HELMHOLTZ ASSOCIATION
Alliance on Systems Biology
HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE UNIVERSITÄT MÜNCHEN

STUDYING THE DYNAMICS OF STOCHASTIC BIOCHEMICAL PROCESSES USING GENERALISED MOMENT CLOSURE APPROXIMATIONS

ATEFEH KAZEROONIAN

December 2017

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Mathematik — Lehrstuhl M12 (Mathematische Modellierung biologischer Systeme)

Studying the Dynamics of Stochastic Biochemical Processes Using Generalised Moment Closure Approximations

Atefeh Kazeroonian

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Prof. Dr. Silke Rolles

Prüfer der Dissertation:

1. Prof. Dr. Fabian Theis
2. Prof. Dr. Manfred Claassen
3. Prof. Dr. Oliver Junge

Die Dissertation wurde am 10.08.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 08.05.2018 angenommen.

Acknowledgement

First of all, I would like to thank Fabian Theis and Jan Hasenauer for giving me the opportunity to work on this thesis, and for their supervision and helpful insights throughout my PhD work. I am also grateful to Manfred Claassen who gave me valuable feedback whenever I asked for it.

I also like to thank Fabian for his kind support and all the opportunities he provided me with to grow as a researcher. Fabian, thank you for creating a friendly and warm atmosphere at ICB. I surely learnt a lot from you as a leader.

Furthermore, I would like to thank Christiane Fuchs, Carsten Marr and the rest of my colleagues at ICB for interesting scientific discussions and feedback; it certainly is a great chance to work with people who can help one develop further professionally.

Next, I would like to thank my friends at ICB for making my PhD times a memorable part of my life, filled with precious memories. I certainly had more fun with you all than one can reasonably handle in one PhD life. Thinking about it all, I can only smile and cherish many moments we lived together. You know who you are.

I want to thank my family, and most of all my parents, whose love, understanding, devotion and care is beyond me. It tears me up even imagining how anyone could so selflessly be there for someone else, the way you have always been there for me. I find peace with you. My little brother, you are probably my favourite person in the entire world; I hope you know how much you mean to me.

And finally, Babak, there are no words with which I can possibly describe what you are to me. All I know is that everything I achieved in my life, I owe more to you than myself. You are the only one who truly knows me; all of me.

Abstract

Systems biology exploits experimental and computational techniques to explain the behaviour of biological systems as a whole. Complex behaviours of biological systems, e.g., single cells, arise from the interaction of underlying mechanisms that have diverse and complicated functionalities. These underlying mechanisms are extensively described by means of mechanistic models. The intrinsic stochasticity in many cellular processes has been shown to have functional roles to increase the robustness of cellular organisms. Therefore, capturing the stochasticity is crucial for a true understanding of the behaviour of biological systems. A broad range of cellular processes are modelled by means of biochemical reaction networks. The exact temporal evolution of the probability distribution over the state space of a biochemical reaction network is governed by the chemical master equation (CME). Due to the intractability of solving the CME, or approximations to it, for most realistic systems, a field of systems biology research has been devoted to developing mesoscopic descriptions which describe the stochastic biochemical kinetics in terms of the statistical moments of the solution of the CME.

In this thesis, my aim was to establish robust, reliable and feasible mesoscopic approximative methods that could be used in the formalism of systems biology to learn about the underlying mechanisms of stochastic biochemical processes. To accomplish this goal, I asked what key obstacles were in the way of obtaining mesoscopic descriptions for generic biochemical reaction networks. A key bottleneck for the feasibility of mesoscopic approaches is their size scalability with respect to the number of species in the reaction network: even if only the second-order moments of the CME solution are of interest, the standard mesoscopic approaches are of quadratic scaling which is prohibitive for their applicability to large-scale biochemical reaction networks. In this thesis, I proposed a model reduction based on the topological structure of the reaction network that yields near-linear scalability, and therefore, enables capturing of heterogeneity in large-scale biological processes. In addition, complex network structures, e.g., nonlinear kinetics that do not follow the law of mass action and/or copy-number scale separation, require special

treatments to ensure reliable approximation of the system dynamics. I addressed these challenges by developing special moment closure approximations—e.g., employing Taylor series expansion for the handling of non-mass action kinetics—which manifested superior performance over standard moment approximation methods in our simulation studies.

Having achieved feasible mathematical descriptions for a wide range of realistic biological processes, I asked how this framework can be efficiently utilised for answering systems biology questions. Due to the variety of modelling approaches, along with the absence of corresponding a priori error estimates, the optimal choice of the descriptive method cannot easily be identified. In this regard, I believed that the systems biology research would benefit from a comprehensive platform where a variety of relevant approaches can be easily simulated and compared to assess their performances for the problem at hand. To this end, I, together with my colleagues, developed a simulation platform enabling efficient simulation and comprehensive comparisons across a broad range of modelling approaches. This simulation toolbox greatly enhances the accessibility of available modelling approaches and facilitates the integration of the resulting descriptive models for the inference of stochastic biochemical kinetics. In particular, for the latter, I investigated how mesoscopic approximations can be used to better inform the inference of underlying mechanisms. Our studies indicated that the information contained in higher-order statistical moments, can increase certainty and predictive power of parameter estimation results. Taking all the contributions above together, I addressed various missing parts for achieving reliable and predictive mesoscopic descriptions that can improve our understanding of the behaviour of biological systems.

List of contributed articles

i) A. Kazeroonian, J. Hasenauer, and F. J. Theis. **Parameter Estimation for Stochastic Biochemical Processes: A Comparison of Moment Equation and Finite State Projection.** *In Proceedings of* 10^{th} *International Workshop on Computational Systems Biology (WCSB), Tampere, Finland, pages 66-73, 2013.*

ii) A. Kazeroonian, F. J. Theis, and J. Hasenauer. **Modeling of stochastic biological processes with non-polynomial propensities using non-central conditional moment equation.** *IFAC Proceedings Volumes, Volume 47, Issue 3, 2014, Pages 1729-1735.*

iii) A. Kazeroonian, F. Fröhlich, A. Raue, F. J. Theis, J. Hasenauer. **CERENA: ChEmical REaction Network Analyzer – A Toolbox for the Simulation and Analysis of Stochastic Chemical Kinetics.** *PLOS ONE 11(1): e0146732, 2016.*

iv) A. Kazeroonian, F. J. Theis, J. Hasenauer. **A scalable moment-closure approximation for large-scale biochemical reaction networks**. *Bioinformatics 2017; 33 (14): i293-i300.*

Contents

Chapter 1

Introduction

Biological systems are complex systems that are comprised of many elements with diverse and complicated functionalities. Systems biology combines experimental and computational techniques to explain the collective behaviour of biological systems as a result of the interaction of underlying mechanisms [Kitano, 2002]. Starting from early works such as [Hodgkin and Huxley, 1952], an active field in systems biology research has been the application of mathematics to describe the biological processes, e.g., gene expression and signal transduction, at the single cell level [Elowitz et al., 2002, Raj et al., 2006, Schöberl et al., 2002]. Mechanistic models are a useful asset in this regard that utilise mathematics to explain the dynamics of single cells and learn about the underlying mechanisms that give rise to particular phenomena. A careful consideration of the properties of the biological system at hand, and a clear definition of the research questions to be answered, are essential for the choice of a suitable mechanistic model out of the broad range of available choices. Stochastic models are required for the treatment of many biological questions, including those considered in this thesis work.

Many cellular processes, such as gene expression, signal transduction and cell fate decisions, are subject to intrinsic stochasticity due to stochastic events, e.g., bursty gene expression [Raser and O'Shea, 2004, Raj and van Oudenaarden, 2008]. As a result, isogenic cells can behave differently under the same conditions, for instance in response to a stimulus, and give rise to heterogeneous cell populations [Elowitz et al., 2002]. The stochasticity, sometimes referred to as *noise*, in cellular mechanisms has been shown to have a crucial functional role, for example to increase their robustness in changing environments [Raj and van Oudenaarden, 2008, Eldar and Elowitz, 2010]. Hence, in modelling of biological systems, capturing the intrinsic noise is essential for a true understanding

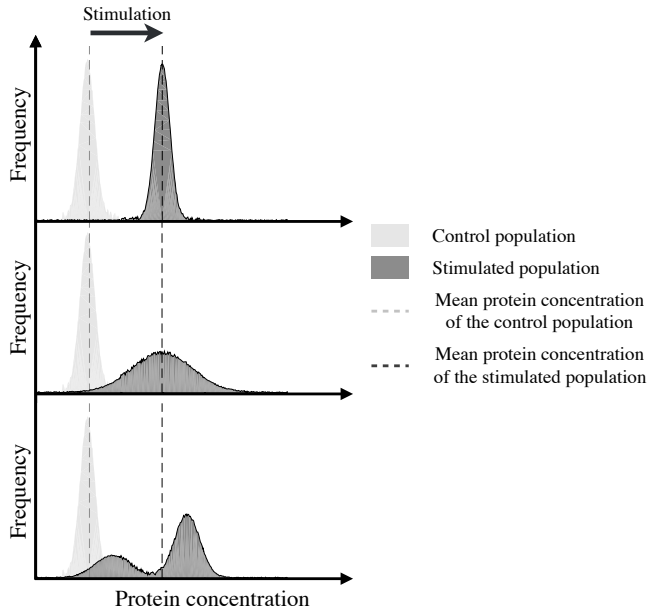


Figure 1.1: **Qualitatively different responses of cell populations to a stimulus might seem identical on the population average level.** Top and middle panels, respectively, show a cell population that responds homogeneously and heterogeneously to a stimulus. Bottom panel depicts a cell population comprised of two subpopulations with distinct responses to the same stimulus. The population-averaged response (dashed line), however, is indistinguishable for the three scenarios.

of biological systems and providing insight into their underlying mechanisms. Figure 1.1 demonstrates how qualitatively distinct behaviours can become indistinguishable when neglecting the heterogeneity and merely analysing the average behaviour of the biological system.

In my doctoral research, I wanted to use mathematical models to capture the dynamics of stochastic biochemical processes that are relevant for typical systems biology questions. In particular, I wanted to use reliable and feasible modelling approaches that could be applied to generic biochemical reaction networks, and efficiently be utilised for further analyses such as parameter estimation and model selection. Studying several biochemical processes, e.g., models of gene expression and signal transduction, I encountered biological scenarios that available approaches failed to handle reliably. In the following, I first provide an introductory overview of these modelling approaches for describing stochastic biochemical kinetics, and then point out the shortcomings of these approaches that motivated the key questions of this thesis. In Section 1.2, I give a brief overview of how I addressed these questions in the course of my PhD work.

1.1 Dynamics of biochemical reaction networks

Many biological processes, such as gene regulation, signal transduction and metabolism, are modelled as (bio)chemical reaction networks (CRNs) consisting of biochemical species that undergo chemical reactions. The instantaneous abundance of species in the network, that defines the state of the CRN, changes via the firing of chemical reactions [Klipp et al., 2005]. Due to the probabilistic nature of chemical reactions, the state of a CRN evolves stochastically in time. The temporal evolution of the state of a CRN is a *memoryless* stochastic process that can take on discrete values, i.e. molecular counts. Since the evolution of this process occurs in continuous time, its dynamics is often modelled by continuous-time Markov chains (CTMCs) [Norris, 1997] (see Section 2.1.2). Given an initial condition, at each point in time, the CTMC can be found in any of a set of all possible states with an associated probability. The set of all states that can be reached via the firing of feasible chemical reactions, constitute the state space of a CTMC corresponding to a CRN. Assuming a well-mixed and thermally equilibrated reaction environment, the temporal evolution of the probability distribution over the state space of a CTMC is exactly described by the Chemical Master Equation (CME) [Gillespie, 1992a]. Apart from special cases [Jahnke and Huisinga, 2007], finding a solution to the CME is usually intractable due to the infinite, or very large, state space of the corresponding CTMC.

To circumvent solving the CME directly, several approaches have been introduced to approximately capture the statistics of the CTMCs. The Chemical Langevin Equation, a set of Stochastic Differential Equations (SDEs), approximates the dynamics of Markov jump processes by assuming a continuous memoryless noise process. The CTMC may be approximated as a diffusion process by the Fokker-Planck equation. Other approximations for the solution of the CME include using uniformization methods [Mateescu et al., 2010, Sidje et al., 2007], quantised tensor trains [Kazeev et al., 2014], and the so-called product-approximation of Jahnke [2011]. Additionally, Munsky and Khammash [2006] introduced the Finite State Projection (FSP), an error-aware method that truncates the state space of the CME into a finite subspace including the non-negligible part of the probability mass.

Alternatively, realisations of the CTMCs can be simulated by means of the Gillespie's algorithm, aka Stochastic Simulation Algorithm (SSA) [Gillespie, 1977], according to the exact statistics given by the CME. As this algorithm becomes inefficient when the numbers of molecules of species are large and/or reaction kinetics are fast, many exact and approximate variants of the SSA have been proposed over the years. To speed up the stochastic simulations, the variants of SSA exploit algebraic tricks in implementing exact stochastic simulations [Cao et al., 2004, Ramaswamy et al., 2009, Auger et al., 2006],

Poisson statistics [Gillespie, 2001, Rathinam et al., 2003], time-scale separation [Cao et al., 2005, Haseltine and Rawlings, 2002], etc. The realisations simulated by the SSA or its variants can be used to analyse the dynamics of CTMCs by reconstructing the probability distribution over its state space, or calculating the statistical moments of it. All of the above-mentioned approaches yield a microscopic description of the stochastic process by means of the full probability distribution over its state space. This detailed information is gained at a high computational cost, e.g., the cost of solving a large ODE system in FSP (see Section 2.2.1), or generating a statistically representative ensemble of realisations in SSA (see Section 2.1.4).

Traditionally, biochemical reaction networks have also been studied in terms of their *average* dynamics using macroscopic descriptions such as the Reaction Rate Equations (RRE) (see Section 2.2.4). The RRE provides the temporal evolution of the expected value of the state of the system. In contrast to microscopic approaches, the RRE possesses a low computational cost that scales linearly with the number of species (see Section 2.2.4), making it a popular tool in mathematical modelling of biochemical reaction networks [Klipp et al., 2005, Fall et al., 2010]. Although a valid approximation in the limit of large molecule numbers, the RRE fails to capture the true dynamics of the system if the size of the intrinsic noise is large compared to the average state of the system [Grima, 2010]. Therefore, a macroscopic description lacks any description of the intrinsic noise and, also possibly, accurate representation of the average dynamics.

To work around the high computational cost of microscopic descriptions on the one hand, and overcome the shortcomings of macroscopic descriptions on the other hand, one usually resorts to mesoscopic descriptions for modelling of stochasticity in biological systems of realistic sizes.

Mesoscopic descriptions of biochemical reaction networks

Mesoscopic descriptions provide information about the propagation of intrinsic stochasticity in the system by modelling the temporal evolution of the statistical moments of the probability distribution over the state space of the CTMCs. This level of resolution is particularly advantageous if the biological system is far from the large molecule number regime and/or the magnitude of intrinsic noise is large due to nonlinear dynamics. In those cases, even if one is mainly interested in the average behaviour of the system—a quantity which is conveniently approximated by the macroscopic descriptions—the inclusion of covariances and higher-order moments enhances the approximation accuracy [Engblom, 2006].

Discarding the full probability distribution and merely modelling a few moments of it, mesoscopic descriptions possess remarkably reduced computational complexity compared to microscopic counterparts. In this way, mesoscopic approaches form an interesting modelling class which enables capturing of heterogeneity in the dynamics of stochastic systems with affordable computational cost. Two well-established mesoscopic descriptions, namely the system size expansion (SSE) [van Kampen, 2007, Grima, 2010, Thomas et al., 2013] and the moment closure approximation (MA) [Engblom, 2006, Lee et al., 2009] are considered in this thesis, with the main focus being on the MA and its extensions. There exist several other mesoscopic approaches, such as hybrid descriptions [Hellander and Lötstedt, 2007, Menz et al., 2012, Jahnke, 2011], which are not in the scope of this thesis.

Moment Closure Approximation

Researchers have long tried to go beyond the macroscopic descriptions by incorporating information about the heterogeneity in terms of statistical moments. In population modelling, Whittle [1957], James H. Matis [1996], Nåsell [2003] and Keeling [2000] approximated the moments of the equilibrium distribution by assuming normal or lognormal distributions. Hespanha and Singh [2005] introduced a stochastic method to truncate the infinitely large system of ODEs governing the dynamics of the moments of the state of chemical reaction networks; they, however, did not provide a general formulation of this approach. Engblom [2006] derived the first general formulation of the so-called *moment equations* for the temporal evolution of the moments of arbitrary orders. Later, Lee et al. [2009] derived another general formulation for the moment equations using a simpler notation, where they also discussed the numerical simulation of the moment equations.

The moment equations for CRNs with linear propensities, i.e. those with at most monomolecular reactions and mass-action kinetics (see 2.1.1), are exact. In CRNs with bimolecular reactions or non-mass action kinetics, the moment equations are generally not closed, i.e. the evolution equations for a moment depend on moments of higher orders. To be able to solve the moment equations, one therefore needs to truncate the moment equations at a desired order by means of a moment closure technique. Moment closure techniques employ distribution assumptions or algebraic constraints [Engblom, 2006, Hespanha, 2007, Lee et al., 2009, Ruess et al., 2011, Singh and Hespanha, 2011, Grima, 2012, Ale et al., 2013, Schnoerr et al., 2014] to approximate the higher-order moments in terms of lower-order moments. The application of moment closure introduces an approximation error into the otherwise exact moment equations.

1.2 Overview and contribution of this thesis

Due to the diversity and complexity of biological systems, scenarios arise that are challenging for standard mesoscopic descriptions. *i)* For instance, although mass-action kinetics in biochemical reaction networks is easily interpretable within the formulation of the above-mentioned mathematical models, special treatments are needed for reliable approximations of general non-mass action kinetics, e.g., Michaelis-Menten kinetics. *ii)* In addition, similar to other fields of computational research, computational complexity in mesoscopic modelling is an essential concern that prohibits the capturing of stochasticity in large-scale processes. *iii)* Furthermore, due to the unavailability of *a priori* error bounds for these mesoscopic approximations, a problem-specific optimal choice of modelling approach is generally unclear. Given the variety of modelling approaches and their degrees of freedom, for instance the choice of moment order and closure technique for moment-closure approximations, I realised a necessity for a unifying, accessible framework that would allow for efficient simulations and performance comparisons across various approaches. *iv)* Finally, as descriptive models are extensively used in the inverse problem of inferring the underlying mechanisms of biological systems, I wanted to utilise the information captured by mesoscopic descriptions for more informative inference. I addressed the above-mentioned issues in the main research projects of my PhD thesis, summarised in Figure 1.2. In the following, I provide a brief description of my contributions, while the corresponding articles are provided in Appendices A–D. The following works resulted in reliable and feasible mesoscopic approaches that can be efficiently applied to analyse the dynamics of generic stochastic biochemical processes.

- **Enhancing the accuracy of mesoscopic modelling in dealing with non-mass action kinetics.** The accuracy of the moment approximation method is directly influenced by the choice of the moment closure technique. Although common well-known closure techniques succeed in providing satisfactory results in many cases, several scenarios, some of which are quite ubiquitous in the modelling of biological systems, render the closure of the MA equations for satisfactory approximation accuracy challenging. A common scenario which challenges the choice of a proper moment closure technique involves reactions with non-mass action kinetics. If the reaction propensities are non-polynomial functions, for instance according to Michaelis-Menten kinetics, infinitely many higher-order moments will appear in the (conditional) moment equations. Thus, we need to find a way to approximate the propensity functions and close the (conditional) moment equations. Milner et al. [2011] proposed an approach for the moment equations in the special case of ratio-

nal propensities. In this thesis, I proposed the use of Taylor series expansion for approximating propensity functions of general form in the framework of conditional moment equations [Kazeroonian et al., 2014]. In this work, I investigated the choice of closure techniques and truncation orders that are consistent with the approximation made by Taylor series expansion. This study showed that the expansion order can be used to tune the approximation error and yield desirable accuracy.

- **Model reduction based on topological network structure to enable mesoscopic description of large-scale biochemical reaction networks.** Mesoscopic descriptions possess significantly fewer state variables than the CME or microscopic approximations of it such as the FSP. However, even in the lowest-order mesoscopic approximations, the number of state variables of the governing equations scales quadratically with the number of species in the CRN. Hence, available methods are infeasible for large-sized biochemical reaction networks. To enable mesoscopic description of large-scale CRNs, I investigated the possibility of exploiting the topological structure of the CRN for model reduction. Studying a variety of recurrent motifs in biological networks, as well as several published signalling and metabolic pathways, I proposed a scalable moment closure approximation method whose complexity predominantly depends on local properties, such as the average node degree, instead of the total number of species [Kazeroonian et al., 2017]. This model reduction resulted in significantly improved scalability of mesoscopic descriptions. Higher-degree extensions of the proposed reduction scheme enabled systematic reduction of the approximation error until a satisfactory approximation quality is achieved.

- **Comprehensive platform for efficient simulations and performance comparisons to enable optimal choice of modelling approaches.** The spectrum of modelling approaches for describing the dynamics of biochemical reaction networks spans different levels of description, namely, microscopic, mesoscopic and macroscopic descriptions. Various approaches in this spectrum provide different levels of information about the system and possess different computational complexities. Even methods with similar descriptive power and complexity can differ in their performance for a specific problem. For instance, the MA and the SSE both provide similar information about the CRN, with ODE systems of roughly the same size, however, their solutions can be of different accuracies [Kazeroonian et al., 2016, Fröhlich et al., 2016]. For the MA alone, the choice of the closure scheme and the truncation order of moment equations for the best outcome, i.e. minimum approximation error in describing the moments of interest, is non-trivial and highly dependent on the kinetics of the system.

Since a priori error bounds for most of the approximative approaches in this spectrum are not available, simulation studies have to be conducted to test the quality of different approximations and select the proper modelling approach. Despite such a necessity, a comprehensive comparative study of available descriptions has received little attention. More importantly, a platform to facilitate comparison of modelling approaches in terms of their approximation accuracy and numerical efficiency has been missing. In addition, for a subset of these methods, efficient implementations have not been publicly available. To address these needs, I led the development and implementation of an accessible and comprehensive software toolbox CERENA (http://cerenadevelopers.github.io/CERENA/). Being an open-source MATLAB toolbox, CERENA allows for the automatic generation of different model types and their numerical simulation [Kazeroonian et al., 2016]. This platform facilitates efficient integration of simulation results (including sensitivity analysis) for further analyses such as parameter estimation.

- **Exploiting mesoscopic descriptions for more informative inference of stochastic biochemical kinetics.** Mechanistic models, generally, depend on some parameters, e.g., the kinetic constants of chemical reactions and initial abundance of species. The predicted behaviour of the biological system given by a mechanistic model, therefore, is dependent on the values of corresponding parameters and can show qualitative differences in different regions of parameter space. Thus, obtaining reliable predictive models is only possible through obtaining reliable parameter values. However, the true values of these parameters are mostly unknown and cannot be measured directly. Therefore, an increasing interest in the inverse problem of estimating the parameter values from experimental data has arisen in the systems biology community. In parameter estimation, experimental data about the quantities of interest are used to calibrate the mathematical model describing the dynamics of the biological system. Due to the finite amount of experimental data—and other factors such as measurement noise—there is uncertainty involved with the estimated parameter values. Decreasing this uncertainty is crucial for increasing the predictive power of mechanistic models. This can be achieved by incorporating more experimental data and/or exploiting more information from the available data. For instance, population-averaged data might yield non-informative or misleading inference, as they lack information about heterogeneity that could otherwise provide valuable insights into the underlying mechanisms (see Figure 1.1).

  To exploit information beyond the population average, mesoscopic descriptions can be employed to describe statistical moments of data, e.g., from population snapshot data, [Zechner et al., 2012, Milner et al., 2013, Ruess and Lygeros, 2015]. I investi-

gated the impact of the added information by more summary statistics on inference results, by employing a parameter estimation framework based on moment-closure approximations of various orders [Kazeroonian et al., 2013]. This study indicated that the incorporation of higher-order statistical moments (significantly) decreases the uncertainty of estimated parameter values and increases the predictive power of the model.

Detailed descriptions of the above-mentioned manuscripts and my contributions are provided in Chapter 3. I am the first author of these contributions and was in charge of the preparation.

1.2.1 Other contributions

In my doctoral work, I also contributed to two other projects and the corresponding publications: i) the development of the method of conditional moments, a hybrid microscopic-mesoscopic approach for the handling of copy-number scale separation in biochemical processes [Hasenauer et al., 2014b]; and ii) the comparison of system size expansion and moment closure approximations for the inference of stochastic chemical kinetics [Fröhlich et al., 2016]. A brief summary of these articles is given below.

1. J. Hasenauer, V. Wolf, A. Kazeroonian, and F. J. Theis. **Method of conditional moments (MCM) for the Chemical Master Equation.** *Journal of Mathematical Biology. 69(3): 687-735, 2014.*

   In this work, we introduced a novel approach, namely the method of conditional moments (MCM), for describing the statistics of the solution of the chemical master equation. To ensure reliable approximations in the presence of copy-number scale separation, the MCM employs a hybrid stochastic-deterministic description. The state of low copy-number species is modelled in terms of their associated marginal probabilities, while the state of medium/high copy-number species is captured in terms of the statistical moments of their corresponding distributions. Furthermore, to account for potentially distinct dynamics in various states of low copy-number species, the state distributions of medium/high copy-number species are conditioned on the state of low copy-number species. This allows for the capturing of complex correlation structures, e.g., as in multi-attractor and oscillatory systems. The MCM was shown to improve upon other hybrid approaches, as well as the standard moment closure approximation, where microscopic effects, such as fluctuations in gene expression, rise to macroscopic differences in the behaviour of the biological system.

My contribution: I contributed to the mathematical derivation of proper initial conditions for the MCM. Furthermore, I performed the simulation study illustrating the properties of the MCM and assessed the numerical error introduced by approximating the DAE that describes the conditional moments by an ODE. Finally, I contributed to the writing of the manuscript.

2. F. Fröhlich, P. Thomas, A. Kazeroonian, F. J. Theis, R. Grima, J. Hasenauer. **Inference for Stochastic Chemical Kinetics Using Moment Equations and System Size Expansion.** *PLOS Computational Biology 12(7): e1005030, 2016.*

   In this article, we introduced efficient gradient-based methods for parameter estimation and uncertainty analysis using moment closure approximations (MA) and system size expansion (SSE). Using these methods, we compared the parameter estimation accuracy and identifiability achieved using different mesoscopic modelling approaches. Studying the Epo-induced JAK/STAT signaling, we showed that MA and SSE yield an improved parameter identifiability compared to reaction rate equations, even if merely population-average data are used. Furthermore, we studied various volume regimes to identify those in which the estimation results are more reliable.

   My contribution: I contributed to the mathematical derivation of the moment closure approximations used in this study. For this, I employed the modelling and analysis toolbox CERENA developed in this thesis (see Chapter 3 and [Kazeroonian et al., 2016] for more details regarding CERENA).

1.3 Outline

This is a cumulative dissertation based on the research work that is published in my first-author articles. I lay the background for these articles in Chapter 2 by introducing the fundamental notions of chemical reaction networks and stochastic chemical kinetics described by continuous-time Markov chains and the chemical master equation. In Section 2.2, I briefly describe several microscopic, mesoscopic and macroscopic descriptions that are used in this thesis for the analysis of biochemical reaction networks. In Section 2.3, the methods used for parameter estimation and uncertainty analysis are outlined. In Chapter 3, I give a summary of all the contributed articles. Chapter 4 concludes this dissertation and discusses possible extensions and continuations of this work. Finally, the full text of my first-author articles are attached in the Appendices A–D.

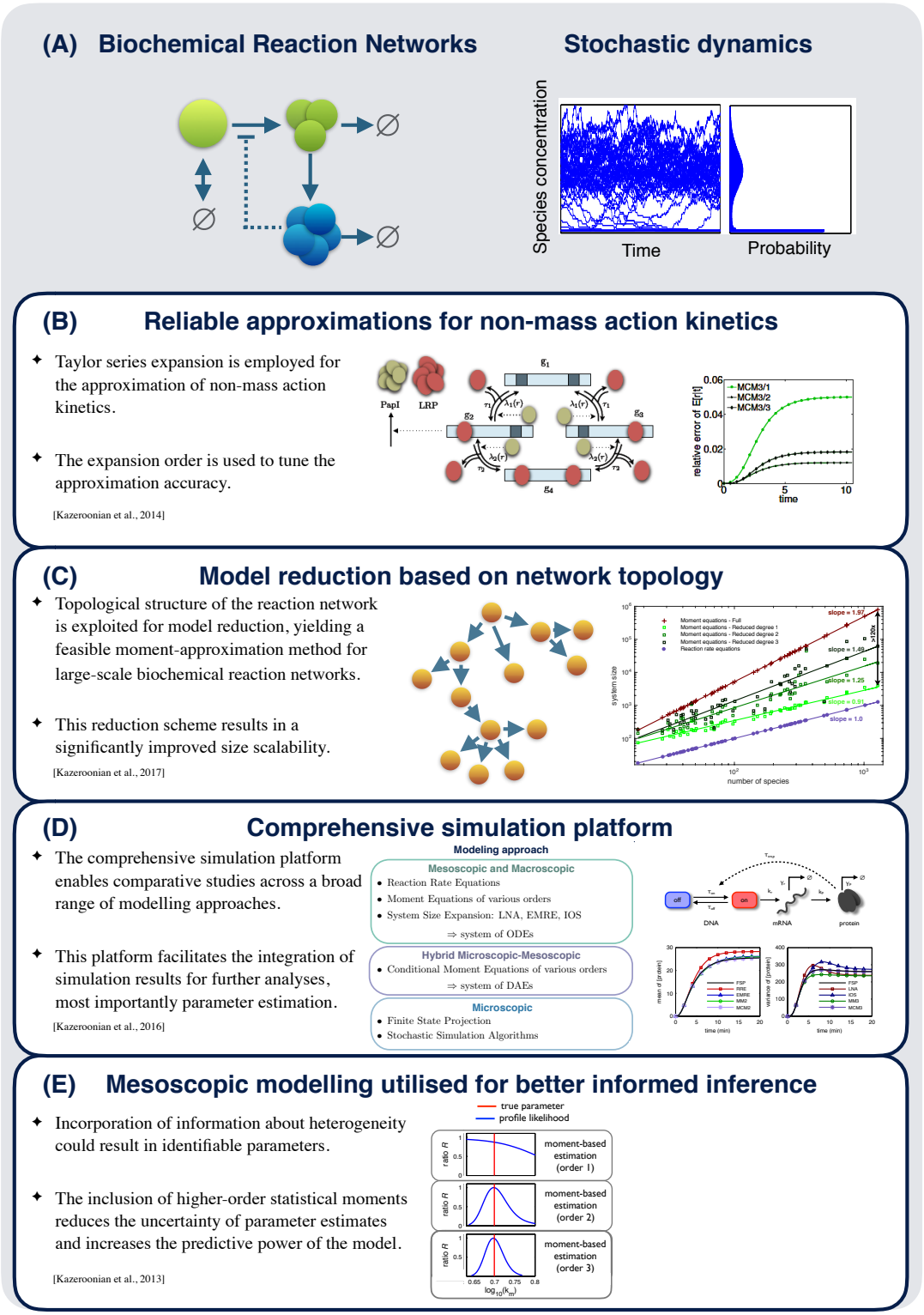


Figure 1.2: **Overview of the thesis.** (A) This thesis focuses on simulating the stochastic dynamics of cellular processes that are modelled as biochemical reaction networks. (B, C, D, E) The key contributions of the thesis that resulted in the corresponding first-author articles. The summaries of these contributions, as well as the full-text articles can be found in Chapter 3 and Appendices A–D respectively.

Chapter 2

Methods

This thesis work mainly focuses on biochemical reaction networks that exist at the heart of many biological processes. The dynamics of biochemical reaction networks is intrinsically stochastic due to the discrete nature of matter and chemical reactions. In this chapter, I give an overview of the methodologies for the analysis of stochastic biochemical kinetics, on which this thesis work has been founded.

2.1 Stochastic chemical kinetics

Stochastic chemical kinetics are mainly modelled using continuous time Markov chains (CTMCs). The probabilistic evolution of CTMCs can be exactly described by the Chemical Master Equation (CME) [Gillespie, 1992a]. Alternatively, individual trajectories of the CTMCs can be simulated using Stochastic Simulation Algorithm (SSA) [Gillespie, 1977] and its variants. In the following, I lay out the background information on chemical reaction networks, continuous-time Markov chains, the Chemical Master Equation and the Stochastic Simulation Algorithm. This information is the foundation on which the later sections will be built.

2.1.1 Chemical Reaction Networks

A (bio)chemical reaction network (CRN) is a system comprising of several distinct entities, or *chemical species*, that can be present in various abundances. A chemical species is an ensemble of chemically identical molecules, e.g., a specific protein. The chemical species

undergo *chemical reactions* [McNaught and Wilkinson, 1997]. A chemical reaction is a transition via which, for instance, the chemical species are synthesised, degraded, or converted into other species. The chemical reactions, therefore, change the abundance of chemical species, and consequently, the configuration of the system in terms of molecular copy numbers changes over time. Given a certain initial configuration, the state of the system, that is the number of molecules of chemical species, evolves randomly in time with statistics given by the kinetics of the chemical reactions (see Section 2.1.4).

Consider a chemical reaction network of n_s chemical species, $S_1, S_2, \cdots, S_{n_s}$, and n_r chemical reactions, $R_1, R_2, \cdots, R_{n_r}$. The reaction R_j changes the configuration of the system as below

$$R_j : \sum_{i=1}^{n_s} \nu_{i,j}^{-} S_i \xrightarrow{k_j} \sum_{i=1}^{n_s} \nu_{i,j}^{+} S_i.$$

The parameter k_j is the *kinetic constant* of reaction R_j. The kinetic constants are influenced by chemical characteristics of the reactions, as well as the environment in which the reactions take place, such as the temperature. The coefficient $\nu_{i,j}^{-} \in \mathbb{N}_0$, called the *stoichiometric coefficient of reactants*, denotes the number of S_i molecules that are consumed in reaction R_j. Similarly, the *stoichiometric coefficient of products*, $\nu_{i,j}^{+} \in \mathbb{N}_0$, denotes the number of S_i molecules that are produced in reaction R_j [Klipp et al., 2005]. The net change in the abundance of species S_i upon firing of reaction R_j is, therefore, given by the overall stoichiometric coefficient $\nu_{i,j} = \nu_{i,j}^{+} - \nu_{i,j}^{-} \in \mathbb{Z}$. In this way, the overall stoichiometry of the reaction R_j can be summarised in the vector $\nu_j = (\nu_{1,j}, \nu_{2,j}, \cdots, \nu_{n_s,j})^T \in \mathbb{Z}^{n_s}$.

The state of the system at time t is represented by a vector $\mathbf{X_t} = (X_{1,t}, X_{2,t}, \ldots, X_{n_s,t})^T \in \mathbb{N}_0^{n_s}$, in which $X_{i,t}$ is the number of molecules of species S_i at time t. Upon firing of the reaction R_j, the state of the system changes according to the stoichiometry of R_j:

$$\mathbf{X}_t \to \mathbf{X}_t + \nu_j. \tag{2.1}$$

The probability of the occurrence of reaction R_j per unit time is called the *propensity* of reaction R_j , denoted by $a_j(\mathbf{X}_t) : \mathbb{N}_0^{n_s} \to \mathbb{R}_+$, and is a function of the current state of the reaction network as well as the kinetics of the reaction. The statistics of the time spent between firing of consecutive reactions and the index of the next reaction to fire are given by the reaction propensities [Feller, 1940] (see Section 2.1.4 for more details). The propensities can have various functional forms depending on the specific kinetics of the reactions. In the simplest case, the reaction kinetics follow the *law of mass action* [Gillespie, 1977], which basically states the probability of a reaction to happen is proportional to the probability of the required molecules of reactants to meet and collide. For instance,

Table 2.1: Reaction propensities according to the law of mass action.

Reaction Order	Reaction Type	Propensity
0	$\emptyset \xrightarrow{k_j}$ product	k_j
1	$S_i \xrightarrow{k_j}$ product	$k_j X_i$
2	$S_i + S_l \xrightarrow{k_j}$ product	$k_j X_i X_l$
2	$S_i + S_i \xrightarrow{k_j}$ product	$\frac{1}{2} k_j X_i (X_i - 1)$

in the case of a bimolecular reaction (see Table 2.1), the probability of a distinct pair of S_i and S_l molecules to meet is proportional to the number of distinct pairs of S_i and S_l molecules, i.e. $\binom{X_i}{1}\binom{X_l}{1} = X_i X_l$, under well-mixed and thermal equilibrium assumptions. The proportionality factor is given by the kinetic constant of the reaction. The propensities for reactions of order 0, 1 and 2, according to the law of mass action, are given in Table 2.1. It is worth noting that in the case of a dimerisation reaction, the probability for two *distinct* molecules of the species S_i to meet is given by $\binom{X_i}{2} = \frac{1}{2} X_i (X_i - 1)$. The reaction kinetics may not follow the law of mass action, and have more complicated forms instead. Usually, this results in situations where not all detailed molecular interactions are modelled as individual reactions, but instead lumped together in one reaction. The most common example is the Michaelis-Menten kinetics [Michaelis and Menten, 1913] describing the rate of enzymatic reactions of some substrate S_i as $\frac{V_{\max} X_i}{K_M + X_i}$ with $V_{\max}$ and K_M being constants and X_i being the number of molecules of S_i. It is worth noting that reactions with three (or more) reactants are improbable as they require three (or more) randomly chosen molecules to come into simultaneous contact. Therefore, trimolecular reactions do not represent regular elementary events, and instead are sometimes used as approximations to a sequence of multiple reactions [Gillespie, 1992b]. Thus in this thesis, we only consider reactions with at most 2 reactants.

As mentioned earlier, the temporal evolution of the state of the system is determined by the sequence of chemical reactions to fire and the times of those events. Consequently, $\mathbf{X}_t$ is a random process due to the probabilistic nature of the chemical reactions. At each time point t, the time until the next reaction and the index of the next reaction are randomly distributed with probability density functions that only depend on the reaction propensities, which are in turn functions of the current state of the system $\mathbf{X}_t$. Therefore, statistically speaking, the knowledge about the current state is sufficient to determine the probability distribution of the state of the system at a later time $t+\Delta t$, meaning that $\mathbf{X}_t$ is a Markov process (see Section 2.1.2 for more information). Since the number of molecules of chemical species is a non-negative integer-valued quantity, the state of the system can

only take on specific discrete values. Thus, $\mathbf{X}_t$ can only *jump* between states in the state space and is, therefore, a *Markov jump process*. Additionally, since the state transitions can happen at any (real-valued) time, $\mathbf{X}_t$ belongs to a particular class of Markov chains, namely *Continuous-Time Markov Chains (CTMCs)* [Norris, 1998]. The random jumps of $\mathbf{X}_t$ occur according to transition probabilities between possible states. The formal definitions of the random process $\mathbf{X}_t$ and the corresponding transition probabilities—that are determined by reaction propensities—are given in the next section.

2.1.2 Markov processes

The temporal evolution of the exact state of many processes cannot be determined, e.g., due to the intrinsic randomness of the system, or insufficient knowledge about the system. Such processes evolve randomly according to some probabilistic rules, and merely the statistic of the process can be determined instead of the exact trajectory of the system.

Markov processes are a simple type of random processes that possess the so-called *Markov property* [Norris, 1998, van Kampen, 2007]. The Markov property states that the future state of the system is only dependent on the current state and not on the preceding states. This property indicates that the Markov processes are memoryless, as the history of the process does not influence the future evolution of it.

More precisely, if we consider a succession of time points, $t_1 < t_2 < \cdots < t_N < t_{N+1}$, and denote the state of a Markov process at those time points by $\{\mathbf{X}_{t_1}, \mathbf{X}_{t_2}, \cdots, \mathbf{X}_{t_N}, \mathbf{X}_{t_{N+1}}\}$, then the Markov property states:

$$P(\mathbf{X}_{t_{N+1}}|\mathbf{X}_{t_1}, \cdots, \mathbf{X}_{t_N}) = P(\mathbf{X}_{t_{N+1}}|\mathbf{X}_{t_N}). \tag{2.2}$$

Hence, the state of the process at the next time point, t_{N+1}, only depends on the state at t_N and is independent from all preceding states. Using the identity (2.2) on the conditional probability, one can obtain the following for the joint probability of the trajectory $\{\mathbf{X}_{t_1}, \mathbf{X}_{t_2}, \cdots, \mathbf{X}_{t_N}\}$:

$$P(\mathbf{X}_{t_1}, \cdots, \mathbf{X}_{t_N}) = P(\mathbf{X}_{t_1})P(\mathbf{X}_{t_2}|\mathbf{X}_{t_1}) \cdots P(\mathbf{X}_{t_N}|\mathbf{X}_{t_{N-1}}) = P(\mathbf{X}_{t_1}) \prod_{i=1}^{N-1} P(\mathbf{X}_{t_{i+1}}|\mathbf{X}_{t_i}). \tag{2.3}$$

Markov jump processes

A special type of Markov processes are those whose state is a subset of a d-dimensional integer lattice $\Psi \in \mathbb{Z}^d$, in other words, whose state can take on only discrete values. These are called *Markov Jump Processes (MJP)* since the system can only *jump* from one state to another. The state space Ψ may be infinite, or finite due to some constraints on the dynamics of the system.

For simplicity, we consider a finite MJP with a state space of n different states $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$. At each time point t_i, we assign a probability vector over all possible states, $\mathbf{P}(t_i) = (P(\mathbf{X}_{t_i} = \mathbf{x}_1), P(\mathbf{X}_{t_i} = \mathbf{x}_2), \cdots, P(\mathbf{X}_{t_i} = \mathbf{x}_n))^T$. Using the Markov property, the probability vector at the next time point t_{i+1} is obtained via

$$\mathbf{P}(t_{i+1}) = \mathbf{W}(t_{i+1})\mathbf{P}(t_i), \tag{2.4}$$

where $\mathbf{W}(t)$ is a matrix whose elements represent the probabilities of transitioning from one point into another in the state space:

$$[\mathbf{W}(t_{i+1})]_{kl} = P(\mathbf{X}_{t_{i+1}} = \mathbf{x}_k | \mathbf{X}_{t_i} = \mathbf{x}_l). \tag{2.5}$$

Hence, $\mathbf{W}(t)$ is called *transition matrix* of the system at time t [Norris, 1998, van Kampen, 2007]. From the definition (2.5), one can deduce the following properties for the transition matrix:

- All elements of $\mathbf{W}$ are non-negative.
- All columns of $\mathbf{W}$ add up to unity, since each column represents the transition probability from a given state to all possible states.

