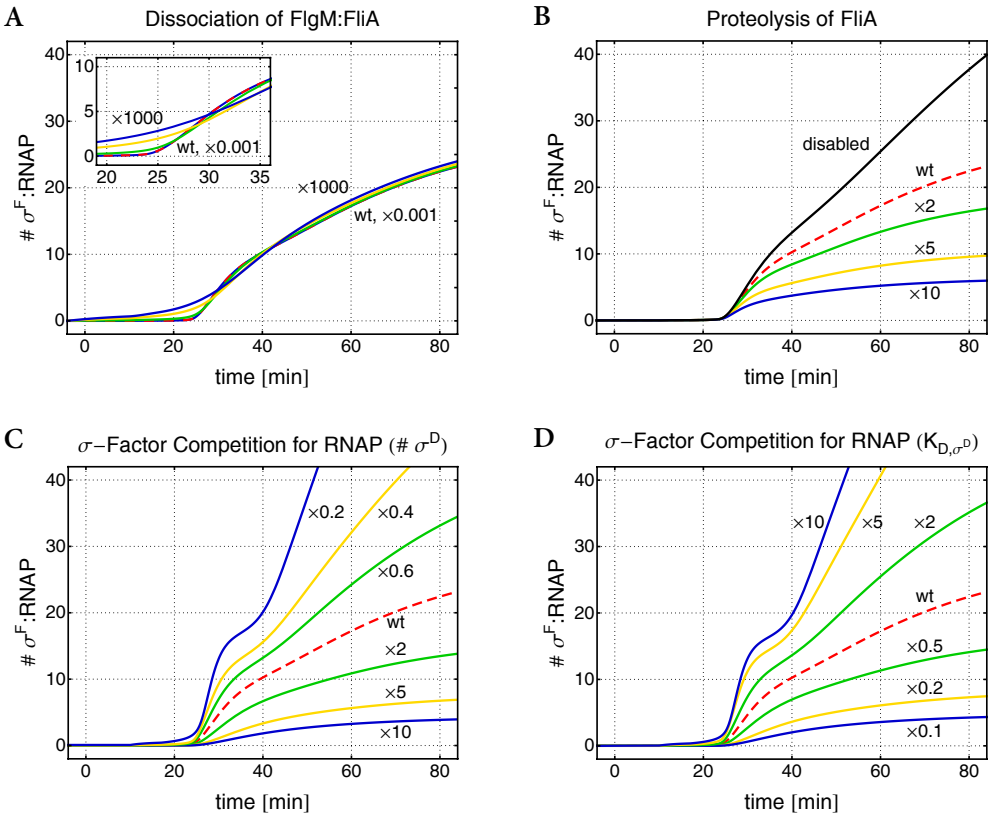

Figure 5.6: *Predicted σ^F:RNAP levels for the wild type (red, dashed line) and different in silico mutants (solid lines). (A) Dissociation constant of the FlgM:FliA complex:* 1000*-fold decrease (dashed blue), and* 100*-fold (green),* 500*-fold (yellow) and* 1000*-fold (solid blue) increase. (B) Rate of FliA proteolysis:* 2*-fold (green),* 5*-fold (yellow) and* 10*-fold (blue) increase plus proteolysis disabled (black). (C) Total level of the sigma factor σ^D, from highest to lowest levels:* 0.2*-fold (blue),* 0.4*-fold (yellow) and* 0.6*-fold (green) decrease, and* 2*-fold (green),* 5*-fold (yellow) and* 10*-fold (blue) increase. (D) Dissociation constant of the σ^D:RNAP complex, from highest to lowest levels:* 10*-fold (blue),* 5*-fold (yellow) and* 2*-fold (green) increase, and* 0.5*-fold (green),* 0.2*-fold (yellow) and* 0.1*-fold (blue) decrease.*

FliA Proteolysis and Sigma Factor Competition for RNAP Modulate Intensity of Class 3 Gene Expression

The intensity of class 3 gene expression is directly related to the available level of σ^F:RNAP. Figure 5.6B shows the changes in σ^F:RNAP levels resulting from an increase in the FliA (σ^F) proteolysis rate constant. A higher proteolysis rate results in lower levels of σ^F and consequently decreases the level of available σ^F:RNAP. In contrast, absence of proteolysis resulted in slightly increased levels of σ^F:RNAP.

In silico experiments that directly alter the competition of sigma factors for RNAP revealed a similar relation. In Figure 5.6C, we analyzed the robustness with respect to changes in the level of σ^D competing for RNAP. We infer that the level of σ^D is negatively correlated to the intensity of class 3 gene expression: Higher values of σ^D decrease the intensity of class 3 expression, whereas lower values of σ^D increase the intensity. Analogous results can be observed for alterations in the dissociation constant $K_{D,\sigma^D} = k_d/k_{a,\sigma^D}$ of the σ^D:RNAP complex. The larger K_{D,σ^D} the larger the intensity of class 3 expression, and vice versa (see Figure 5.6D).

In all three cases, the considered *in silico* settings modulate the steepness of increase in σ^F:RNAP levels and hence the intensity of class 3 expression. Most notably when altering the sigma factor competition for RNAP, to which the regulatory mechanism is very sensitive.

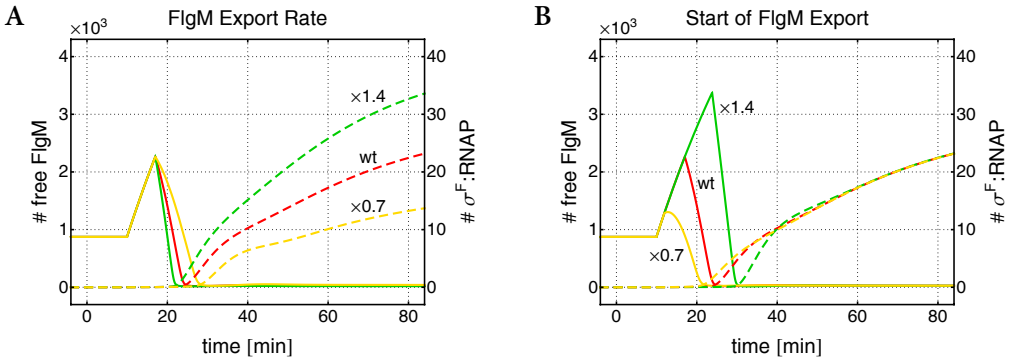

Figure 5.7: *Predicted levels of free FlgM (solid lines, left scale) and σ^F:RNAP (dashed lines, right scale) for the wild type (red) and different in silico mutants (yellow and green). (A) Effective FlgM export rate constant k_{export} and (B) starting time t_{export} of FlgM export, i.e., completion of type III secretion systems:* 0.7*-fold (yellow),* 1*-fold (red) and* 1.4*-fold (green) change.*

Pool of Free FlgM Precisely Controls Free FliA Level and Acts as a Molecular Timer

From the results shown in Figure 5.5 we inferred that class 3 initiation is coupled to the reduction of the pool of accumulated free FlgM. Since the timing of this reduction directly depends on the FlgM export rate k_{export} and the time t_{export} when the type III secretion system is completed and export starts, we expected a change in the initiation of class 3 gene expression due to alterations in these two parameters. As illustrated in Figure 5.7, the checkpoint mechanism is indeed sensitive to k_{export} and t_{export}. Perturbations in these two parameters affect both the point in time when the increase in σ^F:RNAP starts and the rate at which σ^F:RNAP increases.

In addition to σ^F:RNAP levels, the levels of free FlgM are shown in Figure 5.7. Changes in the above processes have a direct influence on the accumulation of free FlgM and the decay of the FlgM pool. As can be nicely seen in Figure 5.7, the number of σ^F:RNAP complexes does not start to increase until excessive *free* FlgM is exported from the cell. In this sense, the intra-cellular pool of free FlgM acts as a *molecular timer* that precisely controls the start of σ^F:RNAP formation and thus initiation of class 3 expression.

High Sensitivity of Effective Synthesis Rates does Not Necessarily Imply High Sensitivity with respect to Parameters of the Subsumed Transcriptional and Translational Processes

Changes in the synthesis rate of FlgM or FliA alter the ratio between their levels in the system, thus re-weighting the pool of free FlgM and its function as a molecular timer. Therefore, we expected that the start of class 3 gene expression will be affected by changes in the effective synthesis rates k_{FlgM} and k_{FliA}, shown in Figure 5.8A+B. An increased synthesis rate of FlgM resulted in a larger pool of free FlgM, thus increasing the delay between start of export and class 3 initiation (1.4-fold change, green). Contrarily, a decreased synthesis rate of FlgM resulted in a shorter delay (0.7-fold change, yellow). The opposite relation holds for FliA. A higher FliA synthesis rate increased the level of FlgM:FliA complex and thus effectively diminished the pool of *free* FlgM, which shortens the delay between start of export and class 3 initiation (1.4-fold change, green). Again, the opposite occurred for a decreased synthesis rate of FliA (0.7-fold change, yellow). The results illustrated in Figure 5.8A+B suggest that the checkpoint mechanism is very sensitive to changes in the effective synthesis rates. Since alterations of 30–40 % might be expected on a population level, this high sensitivity of the regulatory network was unexpected.

Since the synthesis rates of FlgM and FliA are effective rates subsuming complex reaction events of gene transcription and translation (see Section 5.3), we subsequently analyzed the sensitivity of the regulatory network to changes in the underlying transcriptional and translational parameters. Based on eqs. (5.3)–(5.10), we studied whether the same sensitivity on σ^F:RNAP levels can be observed when altering the parameters of the processes that are aggregated in the effective synthesis rates. All parameter values of the detailed transcription and translation model are listed in Table C.3.

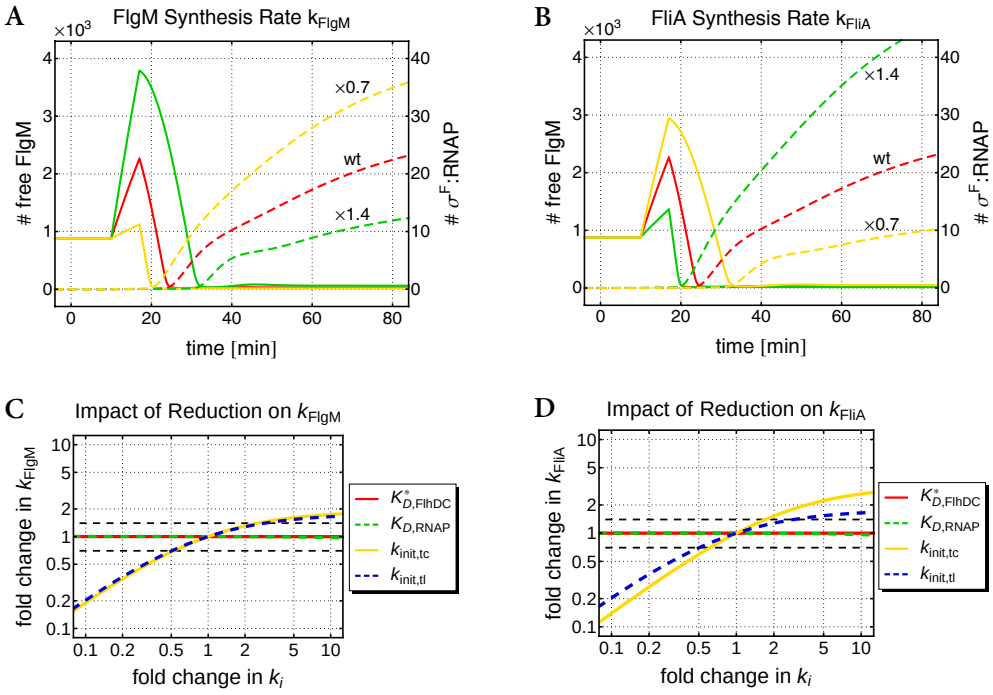

Figure 5.8: *Sensitivity of the core regulatory mechanism to the effective synthesis rates of FlgM and FliA. Top: Predicted levels of free FlgM (solid lines, left scale) and σ^F:RNAP (dashed lines, right scale) for the wild type (red) and different in silico mutants (yellow and green). (A) FlgM synthesis rate k_{FlgM} and (B) FliA synthesis rate k_{FliA}: 0.7-fold (yellow), 1-fold (red) and 1.4-fold (green) change. Bottom: Log-log plot of changes in the effective synthesis rates of FlgM (C) and FliA (C) with respect to changes in subsumed rate constants of the detailed class 2 transcriptional and translational processes. These are $K^*_{D,FlhDC} = K_{D,FlhDC}/FlhDC$, i.e., the ratio between the dissociation constant $K_{D,FlhDC}$ of the master regulator FlhDC for DNA-binding and its total level; the dissociation constant $K_{D,RNAP} = k_{d,RNAP}/k_{a,RNAP}$ of the σ^D:RNAP complex to the class 2 promoter; and the initiation rate constants of transcription $k_{init,tc}$ and translation $k_{init,tl}$. The dashed black horizontal lines mark the 0.7-fold and 1.4-fold changes in k_{FlgM} and k_{FliA} considered in (A) and (B).*

As can be inferred from Figure 5.8C+D, a high sensitivity of σ^F:RNAP levels to alterations in the effective synthesis rates k_{FlgM} and k_{FliA} does not necessarily translate into a high sensitivity with respect to alterations in the subsumed parameters of gene expression. The dashed black horizontal lines correspond to 0.7- and 1.4-fold changes in the effective synthesis rates of FlgM and FliA. We infer that a 10-fold decrease or increase in the dissociation constant $K_{D,RNAP} = k_{d,RNAP}/k_{a,RNAP}$ of the σ^D:RNAP complex or in the ratio $K^*_{D,FlhDC} = K_{D,FlhDC}/\mathrm{FlhDC}$ for the master regulator FlhDC does not have any significant impact on the effective synthesis rates of FlgM or FliA. We further infer that a 10-fold increase in the initiation rate constants $k_{init,tc}$ and $k_{init,tl}$ of class 2 transcription and translation, respectively, does only result in a roughly 2-fold increase in the effective synthesis rates. The most pronounced impact was observed when reducing the above initiation rate constants by a factor of 10, resulting in a similar decrease in the effective synthesis rates k_{FlgM} and k_{FliA}. All these subprocesses hence operate close to saturation, which becomes only apparent on the detailed level and can not be inferred from the corresponding lumped rates itself.

5.5 Discussion

We analyzed the FlgM–FliA regulatory network based on a deterministic model and quantitative experimental data. To incorporate the feedback between middle and late gene expression, we derived protein synthesis rates from a detailed stochastic model of the key transcriptional and translational processes. Our reduction approach effectively eliminated these discrete-stochastic submodels. The resulting functional relationship between the effective synthesis rates and the detailed parameters allowed for a careful sensitivity analysis of the system.

The *in silico* levels of intra- and extra-cellular FlgM in the wild type as well as the levels of FliA in the *flgM*$^-$ mutant are in excellent agreement with the *in vivo* data (Figure 5.3). For the wild type, the model slightly underestimates the experimentally observed steep increase in FliA after 35 min (Figure 5.3C). In addition to the FlgM–FliA checkpoint, there exists a number of interlocking positive and negative feedback loops with the potential to further modulate class 2 and class 3 expression [200, 220, 221]. While we observed no impact when reducing FlhDC concentration by a factor of 10 (Figure 5.8), a more significant decrease will eventually slow down class 2 gene expression. From eqs. (5.4) and (5.5) we infer that K_{tc} increases with decreasing FlhDC, such that eventually k_{mRNA} will be proportional to $\mathrm{FlhDC}/(\mathrm{FlhDC} + K_{D,FlhDC})$. For small enough levels, class 2 expression will thus be con-

trolled by FlhDC. This switch like dependence of class 2 expression on FlhDC concentration might also serve as an explanation for the observed difference in t_{class2} between wild type and *flgM*$^-$ mutant, since in the latter case FlhDC concentration might already be subject to FliA-controlled negative regulation.

A feedback loop with similar characteristics as the FlgM–FliA checkpoint mechanism is the FliT–FliD regulatory system [220, 221]. FliT is the secretion chaperone for the filament capping protein FliD. Upon completion of the type III secretion system, FliD is secreted to the tip of the hook where it facilitates polymerizaton of the flagellar filament [221]. The depletion of FliD from the cytoplasm eventually results in increased levels of free FliT. Free FliT subsequently binds to the FlhC subunit of FlhDC and thereby inhibits transcription of the middle genes whose products are no longer required for the assembly process [220]. We expect that incorporation of the FliT–FliD feedback in the model will result in a larger class 2 synthesis rate of FliA in the parameter estimation process, and thus in a stronger initial increase in FliA levels during class 2 expression. Increasing levels of free FliT after export of FliD would then subsequently slow down class 2 gene expression. Importantly, the FliT–FliD regulatory system is expected to exhibit the same characteristics as the FlgM–FliA checkpoint mechanism. The herein presented analysis and results can thus serve as a starting point for future experimental design and theoretical studies.

Sensitivity analysis of the FlgM–FliA regulatory mechanism revealed that the system is robust to alterations in most of the parameters (Figures 5.6A and 5.8C+D). At the same time, it is very sensitive to alterations in those input signals that are exploited by *E. coli* to adapt and tune flagellar synthesis in face of a changing environment. These correspond either to parameters that allow the cell to tune initiation of class 3 expression, e.g., in terms of strength or start of export (Figure 5.7), or that serve as the entry point of other master regulators. The increase in σ^F:RNAP is most notably affected by sigma factor competition (Figure 5.6C+D). This tuning point allows for a direct, efficient and instantaneous alteration of flagellar synthesis, which is important, for instance, in the transition from the motile-planktonic to the stationary phase 'lifestyle', induced by the sigma factor σ^S [222, 223].