Therefore, $\mathbf{W}$ is a left stochastic matrix. If $\mathbf{W}(t)$ is constant for all times t, the Markov process is said to be *homogenous*. Accordingly, a homogeneous MJP starting from the initial probability distribution $\mathbf{P}(t_1)$ evolves as

$$\mathbf{P}(t_N) = \mathbf{W}^{N-1}\mathbf{P}(t_1). \tag{2.6}$$

Continuous-time discrete-state Markov chains

So far, we assumed a series of successive time points, $t_1 < t_2 < \cdots < t_N < t_{N+1}$, to analyse the dynamics of a Markov process. However, such a selection of discrete

time points is not natural for many processes that evolve continuously in time. These processes mark a special class, namely *Continuous-Time Markov Chains (CTMCs)*. At each time point, we can define a probability distribution $\mathbf{P}(t)$ comprising of the probability of the CTMC possessing a particular configuration $\mathbf{x}$, $P(\mathbf{X}_t = \mathbf{x})$, for all $\mathbf{x} \in \Psi$. Again, assuming a finite state space of n different states $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, the vector $\mathbf{P}(t) = (P(\mathbf{X}_t = \mathbf{x}_1), P(\mathbf{X}_t = \mathbf{x}_2), \cdots, P(\mathbf{X}_t = \mathbf{x}_n))^T$ represents all state probabilities at time t. The probability distribution $\mathbf{P}(t)$ evolves in continuous time and its rate of change at any time t is given by

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{Q}(t)\mathbf{P}(t), \quad \mathbf{P}(t_0) = \mathbf{P}_0, \tag{2.7}$$

where $\mathbf{Q}(t)$ is called the *transition rate matrix* of the system, denoting the change in the probability distribution per unit time [Norris, 1998, van Kampen, 2007]. The matrix $\mathbf{Q}$ has the following properties:

- All diagonal elements of $\mathbf{Q}$ are finite negative values, i.e., $0 \leq -[Q]_{kk} < \infty$ for all k.
- All off-diagonal elements of $\mathbf{Q}$ are non-negative, i.e., $[Q]_{kl} \geq 0$ for all $k \neq l$.
- All columns of $\mathbf{Q}$ add up to zero, i.e., $\sum_k [Q]_{kl} = 0$ for all l.

The initial condition $\mathbf{P}_0$ denotes the initial probability distribution of the process. The differential equation (2.7) can be written and solved if the state space Ψ is finite. In the case of infinite state spaces, a truncation is required to obtain a finite set on which the problem (2.7) is tractable. However, special cases exist in which equation (2.7) can be analytically solved on infinite state spaces [Jahnke and Huisinga, 2007]. This class of Markov processes are the main focus of this thesis, as they are extensively used to model stochastic chemical kinetics. In the next section, I briefly outline how the CTMCs are used to derive the governing equations for the dynamics of biochemical reaction networks.

2.1.3 Chemical Master Equation

As pointed out earlier, the temporal evolution of the state of a biochemical reaction network fulfils the Markov property, as the statistics of the future configuration only depends on the current configuration of the reaction network. The state of the CRN, $\mathbf{X_t} = (\mathbf{X}_{1,t}, \mathbf{X}_{2,t}, \ldots, \mathbf{X}_{n_s,t})^T \in \mathbb{N}_0^{n_s}$, represents the counts of species and therefore, can only take non-negative integer values, making $\mathbf{X}_t$ an MJP. Furthermore, since a chemical reaction network evolves in continuous time, $\mathbf{X}_t$ can be regarded as a CTMC.

We consider the probability distribution $\mathbf{P}(t)$ consisting of elements $P(\mathbf{X}_t = \mathbf{x})$ over all states $\mathbf{x} \in \Psi$. To use Eq. (2.7) and obtain the governing equation for $\mathbf{P}(t)$, one needs to know the transition probabilities for all *adjacent states* in the state space. By adjacent states, I refer to those states that can be immediately reached from one another. Since the state $\mathbf{X}_t$ is only changed via the firing of chemical reactions, the adjacent states to the current configuration are those that are reached via one firing of any of the feasible reactions. If the reaction network has the configuration $\mathbf{x}$ at time t, the transition probability from $\mathbf{x}$ at time t, into $\mathbf{x} + \nu_j$ at time $t + \Delta t$, for an infinitesimal time interval Δt, is proportional to the probability of firing of reaction R_j in Δt: $a_j(\mathbf{x})\Delta t$ for $\Delta t \ll 1$.

The change in the probability associated with a state $\mathbf{x}$, $P(\mathbf{X}_t = \mathbf{x})$, results from the influx from/outflux into adjacent states. More precisely, assume that the probability of the stochastic process to realise state $\mathbf{x}$ at time t is $P(\mathbf{X}_t = \mathbf{x})$. The change in $P(\mathbf{X}_t = \mathbf{x})$ in an infinitesimal time interval $\Delta t \ll 1$ is given by

$$\Delta P(\mathbf{x}, t) = P(\mathbf{x}, t + \Delta t) - P(\mathbf{x}, t) = \sum_{j=1}^{n_r} P_{j_{\text{in}}} - P_{j_{\text{out}}}, \tag{2.8}$$

where $P_{j_{\text{in}}}$ is the probability of arriving at state $\mathbf{x}$ via one firing of reaction R_j in Δt, and $P_{j_{\text{out}}}$ is the probability of leaving the state $\mathbf{x}$ via one firing of reaction R_j in Δt. The influx of probability from a state $\mathbf{x} - \nu_j$ in an infinitesimal time interval Δt, $P_{j_{\text{in}}}$, is given by 1) the probability of the system realising configuration $\mathbf{x} - \nu_j$ at time t, and 2) the probability of the firing of reaction R_j in Δt given that the system is at state $\mathbf{x} - \nu_j$. Similarly, the outflux into a state $\mathbf{x} + \nu_j$ in an infinitesimal time interval Δt, $P_{j_{\text{out}}}$, is given by 1) the probability of the system realising configuration $\mathbf{x}$ at time t, and 2) the probability of the firing of reaction R_j in Δt given that the system is at state $\mathbf{x}$. Consequently, the net change in $P(\mathbf{X}_t = \mathbf{x})$ is obtained by the superposition of influx and outflux of all reactions:

$$\Delta P(\mathbf{x}, t) = \sum_{j=1}^{n_r} \Big(P(\mathbf{x} - \nu_j, t) a_j(\mathbf{x} - \nu_j)\Delta t - P(\mathbf{x}, t) a_j(\mathbf{x})\Delta t \Big). \tag{2.9}$$

Here, it is assumed that the reaction propensities are *proper*: if $\mathbf{x} \ngeq \nu_j^-$ (i.e., $\exists i : x_i < \nu_{i,j}^-$) then $a_j(\mathbf{x}) = 0$, imposing that the reactions can only take place if sufficiently many reactants are available. Taking the limit $\Delta t \to 0$ yields the following differential equation, governing the temporal evolution of $P(\mathbf{x}, t)$:

$$\frac{\partial}{\partial t} P(\mathbf{x}, t) = \sum_{j=1}^{n_r} \Big(a_j(\mathbf{x} - \nu_j) P(\mathbf{x} - \nu_j, t) - a_j(\mathbf{x}) P(\mathbf{x}, t) \Big). \tag{2.10}$$

Eq. (2.10) is called the *Chemical Master Equation (CME)* [Gillespie, 1992a], and describes the exact statistics of the CTMCs corresponding to the temporal evolution of biochemical reaction networks. The CME is also known as the forward Kolmogorov equation (for more details see van Kampen [2007]).

The state space of the CME is mostly infinitely large and therefore, finding a direct solution to the CME is mostly intractable. For this reason, several approximative approaches have been proposed over the last decades to approximate the solution of the CME. A few of these methods are described in the next sections.

2.1.4 Stochastic Simulation Algorithm

An alternative approach to analyse the dynamics of stochastic processes is to simulate realisations of the process. The well-known Stochastic Simulation Algorithm (SSA) [Gillespie, 1977] is such an approach whose main idea is to generate statistically representative trajectories of the CTMCs. To ensure this, various trajectories of a CTMC are generated such that the frequency of generating a trajectory is proportional to the probability of the CTMC realising that trajectory according to the chemical master equation.

The idea of the SSA is simple and intuitive. Starting from an ensemble of initial configurations, randomly drawn from the initial probability distribution of the CTMC, the SSA tracks changes to the configuration of the system which result from the occurrence of chemical reactions. The necessary information for simulating exact trajectories of the underlying CTMC are encoded in the reaction propensities, as the firing of the chemical reactions is the only mechanism through which the state of the system changes.

To move forward starting from a given configuration, one needs to know i) the time until the next reaction happens, and ii) the index of the reaction which happens next. For all possible reactions, one can write a joint probability density function for the time of the next reaction, τ, and the next reaction being a particular one, R_j: $\mathcal{P}(\tau, j)$. In this way, $\mathcal{P}(\tau, j)\Delta\tau$ is the probability that the next reaction to fire is reaction R_j and that this reaction fires in the infinitesimal time interval $(\tau, \tau + \Delta\tau)$. This joint probability can be written as the product of the two independent probabilities [Gillespie, 1977]

$$\mathcal{P}(\tau, j)\Delta\tau = \mathcal{P}_1(\tau)\mathcal{P}_2(j)\Delta\tau, \tag{2.11}$$

where

- $\mathcal{P}_1(\tau)$ is the probability density function that no reaction happens in the next time interval of length τ.

- $\mathcal{P}_2(j)$ is the probability density function of the reaction R_j to fire in the time interval $(\tau, \tau + \Delta\tau)$. By definition, this is equal to the propensity of reaction R_j given the state of the system at time τ.

Given that the system is in state $\mathbf{x}$ at τ', i.e. $\mathbf{X}'_\tau = \mathbf{x}$, for any infinitesimal time interval $\Delta\tau'$, the probability that *any* reaction fires in $(\tau', \tau' + \Delta\tau')$ is given by the sum of the propensities of all possible reactions times the time interval: $\sum_j a_j(\mathbf{x})\Delta\tau'$. Accordingly, one can write:

$$\begin{aligned} \mathcal{P}_1(\tau' + \Delta\tau') &= \mathcal{P}_1(\tau') \left(1 - \sum_j a_j(\mathbf{x})\Delta\tau'\right), \\ \xrightarrow{\Delta\tau' \to 0} \quad \frac{d\mathcal{P}_1(\tau')}{d\tau'} &= -\mathcal{P}_1(\tau') \sum_j a_j(\mathbf{x}) \end{aligned} \tag{2.12}$$

which yields:

$$\mathcal{P}_1(\tau) = \exp\left(-\sum_j a_j(\mathbf{x})\tau\right). \tag{2.13}$$

Using the independence assumption (2.11), and substituting $\mathcal{P}_1(\tau)$ by Eq. (2.13) and $\mathcal{P}_2(j)$ by the propensity $a_j(\mathbf{x})$, we can calculate the probability density function of the time until the next firing of reaction R_j as

$$\mathcal{P}(\tau, j) = \exp\left(-\sum_j a_j(\mathbf{x})\tau\right) a_j(\mathbf{x}). \tag{2.14}$$

The first-reaction method for implementing the SSA proceeds by sampling n_r random variables for the time until the next firing of all possible reactions from the corresponding probability distributions according to (2.14). Taking the smallest of the resulting times, one finds the time and the index of the next reaction to fire. Subsequently, the state of the process $\mathbf{X}$ is updated according to the stoichiometry of that reaction, i.e. $\mathbf{X}_{\tau+\boldsymbol{\Delta}\tau'} := \mathbf{x} + \nu_j$, where j is the index of the next reaction to fire, and τ is the time of the firing.

To speed up this algorithm, one can sample the reaction time and the reaction index independently, to avoid the unnecessary sampling of n_r random variables. Taking the sum of $\mathcal{P}(\tau, j)$ in (2.14) over all reactions j, we obtain the following probability density function

for the time until *any* reaction fires:

$$\mathcal{P}(\tau) = \exp\left(-a_0(\mathbf{x})\tau\right) a_0(\mathbf{x}). \tag{2.15}$$

where $a_0(\mathbf{x}) = \sum_j a_j(\mathbf{x})$. Thus, the time until the next reaction is exponentially distributed with parameter $a_0(\mathbf{x})$. The probability of the next reaction to be R_j is simply $\frac{a_j(\mathbf{x})}{a_0(\mathbf{x})}$. This more efficient SSA, called the next-reaction method [Gibson and Bruck, 2000], consists of sampling two random variables for the time and the index of the next reaction from their corresponding distributions. The state update step is done similarly to the first-reaction method.

The SSA generates trajectories of the CTMC for the time interval of interest. An estimate to the probability $P(\mathbf{x}, t)$ is given by the frequency of the trajectories that visit state $\mathbf{x}$ at time t. The SSA results can also be used to estimate the statistics of the CTMC by the Monte-Carlo integration [Newman and Barkema, 1999]. For example, the mean and variance of the process are approximated by

$$\begin{aligned} \mathbf{m}(t) &= \frac{1}{N}\sum_{l=1}^{N} \mathbf{x}_l(t), \\ \mathbf{C}(t) &= \frac{1}{N-1}\sum_{l=1}^{N} \Big(\mathbf{x}_l(t) - \mathbf{m}(t)\Big)\Big(\mathbf{x}_l(t) - \mathbf{m}(t)\Big)^T, \end{aligned} \tag{2.16}$$

with N being the number of trajectories, and $\mathbf{x}_l(t)$ being the state of the CTMC at time t in the l^{th} trajectory. The estimators for probability distribution and the statistical moments are unbiased and converge to the true moments of the CTMC. However, to ensure estimates with low variances, generally a large number of realisations are required. As a result, although SSAs are simple and powerful methods for exact simulation of CTMCs, their high computational cost limits their applicability.

Many variants of the SSA have been proposed over the years to improve its efficiency or generality. In this thesis, in addition to the next-reaction SSA, a modified next-reaction method [Anderson, 2007] was used for the simulation of systems with time-dependent propensities where the sampling of only one random variable is required per reaction event.

2.2 Approximative methods for the stochastic chemical kinetics

The exact description of statistics of stochastic chemical kinetics, as is given by the chemical master equation, is mostly intractable due to the large or infinite state space of the CME. To circumvent this problem, several approximative methods have been introduced over the past decades for the description of the dynamics of chemical reaction systems. This section, briefly introduces the approximations that have been in the scope of this thesis.

2.2.1 Finite State Projection

In a reaction network of n_s biochemical species, the CME is defined for all states $\mathbf{x} \in \Psi \subset \mathbb{N}_0^{n_s}$ which can be reached via chemical reactions. The set Ψ is generally very large, or infinite, and as a result, a direct solution of the CME is mostly intractable. However, the subset of states with non-negligible probability mass is usually significantly smaller than the set of all reachable states Ψ. Motivated by this notion, the *Finite State Projection (FSP)*, proposed by Munsky and Khammash [2006], merely solves the CME on a truncated state space $\hat{\Psi} \subset \Psi$ which possesses a sufficiently large portion of the probability mass.

In Eq. (2.10), we presented the CME equation for every individual state $\mathbf{x} \in \Psi$. If we enumerate all states in Ψ as $\mathcal{X}_\Psi := (\mathbf{x}_1, \mathbf{x}_2, \cdots)^T$, and define the probability vector $\mathbf{P}(t) := (P(\mathbf{x}_1, t), P(\mathbf{x}_2, t), \cdots)^T$, the CME for all states $\mathbf{x} \in \Psi$ is written as

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{A}\,\mathbf{P}(t). \tag{2.17}$$

The matrix $\mathbf{A}$ is a linear operator defined on the sequence $\mathcal{X}_\Psi$ as

$$A_{kl} = \begin{cases} -\sum_{j=1}^{n_r} a_j(\mathbf{x}_k) & \text{for } k = l \\ a_j(\mathbf{x}_k) & \text{for all } l \text{ such that } \mathbf{x}_l = \mathbf{x}_k + \nu_j \\ 0 & \text{otherwise} \end{cases} \tag{2.18}$$

If we now define a finite ordered index set $J := \{j_1, j_2, \cdots, j_n\}$ corresponding to the states $\mathcal{X}_J = \{\mathbf{x}_{j_i}, \mathbf{x}_{j_2}, \cdots, \mathbf{x}_{j_n}\}$, the approximation of CME for the subset of states defined by the index set J is given by

$$\frac{d\mathbf{P}_J(t)}{dt} = \mathbf{A}_J\,\mathbf{P}_J(t), \tag{2.19}$$

where $\mathbf{P}_J(t) := (P(\mathbf{x}_1,t), P(\mathbf{x}_2,t), \cdots, P(\mathbf{x}_{j_n},t))^T$, and the matrix $\mathbf{A}_J$ is obtained by choosing the rows and columns of $\mathbf{A}$ according to J. In (2.19), the incoming probability from the states $\mathbf{x} \in \mathcal{X}_\Psi \setminus \mathcal{X}_J$ is disregarded. Since Eq. (2.19) is finite, the solution $\mathbf{P}_J(t)$ is calculated via

$$\mathbf{P}_J(t) = \exp(\mathbf{A}_J t)\mathbf{P}_J(0). \tag{2.20}$$

Equation (2.19) approximates the solution of the CME on a truncated state space $\hat{\Psi}$ defined by the index set J. Eq. 2.19 can be interpreted as solving the CME for a system consisting of states $\mathcal{X}_J$, while all other states are lumped together into a sink state that only absorbs probability from the system, without returning any probability back to the system. Munsky and Khammash [2006] show that the solution $\mathbf{P}_J$ provides a lower bound on the solution of the CME. The sum of the probabilities remained in the system, $\sum_{\mathbf{x}\in\hat{\Psi}} P_J(\mathbf{x},t)$, is an indicator of the approximation error by truncating the state space. If no truncation is made, this sum is equal to 1. Therefore, the difference $1 - \sum_{\mathbf{x}\in\hat{\Psi}} P_J(\mathbf{x},t)$ indicates the amount of probability leaked out of the system. This approximation error can be monotonically decreased by expanding the state space of the FSP.

The applicability of FSP depends on the number of states with a significant probability mass. There exist novel algorithms than can solve the FSP with some million states [Mateescu et al., 2010] and higher [Kazeev et al., 2014]. The latter uses the quantised tensor train representation of the CME and achieves a computational complexity that scales linearly with the number of biochemical species.

2.2.2 Moment closure approximation method

Microscopic descriptions, such as the FSP (Section 2.2.1), quickly become infeasible as the number of species in the chemical reaction network grows. To circumvent such demanding computational cost, several approaches have been developed that focus on representing the solution of the CME in terms of its statistical moments, and thereby providing a mesoscopic description of the dynamics of the CRNs. The *moment closure approximation method (MA)* [Engblom, 2006, Lee et al., 2009] is such a mesoscopic approximation that yields a system of ordinary differential equations for the temporal evolution of the moments of the state of a CRN.

Here, I lay out the derivation of the moment equations, using a notation similar to [Lee et al., 2009]. The central moments of the probability distribution over the state space of

the CTMC, $\mathbf{P}(\mathbf{x},t)$, are defined as below:

$$\begin{aligned} \text{mean} \quad & \mathbf{m} = (m_1,\ldots,m_{n_s})^T = \sum_{\mathbf{x}\geq\mathbf{0}} \mathbf{x}\mathbf{P}(\mathbf{x},t), \\ \text{covariance} \quad & \mathbf{C} = \sum_{\mathbf{x}\geq\mathbf{0}} (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T \mathbf{P}(\mathbf{x},t), \\ \text{higher-order moments} \quad & C_{\mathbf{I}} = \sum_{\mathbf{x}\geq\mathbf{0}} (\mathbf{x}-\mathbf{m})^{\mathbf{I}} \mathbf{P}(\mathbf{x},t), \end{aligned} \tag{2.21}$$

where the sums go over all $\mathbf{x} \in \mathbb{N}_0^{n_s}$. Here, the following product notation is used:

$$(\mathbf{x}-\mathbf{m})^{\mathbf{I}} := \prod_{i=1}^{n_s} (x_i - m_i)^{I_i}, \tag{2.22}$$

where $\mathbf{I} = (I_1,\ldots,I_{n_s})$ is a vector of non-negative integers. The order of the moment $\mathbf{C_I}$ is defined to be $M_{\mathbf{I}} = \sum_{i=1}^{n_s} I_i$. The CME can be used to derive the governing equations for the dynamics of the above moments, by using the following lemma:

Lemma 1. Let $\mathbf{P}(\mathbf{x}|t)$ satisfy a proper Chemical Master Equation (2.10). Then, the time evolution of the expectation of a polynomial test-function $T_{\mathbf{I}}(\mathbf{x})$: $\mathbb{N}_0^{n_s} \to \mathbb{R}$ is governed by the following ODE:

$$\frac{\partial}{\partial t}\mathbb{E}[T_{\mathbf{I}}(\mathbf{x})] = \sum_{\mathbf{x}\geq\mathbf{0}} T_{\mathbf{I}}(\mathbf{x}) \frac{\partial}{\partial t}\mathbf{P}(\mathbf{x}|t) = \sum_{j=1}^{n_r} \mathbb{E}\big[(T_{\mathbf{I}}(\mathbf{x}-\nu_j) - T_{\mathbf{I}}(\mathbf{x}))a_j(\mathbf{x})\big]. \tag{2.23}$$

The proof of Lemma 1 is provided in [Engblom, 2006]. Choosing appropriate test functions $T_{\mathbf{I}}(\mathbf{x})$ and using the Taylor expansion of the propensities $a_j(\mathbf{x})$, one can derive the ODEs for the temporal evolution of the moments. In the simplest case, $T_{\mathbf{e}_i}(\mathbf{x}) = x_i$ is chosen for which Lemma 1 yields the temporal evolution of the mean $\mathbb{E}[T_{\mathbf{e}_i}(\mathbf{x})] = \mathbb{E}[x_i] = m_i$. Here, $\mathbf{e}_i$ is a vector of length n_s in which the i^{th} element is 1 and all other elements are zero. Similarly, choosing $T_{\mathbf{e}_i+\mathbf{e}_k}(\mathbf{x}) = (x_i - m_i)(x_k - m_k)$, Lemma 1 provides the ODEs governing the temporal evolution of the covariance C_{ik}.

Given that the chemical reactions are at most bimolecular, so that the third- and higher-order terms in their Taylor series expansion vanish, the moment equations for the mean,

covariance and higher-order central moments can be written as:

$$\frac{\partial m_i(t)}{\partial t} = \sum_{j=1}^{n_r} \nu_{i,j} \left(a_j(\mathbf{m}(t)) + \frac{1}{2} \sum_{k_1,k_2} \frac{\partial^2 a_j(\mathbf{m}(t))}{\partial x_{k_1} \partial x_{k_2}} C_{k_1 k_2}(t) \right), \tag{2.24}$$

$$\begin{aligned}\frac{\partial C_{i_1 i_2}(t)}{\partial t} = &\sum_{j=1}^{n_r} \left(\nu_{i_1,j} \sum_k \frac{\partial a_j(\mathbf{m}(t))}{\partial x_k} C_{i_1 k} + \nu_{i_2,j} \sum_k \frac{\partial a_j(\mathbf{m}(t))}{\partial x_k} C_{i_2 k} \right. \\ &\left. + \nu_{i_1 j} \nu_{i_2 j} \left(a_j(\mathbf{m}(t)) + \frac{1}{2} \sum_{k_1,k_2} \frac{\partial^2 a_j(\mathbf{m}(t))}{\partial x_{k_1} \partial x_{k_2}} C_{k_1 k_2}(t) \right) \right) \\ &+ \sum_{j=1}^{n_r} (\nu_{i_1 j} \sum_{k_1,k_2} \frac{\partial^2 a_j(\mathbf{m}(t))}{\partial x_{k_1} \partial x_{k_2}} C_{i_1 k_1 k_2}(t) + \nu_{i_2 j} \sum_{k_1,k_2} \frac{\partial^2 a_j(\mathbf{m}(t))}{\partial x_{k_1} \partial x_{k_2}} C_{i_2 k_1 k_2}(t) \Bigg),\end{aligned} \tag{2.25}$$

$$\begin{aligned}&\frac{\partial C_{[I_1,\cdots,I_{n_s}]}(t)}{\partial t} = \\ &\sum_{j=1}^{n_r} a_j(\mathbf{m}(t)) \sum_{\substack{l_1,l_2,\cdots,l_{n_s} \\ l_1+l_2+\cdots+l_{n_s} \neq M}} \binom{I_1}{l_1} \cdots \binom{I_{n_s}}{l_{n_s}} \nu_{1,j}^{I_1-l_1} \cdots \nu_{n_s,j}^{I_{n_s}-l_{n_s}} C_{[l_1,\cdots,l_{n_s}]}(t) \\ &+ \sum_{j=1}^{n_r} \sum_k \frac{\partial a_j(\mathbf{m}(t))}{\partial x_k} \sum_{\substack{l_1,l_2,\cdots,l_{n_s} \\ l_1+l_2+\cdots+l_{n_s} \neq M}} \binom{I_1}{l_1} \cdots \binom{I_{n_s}}{l_{n_s}} \nu_{1,j}^{I_1-l_1} \cdots \nu_{n_s,j}^{I_{n_s}-l_{n_s}} C_{[l_1,\cdots,l_k+1,\cdots,l_{n_s}]}(t) \\ &+ \frac{1}{2} \sum_{j=1}^{n_r} \sum_{k_1,k_2} \frac{\partial^2 a_j(\mathbf{m}(t))}{\partial x_{k_1} \partial x_{k_2}} \times \\ &\quad \sum_{\substack{l_1,l_2,\cdots,l_{n_s} \\ l_1+l_2+\cdots+l_{n_s} \neq M}} \binom{I_1}{l_1} \cdots \binom{I_{n_s}}{l_{n_s}} \nu_{1,j}^{I_1-l_1} \cdots \nu_{n_s,j}^{I_{n_s}-l_{n_s}} C_{[l_1,\cdots,l_{k_1}+1,\cdots,l_{k_2}+1,\cdots,l_{n_s}]}(t) \\ &- \sum_{k=1}^{n_s} I_k \frac{\partial m_k(t)}{\partial t} C_{[I_1,\cdots,I_k-1,\cdots,I_{n_s}]}(t).\end{aligned} \tag{2.26}$$

Here, $C_{ij}(t)$ denotes the $(i,j)^{\text{th}}$ element of the covariance matrix $\mathbf{C}$ in (2.21), and $C_{i_1 k_1 k_2}(t)$ denotes the third-order moment $C_{[I_1,\cdots,I_{n_s}]}(t)$ where $I_{i_1} = 1, I_{k_1} = 1, I_{k_2} = 1$ and all other elements are zero. The notation $C_{[I_1,\cdots,I_{n_s}]}$ is used according to (2.21). The details of derivation can be found in [Lee et al., 2009, Appendix]. If the reactions have more than 2 reactants, or if they do not follow the law of mass-action such that the third- and higher-order derivatives of the propensity function do not vanish, then moments of order $M+2$ or higher will appear in the moment equations of order M; in some cases, e.g., if the propensities are rational functions, infinitely many higher-order moments will appear [Milner et al., 2011].

Moment closure

The moment equations are generally not closed, meaning that the time evolution of moments of order M depends on moments of orders higher than M. For instance, the covariance equation (2.25) includes third-order moments ($C_{i_1k_1k_2}$ and $C_{i_2k_1k_2}$). In this way, the moment equations form an infinitely large coupled system of ODEs that cannot be integrated. Therefore, to enable numerical or analytical solution of the moment equations, one needs to truncate and close the ODE system by applying a so-called *moment closure* technique. Moment closure introduces an error to the otherwise exact moment equations (2.24)-(2.26), as it merely approximates the higher-order moments in terms of the lower-order moments to truncate the infinite set of moment equations. The magnitude of this error depends on the closure scheme used. For a specific class of systems, that includes the systems with linear propensities, e.g., monomolecular reactions with mass-action kinetics, the second-order derivative of the propensities vanishes, and as a result, the moment equations will not depend on higher-order moments. Therefore, the moment equations for systems with linear propensities are closed and exact.

Using a moment closure technique, one truncates the moment equations at a desirable order M, and approximates the higher-order moments as functions of moments of orders smaller than or equal to M [Engblom, 2006, Hespanha, 2007, Lee et al., 2009, Ruess et al., 2011, Singh and Hespanha, 2011]. Several approaches are available for the approximation of higher-order moments. The three most common approaches are: 1) Making specific assumptions, e.g., normality, about the distribution; 2) Approximating higher-order moments in such a way that the derivatives in the resulting system of equations matches that of the original system of equations; 3) Making specific assumptions about the process, for instance assuming that the process is in the regime where the macroscopic approximation is valid.

In this thesis, I mainly apply four well-know closure schemes that are detailed below:

- **Low-dispersion closure [Hespanha, 2008].** If the distribution is tightly clustered around the mean, i.e., the standard deviation is much smaller than the mean, then the higher-order central moments of the distribution can be negligible compared to the lower-order moments. The low-dispersion closure of M^{th}-order, therefore, sets the moments of order $M+1$ and higher to zero. For instance, the low-dispersion closure sets the third-order term $C_{i_1k_1k_2}$ to zero in (2.25). If the distribution is rather symmetrical, the odd-order moments become quite negligible. Hence, the low-dispersion closure applied on even-order moment equations usually yields more

accurate results.

- **Zero-cumulants [Matis and Kiffe, 1999].** Cumulants of a distribution are statistical quantities that are functions of the moments of that distribution. For instance, the first four cumulants of a univariate distribution, $\kappa_1, \kappa_2, \kappa_3, \kappa_4$, are written as functions of the first four central moments:

$$\kappa_1 = m, \quad \kappa_2 = C_2, \quad \kappa_3 = C_3, \quad \kappa_4 = C_4 - 3C_2^2.$$

For a multivariate distribution, the cumulants between random variables $X_1, \cdots, X_k$ can be represented in terms of the non-central moments via the following:

$$\kappa(X_1, \cdots, X_k) = \sum_{\pi} (|\pi| - 1)!(-1)^{|\pi|-1} \prod_{B \in \pi} E\left(\prod_{i \in B} X_i\right), \tag{2.27}$$

where the first sum runs over all partitions π in the set $\{1, \cdots, k\}$, and the first product runs over all blocks B in a partition π, and $|\pi|$ denotes the number of blocks in partition π.

By definition, the third- and all higher-order cumulants of a Normal distribution are zero. Therefore, by applying the zero-cumulants closure to the moment equations of order M, one assumes that the distribution of interest is similar to a Normal distribution and sets all cumulants of order $M+1$ and higher to zero. This equality yields an expression of the higher-order moments in terms of lower-order moments. For instance, applying the zero-cumulants closure to the third-order moment equations for a univariate distribution, one obtains: $\kappa_4 = 0 \quad \Rightarrow \quad C_4 = 3C_2^2$. Similar to the low-dispersion closure, applying zero-cumulants closure sets $C_{i_1 k_1 k_2} = 0$ in (2.25).

- **Derivative-matching/Log-normal closure [Singh and Hespanha, 2007].**

Derivative-matching closure aims at representing the higher-order moments in terms of the lower-order moments in such a way that minimises the difference between the derivatives of the original ODE system and the closed ODE system. More specifically, a moment of order $M+1$ is represented as a separable function of lower-order moments:

$$C_{M+1} = C_M^{\gamma_M} C_{M-1}^{\gamma_{M-1}} \cdots C_1^{\gamma_1}. \tag{2.28}$$

The coefficients $\gamma_1, \cdots, \gamma_M$ are chosen in such a way that the relative error of the derivatives is minimised:

$$\left| \frac{\frac{\partial^l \tilde{C}}{\partial t^l} - \frac{\partial^l C}{\partial t^l}}{\frac{\partial^l C}{\partial t^l}} \right|, \tag{2.29}$$

where $\tilde{C}$ is the vector of approximate moments after applying moment closure. Even though the Derivative-matching closure does not make an explicit assumption about the distribution, the resulting expressions for the higher-order moments match those of a log-normal distribution. Applying derivative-matching closure to (2.25), one obtains

$$C_{i_1k_1k_2} = \frac{C_{i_1k_1}C_{i_1k_2}C_{k_1k_2} + C_{i_1k_1}C_{i_1k_2}m_{k_1}m_{k_2} + C_{i_1k_1}C_{k_1k_2}m_{i_1}m_{k_2} + C_{i_1k_2}C_{k_1k_2}m_{i_1}m_{k_1}}{m_{i_1}m_{k_1}m_{k_2}}.$$

- **Mean-field closure.** The Mean-field closure assumes independence between different random variables. Consequently, the joint non-central moments can be represented as products of moments of individual variable, i.e.,

$$\hat{C}_{\mathbf{I}} = \mathbb{E}[X_1^{I_1}X_2^{I_2}\dots X_N^{I_N}] = \mathbb{E}[X_1^{I_1}]\mathbb{E}[X_2^{I_2}]\dots\mathbb{E}[X_N^{I_N}] = \hat{C}_{I_1\mathbf{e}_1}\hat{C}_{I_2\mathbf{e}_2}\dots\hat{C}_{I_N\mathbf{e}_N}, \quad (2.30)$$

  where $\hat{C}_{\mathbf{I}}$ is the non-central moment of order $M_{\mathbf{I}} = \sum_{i=1}^{N} I_i$, and $\mathbf{e}_i$ is a vector of length N in which the i^{th} element is 1 and all other elements are zero. Using conversion relations between central and non-central moments, the mean-field closure can be applied to central moment equations. According to mean-field closure, the third-order moment in (2.25) is approximated as $C_{i_1k_1k_2} = m_{i_1}m_{k_1}m_{k_2}$.

Prior knowledge about the distribution $\mathbf{P}(\mathbf{x}, t)$ can be used to guide the choice of moment closure. For instance, if the counts of species in the biochemical reaction network are expected to be normally distributed, then the zero-cumulants closure could be a proper choice. Since such a knowledge is usually not available, the derivative-matching closure can be used that does not make an explicit distribution assumption. However, as a priori error bounds for various moment closures are not available, the optimal choice of moment closure cannot be guaranteed beforehand, and simulation comparisons can be used for this purpose. For instance, in simulating a three-stage model of gene expression (Figure 2.1), the third-order moment approximation with derivative-matching closure correctly predicts the mRNA and protein levels. Predictions given by the third-order moment approximations with low-dispersion and zero-cumulants closures, however, deviate considerably from the reference solution given by the Stochastic Simulation Algorithm. In this simulation study, increasing the order of (conditional) moment approximation method almost consistently decreases the relative error in predicting the protein concentration for various moment closures. In many applications, low-dispersion closure is used for its simplicity and numerical efficiency.

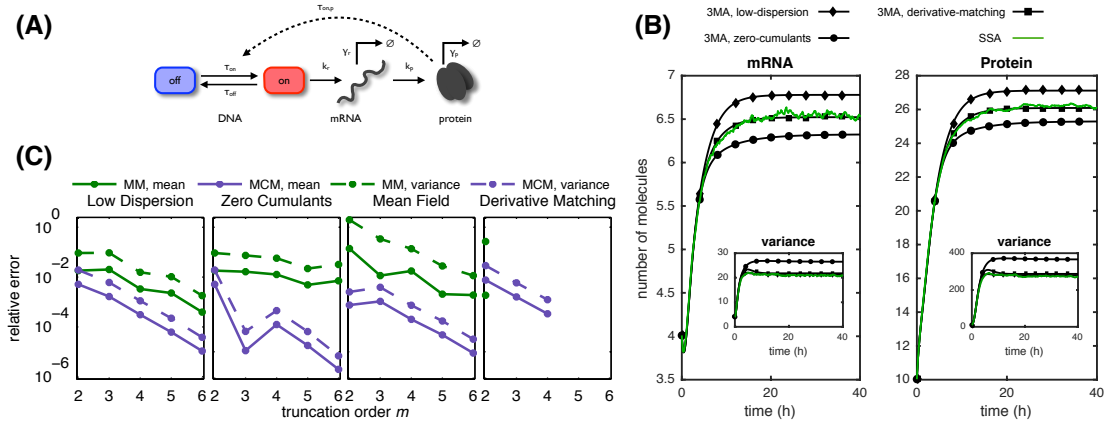


Figure 2.1: **Performance of different moment closure techniques in predicting the mRNA and protein levels in a three-stage model of gene expression.** (A) Schematic of the three-stage model of gene expression used for a comparative study of different moment closures. (B) mRNA and protein levels predicted by third-order moment approximations with derivative-matching, low-dispersion and zero-cumulants closures, compared to the reference solution given by the Stochastic Simulation Algorithm. (C) Influence of the order of moment approximation method on the performance of several moment closure techniques. (A, C) Figures taken from [Kazeroonian et al., 2016].

2.2.3 System size expansion

The *System Size Expansion (SSE)* [van Kampen, 2007, Grima, 2010, Thomas et al., 2013] is a systematic approximation method that yields mesoscopic approximations to solution of the CME. The SSE provides a power series expansion, derived from the CME, whose truncation order determines the order of the approximation error. The idea of the SSE, developed by van Kampen [2007], is based on deriving a series expansion in powers of a small parameter Ω^{-1}. To derive an expansion series which describes dynamics of the internal noise or stochasticity to various orders, the chosen parameter Ω^{-1} must govern the size of the fluctuations. In particular, we desire that the size of the fluctuations are small for small Ω^{-1} or large Ω, such that the contribution of the expansion terms for describing the dynamics of noise decreases with their order.

To choose such a parameter, we note that the internal noise in chemical reaction networks stems from the discrete nature of molecular species and chemical reaction. Contrarily, the macroscopic features of reaction network are resulted from the collective behaviour of all molecules together [van Kampen, 2007]. Therefore, it is intuitive to expect that the *size of the system*, e.g., the cellular volume in which the biochemical reactions occur, determines the importance of the fluctuations. Therefore, the *inverse of the size parameter* is chosen

as the expansion parameter Ω^{-1}.

To have the system size appear explicitly in the CME, we rewrite the CME (2.10) as:

$$\frac{\partial}{\partial t}P(\mathbf{x},t) = \Omega \sum_{j=1}^{n_r} \Big(\hat{f}_j(\mathbf{x}-\nu_j)P(\mathbf{x}-\nu_j,t) - \hat{f}_j(\mathbf{x})P(\mathbf{x},t) \Big), \tag{2.31}$$

where $\Omega \hat{f}_j(\mathbf{x}) = a_j(\mathbf{x})$ is the propensity of reaction R_j. The limit of the function $\hat{f}_j$ as Ω tends to infinity is the familiar macroscopic reaction rate used in Reaction Rate Equations (see Section 2.2.4).

In studying stochastic systems, it is commonly observed that, in a collection of N particle, e.g., N molecules, the fluctuations are of order $N^{\frac{1}{2}}$. More specifically, it is usually expected that $\mathbf{P}(\mathbf{x},t)$ has a sharp maximum around the macroscopic value $\mathbf{x} = \Omega\phi(t)$, where $\phi(t)$ is the concentration of species as given by the macroscopic reaction rate equations (see Section 2.2.4). The width of the distribution $\mathbf{P}(\mathbf{x},t)$ around this macroscopic value is then assumed to be of order $\mathbf{x}^{1/2} \sim \Omega^{1/2}$. Following this rule of thumb, we use the following ansatz as the basic step towards the derivation of the SSE:

$$\mathbf{x} = \Omega\phi(t) + \Omega^{\frac{1}{2}}\xi, \tag{2.32}$$

Parameter ξ represents the fluctuations around these macroscopic concentrations. Using ansatz (2.32), we transform the probability function $P(\mathbf{x},t)$ into the probability function of ξ:

$$P(\mathbf{x},t) = P(\Omega\phi(t) + \Omega^{\frac{1}{2}}\xi) = \Pi(\xi,t). \tag{2.33}$$

According to the transformation (2.33), one can calculate derivatives of $\Pi(\xi,t)$ in terms of $P(\mathbf{x},t)$, and rewrite the CME in terms of the variable ξ. The resulting ODE system, which is an expansion series in powers of Ω, constitute the SSE equations. Details of the derivation is provided in [van Kampen, 2007].

The lowest order of the SSE, reproduces the macroscopic reaction rate equations:

$$\frac{\partial \phi_i}{\partial t} = \sum_{j=1}^{n_r} \nu_{ij} f_j(\phi), \tag{2.34}$$

where $f_j(\phi) = \lim_{\Omega\to\infty} \hat{f}_j(\Omega\phi)$ is the macroscopic rate function.

The next order term in SSE, called the *Linear Noise Approximation (LNA)*, describes the covariance of the fluctuations around the macroscopic concentrations given by (2.34):

$$\frac{\partial \Sigma_{ik}}{\partial t} = \sum_{j=1}^{n_r} \sum_{l=1}^{n_s} \left(\nu_{ij} \frac{\partial f_j(\phi)}{\partial \phi_l} \Sigma_{lk} + \nu_{kj} \frac{\partial f_j(\phi)}{\partial \phi_l} \Sigma_{li} \right) + \frac{1}{\Omega} \sum_{j=1}^{n_r} \nu_{ij} \nu_{kj} f_j(\phi). \tag{2.35}$$

For CMEs with unimodal solutions, the LNA can be interpreted as a Gaussian approximation of the probability density function of the CME.

The RRE and LNA are exact for all CRNs including at most monomolecular reactions, as well as a small class of bimolecular reactions [Grima, 2015]. In cases where (2.34) and (2.35) are not exact, one can take higher orders of the SSE into account to systematically correct the mean concentrations and the covariance of fluctuations predicted by RRE and LNA respectively. For instance, the *Effective Mesoscopic Rate Equation (EMRE)* [Grima, 2010], provides a more accurate approximation for mean concentrations:

$$\frac{\partial \mu_i}{\partial t} = \frac{\partial \phi_i}{\partial t} + \sum_{j=1}^{n_r} \nu_{ij} \left(\sum_{k=1}^{n_s} \frac{\partial f_j(\phi)}{\partial \phi_k} (\mu_k - \phi_k) + \frac{1}{2} \sum_{k,l=1}^{n_s} \frac{\partial^2 f_j(\phi)}{\partial \phi_k \phi_l} \Sigma_{kl} - \frac{1}{2} \sum_{k=1}^{n_s} \frac{\phi_k}{\Omega} \frac{\partial^2 f_j(\phi)}{\partial \phi_k^2} \right). \tag{2.36}$$

The EMRE yields correction terms, of order Ω^{-1}, to the solution of the RRE. Similarly, the next order of the SSE, the *Inverse Omega Square (IOS)* [Thomas et al., 2013], describes the covariance of fluctuations around the mean μ, predicted by the EMRE. In this way IOS provides correction terms of order Ω^{-2} to the solution of the LNA. The SSE methods tend to be more accurate for systems of small and medium volumes [Ramaswamy et al., 2012].

2.2.4 Macroscopic rate equation

All the methods described so far concern with microscopic or mesoscopic description of the dynamics of the CRNs where the noise is taken into account. In the limit of large molecule numbers where the importance of fluctuations are small, however, it is very common to resort to a macroscopic description of the system where merely the average behaviour is studied.