Before successful completion of the first type III secretion systems and subsequent start of FlgM export, FliA is sequestered in the FlgM:FliA complex in its inactive form. Since only free FliA can bind to the core enzyme RNAP, control of initiation of class 3 expression is implemented by maintaining a certain ratio between FlgM and FliA. Upon start of export, the pool of free FlgM is gradually degraded until a change in the FlgM–FliA ratio results in sufficient many free FliA molecules to enable class 3 initiation (Figure 5.5A). In the absence of

FlgM, this delay is not present and class 3 expression is closely following the increasing FliA levels (Figure 5.5B). In [199] experimental evidence is given that when pre-existing flagella are present, newly synthesized FlgM is already exported before new basal bodies have been completed. In the proposed model, an increase in the number of successfully completed hook basal bodies would correspond to an increase in the FlgM export rate. As shown in Figure 5.7, this directly decreases the delay between start of export and class 3 initiation, as less free FlgM is accumulated. The dynamic control of the pool of free FlgM analyzed herein could serve as a mechanistic explanation of the above experimental observation, highlighting the fact that the relative ratio of FlgM to FliA is important for the functionality of the checkpoint mechanism.

The σ^F:RNAP level also showed a high sensitivity to changes in the effective synthesis rates of FlgM and FliA (Figure 5.8A+B). This theoretical finding was unexpected, since even small alterations of 30–40 % would imply a large impact on the timing and intensity of class 3 gene expression. A sensitivity analysis with respect to changes in parameters of the detailed transcription and translation model (Figure 5.2) underlying the effective synthesis rates gave further insight: Alterations in the detailed processes do not necessarily imply alterations of the same order of magnitude in the effective synthesis rates. We find that the initiation rate constants of transcription and translation show the highest sensitivity, but only when reduced (Figure 5.8C+D). This suggests that potentially large variations in the flagella synthesizing population (as suggested by the reduced model) are actually avoided by operating transcription and translation of *fliA* and *flgM* close to saturation. On the other hand, sensitivity against reduced rates of synthesis can effectively be used to integrate other stress signals.

In general, the functional relationships between k_{FlgM} and k_{FliA} and the detailed parameters as given by eqs. (5.3)–(5.10) and depict in Figure 5.8C+D illustrate that an observed high sensitivity with respect to effective rates does not directly carry over to the detailed parameters, it breaks down into more diverse relations. This phenomenon is not restricted to the present analysis. As a consequence, care has to be taken when interpreting results of robustness analysis of general reaction kinetic models with respect to lumped or effective parameters.

Motility and biofilm formation in *E. coli* are both under control of regulatory feed forward cascades with mutual interaction and cross-regulation at different levels [197, 223]. In this context the derived model also constitutes a first step towards a more comprehensive model of life style adaptation in *E. coli*. While experimental data on gene transcription or promoter activity may provide further insight into the temporal hierarchy of gene expression, more quantitative data in terms of number of molecules combined with mathematical modeling is needed to analyze the regulatory processes on the protein interactions level.

6

Summary & Conclusion

In this thesis, we presented different approaches that couple the discrete–stochastic with the continuous–deterministic formulation of biochemical reaction networks. We demonstrated how such hybrid approaches can be employed for the efficient simulation and thorough analysis of systems of biological interest, such as viral dynamics and gene regulatory networks.

While current hybrid methods are almost exclusively *indirect*, i.e., relying on the computation of ensembles of stochastic realizations, we were able to derive and theoretically justify a novel hybrid stochastic–deterministic approach to solve the CME *directly*. In our direct hybrid approach presented in Chapter 4, all problematic species that are present in large numbers are effectively eliminated from the discrete state space by replacing them with related expected levels. This approach can be interpreted as taking 'partial' expectations over the large copy number subspace. In contrast to an indirect hybrid approach, the resulting evolution equations explicitly incorporate the impact of changes in the singled out discrete distribution on the expectations of continuous variables. We provided the conditions under which such partial averaging of the CME is exact and, most importantly, by using multi-scale expansion techniques, we further demonstrated that it can be generalized to approximate the dynamics of more complex systems. A direct solution method does not suffer from the disadvantages of an indirect, MC-based approach. We demonstrated the performance of our direct hybrid stochastic–deterministic approach on a viral infection kinetics model and a transcriptional regulatory network.

The numerical costs of solving the derived hybrid model equations scale with the size of the singled out discrete subspace associated with species present in low copy numbers. Hence, we expect our direct hybrid solution to be particularly suitable for networks including a few 'discrete' species only. Otherwise, an indirect hybrid method may be applied, such as the deterministic–stochastic simulation algorithm presented in Section 3.1. We believe that the application of this method to the HIV-dynamics model studied in Section 3.2 clearly illustrated the advantages obtained by a hybrid system representation: Neither a purely stochastic (due to numerical complexity) nor a purely deterministic model (due to lack of stochastic effects)

would have allowed to perform our *in silico* study. Using hybrid simulations, however, we were able to design and test (in terms of the mathematical model) a novel treatment strategy that can lead to significant improvements in comparison to conventional treatment strategies. We discussed and demonstrated superiority of our pro-active treatment switching strategy for different scenarios; nevertheless, the clinical validity of our finding remains to be shown.

If a purely deterministic formulation is found to be adequate or unavoidable (by reasons of model and/or numerical complexity), our results presented in Section 5.4 clearly illustrated that nonetheless care has to be taken when interpreting results of a sensitivity analysis with respect to lumped or effective parameters. The reduction of the detailed stochastic model of gene expression performed in Section 5.3 not only allowed us to derive effective protein synthesis rates as typically used in deterministic models, but also made very explicit the assumptions underlying such rates. Effectively, we eliminated the transcriptional and translational subprocesses, but disclosed their functional relationship to the effective synthesis rates. A step towards an intermediate model of gene expression would be to explicitly incorporate mRNA species. Since we considered the processes involved in transcription and translation separately, the presented derivation can be directly used to formulate effective mRNA synthesis rates and rates of mRNA translation.

We successfully demonstrated our reduction approach on the deterministic FlgM–FliA interactions model presented in Section 5.2. The proposed derivation of effective synthesis rates, enabled us to mechanistically incorporate the feedback between middle and late gene expression in the deterministic model. Thus, for the first time, we could perform a detailed quantitative *in silico* study of the timing and robustness of the FlgM–FliA checkpoint mechanism in the flagellar gene regulation cascade of *E. coli*. Sensitivity analysis pointed out that the system is robust to most of the parameters, but still very sensitive to those input signals that are exploited by *E. coli* to tune flagellar synthesis in view of a changing environment. Since similar or analogous checkpoint mechanisms are present in many other motile bacteria, we expect our findings to have implications beyond the present study.

It is evident that stochastic effects play a crucial role in biological processes, requiring efficient mathematical modeling approaches that are capable to capture these effects. One of the most promising directions is the development of hybrid methods, as these are flexible enough to study stochasticity in large and ever more complex biological models that integrate processes of different scales like gene expression, regulation and metabolism [92]. In this thesis, we introduced the concept of a new hybrid methodology to directly couple the CME with the

ODE formulation of biochemical reaction networks. There are some important open questions related to such kind of direct hybrid models that will be subject of future research. These include: Where are the limits of our hybrid approach, i.e., if there are species with moderate copy numbers in between low and large copy numbers, when will their fluctuations destroy the approximation quality of the model? In order to decide whether the asymptotic assumptions underlying our hybrid model are valid, how can we estimate the value of the scaling parameter ε for a given chemical reaction network? How can one construct an efficient and robust numerical scheme that allows to adaptively change the low copy number subspace on the fly during numerical integration based on some prescribed accuracy requirements? We believe that the mathematical framework and fundamental understanding presented in this work provides essential tools to tackle such problems.

Appendix

A

Description of the HIV-Dynamics Model

In [138], we introduced a novel HIV-dynamics model of viral replication, mutation and drug interference that allows the mechanistic integration of all novel and some developmental HIV drugs. We extended this model for the studies presented in Section 3.2 by the compartment of very long lived, latently infected T-cells $\mathrm{T_L}$. The resulting model comprises uninfected T-cells $\mathrm{T_U}$ and macrophages $\mathrm{M_U}$, free non-infectious virus $\mathrm{V_{NI}}$, free infectious virus $\mathrm{V_I}(i)$ of mutant strain $i = 1, 2, \ldots$, and infected cell types belonging to each mutant strain i: infected T-cells $\mathrm{T_1}(i)$ and macrophages $\mathrm{M_1}(i)$ *prior* to proviral genomic integration, and infected T-cells $\mathrm{T_2}(i)$ and $\mathrm{T_L}(i)$ and macrophages $\mathrm{M_2}(i)$ *after* proviral genomic integration. The deterministic formulation of the model is given by the following system of ODEs:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{T_U} &= \lambda_\mathrm{T} + \delta_\mathrm{PIC,T} \cdot \mathrm{T_1}(i) - \Big(\delta_\mathrm{T} + \sum_i \beta_\mathrm{T}(i,d) \cdot \mathrm{V_I}(i)\Big) \cdot \mathrm{T_U}, \\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{M_U} &= \lambda_\mathrm{M} + \delta_\mathrm{PIC,M} \cdot \mathrm{M_1}(i) - \Big(\delta_\mathrm{M} + \sum_i \beta_\mathrm{M}(i,d) \cdot \mathrm{V_I}(i)\Big) \cdot \mathrm{M_U}, \\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{T_1}(i) &= \beta_\mathrm{T}(i,d) \cdot \mathrm{V_I}(i) \cdot \mathrm{T_U} - \Big(\delta_{\mathrm{T_1}} + \delta_\mathrm{PIC,T} + k_\mathrm{T}(i,d)\Big) \cdot \mathrm{T_1}(i), \\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{M_1}(i) &= \beta_\mathrm{M}(i,d) \cdot \mathrm{V_I}(i) \cdot \mathrm{M_U} - \Big(\delta_{\mathrm{M_1}} + \delta_\mathrm{PIC,M} + k_\mathrm{M}(i,d)\Big) \cdot \mathrm{M_1}(i), \\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{T_2}(i) &= \sum_j (1-p) \cdot k_\mathrm{T}(j,d) \cdot \mathrm{T_1}(j) \cdot r_{j\to i} + \alpha \cdot \mathrm{T_L}(i) - \delta_{\mathrm{T_2}} \cdot \mathrm{T_2}(i), \\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{T_L}(i) &= \sum_j p \cdot k_\mathrm{T}(j,d) \cdot \mathrm{T_1}(j) \cdot r_{j\to i} - \Big(\alpha + \delta_{\mathrm{T_L}}\Big) \cdot \mathrm{T_L}(i), \\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{M_2}(i) &= \sum_j k_\mathrm{M}(j,d) \cdot \mathrm{M_1}(j) \cdot r_{j\to i} - \delta_{\mathrm{M_2}} \cdot \mathrm{M_2}(i), \\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{V_I}(i) &= N_\mathrm{T}(i,d) \cdot \mathrm{T_2}(i) + N_\mathrm{M}(i,d) \cdot \mathrm{M_2}(i) \\
&\quad - \Big(CL + \Big(CL_\mathrm{T}(i,d) + \beta_\mathrm{T}(i,d)\Big) \cdot \mathrm{T_U} + \Big(CL_\mathrm{M}(i,d) + \beta_\mathrm{M}(i,d)\Big) \cdot \mathrm{M_U}\Big) \cdot \mathrm{V_I}(i), \\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{V_{NI}} &= \sum_i \Big[\Big(\widehat{N}_\mathrm{T} - N_\mathrm{T}(i,d)\Big) \cdot \mathrm{T_2}(i) + \Big(\widehat{N}_\mathrm{M} - N_\mathrm{M}(i,d)\Big) \cdot \mathrm{M_2}(i)\Big] - CL \cdot \mathrm{V_{NI}}.
\end{aligned}
\tag{A.1}
$$

The model accounts for a production of uninfected target cells T_U and M_U by the immune system with constant rates λ_T and λ_M, respectively. All cell types get destroyed by the immune system, where the parameters δ_T, δ_M, δ_{T_1}, δ_{M_1}, δ_{T_2}, δ_{T_L} and δ_{M_2} denote the corresponding death rate constants of T_U, M_U, T_1, M_1, T_2, T_L and M_2 cells, respectively.

Target cells T_U and M_U can become successfully infected by infectious virus $V_I(i)$ of mutant strain i with rate constants $\beta_T(i,d)$ and $\beta_M(i,d)$, respectively, under treatment d, creating early infected cells $T_1(i)$ and $M_1(i)$. Infection by virus $V_I(i)$ can also be unsuccessful after the irreversible step of fusion, with rate constants $CL_T(i,d)$ and $CL_M(i,d)$ under treatment d, eliminating the virus and rendering the target cells T_U and M_U, respectively, uninfected. The parameters $\delta_{PIC,T}$ and $\delta_{PIC,M}$ refer to the intracellular degradation of essential components of the pre-integration complex, e.g., by the proteasome of early infected cells T_1 and M_1, respectively, returning these cells to an uninfected stage.

Viral genome of mutant strain i is irreversibly integrated into the DNA of infected T-cells T_1 and macrophages M_1 under treatment d with rate constant $k_T(i,d)$ and $k_M(i,d)$, respectively, creating post-integration, infected cells T_2, T_L and M_2. Latently infected T-cells T_L (created with probability p) do not express viral genes, but can become activated with rate constant α, transforming T_L into a productively infected T-cell T_2. The parameters $\widehat{N}_T$ and $\widehat{N}_M$ denote the number of *total* (infectious and non-infectious) virus released per day per late infected T-cell T_2 and macrophage M_2, respectively. $N_T(i,d)$ and $N_M(i,d)$ give the number of *infectious* virus $V_I(i)$ of mutant strain i released under treatment d per day per $T_2(i)$ and $M_2(i)$, respectively. Free infectious and non-infectious virus, V_I and V_{NI}, respectively, gets cleared by the immune system with rate constant CL. The parameter $r_{j\to i}$ denotes the probability to mutate from strain j to strain i, defined below.

Integration of Mutation Dynamics: The overall model comprises a complete mutagenic pathway. In HIV-dynamics, genomic mutation occurs during the reverse transcription process [224]. The reverse transcriptase of HIV lacks a proof reading mechanism in contrast to host polymerase enzymatic reactions. However, viral proteins are only produced from newly mutated viral genome after its integration into the host DNA, and the proteins required for stable integration originate from the founder virus. Phenotypically, drug-resistance of new mutants will thus only be observed after integration, i.e., in the infectious stages T_2 and M_2.

In general, the model includes a total number of 2^L different viral strains i, containing point mutations in any possible pattern of the considered number L of positions that may mutate. In Figure A.1A, a mutagenic pathway for the example of two distinct mutations $L = 2$

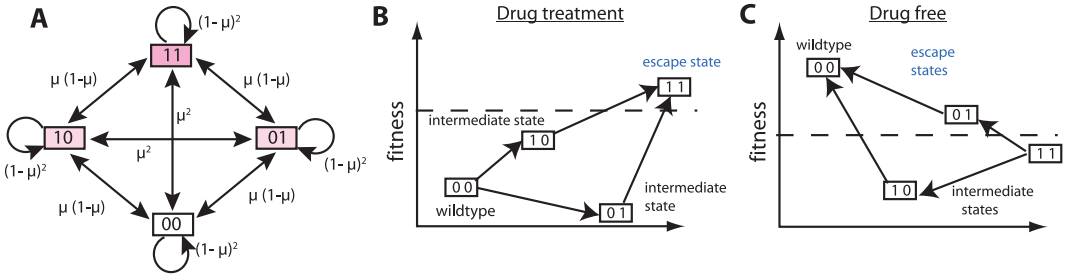

Figure A.1: *Fitness and possible mutational pathways in the HIV-dynamics model if two mutations are considered. A: General transition pathways between wild type (*00*) and a fully drug-resistant strain (*11*) that involves two part-resistant intermediates (*10, 01*). B: Fitness in the presence of a drug. C: Fitness in the absence of drugs. The dashed line indicates the critical fitness that allows the strain to survive (i.e.,* $R_0(i,d) = 1$*).*

is shown. Each mutant can mutate into every other mutant in one step. As can be seen from the example depict in Figure A.1A, the probability $r_{j\to i}$ to mutate from a strain j into another strain i is given by

$$r_{j\to i} := \mu^{h(i,j)} \cdot (1-\mu)^{Lh(i,j)}$$

where μ is the mutation probability per base pair during reverse transcription ($\mu \approx 2.2 \times 10^{-5}$, see [224]), $h(i,j)$ denotes the hamming distance between strain i and strain j, and L is the total number of different positions subject to mutation ($L = 6$ in the later numerical experiments).

Integration of Phenotypic Fitness and Drug Interference: The phenotype of each mutant strain i is modeled by introducing a selective disadvantage $s(i)$, which denotes the loss of functionality (e.g., in the activity of some viral enzyme that is affected by the mutation) relative to the wild type, and a strain specific inhibitory activity $\eta(i,d)$ of a treatment d against the mutant strain i. For example, the strain specific infection rate i under a certain treatment d is given by $\beta(i,d) = (1-\eta(i,d)) \cdot (1-s(i)) \cdot \beta(wt,\varnothing)$, where $\beta(wt,\varnothing)$ denotes the infection rate constant of the wild type in the absence of drug (parameters listed in Table A.1). The strain specific inhibitory activity is calculated via $\eta(i,d) = \eta(wt,d) \cdot \mathrm{res}(i,d)$, where the efficacy of the drugs against the wild type $\eta(wt,d)$ is generally stated in the corresponding tables and figures (Figures 3.2 and 3.5, and Table 3.1) and the resistance of a particular mutant $\mathrm{res}(i,d)$ was either set to 1 (100 % susceptible) or 0.01 (99 % resistant), if the particular mutant i conferred resistance to the particular drug d. For further details on the integration of the

Table A.1: *Parameter values of the HIV-dynamics model, referring to the wild type 'wt' in the absence of drug treatment '∅'. All parameters are in units day^{-1}, except p, ρ_{rev} and $b \cdot q \cdot \rho_{PR}$ (unit less), and μ in (rev. trans. · base)$^{-1}$. $CL_{T/M}(wt, \varnothing) = (1/\rho_{rev} - 1) \cdot \beta_{T/M}(wt, \varnothing)$ and $N_{T/M}(wt, \varnothing) = b \cdot q \cdot \rho_{PR} \cdot \widehat{N}_{T/M}$.*

Parameter	Value	Reference	Parameter	Value	Reference
λ_{T}	2×10^9	[226]	λ_{M}	6.9×10^7	[227]
$\delta_{\mathrm{T/T_1}}$	0.02	[227]	$\delta_{\mathrm{M/M_1}}$	0.0069	[227]
$\delta_{\mathrm{T_2}}$	1	[184]	$\delta_{\mathrm{M_2}}$	0.09	[138]
CL	23	[184]	$\delta_{\mathrm{T_L}}$	0.0001	[163, 228]
$\delta_{\mathrm{PIC,T}}$	0.35	[229, 230]	$\delta_{\mathrm{PIC,M}}$	0.0035	[138]
α	0.001	[228]	p	8×10^{-6}	[228]
μ	2.2×10^{-5}	[224]	ρ_{rev}	0.33	[230, 231]
$k_{\mathrm{T}}(wt, \varnothing)$	0.35	[230]	$k_{\mathrm{M}}(wt, \varnothing)$	0.07	[138]
$\beta_{\mathrm{T}}(wt, \varnothing)$	8×10^{-12}	[232]	$\beta_{\mathrm{M}}(wt, \varnothing)$	10^{-14}	[138]
$\widehat{N}_{\mathrm{T}}$	1000	[227]	$\widehat{N}_{\mathrm{M}}$	100	[227]
$b \cdot q \cdot \rho_{\mathrm{PR}}$	0.67	[138]			

distinct molecular effects of different drugs in the model, we refer the interested reader to the original publications [138, 225].

Parameterization of the Model

All parameter values have been chosen according to previous studies and are listed in Table A.1. The particular viral decay dynamics after application of distinct drug classes was validated in [138]. For the above parameter values, the model (A.1) reproduces an average frequency of latently infected cells of 2.6×10^{-5} $CD4^+$ T-cells (reference range: 8.2×10^{-7} to 2.05×10^{-4} $CD4^+$ T-cells [152, 154, 174, 175]), a total of 4.5×10^6 latently infected cells (reference: [154]), with a half-life of 20.6 months (average value reported in [152, 233–236]: 21 months) and a plasma viremia of ≈ 1 HIV RNA/mL [237] from the latent reservoir.

Reproductive Capacity and Reproductive Numbers

In [138], we introduced the *reproductive capacity* $R_{\mathrm{cap}}(d)$, that can be envisaged as the amount of infectious offspring that the whole viral population is expected to produce under some treatment d during one round of replication. It allows to evaluate the infection state from

the perspective of a potential treatment d during simulation and was utilized in Section 3.2 to obtain switching times that maximize the benefit of an initial induction regimen for the following maintenance regimen. The reproductive capacity $R_{\text{cap}}(d)$ of the entire quasi-species ensemble under treatment d is defined as the weighted sum of the basic reproductive numbers of all pathogenic stages of mutant strains [138], i.e.,

$$R_{\text{cap}}(d) = \sum_i \Big[\mathrm{V_I}(i) \cdot R_{\mathrm{V_I}}(i,d) + \mathrm{T_1}(i) \cdot R_{\mathrm{T_1}}(i,d) + \mathrm{M_1}(i) \cdot R_{\mathrm{M_1}}(i,d) + \mathrm{T_2}(i) \cdot R_{\mathrm{T_2}}(i,d) + \mathrm{T_L}(i) \cdot R_{\mathrm{T_L}}(i,d) + \mathrm{M_2}(i) \cdot R_{\mathrm{M_2}}(i,d) \Big],$$

where $R_{\mathrm{V_I}}(i,d)$, $R_{\mathrm{T_1}}(i,d)$, $R_{\mathrm{M_1}}(i,d)$, $R_{\mathrm{T_2}}(i,d)$, $R_{\mathrm{T_L}}(i,d)$ and $R_{\mathrm{M_2}}(i,d)$ are the strain-specific reproductive numbers of the different infective compartments.

The basic *reproductive number* R_0 is a well-characterized quantity in epidemiology, denoting the expected number of infections caused per infected individual/cell [238]. An infection will spread if $R_0 > 1$; it will die out if $R_0 < 1$. The strain-specific reproductive number $R_0(i,d)$ characterizes the fitness of a viral strain i in a pharmacologically modified environment, specified by a drug treatment d. This is illustrated in Figure A.1 for the example of two distinct mutations: Panel B shows the fitness landscape in the presence of some drug d, and panel C shows the fitness landscapes in drug absence; the critical fitness $R_0(i,d) = 1$ is indicated by dashed lines. We used the *survival function* approach [239] to calculate the strain-specific reproductive numbers, which is of particular value, since it captures the possible event of mutation for all infective compartments.

For the above model, the reproductive numbers are given as follows: The reproductive number $R_{\mathrm{V_I}}(i,d)$ of a single infective virus $\mathrm{V_I}(i)$ of strain i under treatment d is given by

$$R_{\mathrm{V_I}}(i,d) = \frac{\beta_{\mathrm{T}}(i,d) \cdot \mathrm{T_U} \cdot k_{\mathrm{T}}(i,d) \cdot \left(1 - p \cdot \frac{\delta_{\mathrm{T_L}}}{\delta_{\mathrm{T_L}} + \alpha}\right) \cdot N_{\mathrm{T}}(i,d)}{c_{\mathrm{V_I}}(i,d) \cdot c_{\mathrm{T}}(i,d) \cdot \delta_{\mathrm{T_2}}} + \frac{\beta_{\mathrm{M}}(i,d) \cdot \mathrm{M_U} \cdot k_{\mathrm{M}}(i,d) \cdot N_{\mathrm{M}}(i,d)}{c_{\mathrm{V_I}}(i,d) \cdot c_{\mathrm{M}}(i,d) \cdot \delta_{\mathrm{M_2}}},$$

with constants

$$c_{\mathrm{V_I}}(i,d) = CL + \left(CL_{\mathrm{T}}(i,d) + \beta_{\mathrm{T}}(i,d)\right) \cdot \mathrm{T_U} + \left(CL_{\mathrm{M}}(i,d) + \beta_{\mathrm{M}}(i,d)\right) \cdot \mathrm{M_U},$$
$$c_{\mathrm{T/M}}(i,d) = \delta_{\mathrm{T/M}} + \delta_{\mathrm{PIC,T/M}} + k_{\mathrm{T/M}}(i,d).$$

Infected cells are also pathogens, which can lead to a rebound of the disease even in the absence of any virus. The basic reproductive numbers $R_{\mathrm{T}_1}(i,d)$ and $R_{\mathrm{M}_1}(i,d)$ of the infectious stages $\mathrm{T}_1(i)$ and $\mathrm{M}_1(i)$, respectively, associated with viral strain i under treatment d are given by

$$R_{\mathrm{T}_1}(i,d) = \frac{k_{\mathrm{T}}(i,d)\cdot\left(1-p\cdot\frac{\delta_{\mathrm{T_L}}}{\delta_{\mathrm{T_L}}+\alpha}\right)\cdot N_{\mathrm{T}}(i,d)}{c_{\mathrm{T}}(i,d)\cdot\delta_{\mathrm{T}_2}}\cdot r_1(i,d)$$

and

$$R_{\mathrm{M}_1}(i,d) = \frac{k_{\mathrm{M}}(i,d)\cdot N_{\mathrm{M}}(i,d)}{c_{\mathrm{M}}(i,d)\cdot\delta_{\mathrm{M}_2}}\cdot r_1(i,d),$$

where

$$r_1(i,d) = \frac{\beta_{\mathrm{T}}(i,d)\cdot\mathrm{T_U}+\beta_{\mathrm{M}}(i,d)\cdot\mathrm{M_U}}{c_{\mathrm{V_I}}(i,d)}.$$

Finally, the reproductive numbers $R_{\mathrm{T}_2}(i,d)$, $R_{\mathrm{T_L}}(i,d)$ and $R_{\mathrm{M}_2}(i,d)$ of the infectious stages $\mathrm{T}_2(i)$, $\mathrm{T_L}(i)$ and $\mathrm{M}_2(i)$ associated with viral strain i under treatment d are given by

$$R_{\mathrm{T}_2}(i,d) = \frac{N_{\mathrm{T}}(i,d)}{\delta_{\mathrm{T}_2}}\cdot r_2(i,d), \qquad R_{\mathrm{T_L}}(i,d) = \frac{\alpha}{\delta_{\mathrm{T_L}}+\alpha}\cdot R_{\mathrm{T}_2}(i,d)$$

and

$$R_{\mathrm{M}_2}(i,d) = \frac{N_{\mathrm{M}}(i,d)}{\delta_{\mathrm{M}_2}}\cdot r_2(i,d),$$

where

$$r_2(i,d) = \left(\frac{k_{\mathrm{T}}(i,d)\cdot\mathrm{T_U}\cdot\beta_{\mathrm{T}}(i,d)}{c_{\mathrm{V_I}}(i,d)\cdot c_{\mathrm{T}}(i,d)}+\frac{k_{\mathrm{M}}(i,d)\cdot\mathrm{M_U}\cdot\beta_{\mathrm{M}}(i,d)}{c_{\mathrm{V_I}}(i,d)\cdot c_{\mathrm{M}}(i,d)}\right).$$

B

Laplace's Method of Integral Approximation

Laplace's method [240] is a general technique used for the asymptotic approximation of integrals of the form

$$I_\Omega := \int_R f(x) \mathrm{e}^{\Omega\phi(x)} \,\mathrm{d}x \quad \text{as} \quad \Omega \to \infty,$$

called Laplace integrals, where f and ϕ are real-valued, continuous functions defined on some region $R \subseteq \mathbb{R}^d$, such that I_Ω is absolutely convergent for some value of $\Omega \in \mathbb{R}^+$. For simplicity, we consider in the following only the scalar case $d = 1$, such that $R =: [a; b]$ is some interval of the extended real line. We further assume that f and ϕ are sufficiently smooth functions. Analogous results also hold for higher dimensions and under weaker assumptions on f and ϕ, cf. [241–244]. The outline below mainly follows corresponding sections in [65, 245].

Laplace's method is based on the observation that if the continuous function ϕ has a unique absolute maximum on R, at $x = \tilde{x}$, say, and $f(\tilde{x}) \neq 0$, then only the immediate neighborhood around $\tilde{x}$ originates to the full asymptotic expansion of I_Ω. This follows from the fact that $\mathrm{e}^{\Omega\phi(x)}$ decays exponentially rapidly away from $\tilde{x}$, because $\phi(x) < \phi(\tilde{x})$ for all $x \neq \tilde{x}$ in R, and thus $\mathrm{e}^{\Omega\phi(x)} = \mathrm{e}^{-\Omega(\phi(\tilde{x})-\phi(x))}\mathrm{e}^{\Omega\phi(\tilde{x})} \ll \mathrm{e}^{\Omega\phi(\tilde{x})}$ as $\Omega \to \infty$.[1] Hence, we may write the integral I_Ω as

$$I_\Omega = \mathrm{e}^{\Omega\phi(\tilde{x})} \int_R f(x) \mathrm{e}^{\Omega\tilde{\phi}(x)} \,\mathrm{d}x \quad \text{as} \quad \Omega \to \infty,$$

where $\tilde{\phi}(x) := \phi(x) - \phi(\tilde{x})$ with $\tilde{\phi}(\tilde{x}) = 0$ and $\tilde{\phi}(x) < 0$ for all $x \neq \tilde{x}$ in R. In the following, we study the asymptotic expansion of $\tilde{I}_\Omega$ given by

$$\tilde{I}_\Omega := \int_R f(x) \mathrm{e}^{\Omega\tilde{\phi}(x)} \,\mathrm{d}x \quad \text{as} \quad \Omega \to \infty.$$

We note that if I_Ω converges absolutely for some value of Ω, say Ω_0, then so does $\tilde{I}_\Omega$, because $|\tilde{I}_{\Omega_0}|$ will differ from $|I_{\Omega_0}|$ only by the constant factor $\mathrm{e}^{-\Omega_0\phi(\tilde{x})}$, i.e., $|\tilde{I}_{\Omega_0}| = \mathrm{e}^{-\Omega_0\phi(\tilde{x})}|I_{\Omega_0}|$. Moreover, if $\Omega \geq \Omega_0$, then $\mathrm{e}^{\Omega\tilde{\phi}(x)} \leq \mathrm{e}^{\Omega_0\tilde{\phi}(x)}$ for all x in R, since by definition $\tilde{\phi}$ is a nonpositive function on R. Hence, $\tilde{I}_\Omega$ is also absolutely convergent for all $\Omega \geq \Omega_0$.[2]

[1] Here and in the following, we use the Vinogradov notations $f(x) \ll g(x)$ and $g(x) \gg f(x)$ as $x \to x_0$, which are equivalent to $f(x) = \mathcal{O}(g(x))$ as $x \to x_0$.

[2] Conversely, as $\Omega \to \infty$, existence of I_Ω would be guaranteed by existence of $\tilde{I}_\Omega$ only if $\phi(\tilde{x}) \leq 0$.

In general, there are three different cases to consider[3]:

(a) The maximum of ϕ occurs in the interior of $R = [a; b]$ and is a stationary point of ϕ, i.e., $a < \tilde{x} < b$, and $\phi'(\tilde{x}) = 0$ and $\phi''(\tilde{x}) \leq 0$.

(b) The maximum of ϕ occurs on the boundary of $R = [a; b]$ and is a stationary point of ϕ, i.e., $\tilde{x} = a$ or $\tilde{x} = b$, and $\phi'(\tilde{x}) = 0$ and $\phi''(\tilde{x}) \leq 0$.

(c) The maximum of ϕ occurs on the boundary of $R = [a; b]$, but is not a stationary point of ϕ, i.e., $\tilde{x} = a$ or $\tilde{x} = b$, and $\phi'(\tilde{x}) \neq 0$.

In each of these cases, the asymptotic expansion of $\tilde{I}_\Omega$ proceeds in the following steps:

1. The dominant contribution to $\tilde{I}_\Omega$ originates from the immediate neighborhood about the maximum $\tilde{x}$. We can reduce the range of integration to this local region for the costs of only an exponentially small error as $\Omega \to \infty$.

2. We assume that $f(x)$ and $\tilde{\phi}(x) = \phi(x) - \phi(\tilde{x})$ can be approximated in the local domain about the maximum $\tilde{x}$ by their Taylor series expansions about $\tilde{x}$, i.e.,

$$f(\tilde{x}+h) = \sum_{k=0}^{n} \frac{f^{(k)}(\tilde{x})}{k!} h^k + \mathcal{O}(h^{n+1}) \quad \text{and} \quad \tilde{\phi}(\tilde{x}+h) = \sum_{k=1}^{n} \frac{\phi^{(k)}(\tilde{x})}{k!} h^k + \mathcal{O}(h^{n+1}),$$

   where $f^{(k)}$ and $\phi^{(k)}$ denotes the k^{th} derivative of f and ϕ, respectively, and $h := x - \tilde{x}$. (Note that $\tilde{\phi}(\tilde{x}) = 0$.) The order used for the Taylor expansions of f and $\tilde{\phi}$ will dictate the order of the asymptotic expansion of $\tilde{I}_\Omega$ with respect to Ω.

3. The resulting integrals can be solved analytically by extending their integration limits to infinity. Again, this results in only an exponentially small error as $\Omega \to \infty$.

Below, we study these steps in more detail for case (a). For case (b) and (c), we briefly point out technical differences that arise compared to case (a).

At first glance, it might seem foolish in the procedure outlined above to first restrict integration to some small region and then extending it to infinity afterwards. However, restriction of the full integral allows to replace f and $\tilde{\phi}$ (or ϕ, respectively) by their Taylor or some other asymptotic series expansions about $\tilde{x}$. As we will see shortly, this results in an

[3]For the cases (a) and (b), where $\tilde{x}$ is a stationary point of ϕ, we assume here that this maximum is non-degenerate, i.e., $\phi''(\tilde{x}) < 0$. If more generally $\phi''(\tilde{x}) = \ldots = \phi^{(p-1)}(\tilde{x}) = 0$ and $\phi^{(p)}(\tilde{x}) \neq 0$, then p must be even and $\phi^{(p)}(\tilde{x}) < 0$ (otherwise $\tilde{x}$ would not be a maximum of ϕ), and we would find similar results as derived below.

expansion of the integral $\tilde{I}_\Omega$ into a series of Gaussian integrals. To evaluate these integrals, it is then convenient to extend integration to infinity. Each change in the limits of integration introduces only a small error that decays exponentially rapidly as $\Omega \to \infty$.

Case (a): $a < \tilde{x} < b$, and $\phi'(\tilde{x}) = 0$ and $\phi''(\tilde{x}) \leq 0$.

Step 1: We split the interval $R = [a; b]$ into the local region about $\tilde{x}$, i.e., $R_h := [\tilde{x} - h; \tilde{x} + h]$, and the remaining interval given by $R \setminus R_h = [a; \tilde{x} - h] \cup [\tilde{x} + h; b]$, where the value of $h \in \mathbb{R}^+$ is arbitrary, but necessarily $0 < h \leq \min(\tilde{x} - a, b - \tilde{x})$, such that

$$\tilde{I}_\Omega = \int_{R_h} f(x) \mathrm{e}^{\Omega\tilde{\phi}(x)} \, \mathrm{d}x + \int_{R \setminus R_h} f(x) \mathrm{e}^{\Omega\tilde{\phi}(x)} \, \mathrm{d}x \quad \text{as} \quad \Omega \to \infty. \tag{B.1}$$

Since $\tilde{\phi}$ is continuous and $\tilde{\phi}(x) < \tilde{\phi}(\tilde{x})$ for all $x \neq \tilde{x}$ in R, we can choose h so small that $\tilde{\phi}(x) \leq \tilde{\phi}(\tilde{x} + h)$ for all $x \in R \setminus R_h$. The contribution of the second integral in eq. (B.1) is then of the order $\mathcal{O}\big(\mathrm{e}^{\Omega\tilde{\phi}(\tilde{x}+h)}\big)$, because for all $\Omega \geq \Omega_0$ it holds

$$\begin{aligned}
\int_{R \setminus R_h} f(x) \mathrm{e}^{\Omega\tilde{\phi}(x)} \, \mathrm{d}x &\leq \int_{R \setminus R_h} |f(x)| \mathrm{e}^{\Omega\tilde{\phi}(x)} \, \mathrm{d}x = \int_{R \setminus R_h} |f(x)| \mathrm{e}^{(\Omega-\Omega_0)\tilde{\phi}(x)} \mathrm{e}^{\Omega_0\tilde{\phi}(x)} \, \mathrm{d}x \\
&\leq \mathrm{e}^{(\Omega-\Omega_0)\tilde{\phi}(\tilde{x}+h)} \int_{R \setminus R_h} |f(x)| \mathrm{e}^{\Omega_0\tilde{\phi}(x)} \, \mathrm{d}x \\
&\leq C \mathrm{e}^{\Omega\tilde{\phi}(\tilde{x}+h)},
\end{aligned}$$

where $C := \mathrm{e}^{-\Omega_0\tilde{\phi}(\tilde{x}+h)} |\tilde{I}_{\Omega_0}| > 0$ is a constant. Hence, the integral $\int_{R \setminus R_h} f(x) \mathrm{e}^{\Omega\tilde{\phi}(x)} \, \mathrm{d}x$ decays exponentially as $\Omega \to \infty$, provided that $h > 0$, and thus $\tilde{\phi}(\tilde{x} + h) < \tilde{\phi}(\tilde{x}) = 0$.