A macroscopic approximation describes the dynamics of the macroscopic variable, $\mathbf{m} = \mathbb{E}[\mathbf{x}]$, i.e., the expected value of the state of the system. As it was shown in Section 2.1.3, the CME was derived by taking into account every single update in the state of the CRN, and the consequent update in the reaction propensities. Contrarily, we note that in the

limit of large molecule numbers, the effect of every single state update on the propensities is negligible. Therefore, we can assume the propensity of a reaction R_j, $a_j(\mathbf{m})$, remains constant in a small time interval Δt. This implies that the time spent between firings of the reaction R_j in Δt are exponentially distributed, and that the reaction R_j can on average fire $a_j(\mathbf{m})\Delta t$ in the interval Δt. Thus, the number of firings of R_j in Δt, N_j, is Poisson distributed with parameter $\lambda = a_j(\mathbf{m})\Delta t$:

$$N_j \sim \text{Pois}(a_j(\mathbf{m})\Delta t) \Rightarrow \mathbb{E}[N_j] = a_j(\mathbf{m})\Delta t. \tag{2.37}$$

Knowing the expected number of the firings of each reaction from (2.37), one can obtain the change in the macroscopic state of the system $\Delta\mathbf{m}$ in Δt:

$$\Delta\mathbf{m} = \sum_{j=1}^{n_r} \nu_j \mathbb{E}[N_j] = \sum_{j=1}^{n_r} \nu_j a_j(\mathbf{m})\Delta t, \tag{2.38}$$

or in matrix form

$$\Delta\mathbf{m} = \mathbf{S}\,\mathbf{a}(\mathbf{m})\Delta t, \tag{2.39}$$

where $\mathbf{S}$ is the stoichiometry matrix

$$\mathbf{S} = \begin{pmatrix} \nu_{1,1} & \nu_{1,2} & \cdots & \nu_{1,n_r} \\ \nu_{2,1} & \nu_{2,2} & \cdots & \nu_{2,n_r} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_{n_s,1} & \nu_{n_s,2} & \cdots & \nu_{n_s,n_r} \end{pmatrix}, \tag{2.40}$$

and $\mathbf{a}(\mathbf{m})$ is the vector of propensities of all reactions.

Taking the limit of $\Delta t \to 0$, we arrive at

$$\lim_{\Delta t \to 0} \frac{\Delta\mathbf{m}}{\Delta t} = \frac{\partial \mathbf{m}}{\partial t} = \mathbf{S}\,\mathbf{a}(\mathbf{m}). \tag{2.41}$$

Equation (2.41) is the macroscopic rate equation which describes the time evolution of the macroscopic state of the CRN. Equation (2.41) is an ODE system of size n_s, meaning that the size of the macroscopic rate equation scales linearly with the number of species.

It is customary to present the macroscopic equation in terms of the concentration of species $\phi = \frac{\mathbf{m}}{\Omega}$ with Ω being the volume of the compartment in which the reactions take place. Rewriting (2.41) in terms of ϕ, we arrive at the *Reaction Rate Equation (RRE)*:

$$\frac{\partial \phi}{\partial t} = \mathbf{S}\,\mathbf{f}(\phi), \tag{2.42}$$

Table 2.2: Macroscopic rate functions according to the law of mass action.

Reaction Order	Reaction Type	Macroscopic rate function
0	$\emptyset \xrightarrow{k'_j}$ product	k_j
1	$S_i \xrightarrow{k'_j}$ product	$k_j \phi_i$
2	$S_i + S_l \xrightarrow{k'_j}$ product	$k_j \phi_i \phi_l$
2	$S_i + S_i \xrightarrow{k'_j}$ product	$k_j \phi_i^2$

where $\mathbf{f}(\phi)$ is the vector of the macroscopic rate functions. Assuming the law of mass action, the macroscopic rate functions for reactions of at most order 2 are listed in Table 2.2. The parameter k'_j denotes the macroscopic rate constant of the reaction. The relationship of the macroscopic rate constants to the kinetic constants introduced in Section 2.1.1 is provided in the Supplement of [Kazeroonian et al., 2016]. The macroscopic rate functions in terms of concentrations are similar to reaction propensities in terms of molecule numbers. Only, in the case of two reactant molecules of the same species, $(X_i)(X_i - 1)$ is approximated by ϕ_i^2 which is justified in the limit of large molecule numbers.

For reaction networks with constant and linear propensities, the solution of the RRE (2.42) yields the exact mean concentrations. However, for most bimolecular reactions an approximation error is introduced. The reason is the implicit assumption that the expectation of a propensity equals the propensity evaluated at the expected value of the state, i.e., $\mathbb{E}[a(\mathbf{x})] = a(\mathbb{E}[\mathbf{x}]) = a(\mathbf{m})$. In such cases, the solution of RRE is only reflective of the true mean concentration in the limit of large molecule numbers [Grima, 2015].

Alternatively, the RRE is derived/reproduced as the lowest-order of the SSE (see Section 2.2.3).

2.3 Parameter estimation

In previous sections, I provided a brief overview of several methods which yield mathematical models for describing the dynamics of biochemical reaction networks. All the mentioned mathematical models depend on some parameters, such as kinetic constants or initial conditions, and thus, can only yield reliable predictions for the behaviour of the CRNs if reliable parameter values are provided. Since in general the parameters are not known and cannot be directly measured, experimental measurements are usually used to calibrate the mathematical models and estimate the parameter values.

Experimental data about the behaviour of biological systems can be obtained via various measurement techniques, and therefore, can provide various levels of information. For instance population average data, e.g., obtained by Western Blots, provide the average value of the measured quantity, e.g., concentration of a species, in a population of cells.

In this thesis, population snapshot data have been used for parameter estimation in CRNs. Population snapshot data, obtained via flow or (non time-lapse) cytometry, includes the values of the measured quantities in individual cells in a population. This type of data may be used in a distribution-based parameter estimation framework, as in Section 2.3.3, to exploit the maximum information provided by the data, and thus, minimise uncertainty of the estimates. The high computational cost of this framework, however, makes it infeasible for many realistic applications. To circumvent this obstacle, merely the statistical moments of the data can be exploited by a moment-based estimation framework (Section 2.3.2), thereby greatly reducing the computational cost. In the following, these likelihood-based approaches are described.

2.3.1 Likelihood-based parameter estimation

In this work, I have used likelihood-based approaches for parameter estimation. Alternatively, one can use Bayesian framework for parameter inference. Given a measured data $\mathcal{D}$ and a model $\mathcal{M}$, the likelihood-based parameter estimation method aims at finding the set of the model parameters θ for which it is most likely to observe $\mathcal{D}$ given the model $\mathcal{M}$.

Consider a model $\mathcal{M}$ with state variables $\mathbf{x}$ and observables $\mathbf{y}$:

$$\mathcal{M}(\theta) = \begin{cases} \dot{\mathbf{x}} = f(\mathbf{x}, \theta, t), \quad \mathbf{x}(t_0) = \mathbf{x}_0(\theta), \\ \mathbf{y} = h(\mathbf{x}, \theta, t), \end{cases} \tag{2.43}$$

and a dataset $\mathcal{D}$ consisting of noise-corrupted measurements $\bar{y}_k$ of the observables $\mathbf{y}$ at time points t_k, $k = 1, \ldots, n_t$:

$$\mathcal{D} = \{(t_k, \bar{y}_k)\}_{k=1}^{n_t}. \tag{2.44}$$

The likelihood of data $\mathcal{D}$ given a set of parameters θ is then defined as

$$\mathcal{L}_{\mathcal{D}}(\theta) = P(\mathcal{D}|\theta) = \prod_{k=1}^{n_t} P(\bar{y}_k | y(t_k, \theta), \sigma_k). \tag{2.45}$$

If there are n_y independent observables, $\bar{y}_k = \{\bar{y}_{i,k}\}_{i=1}^{n_y}$, the likelihood (2.45) is rewritten as

$$\mathcal{L}_{\mathcal{D}}(\theta) = \prod_{k=1}^{n_t} \prod_{i=1}^{n_y} P(\bar{y}_{i,k}|y_i(t_k,\theta),\sigma_{i,k}), \tag{2.46}$$

where, $y_i(t_k,\theta)$ are the model predictions for i^{th} observable at time t_k, and $\sigma_{i,k}$ denotes the measurement noise. The *maximum likelihood estimate (MLE)* is then defined as the set of parameters θ^{MLE} that maximises the likelihood function $\mathcal{L}_{\mathcal{D}}(\theta)$:

$$\begin{aligned} \theta^{\text{MLE}} =& \underset{\theta}{\text{argmax}}\ \mathcal{L}_{\mathcal{D}}(\theta) \\ & \text{subject to } \mathcal{M}(\theta). \end{aligned} \tag{2.47}$$

To improve the numerical robustness and convergence of optimisers, usually the negative log-likelihood $J(\theta) = -\log \mathcal{L}_{\mathcal{D}}(\theta)$ is minimised to obtain the MLE:

$$\theta^{\text{MLE}} = \arg\min_{\theta\in\Theta} J(\theta), \tag{2.48}$$

with Θ being the search region in the parameter space. To further improve the numerical properties of the optimisation problem (2.48), usually log-transformed parameters $\xi = \log(\theta)$ are used [Raue et al., 2013].

Likelihood function for additive normally distributed measurement noise. In case of additive measurement noise that is normally distributed, i.e., $\bar{y}_{i,k} = y_i(t_k,\theta) + \epsilon_{i,k}$ with $\epsilon_{i,k} \sim \mathcal{N}(0, \sigma_{i,k}^2(\theta))$, the conditional probability of data given model parameters is given by

$$P(\bar{y}_{i,k}|y_i(t_k,\theta),\sigma_{i,k}^2) = \mathcal{N}(\bar{y}_{i,k}|y_i(t_k,\theta),\sigma_{i,k}^2(\theta)). \tag{2.49}$$

Consequently, we obtain the following for the likelihood and the negative log-likelihood functions:

$$\begin{aligned} \mathcal{L}_{\mathcal{D}}(\theta) &= \prod_{k=1}^{n_t} \prod_{i=1}^{n_y} \frac{1}{\sqrt{2\pi}\sigma_{i,k}(\theta)} \exp\left\{-\frac{1}{2}\left(\frac{\bar{y}_{i,k} - y_i(t_k,\theta)}{\sigma_{i,k}(\theta)}\right)^2\right\}, \\ J(\theta) &= \frac{1}{2}\sum_{k=1}^{n_t}\sum_{i=1}^{n_y} \log\left(2\pi\sigma_{i,k}^2(\theta)\right) + \left(\frac{\bar{y}_{i,k} - y_i(t_k,\theta)}{\sigma_{i,k}(\theta)}\right)^2. \end{aligned} \tag{2.50}$$

To find the MLE parameters θ^{MLE}, the optimisation problem (2.48) can be efficiently solved using gradient-based optimisation algorithms [Raue et al., 2013, Coleman and Li, 1992, 1996]. For this purpose the gradient of the objective function $J(\theta)$ needs to be evaluated. The gradient of $J(\theta)$ is a function of the gradient of the observables with

respect to model parameters. For instance, in the case of normally-distributed additive noise, the gradient of the objective function is derived as

$$\begin{aligned}\frac{\partial J}{\partial \theta_l} &= \frac{1}{2}\sum_{k=1}^{n_t}\sum_{i=1}^{n_y}\frac{\partial}{\partial \theta_l}\log\left(2\pi\sigma_{i,k}^2(\theta)\right) + \frac{\partial}{\partial \theta_l}\left(\frac{\bar{y}_{i,k}-y_i(t_k,\theta)}{\sigma_{i,k}(\theta)}\right)^2 \\ &= \frac{1}{2}\sum_{k=1}^{n_t}\sum_{i=1}^{n_y}\frac{1}{\sigma_{i,k}^2(\theta)}\left(1-\frac{(\bar{y}_{i,k}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)}\right)\frac{\partial\sigma_{i,k}^2(\theta)}{\partial\theta_l} - 2\frac{\bar{y}_{i,k}-y_i(t_k,\theta)}{\sigma_{i,k}^2(\theta)}\frac{\partial y_i(t_k,\theta)}{\partial\theta_l}.\end{aligned} \tag{2.51}$$

To calculate $\nabla_\theta J$, the gradient of the observables with respect to parameters, $\nabla_\theta y_i(t_k,\theta)$, needs to be evaluated. The latter can be obtained using forward sensitivity analysis [Hindmarsh et al., 2005] in a robust and computationally efficient way.

Forward sensitivity analysis

Having defined a likelihood function, the parameters of the model $\mathcal{M}$ can be estimated by using various optimisation methods for the optimisation problem (2.48). To derive the model $\mathcal{M}$, in this thesis, the methods described in Section 2.2 were used, i.e. the FSP, MA, SSE, and RRE, as well as the method of conditional moments [Hasenauer et al., 2014b]. All these methods yield systems of differential equations, whose parameters can be estimated efficiently using gradient-based optimisation methods [Raue et al., 2013]. A naive way of approximating the gradient of the likelihood function with respect to the parameters is to use finite difference methods. However, for more robust and computationally more efficient results one resorts to methods based on sensitivity equations [Raue et al., 2013]. Sensitivities indicate the change in a functional of the process, e.g., the observables, in response to a change in parameter values. The forward sensitivity analysis [Hindmarsh et al., 2005] provide the time-dependent sensitivity of the state-variables of the differential equations with respect to the parameters. Consider the model $\mathcal{M}$ (2.43), which is an n-dimensional ODE system

$$\begin{aligned}\dot{\mathbf{x}} &= f(\mathbf{x},\theta,t), \quad \mathbf{x}(t_0) = \mathbf{x}_0(\theta) \\ \mathbf{y} &= h(\mathbf{x},\theta,t),\end{aligned} \tag{2.52}$$

or an n-dimensional DAE system

$$\begin{aligned}F(\dot{\mathbf{x}},\mathbf{x},\theta,t) &= 0, \quad \mathbf{x}(t_0) = \mathbf{x}_0(\theta), \quad \dot{\mathbf{x}}(t_0) = \dot{\mathbf{x}}_0(\theta) \\ \mathbf{y} &= h(\mathbf{x},\theta,t),\end{aligned} \tag{2.53}$$

with $\mathbf{x} \in \mathbb{R}^n$ being the state variables of the ODE/DAE system, $\mathbf{y} \in \mathbb{R}^{n_o}$ being the vector of observables, and $\theta \in \mathbb{R}^{n_\theta}$ being the set of parameters. The forward sensitivities for these systems are defined as

$$\begin{aligned} \mathbf{S}^x(t) &= \left(\mathbf{s}_1^x(t), \mathbf{s}_2^x(t), \cdots, \mathbf{s}_{n_\theta}^x(t)\right) \in \mathbb{R}^{n \times n_\theta}, \\ \mathbf{s}_i^x(t) &= \frac{\partial \mathbf{x}(t)}{\partial \theta_i}, \quad \text{for } i = 1, 2, \cdots, n_\theta, \end{aligned} \tag{2.54}$$

in which $\mathbf{s}_i^x(t) \in \mathbb{R}^n$ is the sensitivity of the state variables $\mathbf{x}$ with respect to the i^{th} parameter, θ_i.

The forward sensitivity equations for the ODE system (2.52) are

$$\begin{aligned} \dot{\mathbf{s}}_i &= \frac{\partial f}{\partial x}\mathbf{s}_i + \frac{\partial f}{\partial \theta_i}, \quad \text{for } i = 1, \cdots, n_\theta, \\ \mathbf{s}_i(t_0) &= \frac{\partial \mathbf{x}_0(\theta)}{\partial \theta_i}. \end{aligned} \tag{2.55}$$

The forward sensitivity equations for the DAE system (2.53) are

$$\begin{aligned} &\frac{\partial F}{\partial x}\mathbf{s}_i + \frac{\partial F}{\partial \dot{x}}\dot{\mathbf{s}}_i + \frac{\partial F}{\partial \theta_i} = 0, \quad \text{for } i = 1, \cdots, n_\theta, \\ &\mathbf{s}_i(t_0) = \frac{\partial \mathbf{x}_0(\theta)}{\partial \theta_i}, \quad \dot{\mathbf{s}}_i(t_0) = \frac{\partial \dot{\mathbf{x}}_0(\theta)}{\partial \theta_i}. \end{aligned} \tag{2.56}$$

Therefore, the computation of the state sensitivities with respect to all parameters involves solving a system of $n(1 + n_\theta)$ differential equations.

The sensitivity of the observables $\mathbf{y}$ is defined as

$$\begin{aligned} \mathbf{S}^y(t) &= \left(\mathbf{s}_1^y(t), \mathbf{s}_2^y(t), \cdots, \mathbf{s}_{n_\theta}^y(t)\right) \in \mathbb{R}^{n_o \times n_\theta}, \\ \mathbf{s}_i^y(t) &= \frac{\partial \mathbf{y}(t)}{\partial \theta_i}, \quad \text{for } i = 1, 2, \cdots, n_\theta, \end{aligned} \tag{2.57}$$

where $\mathbf{s}_i^y(t) \in \mathbb{R}^{n_o}$ denotes the sensitivity of the observables $\mathbf{y}$ with respect to the i^{th} parameter, θ_i. Using the solution of the forward sensitivity equations, the sensitivities of the observables can be computed via

$$\mathbf{s}_i^y = \frac{\partial h}{\partial x}\,\mathbf{s}_i^x + \frac{\partial h}{\partial \theta_i}, \tag{2.58}$$

where $\frac{\partial h}{\partial x} = \left(\frac{\partial h_j}{\partial x_k}\right)_{jk} \in \mathbb{R}^{n_o \times n}$.

2.3.2 Moment-based likelihood function for population snapshot data

Consider the population snapshot data $\mathcal{D}_k = \left\{\left(t_k, \bar{y}_k^{(s)}\right)\right\}_{s=1}^{S_k}$ collected at measurement times t_k for $k = 1, \ldots, n_t$ by sampling S_k cells from the cell population. The statistics of the population snapshot data, e.g., statistical moments of various orders, can be used for a more informative parameter estimation in comparison with population average data [Zechner et al., 2012, Milner et al., 2013, Ruess and Lygeros, 2015]. For instance, the mean and variance of observables can be estimated from the population snapshot data as

$$\bar{\mu}_{y,k} = \frac{1}{S_k}\sum_{s=1}^{S_k} \bar{y}_k^{(s)},$$
$$\bar{C}_{yy,k} = \frac{1}{S_k}\sum_{s=1}^{S_k} \left(\bar{y}_k^{(s)} - \bar{\mu}_y(t_k)\right)^2.$$

These estimated moments can be compared to moments predicted by the model $\mathcal{M}$, e.g., using moment closure approximation method or system size expansion, to estimate the parameters. The sample sizes S_k of population snapshot data are usually quite large, e.g., in the order of 10^4 for flow cytometry. Thus, according to the central limit theorem, the estimated moments in (2.59) are expected to be normally distributed around the true moments [Zechner et al., 2012]. Hence, the likelihood of observing the empirical moments $\bar{\mu}_{y,k}$ and $\bar{C}_{yy,k}$ given the predicted moments $\mu_y(t_k, \theta)$ and $C_{yy}(t_k, \theta)$ is

$$\mathcal{L}_{\mathcal{D},\bar{\mu}_y}(\theta) = \prod_{k=1}^{n_t} \mathcal{N}\left(\bar{\mu}_{y,k} | \mu_y(t_k,\theta), \sigma^2_{\bar{\mu}_{y,k}}\right), \tag{2.59}$$
$$\mathcal{L}_{\mathcal{D},\bar{C}_{yy}}(\theta) = \prod_{k=1}^{n_t} \mathcal{N}\left(\bar{C}_{yy,k} | C_{yy}(t_k,\theta), \sigma^2_{\bar{C}_{yy,k}}\right). \tag{2.60}$$

Similar likelihood functions can be derived higher-order moments predicted by the model as well. Assuming independence of estimators for various moments, the overall likelihood function is obtained as the product of the likelihood of individual moments. For instance, if only mean and variance are employed for parameter estimation, then the overall likelihood function is given by $\mathcal{L}_{\mathcal{D}}(\theta) = \mathcal{L}_{\mathcal{D},\bar{\mu}_y}(\theta) \cdot \mathcal{L}_{\mathcal{D},\bar{C}_{yy}}(\theta)$, and, consequently, the negative log-likelihood is given by

$$\begin{aligned} J(\theta) =& \frac{1}{2}\sum_{k=1}^{n_t}\sum_{i=1}^{n_y} \log\left(2\pi\sigma^2_{\bar{\mu}_{y,k}}(\theta)\right) + \left(\frac{\bar{\mu}_{y,k} - \mu_y(t_k,\theta)}{\sigma_{\bar{\mu}_{y,k}}(\theta)}\right)^2 \\ &+ \frac{1}{2}\sum_{k=1}^{n_t}\sum_{i=1}^{n_y} \log\left(2\pi\sigma^2_{\bar{C}_{yy,k}}(\theta)\right) + \left(\frac{\bar{C}_{yy,k} - C_{yy}(t_k,\theta)}{\sigma_{\bar{C}_{yy,k}}(\theta)}\right)^2. \end{aligned} \tag{2.61}$$

The variance of estimators for the mean and variance are calculated as below:

$$\begin{aligned}\sigma^2_{\bar{\mu}_{y,k}} &= \mathbb{E}\left[(\bar{\mu}_{y,k} - \mu_y(t_k,\theta))^2\right] = \mathbb{E}\left[\left(\frac{1}{S_k}\sum_{s=1}^{S_k}\bar{y}_k^{(s)} - \mu_y(t_k,\theta)\right)^2\right] + \mathbb{E}\left[\epsilon^2_{k,T}\right] \\ &= \frac{1}{N}C_{yy}(t_k,\theta) + \sigma^2_{k,T}, \\ \sigma^2_{\bar{C}_{yy,k}} &= \mathbb{E}\left[\left(\bar{C}_{yy,k} - C_{yy}(t_k,\theta)\right)\right)^2\right] = \frac{1}{S_k}\left(C_{yyyy}(t_k,\theta) - \frac{S_k-3}{S_k-1}C^2_{yy}(t_k,\theta)\right),\end{aligned} \tag{2.62}$$

where $C_{yyyy}(t_k,\theta)$ denotes the fourth-order moment of $y(t_k,\theta)$, and $\sigma^2_{k,T}$ denotes the technical measurement noise. As the simulation of the fourth-order moment might be computationally demanding, the empirical fourth-order moment calculated from the population snapshot data may be used. The derivation of the variance of estimators can be found in [Fröhlich et al., 2016, Zechner et al., 2012]. Taking additional moments into account for the evaluation of the overall likelihood function results in a more informative estimation for the model parameters.

2.3.3 FSP-based likelihood function for population snapshot data

Consider the same population snapshot data $\mathcal{D}_k = \left\{\left(t_k, \bar{y}_k^{(s)}\right)\right\}_{s=1}^{S_k}$ as in Section 2.3.2. If the probability distribution of the observables, $p(y|t,\theta)$, is known, the likelihood of observing data $\mathcal{D}$ can be written as

$$\mathcal{L}_{\mathcal{D}}(\theta) = c\prod_{k=1}^{N}\prod_{s=1}^{S_k} p(y = \bar{y}^{(s)}(t_k)|t_k,\theta), \tag{2.63}$$

where the probabilities for all data points are evaluated and multiplied.

The finite state projection can be used to obtain the probability distribution of observables, $p(y|t,\theta)$. The FSP provides an approximation to the solution of the CME for a given parameter set θ, and therefore, can yield the probability distribution of the states, $p(x|t,\theta)$. Using the solution of the FSP, $p(x|t,\theta)$, and the conditional probability of observables given the states, $p(y|x)$, the total probability distribution for observables y is obtained by marginalising over all states $x \in \hat{\Psi}$,

$$p(y|t_k,\theta) = \sum_{x\in\hat{\Psi}} p(y|x)\; p(x|t_k,\theta), \tag{2.64}$$

where $\hat{\Psi}$ is the state space of the FSP (see Section 2.2.1). Assuming noise-free measurements, the observable y is a deterministic function of x, $y = h(x)$, thus

$$p(y|x) = \begin{cases} 1 & \text{if } y = h(x) \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the sum (2.64) is simplified to

$$p(y|t_k, \theta) = \sum_{\substack{x \in \hat{\Psi} \\ h(x)=y}} p(x|t_k, \theta). \tag{2.65}$$

The probability distribution (2.65) is the distribution from which the observations $\bar{y}_k^{(s)}$ are drawn. Substituting (2.65) in (2.63), the FSP-based likelihood function $\mathcal{L}_{\mathcal{D}}^{\text{FSP}}(\theta)$ is obtained. The maximum likelihood estimation problem using the FSP is formulated as:

$$\begin{aligned} &\underset{\theta}{\text{maximize}} \ \log \mathcal{L}_{\mathcal{D}}^{\text{FSP}}(\theta) \\ &\text{subject to } \Sigma^{\text{FSP}}(\theta), \end{aligned} \tag{2.66}$$

where $\Sigma^{\text{FSP}}(\theta)$ denotes the ODE system of the FSP. The constant c in (2.63) only depends on the data and can be disregarded for the optimisation problem (2.66). Unlike the moment-based ML estimator in Section 2.3.2, the FSP-based estimator uses all available information in the data as the likelihood of each single data point is taken into account. More details on the derivation of this FSP-based likelihood function are provided in [Hasenauer et al., 2011, Nüesch, 2010].

2.3.4 Identifiability and uncertainty analysis

The maximum likelihood estimators return the parameter vector θ^{MLE} that is most likely to generate the observed data. Due to the potential measurement noise and the finite sample size of the measurement data, however, the θ^{MLE} usually does not coincide with the true parameters of the system. Moreover, depending on the definition of the model $\mathcal{M}$, the set of observables $\{y_i(t_k, \theta)\}$, and the measurement data $\mathcal{D}$ it may not be possible to estimate (a subset of) parameters with acceptable/desirable certainty. Even assuming perfect measurement data (i.e. noise-free and continuous in time), it may not be possible to determine unique values for parameters. Such cases give rise to *non-identifiable parameters* [Raue et al., 2009, 2010]. Therefore, the MLE alone will not give a valuable insight about the true parameter values, unless these issues are addressed. Thus, one needs to know if the parameter are at all *identifiable*, and how certain the parameter estimates

are. For the latter, one usually considers the corresponding *confidence intervals* [Meeker and Escobar, 1995]. The *significance level* α of the calculated confidence interval states that if the estimation procedure were to be repeated on multiple samples, the calculated confidence interval would contain the true parameter value in $(1-\alpha)\%$ of the times.

To assess identifiability of parameters, two non-identifiability types can be considered:

- *Structural identifiability* [Raue et al., 2009, 2010] solely considers the structure of the model (including the definition of observables) and answers the question wether a certain parameter can be uniquely estimated given perfect (i.e. noise-free and continuous in time) measurement data.

- *Practical identifiability* [Raue et al., 2009] investigates the identifiability of parameters given a specific dataset $\mathcal{D}$. Given the data $\mathcal{D}$, a parameter is said to be identifiable if the corresponding confidence intervals are finite, or desirably tight.

Calculation of confidence intervals based on profile likelihoods

Several approaches for the calculation of confidence intervals exist, for instance those based on the local approximations of the likelihood function. In this thesis, I use *profile likelihoods* [Murphy and van der Vaart, 2000, Raue et al., 2009] to for the evaluation of confidence intervals. Given the likelihood function $\mathcal{L}_{\mathcal{D}}(\theta)$, the profile likelihood of parameter θ_i is

$$\mathrm{PL}(\theta_i) = \max_{\theta_{j\neq i}} \mathcal{L}_{\mathcal{D}}(\theta). \tag{2.67}$$

For a given value of θ_i, the profile likelihood $\mathrm{PL}(\theta_i)$ yields the maximal likelihood by optimising the likelihood function with respect to all other parameters. The profile likelihood is used to calculate the likelihood ratio $R_i = \mathrm{PL}(\theta_i)/\mathcal{L}_{\mathcal{D}}(\theta^{\mathrm{MLE}})$. The likelihood ratio R_i equals one at the MLE θ_i^{MLE} and, if the parameter θ_i is identifiable, approaching zero for large $|\theta_i - \theta_i^{\mathrm{MLE}}|$. Therefore, R_i is an indicator of the uncertainty of the parameter θ_i Using profile likelihood $\mathrm{PL}(\theta_i)$, the confidence interval of parameter θ_i is evaluated as

$$\mathrm{CI}_i^{\alpha} = \left\{ \theta_i \,\middle|\, \frac{\mathrm{PL}(\theta_i)}{\mathcal{L}_{\mathcal{D}}(\theta^{\mathrm{MLE}})} > \exp\left(-\frac{\Delta_\alpha}{2}\right) \right\}, \tag{2.68}$$

where $\Delta\alpha$ is the α-th percentile of the χ^2-squared distribution with n_θ degrees of freedom,

$$\int_0^{\Delta\alpha} \chi^2(\xi|n_\theta)d\xi = \alpha. \tag{2.69}$$

This confidence interval is linked to a rejection test based on the likelihood-ratio in this way: Assume that a hypothesis testing procedure is given which, for any parameter θ_0, tests the null hypothesis $\theta = \theta_0$ against the alternative hypothesis $\theta \neq \theta_0$. The confidence interval CI_i^α includes all parameters θ_0 for which the null hypothesis is not rejected at significance level α. For further details, we refer to [Raue et al., 2009] and [Meeker and Escobar, 1995].

Chapter 3

Summary of Contributed Articles

In this chapter, I provide detailed summaries of the four articles which constitute this publication-based dissertation. I am the sole first author of all of these articles and was in charge of their preparation. A detailed description of my contributions for each publication is provided below. All articles are peer-reviewed, published in international well-established journals or in proceedings of established scientific conferences, and are not used in any other publication-based dissertation. The articles are sorted in chronological order. The full text of these articles will follow in Appendices A-D.

1. A. Kazeroonian, J. Hasenauer, and F. J. Theis. **Parameter Estimation for Stochastic Biochemical Processes: A Comparison of Moment Equation and Finite State Projection.** *In Proceedings of* 10^{th} *International Workshop on Computational Systems Biology (WCSB), Tampere, Finland, pages 66-73, 2013.*

   The inference of the parameters of biochemical reaction networks, e.g., binding affinities and degradation rates, from experimental data is a highly relevant problem in systems and mathematical biology (see Section 2.3). To achieve accurate predictive models, high quality parameter estimates with low uncertainties are essential. Uncertainty of estimates for the parameter values can be reduced by (i) incorporating more experimental data in the parameter estimation process or (ii) improving the extraction of information from the available data. The latter is often preferable as it does not require additional biological experiments, which are often time-consuming and expensive.

   The experimental data for studying heterogeneous cell populations is often collected via techniques, such as flow and mass cytometry, that provide single cell data. The

distribution of single cell measurements was shown to be a source of substantial information about the parameters of a biochemical network [Munsky et al., 2009]. In particular, it was established that even only the use of variance information, in addition to the mean, substantially improves the parameter estimation compared to using the mean alone [Munsky et al., 2009]. Building on these findings, I carried out a novel study which investigated how much information about parameter values was encoded in the higher-order moments, and in the full distribution of single cell measurements.

For this task, I considered mesoscopic modelling, and in particular moment closure approximations (MA), to be a proper tool to explain the higher-order moments of distributions of single cell data. It was explained in the Methods chapter (Section 2.2.2) that the MA describes statistical moments with affordable computational complexity–unlike microscopic modelling approaches, such as the finite state projection (FSP) (see Section 2.2.1).

To assess the information about the parameters encoded in statistical moments of different orders, I derived MAs of various orders–in most previous studies only first- and second-order moments were employed–for a gene expression model with affine propensities. This biochemical reaction network possesses only first-order reactions (see Section 2.1.1). In such a reaction network, the moment equations are exact as no moment closure is needed (see Section 2.2.2). Therefore, I could examine the effect of additional higher-order moments without the interference of approximation errors introduced by moment closure. I used simulated population snapshot data to fit the MAs and estimate the parameters of the reaction network. For this purpose, I devised an objective function assuming independent, normally distributed measurement noise for the statistical moments (see Section 2.3.2). I estimated the error variance for each moment with a bootstrap approach. In addition, I performed parameter estimation using the FSP which uses the full distribution of the data (see Section 2.3.3), and compared the resulting parameter uncertainties with those obtained by MA-based estimation.

In this study, I showed that the parameter uncertainty systematically decreases with the incorporation of higher-order moments. For the considered application, the most significant improvement of parameter estimation accuracy was observed by including the second-order moments: If merely the first-order MA was employed, all parameters of the model were practically non-identifiable. In this case, although the model simulation could be perfectly fit to the mean of the observed data, the higher-order moments could not be correctly predicted by using this set of fitted parameters. As soon as at least second-order moments were used, all parameters

became identifiable. In addition to the assessment of the estimation accuracy, I outlined how higher-order moments can be used for validating parameter estimates. As MA can be written for any arbitrary order of moments, higher-order moments of the same single cell data that are not used for parameter estimation, can be employed for model validation. Thus, in contrast to standard validation approaches, this MA-based approach does not require independent experimental data for model validation. This is a powerful concept that is transferable to a broad spectrum of applications.

In addition to the scientific contributions, I was the author in charge of the preparation of this publication. I wrote the first complete draft of the paper, and iterated it with Jan Hasenauer and Fabian Theis.

2. A. Kazeroonian, F. J. Theis, and J. Hasenauer. **Modeling of stochastic biological processes with non-polynomial propensities using non-central conditional moment equation.** *IFAC Proceedings Volumes, Volume 47, Issue 3, 2014, Pages 1729-1735.*

   As outlined in previous chapters, stochastic biochemical kinetics can be efficiently modelled by moment closure approximations and conditional moment equations (see Section 2.2.2 and 1.2.1). This is especially important for biochemical reaction networks for which simulating the CME or running the SSA is computationally infeasible (see Section 2.1.3 and 2.1.4). For many systems with bimolecular reactions, moment equations and conditional moment equations are not closed and moment closure is necessary (see Section 2.2.2). The derivation of the moment equations and the application of moment closure depend on the reaction kinetics (see Section 2.1.1). If the reaction propensities follow the law of mass-action, a finite number (one or two) of higher-order moments appear in the moment equations of any order, which need to be approximated by means of a moment closure technique. However, many biological processes in practice are modelled using non-mass action kinetics. Frequently used kinetics in this regard, such as Michaelis-Menten kinetics, Hill kinetics, and substrate inhibition kinetics, result in non-polynomial reaction propensities. This means that infinitely many higher-order moments need to be taken into account for deriving the equations for the temporal evolution of the moments. As such a treatment is not feasible, non-mass-action kinetics are not directly interpretable in the framework of moment equations and conditional moment equations. Milner et al. [2011] proposed an approach for moment closure approximations in the case of rational propensity functions; however, a general approach for the treatment of arbitrary (non-polynomial) propensity functions was missing.

To enable moment closure approximations for arbitrary biochemical kinetics, I proposed a novel systematic approximation by means of Taylor series expansion. By truncating the Taylor series expansion, I approximated a non-polynomial propensity function with a finite number of terms that are readily interpretable in the derivation of (conditional) moment equations. The approximation error introduced by this expansion can be controlled via the truncation order. Following the approximation by Taylor series expansion, the resulting higher-order moments need to be approximated by means of moment closure; this introduces another layer of approximation. I analysed several closure schemes and analytically showed that the low-dispersion closure is consistent with the assumptions made by the truncation of the Taylor series expansion. The overall approximation accuracy of the resulting (conditional) moment equations is determined by the truncation order of the Taylor series, and the order of the moment closure. Therefore, I next investigated the interplay of these two sources of approximation error to achieve the best approximation accuracy. Since analytical bounds for the corresponding errors are not available, I opted for a simulation study describing the dynamics of a biochemical reaction network with Michaelis-Menten kinetics including both low-copy and high-copy number species. Such a system is well described by a hybrid stochastic-deterministic approach, and therefore, I employed the method of conditional moments (see Section 1.2.1) for this simulation study. This was the first study where non-mass action kinetics were considered in the framework of conditional moment equations. I assessed the resulting accuracy by comparing the simulation results to those obtained by using the finite state projection (taken as the "ground truth") (see Section 2.2.1). Simulating various orders of conditional moment equations, with varying truncation orders of Taylor series expansion, I showed that the choice of these two orders can be tuned to reduce the overall approximation error.

In addition to the scientific contributions, I was the author in charge of the preparation of this publication. I wrote the first complete draft of the paper, and iterated it with Jan Hasenauer and Fabian Theis.

3. A. Kazeroonian, F. Fröhlich, A. Raue, F. J. Theis, J. Hasenauer. **CERENA: ChEmical REaction Network Analyzer – A Toolbox for the Simulation and Analysis of Stochastic Chemical Kinetics.** *PLOS ONE 11(1): e0146732, 2016.*

The realm of modelling approaches for stochastic biochemical kinetics is broad with freedom in choosing the resolution. A collection of microscopic, mesoscopic and macroscopic approaches were introduced in Section 2.2. It was noted that depend-

ing on the level of description (e.g., microscopic vs. macroscopic), the associated computational complexity of simulating the biochemical reaction network varies. In addition, the approximation accuracy of the mentioned modelling approaches, even among those in the same class, e.g., mesoscopic descriptions, differs. Many of the modelling approaches have an additional degree of freedom in choosing model-specific options–e.g., the choice of moment closure order and technique for MA and MCM–which can substantially influence their performance. To add one more layer of complexity, the relative performance of different modelling approaches depends on the properties of the biological system at hand, rendering the choice of optimal modelling approach problem-specific. As several of these modelling approaches lack a priori estimates on error bounds, their performance for a given biochemical network is not predictable. Due to these issues, in general, no automatic solution for a problem-specific choice of optimal modelling approach can be obtained, and instead, simulation and comparative studies are required. Given that public implementations of many of these approaches were not available, we realised that there was a need for a software toolbox that allows for comprehensive comparison of a broad range of modelling approaches with distinct properties.

To this end, I decided to develop a novel, unifying framework which offers efficient derivation and simulation of a wide range of modelling approaches with a user-friendly interface. I led the development of CERENA, a unique open-source MATLAB toolbox for the modelling and simulation of biochemical reaction networks. CERENA encompasses the collection of the modelling approaches introduced in Section 2.2. In particular, I developed efficient implementations for deriving the moment and the conditional moment equations, and extended the implemented methods to handle general kinetics, i.e. non-mass action kinetics and time-dependent propensities. I implemented a unifying interface for all the included modelling approaches, and enabled proper interpretation of chemical reaction networks in the framework of each approach. I also enabled the import of reaction networks described in the Systems Biology Markup Language (SBML) format. I linked CERENA to efficient numerical solvers, via AMICI–a tool developed by Fabian Fröhlich in another project [Frohlich et al., 2017]. This ensures fast and efficient simulation of a given biochemical reaction network using various methods. Through this linking, CERENA also offers forward and adjoint sensitivity analyses to facilitate further studies such as parameter estimation and uncertainty analysis. Finally, I implemented automatic evaluation and visualisation of the statistical moments of interest based on the solution of various approaches.

To showcase the advantages of such a software toolbox, both in guiding the optimal

choice of modelling approach, as well as the efficiency and feasibility of simulating stochastic kinetics, I conducted comparative studies on two different biochemical networks: a model of gene expression incorporating a feedback loop and a model of JAK/STAT signalling pathway. Comparisons with the SSA and FSP results (treated as the "ground truth") revealed that one could find a method with desirable approximation accuracy from the pool of possible methods. Finally, I showed the significantly reduced computational cost of the simulations enabled in this toolbox compared to available alternatives. Overall, in this work, I could show that the unique collection of modelling approaches in CERENA, together with efficient numerical simulations, enables smooth selection of appropriate problem-specific modelling approaches. CERENA is freely available from
`http://cerenadevelopers.github.io/CERENA/`.

In addition to the scientific contributions, I was the author in charge of the preparation of this publication. I planned the project with the other authors. I led the implementation of the MATLAB toolbox CERENA, which provides a link to the software tool AMICI (previously CVODEwrap) developed by Fabian Fröhlich [Frohlich et al., 2017]. Finally, I wrote the first complete draft of the manuscript and iterated the manuscript with the other authors, in particular Jan Hasenauer.

4. A. Kazeroonian, F. J. Theis, J. Hasenauer. **A scalable moment-closure approximation for large-scale biochemical reaction networks.** *Bioinformatics 2017; 33 (14): i293-i300.*

As described in details in Sections 2.2.2 and 2.2.3, mesoscopic models are powerful tools for simulating stochastic chemical kinetics, as they provide a tradeoff between resolution and computational cost. On the one hand, they provide information about stochasticity and heterogeneity by modelling (a few) statistical moments of the state of a stochastic process. On the other hand, since they consist of a system of ODEs for the statistical moments of a few orders, they can be simulated efficiently using well-established efficient numerical solvers. In particular, the simulation of mesoscopic models is usually far more efficient than microscopic models such as SSA (see Section 2.1.4). However, when it comes to the scaling of the number of state variables of the mesoscopic models with the number of species in a biochemical reaction network, it is immediately noticed that mesoscopic description of realistic biological systems could be challenging. For instance, the number of the state variables of the second-order moment closure approximation scales quadratically with the number of species in a reaction network. Thus, as the number of species increases, even the second-order MA becomes infeasible. Given that there are many relevant biochemical

pathways, e.g., signalling or metabolic pathways, that consist of hundreds or thousands of species, I noticed an important shortcoming of the available approaches for modelling of realistic stochastic biochemical networks.

To address this issue, I explored the possibility of a reliable reduction scheme for mesoscopic models, that would enable mesoscopic description of large-scale reaction networks. Focusing on the second-order moments, I inspected the structure of the covariance matrix in a few signalling pathways, and observed that the covariances tend to zero as the network distance between species increases. Motivated by the observation of strong local dependencies in biological networks, while long-range correlations are mostly not present, I had the idea to exploit the structure of biochemical reaction networks to establish a novel reduction scheme.

To this end, I defined a dependency matrix that encodes the directed connectivities in a reaction network. This dependency matrix can be derived from the topology of the network, i.e. the stoichiometry matrix and the reaction propensities. Using the dependency matrix, I proposed an automatic scheme for identifying the most relevant covariances that correspond to the strongest (direct) dependencies in the network. In this way, I defined the novel scalable moment closure approximation (sMA) that only describes the selected covariances, while the remaining covariances are set to zero or approximated by moment closure.

To assess the size reduction gained by the sMA, I analytically calculated the size of the sMA for a series of network motifs commonly found in biology. In addition, I considered general and scale-free pathway topologies, and calculated a novel approximation for the scaling of the expected size of the sMA. These analytical results showed almost linear scaling of the size of sMA with the number of species in the network. In order to validate and test the performance of sMA on realistic biological networks, I inspected the network topology of a large number of published pathways of varying sizes using sMA, MA and RRE (see Section 2.2.4). The results revealed that also in practice the sMA yields a semi-linear scaling, similar to the RRE, but with the added advantage that it captures heterogeneity by modelling a subset of second-order moments. Finally, I examined the approximation quality of sMA by simulating two signalling pathways with curated parameter values. I showed that the prediction of the sMA is mostly in agreement with that of the full MA. For models with non-mass action kinetics (see Section 2.1.1), I extended the proposed reduction scheme to higher degrees, and showed that the approximation error could be tuned by changing the degree of the reduction. I implemented the scalable moment closure approximation in the toolbox CERENA that I developed as part of my

PhD work [Kazeroonian et al., 2016] (see the summary of the previous article for more information regarding CERENA).