However, as we will see shortly, in the next step we further need $h \ll \Omega^{-1/3}$, such that $h \to 0$ as $\Omega \to \infty$. Since $\tilde{\phi}(\tilde{x} + h) \sim -h^2 \frac{|\phi''(\tilde{x})|}{2}$, we thus require that $\Omega h^2 \gg 1$, i.e., $h \gg \Omega^{-1/2}$ as $\Omega \to \infty$, to guarantee that the second integral in eq. (B.1) still decays exponentially rapid as $h \to 0$. Putting things together, we have

$$\tilde{I}_\Omega = \int_{R_h} f(x) \mathrm{e}^{\Omega\tilde{\phi}(x)} \, \mathrm{d}x + \mathcal{O}\big(\mathrm{e}^{-\Omega h^2 c}\big) \quad \text{as} \quad \Omega \to \infty,$$

with $\Omega^{-1/2} \ll h \ll \Omega^{-1/3}$ and $c := \frac{|\phi''(\tilde{x})|}{2} > 0$.

Step 2: On the interval $R_h = [\tilde{x} - h; \tilde{x} + h]$, we use the Taylor series expansions of f and $\tilde{\phi}$ about $\tilde{x}$, such that

$$\tilde{I}_\Omega \sim \int_{-h}^{+h} \left(\sum_{k=0}^{n} \frac{f^{(k)}(\tilde{x})}{k!} z^k + \mathcal{O}(z^{n+1}) \right) \exp\left\{ \Omega \left(\sum_{k=2}^{n} \frac{\phi^{(k)}(\tilde{x})}{k!} z^k + \mathcal{O}(z^{n+1}) \right) \right\} \mathrm{d}z \tag{B.2}$$

as $\Omega \to \infty$, where we substituted $z := x - \tilde{x}$. (Note that $\phi'(\tilde{x}) = 0$.) We expand the exponential function in the integrand of $\tilde{I}_\Omega$ as

$$\begin{aligned}\exp\left\{\Omega\left(\sum_{k=2}^{n}\frac{\phi^{(k)}}{k!}z^k + \mathcal{O}(z^{n+1})\right)\right\} &= \exp\left\{-\Omega z^2 c\right\}\exp\left\{\Omega z^3\left(\sum_{k=3}^{n}\frac{\phi^{(k)}}{k!}z^{k-3} + \mathcal{O}(z^{n-2})\right)\right\} \\ &= \exp\left\{-\Omega z^2 c\right\}\left\{1 + \Omega z^3\left(\sum_{k=3}^{n}\frac{\phi^{(k)}}{k!}z^{k-3} + \mathcal{O}(z^{n-2})\right) + \cdots\right\},\end{aligned} \tag{B.3}$$

where the derivatives of ϕ are all evaluated at $\tilde{x}$ and $c = \frac{|\phi''(\tilde{x})|}{2}$. This is where we need that $h \ll \Omega^{-1/3}$, i.e., $\Omega h^3 \ll 1$ as $\Omega \to \infty$, such that the above series converges for all $z \in [-h; +h]$. Substituting expansion (B.3) in eq. (B.2) and keeping only the terms of leading order gives

$$\tilde{I}_\Omega \sim f(\tilde{x})\int_{-h}^{+h} \mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z \quad \text{as} \quad \Omega \to \infty. \tag{B.4}$$

Step 3: By extending the integration limits in eq. (B.4) to infinity, we find the Gaussian integral[4]

$$\int_{-\infty}^{+\infty} \mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z = \sqrt{\frac{\pi}{\Omega c}}. \tag{B.5}$$

This again introduces only an exponentially small error, since $\mathrm{e}^{-\Omega z^2 c} \ll 1$ for every $z \neq 0$ as $\Omega \to \infty$. More precisely, we have

$$\begin{aligned}\int_{-\infty}^{-h} \mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z + \int_{h}^{\infty} \mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z &= 2\int_{h}^{\infty} \mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z = 2\int_{h}^{\infty}\left(-\frac{1}{2\Omega z c}\right)\frac{\mathrm{d}}{\mathrm{d}z}\left[\mathrm{e}^{-\Omega z^2 c}\right]\mathrm{d}z \\ &= -\frac{1}{\Omega c}\left\{\frac{\mathrm{e}^{-\Omega z^2 c}}{z}\bigg|_{h}^{\infty} + \int_{h}^{\infty}\frac{1}{z^2}\mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z\right\} \\ &< \frac{1}{\Omega c}\left\{\frac{\mathrm{e}^{-\Omega h^2 c}}{h} + \mathrm{e}^{-\Omega h^2 c}\int_{h}^{\infty}\frac{1}{z^2}\,\mathrm{d}z\right\} = \frac{2\mathrm{e}^{-\Omega h^2 c}}{\Omega h c},\end{aligned}$$

which is indeed exponentially smaller than $\int_{-\infty}^{+\infty} \mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z$ as $\Omega \to \infty$. Thus, we finally derived Laplace's approximation, given by

$$\tilde{I}_\Omega \sim f(\tilde{x})\sqrt{\frac{2\pi}{\Omega|\phi''(\tilde{x})|}} \quad \text{as} \quad \Omega \to \infty, \tag{B.6}$$

[4]The standard trick to evaluate the Gaussian integral (B.5) is to compute its square, interpret this as a double integral in the plane and transform it to polar coordinates r and φ, i.e., $\left(\int_{-\infty}^{+\infty} \mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z\right)^2 = \int_{-\infty}^{+\infty} \mathrm{e}^{-\Omega y^2 c}\,\mathrm{d}y \cdot \int_{-\infty}^{+\infty} \mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z = \iint_{\mathbb{R}^2} \mathrm{e}^{-\Omega(y^2+z^2)c}\,\mathrm{d}y\,\mathrm{d}z = \int_0^\infty \int_0^{2\pi} r\mathrm{e}^{-\Omega r^2 c}\,\mathrm{d}\varphi\,\mathrm{d}r = 2\pi\int_0^\infty r\mathrm{e}^{-\Omega r^2 c}\,\mathrm{d}r = -\frac{\pi}{\Omega c}\mathrm{e}^{-\Omega r^2 c}\Big|_0^\infty = \frac{\pi}{\Omega c}$, from which follows that $\int_{-\infty}^{+\infty} \mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z = \sqrt{\frac{\pi}{\Omega c}}$. The two coordinate systems are here related by $y = r\cos(\varphi)$ and $z = r\sin(\varphi)$, such that $r^2 = y^2 + z^2$.

or, regarding the original integral $I_\Omega = \mathrm{e}^{\Omega\phi(\tilde{x})}\tilde{I}_\Omega$, by

$$I_\Omega \sim f(\tilde{x})\mathrm{e}^{\Omega\phi(\tilde{x})}\sqrt{\frac{2\pi}{\Omega|\phi''(\tilde{x})|}} \quad \text{as} \quad \Omega \to \infty.$$

Higher Order Terms: Using the same procedure as above, but considering higher terms in the Taylor expansions of f and $\tilde{\phi}$, e.g.,

$$\begin{aligned}\tilde{I}_\Omega &\sim \int_{-\infty}^{+\infty}\left\{f + f'z + \frac{f''}{2!}z^2\right\}\exp\left\{\Omega\left(\frac{\phi''}{2!}z^2 + \frac{\phi^{(3)}}{3!}z^3 + \frac{\phi^{(4)}}{4!}z^4 + \frac{\phi^{(5)}}{5!}z^5\right)\right\}\mathrm{d}z \\ &\sim \int_{-\infty}^{+\infty}\left\{f + f'z + \frac{f''}{2}z^2\right\} \\ &\qquad \times \exp\{-\Omega z^2 c\}\left\{1 + \Omega\left(\frac{\phi^{(3)}}{6}z^3 + \frac{\phi^{(4)}}{24}z^4 + \frac{\phi^{(5)}}{120}z^5 + \Omega\frac{\phi^{(3)}\phi^{(3)}}{36}z^6 + \dots\right)\right\}\mathrm{d}z,\end{aligned} \tag{B.7}$$

as $\Omega \to \infty$, where f, ϕ and their derivatives are all evaluated at $\tilde{x}$, and $c = \frac{|\phi''(\tilde{x})|}{2}$, we find a series of Gaussian integrals in the expansion of $\tilde{I}_\Omega$, for which generally holds[5]

$$\Omega^p\int_{-\infty}^{+\infty} z^n\mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z = \begin{cases}0 & \text{if } n \text{ is odd,}\\ \dfrac{\Gamma(k+1/2)}{c^{k+1/2}}\Omega^{p-k-1/2} & \text{if } n = 2k \text{ is even,}\end{cases} \tag{B.8}$$

where p, n and $k \in \mathbb{N}_0$, and Γ denotes the gamma function with

$$\Gamma(k+1/2) = \sqrt{\pi}\prod_{j=1}^{k}\frac{2j-1}{2}.$$

Therefore, all terms in eq. (B.7) that are of odd order with respect to z will not contribute to the full expansion of $\tilde{I}_\Omega$. All other terms, where $n = 2k$ is even, will contribute to $\tilde{I}_\Omega$, with terms of order $\mathcal{O}(\Omega^p z^{2k})$ integrating to terms of order $\mathcal{O}(\Omega^{p-k-1/2})$. Considering only the term of leading order gives the classical Laplace approximation (B.6). If we also consider those terms in eq. (B.7) that integrate to terms of order $\mathcal{O}(\Omega^{-3/2})$ (i.e., related to $p - k = -1$ in eq. (B.8)), we find the next order in the expansion of $I_\Omega = \mathrm{e}^{\Omega\phi(\tilde{x})}\tilde{I}_\Omega$, i.e.,

$$I_\Omega = \mathrm{e}^{\Omega\phi(\tilde{x})}\sqrt{\frac{2\pi}{\Omega\,|\phi''|}}\left(f + \Omega^{-1}\left(\frac{f''}{2\,|\phi''|} + \frac{f'\phi^{(3)}}{2\,|\phi''|^2} + \frac{f\phi^{(4)}}{8\,|\phi''|^2} + \frac{5f\phi^{(3)}\phi^{(3)}}{24\,|\phi''|^3}\right) + \mathcal{O}(\Omega^{-2})\right),$$

as $\Omega \to \infty$.

[5] Obviously, if n is odd, then $\int_{-\infty}^{+\infty} z^n\mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z = 0$. If $n = 2k$ is even, consider the k^{th} partial derivative of $\mathrm{e}^{-\Omega z^2 c}$ with respect to c, which gives $\int_{-\infty}^{+\infty} z^{2k}\mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z = (-\Omega)^{-k}\int_{-\infty}^{+\infty}\frac{\partial^k}{\partial c^k}\left[\mathrm{e}^{-\Omega z^2 c}\right]\mathrm{d}z = (-\Omega)^{-k}\frac{\partial^k}{\partial c^k}\left[\int_{-\infty}^{+\infty}\mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z\right] = (-\Omega)^{-k}\frac{\partial^k}{\partial c^k}\left[\sqrt{\frac{\pi}{\Omega c}}\right] = \Omega^{-k-1/2}\left(\sqrt{\pi}\prod_{j=1}^{k}\frac{2j-1}{2}\right)c^{-k-1/2} = \Gamma(k+1/2)/c^{k+1/2}\Omega^{-k-1/2}$.

Case (b): $\tilde{x} = a$ or $\tilde{x} = b$, and $\phi'(\tilde{x}) = 0$ and $\phi''(\tilde{x}) \leq 0$.

If $\tilde{x}$ lies on the boundary of $R = [a; b]$, we would restrict the integration domain in the first step to either $[\tilde{x}; \tilde{x} + h]$, if $\tilde{x} = a$, or $[\tilde{x} - h; \tilde{x}]$, if $\tilde{x} = b$. After substituting the Taylor series expansions of f and $\tilde{\phi}$ about $\tilde{x}$, only that integration limit where $\tilde{x}$ is *not* located at can be expanded to (plus/minus) infinity for costs of an exponentially vanishing error as $\Omega \to \infty$. Thus, we would compute the Gaussian integral $\int_0^\infty \mathrm{e}^{-\Omega z^2 c}\,\mathrm{d}z = \frac{1}{2}\sqrt{\frac{\pi}{\Omega c}}$, resulting in

$$I_\Omega = \mathrm{e}^{\Omega\phi(\tilde{x})}\tilde{I}_\Omega \sim f(\tilde{x})\mathrm{e}^{\Omega\phi(\tilde{x})}\sqrt{\frac{\pi}{2\Omega|\phi''(\tilde{x})|}} \quad \text{as} \quad \Omega \to \infty.$$

Case (c): $\tilde{x} = a$ or $\tilde{x} = b$, and $\phi'(\tilde{x}) \neq 0$.

The situation is slightly different if $\phi'(\tilde{x}) \neq 0$ and (necessarily) $\tilde{x}$ is an endpoint of $R = [a; b]$. Consider the case where the supremum of ϕ lies on the lower bound of R, i.e., $\tilde{x} = a$. Then, $\phi'(\tilde{x}) < 0$ and

$$\begin{aligned} I_\Omega = \mathrm{e}^{\Omega\phi(\tilde{x})}\tilde{I}_\Omega &\sim \mathrm{e}^{\Omega\phi(\tilde{x})}\int_{\tilde{x}}^{\tilde{x}+h} f(x)\mathrm{e}^{\Omega\tilde{\phi}(x)}\,\mathrm{d}x \\ &\sim f(\tilde{x})\mathrm{e}^{\Omega\phi(\tilde{x})}\int_0^h \mathrm{e}^{-\Omega|\phi'(\tilde{x})|z}\,\mathrm{d}z \\ &\sim f(\tilde{x})\mathrm{e}^{\Omega\phi(\tilde{x})}\int_0^\infty \mathrm{e}^{-\Omega|\phi'(\tilde{x})|z}\,\mathrm{d}z = \frac{f(\tilde{x})\mathrm{e}^{\Omega\phi(\tilde{x})}}{\Omega|\phi'(\tilde{x})|} \quad \text{as} \quad \Omega \to \infty. \end{aligned}$$

If $\tilde{x} = b$, then $\phi'(\tilde{x}) > 0$ and a similar computation as above gives

$$I_\Omega = \mathrm{e}^{\Omega\phi(\tilde{x})}\tilde{I}_\Omega \sim \frac{f(\tilde{x})\mathrm{e}^{\Omega\phi(\tilde{x})}}{\Omega\phi'(\tilde{x})} \quad \text{as} \quad \Omega \to \infty.$$

C

Parameterization of the FlgM–FliA Model

Experimental Measurements

In Barembruch and Hengge [195], the flagellar cascade was induced by activating *flhDC* expression from an inducible promoter, which produces FlhDC levels comparable to those in a wild type strain. Subsequently, molecular levels of FliA and intra- and extra-celular FlgM were measured at several time points: immediately upon induction, and 5, 20, 35, 50 and 80 min after induction. Table C.1 lists the number of molecules measured in two independent experiments for the wild type and the *flgM*$^-$ mutant that are compared to our *in silico* predictions in Figure 5.3.

Table C.1: *Measured levels of FliA, intra- and extra-cellular FlgM. The number of molecules of total intra- and extra-cellular FlgM and total FliA per cell in wild type, and the total number of FliA molecules per cell in flgM*$^-$ *mutant have each been measured in two independent experiments.*

Time Point (in min)	0	5	20	35	50	80
			Wild Type			
Total FlgM	848	975	4,027	4,239	5,977	7,630
	2,036	1,697	5,087	4,875	6,571	8,690
FlgM$_{extern}$	1,484	424	636	11,657	18,312	26,706
	2,544	2,332	1,696	12,293	22,044	34,336
Total FliA	345	345	3,537	7,334	7,248	8,743
	489	575	5,465	7,594	7,622	8,859
			***flgM*$^-$ Mutant**			
Total FliA	35	29	431	2,532	3,596	4,026
	95	173	863	2,818	3,940	4,170

Baseline levels of FliA, FlgM and $FlgM_{extern}$ were calculated as the average values of the corresponding measurements at time 0 and 5 min, resulting in 1, 153 nM for total FlgM, 1, 408 nM for $FlgM_{extern}$ and 364 nM for total FliA.

Parameterization of the Central FlgM–FliA Interactions Model

The central model of FlgM–FliA interactions shown in Figure 5.1 was parameterized based on *in vivo* data from *E. coli* or related bacteria (this applied to the majority of parameters, including all key parameters). When *in vivo* data were not available, data were taken from *in vitro* measurements or parameters were estimated based on our experimental measurements [195]. All parameter values are listed in Table C.2.

In accordance with the experimental conditions in [195], the cellular volume of *E. coli* was chosen as 2×10^{-15} l. The cellular volume was used to convert concentrations from molar units to number of molecules, as determined experimentally and shown in all figures.

The association and dissociation rate constants of the FlgM–FliA binding kinetics have been measured as $k_{on} = 0.053\,(\text{nM} \cdot \text{min})^{-1}$ and $k_{off} = 0.096\,\text{min}^{-1}$, respectively [215]. The *in vivo* half-life of FliA has been measured in [195]. Since in the *flgM*$^-$ mutant FliA is not protected against proteolysis, the rate constant of proteolysis has been determined from the half-life measured in *flgM*$^-$, resulting in $k_{prot} = 0.06\,\text{min}^{-1}$.

The results in [208] demonstrate that the export of FlgM is mediated by FliA, acting as its type III secretion chaperone. We modeled the export of FlgM by the type III secretion system as a first order reaction with the FlgM:FliA complex as reactant, and extra-cellular FlgM and free (intra-cellular) FliA as products. The rate constant associated with this export process is difficult to determine, since it depends on both, the number of available secretion apparatuses and the unknown export rate per secretion apparatus. Thus, we estimated the effective export rate constant k_{export} and the start of export t_{export} based on the experimental data of external anti-sigma factor $FlgM_{extern}$ [195]. This resulted in $t_{export} = 17$ min and an export rate constant of $0.16\,\text{min}^{-1}$. The dilution rate constant k_{dil} was determined from a cell-cycle length of 24 min (in accordance with the growth conditions in [195]), resulting in $k_{dil} = \ln(2)/24 \approx 0.029\,\text{min}^{-1}$.

The average σ^D (σ^{70}) level in *E. coli* has been reported as 14, 000 nM (approximately 17, 000 molecules) [246]. The average level of RNAP was set to 1, 700 nM, which is 20 % of the average total RNAP level reported in [247] to account for the fact that a large number of core enzymes is engaged in other pathways which are not part of the model [246, 247].

Table C.2: *Parameter values of the central FlgM–FliA interactions model. Initial concentration for wild type: $FlgM(0) = 725\,nM$, $FlgM{:}FliA(0) = 310\,nM$, $\sigma^D(0) = 12,300\,nM$, $\sigma^D{:}RNAP(0) = 1,700\,nM$, all others zero; and for $flgM^-$ mutant: $FliA(0) = 100\,nM$, $\sigma^D(0) = 12,300\,nM$, $\sigma^D{:}RNAP(0) = 1,700\,nM$, all others zero. These values are based on numerically estimated steady states. To convert the concentrations of molar unit to number of molecules a cellular volume of 2×10^{-15} l was used.*

Parameter	Value	Unit	Reference
	Wild Type		
$k_{\mathrm{FlgM}}(t)$	$\begin{cases} 30, & \text{if } t \le t_{\mathrm{class2}} \\ 500 + \dfrac{1100 \cdot \sigma^{\mathrm{F}}\text{:RNAP}}{10 + \sigma^{\mathrm{F}}\text{:RNAP}}, & \text{otherwise} \end{cases}$	nM/min	[195]
$k_{\mathrm{FliA}}(t)$	$\begin{cases} 9 & \text{if } t \le t_{\mathrm{class2}} \\ 300 & \text{otherwise} \end{cases}$	nM/min	[195]
k_{prot}	0.06	1/min	[195]
k_{on}	0.053	$1/(\mathrm{nM} \cdot \mathrm{min})$	[215]
k_{off}	0.0096	1/min	[215]
$k_{\mathrm{export}}(t)$	$\begin{cases} 0, & \text{if } t \le t_{\mathrm{export}} \\ 0.16, & \text{otherwise} \end{cases}$	1/min	[195]
$k_{\mathrm{a},\sigma^{\mathrm{D}}}$	0.087	$1/(\mathrm{nM} \cdot \mathrm{min})$	[216]
$k_{\mathrm{a},\sigma^{\mathrm{F}}}$	0.029	$1/(\mathrm{nM} \cdot \mathrm{min})$	[215]
k_{d}	0.023	1/min	[215]
k_{dil}	0.029	1/min	[195]
t_{class2}	10	min	[195]
t_{export}	17	min	[195]
	***flgM*$^-$ Mutant** (if different to wild type parameter)		
$k_{\mathrm{FlgM}}(t)$	0	nM/min	[195]
t_{class2}	18.5	min	[195]

The association and dissociation rate constants of σ^F (FliA) binding to RNAP have been measured as $k_{a,\sigma^F} = 0.029\,(\mathrm{nM}\cdot\mathrm{min})^{-1}$ and $k_d = 0.023\,\mathrm{min}^{-1}$, respectively [215]. The dissociation constant K_{D,σ^D} of the σ^D:RNAP complex has been quantified in [216] as $K_{D,\sigma^D} = 0.26\,\mathrm{nM}$. Since the single association/dissociation rate constants can not be derived from these measurements, we set the value of the dissociation rate constant of σ^D identical to that of σ^F and determined the association rate constant by $k_{a,\sigma^D} = k_d/K_{D,\sigma^D} = 0.087\,(\mathrm{nM}\cdot\mathrm{min})^{-1}$. We observed that varying the value of k_d for σ^D did not had any significant effect on the competition for RNAP (data not shown).

The time t_{class2} at which class 2 expression started was not measured experimentally. Examining the experimental data of FliA in wild type and *flgM*⁻ mutant revealed that the largest variability (variation from mean value) was observed for the 20 min measurement. This is the phase when FliA numbers increase due to initiation of class 2 expression. While for the wild type, the mean of the 20 min corresponds to roughly 1/2 of the final level at 80 min, it is only 1/6 of the final level in the *flgM*⁻ mutant. Assuming the same FliA synthesis rates in wild type and *flgM*⁻ mutant, this was considered as supporting evidence that the start of class 2 expression was earlier in the wild type compared to the *flgM*⁻ mutant. Parameter estimation based on the experimental data for the wild type and the *flgM*⁻ mutant resulted in $t_{class2} = 10\,\mathrm{min}$ and $t_{class2} = 18.5\,\mathrm{min}$, respectively.

Protein synthesis rates of FlgM and FliA were parameterized in the form

$$k_{prot} = \begin{cases} k_{basal} & \text{if } t \le t_{class2} \\ k_{class2} + k_{class3}(\sigma^F\text{:RNAP}) & \text{otherwise,} \end{cases}$$

were k_{basal} denotes a basal synthesis rate before induction, and k_{class2} and k_{class3} are the class 2 and 3 synthesis rates, respectively, after induction at time t_{class2} by the master regulator FlhDC. The basal synthesis rates k_{basal} were estimated from the experimental data such that the model predictions were in agreement with the measured baseline levels of total FlgM and FliA. This resulted in $k_{basal} = 30\,\mathrm{nM/min}$ for FlgM and $k_{basal} = 9\,\mathrm{nM/min}$ for FliA. Class 2 gene expression was assumed to be rapidly saturated with start at time t_{class2} (as supported by the detailed transcription and translation model, see below and Figure 5.4), and hence modeled by zero-order rates $k_{class2} \approx V_{max}$. These k_{class2} rates were determined by fitting the proposed model to the experimental data of FlgM and FliA. This resulted in $k_{class2} = 500\,\mathrm{nM/min}$ for FlgM and $k_{class2} = 300\,\mathrm{nM/min}$ for FliA. For FlgM we further determined the σ^F:RNAP induced class 3 gene expression rate according to eq. (5.13) by using the experimental data of FlgM for the wild type, resulting in $V_{max} = 1,100\,\mathrm{nM/min}$ and $K_M = 10\,\mathrm{nM}$.

Parameterization of the Detailed Transcription and Translation Model

Based on measured levels of FlhC and FlhD, the FlhDC level in *E. coli* has been estimated in [248] to be of approximately 200 nM (≈ 250 molecules) per differentiated cell. The dissociation constants of FlhDC-binding to the target sites of the class 2 flagellar operons of *E. coli* were measured in [249], i.e., 12 nM for *flgB* and 25 nM for *fliA*, respectively.

The average dissociation rate constant of RNAP from the positions of promoters on DNA has been measured in [250] as $k_{\mathrm{d,RNAP}} = 40\,\mathrm{min}^{-1}$. The corresponding association rate constant $k_{\mathrm{a,RNAP}} = 4\,(\mathrm{nM}\cdot\mathrm{min})^{-1}$ was estimated with respect to the effective binding constant $K_{\mathrm{B,RNAP}} = k_{\mathrm{a,RNAP}}/k_{\mathrm{d,RNAP}}$ of RNAP to DNA, which is approximately $0.1\,\mathrm{nM}^{-1}$ [251].

The average elongation kinetics of transcription and translation in *E. coli* have been reported in [247] as $k_{\mathrm{elong,tc}} = 3,300$ nucleotides per minute and $k_{\mathrm{elong,tl}} = 1,300$ amino acid residues per minute, respectively. According to the results in [219], the average distance between two traversing RNAP molecules is approximately $n_{\mathrm{tc}} = 100$ nucleotides. The average distance of ribosomes on an mRNA has been reported in [247] as $n_{\mathrm{tl}} = 14$ amino acids.

For the effective class 2 synthesis rates, see eqs. (5.11) and (5.12), the above parameter values of the detailed transcription and translation model gave upper bounds for the Michaelis–Menten constants, i.e.,

$$K_{\mathrm{M_2}} < 10 \cdot \left(1 + \frac{12}{\mathrm{FlhDC}}\right) \mathrm{nM},$$

in the case of FlgM, and

$$K_{\mathrm{M_2}} < 10 \cdot \left(1 + \frac{25}{\mathrm{FlhDC}}\right) \mathrm{nM},$$

for FliA. Hence, σ^{D}:RNAP $\approx 1700\,\mathrm{nM} \gg K_{\mathrm{M_2}}$ for both, FlgM and FliA, which clearly supported the assumption of a rapid saturation of class 2 synthesis after induction at time $t_{\mathrm{class_2}}$.

The degradation rates of mRNAs were estimated by corresponding half-lifes measured in [252], i.e., $k_{\mathrm{deg}} = \ln(2)/0.45 \approx 1.5\,\mathrm{min}^{-1}$ for class 2 and $k_{\mathrm{deg}} = \ln(2)/0.98 \approx 0.7\,\mathrm{min}^{-1}$ for class 3 transcripts, respectively. The translation initiation rate was set to $k_{\mathrm{init,tl}} = 120\,\mathrm{min}^{-1}$ accounting for an average number of approximately one hundred translations per mRNA as reported in [247].

Measured values of the initiation rates of class 2 and class 3 transcription were not available. However, having all other parameters fixed, we used the effective synthesis rates of FlgM and FliA, reported in Table C.2, and eqs. (5.3)–(5.10) to compute the transcription initiation rates $k_{\mathrm{init,tc}}$. This resulted in $k_{\mathrm{init,tc}} = 37\,\mathrm{min}^{-1}$ for FlgM and $k_{\mathrm{init,tc}} = 15\,\mathrm{min}^{-1}$ for FliA. These

values are comparable to the average transcription initiation rate of ribosomal genes, reported in [247] as $53\,\mathrm{min}^{-1}$.

All parameter values of the detailed transcription and translation model are listed in Table C.3. These values were used for evaluating the reduced model (see Subsection 5.4.2), and for the detailed sensitivity analysis of the effective synthesis rates of FlgM and FliA shown in Figure 5.8C+D.

Table C.3: *Parameter values of gene expression in the FlgM–FliA regulatory network.*

Rate Constant	Value	Unit	Reference
General Parameters			
FlhDC	200	nM	[248]
$k_{\mathrm{a,RNAP}}$	4	$1/(\mathrm{nM}\cdot\mathrm{min})$	[251]
$k_{\mathrm{d,RNAP}}$	40	1/min	[250]
$k_{\mathrm{elong,tc}}$	3, 300	nucleotides/min	[247]
n_{tc}	100	nucleotides	[219]
$k_{\mathrm{init,tl}}$	120	1/min	[247]
$k_{\mathrm{elong,tl}}$	1, 300	aa residues/min	[247]
n_{tl}	14	aa residues	[247]
Class 2 Expression of FlgM			
$K_{\mathrm{D,FlhDC}}$	12	nM	[249]
$k_{\mathrm{init,tc}}$	37	1/min	see text
k_{deg}	1.5	1/min	[252]
Class 2 Expression of FliA			
$K_{\mathrm{D,FlhDC}}$	25	nM	[249]
$k_{\mathrm{init,tc}}$	15	1/min	see text
k_{deg}	1.5	1/min	[252]
Class 3 Expression of FlgM			
$k_{\mathrm{init,tc}}$	37	1/min	see text
k_{deg}	0.7	1/min	[252]

Deutsche Zusammenfassung (German Summary)

Traditionell beruhen quantitative Modelle von Reaktionsnetzwerken auf der Sicht der klassischen chemischen Kinetik. Unter der Annahme des thermodynamischen Grenzfalls (unendlicher Molekülanzahlen-/Volumenlimes) werden Reaktionen hierbei vereinfacht als kontinuierliche, deterministische Prozesse modelliert. In zellulären Systemen, die Prozesse wie Genexpression oder Signaltransduktion beinhalten, zeigt sich jedoch, dass zu beobachtende diskrete Fluktuationen in geringen Molekülanzahlen von entscheidender Bedeutung sind. In diesen Fällen ist eine Modellierung basierend auf der stochastischen Reaktionskinetik erforderlich, in der Reaktionen als diskrete Zufallsprozesse beschrieben werden. Die zeitliche Entwicklung der Wahrscheinlichkeitsverteilung an Molekülanzahlen ist hierbei durch die chemische Mastergleichung (CME) gegeben, welche jedoch aufgrund ihrer hohen Dimensionalität im Allgemeinen nicht direkt gelöst werden kann. Stattdessen ist es üblich eine indirekte Lösung der CME durch Realisierungen des zugrundeliegenden Markov-Sprungprozesses zu approximieren. Ein weitverfolgtes Ziel ist nun die Entwicklung solcher indirekten Methoden, die die Simulation von komplexen, mehrskaligen Reaktionsnetzwerken ermöglichen.

Gegenstand dieser Arbeit ist die vielversprechende Entwicklung von sogenannten hybriden Methoden, in denen schnelle Reaktionen assoziert mit hohen Molekülanzahlen kontinuierlich–deterministisch und komplementäre Reaktionen diskret–stochastisch modelliert werden. Wir demonstrieren den Nutzen einer hybriden Systembeschreibung an einem integrativen Modell der Replikationsdynamik des Humane Immundefizienz-Virus (HIV). Mithilfe hybrider Simulationen ist es uns möglich eine neuartige Behandlungsstrategie für HIV-Patienten zu entwerfen und zu validieren, die zu wesentlichen Verbesserungen gegenüber konventionellen Behandlungsstrategien führen kann.

Während derzeitige hybride Methoden fast ausschließlich indirekte Näherungslösungen liefern, wird in dieser Arbeit ein neuer hybrider Zugang zur direkten Lösung der CME entwickelt. Anhand eines Mehrskalenansatzes werden Evolutionsgleichungen hergeleitet, die eine CME auf reduziertem Zustandsraum mit Evolutionsgleichungen der deterministisch approximierten Variablen koppeln. Hierdurch wird die Beeinflußung der Dynamik von deterministischen Komponenten durch Veränderungen in der Wahrscheinlichkeitsverteilung des stochastischen Teilsystems offensichtlich und kann, im Gegensatz zu indirekten hybriden Methoden, explizit berücksichtigt werden. Wir illustrieren und diskutieren unseren direkten hybriden Lösungsansatz an Modellsystemen von biologischem Interesse.

Im letzten Teil dieser Arbeit leiten wir effektive Proteinsyntheseraten, wie sie üblicherweise in deterministischen Modellen genutzt werden, über Reduktion eines detaillierten, stochastischen Genexpressionsmodells her. Wir nutzen unseren Reduktionsansatz um ein Modell der Proteininteraktionen bei der flagellaren Genregulation in *Escherichia coli* abzuleiten. Die erhaltenen funktionalen Zusammenhänge von Transkriptions- und Translationsprozessen zu den Syntheseraten zeigen hierbei auf, dass sich eine hohe Sensitivität hinsichtlich effektiver Raten nicht zwangsläufig auf zugrundeliegende Subprozesse übertragt.

Eidesstattliche Erklärung (Declaration)

Ich versichere hiermit an Eides Statt, dass diese Arbeit von niemand anderem als meiner Person verfasst worden ist. Alle verwendeten Hilfsmittel wie Berichte, Bücher, Internetseiten oder ähnliches sind im Literaturverzeichnis angegeben. Zitate aus fremden Arbeiten sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

Berlin, den 11. Juli 2012

Stephan Menz

References

[1] J. D. Murray. *Mathematical Biology*, volume 1 & 2. Springer-Verlag, 3rd edition, 2003. 1

[2] U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall/CRC, 2006.

[3] Z. Szallasi, J. Stelling, and V. Periwali. *System Modelling in Cellular Biology: From Concepts to Nuts and Bolts*. MIT Press, 2006.