In addition to the scientific contributions, I was the author in charge of the preparation of this publication. I proposed the ideas for the derivation of the sMA. Finally, I wrote the first complete draft of the manuscript and iterated the manuscript with Jan Hasenauer and Fabian Theis.

Chapter 4

Discussion and Outlook

The ultimate aim of systems biology centres around a holistic understanding of biology, at a global scale, where collective behaviour of biological systems are explained as a result of underlying mechanisms. A part of this holistic view is concerned with the study of how single cells respond to stimuli, process information and make decisions. Processes involved in the single cell dynamics, e.g., gene expression, signal transduction and metabolism, are subject to stochasticity. This *intrinsic noise* of biological processes is not a nuisance and in contrast, can have roles in the functioning of cell populations. Therefore, in the study of single cells, capturing this stochasticity provides more insights into the underlying mechanisms—the information that would be obscured if one merely analysed the population-averaged behaviour.

In my doctoral thesis, my goal was to utilise mathematical models to capture heterogeneity in realistic biochemical processes, and obtain reliable descriptive models to learn about the underlying mechanisms of biological systems. Although microscopic descriptions, such as Stochastic Simulation Algorithms, and the Chemical Master Equation on finite state spaces, can provide detailed information about stochastic systems, their high computational complexities make them infeasible for realistic processes. Mesoscopic descriptions capture the heterogeneity in cell populations by modelling a few statistical moments of the probability distribution over the state space of the stochastic process, and thereby drastically reduce the computational cost. However, due to complexity and diversity of biochemical processes (in terms of their kinetics, sizes, etc.), there are ubiquitous biological scenarios that pose challenges to the applicability of standard mesoscopic approaches.

Firstly, size scalability of standard mesoscopic approaches is prohibitive for large-scale reaction networks, such as metabolic pathways that may have hundreds to thousands of

species. To address this issue, in this thesis, I proposed a model reduction exploiting the topological structure of the reaction network. Our conducted simulation studies on several published pathways showed promising results for drastically reducing the computational cost while maintaining satisfactory approximation accuracy. Secondly, nonlinear kinetics, or copy-number regime of the species in the reaction network calls for special treatments to achieve reliable approximations. In my thesis, I extended the moment-closure approximations for the handling of non-mass action kinetics of general form. I also contributed to the development of the method of conditional moments for the handling of copy-number scale separation. These extensions showed superior performances over standard treatments in simulation studies.

Furthermore, as error bounds for mesoscopic approximative methods are not known a priori, the optimal choice for modelling approach, order of approximation and moment closure techniques is only accessible through simulation and comparative studies. Motivated by this need, in this thesis, we developed a comprehensive simulation platform that enables efficient simulation and performance comparisons across multiple modelling approaches. This unifying framework, also facilitates the integration of mechanistic models in the systems biology framework for inference/parameter estimation and model selection. To this end, in this thesis work, we conducted studies for utilising moment-closure approximations (and system size expansions) to provide more insight into the inference of biological systems. Our results indicated the added benefit of higher-order moments in reducing the uncertainty of parameter estimates and increasing the predictive power of mechanistic models.

This thesis work resulted in a feasible framework for reliable descriptions of realistic stochastic biochemical processes that can be used for understanding the behaviour of biological systems. This work can be extended further to enable the analysis of stochastic biochemical kinetics in a wider range of applications. Several possible extensions in this regard are discussed in the following. These ideas describe further improvements that would allow for more efficient use of mesoscopic approaches to answer fundamental questions of systems biology.

4.1 Outlook 1: Potential advantages of mesoscopic approaches in multi-scale modelling

Biological systems are comprised of mechanisms that operate across a broad range of spatial and temporal scales. The spatial scales include molecular, cellular, tissue, organ,

organism and population levels, while the temporal scales range from microseconds to years [Dada and Mendes, 2011]. The intra- and inter-scale interactions of these mechanisms give rise to complex behaviours of biological systems, such as their growth and development. Therefore, understanding of complex biological functions can only be achieved by models that integrate the mechanisms on various spatial and temporal scales. In the recent years, the advent of high-throughput experimental technologies, together with powerful computational techniques, smoothed the way for developing multi-scale modelling and inference for the above-mentioned purpose [Martins et al., 2010, Dada and Mendes, 2011, Walpole et al., 2013].

The aim of multi-scale modelling is to explain a macroscopic behaviour by combining models that describe individual scales/aspects of the system. Multi-scale models are usually derived by putting together models from various model classes such as Agent-based models, Ordinary/Partial Differential Equations, Boolean models, etc. The inclusion and coupling of several modelling classes make the simulation of multi-scale models computationally complex, and call for efficient computational techniques. For instance, the heterogeneous multi-scale method (HMM) [E and Engquist, 2003] is an efficient computational framework which couples macroscopic and microscopic descriptions. To achieve an efficient coupling, the HMM uses microscopic descriptions to provide the necessary information for macroscopic descriptions in regions where the macroscopic model is not valid or explicitly defined. Apart from the simulation, inference of multi-scale models is usually challenging as usually a large number of simulations are required due to the stochasticity of involved processes. Furthermore, to obtain reliable parameter estimates, reproducibility of the results has to be ensured [Hasenauer et al., 2015]. In this regard, the mesoscopic approaches considered in this thesis, namely the SSE, MA, MCM and sMA, can be employed in multi-scale inference problems to enable feasible capturing of stochasticity and circumvent the need for repeated stochastic simulations. In addition, as these mesoscopic methods belong to the ODE model class, efficient deterministic optimisers for ODEs can be used for which the reproducibility can generally be easily achieved.

4.2 Outlook 2: Incorporation of deterministic variability

In this thesis, I mainly focused on capturing the intrinsic stochasticity in biological processes. However, heterogeneity in cell populations can also arise from deterministic variability among individual cells. For instance, cellular organisms may form subpopulations with distinct phenotypic properties. The distinct subpopulations can respond differently to external stimuli, and as a result, can increase the robustness of cell populations in

fluctuating environments. The heterogeneity has been shown to have functional roles in the emergence of complex behaviours in cellular mechanisms [Eldar and Elowitz, 2010]. Therefore, both intrinsic stochasticity and extrinsic variability, e.g., distinct properties of cellular subpopulations, need to be taken into account to obtain a holistic understanding of biological systems.

Heterogeneity in the presence of subpopulations is commonly analysed by means of mixture models where a cell population is modelled as the weighted sum of underlying subpopulations. Recently, Hasenauer et al. [2014a] introduced the ODE-constrained mixture-modelling framework (ODE-MM) which improves upon the standard mixture-modelling by exploiting the predictive power of reaction rate equations (RRE) to describe the underlying reaction networks. Using the RRE, the average dynamics of individual subpopulations are mechanistically modelled while further properties, e.g., the variance of the mixture components corresponding to individual subpopulations, are modelled as unknown parameters. The ODE-MM framework can unravel the underlying subpopulation structure in experimental data, and also provide insight into the sources of variability [Hasenauer et al., 2014a, Loos et al., 2016]. A natural extension of the ODE-MM can be achieved by utilising mesoscopic approaches, such as the MA and SSE to incorporate more mechanistic knowledge about the subpopulation dynamics. In this way, in addition to the average dynamics, the covariance structure of individual subpopulations are described and used for parameter estimation. The ODE-MM can be further extended by employing the MCM such that the underlying subpopulations are modelled as conditional distributions resulted by the MCM solution. In this case, the weighting of the subpopulations is also mechanistically obtained in terms of the marginal probabilities associated with individual conditional distributions [Hasenauer et al., 2014b].

4.3 Outlook 3: Exploiting autocorrelation information

In this thesis, I considered the application of moment closure approximation for parameter estimation using population snapshot data. This data type provides information about the dynamics of cell populations with the resolution of single cells; however, individual cells are not tracked over time. Time-lapse microscopy data, on the other hand, contains trajectories for individual cells, and therefore, provides more information about the dynamics of single cells and cell populations. Incorporating the temporal correlation information encoded in the trajectories of single cells in time-lapse microscopy can enable a more informative parameter estimation approach with decreased uncertainty. As an example, in pedigree analysis, the incorporation of temporal correlations in trajectories could en-

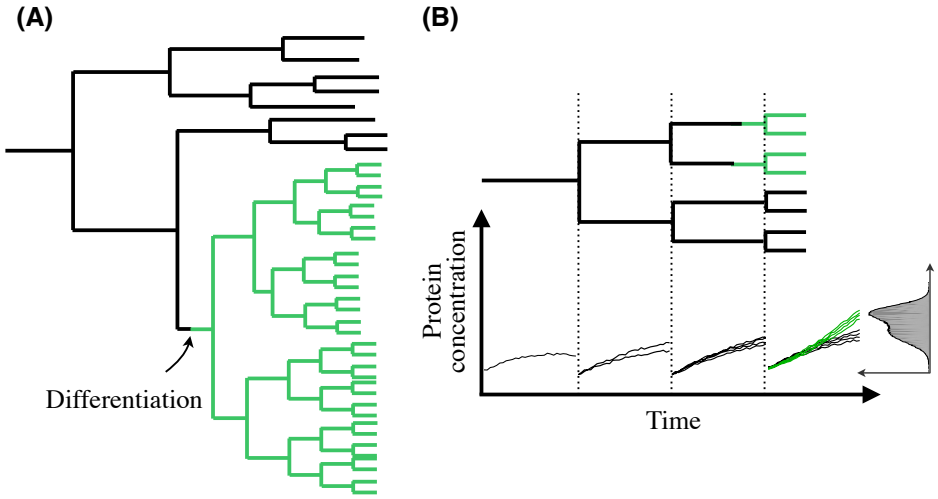


Figure 4.1: **Single cell traces provide temporal correlation information that is not captured in population snapshot data.** (A) A sample single cell derived progeny depicting the impact of differentiation on cell division times. (B) A sample progeny in which a particular protein is up-regulated after differentiation.

able inference about inheritance mechanisms, and detection of distinct cell subsets. This information is not accessible via population snapshot data (Figure 4.1).

In order to use time-lapse data in the framework of moment-based parameter estimation, an appropriate likelihood function needs to be established to utilise the temporal structure of the data. For this purpose, governing equations for the autocorrelation of the state of the stochastic process should be derived. Lestas et al. [2008] consider the autocorrelation equations for linear propensities and at steady state. A theoretical extension of this thesis work can focus on the derivation of the governing equations for autocorrelation in MA and MCM, and corresponding likelihood functions to utilise time-lapse microscopy data.

Bibliography

A. Ale, P. Kirk, and M. P. H. Stumpf. A general moment expansion method for stochastic kinetic models. *J. Chem. Phys.*, 138(17):174101, May 2013. doi: 10.1063/1.4802475.

D. F. Anderson. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *Journal of Chemical Physics*, 127(214107), Dec. 2007. doi: 10.1063/1.2799998.

Anne Auger, Philippe Chatelain, and Petros Koumoutsakos. R-leaping: Accelerating the stochastic simulation algorithm by reaction leaps. *The Journal of Chemical Physics*, 125 (8):084103, 2006. doi: http://dx.doi.org/10.1063/1.2218339. URL `http://scitation.aip.org/content/aip/journal/jcp/125/8/10.1063/1.2218339`.

Yang Cao, Hong Li, and Linda Petzold. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *The Journal of Chemical Physics*, 121(9): 4059–4067, 2004. doi: http://dx.doi.org/10.1063/1.1778376. URL `http://scitation.aip.org/content/aip/journal/jcp/121/9/10.1063/1.1778376`.

Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. The slow-scale stochastic simulation algorithm. *The Journal of Chemical Physics*, 122(1):014116, 2005. doi: http://dx.doi.org/10.1063/1.1824902. URL `http://scitation.aip.org/content/aip/journal/jcp/122/1/10.1063/1.1824902`.

T. F. Coleman and Y. Li. On the convergence of reflective Newton Methods for large-scale nonlinear minimization subject to bounds. *Math. Prog.*, pages 1–36, 1992.

T. F. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.*, 6:418–445, 1996.

J. O. Dada and P. Mendes. Multi-scale modelling and simulation in systems biology. *Integr. Biol.*, 3:86–96, 2011. doi: 10.1039/c0ib00075b.

Weinan E and B. Engquist. The heterogenous multi-scale methods. *Commun. Math. Sci.*, 1:87–133, 2003.

A. Eldar and M. B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(9): 1–7, Sept. 2010. doi: 10.1038/nature09326.

M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, Aug. 2002. doi: 10.1126/science.1070919.

S. Engblom. Computing the moments of high dimensional solutions of the master equation. *Appl. Math. Comp.*, 180:498–515, 2006. doi: 10.1016/j.amc.2005.12.032.

C. P. Fall, E. S. Marland, J. M. Wagner, and J. J. Tyson, editors. *Computational cell biology*, volume 20 of *Interdisciplinary Applied Mathematics*. Springer, Berlin / Heidelberg, 2010. URL `http://www.computationalcellbiology.net/Default.htm`.

W. Feller. On the integro-differential equation of purely discontinous Markoff processes. *Trans. of the American Mathematical Society*, 48:4885–4915, 1940.

F. Fröhlich, P. Thomas, A. Kazeroonian, F. J. Theis, R. Grima, and J. Hasenauer. Inference for stochastic chemical kinetics using moment equations and system size expansion. *PLoS Comput. Biol.*, 12(7):e1005030, July 2016. doi: 10.1371/journal.pcbi.1005030.

Fabian Frohlich, Fabian J Theis, Joachim O Radler, and Jan Hasenauer. Parameter estimation for dynamical systems with discrete events and logical operations. *Bioinformatics*, 33(7):1049–1056, Apr 2017. ISSN 1367-4811 (Electronic); 1367-4803 (Linking). doi: 10.1093/bioinformatics/btw764.

Michael A. Gibson and Jehoshua Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The Journal of Physical Chemistry A*, 104(9):1876–1889, 2000. doi: 10.1021/jp993732q. URL `http://dx.doi.org/10.1021/jp993732q`.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361, Dec. 1977. doi: 10.1021/j100540a008.

D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188 (1):404–425, Sept 1992a. doi: 10.1016/0378-4371(92)90283-V.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reaction systems. *J. Chem. Phys.*, 115:1716–1733, 2001. doi: 10.1063/1.1378322.

D.T. Gillespie. *Markov Processes: An Introduction for Physical Scientists*. Academic Press, 1992b. ISBN 9780122839559. URL `https://books.google.de/books?id=blinSGpXHtEC`.

R. Grima. An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions. *J. Chem. Phys.*, 133(035101), July 2010. doi: 10.1063/1.3454685.

R. Grima. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *J. Chem. Phys.*, 136(15):154105, 2012. doi: 10.1063/1.3702848.

Ramon Grima. Linear-noise approximation and the chemical master equation agree up to second-order moments for a class of chemical systems. *Phys. Rev. E*, 92:042124, Oct 2015. doi: 10.1103/PhysRevE.92.042124. URL `http://link.aps.org/doi/10.1103/PhysRevE.92.042124`.

E. L. Haseltine and J. B. Rawlings. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *J. Chem. Phys.*, 117(15):6959–6969, Oct. 2002. doi: 10.1063/1.1505860.

J. Hasenauer, N. Radde, M. Doszczak, P. Scheurich, and F. Allgöwer. Parameter estimation for the CME from noisy binned snapshot data: Formulation as maximum likelihood problem. Extended abstract at *Conf. of Stoch. Syst. Biol.*, Monte Verita, Switzerland, July 2011.

J. Hasenauer, C. Hasenauer, T. Hucho, and F. J. Theis. ODE constrained mixture modelling: A method for unraveling subpopulation structures and dynamics. *PLoS Comput. Biol.*, 10(7):e1003686, July 2014a. doi: 10.1371/journal.pcbi.1003686.

J. Hasenauer, V. Wolf, A. Kazeroonian, and F. J. Theis. Method of conditional moments (MCM) for the chemical master equation. *J. Math. Biol.*, 69(3):687–735, Sept. 2014b. doi: 10.1007/s00285-013-0711-5.

J. Hasenauer, N. Jagiella, S. Hross, and F. J. Theis. Data-driven modelling of biological multi-scale processes. *Journal of Coupled Systems and Multiscale Dynamics*, 3(2):101–121, Sept. 2015. doi: 10.1166/jcsmd.2015.1069.

A. Hellander and P. Lötstedt. Hybrid method for the Chemical Master Equation. *J. Comput. Phys.*, 227:100–122, 2007. doi: 10.1016/j.jcp.2007.07.020.

J. Hespanha. Moment closure for biochemical networks. In *Proc. Int. Symp. on Communications, Control and Signal Processing*, pages 42–147, 2008. doi: 10.1109/ISCCSP.2008.4537208.

J. P. Hespanha. Modeling and analysis of stochastic hybrid systems. *IEE Proc. Control Theory & Applications,* Special Issue on Hybrid Systems, 153(5):520–535, 2007. doi: 10.1049/ip-cta:20050088.

J. P. Hespanha and A. Singh. Stochastic models for chemically reacting systems using polynomial stochastic hybrid systems. *International Journal of Robust and Nonlinear Control*, 15(15):669–689, 2005. ISSN 1099-1239. doi: 10.1002/rnc.1017. URL `http://dx.doi.org/10.1002/rnc.1017`.

A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward. SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers. *ACM T. Math. Software*, 31(3):363–396, Sept. 2005.

A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117(4):500–544, Aug. 1952.

T. Jahnke. On reduced models for the chemical master equation. *Multiscale Model. Simul.*, 9(4):1646–1676, 2011.

T. Jahnke and W. Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *J. Math. Biol.*, 54(1):1–26, Jan. 2007. doi: 10.1007/s00285-006-0034-x.

Thomas R. Kiffe James H. Matis. On approximating the moments of the equilibrium distribution of a stochastic logistic model. *Biometrics*, 52(3):980–991, 1996. ISSN 0006341X, 15410420. URL `http://www.jstor.org/stable/2533059`.

V. Kazeev, M. Khammash, M. Nip, and C. Schwab. Direct solution of the Chemical Master Equation using quantized tensor trains. *PLoS Comput. Biol.*, 10(3):e1003359, Mar. 2014. doi: 10.1371/journal.pcbi.1003359.

A. Kazeroonian, J. Hasenauer, and F. J. Theis. Parameter estimation for stochastic biochemical processes: A comparison of moment equation and finite state projection. In R. Autio, I. Shmulevich, K. Strimmer, C. Wiuf, S. Sarbu, and O. Yli-Harja, editors, *Proceedings of 10th International Workshop on Computational Systems Biology*, pages 66–73, Tampere, Finland, June 2013. Tampere International Center for Signal Processing.

A. Kazeroonian, F. J. Theis, and J. Hasenauer. Modeling of stochastic biological processes with non-polynomial propensities using non-central conditional moment equation. In *Proc. of the 19th IFAC World Congress*, volume 19, pages 1729–1735, Cape Town, South Africa, August 2014.

A. Kazeroonian, F. Fröhlich, A. Raue, F. J. Theis, and J. Hasenauer. CERENA: ChEmical REaction Network Analyzer–A toolbox for the simulation and analysis of stochastic

chemical kinetics. *PLoS ONE*, 11(1):e0146732, Jan. 2016. doi: 10.1371/journal.pone.0146732.

Atefeh Kazeroonian, Fabian J. Theis, and Jan Hasenauer. A scalable moment-closure approximation for large-scale biochemical reaction networks. *Bioinformatics*, 33(14): i293–i300, 2017. doi: 10.1093/bioinformatics/btx249. URL `+http://dx.doi.org/10.1093/bioinformatics/btx249`.

M J Keeling. Multiplicative moments and measures of persistence in ecology. *J Theor Biol*, 205(2):269–281, Jul 2000. ISSN 0022-5193 (Print); 0022-5193 (Linking). doi: 10.1006/jtbi.2000.2066.

H. Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, Mar. 2002.

E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems biology in practice*. Wiley-VCH, Weinheim, 2005. ISBN 978-3-527-31078-4.

C. H. Lee, K. H. Kim, and P. Kim. A moment closure method for stochastic reaction networks. *J. Chem. Phys.*, 130(13):134107, Apr. 2009. doi: http://dx.doi.org/10.1063/1.3103264.

I. Lestas, J. Paulsson, N. E. Ross, and G. Vinnicombe. Noise in gene regulatory networks. *IEEE Trans. Autom. Control*, 53:189–200, Jan. 2008.

Carolin Loos, Anna Fiedler, and Jan Hasenauer. *Parameter Estimation for Reaction Rate Equation Constrained Mixture Models*, pages 186–200. Springer International Publishing, Cham, 2016. ISBN 978-3-319-45177-0. doi: 10.1007/978-3-319-45177-0_12. URL `http://dx.doi.org/10.1007/978-3-319-45177-0_12`.

M. L. Martins, S. C. Ferreira Jr., and M. J. Vilela. Multiscale models for biological systems. *Curr. Opin. Colloid Interface Sci.*, 15(1–2):18–23, Apr. 2010. doi: 10.1016/j.cocis.2009.04.004.

M. Mateescu, V. Wolf, F. Didier, and T.A. Henzinger. Fast adaptive uniformisation of the chemical master equation. *IET. Syst. Biol.*, 4(6):441–452, 2010.

H. J. Matis and T. R. Kiffe. Effects of immigration on some stochastic logistic models: a cumulant truncation analysis. *Theor. Popul. Biol.*, 56(2):139–161, Oct. 1999.

A. D. McNaught and A. Wilkinson. *IUPAC Compendium of chemical terminology*. Blackwell Science, 2nd edition, 1997. doi: 10.1351/gooldbook.

W. Q. Meeker and L. A. Escobar. Teaching about approximate confidence regions based on maximum likelihood estimation. *Am. Stat.*, 49(1):48–53, Feb 1995.

S. Menz, J. C. Latorre, C. Schütte, and W. Huisinga. Hybrid stochastic deterministic solution of the Chemical Master Equation. *SIAM J. on Multiscale Model. Simul.*, 10 (4):1232–1262, Oct. 2012. doi: 10.1137/110825716.

L. Michaelis and M. Menten. Die Kinetik der Invertinwirkung. *Biochem. Z.*, 49:333–369, 1913.

P. Milner, C. S. Gillespie, and D. J. Wilkinson. Moment closure approximations for stochastic kinetic models with rational rate laws. *Math. Biosci.*, 231(2):99–104, June 2011. doi: 10.1016/j.mbs.2011.02.006.

P. Milner, C. S. Gillespie, and D. J. Wilkinson. Moment closure based parameter inference of stochastic kinetic models. *Stat. Comp.*, 23(2):287–295, Mar. 2013. doi: 10.1007/s11222-011-9310-8.

B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.*, 124(4):044104, Jan. 2006. doi: 10.1063/1.2145882.

B. Munsky, B. Trinh, and M. Khammash. Listening to the noise: random fluctuations reveal gene network parameters. *Mol. Syst. Biol.*, 5(318), Oct. 2009. doi: 10.1038/msb.2009.75.

S. A. Murphy and A. W. van der Vaart. On profile likelihood. *J. Am. Stat. Assoc.*, 95 (450):449–485, June 2000.

I. Nåsell. Moment closure and the stochastic logistic model. *Theoretical Population Biology*, 63(2):159–168, 2003.

MEJ Newman and GT Barkema. *Monte Carlo Methods in Statistical Physics chapter 1-4*. Oxford University Press: New York, USA, 1999.

J. R. Norris. Continuous-time markov chains i. In *Markov Chains*, pages 60–107. Cambridge University Press, 1997. ISBN 9780511810633. URL `http://dx.doi.org/10.1017/CBO9780511810633.004`. Cambridge Books Online.

James R. Norris. *Markov chains.* Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 1998. ISBN 978-0-521-48181-6.

T. Nüesch. Finite state projection-based parameter estimation algorithms for stochastic chemical kinetics. Master thesis, Swiss Federal Institute of Technology, Zürich, 2010.

A. Raj and A. van Oudenaarden. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, 135(2):216–226, Oct 2008.

A. Raj, C.S. Peskin, D. Tranchina, D.Y. Vargas, and S. Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.*, 4(10):e309, Oct. 2006.

R. Ramaswamy, N. González-Segredo, and I. F. Sbalzarini. A new class of highly efficient exact Stochastic Simulation Algorithms for chemical reaction networks. *J. Chem. Phys.*, 130(2009):244104, 2009. doi: 10.1063/1.3154624.

R. Ramaswamy, N. González-Segredo, I. F. Sbalzarini, and R. Grima. Discreteness-induced concentration inversion in mesoscopic chemical systems. *Nat. Comm.*, 3(779), Apr. 2012. doi: DOI:10.1038/ncomms1775.

J. M. Raser and E. K. O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–1814, June 2004. doi: 10.1126/science.1098641.

M. Rathinam, L.R. Petzold, Y. Cao, and D. T. Gillespie. Stiffness in stochastic chemically reaction systems: The implicit tau-leaping method. *J. Chem. Phys.*, 119(24):12784–12794, 2003. doi: 101063/1.1627296.

A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinf.*, 25(25):1923–1929, May 2009.

A. Raue, V. Becker, U. Klingmüller, and J. Timmer. Identifiability and observability analysis for experimental design in nonlinear dynamical models. *Chaos*, 20(045105), Dec. 2010. doi: 10.1063/1.3528102.

A. Raue, M. Schilling, J. Bachmann, A. Matteson, M. Schelke, D. Kaschek, S. Hug, C. Kreutz, B. D. Harms, F. J. Theis, U. Klingmüller, and J. Timmer. Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE*, 8(9):e74335, Sept. 2013. doi: 10.1371/journal.pone.0074335.

J. Ruess, A. Milias, S. Summers, and J. Lygeros. Moment estimation for chemically reacting systems by extended Kalman filtering. *J. Chem. Phys.*, 135(165102), Oct. 2011. doi: 10.1063/1.3654135.

Jakob Ruess and John Lygeros. Moment-based methods for parameter inference and experiment design for stochastic biochemical reaction networks. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 25(2):8, 2015.

D. Schnoerr, G. Sanguinetti, and R. Grima. Validity conditions for moment closure approximations in stochastic chemical kinetics. *J. Chem. Phys.*, 141(8):084103, Aug. 2014.

B. Schöberl, C. Eichler-Jonsson, E. D. Gilles, and G. Muller. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotechnol.*, 20:370–375, 2002.

R. Sidje, K. Burrage, and S. MacNamara. Inexact uniformization method for computing transient distributions of Markov chains. *SIAM J. Sci. Comput.*, 29(6):2562–2580, 2007. ISSN 1064-8275.

A. Singh and J. Hespanha. A derivative matching approach to moment closure for the stochastic logistic model. *Bull. Math. Biol.*, 69(6):1909—1925, Aug. 2007.

A. Singh and J. P. Hespanha. Approximate moment dynamics for chemically reacting systems. *IEEE Trans. Autom. Control*, 56(2):414–418, Feb. 2011. doi: 10.1109/TAC.2010.2088631.

P. Thomas, H. Matuschek, and R. Grima. How reliable is the linear noise approximation of gene regulatory networks? *BMC Genomics*, 14(Suppl 4)(S5), Oct. 2013. doi: 10.1186/1471-2164-14-S4-S5.

N. G. van Kampen. *Stochastic processes in physics and chemistry.* North-Holland, Amsterdam, 3rd edition, 2007.

J. Walpole, J. A. Papin, and S. M. Peirce. Multiscale computational models of complex biological systems. *Annu. Rev. Biomed. Eng.*, 15:137–154, 2013.

P. Whittle. On the use of the normal approximation in the treatment of stochastic processes. *J. R. Stat. Soc. B*, 19(2):268–281, 1957.

C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl. Moment-based inference predicts bimodality in transient gene expression. *Proc. Natl. Acad. Sci. U S A*, 109(21):8340–8345, May 2012. doi: 10.1073/pnas.1200161109.

Appendix A

Parameter Estimation for Stochastic Biochemical Processes: A Comparison of Moment Equation and Finite State Projection. *International Workshop on Computational Systems Biology (WCSB), 2013.*

This is a pre-copyedited, author-produced PDF of an article accepted for publication in Proceedings of 10^{th} International Workshop on Computational Systems Biology (WCSB) following peer review. The version of record
A. Kazeroonian, J. Hasenauer, and F. J. Theis. **Parameter Estimation for Stochastic Biochemical Processes: A Comparison of Moment Equation and Finite State Projection.** *In Proceedings of* 10^{th} *International Workshop on Computational Systems Biology (WCSB), Tampere, Finland, pages 66-73, June 2013.*
is available online at:
http://www.cs.tut.fi/wcsb13/WCSB2013_proceedings.pdf

PARAMETER ESTIMATION FOR STOCHASTIC BIOCHEMICAL PROCESSES: A COMPARISON OF MOMENT EQUATION AND FINITE STATE PROJECTION

Atefeh Kazeroonian[1], *Jan Hasenauer*[1,2], *and Fabian Theis*[1,2]

[1]Institute of Computational Biology, Helmholtz Center Munich,
Ingolstädter Landstr. 1, 85764 Neuherberg, Germany

[2]Department of Mathematics, University of Technology Munich,
Boltzmannstr. 3, 85747 Garching, Germany

{atefeh.kazeroonian,jan.hasenauer,fabian.theis}@helmholtz-muenchen.de

ABSTRACT

Many biochemical processes exhibit intrinsic stochastic fluctuations. These intrinsic fluctuations can be modeled using the chemical master equation (CME). The estimation of the parameters of the CME is challenging because the CME is a high or infinite dimensional system.

We compare two approaches currently used to estimate parameters of CMEs from population snapshot data. The first approach relies on a truncation of the CME, the finite state projection, and uses the data directly. The second method relies on moment equations – dynamical systems computing the moments of the CME solution – and merely uses the moments of the data. The second method is computationally more efficient, however, it cannot use all information contained in the data. In this manuscript, we assess the statistical power of the individual approaches and study moment equations of different order. Furthermore, we refine the likelihood function for the moment equation and introduce a novel validation method.

We performed a comparative study of the commonly used 3-stage model of gene expression. Using maximum likelihood estimates and a rigorous uncertainty quantification based on profile likelihoods, we show that the finite state projection approach is statistically more powerful than approaches based on moment equation. Nevertheless, even in case of partial observations, the first and second moments of the CME solution are highly informative and permit parameter identifiability. These findings, in combination with the novel tools for validation and uncertainty analysis, improve the insight into the problem class.

1. INTRODUCTION

In recent years, a multitude of studies have shown that many biochemical processes in prokaryotic and eukaryotic cells exhibit intrinsic stochastic fluctuations [1]. These fluctuations arise from low copy-number effects and are particularly significant for transcription and translation [2]. It is now known that these fluctuations are in many cases required for cellular function, e.g., for robust decision making on the population level [1].

The stochastic dynamics of biological processes can be described using continuous-time discrete-state Markov chains (CTMCs). The statistics of these Markov chains are governed by the chemical master equation (CME). Individual realizations of the process can be obtained via stochastic simulation algorithms (SSAs) [3, 4]. The stochastic process can be studied by analyzing statistics of many such realizations. Alternatively, the CME can be simulated using the finite state projection (FSP) method [5], which relies on truncation of the state space of the CME. While SSAs and the FSP are in principle capable of resolving all details of the dynamics of the CME, they impose a significant computational cost. This computational cost already becomes intractable for many small-scale systems. As an alternative, the method of moments (MM) [6, 7, 8] can be employed to capture the overall statistics of the process, such as mean and variance of individual species as well as covariances.

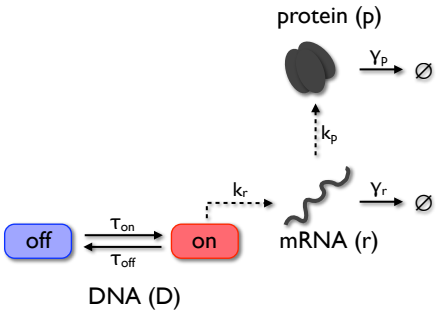
While the SSA, the FSP, and the MM all have advantages and disadvantages, a joint property is that they require accurate parameter values. The models and simulations are only predictive if good estimates of the reaction rates are available. Several estimation methods, relying on different models, were proposed (see, e.g., [9] and references therein), however, in most studies only the optimal parameter estimate has been considered, and the methods have not been compared. In this manuscript, we study the parameter estimates and confidence intervals obtained using FSP and MM. We present the individual likelihood functions and evaluate the informativeness using profile likelihoods. This is done for the widely used 3-stage model of gene expression [2], which is depicted in Figure 1.

2. METHODS

2.1. Modeling and simulation

2.1.1. Chemical master equation

The time evolution of the state $X = (X_1, \ldots, X_{n_s})^{\mathrm{T}} \in \mathbb{N}_0^{n_s}$ of stochastic biochemical reaction networks is mostly described using CTMCs. The statistics of CTMCs are



Moment equation (order 1):

$$\dot{\mu}_{\mathrm{D_{off}}} = \tau_{\mathrm{off}}\mu_{\mathrm{D_{on}}} - \tau_{\mathrm{on}}\mu_{\mathrm{D_{off}}}$$
$$\dot{\mu}_{\mathrm{D_{on}}} = \tau_{\mathrm{on}}\mu_{\mathrm{D_{off}}} - \tau_{\mathrm{off}}\mu_{\mathrm{D_{on}}}$$
$$\dot{\mu}_{\mathrm{r}} = k_{\mathrm{r}}\mu_{\mathrm{D_{on}}} - \gamma_{\mathrm{r}}\mu_{\mathrm{r}}$$
$$\dot{\mu}_{\mathrm{p}} = k_{\mathrm{p}}\mu_{\mathrm{r}} - \gamma_{\mathrm{p}}\mu_{\mathrm{p}}$$

Figure 1. **Three-stage gene expression model.** (left) Schematic of the 3-stage gene expression model shows two DNA states (on, off), mRNAs and proteins. Transitions as well as synthesis and degradation reactions are shown as arrows. (right) Moment equations for means and variances of the individual species. The subscripts indicate the dependency, e.g., μ_r is the mean mRNA number.

governed by the CME. For a process with n_r chemical reactions,

$$R_k : \sum_{i=1}^{n_s} \nu_{ik}^- X_i \to \sum_{i=1}^{n_s} \nu_{ik}^+ X_i,$$

with reaction stoichiometries ν_k^-, ν_k^+, and $\nu_k = \nu_k^+ - \nu_k^-$, and reaction propensities $a_k(X, \theta)$, the CME is

$$\frac{\partial}{\partial t}p(x;t) = \sum_{\substack{k=1 \\ x \geq \nu_k^+}}^{n_r} a_k(x-\nu_k,\theta)p(x-\nu_k;t) - \sum_{k=1}^{n_r} a_k(x,\theta)p(x;t).$$

The solution of the CME depends on the parameters θ, which are for instance reaction rates.

The CME is defined for all reachable states $x \in \Omega \subset \mathbb{N}_0^{n_s}$, where n_s is the number of biochemical species. The set of reachable states Ω is in general very large, or infinite, rendering a direct solution of the full CME infeasible. Fortunately, the set of states with a significant probability mass is often small. This is exploited by the FSP, a direct method for approximating the solution of the CME [5] with pre-specified accuracy. Therefore, a subset Ω^{FSP} of the set of reachable states Ω is chosen. The time evolution of $p(x;t)$ with $x \in \Omega^{\mathrm{FSP}}$ is described by the CME, but influxes from states $x - \nu_k \notin \Omega^{\mathrm{FSP}}$ are removed. Probabilities $p(x;t)$ resulting from the simulation of this truncated system, which can be shown to be a lower bound for the actual probabilities of the CME, converge to the actual probabilities by growing Ω^{FSP} until the pre-specified accuracy is met.

A requirement for the application of the FSP is that the number of states with a significant probability mass is not too large. Novel algorithms can handle some million states [10]. Beyond this, the direct numerical simulation becomes infeasible.

2.1.2. Method of moments

In situations where the FSP is no longer applicable, the method of moments can be employed to approximate the solution of the CME [6]. The MM, also called moment equation, does not reproduce the exact solution of the CME. Instead, it computes the moments of $p(x;t)$, i.e. mean

$$\mu_i(t) = \sum_{x\in\Omega} x_i p(x;t),$$

variance

$$C_{ij}(t) = \sum_{x\in\Omega} (x_i - \mu_i(t))(x_j - \mu_j(t))p(x;t),$$

and higher-order moments [6]. The dynamics of the moments are governed by a set of ordinary differential equations (ODEs). Given that chemical reactions are at most bimolecular, the ODEs for the mean and the variance are

$$\frac{d\mu_i}{dt} = \sum_{k=1}^{n_r} \nu_{ik} \left(a_k(\mu,\theta) + \frac{1}{2} \sum_{l_1,l_2} \frac{\partial^2 a_k(\mu,\theta)}{\partial x_{l_1} \partial x_{l_2}} C_{l_1 l_2} \right),$$

$$\frac{dC_{ij}}{dt} = \sum_{k=1}^{n_r} \left(\nu_{ik} \sum_l \frac{\partial a_k(\mu,\theta)}{\partial x_l} C_{il} + \nu_{jk} \sum_l \frac{\partial a_k(\mu,\theta)}{\partial x_l} C_{jl} + \nu_{ik}\nu_{jk} \left(a_k(\mu,\theta) + \frac{1}{2} \sum_{l_1,l_2} \frac{\partial^2 a_k(\mu,\theta)}{\partial x_{l_1} \partial x_{l_2}} C_{l_1 l_2} \right) \right)$$
$$+ \sum_{k=1}^{n_r} \left(\nu_{ik} \sum_{l_1,l_2} \frac{\partial^2 a_k(\mu,\theta)}{\partial x_{l_1} \partial x_{l_2}} C_{i l_1 l_2} + \nu_{jk} \sum_{l_1,l_2} \frac{\partial^2 a_k(\mu,\theta)}{\partial x_{l_1} \partial x_{l_2}} C_{j l_1 l_2} \right),$$

in which $C_{il_1l_2}$ and $C_{jl_1l_2}$ are third order moments according to notation used in [6]. The governing equation for arbitrary moment orders can be found in [6, Equation (2.46)]. If all reactions are at most mono-molecular, the moment equation is closed, meaning that the evolution of moments of order m does not depend on moments of order greater than m. In this case, the moment equations are exact. If bimolecular chemical reactions are present, the moment equation ODEs are not closed, and the evaluation of a moment of order m requires the moments of order $m+1$ [6]. Moment closure techniques must be employed [11], and the resulting moments will only be an approximation of the true moments of the solution of the CME.

Moment equations are in general low-dimensional compared to the CME. Thus, they can generally be solved more efficiently. However, a drawback is that a finite number of moments does not allow the reconstruction of the underlying distribution $p(x;t)$. Hence, information is lost.

2.2. Parameter estimation

In this work, we considered population snapshot data $\mathcal{D}_k = \left\{ \left(\bar{Y}^{(s)}(t_k), t_k \right) \right\}_{s=1}^{S_k}$, $k = 1, \ldots, N$, obtained by sampling cells $s = 1, \ldots, S_k$ from the cell population and measuring one (or more) properties of these cells, e.g., using flow cytometry or microscopy. For notational simplicity, we assume that one observable, $\bar{Y} = h(X)$, can be measured. The observation function h describes the type of measurement; in the most simple case $h(X) = X_i$. The measurement is assumed to be noise-free as we later want to assess the informativeness of single-cell data vs. the moments.

Given a realization X at a certain time t_k, the probability of observing $\bar{Y}$ at time t_k is $p(y = \bar{Y}; x = X)$. The total probability to observe $\bar{Y}$ at time t_k is obtained by taking into account all possible realizations $X \in \Omega$ of the process. Given that the number of molecules is a discrete variable, this total probability is obtained by marginalizing over the state space Ω,

$$p(y; t_k, \theta) = \sum_{x \in \Omega} p(y; x)\ p(x; t_k, \theta),$$

where $p(x; t_k, \theta)$ is the solution of the CME. Bearing in mind that we do not consider any measurement noise, y is a deterministic function of x, $y = h(x)$, thus

$$p(y|x) = \begin{cases} 1 & \text{if } y = h(x) \\ 0 & \text{otherwise,} \end{cases}$$

so the sum simplifies to

$$p(y; t_k, \theta) = \sum_{\substack{x \in \Omega \\ h(x) = y}} p(x; t_k, \theta).$$

Following the argumentation above, the probability distribution $p(y; t_k, \theta)$ is the distribution from which the observations are drawn. Thus,

$$p(y = \bar{Y}^{(s)}(t_k)) = p(y; t_k, \theta), \quad s = 1, \ldots, S_k.$$

In the following, we compare two classes of likelihood functions for these data, namely an FSP-based likelihood function and a moment-based likelihood function with respect to their statistical power. As mentioned before, we do not consider any measurement noise in this comparison, but the inclusion of noise in the presented procedure would be rather straightforward.

2.2.1. FSP-based estimation

As outlined earlier, for CTMCs with a small effective state space, the FSP can be used to approximate the solution of the CME for a given parameter set θ. Using this approximation of the probability distribution of the hidden state, $p(x; t, \theta)$, and the corresponding approximation of the probability distribution of the observable, $p(y; t, \theta)$, the likelihood of the stochastic process,

$$\mathcal{L}_{\mathcal{D}}^{\text{FSP}}(\theta) = c \prod_{k=1}^{N} \prod_{s=1}^{S_k} p(y = \bar{Y}^{(s)}(t_k); t_k, \theta),$$

can be evaluated. Basically, the probabilities are evaluated and multiplied for all observed states. The constant c depends only on the data and can be neglected for optimization purposes. For a detailed introduction of this FSP-based likelihood function, we refer to [12, 13]. Given the FSP-based likelihood function, the estimation problem can be formulated. The FSP-based maximum likelihood (ML) estimation problem is:

$$\begin{aligned} &\underset{\theta}{\text{maximize}}\ \log \mathcal{L}_{\mathcal{D}}^{\text{FSP}}(\theta) \\ &\text{subject to } \Sigma^{\text{FSP}}(\theta), \end{aligned}$$

in which $\Sigma^{\text{FSP}}(\theta)$ denotes the finite-dimensional ODE model resulting from the FSP of the CME on the subset Ω^{FSP}. To reduce numerical problems, the problem is formulated using the log-likelihood function $\log \mathcal{L}^{\text{FSP}}_{\mathcal{D}}(\theta)$. Furthermore, we optimize the logarithm of the parameters $\xi = \log_{10}(\theta)$ to ensure positivity and improve the performance of the optimization routines. The optimal solution of the FSP-based ML estimation problem is the parameter vector for which the likelihood of observing the single cell data is maximized. This estimator uses all available information.