[4] E. Klipp, W. Liebermeister, C. Wierling, A. Kowald, H. Lehrach, and R. Herwig. *Systems Biology: A Textbook*. Wiley-VCH, 2009. 1, 2, 5, 6

[5] B. M. Rao, D. A. Lauffenburger, and K. D. Wittrup. Integrating cell-level kinetic modeling into the design of engineered protein therapeutics. *Nature Biotechnology*, 23 (2):191–194, February 2005. 1

[6] D. Fell. *Understanding the Control of Metabolism*. Portland Press, 1st edition, 1996. 1

[7] R. Heinrich and S. Schuster. *The Regulation Of Cellular Systems*. Chapman & Hall, New York, 1996. 1, 2

[8] A. Cornish-Bowden. *Fundamentals of Enzyme Kinetics*. Portland Press, 3rd edition, 2004. 1

[9] H. Bolouri. *Computational Modeling of Gene Regulatory Networks*. Imperial College Press, 1st edition, 2008. 1

[10] D. T. Gillespie. Stochastic Simulation of Chemical Kinetics. *Annual Review of Physical Chemistry*, 58:35–55, 2007. 1, 2, 3, 17, 18, 19, 20, 21, 35

[11] I. R. Epstein and J. A. Pojman. *An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns, and Chaos*. Oxford University Press, Oxford, 1998. 2

[12] S. K. Upadhyay. *Chemical Kinetics and Reaction Dynamics*. Springer, New York, 1st edition, 2006. 2, 22, 31, 32

[13] C. E. Mortimer and U. Müller. *Chemie – Das Basiswissen der Chemie*. Thieme, 9th edition, Juli 2007. 2, 22, 31, 32

[14] D. T. Gillespie and L. R. Petzold. Numerical Simulation for Biochemical Kinetics. In Z. Szallasi, J. Stelling, and V. Periwal, editors, *System Modeling in Cellular Biology*, chapter 16, pages 331–354. The MIT Press, April 2006. 2, 4

[15] D. T. Gillespie. The chemical Langevin equation. *Journal of Chemical Physics*, 113(1): 297–306, July 2000. 2, 11, 21, 27, 28, 30

[16] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland Personal Library, 3rd edition, 2007. 2, 27, 28

[17] R. Srivastava, L. You, J. Summers, and J. Yin. Stochastic vs. Deterministic Modeling of Intracellular Viral Kinetics. *Journal of Theoretical Biology*, 218(3):309–321, October 2002. 2, 99, 100

[18] M. S. Samoilov and A. P. Arkin. Deviant effects in molecular reaction pathways. *Nature Biotechnology*, 24(10):1235–1240, October 2006. 2, 3, 34, 46, 47, 50

[19] H. H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *PNAS USA*, 94(3):814–819, February 1997. 2

[20] A. Arkin, J. Ross, and H. H. McAdams. Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage λ-Infected *Escherichia coli* Cells. *Genetics*, 149(4):1633–1648, August 1998.

[21] H. H. McAdams and A. Arkin. It's a noisy business! Genetic regulation at the nanomolar scale. *Trends in Genetics*, 15(2):65–69, February 1999.

[22] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic Gene Expression in a Single Cell. *Science*, 297:1183–1186, 2002.

[23] W. J. Blake, M. Kærn, C. R. Cantor, and J. J. Collins. Noise in eukaryotic gene expression. *Nature*, 422:633–637, March 2003.

[24] J. M. Raser and E. K. O'Shea. Control of Stochasticity in Eukaryotic Gene Expression. *Science*, 304:1811–1814, June 2004. 2

[25] B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *Journal of Chemical Physics*, 124(4):044104, January 2006. 2, 25, 26, 27, 95

[26] P. Deuflhard, W. Huisinga, T. Jahnke, and M. Wulkow. Adaptive discrete Galerkin Methods Applied to the Chemical Master Equation. *SIAM Journal on Scientific Computing*, 30(6):2990–3011, 2008. 25

[27] S. Engblom. Galerkin Spectral Method Applied to the Chemical Master Equation. *Communications in Computational Physics*, 5(5):871–896, May 2009.

[28] T. Jahnke. An Adaptive Wavelet Method for the Chemical Master Equation. *SIAM Journal on Scientific Computing*, 31(6):4373–4394, 2010. 25

[29] V. Wolf, R. Goel, M. Mateescu, and T. A. Henzinger. Solving the chemical master equation using sliding windows. *BMC Systems Biology*, 4:42, 2010. 2, 25, 27, 95

[30] D. T. Gillespie. A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. *Journal of Computational Physics*, 22(4): 403–434, December 1976. 2, 3, 12, 17, 19, 31

[31] D. T. Gillespie. Exact Stochastic Simulation of Coupled Chemical Reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, 1977. 3, 17, 19

[32] M. A. Gibson and J. Bruck. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *Journal of Physical Chemistry A*, 104(9): 1876–1889, 2000. 3, 19

[33] D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115(4):1716–1733, July 2001. 3, 20, 21

[34] M. Rathinam, L. R. Petzold, Y. Cao, and D. T. Gillespie. Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. *Journal of Chemical Physics*, 119 (24):12784–12794, 2003. 3, 20, 21

[35] M. Hegland, C. Burden, L. Santoso, S. MacNamara, and H. Booth. A solver for the stochastic master equation applied to gene regulatory networks. *Journal of Computational and Applied Mathematics*, 205(2):708–724, 2007. 2, 27

[36] G. S. Fishmann. *Monte Carlo: concepts, algorithms, and applications*. Springer Series in Operations Research. Springer, 1996. 3

[37] J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer, 2001. 3

[38] C. Schütte, J. Walter, C. Hartmann, and W. Huisinga. An averaging principle for fast degrees of freedom exhibiting long-term correlations. *Multiscale Modeling and Simulation*, 2(3):501–526, 2004. 3

[39] T. Tian and K. Burrage. Binomial Leap Methods for Simulating Stochastic Chemical Kinetics. *Journal of Chemical Physics*, 121(21):10356–10364, 2004. 3, 20, 21

[40] A. Chatterjee, D. G. Vlachos, and M. A. Katsoulakis. Binomial ditribution based τ-leap accelerated stochastic simulation. *Journal of Chemical Physics*, 122(2):024112, 2005. 21

[41] Y. Cao and L. Petzold. Trapezoidal tau-leaping formula for the stochastic simulation of chemically reacting systems. *Proceedings of Foundations of Systems Biology in Engineering (FOSBE 2005)*, pages 149–152, 2005. 21

[42] Y. Cao, D. T. Gillespie, and L. R. Petzold. The adaptive explicit-implicit tau-leaping method with automatic tau selection. *Journal of Chemical Physics*, 126(22):224101, 2007. 21

[43] T. Ahn and A. Sandu. Implicit Second Order Weak Taylor Tau-Leaping Methods for the Stochastic Simulation of Chemical Kinetics. *Procedia Computer Science*, 4:2297–2306, 2011. 3, 20, 21

[44] M. Samailov, S. Plyasnuov, and A. P. Arkin. Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *PNAS*, 102 (7):2310–2315, February 2005. 3

[45] F. Baras, M. M. Mansour, and J. E. Pearson. Microscopic simulation of chemical bistability in homogeneous systems. *The Journal of Chemical Physics*, 105(18):8257–8261, 1996. 3

[46] C. V. Rao and P. Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *The Journal of Chemical Physics*, 118(11): 4999–5010, 2003. 3, 20, 22, 23

[47] Y. Cao, D. T. Gillespie, and L. R. Petzold. The slow-scale stochastic simulation algorithm. *Journal of Chemical Physics*, 122(1):014116, 2005. 22, 23

[48] Y. Cao, D. T. Gillespie, and L. R. Petzold. Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems. *Journal of Chemical Physics*, 206(2):395–411, July 2005. 22

[49] J. Goutsias. Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems. *Journal of Chemical Physics*, 122(18):184102, May 2005. 22, 23, 104

[50] E. L. Haseltine and J. B. Rawlings. On the origins of approximations for stochastic chemical kinetics. *Journal of Chemical Physics*, 123(16):164115, 2005. 4, 54, 57, 60

[51] A. Samant and D. G. Vlachos. Overcoming stiffness in stochastic simulation stemming from partial equilibrium: A multiscale Monte Carlo algorithm. *Journal of Chemical Physics*, 123(14):144114, October 2005. 22, 23

[52] W. E, D. Liu, and E. Vanden-Eijnden. Nested stochastic simulation algorithms for chemical kinetic systems with multiple time scales. *Journal of Computational Physics*, 221(1):158–180, 2007. 3, 20, 22, 23

[53] E. L. Haseltine and J. B. Rawlings. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *Journal of Chemical Physics*, 117(15):6959–6969, 2002. 4, 54, 57, 60

[54] K. Burrage, T. Tian, and P. Burrage. A multi-scaled approach for simulating chemical reaction systems. *Progress in Biophysics & Molecular Biology*, 85(2–3):217–234, 2004. 54, 60

[55] M. Bentele and R. Eils. General Stochastic Hybrid Method for the Simulation of Chemical Reaction Processes in Cells. In V. Danos and V. Schachter, editors, *Computational Methods in Systems Biology*, volume 3082 of *Lecture Notes in Computer Science*, pages 248–251. Springer-Verlag Berlin Heidelberg, 2005. 60

[56] H. Salis and Y. Kaznessis. Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. *Journal of Chemical Physics*, 122(5):054103, February 2005. 4, 53, 54, 55, 56, 57, 60

[57] T. R. Kiehl, R. M. Mattheyses, and M. K. Simmons. Hybrid simulation of cellular behavior. *Bioinformatics*, 20(3):316–322, 2004. 4, 54, 57, 60

[58] N. A. Neogi. Dynamic Partitioning of Large Discrete Event Biological Systems for Hybrid Simulation and Analysis. In *Lecture Notes in Computer Science – Hybrid Systems: Computation and Control*, volume 2993, pages 229–248. Springer-Verlag Berlin Heidelberg, 2004. 54, 60

[59] K. Takahashi, K. Kaizu, B. Hu, and M. Tomita. A multi-algorithm, multi-timescale method for cell simulation. *Bioinformatics*, 20(4):538–546, 2004. 54, 57, 60

[60] K. Vasudeva and U. S. Bhalla. Adaptive stochastic-deterministic chemical kinetic simulations. *Bioinformatics*, 20(1):78–84, 2004. 54, 60

[61] A. Alfonsi, E. Cancès, G. Turinici, B. D. Ventura, and W. Huisinga. Adaptive simulation of hybrid stochastic and deterministic models for biochemical systems. *European Series in Applied and Industrial Mathematics (ESAIM): Proceedings*, 14:1–13, September 2005. 53, 55, 56, 60

[62] M. Griffith, T. Courtney, J. Peccoud, and W. H. Sanders. Dynamic partitioning for hybrid simulation of the bistable HIV-1 transactivation network. *Bioinformatics*, 22 (22):2782–2789, September 2006. 4, 54, 55, 57, 60

[63] T. G. Kurtz. The Relationship between Stochastic and Deterministic Models of Chemical Reactions. *Journal of Chemical Physics*, 57(7):2976–2978, April 1972. 4, 34, 39

[64] D. T. Gillespie. Deterministic Limit of Stochastic Chemical Kinetics. *Journal of Physical Chemistry B*, 113(6):1640–1644, 2009. 4, 35, 36, 37

[65] C. M. Bender and S. A. Orszag. *Advanced Mathematical Methods for Scientist and Engineers*. Springer, New York, 1999. 5, 12, 46, 47, 48, 147

[66] T. A. Henzinger, M. Mateescu, L. Mikeev, and V. Wolf. Hybrid Numerical Solution of the Chemical Master Equation. In *Proceedings of the 8th International Conference on Computational Methods in Systems Biology (CMSB'10)*, pages 55–65, 2010. 5, 91, 108

[67] T. Jahnke. On Reduced Models for the Chemical Master Equation. *Multiscale Modeling and Simulation*, 9(4):1646–1676, December 2011. 5, 108

[68] T. Quaiser, A. Dittrich, F. Schaper, and M. Mönnigmann. A simple work flow for biologically inspired model reduction - application to early JAK-STAT signaling. *BMC Systems Biology*, 5(30), February 2011. 5, 111

[69] U. Maas and S. B. Pope. Simplifying chemical kinetics: Intrinsic low-dimensional manifolds in composition space. *Combustion and Flame*, 88(3–4):239–264, March 1992. 5, 111

[70] S. H. Lam and D. A. Goussis. The csp method for simplifying kinetics. *International Journal of Chemical Kinetics*, 26(4):461–486, April 1994.

[71] I. Surovtsova, N. Simus, T. Lorenz, A. König, S. Sahle, and U. Kummer. Accessible methods for the dynamic time-scale decomposition of biochemical systems. *Bioinformatics*, 25(21):2816–2823, 2009. 5, 111

[72] I. Smets, K. Bernaerts, J. Sun, K. Marchal, J. Vanderleyden, and J. V. Impe. Sensitivity function-based model reduction: A bacterial gene expression case study. *Biotechnology and Bioengineering*, 80(2):195–200, October 2002. 5, 111

[73] D. Degenring, C. Froemel, G. Dikta, and R. Takors. Sensitivity analysis for the reduction of complex metabolism models. *Journal of Process Control*, 14(7):729–745, October 2004. 5, 111

[74] B. C. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, AC-26(1):17–32, February 1981. 5, 111

[75] J. Hahn and T. F. Edgar. An improved method for nonlinear model reduction using balancing of empirical gramians. *Computers & Chemical Engineering*, 26(10):1379–1397, October 2002.

[76] W. Liebermeister, U. Baur, and E. Klipp. Biochemical network models simplified by balanced truncation. *The FEBS Journal*, 272(16):4034–4043, 2005. 5, 111

[77] G. Yagil and E. Yagil. On the Relation between Effector Concentration and the Rate of Induced Enzyme Synthesis. *Biophysical Journal*, 11(1):11–27, January 1971. 6

[78] T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, 403:339–342, January 2000.

[79] P. B. Warren and P. R. ten Wolde. Chemical Models of Genetic Toggle Switches. *Journal of Physical Chemistry B*, 109(14):6812–6823, 2004. 6

[80] International Union of Pure and Applied Chemistry. Glossary of Terms Used in Physical Organic Chemistry. *Pure and Applied Chemistry*, 66(5):1077–1184, 1994. 9, 10, 11, 31, 32

[81] International Union of Pure and Applied Chemistry. A Glossary of Terms Used in Chemical Kinetics, including Reaction Dynamics. *Pure and Applied Chemistry*, 68(1): 149–192, 1996. 9, 10, 11, 31, 32

[82] D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188 (1–3):404–425, September 1992. 11, 12, 13, 24

[83] J. F. C. Kingman. *Poisson Processes*. Number 3 in Oxford Studies in Probability. Clarendon Press, Oxford, 1993. 14, 15, 41

[84] G. F. Lawler. *Introduction to Stochastic Processes*. Chapman & Hall, New York, 1st edition, 1995.

[85] S. M. Ross. *Stochastic Processes*. John Wiley & Sons, 2nd edition, 1996. 14, 15, 29

[86] D. F. Anderson. Incorporating postleap checks in tau-leaping. *Journal of Chemical Physics*, 128(5):054103, 2008. 14, 16

[87] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*, chapter Random Time Changes, pages 306–336. John Wiley & Sons, 2005. 14, 16, 40

[88] D. F. Anderson and T. G. Kurtz. *Design and Analysis of Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology*, chapter Continuous time Markov chain models for chemical reaction networks, pages 3–42. Springer, New York, 1st edition, 2011. 16, 17, 20, 40

[89] D. T. Gillespie. *Markov Processes: An Introduction for Physical Scientists*. Academic Press, 1992. 18, 28, 30, 59

[90] M. A. Gibson. *Computational Methods for Stochastic Biological Systems*. PhD thesis, California Institute of Technology, Pasadena, California, `http://resolver.caltech.edu/CaltechETD:etd-05132005-154222`, April 2000. 19

[91] T. E. Turner, S. Schnell, and K. Burrage. Stochastic approaches for modelling in vivo reactions. *Computational Biology and Chemistry*, 28(3):165–178, 2004. 19

[92] J. Pahle. Biochemical simulations: stochastic, approximate stochastic and hybrid approaches. *Briefings in Bioinformatics*, 10(1):53–64, January 2009. 19, 54, 56, 136

[93] E. S. Zeron and M. Santillán. Distributions for negative-feedback-regulated stochastic gene expression: Dimension reduction and numerical solution of the chemical master equation. *Journal of Theoretical Biology*, 264(2):377–385, 2010. 20, 22, 27

[94] Y. Cao, D. T. Gillespie, and L. R. Petzold. Efficient stepsize selection for the tau-leaping simulation method. *Journal of Chemical Physics*, 124(4):044109, 2006. 21

[95] Y. Cao, D. T. Gillespie, and L. R. Petzold. Avoiding negative populations in explicit Poisson tau-leaping. *Journal of Chemical Physics*, 123(5):054104, 2005. 21

[96] X. Peng, W. Zhou, and Y. Wang. Efficient binomial leap method for simulating chemical kinetics. *Journal of Chemical Physics*, 126(22):224109, 2007. 21

[97] M. F. Pettigrew and H. Resat. Multinomial tau-leaping method for stochastic kinetic simulations. *Journal of Chemical Physics*, 126(8):084101, 2007. 21

[98] Y. Hu and T. Li. Highly accurate tau-leaping methods with random corrections. *Journal of Chemical Physics*, 130(12):124109, 2009. 21

[99] G. E. Briggs and J. B. S. Haldane. A Note on the Kinetics of Enzyme Action. *Biochemical Journal*, 19:338–339, 1925. 22

[100] L. A. Segel. On the validity of the steady state assumption of enzyme kinetics. *Bulletin of Mathematical Biology*, 50(6):579–593, 1988. 22

[101] L. A. Segel and M. Slemrod. The Quasi-Steady-State Assumption: A Case Study in Pertubation. *SIAM Review*, 31(3):446–477, September 1989. 22

[102] C. Gadgil, C. H. Lee, and H. G. Othmer. A stochastic analysis of first-order reaction networks. *Bulletin of Mathematical Biology*, 67(5):901–946, September 2005. 25

[103] T. Jahnke and W. Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54(1):1–26, January 2007. 25, 26

[104] S. Peleš, B. Munsky, and M. Khammash. Reduction and solution of the chemical master equation using time scale separation and finite state projection. *Journal of Chemical Physics*, 125(20):204104, 2006. 25, 27, 95

[105] K. Burrage, M. Hegland, F. MacNamara, and B. Sidje. A Krylov-based Finite State Projection algorithm for solving the chemical master equation arising in the discrete modelling of biological systems. In A. N. Langville and W. J. Stewart, editors, *MAM 2006: Markov Anniversary Meeting*, pages 21–38. Boson Books, Charleston, South Carolina, USA, 2006. 26

[106] B. Drawert, M. J. Lawson, L. Petzold, and M. Khammash. The Diffusive Finite State Projection Algorithm for Efficient Simulation of the Stochastic Reaction-Diffusion Master Equation. *Journal of Chemical Physics*, 132(7):074101, 2010. 25

[107] R. S. Varga. *Geršgorin and His Circles*. Springer, Berlin, 1st edition, 2004. 26

[108] T. Jahnke and W. Huisinga. A dynamical low-rank approach to the chemical master equation. *Bulletin of Mathematical Biology*, 70(8):2283–2302, August 2008. 27

[109] V. Sunkara and M. Hegland. An optimal Finite State Projection Method. *Procedia Computer Science*, 1(1):1579–1586, 2010. 27, 95

[110] C. W. Gardiner. *Handbook of Stochastic Models for Physics, Chemistry and the Natural Sciences*. Springer Series in Synergetics. Springer, 3rd edition, 2004. 27, 28

[111] H. A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940. 27

[112] J. E. Moyal. Stochastic Processes and Statistical Physics. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 11:150–210, 1949. 27

[113] R. Grima, P. Thomas, and A. V. Straube. How accurate are the non-linear chemical Fokker-Planck and chemical Langevin equations? *Journal of Chemical Physics*, 135(8): 084103, 2011. 28, 30

[114] H. Risken. *The Fokker-Planck Equation: Methods of Solutions and Applications*. Springer Series in Synergetics. Springer-Verlag Berlin Heidelberg, 2nd edition, 1996. 28

[115] R. Khanin and D. J. Higham. Chemical Master Equation and Langevin Regimes for a gene transcription model. *Theoretical Computer Science*, 408(1):31–40, 2008. 30

[116] J. Wilkie and Y. M. Wong. Positivity preserving chemical Langevin equations. *Chemical Physics*, 353(1–3):132–138, 2008.

[117] L. Szpruch and D. J. Higham. Comparing hitting time behaviour of Markov jump processes and their diffusion approximations. *Multiscale Modeling and Simulation*, 8 (2):605–621, 2010. 30

[118] C. M. Guldberg and P. Waage. Studies Concerning Affinity. *Forhandlinger i Videnskabs-Selskabet i Christiania*, 35, 1864. 31

[119] P. Deuflhard and F. Bornemann. *Scientific Computing with Ordinary Differential Equations*. Texts in Applied Mathematics, Vol. 42. Springer, 2002. 33, 57, 62

[120] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer, 2nd edition, 2004.

[121] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. Wiley, 2nd edition, 2008.

[122] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer, 2nd edition, 2011. 33, 57, 62

[123] T. G. Kurtz. Solutions of Ordinary Differential Equations as Limits of Pure Jump Markov Processes. *Journal of Applied Probability*, 7(1):49–58, April 1970. 34, 39

[124] T. G. Kurtz. Limit Theorems for Sequences of Jump Markov Processes Approximating Ordinary Differential Processes. *Journal of Applied Probability*, 8(2):344–356, 1971. 34, 39

[125] L. Bortolussi. Limit Behavior of the Hybrid Approximation of Stochastic Process Algebras. *Analytical and Stochastic Modeling Techniques and Applications ASMTA*, 6148:367–381, 2010. 40

[126] R. Durrett. *Probability: theory and examples*. Duxbury Press, 1995. 40

[127] S. Engblom. Computing the Moments of High Dimensional Solutions of the Master Equation. *Applied Mathematics and Computation*, 180(2):498–515, 2006. 43, 44, 45

[128] D. T. Gillespie and M. Mangel. Conditioned averages in chemical kinetics. *Journal of Chemical Physics*, 75(2):704–709, July 1981. 45

[129] J. Ruess, A. Milias-Argeitis, S. Summers, and J. Lygeros. Moment estimation for chemically reacting systems by extended Kalman filtering. *Journal of Chemical Physics*, 135 (16):165102, 2011. 45

[130] J. P. Hespanha. Moment closure for biochemical networks. *Proceedings of the IEEE 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP), St. Julians, Malta*, pages 142–147, 2008 (IEEE, New York, 2008). 45

[131] A. Singh and J. P. Hespanha. Approximate Moment Dynamics for Chemically Reacting Systems. *IEEE Transactions on Automatic Control*, 56(2):414–418, 2011. 45

[132] R. Kubo, K. Matsuo, and K. Kitahara. Fluctuation and Relaxation of Macrovariables. *Journal of Statistical Physics*, 9(1):51–96, 1973. 47, 49

[133] B. Gaveau, M. Moreau, and J. Tóth. *Variational and Extremum Principles in Macroscopic Systems: Part I - Theory*, chapter 15, pages 315–338. Elsevier Ltd., 1st edition, 2005. 47, 49

[134] M. I. Dykman, E. Mori, J. Ross, and P. M. Hunt. Large fluctuations and optimal paths in chemical kinetics. *Journal of Chemical Physics*, 100(8):5735–5750, April 1994. 47

[135] F. John. *Partial Differential Equations*. Springer, New York, 4th edition, 1981. 48, 49

[136] H. Gang. Stationary solution of master equations in the larg-system-size limit. *Physical Review A*, 36(12):5782–5790, December 1987. 49

[137] E. L. Haseltine. *Systems Analysis of Stochastic and Population Balance Models for Chemically Reacting Systems*. Dissertation, University of Wisconsin-Madison, `http://jbrwww.che.wisc.edu/theses/haseltine.pdf`, February 2005. 53

[138] M. von Kleist, S. Menz, and W. Huisinga. Drug-Class Specific Impact of Antivirals on the Reproductive Capacity of HIV. *PLoS Computational Biology*, 6(3):e1000720, 2010. 53, 62, 64, 68, 141, 144, 145

[139] M. von Kleist, S. Menz, H. Stocker, K. Arasteh, C. Schütte, and W. Huisinga. HIV Quasispecies Dynamics during Pro-Active Treatment Switching: Impact on Multi-Drug Resistance and Resistance Archiving in Latent Reservoirs. *PLoS ONE*, 6(3):e18204, 2011. 53, 66, 68, 70, 72, 74

[140] F. Zhang, M. Yeddanapudi, and P. J. Mosterman. Zero-Crossing Location and Detection Algorithms for Hybrid System Simulation. *Proceedings of the 17th IFAC World Congress, Seoul, Korea*, pages 7967–7972, July 6–11 2008. 62

[141] T. Park and P. I. Barton. State Event Location in Differential-Algebraic Models. *ACM Transactions on Modeling and Computer Simulation*, 6(2):137–165, April 1996. 62

[142] J. M. Esposito, V. Kumar, and G. J. Pappas. Accurate Event Detection for Simulating Hybrid Systems. *Proceedings of the 4th International Workshop on Hybrid Systems: Computation and Control (HSCC 2001), Heidelberg, Germany, Springer-Verlag, Lecture Notes in Computer Science*, 2034:204–217, 2001. 62

[143] R. P. Brent. *Algorithms for Minimization Without Derivatives*. Dover Publications, Inc. 2002. 62

[144] J. Martinet-Picado, E. Negredo, L. Ruiz, A. Shintani, C. R. Fumaz, C. Zala, P. Domingo, J. Vilaró, J. M. Libre, P. Viciana, K. Hertogs, C. Boucher, R. T. D'Aquila, and B. Clotet and the SWATCH Study Team. Alternation of Antiretroviral Drug Regimens for HIV Infection – A Randomized, Controlled Trial. *Annals of Internal Medicine*, 139(2):81–89, July 2003. 63, 64, 75

[145] E. Negredo, R. Paredes, J. Peraire, E. Pedrol, H. Côté, S. Gel, C. R. Fumoz, L. Ruiz, V. Abril, E. Rodriguez de Castro, C. Ochoa, J. Martinez-Picado, J. Montaner, C. Rey-Joly, and B. Clotet; Swatch Study Team. Alternation of antiretroviral drug regimens for HIV infection. Efficacy, safety and tolerability at week 96 of the Swatch Study. *Antiviral Therapy*, 9(6):889–893, December 2004. 63, 64, 75

[146] S. M. Hammer, K. E. Squires, M. D. Hughes, J. M. Grimes, L. M. Demeter, J. S. Currier, J. J. Eron, Jr., J. E. Feinberg, H. H. Balfour, L. R. Deyton, J. A. Chodakewitz, and M. A.

Fischl for the AIDS Clinical Trials Group 320 Study Team. A Controlled Trial of Two Nucleoside Analogues plus Indinavir in Persons with Human Immunodeficiency Virus Infection and CD4 Cell Counts of 200 per Cubic Millimeter or Less. *New England Journal of Medicine*, 337(11):725–733, September 1997. 63

[147] M. Hirsch, R. Steigbigel, S. Staszewski, J. Mellors, E. Sçerpella, B. Hirschel, J. Lange, K. Squires, S. Rawlins, A. Meibohm, and R. Leavitt for the Protocol 039 Study Group. A Randomized, Controlled Trial of Indinavir, Zidovudine, and Lamivudine in Adults with Advanced Human Immunodeficiency Virus Type 1 Infection and Prior Antiretroviral Therapy. *Journal of Infectious Diseases*, 180(3):659–665, September 1999. 63

[148] T.-W. Chun, D. Finzi, J. Margolick, K. Chadwick, D. Schwartz, and R. F. Siliciano. *In vivo* fate of HIV-1-infected T cells: Quantitative analysis of the transition to stable latency. *Nature Medicine*, 1(12):1284–1290, December 1995. 63

[149] T.-W. Chun, L. Stuyver, S. B. Mizell, L. A. Ehler, J. A. M. Mican, M. Baseler, A. L. Lloyd, M. A. Nowak, and A. S. Fauci. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *PNAS USA*, 94(24):13193–13197, November 1997.

[150] D. Finzi, M. Hermankova, T. Pierson, L. M. Carruth, C. Buck, R. E. Chaisson, T. C. Quinn, K. Chadwick, J. Margolick, R. Brookmeyer, J. Gallant, M. Markowitz, D. D. Ho, D. D. Richman, and R. F. Siliciano. Identification of a Reservoir for HIV-1 in Patients on Highly Active Antiretroviral Therapy. *Science*, 278(5341):1295–1300, November 1997.

[151] J. K. Wong, M. Hezareh, H. F. Günthard, D. V. Havlir, C. C. Ignacio, C. A. Spina, and D. D. Richman. Recovery of Replication–Competent HIV Despite Prolonged Suppression of Plasma Viremia. *Science*, 278(5341):1291–1295, November 1997.

[152] D. Finzi, J. Blankson, J. D. Siliciano, J. B. Margolick, K. Chadwick, T. Pierson, K. Smith, J. Lisziewicz, F. Lori, C. Flexner, T. C. Quinn, R. E. Chaisson, E. Rosenberg, B. Walker, S. Gange, J. Gallant, and R. F. Siliciano. Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nature Medicine*, 5(5):512–517, May 1999. 72, 144

[153] C. C. Carter, A. Onafuwa-Nuga, L. A. McNamara, J. Riddell IV, D. Bixby, M. R. Savona, and K. L. Collins. HIV-1 infects multipotent progenitor cells causing cell death and establishing latent cellular reservoirs. *Nature Medicine*, 16(4):446–451, April 2010. 63

[154] T.-W. Chun, L. Carruth, D. Finzi, X. Shen, J. A. DiGiuseppe, H. Taylor, M. Hermankova, K. Chadwick, J. Margolick, T. C. Quinn, Y.-H. Kuo, R. Brookmeyer, M. A. Zeiger, P. Barditch-Crovo, and R. F. Siliciano. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature*, 387(6629):183–188, May 1997. 63, 72, 144

[155] T. Pierson, J. McArthur, and R. F. Siliciano. Reservoirs for HIV-1: Mechanisms for Viral Persistence in the Presence of Antiviral Immune Responses and Antiretroviral Therapy. *Annual Review of Immunology*, 18:665–708, 2000. 63

[156] J. D. Siliciano and R. F. Siliciano. A long-term latent reservoir for HIV-1: discovery and clinical implications. *Journal of Antimicrobial Chemotherapy*, 54(1):6–9, July 2004. 63, 64

[157] O. Lambotte, M.-L. Chaix, B. Gubler, N. Nasreddine, C. Wallon, C. Goujard, C. Rouzioux, Y. Taoufik, and J.-F. Delfraissy. The lymphocyte HIV reservoir in patients on long-term HAART is a memory of virus evolution. *AIDS*, 18(8):1147–1158, May 2004. 63, 64

[158] J. Reekie, A. Mocroft, B. Ledergerber, M. Beniowski, B. Clotet, J. van Lunzen, A. Chiesi, C. Pradier, L. Machala, and J. D. Lundgren for the EuroSIDA Study Group. History of viral suppression on combination antiretroviral therapy as a predictor of virological failure after a treatment change. *HIV Medicine*, 11(7):469–478, August 2010. 63

[159] D. D. Ho, A. U. Neumann, A. S. Perelson, W. Chen, J. M. Leonard, and M. Markowitz. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, 373(6510):123–126, January 1995. 63

[160] W. Zimmerli and P. E. Ochsner. Management of Infection Associated with Prosthetic Joints. *Infection*, 31(2):99–108, March 2003. 63

[161] D. Hortskotte, F. Follath, E. Gutschik, M. Lengyel, A. Oto, A. Pavie, J. Soler-Soler, G. Thiene, and A. von Graevenitz. Esc Guidlines: Guidelines on Prevention, Diagnosis and Treatment of Infective Endocarditis Executive Summary; The Task Force on Infective Endocarditis of the European Society of Cardiology. *European Hert Journal*, 25(3): 267–276, February 2004.

[162] H. M. Blumberg, M. K. Leonard, and R. M. Jasmer. Update on the Treatment of Tuberculosis and Latent Tuberculosis Infection. *JAMA*, 293(22):2776–2784, June 2005. 63

[163] M. E. Curlin, S. Iyer, and J. E. Mittler. Optimal Timing and Duration of Induction Therapy for HIV-1 Infection. *PLoS Computational Biology*, 3(7):e133, July 2007. 63, 144

[164] L. M. Catanzaro, J. C. Slish, R. DiCenzo, and G. D. Morse. Drug interactions with antiretrovirals. *Current HIV/AIDS Reports*, 1(2):89–96, June 2004. 63

[165] A. Carr. Toxicity of antiretroviral therapy and implications for drug development. *Nature Reviews Drug Discovery*, 2(8):624–634, August 2003. 63

[166] R. M. D'Amato, R. T. D'Aquila, and L. M. Wein. Management of Antiretroviral Therapy for HIV Infection: Modelling When to Change Therapy. *Antivir Ther*, 3(3):147–158, 1998. 64, 75

[167] L. Sherr, F. Lampe, S. Norwood, H. Leake-Date, M. Fisher, S. Edwards, G. Arthur, J. Anderson, S. Zetler, M. Johnson, and R. Harding. Successive switching of antiretroviral therapy is associated with high psychological and physical burden. *International Journal of STD & AIDS*, 18(10):700–704, October 2007. 64

[168] Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. Department of Health and Human Services. `http://www.aidsinfo.nih.gov/ContentFiles/AdultandAdolescentGL.pdf` (accessed 01.10.2010). pages 1–161, December 2009. 64, 69, 70, 71

[169] R. Klopfenstein. Numerical differentiation formulas for stiff systems of ordinary differential equations. *RCA Review*, 32:447–462, 1971. 66

[170] L. Shampine and M. Reichelt. The Matlab ODE Suite. *SIAM Journal on Scientific Computing*, 18:1–22, 1997. 66

[171] M. Fabbiani, S. D. Giambenedetto, L. Bracciale, A. Bacarelli, E. Ragazzoni, R. Cauda, P. Navarra, and A. D. Luca. Pharmacokinetic variability of antiretroviral drugs and correlation with virological outcome: 2 years of experience in routine clinical practice. *Journal of Antimicrobial Chemotherapy*, 64(1):109–117, July 2009. 68

[172] F. Clavel and A. J. Hance. HIV Drug Resistance. *New England Journal of Medicine*, 350 (10):1023–1035, March 2004. 68

[173] F. Clavel. HIV Resistance to Raltegravir. *European Journal of Medical Research*, 14 Suppl. 3:47–54, November 2009. 68

[174] J. N. Blankson, D. Finzi, T. C. Pierson, B. P. Sabundayo, K. Chadwick, J. B. Margolick, T. C. Quinn, and R. F. Siliciano. Biphasic Decay of Latently Infected CD4+ T Cells in Acute Human Immunodeficiency Virus Type 1 Infection. *Journal of Infectious Diseases*, 182(6):1636–1642, December 2000. 72, 144

[175] J.-M. Fondere, J.-F. Planas, M.-F. Huguet, V. Baillat, F. Bolos, J. Reynes, and J.-P. Vendrell. Enumeration of latently infected CD4+ T cells from HIV-1-infected patients using an HIV-1 antigen ELISPOT assay. *Journal of Clinical Virology*, 29(1):33–38, January 2004. 72, 144

[176] D. D. Richman. Principles of HIV resistance testing and overview of assay performance characteristics. *Antiviral Therapy*, 5(1):27–31, March 2000. 72, 75

[177] N. Eriksson, L. Pachter, Y. Mitsuya, S.-Y. Rhee, C. Wang, B. Gharizadeh, M. Ronaghi, R. W. Shafer, and N. Beerenwinkel. Viral Population Estimation Using Pyrosequencing. *PLoS Computational Biology*, 4(5):e1000074, April 2008. 72, 73, 75

[178] J. Archer, M. S. Braverman, B. E. Taillon, B. Desany, I. James, P. R. Harrigan, M. Lewis, and D. L. Robertson. Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4-using HIV-1 with ultra-deep pyrosequencing. *AIDS*, 23(10):1209–1218, June 2010.

[179] H. Ji, N. Massé, S. Tyler, B. Liang, Y. Li, H. Merks, M. Graham, P. Sandstrom, and J. Brooks. HIV Drug Resistance Surveillance Using Pooled Pyrosequencing. *PLoS One*, 5(2):e9263, 2010. 73, 75

[180] M. Nijhuis, R. Schuurman, D. de Jong, J. Erickson, E. Gustchina, J. Albert, P. Schipper, S. Gulnik, and C. A. Boucher. Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy. *AIDS*, 13 (17):2349–2359, December 1999. 73

[181] P. Arora and N. M. Dixit. Timing the Emergence of Resistance to Anti-HIV Drugs with Large Genetic Barriers. *PLoS Computational Biology*, 5(3):e1000305, March 2009. 73

[182] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho. HIV-1 Dynamics in Vivo: Virion Clearance Rate, Infected Cell Life-Span, and Viral Generation Time. *Science*, 271(5255):1582–1586, March 1996. 74

[183] A. S. Perelson, P. Essunger, Y. Cao, M. Vesanen, A. Hurley, K. Saksela, M. Markowitz, and D. D. Ho. Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature*, 387(6629):188–191, May 1997.