2.2.2. *Moment-based estimation*

For many processes the approximation of the CME solution using the FSP is not feasible because the number of states with non-negligible probability is too large. In such cases, the moment equation can be employed to approximate the statistics of the CME solution. To employ moment equations for parameter estimation, the statistics of the snapshots are computed, e.g., mean and variance,

$$\bar{\mu}_y(t_k) = \frac{1}{S_k} \sum_{s=1}^{S_k} \bar{Y}^{(s)}(t_k),$$

$$\bar{C}_{yy}(t_k) = \frac{1}{S_k} \sum_{s=1}^{S_k} \left(\bar{Y}^{(s)}(t_k) - \bar{\mu}_y(t_k) \right)^2.$$

These measured moments are compared to moments predicted by the model and the observation function $h(x)$. Since the sample sizes S_k are often quite large – for flow cytometry often in the order of 10^4 – it follows from the central limit theorem that the empirical moments, e.g., $\bar{\mu}_y(t_k)$ and $\bar{C}_{yy}(t_k)$, are almost normally distributed around the true moments [14]. Hence, a normal error model is assumed,

$$\mathcal{L}^{\text{MM}}_{\mathcal{D},\mu_y}(\theta) = \prod_{k=1}^{N} \mathcal{N}\left(\mu_y(t_k,\theta) | \bar{\mu}_y(t_k), \sigma^2_{\bar{\mu}_y}(t_k) \right),$$

$$\mathcal{L}^{\text{MM}}_{\mathcal{D},C_{yy}}(\theta) = \prod_{k=1}^{N} \mathcal{N}\left(C_{yy}(t_k,\theta) | \bar{C}_{yy}(t_k), \sigma^2_{\bar{C}_{yy}}(t_k) \right),$$

where $\mathcal{N}(\cdot|\mu,\sigma^2)$ is the probability density of the normal distribution. Such a likelihood function can be derived for every moment predicted by the model, e.g., also the third and fourth order central moments. Clearly, the consideration of additional, non-redundant moments provides additional information about the model parameters as the individual likelihood functions are multiplied, e.g., if mean and variance are employed then a reasonable likelihood function is

$$\mathcal{L}^{\text{MM}}_{\mathcal{D}}(\theta) = \mathcal{L}^{\text{MM}}_{\mathcal{D},\mu_y}(\theta) \cdot \mathcal{L}^{\text{MM}}_{\mathcal{D},C_{yy}}(\theta).$$

Unfortunately, also the computational complexity of simulating the moment equations increases with each additional moment considered in the model.

The likelihoods $\mathcal{L}^{\text{MM}}_{\mathcal{D},\mu_y}(\theta)$, $\mathcal{L}^{\text{MM}}_{\mathcal{D},C_{yy}}(\theta)$ and those for the higher-order moments require information about the error variance of the respective empirical estimator, e.g., $\sigma^2_{\bar{\mu}_y}$ for $\bar{\mu}_y(t_k)$ and $\sigma^2_{\bar{C}_{yy}}$ for $\bar{C}_{yy}(t_k)$. The variance of the estimators for the first and second order moments can be found in [14]. For third and higher-order moments the calculation of these estimators become increasingly complex, and we did not find respective results in the literature. To circumvent the analytical derivation, we propose to estimate the variance of the empirical estimators using non-parametric bootstrapping [15]. This approach employs a two-step procedure. At first, a sample of size S_k is drawn from $\{\bar{Y}^{(s)}(t_k)\}_{s=1}^{S_k}$ (all $\bar{Y}^{(s)}(t_k)$ have probability $\frac{1}{S_k}$) and the moments of this artificial sample are evaluated. This step is repeated a large number of times, in general more than one thousand times, yielding a large sample for each moment of interest. Therefore, the variance of each moment can easily be computed from the corresponding sample. This sample variance is a reliable measure for the uncertainty, if $S_k \gg 1$. It does not require any distribution assumption for $p(y;t_k,\theta)$ and is easily applicable to any higher-order moments.

Given the likelihood function $\mathcal{L}^{\text{MM}}_{\mathcal{D}}(\theta)$, which is the product of the likelihood functions for the moments of interest, the moment-based ML estimation problem,

$$\begin{aligned} &\underset{\theta \in \mathbb{R}^{n_\theta}_+}{\text{maximize}} \ \log \mathcal{L}^{\text{MM}}_{\mathcal{D}}(\theta) \\ &\text{subject to } \Sigma^{\text{MM}}(\theta), \end{aligned}$$

can be formulated. $\Sigma^{\text{MM}}(\theta)$ is the model used to simulate the moment equations for the moments of interest.

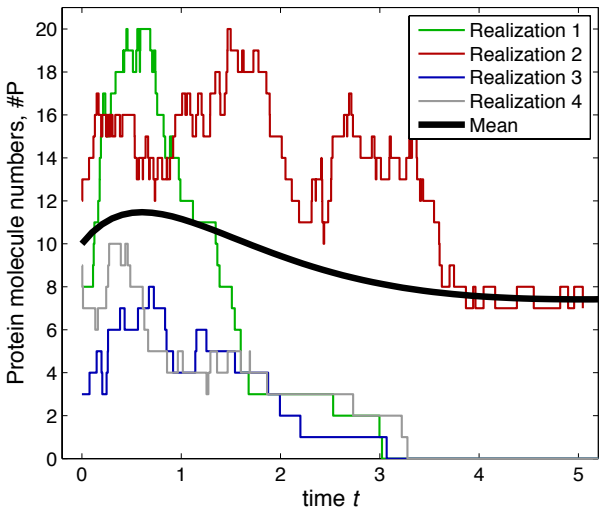
2.2.3. *Identifiability and uncertainty analysis*

As the measurement data are limited and potentially noise corrupted, the parameters can in general not be estimated precisely. To assess the remaining parameter uncertainty and the practical identifiability, we use profile likelihoods [16]. Given the likelihood function $\mathcal{L}_{\mathcal{D}}(\theta)$, the profile likelihood of parameter θ_i is

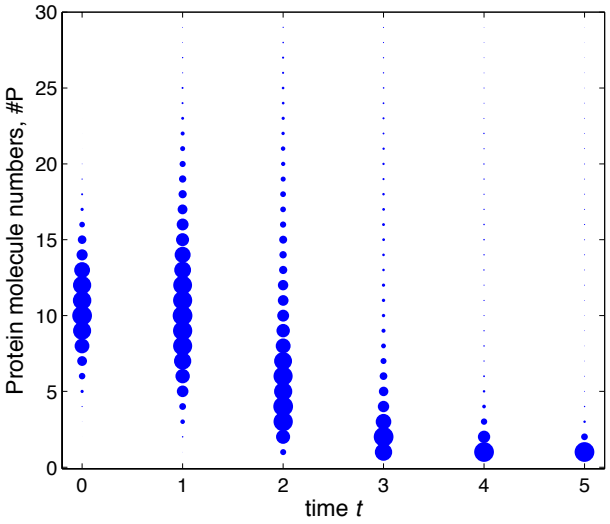
$$\text{PL}(\theta_i) = \max_{\theta_{j \neq i}} \mathcal{L}_{\mathcal{D}}(\theta).$$

This profile likelihood $\text{PL}(\theta_i)$ is the maximal likelihood for a given value of θ_i. Using the profile likelihood, the likelihood ratio $R(\theta_i) = \text{PL}(\theta_i)/\mathcal{L}_{\mathcal{D}}(\hat{\theta})$ can be evaluated, in which $\hat{\theta}$ is the ML estimate. The likelihood ratio R is one at the globally optimal point $\hat{\theta}_i$ and approaches zero for large $|\theta_i - \hat{\theta}_i|$ if the parameter is identifiable. The area under $\text{PL}(\theta_i)$ provides a reasonable measure for the uncertainty of parameter θ_i. For further details, we refer to [16, 17].

In the following, we employ profile likelihoods to assess the information content of the moments of the data in comparison with that of the full distribution of data. More information will result in many identifiable parameters and small parameter uncertainties.



(a) **Four stochastic realizations of the 3-stage model of gene expression.**



(b) **Population snapshot data used for parameter estimation.**

Figure 2. **Dynamics of the 3-stage model of gene expression.** (a) Time-dependent protein number in four representative cells together with the population mean. (b) Population snapshot data obtained by sampling single cell trajectories. The size of the markers in (b) is proportional to the number of observed cells with the corresponding protein number. Due to the long tail of the distribution, the mode of the data seen in (b) differs significantly from the mean of the data depicted in (a).

3. RESULTS AND DISCUSSION

3.1. Parameter estimation for the 3-stage model of gene expression

In this section, we compare the performance of previously mentioned estimation methods, namely, FSP-based and MM-based parameter estimates, using the common 3-stage model of gene expression [2]. A schematic of the process and the corresponding moment equations for mean and variance are shown in Figure 1. The model has six parameters: the transition rate of DNA into the on-state (τ_{on}), the transition rate of DNA into the off-state (τ_{off}), the transcription rate in the on-state (k_{r}), the rate of mRNA degradation (γ_{r}), the translation rate (k_{p}), and the rate of protein degradation (γ_{p}). In the following, we study the problem of estimating these rates from protein measurements. Therefore, we generate artificial data

$$\mathcal{D}_k = \left\{ \left(\bar{Y}^{(s)}(t_k), t_k \right) \right\}_{s=1}^{10^5}, \quad k = 1, \ldots, 10,$$

with $t_k = k$ and $\bar{Y}$ being the number of proteins. For the generation of the artificial data, the parameter vector

$$\begin{aligned} \theta^{\text{true}} &= (\tau_{\text{on}}, \tau_{\text{off}}, k_{\text{r}}, \gamma_{\text{r}}, k_{\text{p}}, \gamma_{\text{p}})^{\text{T}} \\ &= (0.05, 0.05, 5, 1, 4, 1)^{\text{T}} \end{aligned}$$

is used. We refer to this parameter vector θ^{true} as the true parameter vector in the following. Also, no measurement noise is considered in the generation of the data. In the initial state, mRNA and protein numbers follow a Poisson distribution with mean 4 and 10, respectively. The probability to be in the DNA on-state is 0.7. Figure 2 depicts sample paths of the model (Figure 2(a)) as well as the snapshot data (Figure 2(b)) used for parameter estimation. Using these data we estimate $\theta = (\tau_{\text{on}}, \tau_{\text{off}}, k_{\text{r}}, \gamma_{\text{r}}, k_{\text{p}}, \gamma_{\text{p}})^{\text{T}}$.

For FSP-based and moment-based likelihood functions the maximum likelihood estimates are computed and the parameter uncertainty is evaluated. For the moment-based likelihood function we employed different moment orders. The uncertainty of the moments has been determined using the non-parametric bootstrapping approach introduced before.

Figure 3 depicts the model simulation for the ML estimates for the different likelihood functions along with the data. It is clear that for all ML estimates we observe a good agreement with the data used for the estimation. To validate the ML estimates, we employed the higher-order moments of the data, which have not been used for the parameter estimation. We find that all ML estimates, which were obtained using at least the mean and the variance, successfully predict the higher-order moments not used to obtain the ML estimates. Only the ML estimate computed merely from the mean of the data fails. Thus, the information contained in the mean is insufficient. This is confirmed by the profile likelihoods shown in Figure 4, which show that all likelihood functions establish identifiability, except the moment-based likelihood function of order 1. A careful comparison of the profile likelihoods shows that the uncertainty in the estimation of the parameters decreases as more information (more moments) are used. Since the FSP-based likelihood function makes use of all the information, the resulting parameter uncertainties are minimal. If the moment order is increased, the confidence intervals for moment-based likelihood function also become more narrow, however even for moment order 4, the result of the FSP remains superior. Note that for all likelihood functions, the true parameters are con-

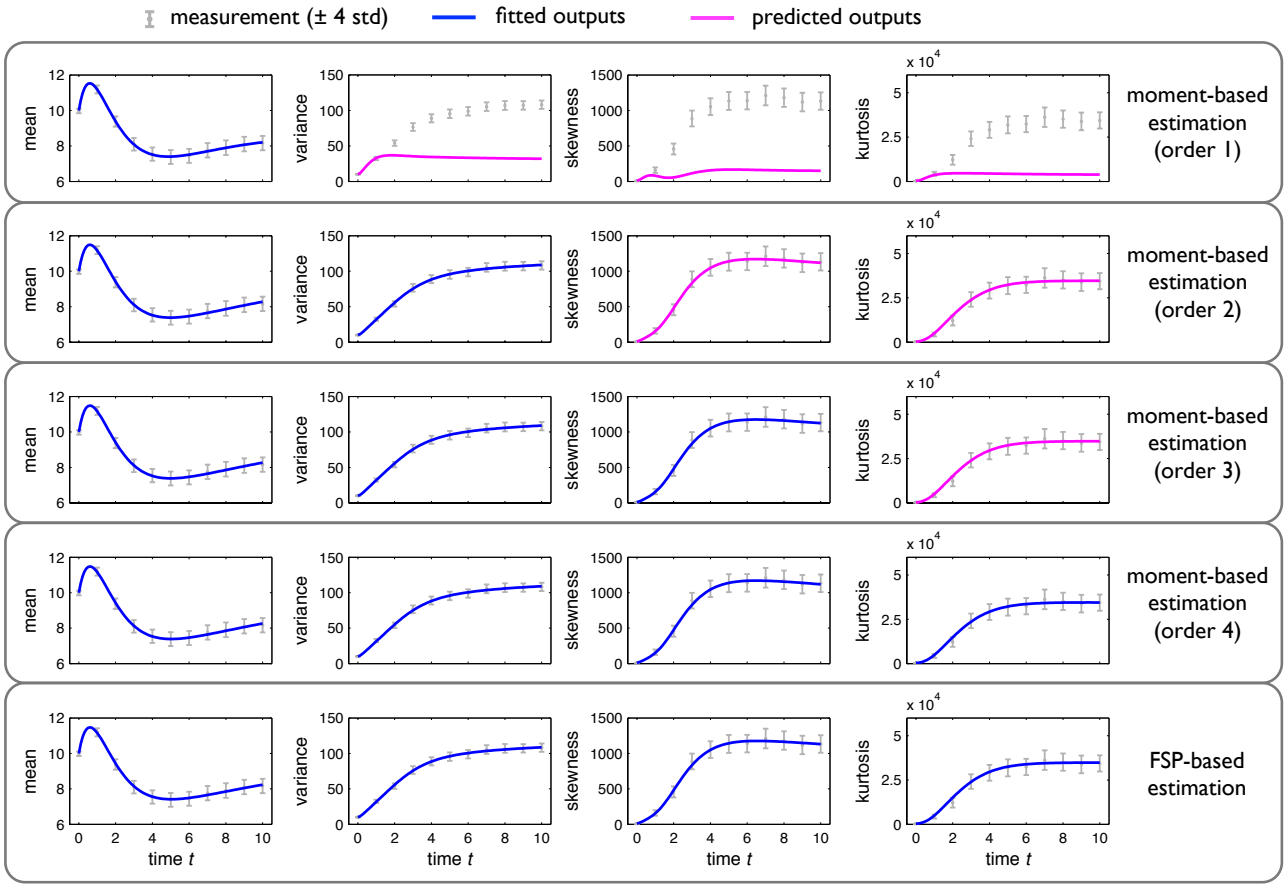


Figure 3. **Model-data comparison for ML estimates obtained using different likelihood functions.** ML estimation has been performed using moment-based likelihood functions of different orders (order 1: mean; order 2: mean and variance; order 3: mean, variance and skewness; and order 4: mean, variance, skewness and kurtosis) and the FSP-based likelihood function. Gray error bars show the mean and 4-σ intervals ($[\mu-4\sigma, \mu+4\sigma]$) of the measurement data. For the different ML estimates the fit is illustrated by showing the model output (blue lines, —) and the measurement data (grey error bars). All models describe the respective data well. To assess the predictive power of the model, the ML estimates are used to predict the higher-order moments (magenta lines, —) which have not been employed for the parameter estimation. The ML estimate computed using moment-based estimation of order 1 fails to provide good prediction, while already information about mean and variance (order 2) is sufficient to obtain a predictive model.

tained in the 95% confidence intervals constructed from the profile likelihoods (not shown).

3.2. Discussion

The computational complexity of the simulation of CTMCs renders the estimation of their parameters challenging. Different methods have been proposed to circumvent this complexity, among other the moment equations [18, 9, 14]. In this work, we evaluate the information contained in the moments of measurement data with respect to parameter estimation (by employing moment-based likelihood function) and compare it with the complete information contained in population snapshot data (by employing FSP-based likelihood function). The practical identifiability and the uncertainty of the parameter estimates are assessed using profile likelihoods. To the best of our knowledge, this is the first profile likelihood-based uncertainty analysis for stochastic processes, probably because the evaluation of the likelihood function is computationally often infeasible. This is not the case if a moment-based estimation is employed.

As a case study, we consider the widely used 3-stage model of gene expression [2]. For this model, we show that measurements of the mean expression do not in general ensure identifiability, but rather that measurements of the variance are required. This is consistent with results by Munsky *et al.* [18] for the two-stage model of gene expression. Information about third and fourth order moments can decrease the uncertainty further, however this reduction is often insignificant. The full information contained in the data, which is exploited by the FSP-based estimation, remains out of reach for the MM-based estimation approach.

Although the FSP-based likelihood function is statistically more powerful, parameter estimation based on the moment equation is the method of choice for processes,

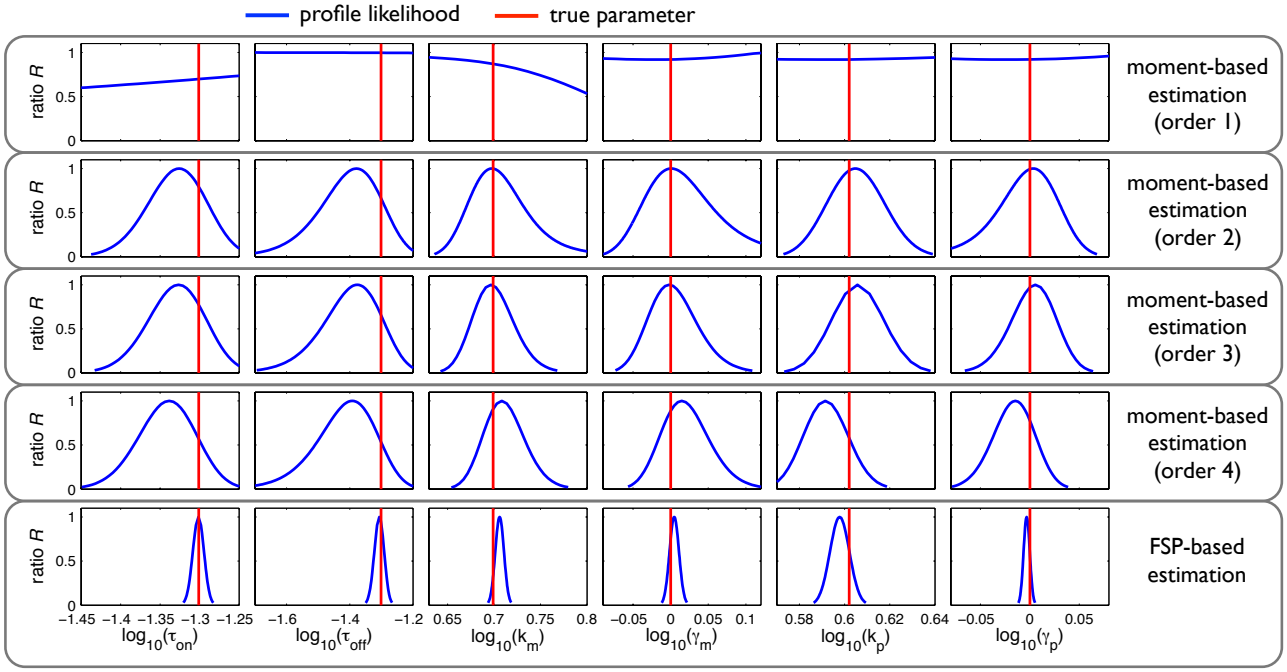


Figure 4. **Parameter uncertainty for different likelihood functions.** The parameter uncertainty and parameter identifiability has been evaluated for moment-based likelihood functions of different orders (order 1: mean; order 2: mean and variance; order 3: mean, variance and skewness; and order 4: mean, variance, skewness and kurtosis) and the FSP-based likelihood function. The profile likelihoods (blue lines, —) indicate that the measurements of the mean do not carry enough information to identify the parameters. Information about mean and variance ensures identifiability, and the uncertainty is slightly reduced if additional moments are used. The FSP-based likelihood function, which exploits all information contained in the data, yields the smallest uncertainties. All confidence intervals (not shown), derived from likelihood profiles, contain the true parameter values (red lines, —), which indicates consistency.

in particular, if the FSP is infeasible. Furthermore, parameter estimation using the moment equation is more efficient. The parameter estimation using the moment equation of order 2 is roughly 30 times faster than the parameter estimation using the FSP. However, it remains to be studied how moment closures, which are required for systems including bimolecular reactions, influence the parameter estimation. If a bias is introduced, as we expect, it should be analyzed how a refinement of the moment equation, e.g., the conditional moment equation [19], can be used to improve the results.

Beyond the profile likelihood-based evaluation of the information encoded in the moments, we introduced a non-parametric bootstrapping approach to evaluate the uncertainty of the empirical estimates of the moments. This approach allows for the construction of likelihood function without additional distribution assumptions. Furthermore, we illustrated how the higher-order moments, which have not been used for parameter estimation, can be used for model validation. This approach is attractive, as models can basically be fitted and validated on the same dataset.

4. AUTHOR'S CONTRIBUTIONS

AK and JH developed the method and analyzed the 3-stage model of gene expression. JH and FJT devised the project. AK, JH and FJT wrote, read and approved the final manuscript.

5. ACKNOWLEDGEMENTS

The authors acknowledge financial support by the European Union within the ERC grant 'LatentCauses' and the BMBF grant 'Virtual Liver' (grant-nr. 315752). The authors would also like to thank Justin Feigelman and Sabine Hug for proofreading the manuscript.

6. REFERENCES

[1] A. Eldar and M. B. Elowitz, "Functional roles for noise in genetic circuits," *Nat.*, vol. 467, no. 9, pp. 1–7, Sept. 2010.

[2] V. Shahrezaei and P. S. Swain, "Analytical distributions for stochastic gene expression," *Proc. Natl. Acad. Sci. U S A*, vol. 105, no. 45, pp. 17256–17261, Nov. 2008.

[3] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *J. Phys. Chem.*, vol. 81, no. 25, pp. 2340–2361, Dec. 1977.

[4] H. E. Samad, M. Khammash, L. Petzold, and D. Gillespie, "Stochastic modelling of gene regulatory networks," *Int. J. Robust Nonlinear Control*, vol. 15, no. 15, pp. 691–711, Oct. 2005.

[5] B. Munsky and M. Khammash, "The finite state projection algorithm for the solution of the chemical master equation," *J. Chem. Phys.*, vol. 124, no. 4, pp. 044104, Jan. 2006.

[6] S. Engblom, "Computing the moments of high dimensional solutions of the master equation," *Appl. Math. Comp.*, vol. 180, pp. 498–515, 2006.

[7] J. P. Hespanha, "Modeling and analysis of stochastic hybrid systems," *IEE Proc. Control Theory & Applications,* Special Issue on Hybrid Systems, vol. 153, no. 5, pp. 520–535, 2007.

[8] J. Ruess, A. Milias, S. Summers, and J. Lygeros, "Moment estimation for chemically reacting systems by extended Kalman filtering," *J. Chem. Phys.*, vol. 135, no. 165102, Oct. 2011.

[9] P. Milner, C. S. Gillespie, and D. J. Wilkinson, "Moment closure based parameter inference of stochastic kinetic models," *Stat. Comp.*, 2012.

[10] M. Mateescu, V. Wolf, F. Didier, and T. Henzinger, "Fast adaptive uniformisation of the chemical master equation," *IET. Syst. Biol.*, vol. 4, no. 6, pp. 441–452, 2010.

[11] A. Singh and J. P. Hespanha, "Approximate moment dynamics for chemically reacting systems," *IEEE Trans. Autom. Control*, vol. 56, no. 2, pp. 414–418, Feb. 2011.

[12] J. Hasenauer, N. Radde, M. Doszczak, P. Scheurich, and F. Allgöwer, "Parameter estimation for the CME from noisy binned snapshot data: Formulation as maximum likelihood problem," Extended abstract at *Conf. of Stoch. Syst. Biol.*, Monte Verita, Switzerland, July 2011.

[13] T. Nüesch, "Finite state projection-based parameter estimation algorithms for stochastic chemical kinetics," Master thesis, Swiss Federal Institute of Technology, Zürich, 2010.

[14] C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl, "Moment-based inference predicts bimodality in transient gene expression," *Proc. Nati. Acad. Sci. U S A*, vol. 109, no. 21, pp. 8340–8345, May 2012.

[15] T. J. DiCiccio and B. Efron, "Bootstrap confidence intervals," *Statist. Sci.*, vol. 11, no. 3, pp. 189–228, 1996.

[16] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer, "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood," *Bioinf.*, vol. 25, no. 25, pp. 1923–1929, May 2009.

[17] W. Q. Meeker and L. A. Escobar, "Teaching about approximate confidence regions based on maximum likelihood estimation," *Am. Stat.*, vol. 49, no. 1, pp. 48–53, Feb 1995.

[18] B. Munsky, B. Trinh, and M. Khammash, "Listening to the noise: random fluctuations reveal gene network parameters," *Mol. Syst. Biol.*, vol. 5, no. 318, Oct. 2009.

[19] J. Hasenauer, V. Wolf, A. Kazeroonian, and F. J. Theis, "Method of conditional moments (MCM) for the chemical master equation," *submitted to the Journal of Mathematical Biology*, 2012.

Appendix B

Modeling of stochastic biological processes with non-polynomial propensities using non-central conditional moment equation.

IFAC Proceedings Volumes, 2014.

This is a pre-copyedited, author-produced version of an article accepted for publication in IFAC Proceedings Volumes following peer review. The version of record
A. Kazeroonian, F. J. Theis, and J. Hasenauer. **Modeling of stochastic biological processes with non-polynomial propensities using non-central conditional moment equation.** *IFAC Proceedings Volumes, Volume 47, Issue 3, 2014, Pages 1729-1735.* is available online at:
http://dx.doi.org/10.3182/20140824-6-ZA-1003.02298

Modeling of stochastic biological processes with non-polynomial propensities using non-central conditional moment equation

Atefeh Kazeroonian, Fabian J. Theis, and Jan Hasenauer

Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany

Department of Mathematics, Technische Universität München, 85748 Garching, Germany

(e-mail: {atefeh.kazeroonian, fabian.theis, jan.hasenauer}@helmholtz-muenchen.de).

Abstract: Biological processes exhibiting stochastic fluctuations are mainly modeled using the Chemical Master Equation (CME). As a direct simulation of the CME is often computationally intractable, we recently introduced the Method of Conditional Moments (MCM). The MCM is a hybrid approach to approximate the statistics of the CME solution. In this work, we provide a more comprehensive formulation of the MCM by using non-central conditional moments instead of central conditional moments. The modified formulation allows for additional insight into the model structure and for extensions to higher-order reactions and non-polynomial propensity functions. The properties of the non-central MCM are analyzed using a model for the regulation of pili formation on the surface of bacteria, which possesses rational propensity functions.

Keywords: stochastic modeling, chemical master equation, moment equations

1. INTRODUCTION

Gene expression, signal transduction and even cell fate decisions have been shown to be subject to stochastic fluctuations [Raser and O'Shea, 2004, Eldar and Elowitz, 2010]. These stochastic fluctuations are often due to the low abundance of DNAs, mRNAs and proteins [Taniguchi et al., 2010]. For decades it was assumed that these fluctuations are a nuisance and disturb correct information processing in cells. However, in recent years it has been shown that stochastic fluctuations are essential for functioning as well as robustness of many processes [Eldar and Elowitz, 2010]. Furthermore, fluctuations can be employed to unravel the underlying signaling mechanisms [Munsky et al., 2009, 2012].

A multitude of approaches have been proposed to model stochastic dynamics in biological systems. Discrete-state continuous-time Markov chains (CTMCs) are the gold standard as they capture the discreteness of the ensemble sizes of chemical species $(S_1, S_2, \dots, S_{n_s})$ as well as the discreteness of chemical reactions,

$$\mathrm{R}_j: \quad \sum_{i=1}^{n_s} \nu_{ij}^- S_i \to \sum_{i=1}^{n_s} \nu_{ij}^+ S_i, \quad j = 1, \dots, n_r.$$

The stoichiometric coefficients ν_{ij}^-, ν_{ij}^+ and $\nu_{ij} = \nu_{ij}^+ - \nu_{ij}^-$ denote the number of molecules of species S_i consumed, produced and net produced, respectively, when the reaction R_j takes place. Accordingly, ν_j^-, ν_j^+ and ν_j describe the overall stoichiometry of reaction R_j.

CTMCs describe the time evolution of the ensemble state $X_t = (X_{1,t}, \dots, X_{n_s,t}) \in \mathbb{N}_0^{n_s}$ of the species $S_1, S_2, \dots, S_{n_s}$ as a jump process. X_t remains constant as long as no reaction occurs. If R_j takes place, the ensemble sizes change according to the stoichiometry of R_j, $X_t \to X_t + \nu_j$. The index j of the next reaction and the time to the next reaction are random with distributions determined by the propensity functions $a_j : \mathbb{N}_0^{n_s} \to \mathbb{R}_+$, $j = 1, \dots, n_r$ [Feller, 1940]. The statistics of the process, i.e. the probabilities $p(x|t) = P(X_t = x)$ that X_t occupies a certain state x at time t, are described by the chemical master equation (CME) [van Kampen, 2007],

$$\frac{\partial}{\partial t} p(x|t) = \sum_{\substack{j=1 \\ x \geq \nu_j^+}}^{n_r} a_j(x - \nu_j) p(x - \nu_j|t) - \sum_{j=1}^{n_r} a_j(x) p(x|t), \quad (1)$$

in which the inequality constraint $x \geq \nu_j^+$ ensures positivity. Associated propensities a_j are "proper", meaning that if $\exists i \in \{1, \dots, n_s\} : X_{i,t} \not\geq \nu_{ij}^-$ then $a_j(X_t) = 0$.

The CME is a system of linear ordinary differential equations (ODEs) which describes the dynamics of CTMCs. Jahnke and Huisinga [2007] derived a closed-form solution of the CME in the case of monomolecular reactions. If the process contains nonlinear propensity functions, in general, numerical approximations are necessary. A multitude of approximation methods have been proposed over the last decades, e.g., error-aware state truncation [Munsky and Khammash, 2006], inexact integration [Sidje et al., 2007], product approximations [Jahnke, 2011], approximation of the CME by the Fokker-Planck equation [Gardiner,

2011], or modeling of the statistical moments of the CME solution [Engblom, 2006]. However, these methods often fail if low- as well as high-copy number species are involved in the biochemical process.

In recent years, several hybrid methods have been introduced to circumvent these shortcomings. These hybrid methods are based on decomposing the system into fast and slow reactions [Haseltine and Rawlings, 2002], or low- and high-copy number species [Hellander and Lötstedt, 2007, Henzinger et al., 2010, Jahnke, 2011, Menz et al., 2012]. For the latter we recently proposed a generalization, the *method of conditional moments (MCM)* [Hasenauer et al., 2013]. The MCM provides a fully stochastic description for the low-copy number species and a moment-based description for the medium/high-copy number species. Thus, it combines concepts from *hybrid stochastic-deterministic modeling* [Jahnke, 2011, Menz et al., 2012] and *moment-based modeling* [Engblom, 2006]. We showed that this allows for an improved approximation quality for common models of transcription-translation process.

In this manuscript, we generalize the MCM to include reactions with rates not obeying the law of mass action. This allows for the consideration of activation and inhibition mechanisms possessing Michaelis-Menten-like characteristics. In addition to this generalization, we state the MCM in terms of non-central moments. This improves the readability and interpretability compared to the central MCM [Hasenauer et al., 2013]. To enhance the MCM further for systems with nonlinear propensity functions, we propose the use of Taylor series expansion (TSE) together with the low-dispersion closure scheme. This approach is evaluated using a model for PapI regulation in *E. coli* [Munsky and Khammash, 2006].

2. APPROACH

Single-molecule fluorescence microscopy techniques, such as fluorescence *in situ* hybridization, revealed that the copy numbers of chemical species spread over several orders of magnitude. In *E. coli*, the mean number of a protein is in general 100- to 1000-fold higher than the mean number of the corresponding mRNA [Taniguchi et al., 2010]. Such naturally occurring scale separations can be exploited to accelerate the simulation of stochastic biochemical processes. Therefore, species $S_1, \dots, S_{n_s}$ are classified as either low- or medium/high-copy number species. The abundances of low-copy number species are collected in Y_t, while the abundances of medium/high-copy number species are collected in Z_t. Thus, without lost of generality $X_t = (Y_t, Z_t)$ and $p(x|t) = p(y, z|t)$.

The CME describes the evolution of the full joint distribution $p(y, z|t)$. In contrast, the MCM employs the decomposition

$$p(y, z|t) = p(z|y, t)p(y|t) \tag{2}$$

which follows from the multiplication axiom. $p(y|t)$ denotes the marginal probability of the low-copy number species being in state y, while $p(z|y, t)$ denotes the conditional probability of the medium/high-copy number species being in state z given that the low-copy number species are in state y. Using this decomposition, the CME can be rewritten as

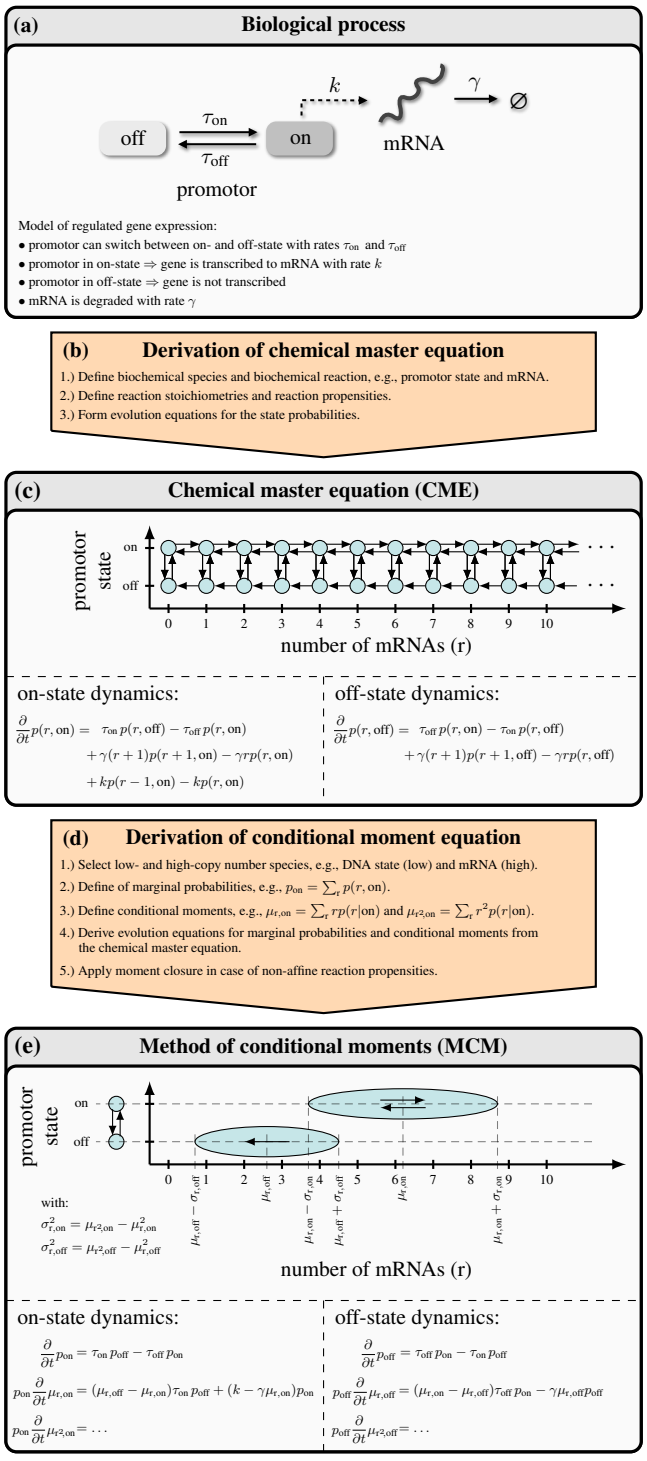


Fig. 1. Illustration of the method of conditional moments using a two-state model for gene expression. (a) Model of gene expression accounting for two promotor states [Munsky et al., 2012]. (b) Procedure to derive the CME. (c) CME of the gene expression model. The discrete state space is visualized along with the possible transitions. Note that we skip the dependence on the time t to simplify the notation. (d) Procedure to derive the conditional moment equation from the CME. (e) Conditional moment equation when modeling the promotor state as low-copy number species and the mRNA as medium/high-copy number species.

$$\frac{\partial}{\partial t}p(y,z|t) = -\sum_{j=1}^{n_r} a_j(y,z)p(z|y,t)p(y|t)$$
$$+\sum_{\substack{j=1\\ y\geq\nu_{j,y}^+\\ z\geq\nu_{j,z}^+}}^{n_r} a_j(y-\nu_{j,y},z-\nu_{j,z})p(z-\nu_{j,z}|y-\nu_{j,y},t)p(y-\nu_{j,y}|t). \tag{3}$$

This decomposition suggests that the dynamics of $p(y|t)$ and $p(z|y,t)$ can be modeled separately [Haseltine and Rawlings, 2005, Hasenauer et al., 2013]. In the MCM, the distribution of the low-copy number species is described in terms of marginal probabilities,

$$p(y|t) = \sum_{z\geq 0} p(y,z|t). \tag{4}$$

For medium/high-copy number species, the non-central moments of $p(z|y,t)$ are considered,

$$\mu_{I,z}(y,t) = \mathbb{E}_z\left[Z^I \middle|\, y,t\right] = \sum_{z\geq 0} z^I p(z|y,t), \tag{5}$$

with I being a non-negative integer-valued vector of length $n_{s,z}$ and $Z^I := \prod_{i=1}^{n_{s,z}} Z_i^{I_i}$. The conditioning on y can be important if transitions between different low-copy number states are slow. The marginal probabilities of discrete states together with the corresponding conditional moments can be used to determine the overall moments of the process, $\bar{\mu}_{I,z}(t)$, i.e. the moments independent of the stochastic states, via

$$\bar{\mu}_{I,z}(t) = \mathbb{E}\left[Z^I|t\right] = \sum_{y\geq 0} \mu_{I,z}(y,t)p(y|t).$$

In the following, we provide exact and approximate evolution equations for $p(y|t)$ and $\mu_{I,z}(y,t)$. Figure 1 provides a visual outline of the method.

3. NON-CENTRAL CONDITIONAL MOMENT EQUATIONS

Upon decomposition of state vector, $x = (y,z)$, evolution equations for marginal probabilities of low-abundance species, $p(y|t)$, as well as non-central moments of high-abundance species conditioned on the state of the low-abundance species, $\mu_{I,z}(y,t)$, have to be determined. Therefore, a governing equation for expectation of an arbitrary polynomial test-function $T(Z)$ is derived to provide MCM equations as special cases.

Lemma 1. Let $p(y,z|t) = p(z|y,t)p(y|t)$ satisfy a proper CME (3) ($\forall x \not\geq \nu_j^- : a_j(x) = 0$), then, for any polynomial test-function $T : \mathbb{N}_0^{n_z} \times \mathbb{R}_+ \to \mathbb{R}$,

$$\frac{\partial}{\partial t}\left(\mathbb{E}_z[T(Z)|\, y,t]\, p(y|t)\right) =$$
$$\sum_{\substack{j=1\\ y\geq\nu_{j,y}^+}}^{n_r} \mathbb{E}_z[T(Z+\nu_{j,z})a_j(y-\nu_{j,y},Z)|\, y-\nu_{j,y},t]\, p(y-\nu_{j,y}|t)$$
$$-\sum_{j=1}^{n_r} \mathbb{E}_z[T(Z)a_j(Z,y)|\, y,t]\, p(y|t). \tag{6}$$

Note that Lemma 1 is only valid if the expectation $\mathbb{E}_z[T(Z)a_j(Z,y)|\, y,t]$ exists. This is generally true for reasonable models of biological processes.

Proof. The time derivative of $\mathbb{E}_z[T(Z)|\, y,t]\, p(y|t)$ is

$$\frac{\partial}{\partial t}\left(\mathbb{E}_z[T(Z)|\, y,t]\, p(y|t)\right) = \frac{\partial}{\partial t}\left(\sum_{z\geq 0} T(z)p(z|y,t)p(y|t)\right)$$
$$= \sum_{z\geq 0} T(z)\frac{\partial}{\partial t}p(z,y|t) + \sum_{z\geq 0} p(z,y|t)\frac{\partial}{\partial t}T(z).$$

The second term vanishes as the time derivative of $T(z)$ is zero. Similar to the proof by Hasenauer et al. [2013], $\frac{\partial}{\partial t}p(z,y|t)$ is substituted according to the CME (3), the order of summations is changed, and z is replaced by $z+\nu_{j,z}$ in the first sum to obtain

$$\frac{\partial}{\partial t}\left(\mathbb{E}_z[T(Z)|\, y,t]\, p(y|t)\right) =$$
$$\sum_{\substack{j=1\\ y\geq\nu_{j,y}^+}}^{n_r} \sum_{z\geq\nu_{j,z}^-} T(z+\nu_{j,z})a_j(y-\nu_{j,y},z)p(z|y-\nu_{j,y},t)$$
$$\times\, p(y-\nu_{j,y}|t) - \sum_{j=1}^{n_r}\sum_{z\geq 0} T(z)a_j(y,z)p(z|y,t)p(y|t).$$

The lower bound $z \geq \nu_{j,z}^-$ can be replaced by $z \geq 0$ as for $z \not\geq \nu_{j,z}^- : a_j(z) = 0$ (due to propensities being proper). Utilizing the definition of conditional expectation $\mathbb{E}_z[T(z)|\, y,t] = \sum_{z\geq 0} T(Z)p(z|y,t)$, the expression above simplifies to the evolution equation stated in Lemma 1, which concludes the proof. □

Setting $T(Z)$ to 1 and Z^I, Lemma 1 yields the governing equations for $p(y|t)$ and $\mu_{I,z}(y,t)$ respectively.

Theorem 2. Let $p(y,z|t) = p(z|y,t)p(y|t)$ satisfy a proper CME (3), the evolution equations for marginal probabilities, $p(y|t)$, and non-central conditional moments, $\mu_{I,z}(y,t)$, are given by the system

$$\frac{\partial}{\partial t}p(y|t) = -\sum_{j=1}^{n_r} \mathbb{E}_z[a_j(Z,y)|\, y,t]\, p(y|t)$$
$$+\sum_{\substack{j=1\\ y\geq\nu_{j,y}^+}}^{n_r} \mathbb{E}_z[a_j(y-\nu_{j,y},Z)|\, y-\nu_{j,y},t]\, p(y-\nu_{j,y}|t),$$
$$p(y|t)\frac{\partial}{\partial t}\mu_{I,z}(y,t) + \mu_{I,z}(y,t)\frac{\partial}{\partial t}p(y|t) =$$
$$\sum_{\substack{j=1\\ y\geq\nu_{j,y}^+}}^{n_r} \mathbb{E}_z\left[a_j(y-\nu_{j,y},Z)(Z+\nu_{j,z})^I \middle|\, y-\nu_{j,y},t\right] \tag{7}$$
$$\times\, p(y-\nu_{j,y}|t) - \sum_{j=1}^{n_r} \mathbb{E}_z\left[a_j(Z,y)Z^I \middle|\, y,t\right] p(y|t).$$

The MCM equations can be written for moments of arbitrary order. In contrast to the central conditional moment equations [Hasenauer et al., 2013], no distinction between first and higher-order moments is necessary, yielding a more compact set of equations. Also this presentation of the MCM is a generalization of the

central MCM since it removes the assumption that the propensity functions should allow for a decomposition of the form $a_j(x,t) = c\, g_j(y,t)h_j(z,t)$. Therefore, the non-central MCM provides a simpler and more general formulation and thus facilitates further investigations.

The resulting set of evolution equations is a DAE system. Initial conditions for $p(y|t)$, $\mu_{I,z}(y,t)$, $\dot{p}(y|t)$ and $\dot{\mu}_{I,z}(y,t)$ can be calculated via (7) given that $\forall y : p(y|t_0) \neq 0$. If this is not fulfilled, the procedure introduced by Hasenauer et al. [2013] can be adopted.

The simulation of the conditional moment equations requires the evaluation of expectations $\mathbb{E}_z[a_j(Z,y)|\,y,t]$ and $\mathbb{E}_z\big[a_j(Z,y)Z^I\big|\,y,t\big]$. This can be done by employing the Taylor series expansion of the propensity function $a_j(z,y)$. More specifically, since the expectation with respect to the random variable Z is sought, $a_j(z,y)$ is merely expanded with respect to z. In principle, any expansion point can be selected for the TSE, however, the vector of conditional means of z, $\mu'_z(y,t) = \sum_{z\geq 0} z\; p(z|y,t) = (\mu_{e_1,z}(y,t),\ldots,\mu_{e_{n_s},z}(y,t))$, is considered in the following:

$$\begin{aligned} a_j(z,y) &= a_j(\mu'_z(y,t),y) \\ &+ \sum_{k=1}^{n_{s,z}} \frac{\partial a_j(\mu'_z(y,t),y)}{\partial z_k}(z_k - \mu_{e_k,z}(y,t)) \\ &+ \frac{1}{2}\sum_{k,l=1}^{n_{s,z}} \frac{\partial^2 a_j(\mu'_z(y,t),y)}{\partial z_k \partial z_l}(z_k - \mu_{e_k,z}(y,t))(z_l - \mu_{e_l,z}(y,t)) \\ &+ \ldots. \end{aligned} \tag{8}$$

The expectation $\mathbb{E}_z[a_j(Z,y)|\,y,t]$ follows as

$$\begin{aligned} \mathbb{E}_z[a_j(Z,y)|\,y,t] &= a_j(\mu'_z(y,t),y) \\ &+ \frac{1}{2}\sum_{k,l=1}^{n_{s,z}} \frac{\partial^2 a_j(\mu'_z(y,t),y)}{\partial z_k \partial z_l} C_{e_k+e_l,z}(y,t) + \ldots, \end{aligned} \tag{9}$$

in which e_i denotes the i^{th} unit vector and $C_{I,z}(y,t) = \sum_{z\geq 0}(z-\mu'_z(y,t))^I\, p(z|y,t)$ represent the central moments. The central moments can be replaced by their equivalent expressions in terms of non-central moments, e.g., $C_{e_k+e_l,z}(y,t) = \mu_{e_k+e_l,z}(y,t) - \mu_{e_k,z}(y,t)\mu_{e_l,z}(y,t)$. In case the TSE (9) is finite, $\mathbb{E}_z\big[a_j(Z,y)Z^I\big|\,y,t\big]$ can be evaluated in a similar manner by writing the TSE of $a_j(Z,y)Z^I$. However, if the TSE (9) is infinite, or intractably high-order, it may be truncated at a specific order N. This truncation introduces a degree of freedom in choosing either of the following approaches for evaluating $\mathbb{E}_z\big[a_j(Z,y)Z^I\big|\,y,t\big]$.

Truncate-multiply approach. To approximate the expectation $\mathbb{E}_z\big[a_j(Z,y)Z^I\big|\,y,t\big]$, first the TSE of $a_j(Z,y)$ (8) is truncated at order N, and then it is multiplied by Z^I. The expectation of the resulting product is

$$\begin{aligned} \mathbb{E}_z\big[a_j(Z,y)Z^I\big|\,y,t\big] &= a_j(\mu'_z(y,t),y)\mu_{I,z}(y,t) \\ &+ \sum_{k=1}^{n_{s,z}} \frac{\partial a_j(\mu'_z(y,t),y)}{\partial z_k}\mathbb{E}_z\big[(Z_k - \mu_{e_k,z}(y,t))Z^I\big|\,y,t\big] \\ &+ \frac{1}{2}\sum_{k,l=1}^{n_{s,z}} \frac{\partial^2 a_j(\mu'_z(y,t),y)}{\partial z_k \partial z_l} \\ &\times \mathbb{E}_z\big[(Z_k - \mu_{e_k,z}(y,t))(Z_l - \mu_{e_l,z}(y,t))Z^I\big|\,y,t\big] + \ldots. \end{aligned} \tag{10}$$

The expectation terms in (10) can easily be expressed in terms of non-central moments.

Multiply-truncate approach. In the multiply-truncate approach, the order of operations is changed. First Z^I is multiplied by $a_j(Z,y)$, then the TSE of the product $a_j(Z,y)Z^I$ is obtained and, if necessary, truncated, yielding the expectation

$$\begin{aligned} \mathbb{E}_z\big[Z^I a_j(Z,y)\big|\,y,t\big] &= (\mu'_z(y,t))^I\, a_j(\mu'_z(y,t),y) \\ &+ \frac{1}{2}\sum_{k,l=1}^{n_{s,z}} \frac{\partial^2\left((\mu'_z(y,t))^I\, a_j(\mu'_z(y,t),y)\right)}{\partial z_k \partial z_l} C_{e_k+e_l,z}(y,t) \\ &+ \ldots. \end{aligned} \tag{11}$$

In the truncate-multiply approach, if the TSE is truncated at order N, (10) contains moments up to order $N+\sum_i I_i$, whereas in the multiply-truncate approach, with the TSE of order N, (11) contains moments up to order N. Thus, for the multiply-truncate approach it may be more plausible to have the truncation order N equal to or greater than the moment order, i.e. $N \geq \sum_i I_i$. In this way, the evolution equation for a moment $\mu_{I,z}(y,t)$ depends on moments of the same order.

4. CLOSURE OF THE CONDITIONAL MOMENT EQUATIONS

The evolution equations for moments up to order M, i.e. $\forall I : \sum_i I_i \leq M$, in general depend on moments of orders $> M$. To simulate the conditional moment equations, these higher-order moments have to be approximated using moment closure. Also, if the propensities are non-polynomial, their TSEs are generally infinite and need to be truncated. Accordingly, the accuracy of conditional moment equations is determined by (1) the error introduced by truncating Taylor series of the propensity functions $a_j(z,y)$, and (2) the error introduced by the moment closure scheme. In the following, these two sources of error are discussed for polynomial and non-polynomial propensity functions.

4.1 Polynomial propensities

If the kinetics obey the law of mass action, all propensities are polynomial functions and their TSEs are finite and their truncation is not necessary. However, higher-order moments still appear.

Under certain conditions the higher-order moments cancel out, yielding a closed set of equations [Hasenauer et al., 2013]. However, in general, closure schemes have to be employed. Moment closure schemes approximate higher-order moments as functions of the lower-order moments, e.g., using distributional assumptions [Engblom, 2006, Singh and Hespanha, 2011]. For instance, the simplest and also most commonly used moment closure is low-dispersion closure which relies on the assumption that the distribution is tightly clustered around the mean, implying that the higher-order central moments are negligible. Accordingly, all higher-order central moments are set to zero,

$$\forall I \text{ with } \sum_i I_i > M : C_{I,z}(y,t) = 0. \tag{12}$$

In case of polynomial propensities, different moment closure schemes can be used with either of the truncate-multiply and multiply-truncate approaches. The error of the approximation is then directly related to the validity of the assumptions made by the closure.

4.2 Non-polynomial propensity functions

In case of non-polynomial propensity functions, the corresponding TSEs are infinite and need to be truncated. In this case, both the errors introduced by the truncation of the TSE and by the moment closure affect the approximation quality of the MCM.

TSEs are truncated by discarding higher-order terms in (9) and (10) or (11). In the truncate-multiply approach, higher-order terms in (9) and (10) are of different natures, i.e. the former are the higher-order central moments while the latter are combinations of non-central moments. Thus, setting them to zero implies different, and inconsistent, assumptions about the moments. However, in the multiply-truncate approach, truncations of the TSEs (9) and (11) both correspond to the same assumption, i.e. that the higher-order central moments are zero. This is conceptually similar to the low-dispersion moment closure, which also sets higher-order central moments to zero. Hence, in the approximation of conditional moment equations with non-polynomial propensities, the low dispersion closure together with the multiply-truncate approach is a promising choice as it ensures consistency.

Interestingly, it can be shown that using the low-dispersion closure, the truncate-multiply approach is identical to the multiply-truncate approach, given that the order of the TSE truncation at least equals the moment order, i.e. $N \geq M$. However, the two approaches are different if $N < M$, or if another moment closure scheme is applied.

5. EXAMPLE: PAPI REGULATION MODEL

In this section, the performance of non-central MCM is assessed using a biological system that describes the regulation of Pap pili formation on the surface of *E. coli* [Munsky and Khammash, 2006]. This biological process involves low- as well as medium/high-copy number species. Therefore, it is challenging for simulation methods that do not account for the differences in the abundance of the species. Furthermore, it demands handling of non-polynomial propensity functions.

Several simulations based on MCM with different setups are carried out and the results are compared to the results obtained by finite state projection (FSP). As shown by Munsky and Khammash [2006], the results of FSP can be assumed to be exact for this problem.

5.1 Biological system

The PapI regulation model (Figure 2) comprises a *pap* operon and two regulatory proteins. The regulatory protein LRP can reversibly bind to either or both of the binding sites on the *pap* operon. The states g_1 to g_4 represent the four possible configurations of the *pap* operon. Pili production can only take place if the operon is in state g_2. Protein PapI decreases the unbinding rate of LRP from

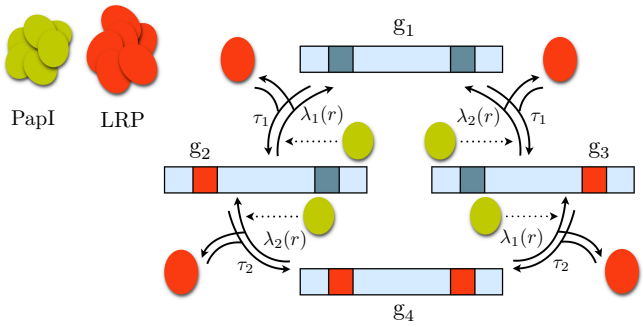


Fig. 2. Schematic of the PapI regulation model. Arrows represent the binding and unbinding of LRP to/from the operon. Dotted arrows indicate the influence of PapI on the reaction rates.

Table 1. Reactions and reaction propensities for the PapI regulation model.

reaction number	stoichiometry	rate
R_1	$\mathrm{g}_1 + l \to \mathrm{g}_2$	$\tau_1 = c_1$
R_2	$\mathrm{g}_2 \to l + \mathrm{g}_1$	$\lambda_1 = c_3 - c_4 \frac{r}{r+1}$
R_3	$\mathrm{g}_1 + l \to \mathrm{g}_3$	$\tau_1 = c_1$
R_4	$\mathrm{g}_3 \to l + \mathrm{g}_1$	$\lambda_2 = c_5 - c_6 \frac{r}{r+1}$
R_5	$\mathrm{g}_2 + l \to \mathrm{g}_4$	$\tau_2 = c_2$
R_6	$\mathrm{g}_4 \to l + \mathrm{g}_2$	$\lambda_2 = c_5 - c_6 \frac{r}{r+1}$
R_7	$\mathrm{g}_3 + l \to \mathrm{g}_4$	$\tau_2 = c_2$
R_8	$\mathrm{g}_4 \to l + \mathrm{g}_3$	$\lambda_1 = c_3 - c_4 \frac{r}{r+1}$
R_9	$\mathrm{g}_2 \to \mathrm{g}_2 + r$	k_r
R_{10}	$r \to \emptyset$	γ_r

the operon, and therefore establishes a positive feedback loop for the production of pili. The total number of LRP molecules (denoted by l) is constant, while the count of PapI molecules (denoted by r) is variable. Reactions and kinetic rates of the model are provided in Table 1.

The operon states are modeled as low-abundance species as there is only a single operon. PapI and LRP proteins are found in relatively larger amounts, therefore, they are considered as medium/high-copy number species and represented by the moments of their distributions. Furthermore, to obtain the MCM equations, the nonlinear kinetic rates, i.e. those in reactions R_2, R_4, R_6 and R_8, should be approximated as polynomials by means of TSE.

5.2 Simulation study

To analyze the impact of the approximation errors of moment closure and truncation of TSE (in either of the truncate-multiply and multiply-truncate approaches) on the accuracy of the MCM simulation, several simulations are carried out. We use the notation MCMi/j to refer to different simulations where i denotes the highest moment order (previously mentioned as M) and j denotes the order of the TSE (previously mentioned as N). For all the simulations, parameter values $(c_1, c_2, c_3, c_4, c_5, c_6, k_r, \gamma_r) = (1, 0.01, 2.5, 2.25, 1.2, 0.2, 10, 1)$ and initial conditions $l = 100, r = 5$, and $p(\mathrm{g}_1) = 1$ are used.

Using the truncate-multiply approach (Figure 3), we find that all MCM simulations generally agree with the FSP in resolving marginal probabilities and conditional moments. However, as Figure 4 shows, there is no consistent trend in

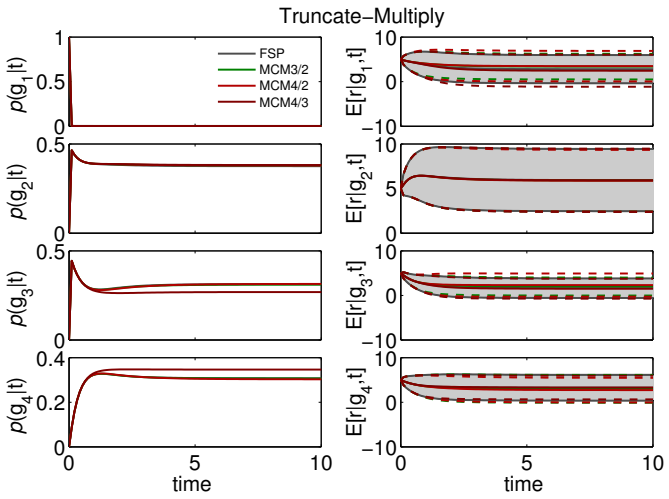


Fig. 3. Marginal probabilities of states of the *pap* operon (left) and conditional means and 1-σ intervals of PapI (right) for FSP, MCM3/2, MCM4/2, and MCM4/3 with the truncate-multiply approach.

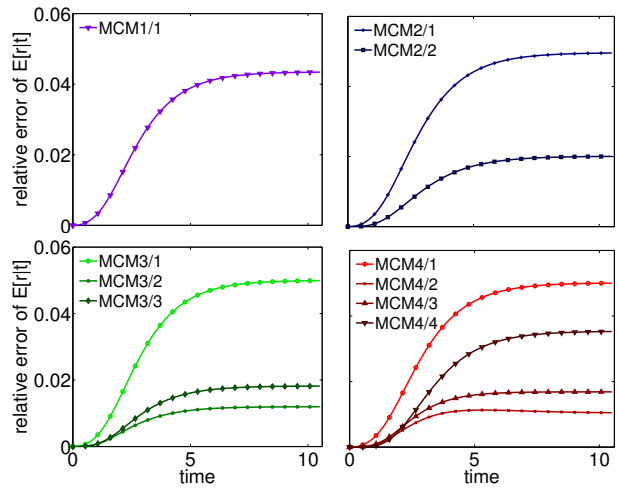


Fig. 4. Relative errors of the mean number of PapI molecules predicted by different MCM simulations with the truncate-multiply approach.

the impact of the moment order and the truncation order on the accuracy of the simulation results.

The results for the overall mean of PapI (Figure 4) suggest that, for most cases, applying a truncation order smaller than the moment order leads to improved approximation quality. For this example, the TSE of order 2 yields the smallest error. Although increasing the moment order improves the results when the truncation order is equal to/greater than two, this is not always the case.

For instance, MCM1/1 performs better than MCM2/1, MCM3/1, and MCM4/1 (Figure 4). Relative errors in Figure 4 are computed with respect to FSP simulation, e.g., $\text{error}_{\text{MCM2/2}} = \text{abs}(\mathbb{E}[r|t]_{\text{MCM2/2}} - \mathbb{E}[r|t]_{\text{FSP}})/\mathbb{E}[r|t]_{\text{FSP}}$.

The same study is repeated for the multiply-truncate approach. In this approach, the highest moment order that appears in the MCM equations corresponds to the minimum of the truncation order and the moment order,

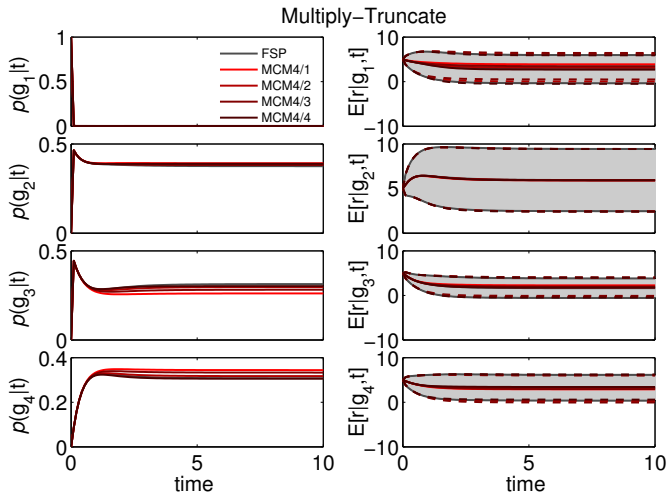


Fig. 5. Marginal probabilities of states of the *pap* operon (left) and conditional means and 1-σ intervals of PapI (right) for FSP, MCM4/1, MCM4/2, MCM4/3, and MCM4/4 with the multiply-truncate approach.

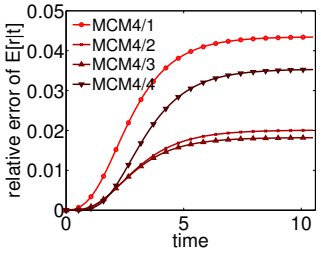


Fig. 6. Relative errors of the mean number of PapI molecules predicted by different MCM simulations with the multiply-truncate approach.

i.e. $\min(M, N)$. Therefore, given that $j < i$, all simulations MCMi/j with the same j are identical. Therefore, only the effect of the truncation order on the accuracy of the MCM simulation has to be investigated. Unfortunately, we again do not find a consistent trend (Figure 6).

To summarize, this example illustrates how MCM can be used to approximate the statistics of stochastic processes with non-polynomial reaction propensities. Surprisingly, no consistent trend was found in the impact of the order of TSE and the order of moment closure on the accuracy of the MCM simulation when low dispersion closure was used.

6. DISCUSSION

In this work, we presented the non-central conditional moment equations, a reformulation and extension of the central MCM [Hasenauer et al., 2013]. Being a hybrid simulation method for systems of stochastic dynamics, the MCM combines stochastic and moment-based descriptions depending on copy-numbers of species. Reformulation in terms of non-central moments facilitated the extension of the MCM to include reactions with non-polynomial kinetic rates. We proposed the use of Taylor series expansion for the approximation of non-polynomial propensity functions. As the truncation of the TSE introduces degrees of

freedom, we compared two alternative approaches for the approximation of the conditional moment equations.

To evaluate the performance of non-central MCM, a model for regulation of Pap pili formation on the surface of *E. coli* was analysed. Our study demonstrated that non-central MCM can handle non-polynomial propensity functions by means of Taylor series expansion. Surprisingly, we found that increasing the order of Taylor series expansion does not always improve the accuracy of simulation.

In situations where the low-dispersion assumption is not physically plausible, the compatibility of more sophisticated closure techniques [Gillespie, 2009, Singh and Hespanha, 2011] with the TSE has to be analyzed. Also, to further enhance the approximation quality, approximation approaches such as sigma-point expansion methods, instead of Taylor series expansion, might be used.

If all propensities are rational, the approach introduced by Milner et al. [2011] for moment equations can also be adapted for the MCM. In this approach, a polynomial system is obtained by multiplying the original system by the product of the propensity denominators, and the TSE can be avoided.

The approximation of the statistics of stochastic processes by the MCM can be used in a variety of applications. In particular, parameter estimation, experimental design and control of stochastic processes can be rendered more efficient [Zechner et al., 2012].

ACKNOWLEDGEMENTS

The authors would like to thank Justin Feigelman and Fabian Fröhlich for proofreading the article. The authors acknowledge financial support from the German Federal Ministry of Education and Research (BMBF) within the Virtual Liver project (Grant No. 0315766) and LungSys II (Grant No. 0316042G), and the European Union within the ERC grant "LatentCauses".

REFERENCES

A. Eldar and M. B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(9):1–7, Sept. 2010. doi: 10.1038/nature09326.

S. Engblom. Computing the moments of high dimensional solutions of the master equation. *Appl. Math. Comp.*, 180:498–515, 2006.

W. Feller. On the integro-differential equation of purely discontinous Markoff processes. *Trans. of the American Mathematical Society*, 48:4885–4915, 1940.

C. W. Gardiner. *Handbook of stochastic methods: For physics, chemistry and natural sciences.* Springer Series in Synergetics. Springer Berlin / Heidelberg, 4th edition, 2011.

C. S. Gillespie. Moment-closure approximations for mass-action models. *IET Syst. Biol.*, 3(1):52–58, 2009.

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

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

J. Hasenauer, V. Wolf, A. Kazeroonian, and F. J. Theis. Method of conditional moments (MCM) for the chemical master equation. *J. Math. Biol.*, pages 1–49, 2013.

A. Hellander and P. Lötstedt. Hybrid method for the Chemical Master Equation. *J. Comput. Phys.*, 227:100–122, 2007.

T. A. Henzinger, L. Mikeev, M. Mateescu, and V. Wolf. Hybrid numerical solution of the chemical master equation. In *Proceedings of the 8th International Conference on Computational Methods in Systems Biology*, pages 55–65, New York, NY, USA, 2010. ACM.

T. Jahnke. On reduced models for the chemical master equation. *Multiscale Model. Simul.*, 9(4):1646–1676, 2011.

T. Jahnke and W. Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *J. Math. Biol.*, 54(1):1–26, 2007.

I. Lestas, G. Vinnicombe, and J. Paulsson. Fundamental limits on the suppression of molecular fluctuations. *Nat.*, 467:174–178, Sept. 2010.

S. Menz, J. C. Latorre, C. Schütte, and W. Huisinga. Hybrid stochastic deterministic solution of the Chemical Master Equation. *SIAM J. on Multiscale Model. Simul.*, 10(4):1232—1262, 2012.

P. Milner, C. S. Gillespie, and D. J. Wilkinson. Moment closure approximations for stochastic kinetic models with rational rate laws. *Mathematical Biosciences*, 231 (2):99 – 104, 2011.

B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.*, 124(4):044104, 2006.

B. Munsky, B. Trinh, and M. Khammash. Listening to the noise: random fluctuations reveal gene network parameters. *Mol. Syst. Biol.*, 5(318), 2009.

B. Munsky, G. Neuert, and A. von Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, 2012.

J. M. Raser and E. K. O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–1814, 2004.

R. Sidje, K. Burrage, and S. MacNamara. Inexact uniformization method for computing transient distributions of Markov chains. *SIAM J. Sci. Comput.*, 29(6): 2562–2580, 2007.

A. Singh and J. P. Hespanha. Approximate moment dynamics for chemically reacting systems. *IEEE Trans. Autom. Control*, 56(2):414–418, 2011.

Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X.S. Xie. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, 2010.

N. G. van Kampen. *Stochastic processes in physics and chemistry.* North-Holland, Amsterdam, 3rd revised edition, 2007.

C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl. Moment-based inference predicts bimodality in transient gene expression. *Proc. Nati. Acad. Sci. USA*, 109(21):8340–8345, 2012.

Appendix C

CERENA: ChEmical REaction Network Analyzer – A Toolbox for the Simulation and Analysis of Stochastic Chemical Kinetics.

PLOS ONE, 2016.

This is an article published in PLOS ONE following peer review. The version of record Kazeroonian A, Fröhlich F, Raue A, Theis FJ, Hasenauer J. **CERENA: ChEmical REaction Network Analyzer – A Toolbox for the Simulation and Analysis of Stochastic Chemical Kinetics.** *PLOS ONE 11(1): e0146732, January 2016.*
is available online at:
https://doi.org/10.1371/journal.pone.0146732

PLOS ONE

RESEARCH ARTICLE

CERENA: ChEmical REaction Network Analyzer—A Toolbox for the Simulation and Analysis of Stochastic Chemical Kinetics

Atefeh Kazeroonian[1,2], Fabian Fröhlich[1,2], Andreas Raue[3], Fabian J. Theis[1,2], Jan Hasenauer[1,2]*

1 Institute of Computational Biology, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany, **2** Department of Mathematics, Chair of Mathematical Modeling of Biological Systems, Technische Universität München, Garching, Germany, **3** Merrimack Pharmaceuticals Inc., Discovery Devision, Cambridge, MA 02139, United States of America

* jan.hasenauer@helmholtz-muenchen.de

Citation: Kazeroonian A, Fröhlich F, Raue A, Theis FJ, Hasenauer J (2016) CERENA: ChEmical REaction Network Analyzer—A Toolbox for the Simulation and Analysis of Stochastic Chemical Kinetics. PLoS ONE 11(1): e0146732. doi:10.1371/journal.pone.0146732

Editor: Dennis Salahub, University of Calgary, CANADA

Received: September 30, 2015

Accepted: December 21, 2015

Published: January 25, 2016

Data Availability Statement: The code for CERENA toolbox is freely available from http://cerenadevelopers.github.io/CERENA/.

Funding: This work was funded by the European Union within the ERC grant "Latent Causes" (AK, FJT), and the German Research Foundation (DFG) through the Graduate School of Quantitative Biosciences Munich (FF). Merrimack Pharmaceuticals Inc. provided support in the form of salaries for AR, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Gene expression, signal transduction and many other cellular processes are subject to stochastic fluctuations. The analysis of these stochastic chemical kinetics is important for understanding cell-to-cell variability and its functional implications, but it is also challenging. A multitude of exact and approximate descriptions of stochastic chemical kinetics have been developed, however, tools to automatically generate the descriptions and compare their accuracy and computational efficiency are missing. In this manuscript we introduced CERENA, a toolbox for the analysis of stochastic chemical kinetics using Approximations of the Chemical Master Equation solution statistics. CERENA implements stochastic simulation algorithms and the finite state projection for microscopic descriptions of processes, the system size expansion and moment equations for meso- and macroscopic descriptions, as well as the novel conditional moment equations for a hybrid description. This unique collection of descriptions in a single toolbox facilitates the selection of appropriate modeling approaches. Unlike other software packages, the implementation of CERENA is completely general and allows, e.g., for time-dependent propensities and non-mass action kinetics. By providing SBML import, symbolic model generation and simulation using MEX-files, CERENA is user-friendly and computationally efficient. The availability of forward and adjoint sensitivity analyses allows for further studies such as parameter estimation and uncertainty analysis. The MATLAB code implementing CERENA is freely available from http://cerenadevelopers.github.io/CERENA/.

Introduction

Biological processes, including chemical reaction networks, are dynamical systems with inherently stochastic dynamics due to the discrete nature of matter [1]. The kinetics of these processes are described by continuous-time Markov chains and can be simulated using stochastic

The specific roles of these authors are articulated in the "author contributions" section.

Competing Interests: The employment of AR by Merrimack Pharmaceuticals Inc. does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

simulation algorithms (SSAs) [2]. The impact of stochastic fluctuations is more pronounced in low copy-number regimes [3] and tends to decrease, but possibly remaining important, as copy-numbers increase [4]. Given the importance of stochasticity in the dynamics of biological systems, e.g., cellular mechanisms and their functions [5], a holistic understanding of cell biology requires accurate capturing of stochastic effects.

Using well-mixed and thermal equilibrium assumptions, the dynamics of chemical reaction networks is exactly described by the Chemical Master Equation (CME) [6]. The solution of the CME yields the probability distribution over the state of the system [1]. Besides special cases [7], the exact solution of the CME is mostly infeasible as CMEs are usually infinite-dimensional systems of differential equations. Several approaches have been developed to approximate the solution of the CME, amongst others variants of finite state projection (FSP) [8–10]. However, high computational complexity is a limiting factor for the applicability of this class of simulation methods.

To reduce computational complexity, a multitude of approaches have been introduced that, instead of approximating the full probability distribution, focus on the statistical moments of it. Various orders of the method of moments (MM) [11] and the system size expansion (SSE) [1, 12] provide information about the mean and higher-order moments of the distribution. These methods yield the reaction rate equations (RRE) as a special case. To improve upon the approximation in the presence of low as well as high copy-number species, hybrid microscopic-mesoscopic approaches such as the method of conditional moments (MCM) [13] and the conditional linear noise approximations [14] have been introduced. All these methods are of reduced computational complexity as they possess significantly fewer state-variables compared to the CME or FSP, thus remain feasible for real-world application problems.

Beyond fast numerical simulation, moment-based descriptions facilitate parameter estimation and model selection for stochastic processes [15, 16]. This is essential for inferring unknown rate constants and pathway topologies from experimental data. In addition to the approximative model, state-of-the-art estimation algorithms strongly benefit from the solution of sensitivity equations [17].

Several well-known open-source software packages are available for stochastic simulations, finite state projection, method of moments, and system size expansion (e.g., [18–26] whose properties are summarized in Fig 1). In addition, there exist web-based simulation platforms, e.g., SHAVE [27]. However, a software package offering a broad collection of simulation methods is still missing. Furthermore, none of the available software provides sensitivity equations, or hybrid approaches such as the method of conditional moments.

In this paper, we introduce CERENA (ChEmical REaction Network Analyzer), a toolbox for the analysis of stochastic chemical kinetics. CERENA includes a variety of methods for the analysis of stochastic biochemical reaction networks, focusing on mesoscopic and macroscopic descriptions, namely RRE, MM, and SSE. Also, CERENA provides the first implementation of MCM, and offers a wide range of options, amongst others variable truncation orders and different closure schemes. In addition, FSP and SSA implemented in CERENA can be used to provide microscopic descriptions of stochastic chemical kinetics. Although efficient implementations of many variants of SSA are available, e.g., in StochKit [18], CERENA is the only package supporting arbitrary, including fast-varying, time-dependent reaction propensities [28]. This variety of descriptions renders CERENA unique compared to other relevant software packages (see Fig 1). To improve applicability of CERENA for realistic systems, the toolbox allows for multiple compartments, non-mass action kinetics, and time-dependent propensities. CERENA is the first toolbox for stochastic modeling to provide forward and adjoint sensitivity equations to facilitate efficient parameter estimation when linked to optimization

				StochKit [18]	StochPy [19]	Dizzy [20]	CMEpy [21]	Copasi [22]	MomentClosure [23]	StochDynTools [24]	MOCA [25]	iNA [26]	CERENA
Methods	Microscopic	SSA	time-independent propensities	✓	✓	✓	-	✓	-	-	-	✓	✓
			time-dependent propensities	-	-	✓†	-	-	-	-	-	-	✓
			FSP	-	-	-	✓	-	-	-	-	-	✓
	Hybrid	MCM	Low Dispersion	-	-	-	-	-	-	-	-	-	✓
			Zero Cumulants	-	-	-	-	-	-	-	-	-	✓
			Mean Field	-	-	-	-	-	-	-	-	-	✓
			Derivative Matching	-	-	-	-	-	-	-	-	-	✓
			User-defined closure	-	-	-	-	-	-	-	-	-	✓
	Macroscopic and Mesoscopic		RRE	-	-	✓	-	✓	-	-	✓	✓	✓
		SSE	LNA	-	-	-	-	✓	-	✓	-	✓	✓
			EMRE	-	-	-	-	-	-	-	-	✓	✓
			IOS	-	-	-	-	-	-	-	-	✓	✓
		MM	Low Dispersion	-	-	-	-	-	-	✓	-	-	✓
			Zero Cumulants	-	-	-	-	-	✓	✓	✓	-	✓
			Mean Field	-	-	-	-	-	-	-	-	-	✓
			Derivative Matching	-	-	-	-	-	-	✓	✓	-	✓
			Quasi Deterministic	-	-	-	-	-	-	✓	-	-	-
			Poisson	-	-	-	-	-	-	-	✓	-	-
			User-defined closure	-	-	-	-	-	-	-	✓	-	✓
Features	Sensitivity Analysis		Forward Sensitivity	-	-	-	-	✓‡	-	-	-	-	✓*
			Adjoint Sensitivity	-	-	-	-	-	-	-	-	-	✓**
	SBML		Import	✓	✓	✓	-	✓	✓	-	-	✓	✓
	Kinetics		Mass Action	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
			Non-Mass Action	✓	✓	✓	✓	✓	-	-	✓	✓	✓
	Equations		Accessibility	-	-	-	✓	✓	✓	✓	-	-	✓

†Dizzy technically allows for the definition of time-dependent propensities, but performs an approximate simulation which assumes that the propensities are constant between reaction events. Therefore, only in cases where the time-dependent propensities vary slowly, the stochastic simulators implemented in Dizzy may provide compatible results [20].

‡Copasi supports the calculation of parameter sensitivities via the method of finite differences.

*CERENA supports the calculation of parameter sensitivities for FSP, MCM, RRE, SSE and MM using forward sensitivity equations.

**CERENA supports the calculation of parameter sensitivities for FSP, RRE, SSE and MM using adjoint sensitivity equations.

Fig 1. Overview of software packages for stochastic modeling and their capabilities.

doi:10.1371/journal.pone.0146732.g001

packages. Ensuring efficient numerical simulation, CERENA enables comprehensive studies for a variety of meso- and macroscopic descriptions.

In the following, we describe the functionality of CERENA and introduce the different approximations. CERENA is then used for a detailed quantitive comparison of different approximation methods, including various moment closures, which was not done before. In particular, the approximation accuracies and computation times are assessed, demonstrating the efficiency of the CERENA implementation.

Methods

In the following, several methods for the modeling of stochastic processes and the corresponding sensitivity analysis are briefly introduced.

Modeling approaches for stochastic biochemical reaction networks

A chemical reaction network, comprising of n_s chemical species and n_r chemical reactions is described using a continuous-time Markov chain (CTMC) [29]. The state vector of this CTMC, $\mathbf{X} \in \mathbb{N}_0^{n_s}$, represents the counts of species, and is changed every time a reaction fires. The probability of observing the CTMC at a particular state $\mathbf{x}$ at time t is denoted by $p(\mathbf{x}|t)$. The time evolution of the probability distribution $p(\mathbf{x}|t)$ is governed by the CME, which is a system of ordinary differential equations (ODEs) (see S1 CERENA Documentation for more details). As solving the CME is mostly infeasible due to the large or infinite number of states $\mathbf{x}$, various approximative methods have been developed. Several methods concentrate on the full distribution $p(\mathbf{x}|t)$ to provide a microscopic description. For mesoscopic and macroscopic descriptions, there exist several methods that focus on representing the solution of the CME in terms of its statistical moments. The microscopic, mesoscopic and hybrid methods implemented in CERENA are briefly introduced in the following.

Stochastic Simulation Algorithm. SSAs generate statistically representative sample paths of the CTMC [2]. An estimate to the probability distribution $p(\mathbf{x}|t)$ is given by the frequency of sample paths that occupy state $\mathbf{x}$ at time t. To estimate the moments of the process, Monte-Carlo integration can be performed. While estimators for probability distribution and moments are unbiased and converge, the sample-sizes required to obtain low-variance estimates are generally large, rendering SSA-based methods computationally demanding.

Finite State Projection. To enable a direct approximation of $p(\mathbf{x}|t)$, FSP [8] reduces the number of state variables of the CME by only considering the states of non-negligible probabilities. The remaining set of ODEs then yields a lower bound for $p(\mathbf{x}|t)$. Growing the state-space of FSP decreases the approximation error at the cost of increased computational complexity.

Reaction Rate Equations. The RRE is the most commonly used modeling approach for biochemical reaction networks. It constitutes a system of ODEs for the time evolution of the mean of the stochastic process in the macroscopic limit. For reaction networks with constant and linear propensities, i.e. those with only zero- or monomolecular reactions, the solution of the RRE is exactly the mean of the stochastic process. For reaction networks with nonlinear propensities, the RRE prediction can be considerably different from the true mean of the process since it neglects the stochastic effects. In such cases, the solution of RRE is reflective of the true mean of the stochastic process only in the limit of large molecule numbers [30].

System Size Expansion. For a systematic approximation of the dynamics of mesoscopic systems, the SSE has been introduced [1]. The SSE is a power series expansion of the CME in the inverse volume of the system. The lowest-order approximation for the mean reproduces the aforementioned RRE. For the covariance, the lowest-order approximation yields the well-known linear noise approximation (LNA), whose validity has been studied in [30] for different classes of reaction systems. Higher-order corrections for the mean and covariance yield the effective mesoscopic rate equation (EMRE) [12] and the inverse omega square (IOS) approximation [26].

Method of Moments. The method of moments (MM) [11] is conceptually similar to SSE in that it also sets a framework for describing the moments of the solution of the CME. A system of ODEs for the exact time evolution of the moments, which constitutes the moment equations, can be derived from the CME. Generally, the equations for the lower-order moments depend on the higher-order moments, rendering moment closure necessary. Commonly used

closure techniques include low dispersion closure, mean field closure, zero cumulants closure, and derivative matching closure [31]. The application of moment closure yields a closed set of approximative equations for the time evolution of the moments.

Method of Conditional Moments. The MCM [13] combines a microscopic description of low copy-number species with a moment-based description of high copy-number species, providing a hybrid approach for approximating the solution of the CME. Since stochastic fluctuations are more dominant for low copy-number species, marginal probability densities for these species are determined. The high-copy number species are merely described in terms of their moments, conditioned on the state of low-copy number species. The MCM equations are derived from the CME, and form a system of differential algebraic equations (DAEs). Similar to the moment equations, the moment closure is generally required to close the set of MCM equations. This hybrid description can yield an improved approximation accuracy [13].

Sensitivity analysis

FSP, RRE, SSE, MM and MCM yield systems of differential equations. The parameters of differential equations can efficiently be inferred using gradient-based optimization methods [17]. While gradients can be approximated using finite differences, methods based on sensitivity equations are known to be more robust and computationally more efficient [17]. CERENA enables first- and second-order forward sensitivity analysis for all ODE-based and DAE-based modeling approaches, as well as adjoint sensitivity analysis [32] for all ODE-based modeling approaches.

Forward sensitivity equations. Forward sensitivity equations provide the time-dependent sensitivity of the state-variables of the differential equations with respect to the parameters. Assuming that the model possesses n state-variables and n_θ parameters, roughly a system of $n(1+n_\theta)$ differential equations is solved to compute the first-order state sensitivities with respect to all parameters. The sensitivity of measured quantities and objective functions can then be computed based on state sensitivities.

Adjoint sensitivity equations. If the sensitivity of few functions with respect to many parameters is required, computing the state sensitivities is unnecessarily demanding. In this case, the adjoint sensitivity equations [32] can be solved to yield a set of adjoint states which are independent of the parameters. These trajectories are then used to calculate the sensitivity with respect to any parameters of interest, with low computational cost. Thus, in applications with high-dimensional parameter spaces and/or few output functionals, calculating adjoint sensitivities tends to be computationally more advantageous. In parameter estimation, the likelihood function can be defined as the sole output functional of the system.

Implementation

CERENA is a MATLAB-based toolbox for the simulation of chemical reaction networks. It provides a collection of methods for the analysis of stochastic processes, focusing on SSE, MM and MCM of various orders. In addition, FSP and SSAs are implemented in CERENA to provide microscopic descriptions of the process, and can also be used to assess the approximation errors of the aforementioned methods. The workflow of the toolbox is laid out in Fig 2. In the following, different aspects of implementation and features of the toolbox are explained. For a detailed list of functions, we refer to the S1 CERENA Documentation. The CERENA toolbox is freely available from http://cerenadevelopers.github.io/CERENA/.

Network specification

To use CERENA, the biochemical reaction network has to be defined in a specific format described in the S1 CERENA Documentation. The definition includes species, compartments,

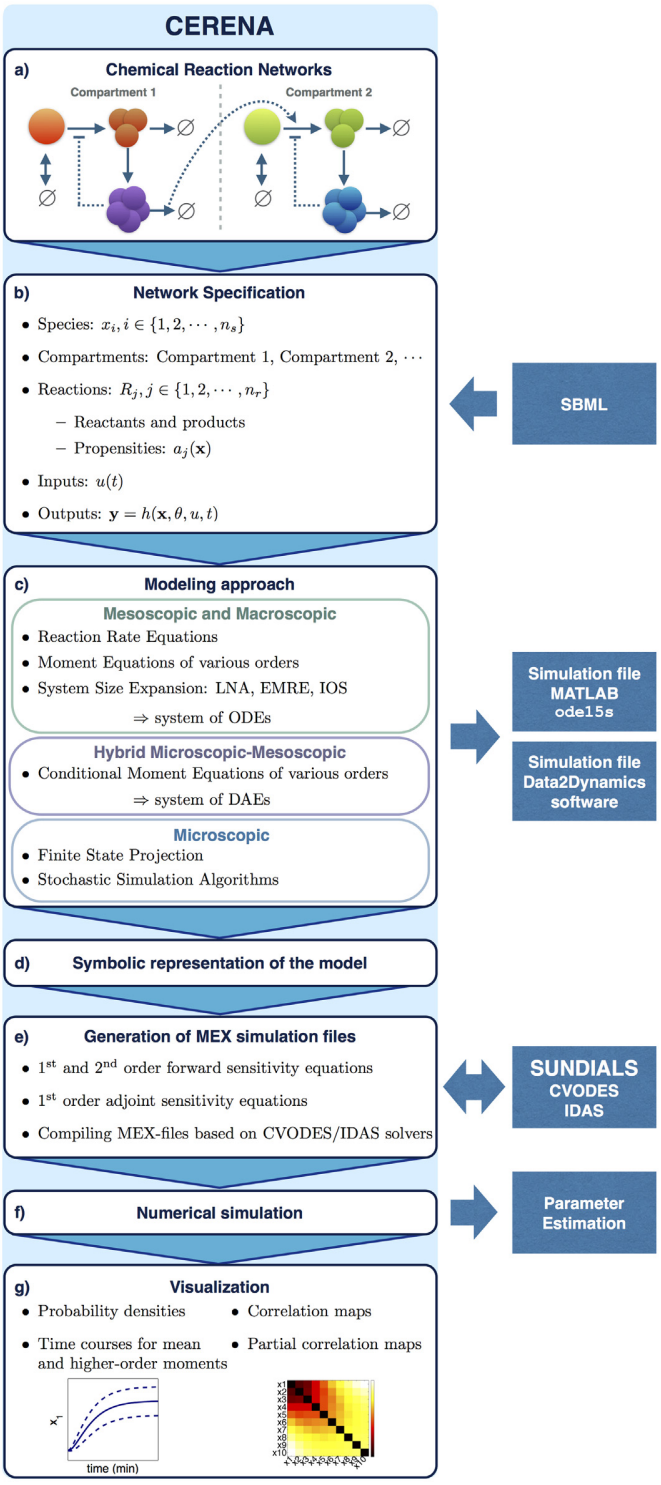


Fig 2. Workflow of CERENA. (a) CERENA can be used to study (multi-compartment) chemical reaction networks. (b) The reaction network can be defined in MATLAB, or alternatively, imported from SBML. (c) The system of equations for different modeling approaches implemented in CERENA is generated, and optionally stored as MATLAB functions for numerical simulation using MATLAB ODE solvers. Furthermore, the representation of the system can be exported to the estimation toolbox Data2Dynamics. (d) The symbolic representation of the system of equations together with the initial conditions is stored in a MATLAB script. (e)

Based on the symbolic representation, 1st and 2nd order sensitivity equations are derived. MEX-files, which use CVODES and IDAS packages of SUNDIALS for the numerical simulation of the models, are compiled. (f) The generated MEX-files are used for numerical simulation, and can be integrated with other software for parameter estimation. (g) Various aspects of the simulation results can be visualized using CERENA.

doi:10.1371/journal.pone.0146732.g002

reactions and their propensities, inputs and observables of the system. Reaction propensities can be time-dependent and may or may not follow the law of mass action. In case of non-mass action kinetics the propensities are approximated, e.g., using Taylor series expansion in MM and MCM [33]. Inputs are used to describe experimental conditions. Alternatively, networks described in the Systems Biology Markup Language (SBML) can be imported.