[184] M. Markowitz, M. Louie, A. Hurley, E. Sun, M. D. Mascio, A. S. Perelson, and D. D. Ho. A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. *Journal of Virology*, 77(8):5037–5038, April 2003. 74, 144

[185] R. J. Smith, J. T. Okano, J. S. Kahn, E. N. Bodine, and S. Blower. Evolutionary Dynamics of Complex Networks of HIV Drug-Resistant Strains: The Case of San Francisco. *Science*, 327(5966):697–701, February 2010. 74

[186] J. Goshn, I. Pellegrin, C. Goujard, C. Deveau, J.-P. Viard, J. Galimand, M. Harzic, C. Tamalet, L. Meyer, C. Rouzioux, and M.-L. Chaix. HIV-1 resistant strains acquired at the time of primary infection massively fuel the cellular reservoir and persist for lengthy periods of time. *AIDS*, 20(2):159–170, January 2006. 74

[187] J. A. Bartlett. Addressing the Challenges of Adherence. *Journal of Acquired Immune Deficiency Syndromes*, 29:S2–S10, February 2002. 74

[188] B. Ramratnam, S. Bonhoeffer, J. Binley, A. Hurley, L. Zhang, J. E. Mittler, M. Markowitz, J. P. Moore, A. S. Perelson, and D. D. Ho. Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. *Lancet*, 354(9192): 1782–1785, November 1999. 75

[189] K. Ball, T. G. Kurtz, L. Popovic, and G. Rempala. Asymptotic analysis of multiscale approximations to reaction networks. *The Annals of Applied Probability*, 16(4):1925–1961, 2006. 83

[190] E. Vanden-Eijnden. FAST COMMUNICATIONS: Numerical Techniques for Multi-Scale Dynamical Systems with Stochastic Effects. *Communications in Mathematical Sciences*, 1(2):385–391, 2003. 86

[191] F. Didier, T. A. Henzinger, M. Mateescu, and V. Wolf. Fast Adaptive Uniformization of the Chemical Master Equation. *Proceedings of HIBI, 2009 International Workshop on High Performance Computational Systems Biology*, pages 118–127, 2009. 91

[192] J. Adler and B. Templeton. The Effect of Environmental Conditions on the Motility of *Escherichia coli*. *Journal of General Microbiology*, 46(2):175–184, February 1967. 112

[193] C. D. Amsler, M. Cho, and P. Matsumura. Multiple Factors Underlying the Maximum Motility of *Escherichia coli* as Cultures Enter Post-Exponential Growth. *Journal of Bacteriology*, 175(19):6238–6244, October 1993.

[194] K. Zhao, M. Liu, and R. R. Burgess. Adaptation in bacterial flagellar and motility systems: from regulon members to 'foraging'-like behavior in E. coli. *Nucleic Acids Research*, 35(13):4441–4452, June 2007.

[195] C. Barembruch and R. Hengge. Cellular levels and activity of the flagellar sigma factor FliA of *Escherichia coli* are controlled by FlgM-modulated proteolysis. *Molecular Microbiology*, 65(1):76–89, July 2007. 113, 115, 121, 122, 123, 126, 153, 154, 155

[196] H. Weber, C. Pesavento, A. Possling, and R. Hengge. Cyclic-di-GMP-mediated signalling within the σ^S network of *Escherichia coli*. *Molecular Microbiology*, 62(4):1014–1034, November 2006.

[197] R. Hengge. Principles of c-di-GMP signalling in bacteria. *Nature Reviews Microbiology*, 7:263–273, 2009. 112, 133

[198] R. M. Macnab. The Bacterial Flagellum: Reversible Rotary Propellor and Type III Export Apparatus. *Journal of Bacteriology*, 181(23):7149–7153, December 1999. 112, 115

[199] S. Kalir, J. McClure, K. Pabbaraju, C. Southward, M. Ronen, S. Leibler, M. G. Surette, and U. Alon. Ordering Genes in a Flagella Pathway by Analysis of Expression Kinetics from Living Bacteria. *Science*, 292(5524):2080–2083, June 2001. 112, 113, 133

[200] S. Saini, P. D. Aldridge, and C. V. Rao. Role of Feedback Loops in Regulating Flagellar Gene Expression Dynamics in *S. Typhimurium*. *Proceedings of Foundations of Systems Biology in Engineering (FOSBE 2009)*, pages 48–51, August 2009. 112, 131

[201] R. M. Macnab. Flagella and motility. In F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. M. Riley, M. Schaechter, and H. E. Umbarger, editors, *Escherichia coli and Salmonella: Cellular and Molecular Biology*, volume 2, pages 123–145. ASM Press, Washington D.C., 2 edition, 1996. 112

[202] G. S. Chilcott and K. T. Hughes. Coupling of Flagellar Gene Expression to Flagellar Assembly in *Salmonella enterica* Serovar Typhimurium and *Escherichia coli*. *Microbiology and Molecular Biology Reviews*, 64(4):694–708, December 2000.

[203] F. F. V. Chevance and K. T. Hughes. Coordinating assembly of a bacterial macromolecular machine. *Nature Reviews Microbiology*, 6:455–465, June 2008. 112, 115

[204] X. Liu and P. Matsumura. The FlhD/FlhC Complex, a Transcriptional Activator of the *Escherichia coli* Flagellar Class II operons. *Journal of Bacteriology*, 176(23):7345–7351, December 1994. 112, 115

[205] T. K. Kundu, S. Kusano, and A. Ishihama. Promoter Selectivity of *Escherichia coli* RNA Polymerase σ^f Holoenzyme Involved in Transcription of Flagellar and Chemotaxis Genes. *Journal of Bacteriology*, 179(13):4264–4269, July 1997. 112

[206] X. Liu and P. Matsumura. An alternative sigma factor controls transcription of flagellar class-III operons in *Escherichia coli*: gene sequence, overproduction, purification and characterization. *Gene*, 164(1):81–84, October 1995. 112

[207] K. T. Hughes, K. L. Gillen, M. J. Semon, and J. E. Karlinsey. Sensing structural intermediates in bacterial flagellar assembly by export of a negative regulator. *Science*, 262 (5137):1277–1280, November 1993. 112

[208] P. D. Aldridge, J. E. Karlinsey, C. Aldridge, C. Birchall, D. Thompson, J. Yagasaki, and K. T. Hughes. The flagellar-specific transcription factor, σ^{28}, is the Type III secretion chaperone for the flagellar-specific anti-σ^{28} factor FlgM. *Genes & Development*, 20(16): 2315–2326, August 2006. 112, 115, 126, 154

[209] J. E. Karlinsey, S. Tanaka, V. Bettenworth, S. Yamaguchi, W. Boos, S.-I. Aizawa, and K. T. Hughes. Completion of the hook–basal body complex of the *Salmonella typhimurium* flagellum is coupled to FlgM secretion and *fliC* transcription. *Molecular Microbiology*, 37(5):1220–1231, September 2000. 113, 123

[210] S. Saini, E. Floess, C. Aldridge, J. Brown, P. D. Aldridge, and C. V. Rao. Continuous control of flagellar gene expression by the σ^{28}–FlgM regulatory circuit in *Salmonella enterica*. *Molecular Microbiology*, 79(1):264–278, November 2011. 113

[211] S. Kalir and U. Alon. Using a Quantitative Blueprint to Reprogram the Dynamics of the Flagella Gene Network. *Cell*, 117(6):713–720, June 2004. 113

[212] J. Stelling, U. Sauer, Z. Szallasi, F. J. D. III, and J. Doyle. Robustness of cellular functions. *Cell*, 118(6):675–685, September 2004. 113

[213] H. Kitano. Towards a theory of biological robustness. *Molecular Systems Biology*, 3 (137):1–7, 2007. 113

[214] R. Liu and H. Ochman. Origins of Flagellar Gene Operons and Secondary Flagellar Systems. *Journal of Bacteriology*, 189(19):7098–7104, October 2007. 113

[215] M. S. Chadsey, J. E. Karlinsey, and K. T. Hughes. The flagellar anit-σ factor FlgM actively dissociates *Salmonella typhimurium* σ^{28} RNA polymerase holoenzyme. *Genes & Development*, 12:3123–3136, 1998. 114, 115, 154, 155, 156

[216] H. Maeda, N. Fujita, and A. Ishihama. Competition among seven *Escherichia coli* σ subunits: relative binding affinities to the core RNA polymerase. *Nucleic Acids Research*, 28(18):3497–3503, 2000. 115, 155, 156

[217] T. Yura, M. Kanemori, and M. T. Morita. The heat shock response: regulation and function. In G. Storz and R. Hengge-Aronis, editors, *Bacterial Stress Response*, pages 3–18. Washington, DC: American Society for Microbiology Press, 2000. 115

[218] R. Hengge-Aronis. Signal Transduction and Regulatory Mechanisms Involved in Control of the σ^s (RpoS) Subunit of RNA Polymerase. *Microbiology and Molecular Biology Reviews*, 66(3):373–395, September 2002. 115

[219] C. Condon, S. French, C. Squires, and C. L. Squires. Depletion of functional ribosomal RNA opreons in *Escherichia coli* causes increased expression of the remaining intact copies. *The EMBO Journal*, 12(11):4305–4315, November 1993. 118, 157, 158

[220] S. Yamamoto and K. Kutsukake. FliT Acts as an Anti-$FlhD_2C_2$ Factor in the Transcriptional Control of the Flagellar Regulon in *Salmonella enterica* Serovar Typhiumurium. *Journal of Bacteriology*, 188(18):6703–6708, September 2006. 131, 132

[221] E. D. Brutinel and T. L. Yahr. Control of gene expression by type III secretory activity. *Current Opinion in Microbiology*, 11(2):128–133, April 2008. 131, 132

[222] R. Hengge. The general stress response in gram-negative bacteria. In G. Storz and R. Hengge, editors, *Bacterial Stress Response*, number 2nd edition, pages 251–289. ASM Press, Washington, DC, 2011. 132

[223] C. Pesavento, G. Becker, N. Sommerfeldt, A. Possling, N. Tschowri, A. Mehlis, and R. Hengge. Inverse regulatory coordination of motility and curli-mediated adhesion in *Escherichia coli*. *Genes & Development*, 22(17):2434–2446, 2008. 132, 133

[224] L. M. Mansky and H. M. Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of Virology*, 69(8):5087–5094, August 1995. 142, 143, 144

[225] M. von Kleist. *Combining Pharmacology and Mutational Dynamics to Understand and Combat Drug Resistance in HIV*. Dissertation, Freie Universität Berlin, `http://www.hamilton.ie/publications/MaxvonKleist_2009_PhDThesis.pdf`, December 2009. 144

[226] X. Wei, S. K. Ghosh, M. E. Taylor, V. A. Johnson, E. A. Emini, P. Deutsch, J. D. Lifson, S. Bonhoeffer, M. A. Nowak, and B. H. Hahn. Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, 373(6510):117–122, January 1995. 144

[227] A. R. Sedaghat, R. F. Siliciano, and C. O. Wilke. Constraints on the dominant mechanism for HIV viral dynamics in patients on raltegravir. *Antiviral Therapy*, 14(2):263–271, 2009. 144

[228] D. S. Callaway and A. S. Perelson. HIV-1 infection and low steady state viral loads. *Bulletin of Mathematical Biology*, 64(1):29–64, January 2002. 144

[229] K. K. Koelsch, L. Liu, R. Haubrich, S. May, D. Havlir, H. F. Günthard, C. C. Ignacio, P. Campos-Soto, S. J. Little, R. Shafer, G. K. Robbins, R. T. D'Aquila, Y. Kawano, K. Young, P. Dao, C. A. Spina, D. D. Richman, and J. K. Wong. Dynamics of total, linear nonintegrated, and integrated HIV-1DNA in vivo and in vitro. *Journal of Infectious Diseases*, 197(3):411–419, February 2008. 144

[230] Y. Zhou, H. Zhang, J. D. Siliciano, and R. F. Siliciano. Kinetics of human immunodeficiency virus type 1 decay following entry into resting CD4+ T cells. *Journal of Virology*, 79(4):2199–2210, February 2005. 144

[231] T. C. Pierson, Y. Zhou, T. L. Kieffer, C. T. Ruff, C. Buck, and R. F. Siliciano. Molecular characterization of preintegration latency in human immunodeficiency virus type 1 infection. *Journal of Virology*, 76(17):8518–8531, Sep 2002. 144

[232] A. R. Sedaghat, J. B. Dinoso, L. Shen, C. O. Wilke, and R. F. Siliciano. Decay dynamics of HIV-1 depend on the inhibited stages of the viral life cycle. *PNAS*, 105(12):4832–4837, March 2008. 144

[233] L. Zhang, B. Ramratnam, K. Tenner-Racz, Y. He, M. Vesanen, S. Lewin, A. Talal, P. Racz, A. S. Perelson, B. T. Korber, M. Markowitz, Y. Guo, M. Duran, A. Hurley, J. Tsay, Y.-C. Huang, C.-C. Wang, and D. D. Ho. Quantifying Residual HIV-1 Replication in Patients

Receiving Combination Antiretroviral Therapy. *New England Journal of Medicine*, 340 (21):1605–1613, May 1999. 144

[234] B. Ramratnam, J. E. Mittler, L. Zhang, D. Boden, A. Hurley, F. Fang, C. A. Macken, A. S. Perelson, M. Markowitz, and D. D. Ho. The decay of the latent reservoir of replication-competent HIV-1 is inversely correlated with the extent of residual viral replication during prolonged anti-retroviral therapy. *Nature Medicine*, 6(1):82–85, January 2000. doi: 10.1038/71577.

[235] J. D. Siliciano, J. Kajdas, D. Finzi, T. C. Quinn, K. Chadwick, J. B. Margolick, C. Kovacs, S. J. Gange, and R. F. Siliciano. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nature Medicine*, 9(6):727–728, Jun 2003.

[236] T.-W. Chun, J. S. Justement, S. Moir, C. W. Hallahan, J. Maenza, J. I. Mullins, A. C. Collier, L. Corey, and A. S. Fauci. Decay of the HIV Reservoir in Patients Receiving Antiretroviral Therapy for Extended Periods: Implications for Eradication of Virus. *Journal of Infectious Diseases*, 195(12):1762–1764, June 2007. 144

[237] J. Coffin, F. Maldarelli, S. Palmer, A. Wiegand, S. Brun, D. Kempf, M. King, G. Hanna, and J. Mellors. Long-term persistence of low-level HIV-1 in patients on suppressive antiretroviral therapy. Abstract 169, 13th Conference on Retroviruses and Opportunistic Infections, 5–8th February 2006, Denver, Colorado. `http://www.retroconference.org/2006/Abstracts/28061.htm`. 144

[238] J. A. P. Heesterbeek. A Brief History of ro and a Recipe for its Calculation. *Acta Biotheoretica*, 50(3):189–204, 2002. 145

[239] J. M. Heffernan, R. J. Smith, and L. M. Wahl. Perspectives on the basic reproductive ratio. *Journal of the Royal Society Interface*, 2(4):281–293, September 2005. doi: 10.1098/rsif.2005.0042. 145

[240] P.-S. Laplace. *Théorie Analytique des Probabilités*, volume 1. Courcier, Paris, 1820 (Reprint in Complete Works, volume 7, Gauthier-Villars, Paris, 1886). 147

[241] N. G. de Bruijn. *Asymptotic Methods in Analysis*. Dover, New York, 1981. 147

[242] W. Fulks and J. O. Sather. Asymptotics II: Laplaces method for multiple integrals. *Pacific J. Math.*, 11(1):185–192, 1961.

[243] J. Wojdylo. On the coefficients that arise from Laplace's method. *Journal of Computational and Applied Mathematics*, 196(1):241–266, November 2006.

[244] W. D. Kirwin. Higher Asymptotics of Laplace's Approximation. to appear in Asymp. Anal., arXiv:math.CA/0810.1700, 2010. 147

[245] G. F. Carrier, M. Krook, and C. E. Pearson. *Functions of a Complex Variable: Theory and Technique*. SIAM, 2005. 147

[246] I. L. Grigorova, N. J. Phleger, V. K. Mutalik, and C. A. Gross. Insights into transcriptional regulation and σ competition from an equilibrium model of RNA polymerase binding to DNA. *PNAS*, 103(14):5332–5337, April 2006. 154

[247] H. Bremer and P. P. Dennis. Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate. In F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. M. Riley, M. Schaechter, and H. E. Umbarger, editors, *Escherichia coli and Salmonella: Cellular and Molecular Biology*, volume 2, pages 1553–1569. ASM Press, Washington D.C., 2nd edition, 1996. 154, 157, 158

[248] L. Claret and C. Hughes. Functions of the Subunits in the $FlhD_2C_2$ Transcriptional Master Regulator of Bacterial Flagellum Biogenesis and Swarming. *Journal of Molecular Biology*, 303(4):467–478, November 2000. 157, 158

[249] G. P. Stafford, T. Ogi, and C. Hughes. Binding and transcriptional activation of non-flagellar genes by the *Escherichia coli* flagellar master regulator $FlhD_2C_2$. *Microbiology*, 151(6):1779–1788, June 2005. 157, 158

[250] Y. Harada, T. Funatsu, K. Murakami, Y. Nonoyama, A. Ishihama, and T. Yanagida. Single-Molecule Imaging of RNA Polymerase-DNA Interactions in Real Time. *Biophysical Journal*, 76(2):709–715, February 1999. 157, 158

[251] L. Bai, T. J. Santangelo, and M. D. Wang. Single-Molecule Analysis of RNA Polymerase Transcription. *The Annual Review of Biophysics and Biomolecular Structure*, 35:343–360, February 2006. 157, 158

[252] J. E. Karlinsey, H.-C. T. Tsui, M. E. Winkler, and K. T. Hughes. Flk Couples *flgM* Translation to Flagellar Ring Assembly in *Salmonella typhimurium*. *Journal of Bacteriology*, 180(20):5384–5397, October 1998. 157, 158