Model derivation and symbolic representation

Following the definition of the biochemical reaction network, a modeling approach and corresponding options, such as approximation order and moment closure technique, can be selected. In addition to the moment closure techniques implemented in CERENA (see Fig 1), user-defined closures can be provided. In case of MM and MCM, it can be specified whether the equations in terms of molecule numbers or concentrations are to be derived. A system of equations corresponding to the selected modeling approach is then derived, and provided as a MATLAB script file including the corresponding initial conditions. This symbolic representation is the basis for the rest of the simulation and analysis. CERENA extensively uses MATLAB Symbolic Math Toolbox for a variety of symbolic manipulations including symbolic differentiation, e.g., in the calculation of Jacobian matrices used to accelerate the numerical simulation.

The models can be exported to Data2Dynamics software [34] for parameter estimation and model selection. In addition, an optional intermediate MATLAB function can be generated for the numerical simulation of the symbolic equations using MATLAB ODE solver `ode15s`.

Derivation of sensitivity equations and numerical integration

Forward and adjoint sensitivity equations for the selected model are derived based on the aforementioned symbolic representation. The complete symbolic representation can then be used to compile simulation files. CERENA uses CVODES and IDAS solvers of the SUNDIALS package [32] which are C implementations of solvers suited for efficient numerical integration of stiff ODEs and DAEs. Although the SUNDIALS package provides a MATLAB interface to the C solvers, the governing equations must be specified as MATLAB code, which adds an overhead to the overall computational cost of numerical simulation. To ensure efficiency, wrappers for CVODES and IDAS, which compile model-specific MEX-files from automatically generated native C code, have been implemented in CERENA. The compiled MEX-files are used for the numerical simulation of the system with given parameter values and time vector. Options for the numerical solvers and sensitivity analysis can be specified as inputs to the MEX simulation files. For efficient numerical simulation, essential capabilities of the SUNDIALS package can be exploited. The compiled MEX-files can be used for subsequent analysis.

Stochastic simulations

The solvers based on differential equations are complemented by SSAs, e.g. to provide reference solutions. In the case of SSAs, realizations of the stochastic process are simulated. CERENA implements next-reaction methods for constant [2] and time-dependent propensities [28]. To the best of our knowledge, an implementation of the modified next-reaction method for systems with time-dependent propensities and delays is not available in other software

packages. This method, implemented in CERENA, is exact for reaction networks with time-dependent propensities whose antiderivatives are available in closed-form. Otherwise, a numerical integration error is introduced. This error can be controlled by adjusting the integration error tolerance of the respective numerical solvers.

Visualization

To facilitate the interpretation of the numerical simulation results, CERENA offers various visualization routines. Time courses for stochastic realizations, as well as mean and higher-order moments of species, can be plotted. Moreover, the full and marginal probabilities can be visualized for SSA, FSP and MCM. To illustrate the interaction between different network components and propagation of stochasticity, correlation and partial correlation maps, including movies of these maps over time, are provided.

Application

In this Section, we present two biological models to demonstrate different features of CERENA, including the improved computational complexity. Furthermore, we exploit the comprehensiveness of CERENA to compare different approximative descriptions.

Three-stage gene expression model

As the first example, we consider the generalized three-stage model of gene expression [3] depicted in Fig 3(a). This model includes a gene with a promotor switching between on- and off-states. Transcription of mRNA takes place if the promotor is in the on-state, and the transcribed mRNA can be translated into protein. The model also incorporates a protein-induced activation of the promoter which establishes a positive feedback loop. Protein and mRNA are subject to degradation. The combination of low-copy number species (the gene) and medium/high-copy number species (mRNA and protein) makes this model an interesting simulation test example.

Comprehensive comparison of approximation accuracy. The accuracy of various approximative descriptions is problem-specific, and therefore, comparisons of different descriptions for a process of interest is interesting in different applications. As demonstrated for this model, CERENA offers an easy-to-use framework for such a comprehensive comparison, thanks to its broad collection of simulation methods.

This process was implemented and simulated in CERENA for the parameter values given in S1 CERENA Documentation, Chapter 1, Table E. Fig 3(b) depicts the simulation results for the mean and the variance of the number of protein molecules obtained using various methods. All methods yield results which agree well with the reference solution, obtained using FSP. The RRE deviates the most from the reference solution. This behavior is expected, especially when the abundance of species is low, as RRE merely provides a macroscopic description of the stochastic process.

As mRNA is only transcribed if the promotor is in the on-state, the conditional distributions of mRNA and protein counts in the on- and off-states differ. These differences are captured by the MCM (Fig 3(c)), which provides information about the probability of different promotor states and the moments of the corresponding conditional distributions of the counts of mRNA and protein.

The accuracy of different descriptions is quantified in terms of the relative errors of the mean and variance with respect to the FSP, e.g., $|\mu_{\mathrm{MCM}}-\mu_{\mathrm{FSP}}|/\mu_{\mathrm{FSP}}$. Fig 4 displays the relative errors of MM and MCM close to steady state (t = 100), for various truncation orders and moment closures. For derivative matching closure, we find that the resulting ODE model

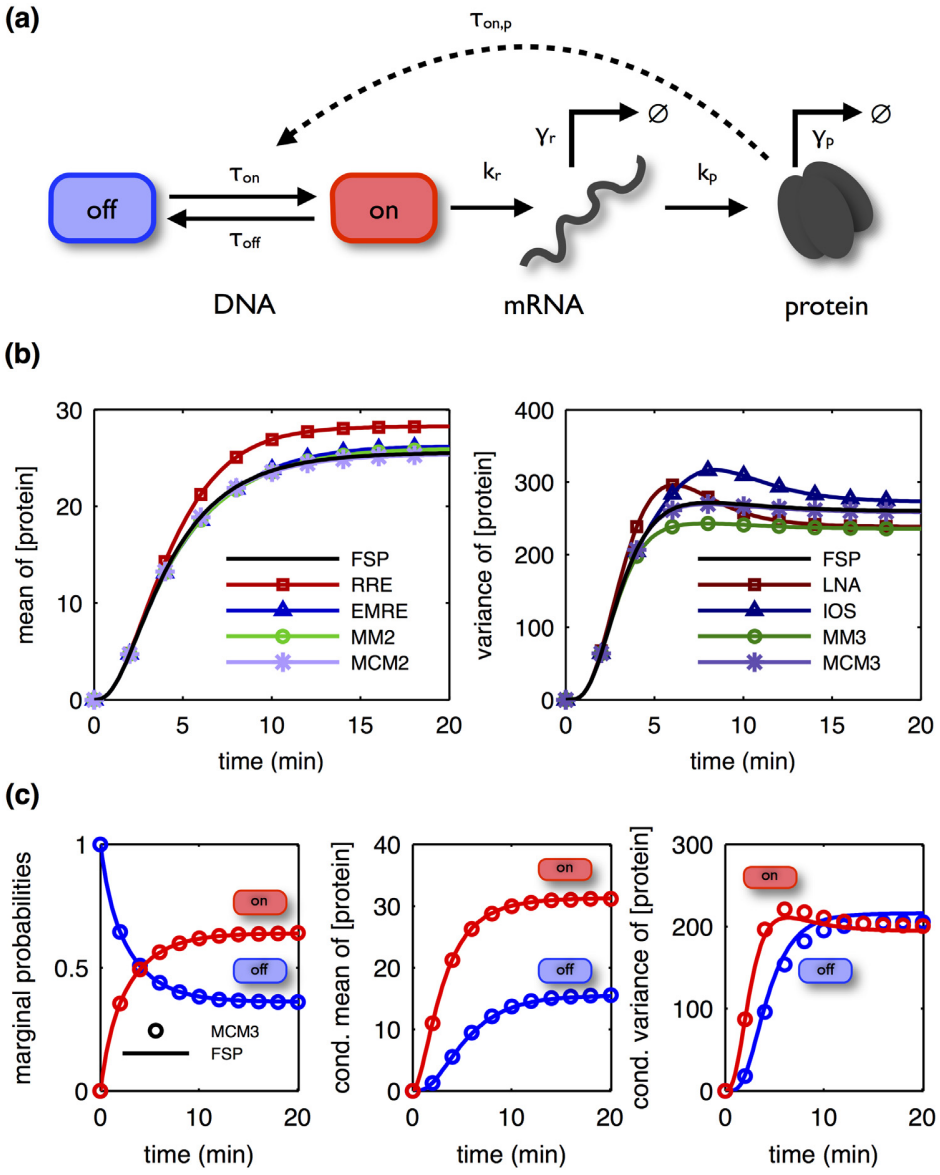


Fig 3. Simulation of the three-stage model of gene expression. (a) Schematic of the three-stage model of gene expression. (b) Mean (left) and variance (right) of the number of protein molecules obtained using different orders of SSE, MM and MCM. (c) Marginal probabilities of promotor states (left), the mean of protein molecule numbers conditioned on the promotor state (middle), and the variance of protein molecule numbers conditioned on the promotor state (right) predicted by MCM of order 3. (b,c) FSP results serve as the reference solution. Low dispersion closure was used for MM and MCM. MM2, MM3, MCM2 and MCM3 denote the second- and third-order MM and the second- and third-order MCM.

doi:10.1371/journal.pone.0146732.g003

cannot be simulated robustly as it diverges for several truncation orders. It is observed that the contribution of higher-order moments tends to enhance the simulation accuracy of lower-order moments. The influence of truncation order on the accuracy varies for different closure schemes.

Improved computational efficiency. A key bottleneck in the analysis of stochastic chemical kinetics is the computational complexity of the numerical simulation. As the number of

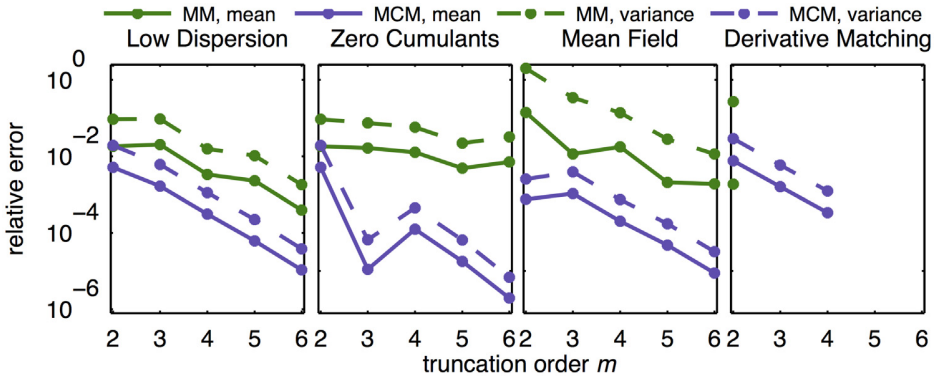


Fig 4. Approximation error of MM and MCM of various orders with various moment closures for the three-stage model of gene expression. Relative errors of mean and variance of the protein concentration at the steady state are depicted for different truncation orders and moment closures. The truncation order *m* means that moments up to order *m* are simulated. For moment orders and closures for which the numerical simulation could not be completed, i.e. derivative matching, no approximation error is reported.

doi:10.1371/journal.pone.0146732.g004

biochemical species or the approximation order increases, the system of differential equations to be solved becomes larger (Fig 5, top panel), indicating the need for efficient numerical simulation schemes. Since the FSP describes the full probability distribution, its system of equations is several orders of magnitude larger than the rest of the methods which merely capture a few moments of the probability distribution (Fig 5, top panel).

We assessed the computation time for implementations in CERENA and compared it to other packages/implementations (Fig 5, bottom panel). It is evident that the combination of

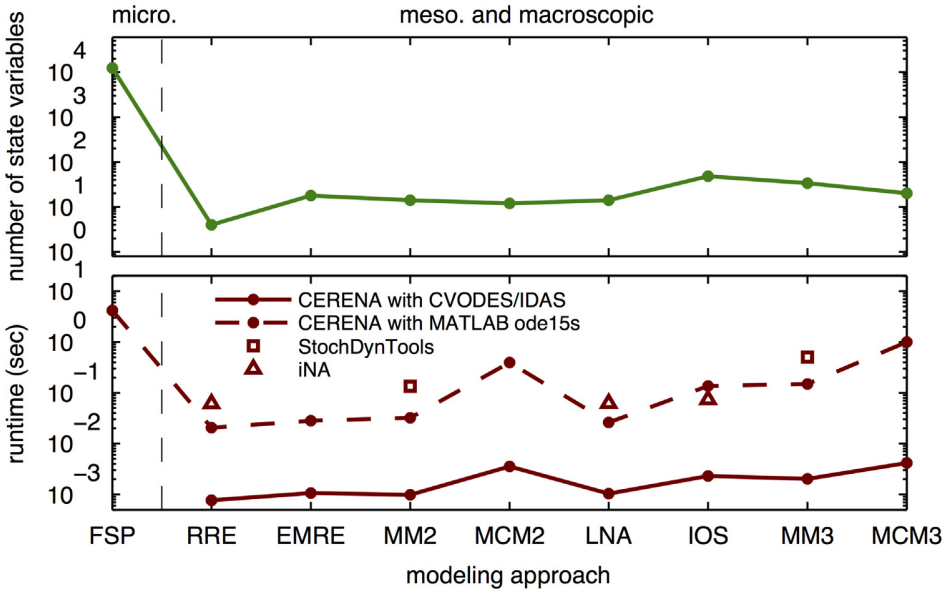


Fig 5. Complexity of different descriptions of the three-stage model of gene expression. Number of state-variables (top) and computation time (bottom). Runtimes are shown for the numerical simulation using CVODES/IDAS wrappers implemented in CERENA and MATLAB solver `ode15s`, as well as for StochDynTools. The computation times were calculated by averaging over at least 10 simulations. For MM and MCM low dispersion closure was used.

doi:10.1371/journal.pone.0146732.g005

CVODES and IDAS packages with corresponding wrappers in CERENA resulted in remarkable speedup, around 10–100 fold, compared to the use of standard MATLAB ODE-solvers, e.g., `ode15s`. Also, other toolboxes, e.g., StochDynTools and iNA, were outperformed by CERENA. A comparison across different methods reveals that the simulation of higher-order descriptions which possess more state-variables tends to be computationally more demanding than the simulation of lower-order descriptions.

JAK-STAT signaling pathway

The second example studied using CERENA is a model of the JAK-STAT signaling pathway introduced by [35]. The model, sketched in Fig 6(a), describes the signaling cascade of STAT protein. Upon activation, the Epo receptor triggers the phosphorylation of cytoplasmic STAT. Dimerization and translocation of phosphorylated STAT into the nucleus, followed by a

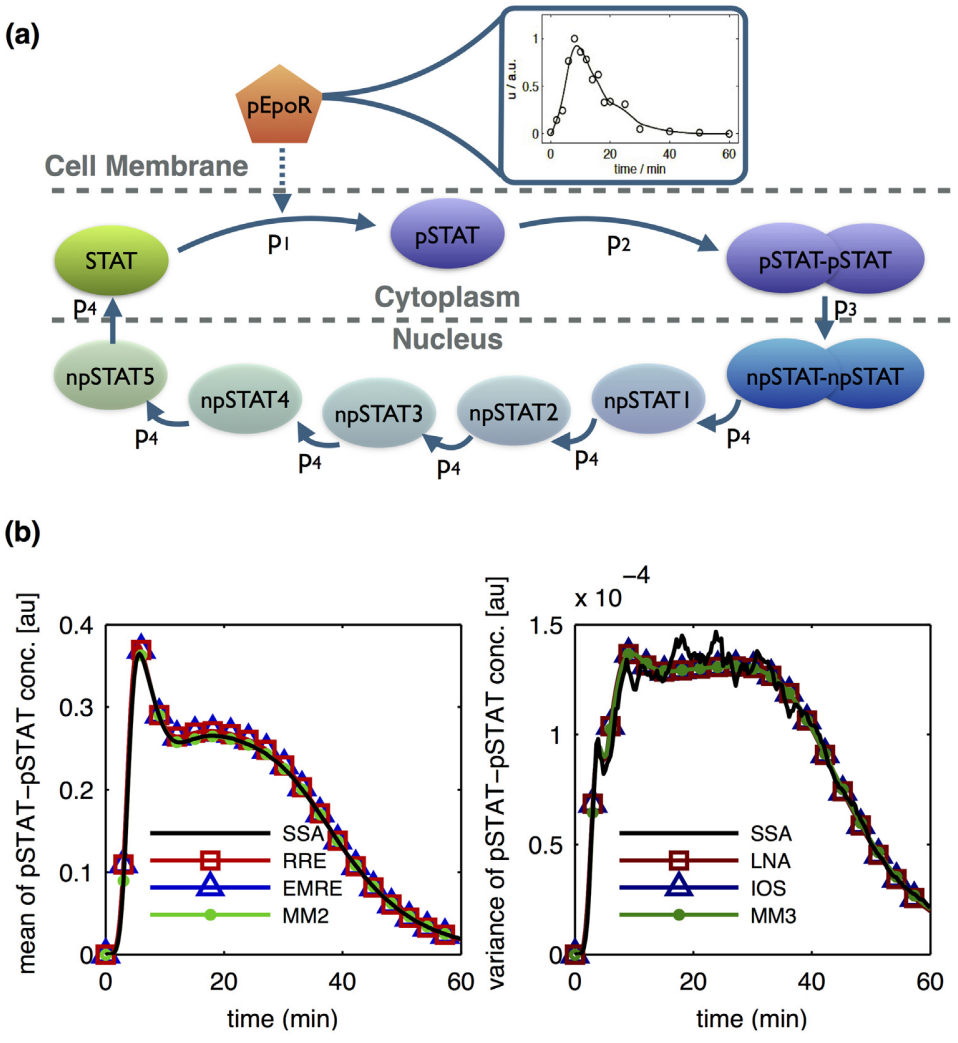


Fig 6. Simulation results for the JAK-STAT signaling pathway. (a) Schematic of the simplified JAK-STAT signaling pathway. The intermediate states npSTAT1 to npSTAT5 are used to model the delayed export of STAT from the nucleus. (b) The mean (left) and variance (right) of dimerized phosphorylated STAT concentration, obtained using several methods. SSA simulation results serve as the reference solution.

doi:10.1371/journal.pone.0146732.g006

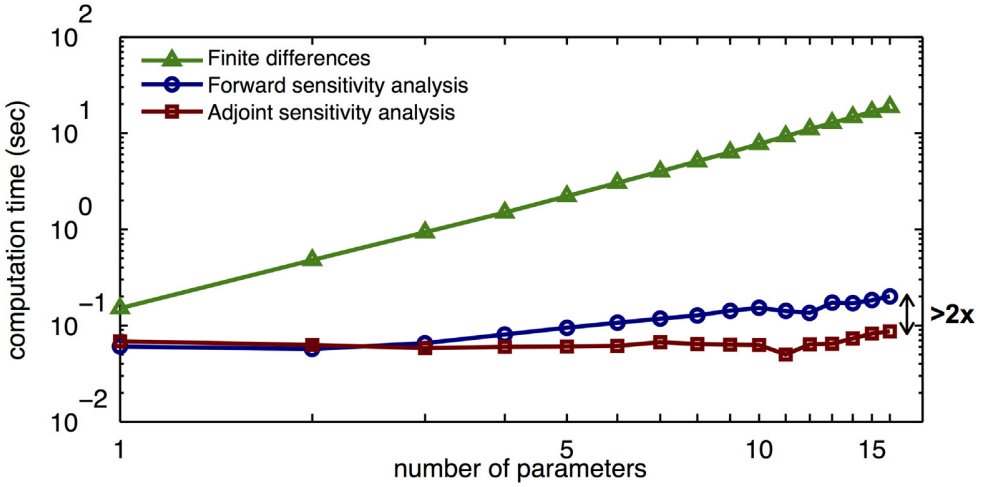


Fig 7. Computation time for different sensitivity analysis methods. The objective function gradient for MM2 simulation is evaluated for an increasing number of parameters. The computation times of finite differences, forward sensitivity analysis, and adjoint sensitivity analysis are shown.

doi:10.1371/journal.pone.0146732.g007

delayed export of STAT from the nucleus complete the loop. The time-dependent concentration of phosphorylated Epo receptor, [pEpoR], functions as an input to the system. The experimental data for the concentration of phosphorylated Epo receptor, cytoplasmic STAT and phosphorylated cytoplasmic STAT are available from previous studies [36].

The JAK-STAT signaling pathway is an interesting application example as it (i) includes two compartments, namely cytoplasm and nucleus, and (ii) involves a time-dependent propensity.

Simulation of multi-compartment systems with time-dependent propensities. We used CERENA to describe the dynamics of JAK-STAT signaling pathway for parameter values given in S1 CERENA Documentation, Chapter 8, Table A. As the copy numbers are relatively high in this pathway, MCM and FSP were not considered. To provide the reference solution, the modified next-reaction method for systems with time-dependent propensities implemented in CERENA was used which enabled handling of the time-dependent input. As seen in Fig 6(b), all methods showed the same qualitative behavior as the reference solution.

Comparison of sensitivity analysis methods. In previous studies, it was shown that the parameters of the JAK-STAT signaling pathway can be estimated efficiently for RRE [35] and EMRE and second-order MM descriptions [37]. These studies used gradient-based optimization methods with gradients being computed using forward sensitivity analysis. Here, we considered a weighted least-squares objective function as used by [37], and compared the performance of finite differences, forward and adjoint sensitivity analyses in gradient calculation for second- and third-order moment equations.

We observed that, even for a small number of parameters, a gain in efficiency is achieved by using forward and adjoint sensitivity analysis methods instead of finite differences (Fig 7). Moreover, the adjoint sensitivity analysis has the best scalability with respect to the number of parameters.

Discussion

A multitude of studies revealed the functional role of cell-to-cell variability in cellular mechanisms [5]. Hence, the analysis of cell-to-cell variability and its implications is crucial for a

holistic understanding of biological systems, indicating the need for corresponding efficient simulation tools. In this work, we introduced CERENA, a user-friendly toolbox for the study of stochastic biological processes. CERENA offers a broad collection of simulation methods for micro-, meso- and macroscopic description of stochastic processes, rendering it unique compared to other software packages. In addition to various orders of the system size expansion and moment equations, the first implementation of the method of conditional moments is provided. CERENA attains generality not only method-wise, but also by imposing the least restrictions on the biological systems. Specifically, (regulatory) processes involving non-mass action kinetics, and/or time-dependent propensities can be analyzed. CERENA is one of the few packages to provide an SSA for the latter case. A key feature, distinguishing CERENA from all other packages for stochastic modeling, is the implementation of forward and adjoint sensitivity analyses for robust and efficient gradient calculations, especially in applications with high-dimensional parameter spaces. This enables feasible gradient-based optimization. To improve the computational efficiency, CERENA uses SUNDIALS solvers to compile numerical simulation MEX-files.

We used CERENA for detailed quantitative comparisons of different modeling approaches on models for three-stage gene expression and Epo-induced JAK-STAT signaling. These applications demonstrated that CERENA (i) offers suitable approximative methods for different biological regimes (or systems in different regimes of copy-numbers), and (ii) renders the comprehensive comparison of approximative descriptions and the subsequent selection straightforward. Also, the implementation of numerical solvers in CERENA proved to be significantly more efficient compared to other packages/implementations. For sensitivity analysis, a further acceleration was achieved by using forward and adjoint sensitivity analyses, with the latter possessing a superior scalability with respect to the number of parameters.

The current version of CERENA allows for the study of population-averaged and population snapshot data by providing time-dependent moments. To that end, a useful advancement could be realized by the integration of CERENA with sophisticated parameter estimation and model selection tools, such as ODE-constrained mixture modeling [38]. Complementarily, the moments obtained using MM, MCM and SSE could be used to compute a distribution approximation [39–41] to provide a more informative comparison with respect to SSA and FSP solutions. An automatic reconstruction of such approximative distributions could be incorporated in future releases of CERENA.

In conclusion, we have shown that CERENA is a comprehensive toolbox for stochastic modeling which maximizes both applicability and computational efficiency. This renders further studies of biological problems of realistic sizes feasible.

Supporting Information

S1 CERENA Documentation. The documentation of CERENA. This documentation includes a more detailed description of the modeling approaches implemented in CERENA, as well as elaborate instructions on using the CERENA toolbox.
(PDF)

Acknowledgments

The authors thank Ramon Grima and Philipp Thomas for discussions regarding the system size expansion.

Funding: This work was funded by the European Union within the ERC grant 'Latent Causes' (A.K., F.J.T.), and the German Research Foundation (DFG) through the Graduate

School of Quantitative Biosciences Munich (F.F.). Merrimack Pharmaceuticals Inc. provided support in the form of salaries for A.R., but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Author Contributions

Conceived and designed the experiments: AK FF FJT JH. Performed the experiments: AK JH. Analyzed the data: AK FF JH. Contributed reagents/materials/analysis tools: AK FF AR JH. Wrote the paper: AK FF FJT JH.

References

1. van Kampen NG. Stochastic processes in physics and chemistry. 3rd ed. Amsterdam: North-Holland; 2007.
2. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1977 Dec; 81(25):2340–2361. doi: 10.1021/j100540a008
3. Shahrezaei V, Swain PS. Analytical distributions for stochastic gene expression. Proc Natl Acad Sci U S A. 2008 Nov; 105(45):17256–17261. doi: 10.1073/pnas.0803850105 PMID: 18988743
4. Ramaswamy R, González-Segredo N, Sbalzarini I, Grima R. Discreteness-induced concentration inversion in mesoscopic chemical systems. Nat Comm. 2012 Apr; 3(779).
5. Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. Nature. 2010 Sept; 467(9):1–7.
6. Gillespie DT. A rigorous derivation of the chemical master equation. Physica A. 1992 Sept; 188(1): 404–425. doi: 10.1016/0378-4371(92)90283-V
7. Jahnke T, Huisinga W. Solving the chemical master equation for monomolecular reaction systems analytically. J Math Biol. 2007 Jan; 54(1):1–26. doi: 10.1007/s00285-006-0034-x PMID: 16953443
8. Munsky B, Khammash M. The finite state projection algorithm for the solution of the chemical master equation. J Chem Phys. 2006 Jan; 124(4):044104. doi: 10.1063/1.2145882 PMID: 16460146
9. Mateescu M, Wolf V, Didier F, Henzinger TA. Fast adaptive uniformisation of the chemical master equation. IET Syst Biol. 2010; 4(6):441–452. doi: 10.1049/iet-syb.2010.0005 PMID: 21073242
10. Kazeev V, Khammash M, Nip M, Schwab C. Direct solution of the Chemical Master Equation using quantized tensor trains. PLOS Comput Biol. 2014 Mar; 10(3):e1003359. doi: 10.1371/journal.pcbi.1003359 PMID: 24626049
11. Engblom S. Computing the moments of high dimensional solutions of the master equation. Appl Math Comp. 2006; 180:498–515. doi: 10.1016/j.amc.2005.12.032
12. Grima R. An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions. J Chem Phys. 2010 July; 133(035101).
13. Hasenauer J, Wolf V, Kazeroonian A, Theis FJ. Method of conditional moments (MCM) for the chemical master equation. J Math Biol. 2014 Aug; 69(3):687–735. doi: 10.1007/s00285-013-0711-5 PMID: 23918091
14. Thomas P, Popovic N, Grima R. Phenotypic switching in gene regulatory networks. Proc Natl Acad Sci U S A. 2014 May; 111(19):6994–6999. doi: 10.1073/pnas.1400049111 PMID: 24782538
15. Zechner C, Ruess J, Krenn P, Pelet S, Peter M, Lygeros J, et al. Moment-based inference predicts bimodality in transient gene expression. Proc Natl Acad Sci U S A. 2012 May; 109(21):8340–8345. doi: 10.1073/pnas.1200161109 PMID: 22566653
16. Milner P, Gillespie CS, Wilkinson DJ. Moment closure based parameter inference of stochastic kinetic models. Stat Comp. 2013 Mar; 23(2):287–295. doi: 10.1007/s11222-011-9310-8
17. Raue A, Schilling M, Bachmann J, Matteson A, Schelke M, Kaschek D, et al. Lessons learned from quantitative dynamical modeling in systems biology. PLOS ONE. 2013 Sept; 8(9):e74335. doi: 10.1371/journal.pone.0074335 PMID: 24098642
18. Sanft KR, Wu S, Roh M, Fu J, Lim RK, Petzold LR. StochKit2: software for discrete stochastic simulation of biochemical systems with events. Bioinf. 2011; 27(17):2457–2458. doi: 10.1093/bioinformatics/btr401
19. Maarleveld TR, Olivier BG, Bruggeman FJ. StochPy: A comprehensive, user-friendly tool for simulating stochastic biological processes. PLOS ONE. 2013 11; 8(11):e79345. doi: 10.1371/journal.pone.0079345 PMID: 24260203

20. Ramsey S, Orrell D, Bolouri H. Dizzy: stochastic simulation of large-scale genetic regulatory networks. J Bioinform Comput Biol. 2005 Apr; 3(2):415–436. doi: 10.1142/S0219720005001144 PMID: 15852513
21. Hegland M, Fletcher-Costin R. CmePy documentation. http://fcostin.github.com/cmepy; 2010.
22. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, et al. COPASI—a COmplex PAthway SImulator. Bioinf. 2006; 22:3067–3074. doi: 10.1093/bioinformatics/btl485
23. Gillespie CS. Moment-closure approximations for mass-action models. IET Syst Biol. 2009 Jan; 3(1): 52–58. doi: 10.1049/iet-syb:20070031 PMID: 19154084
24. Hespanha J. Moment closure for biochemical networks. In: Proc. Int. Symp. on Communications, Control and Signal Processing; 2008. p. 42–147.
25. Schnoerr D, Sanguinetti G, Grima R. Comparison of different moment-closure approximations for stochastic chemical kinetics. The Journal of Chemical Physics. 2015; 143(18). doi: 10.1063/1.4934990 PMID: 26567686
26. Thomas P, Matuschek H, Grima R. Intrinsic Noise Analyzer: A software package for the exploration of stochastic biochemical kinetics using the system size expansion. PLOS ONE. 2013 Jun; 7(6):e38518. doi: 10.1371/journal.pone.0038518
27. Lapin M, Mikeev L, Wolf V. SHAVE: Stochastic Hybrid Analysis of Markov Population Models. In: Proceedings of the 14th International Conference on Hybrid Systems: Computation and Control. HSCC'11. New York, NY, USA: ACM; 2011. p. 311–312. Available from: http://doi.acm.org/10.1145/1967701.1967746.
28. Anderson DF. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. J Chem Phys. 2007 Dec; 127(214107). PMID: 18067349
29. Norris JR. Continuous-time Markov chains I. In: Markov Chains. Cambridge University Press; 1997. p. 60–107. Cambridge Books Online.
30. Grima R. Linear-noise approximation and the chemical master equation agree up to second-order moments for a class of chemical systems. Phys Rev E. 2015 Oct; 92:042124. doi: 10.1103/PhysRevE.92.042124
31. Singh A, Hespanha JP. Approximate moment dynamics for chemically reacting systems. IEEE Trans Autom Control. 2011 Feb; 56(2):414–418. doi: 10.1109/TAC.2010.2088631
32. Hindmarsh AC, Brown PN, Grant KE, Lee SL, Serban R, Shumaker DE, et al. SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers. ACM T Math Software. 2005 Sept; 31(3):363–396. doi: 10.1145/1089014.1089020
33. Kazeroonian A, Theis FJ, Hasenauer J. Modeling of stochastic biological processes with non-polynomial propensities using non-central conditional moment equation. In: Proc. of the 19th IFAC World Congress. vol. 19. Cape Town, South Africa; 2014. p. 1729–1735.
34. Raue A, Steiert B, Schelker M, Kreutz C, Maiwald T, Hass H, et al. Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems. Bioinformatics. 2015 Jul.
35. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinf. 2009 May; 25(25):1923–1929. doi: 10.1093/bioinformatics/btp358
36. Swameye I, Müller TG, Timmer J, Sandra O, Klingmüller U. Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. Proc Natl Acad Sci USA. 2003 Feb; 100(3):1028–1033. doi: 10.1073/pnas.0237333100 PMID: 12552139
37. Fröhlich F, Thomas P, Kazeroonian A, Theis FJ, Grima R, Hasenauer J. Inference for stochastic chemical kinetics using moment equations and system size expansion. submitted. 2015.
38. Hasenauer J, Hasenauer C, Hucho T, Theis FJ. ODE constrained mixture modelling: A method for unraveling subpopulation structures and dynamics. PLOS Comput Biol. 2014 July; 10(7):e1003686. doi: 10.1371/journal.pcbi.1003686 PMID: 24992156
39. Thomas P, Grima R. Approximate probability distributions of the master equation. Phys Rev E. 2015 Jul; 92:012120. doi: 10.1103/PhysRevE.92.012120
40. Andreychenko A, Mikeev L, Wolf V. Reconstruction of multimodal distributions for hybrid moment-based chemical kinetics. Journal of Coupled Systems and Multiscale Dynamics. 2015-06-01T00:00:00; 3(2):156–163. doi: 10.1166/jcsmd.2015.1073
41. Andreychenko A, Bortolussi L, Grima R, Thomas P and Wolf V. Distribution approximations for the chemical master equation: comparison of the method of moments and the system size expansion. submitted. 2015.

Appendix D

A scalable moment-closure approximation for large-scale biochemical reaction networks.

Bioinformatics, 2017.

This is an article published in Bioinformatics following peer review. The version of record Atefeh Kazeroonian, Fabian J. Theis, Jan Hasenauer. **A scalable moment-closure approximation for large-scale biochemical reaction networks.** *Bioinformatics, Volume 33, Issue 14, 15 July 2017, Pages i293-i300. doi: 10.1093/bioinformatics/btx249.* is available online at:
https://doi.org/10.1093/bioinformatics/btx249

Bioinformatics, 33, 2017, i293–i300
doi: 10.1093/bioinformatics/btx249
ISMB/ECCB 2017

OXFORD

A scalable moment-closure approximation for large-scale biochemical reaction networks

Atefeh Kazeroonian,[1,2,3,*] Fabian J. Theis[1,2] and Jan Hasenauer[1,2,*]

[1]Institute of Computational Biology, Helmholtz Zentrum München - German Research Center for Environmental Health, 85764 Neuherberg, Germany, [2]Department of Mathematics, Technische Universität München, 85748 Garching, Germany and [3]Institut für Medizinische Mikrobiologie, Immunologie und Hygiene, Fakultät für Medizin, Technische Universität München, 81675 München, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: Stochastic molecular processes are a leading cause of cell-to-cell variability. Their dynamics are often described by continuous-time discrete-state Markov chains and simulated using stochastic simulation algorithms. As these stochastic simulations are computationally demanding, ordinary differential equation models for the dynamics of the statistical moments have been developed. The number of state variables of these approximating models, however, grows at least quadratically with the number of biochemical species. This limits their application to small- and medium-sized processes.

Results: In this article, we present a scalable moment-closure approximation (sMA) for the simulation of statistical moments of large-scale stochastic processes. The sMA exploits the structure of the biochemical reaction network to reduce the covariance matrix. We prove that sMA yields approximating models whose number of state variables depends predominantly on local properties, i.e. the average node degree of the reaction network, instead of the overall network size. The resulting complexity reduction is assessed by studying a range of medium- and large-scale biochemical reaction networks. To evaluate the approximation accuracy and the improvement in computational efficiency, we study models for JAK2/STAT5 signalling and NFκB signalling. Our method is applicable to generic biochemical reaction networks and we provide an implementation, including an SBML interface, which renders the sMA easily accessible.

Availability and implementation: The sMA is implemented in the open-source MATLAB toolbox CERENA and is available from https://github.com/CERENADevelopers/CERENA.

Contact: jan.hasenauer@helmholtz-muenchen.de or atefeh.kazeroonian@tum.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cellular mechanisms are subject to inherent biological noise that stems from stochastic events such as bursty gene expression. Due to such stochasticity, isogenic cells can behave differently under identical conditions (Elowitz *et al.*, 2002), giving rise to heterogeneous cell populations. Rather than being a nuisance, biological noise has been proven to be crucial in the functioning of biological systems such as microbial populations and biological tissue (Raj and van Oudenaarden, 2008), e.g. increasing their robustness. Studying the stochasticity of biological processes, therefore, can shed light on their underlying mechanisms and is crucial for a better understanding of their behaviour.

Many biological processes, e.g. gene expression and signal transduction, are modelled as networks of chemical species that undergo chemical reactions. The dynamics of chemical reaction networks, i.e. the temporal evolution of the counts of individual species, is usually described by continuous-time discrete-state Markov chains (CTMCs). The statistics of CTMCs are described by the Chemical Master Equation (CME). As the simulation of the CME is computationally intractable for most processes due to their high- or even infinite-dimensional state space, several methods have been proposed to approximate the statistical moments, e.g. moment-closure approximations (MAs) (Engblom, 2006; Lee *et al.*, 2009) and system-size expansions (Grima, 2010; van Kampen, 2007). These

methods yield ordinary differential equations (ODEs) that approximate the temporal evolution of the statistical moments. These ODEs are usually lower-dimensional than the CME, rendering their numerical simulation more tractable. However, already for the analysis of the mean and covariance of the stochastic process, the size of the state space of the approximating models grows quadratically with the number of biochemical species. This limits the application of these methods to small- and medium-scale biochemical reaction networks if the calculation of all statistical moments is required. However interestingly, in a range of applications, including parameter estimation (Fröhlich *et al.*, 2016; Munsky *et al.*, 2009), information about a subset of statistical moments can be sufficient.

In this study, we introduce a scalable second-order moment-closure approximation (s2MA) which is feasible for large-scale biochemical reaction networks. The s2MA is designed for the accurate description of selected statistical moments, including means and variances. We introduce an algorithm that exploits the structure of the reaction network to select the subset of moments which are most relevant for the reliable approximation of means and variances. Using analytical results for toy networks and published biological models, we show the superior scaling of s2MA over other methods for moment approximation, which renders the s2MA tractable for large reaction networks. To assess the accuracy and computational efficiency of s2MA, we simulated several network motifs and models for JAK2/STAT5 and TNF signalling.

2 Approach

We consider a biochemical reaction network of n species, $S_1, \ldots, S_n$, and n_r reactions, $R_1, \ldots, R_{n_r}$. The state of this network is denoted by $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ where X_i is the number of molecules of species S_i. Upon the firing of reaction R_r, the state $\mathbf{X}$ undergoes the transition $\mathbf{X} \xrightarrow{a_r} \mathbf{X} + \boldsymbol{\nu}_r$, in which ν_r and $a_r(\mathbf{X})$ denote the stoichiometry and the propensity of reaction R_r, respectively. Due to the stochastic nature of chemical reactions, the state vector $\mathbf{X}$ evolves stochastically over time. The probability distribution of $\mathbf{X}$ at time t is denoted by $p(\mathbf{x}|t)$ over all possible states $\mathbf{x}$.

The temporal evolution of the statistical moments of $p(\mathbf{x}|t)$ can be approximated using MAs of different orders. The order of an MA is the highest order of the statistical moments which are modelled. The second-order MA (2MA) is an ODE with $n(n+3)/2$ state variables which describes the dynamics of the mean $\mathbf{m} = \sum_{\mathbf{x}} \mathbf{x} p(\mathbf{x}|t)$ and covariance $\mathbf{C} = \sum_{\mathbf{x}} (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T p(\mathbf{x}|t)$:

$$\begin{aligned}
\frac{\partial m_i}{\partial t} &= \sum_r \nu_{ri} \left(a_r(\mathbf{m}) + \frac{1}{2} \sum_{k,l} \left. \frac{\partial^2 a_r}{\partial x_k \partial x_l} \right|_{\mathbf{m}} C_{kl} \right), \\
\frac{\partial C_{ij}}{\partial t} &= \sum_r \left[\nu_{ri}\nu_{rj} a_r(\mathbf{m}) + \sum_k \left. \frac{\partial a_r}{\partial x_k} \right|_{\mathbf{m}} \left(\nu_{ri} C_{jk} + \nu_{rj} C_{ik} \right) \right. \\
&\quad \left. + \frac{1}{2} \sum_{k,l} \left. \frac{\partial^2 a_r}{\partial x_k \partial x_l} \right|_{\mathbf{m}} \left(\nu_{ri}\nu_{rj} C_{kl} + \nu_{ri} C_{jkl} + \nu_{rj} C_{ikl} \right) \right],
\end{aligned} \tag{1}$$

where C_{ikl} denotes the third-order moment of X_i, X_k and X_l. Due to the symmetry $C_{ij} = C_{ji}$ only C_{ij} with $i \leq j$ is considered. As in (1), the evolution equations for second-order moments usually depend on third-order moments. To close the 2MA equations, moment-closure techniques are applied which approximate the third-order moments as functions of first- and second-order moments (Hespanha, 2008). The moment closure introduces an approximation error to the otherwise exact moment equations, as it relies on assumptions about $p(\mathbf{x}|t)$ (e.g. normality or log-normality; Singh and Hespanha, 2006).

The 2MA (1) describes the covariances of all pairs of species and thus possesses $O(n^2)$ state variables. This quadratic scaling with respect to the number of species, n, poses a challenge for the applicability of 2MA to large biological networks that may contain several hundreds up to thousands of species. However, it is usually observed that in large biochemical networks, many pairwise correlations between species are small. This implies a comparably low covariance and a small contribution to the right-hand side of (1). Consequently, for an approximation of the dynamics of the biochemical network, it may not be necessary to model all covariances.

Studying a series of networks, including the JAK2/STAT5 signalling pathway described by Bachmann *et al.* (2011), we observed that species that directly influence each other via a reaction have a stronger pairwise correlation. For the JAK2/STAT5 signalling pathway, depicted in Figure 1A, we found that >50% of the correlation coefficients do not exceed an absolute value of 0.1 (Fig. 1B). Furthermore, the correlation coefficients decrease as the distance between species in the network increases (Fig. 1C). Since in many cases biological networks are sparsely connected and distances between species are relatively large (Fig. 1D), a significant portion of the covariances may be negligible.

Motivated by this observation, we develop a scalable s2MA that models a subset of covariances. The s2MA is designed to provide a good approximation for means and variances of species, as those moments are essential in a range of applications including parameter estimation (Munsky *et al.*, 2009; Fröhlich *et al.*, 2016). Accordingly, the s2MA captures the subset of covariances that are expected to influence the temporal evolution of the means and variances most strongly. In the simplest case, we only consider the covariances $\mathbf{C}^*$ that have a direct influence on the means and variances, i.e. those that appear in their evolution equations for m_i and C_{ii}:

- Covariances C_{ik} for which a reaction R_r exists with $\nu_{ri} \neq 0$ and $\frac{\partial a_r}{\partial x_k} \neq 0$. This is the case if S_k is a modifier or reactant in a reaction producing or consuming S_i.
- Covariances C_{kl} for which a reaction R_r exists with $\nu_{ri} \neq 0$ and $\frac{\partial^2 a_r}{\partial x_k \partial x_l} \neq 0$. This is the case if both, S_k and S_l, are modifiers or reactants in a reaction producing or consuming S_i.

The remaining covariances are set to zero. The resulting MA exploits the network structure and is similar to a recently proposed MA for spatially distributed systems exploiting the neighbourhood structure (Feng *et al.*, 2016). In the following, we present a mathematical formulation of the s2MA as well as extensions to control its size and approximation accuracy.

3 Materials and Methods

To simulate the statistical moments of the trajectories of large-scale stochastic biochemical reaction networks, we introduce scalable moment-closure approximations (sMAs). These sMA are based on the afore-mentioned findings and exploit the structure of the biochemical reaction network. In the following, we present the required graph characteristics and the derivation of the s2MA.

3.1 Graph representation of biochemical reaction networks

The s2MA uses the structure of the reaction network to identify the covariances that are most relevant to accurately approximate the means and variances of species. To establish a simple structure-based procedure, we exploit the graph structure of the biochemical reaction networks. This graph structure is best represented using the Systems Biology Graphical Notation (SBGN) process diagram

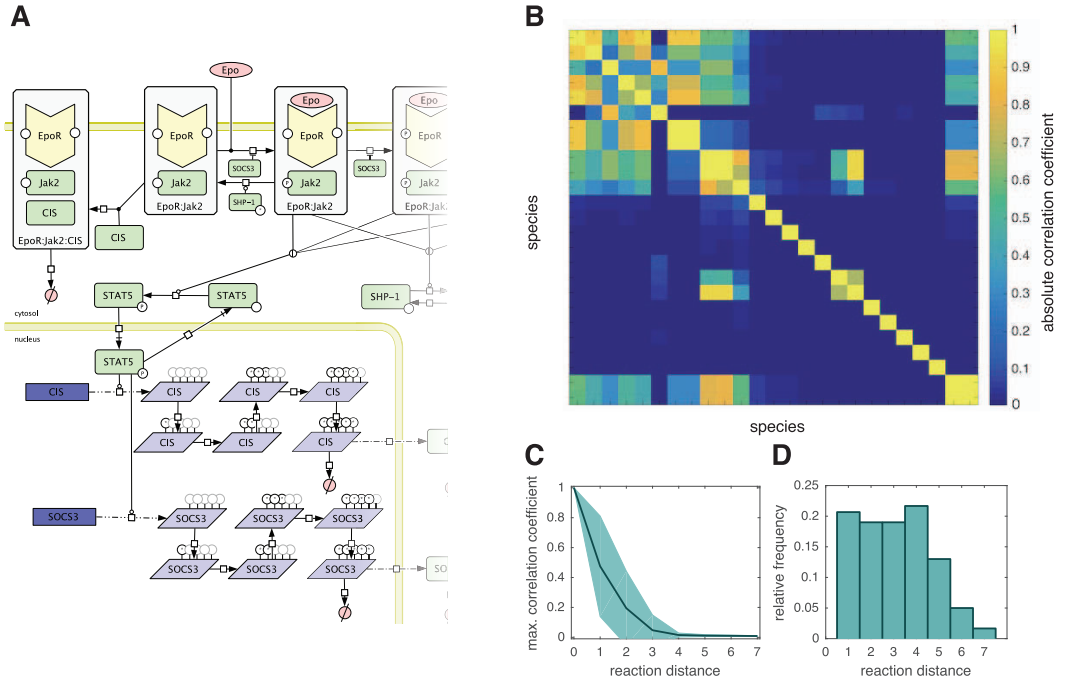


Fig. 1. Correlation coefficients in the simulated JAK2/STAT5 signalling pathway. (**A**) A partial schematic of the JAK2/STAT5 signalling pathway. (**B**) Maximum absolute pairwise correlation coefficients found in the simulation of the JAK2/STAT5 signalling pathway. (**C**) Maximum absolute pairwise correlation coefficients as function of the distance between species. (**D**) Frequency distribution of distance between species pairs

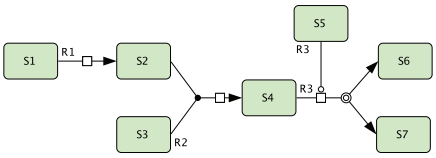


Fig. 2. Illustration of SBGN process diagram of a simple biochemical reaction network. Biochemical species (boxes), biochemical processes (squares) and interactions/dependencies (arcs) are visualised. Label S_i indicates species S_i

(Le Novère *et al.*, 2009). In essence, SBGN process diagram is a graph which consists of *entity nodes* representing biochemical species, *process nodes* representing biochemical reactions and arcs indicating the interactions/dependences. The incoming edges to a process node indicate all the reactants, as well as the modifiers, of the corresponding reaction, while the outgoing edges from a process node mark the products. For instance, reaction R_2 in Figure 2 is a bimolecular reaction where species S_2 and S_3 react to form species S_4. In reaction R_3, species S_5 acts as a modifier that activates the conversion of S_4 into S_6 and S_7. The graph structure is encoded in the propensities and the stoichiometric coefficients and can be easily visualized for Systems Biology Markup Language (SBML) models using software toolboxes such as CellDesigner (Funahashi *et al.*, 2008).

We use the graph representation to define a *dependency matrix* D which summarizes direct dependencies between species in the network. Following the arguments in Section 2, we say that a species S_j directly depends a species S_i, if the evolution equations for the mean or the variance of S_j, i.e., m_j and C_{jj}, depend on moments of S_i. Accordingly, it can be shown that:

- The products of a reaction depend on the reactants and the modifiers.
- The reactants of a reaction depend on the other reactants and the modifier.

This yields the dependency matrix D,

$$D_{ij} = \begin{cases} 1 & \text{if } S_i \text{ directly influences } S_j \\ 0 & \text{otherwise} \end{cases}$$

Note that D is not necessarily symmetric as the defined dependency is a directed property. In the model depicted in Figure 2, S_4 depends on S_2 ($D_{24} = 1$) but not vice versa ($D_{42} = 0$). The dependency matrix D encodes the necessary information for the construction of the s2MA.

3.2 The scalable s2MA

The exact evolution equations for means $\mathbf{m}$ and covariances $\mathbf{C}$ (1) can be written as

$$\begin{aligned} \frac{\partial m_i}{\partial t} &= F_{m,i}(\mathbf{m}, \mathbf{C}, \mathbf{H}), \quad i \in \{1, \ldots, n\} \\ \frac{\partial C_{ij}}{\partial t} &= F_{C,ij}(\mathbf{m}, \mathbf{C}, \mathbf{H}), \quad (i,j) \in I \\ &\qquad \text{with } I = \{(i,j) \in \{1, \ldots, n\}^2 | i \leq j\}. \end{aligned} \tag{2}$$

where $\mathbf{H}$ denotes all moments with orders greater than two. To avoid redundancies caused by the symmetry of the covariances, $C_{ij} = C_{ji}$, we consider only the subset I of covariances. The higher-order moments $\mathbf{H}$ result from reactions with non-linear propensities and their temporal evolution is not to described by (2). To obtain a closed formulation, the higher-order moments $\mathbf{H}$ are approximated by functions of lower-order moments, $\mathbf{H} \approx \bar{H}(\mathbf{m}, \mathbf{C})$, using moment closure techniques. Common techniques include zero-cumulant closure (Matis and Kiffe, 1999), low-dispersion closure (Hespanha, 2008), and

derivative-matching (Singh and Hespanha, 2007). This yields the 2MA,

$$\begin{aligned}\frac{\partial m_i}{\partial t} &= F_{m,i}\big(\mathbf{m},\mathbf{C},\bar{H}(\mathbf{m},\mathbf{C})\big) =: \bar{F}_{m,i}(\mathbf{m},\mathbf{C}), \quad i\in\{1,\dots,n\}\\ \frac{\partial C_{ij}}{\partial t} &= F_{C,ij}\big(\mathbf{m},\mathbf{C},\bar{H}(\mathbf{m},\mathbf{C})\big) =: \bar{F}_{C,ij}(\mathbf{m},\mathbf{C}), \quad (i,j)\in I.\end{aligned}$$

The solution of the 2MA yields an approximation to the moments of the state of the biochemical reaction network. The quality of this approximation depends on the accuracy of the moment closure (Kazeroonian *et al.*, 2016; Schnoerr *et al.*, 2015).

The 2MA possesses $n(n+3)/2$ state variables, thus, it grows quadratically with n. The simplest s2MA, the first-degree s2MA, reduces the growth rate by considering only the covariances on which the temporal evolution of the means and variances depends directly. This reduced set of covariances, $\mathbf{C}_{ij}$ with $(i,j)\in I^{(1)}$, can be determined using the dependency matrix D,

$$I^{(1)} = \left\{(i,j)\in\{1,\dots,n\}^2 \middle| i\le j \wedge \left(D+D^T\right)_{ij}\neq 0\right\}.$$

The covariances C_{ij} with $(i,j)\in I\backslash I^{(1)}$ are not modelled by the first-degree s2MA but can be approximated using the means, the variances and the reduced set of covariances. In this study, we use the low-dispersion closure, $C_{ij}=0$ for $(i,j)\in I\backslash I^{(1)}$.

The approximation quality of the s2MA can be controlled using the cut-off degree. The second-degree s2MA describes the covariances that influence the temporal evolution of the means and variances either directly or via an intermediate step. More precisely, the second-degree s2MA considers the covariances $\mathbf{C}_{ij}$, $(i,j)\in I^{(1)}$ and the covariances which appear in their evolution equations. The set of these covariances, $\mathbf{C}_{ij}$, $(i,j)\in I^{(2)}$, is defined by the second power of the dependency matrix D^2. More generally, we define the δth-degree s2MA (s2MA-δ) which describes the reduced set of covariances $\mathbf{C}_{ij}$ with $(i,j)\in I^{(\delta)}$,

$$I^{(\delta)} = \left\{(i,j)\in\{1,\dots,n\}^2 \middle| i\le j \wedge \left(D^\delta+\left(D^\delta\right)^T\right)_{ij}\neq 0\right\}.$$

The degree $\delta\ge 1$ denotes the maximal intermediate dependency steps between species pairs (S_i, S_j) for which covariances are included in the s2MA. For a given δ, we obtain the s2MA-δ,

$$\begin{aligned}\frac{\partial m_i}{\partial t} &= \bar{F}_{m,i}(\mathbf{m},\mathbf{C}), \quad i\in\{1,\dots,n\}\\ \frac{\partial C_{ij}}{\partial t} &= \bar{F}_{C,ij}(\mathbf{m},\mathbf{C}), \quad (i,j)\in I^{(\delta)}\\ C_{ij}(t) &= 0, \quad (i,j)\in I\backslash I^{(\delta)}.\end{aligned} \tag{3}$$

We focus on the case $\delta=1$, in which merely covariances of interacting species are considered. To capture long-range interactions, we considered $\delta\ge 2$, which can improve the approximation accuracy of the s2MA in biological systems with complex or highly non-linear kinetics. The potentially enhanced approximation accuracy comes at the cost of higher computational complexity as the number of state variables increases with δ. In Section 4, we demonstrate that one can usually find a satisfactory tradeoff between the computational cost and approximation quality for complex biological networks.

3.3 Implementation

We implemented methods for the construction and simulation of the s2MA in the ChEmical REaction Network Analyzer (CERENA), an open source MATLAB toolbox (Kazeroonian *et al.*, 2016). The advanced version of CERENA supports automatic construction of the 2MA and the s2MA using symbolic calculus and allows for a range of moment closure schemes. The proposed construction algorithm circumvents the formulation of the full 2MA to ensure feasibility for large-scale networks. Biochemical reaction networks can be defined in the SBML or in a simple m-file format. For efficient numerical simulation, C-code simulation files are compiled using the Advanced MATLAB Interface for CVODES and IDAS (Fröhlich *et al.*, 2016). This C-code employs sophisticated numerical methods implemented in CVODES (Serban and Hindmarsh, 2005), facilitating the study of a wide range of models. In addition, simulation using MATLAB internal ODE solvers is supported. CERENA is freely available from GitHub (http://cerenadevelopers.github.io/CERENA/) and its functionality is described in a detailed documentation.

4 Results

In the following, we study the properties of the s2MA and illustrate its importance for the study of large-scale biochemical reaction networks. For this purpose, we analyse various network motifs as well as published pathway models for which available methods are computationally demanding or even infeasible.

4.1 Scaling properties

The size of the s2MA for a given network as well as its scaling properties depends on network characteristics. To highlight the scaling properties, we considered reoccurring network motifs and performed a general theoretical assessment. As verification, we inspected published signalling and metabolic pathways with different numbers of biochemical species.

4.1.1 Theoretical scaling for network motifs and generic networks

To study the scaling properties of s2MA, we considered three different network motifs illustrated in Figure 3A–C:

- A *chain of monomolecular reactions* as observed in metabolic processes (Krumsiek *et al.*, 2011) and delay representations (Bachmann *et al.*, 2011).
- A *2D grid of monomolecular reactions* as observed in histone methylation (Zheng *et al.*, 2012).
- A *sequence of bimolecular reactions with a hub* as observed in polymerisation related processes, e.g. prion aggregation (Rubenstein *et al.*, 2007).

For these network motifs, we derived the size of the s2MA-1 and -2 (see Table 1). For all three motifs, we found a linear scaling of the size of the s2MA-1 with respect to the number of species n. The same holds for the s2MA-2 of the chain of monomolecular reactions and the 2D grid of monomolecular reactions. The s2MA-2 of the sequence of bimolecular reactions with a hub is identical to the 2MA as all species are connected via at most one intermediate species (the hub). Accordingly, the analysis of selected motifs suggests that the s2MA allows for a substantial size reduction in the absence of central hubs.

For generic network structures, the scaling of the s2MA depends on the degree distribution $P(d)$ of nodes in the graph representation of the biochemical reaction network (see Section 3.1). By construction, the number of covariances in the s2MA-1 is the sum of node degrees over two,

$$\text{number of covariances in s2MA} - 1 = \frac{1}{2}\sum_{i=1}^{n} d_i = \frac{n\bar{d}}{2},$$

in which d_i denotes the degree of node i and the division by two is required as covariances are associated to two nodes. Introducing the average node degree, $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$, the s2MA-1 describes the temporal evolution of n means, n variances and $\frac{n\bar{d}}{2}$ covariances, and thus possesses $\frac{n}{2}(4+\bar{d})$ state variables. If we assume that there are no long-ranged connections in the network and every node is only connected to a subset of neighbouring nodes, then we can assume that $\bar{d}$ is independent of the size of the network n, and s2MA-1 will scale linearly with the number of species.

The degree distribution in biological systems have been reported to follow a power-law (Albert, 2005), $P(d) \propto d^{-\gamma}$, with an exponent of $2 < \gamma < 3$. Networks with this property are usually referred to as scale-free networks. The expected value of the average node degree in scale-free networks is

$$\mathbb{E}[\bar{d}] = \sum_{i=1}^{n} d_i = \sum_{d=1}^{n-1} d \cdot P(d) = \sum_{d=1}^{n-1} d^{1-\gamma}.$$

Using the lower bound of γ and the upper bound on the partial sums of the harmonic series, we obtain

$$\text{if} \quad \gamma > 2 \quad \Rightarrow \quad \mathbb{E}[\bar{d}] < (\ln(n-1)+1).$$

Evaluating this upper bound, we notice that even for networks with up to $n = 10^4$ species, $\bar{d}$ hardly exceeds 10, making it behave like a constant compared to n. Accordingly, we conclude that the size of the s2MA-1 should scale (only slightly worse than) linearly with the network size.

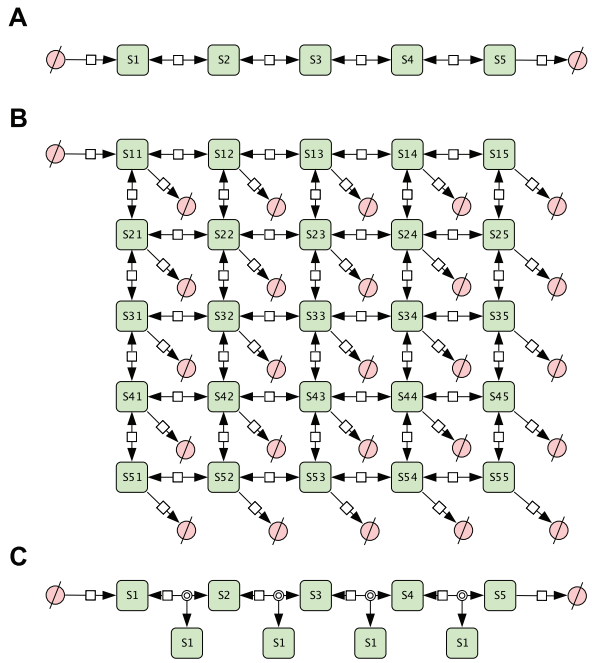


Fig. 3. Illustration of considered network motifs. (**A**) Chain of monomolecular reactions ($n = 5$). (**B**) 2D grid of monomolecular reactions ($n = 25$). (**C**) Chain of bimolecular reactions with a hub ($n = 5$)

Table 1. Comparison of the sizes of the 2MA and the s2MA for different network motifs

Network motif	Number of state variables		
	2MA	s2MA-1	s2MA-2
Chain of monomolecular reactions	$\frac{n(n+3)}{2}$	$3n-1$	$4n-3$
2D grid of monomolecular reactions	$\frac{n(n+3)}{2}$	$4n-\sqrt{n}$	$7n-7\sqrt{n}+1$
Chain of bimolecular reactions	$\frac{n(n+3)}{2}$	$4n-3$	$\frac{n(n+3)}{2}$

4.1.2 Scaling for published biochemical reaction networks

To corroborate the theoretical predictions derived under the assumption of scale-free networks, we studied a collection of 50 published biochemical reaction networks. These networks were extracted from the BioModels, NetPath and Reactome database. They include between 17 and 1277 biochemical species and a range of rate laws. A comprehensive list of the networks is provided in Supplementary Table S1.

We used an extension of the MATLAB toolbox CERENA to generate the s2MAs for the networks and recorded the sizes (Fig. 4). The analysis verified our prediction of a roughly linear relation between the size of the s2MA-1 and the number of species. The s2MA-1, on average, possessed only five times more state variables than the reaction rate equations, ensuring the applicability of the s2MA-1 to large-scale networks. For the largest network, a size reduction by a factor of >120 was achieved compared to the 2MA.

As the consideration of pair-wise correlations between reaction partners might not be sufficient for a particular application, we also assessed the scaling of the s2MA-2 and -3. In agreement with the results for the network motifs, we found that the size of the s2MA of degree ≥ 2 grew stronger than linear, namely with order 1.25 and 1.49. This implies that for realistic pathway structures, also the size of the s2MA of degree 2 and 3 grows substantially slower than the size of the 2MA, facilitating the analysis of stochasticity in large-scale networks.

4.2 Approximation accuracy

The improved scalability of the s2MA is achieved by merely modelling a subset of covariances. In the following section, we will assess the resulting approximation error and its dependence on the degree

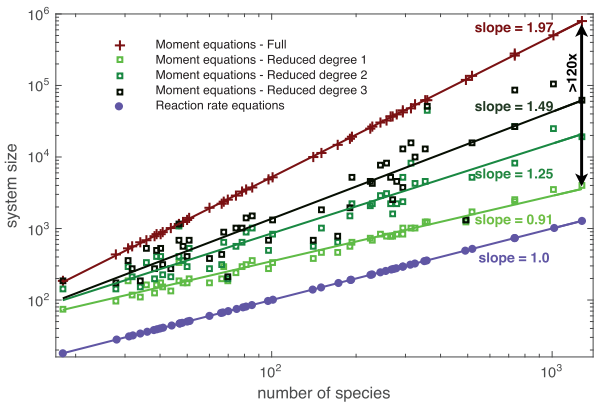


Fig. 4. Scaling of different moment-closure approximations for published networks. Moment-closure approximations for individual networks (markers) and fitted regression curves (lines) are shown

of the s2MA. For this analysis, we consider two network motifs and two published signalling pathways.

4.2.1 Comparison of approximation methods for network motifs

For an initial assessment of the approximation accuracy, we considered the *chain of monomolecular reactions* ($n = 10$) and the *sequence of bimolecular reactions with a hub* ($n = 20$) with mass action kinetics (Fig. 3A and C). The initial conditions and parameter values are reported in Supplementary Tables S2 and S3. As a measure for the approximation accuracy the relative errors in the means and variances were used, e.g.

$$100\% \times \frac{|C_{ii}^{s2MA}(t) - C_{ii}^{2MA}(t)|}{\max_t C_{ii}^{2MA}(t)},$$

in which $C_{ii}^{s2MA}(t)$ and $C_{ii}^{2MA}(t)$ denote the time-dependent variance of species i calculated by s2MA and 2MA, respectively.

The numerical simulation revealed a good agreement of means and variances of 2MA and s2MA-1 (Fig. 5). Neglecting the covariances that are not modelled by the s2MA; however, resulted in a relative error $< 1\%$ for the means and $< 20\%$ for the variances. Given a size reductions of 55.4 and 66.5%, the low relative error supported the validity of the approach.

4.2.2 Comparison of approximation accuracy for s2MA of different degrees on published biochemical reaction networks

To assess the approximation accuracy of s2MAs of different degrees for realistic pathway topologies, we considered the published models of JAK2/STAT5 signalling and TNF signalling. These models were also considered in the scalability analysis (Section 4.1.2).

The model of JAK2/STAT5 signalling describes the activity of the transcription factor STAT5 in response to Epo treatment (Bachmann *et al.*, 2011). STAT5 regulates cell proliferation, differentiation and inflammation. The considered model accounts for 25 biochemical species and includes biochemical reactions with non-mass action kinetics. Its 2MA possesses 350 state variables while the s2MA-1 has less than one-third of the state variables, namely 112. Nonetheless, the simulation revealed a good agreement of 2MA and s2MA-1 for the means and variances (Fig. 6A). The means and variances computed using s2MA-2 and s2MA-3 were essentially indistinguishable from those computed using 2MA. For all s2MAs, we observed a reduction in the computation time comparable to the size reduction.

The model of TNF signalling describes the activation of pro- and antiapoptotic factors, i.e. caspases and NFκB, in response to TNF treatment (Schliemann *et al.*, 2011). Apoptosis is a form of programmed cell death which is relevant, among others, in immune

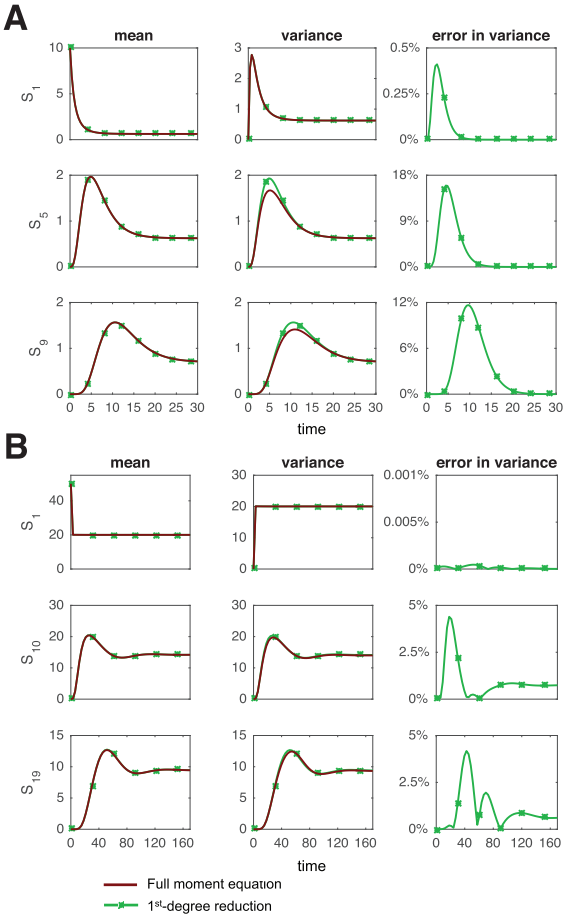


Fig. 5. Approximation accuracy of the s2MA-1 for network motifs. (**A**) The chain of monomolecular reactions with $n = 10$. (**B**) The sequence of bimolecular reactions with $n = 20$. (A, B) Means and variances are depicted along with relative errors in the variances (2MA versus s2MA-1) for several biochemical species

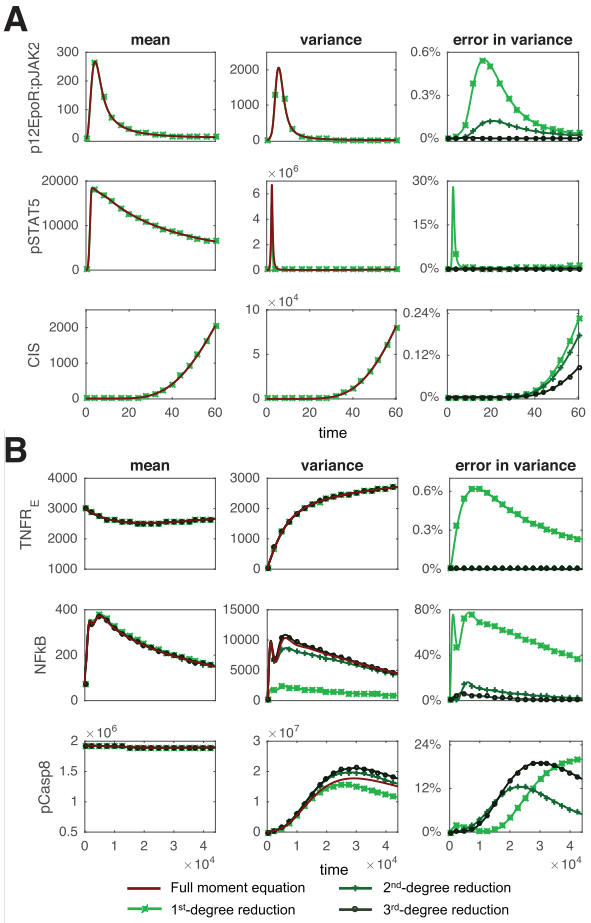


Fig. 6. Approximation accuracy of the s2MA for published pathways. (**A**) The JAK2/STAT5 signalling pathway. (**B**) The TNF signalling network. (A, B) Means and variances computed using the 2MA and the s2MA-1, -2 and -3 are depicted for several biochemical species. For the s2MA of different degrees, the relative error in the variances with respect to the 2MA is provided

response and cancer. The model comprises 47 biochemical species, yielding a 2MA with 1175 state variables. In contrast, the s2MA-1, -2 and -3 possess only 189, 540 and 664 state variables. The numerical simulation of the s2MA-1 was more than 25 times faster than the numerical simulation of the 2MA. The disagreement between s2MA-1 and 2MA, which resulted in a relative error of 100% for some species (Fig. 6B) indicates that also covariances betweens species which do not interact directly might be required for an accurate description of mean and variances. The comparison of the results for s2MA-1, -2 and -3 confirmed that the approximation error decreases as more covariances are taken into account. For s2MA-3, the relative error is below 15%.

In summary, our analysis of network motifs and published networks revealed that the s2MA yields substantially smaller ODE models than the 2MA, indicating a substantial gain in computational efficiency. Moreover, even for models with many species and non-mass action kinetics, a good approximation accuracy was achieved.

5 Discussion

Stochasticity of biochemical reactions is an inherent property of biological processes. It contributes to the establishment of functional cell-to-cell variability and robust decision-making (Eldar and Elowitz, 2010; Raj and van Oudenaarden, 2008). The analysis of the stochastic processes is, however, restricted by the available analytical and numerical methods. In this manuscript, we introduce the scalable second-order moment-closure approximation, the first method to enable the simulation of statistical moments of large-scale stochastic processes. The s2MA exploits the network structure to construct approximate evolution equations for selected process statistics.

To assess and illustrate the properties of s2MA, we studied network motifs and a large collection of published networks. This comprehensive evaluation, which sets this study apart from other studies of moment-closure approximations (e.g. (Feng *et al.*, 2016; Singh and Hespanha, 2006), verified that in practice the size of the first-degree s2MA (s2MA-1) grows linearly with the network size, a scalability that is similar to the reaction rate equations. Accordingly, the s2MA enables the assessment of stochastic dynamics on a new scale. The achieved scalability, however, comes at the cost of an approximation error. The approximation quality can be easily controlled via the degree of the s2MA.

Beyond scalable moment-closure approximations for the calculation of means and variances, structured-based approaches might be used for the evaluation of third-order moments and conditional moments (Hasenauer *et al.*, 2014). Complementarily, an improvement might be achieved by tailored moment-closure schemes which avoid neglecting a large fraction of covariances. A possible formulation, for instance, could be based on partial correlations (Krumsiek *et al.*, 2011) or convergent moments (Zhang *et al.*, 2016). All of these methods would benefit from *a priori* and *a posteriori* error bounds, which are not yet available for moment-closure approximations, such as the s2MA, but are urgently needed.

In summary, we presented a scalable moment-closure approximation for the simulation of stochastic chemical kinetics. This method is beneficial for application problems that require numerical simulations at low computation cost, e.g. parameter estimation (Fröhlich *et al.*, 2016; Munsky *et al.*, 2009). An implementation of the method is provided in the open-source MATLAB toolbox CERENA to facilitate its application and further extensions. This implementation, as well as the concept of structure-based reduction, is applicable to a broad range of problems and will help to improve the analysis of stochastic chemical kinetics.

Acknowledgements

We acknowledge financial support from the Postdoctoral Fellowship Program (PFP) of the Helmholtz Zentrum München.

Conflict of Interest: none declared.

References

Albert,R. (2005) Scale-free networks in cell biology. *J. Cell. Sci.*, **118**, 4947–4957.

Bachmann,J. *et al.* (2011) Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol. Syst. Biol.*, **7**, 516.

Eldar,A. and Elowitz,M.B. (2010) Functional roles for noise in genetic circuits. *Nature*, **467**, 1–7.

Elowitz,M.B. *et al.* (2002) Stochastic gene expression in a single cell. *Science*, **297**, 1183–1186.

Engblom,S. (2006) Computing the moments of high dimensional solutions of the master equation. *Appl. Math. Comp.*, **180**, 498–515.

Feng,C. *et al.* (2016) Automatic moment-closure approximation of spatially distributed collective adaptive systems. In *ACM Transactions on Modeling and Computer Simulation*, Vol. **26**.

Fröhlich,F. *et al.* (2016) Inference for stochastic chemical kinetics using moment equations and system size expansion. *PLoS Comput. Biol.*, **12**, e1005030.

Funahashi,A. *et al.* (2008) CellDesigner 3.5: A versatile modeling tool for biochemical networks. *Proc. IEEE*, **96**, 1254–1265.

Grima,R. (2010) An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions. *J. Chem. Phys.*, **133**, (035101.

Hasenauer,J. *et al.* (2014) Method of conditional moments (MCM) for the chemical master equation. *J. Math. Biol.*, **69**, 687–735.

Hespanha,J. (2008). Moment closure for biochemical networks. In *Proceedings of the 3rd International Symposium on Communications, Control and Signal Processing*, St Julians, 2008, pp. 142–147.

Kazeroonian,A. *et al.* (2016) CERENA: ChEmical REaction Network Analyzer—a toolbox for the simulation and analysis of stochastic chemical kinetics. *PLoS One*, **11**, e0146732.

Krumsiek,J. *et al.* (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.*, **5**,

Le Novère,N. *et al.* (2009) The Systems Biology Graphical Notation. *Nat. Biotechnol.*, **27**, 735–741.

Lee,C.H. *et al.* (2009) A moment closure method for stochastic reaction networks. *J. Chem. Phys.*, **130**, 134107.

Matis,H.J. and Kiffe,T.R. (1999) Effects of immigration on some stochastic logistic models: a cumulant truncation analysis. *Theor. Popul. Biol.*, **56**, 139–161.

Munsky,B. *et al.* (2009) Listening to the noise: random fluctuations reveal gene network parameters. *Mol. Syst. Biol.*, **5**, (318.

Raj,A. and van Oudenaarden,A. (2008) Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, **135**, 216–226.

Rubenstein,R. *et al.* (2007) Dynamics of the nucleated polymerization model of prion replication. *Biophys. Chem.*, **125**, 360–367.

Schliemann,M. *et al.* (2011) Heterogeneity reduces sensitivity of cell death for TNF-stimuli. *BMC Syst. Biol*, **5**, (204.

Schnoerr,D. *et al.* (2015) Comparison of different moment-closure approximations for stochastic chemical kinetics. *J. Chem. Phys.*, **143**, (185101.

Serban,R. and Hindmarsh,A.C. (2005) CVODES: An ODE solver with sensitivity analysis capabilities. *ACM T. Math. Softw.*, **31**, 363–396.

Singh,A. and Hespanha,J. (2007) A derivative matching approach to moment closure for the stochastic logistic model. *Bull. Math. Biol.*, **69**, 1909–1925.

Singh,A. and Hespanha,J.P. (2006). Lognormal moment closures for biochemical reactions. In *Proceedings of the 45th IEEE Conference on Decision and Control (CDC)*, San Diego, CA, 2006, pp. 2063–2068.

van Kampen,N.G. (2007). *Stochastic Processes in Physics and Chemistry, 3rd edn*. North-Holland, Amsterdam.

Zhang,J. *et al*. (2016) A moment-convergence method for stochastic analysis of biochemical reaction networks. *J. Chem. Phys*, **144**, 194109.

Zheng,Y. *et al*. (2012) Total kinetic analysis reveals how combinatorial methylation patterns are established on lysines 27 and 36 of histone H3. *Proc. Natl. Acad. Sci. USA*, **109**, 13549–13554.

Supplement to

A scalable moment-closure approximation for large-scale biochemical reaction networks

Atefeh Kazeroonian [1,2,3,*], Fabian J. Theis [1,2] and Jan Hasenauer [1,2]

[1]Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology, 85764 Neuherberg, Germany

[2]Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, 85748 Garching, Germany.

[3]Technische Universität München, Fakultät für Medizin, Institut für Medizinische Mikrobiologie, Immunologie und Hygiene, 81675 München, Germany.

[*]To whom correspondence should be addressed.

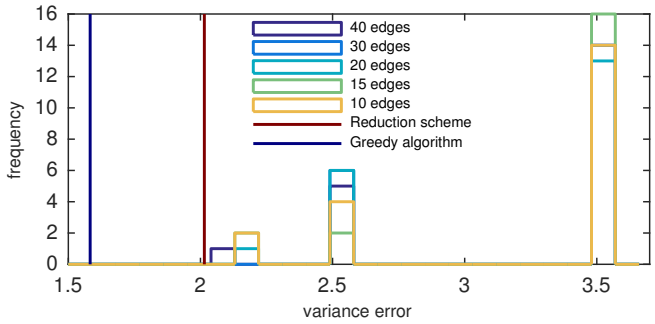
1 Comparison of the s2MA with structure-based, random and greedy selection of covariances

To corroborate our hypothesis that a structure-based selection of covariances is appropriate, we compared the proposed selection scheme to random selection for the simulation of the chain of monomolecular reactions (Supplement Figure 3). For random selection, the index set $I^{(\delta)}$ in Eq. (3) was defined as the union of index pairs corresponding to (1) the variances, $\{(i,i)|i=1,\dots,n\}$, and (2) a random sample drawn from $\{(i,j)\in\{1,\dots,n\}^2|i<j\}$ without replacement. For various numbers of covariances, we sampled the distribution of the error in the variances,

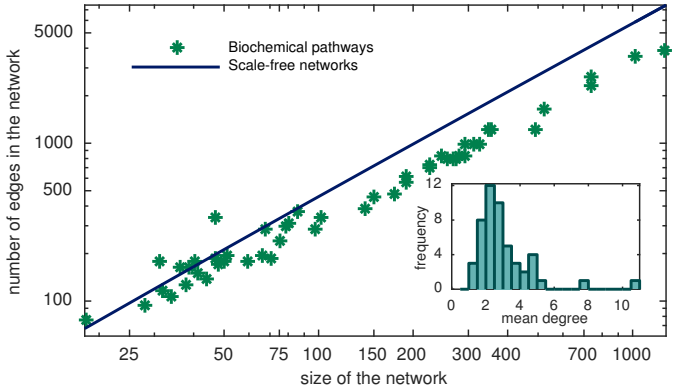
$$\sum_{i=1}^{n}\int_0^T \left(C_{ii}^{\text{rand}}(t)-C_{ii}^{\text{2MA}}(t)\right)^2 dt,$$

in which $C_{ii}^{\text{rand}}(t)$ denotes the time-dependent variance of species i calculated for a random set of covariances. The comparison with the error for the s2MA-1 showed that the structure-based selection achieved a lower approximation error than random selection, even if random selection was allowed to describe a larger number of the covariances (Supplement Figure 1). This implies that the direct dependency, as defined in the *Approach* section, is a good proxy for the relevance of a covariance for a good approximation.

To assess the sub-optimality of the structure-based selection of covariances, we compared it to a greedy approach. The greedy approach started with an ODE which merely includes the evolution equations for means and variances, i.e., Eq. (3) with $I^{(\delta)}=\{(i,i)|i=1,\dots,n\}$, and sequentially added one covariance. In each iteration, all possible choices for the additional covariance were considered and the covariance which resulted in the strongest decrease of the approximation error was included. This procedure was repeated until the ODE had the same size as the s2MA-1. The final model derived using this greedy approach possessed – as expected – a slightly lower approximation error than the s2MA-1 (Supplement Figure 1). The improvement of approximation accuracy, however, came with a substantial computational burden, which might not be feasible for large-scale biochemical reaction networks. Furthermore, the greedy approach is simulation-based and results generally depend on parameter values.



Supplement Figure 1: **Comparison of structure-based covariance selection, random selection and the greedy approach for the chain of monomolecular reactions ($n = 10$).** The integrated error in the variance is shown for the s2MA-1 and reduced MA with randomly/greedy-based selected sets of covariances. For random selection, the frequencies of 20 samples are depicted for various numbers of sample covariances.



Supplement Figure 2: **Number of edges in the simulated pathways.** The number of edges in the simulated pathways is compared to the number of edges in scale-free networks of the same sizes. Parameter γ in Section 4.1.1 is set to 2 to calculate the upper bound on the number of edges in scale-free networks. (inset) The distribution of the average degree in the simulated pathways.

2 Comparison of the published biochemical reaction networks to scale-free networks of same sizes

We calculated the size of the s2MA-1 (that is the number of edges in the network) for scale-free networks of the same size as the studied pathways. Supplement Figure 2 shows that for large networks, the s2MA-1 of scale-free networks is larger than the s2MA-1 of the studied biological pathways. These results suggest that the scale-free assumption can provide a safe upper bound for the connectivities/degree distribution in biochemical reaction networks. Also, to verify the local connectivities assumption, we calculated the average degree of a node in the pathways. Supplement Figure 2 (inset) illustrates that, independently of the size of the network, the average degree hardly exceeds 10.

3 List of published pathways

The Supplement Table 1 provides the list of published biochemical networks that are used for the scalability analysis of the s2MA.

4 Network motifs

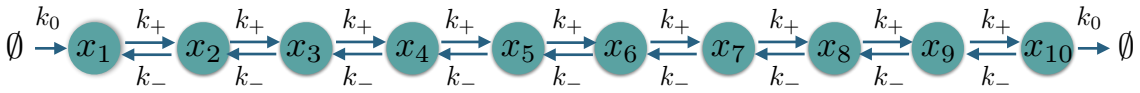
Illustrations of the considered *chain of monomolecular reactions* and *sequence of bimolecular reactions with a hub* are provided by Supplement Figures 3 and 4, respectively. The parameter values and initial conditions used for the simulation are listed in Supplement Tables 2 and 3.

References

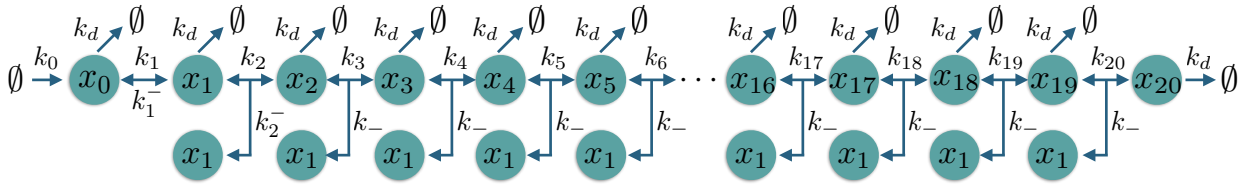
Bachmann, J., Raue, A., Schilling, M., Böhm, M. E., Kreutz, C., Kaschek, D., Busch, H., Gretz, N., Lehmann, W. D., Timmer, J., and Klingmüller, U. (2011). Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol. Syst. Biol.*, **516**(7).

Supplement Table 1: Published models used for the scalability analysis of s2MA.

	Name	**Identifier**	**Source**
1	Androgen receptor (AR) signalling pathway	NetPath_2	NetPath database
2	B Cell Receptor (BCR) signalling pathway	NetPath_12	NetPath database
3	Brain-derived neurotrophic factor (BDNF) signalling pathway	NetPath_76	NetPath database
4	Corticotropin-realising hormone (CRH) signalling pathway	NetPath_129	NetPath database
5	Epidermal growth factor receptor (EGFR1) signalling pathway	NetPath_4	NetPath database
6	Fibroblast growth factor-1 (FGF1) signalling pathway	NetPath_134	NetPath database
7	Gastrin signalling pathway	NetPath_154	NetPath database
8	Hedgehog signalling pathway	NetPath_10	NetPath database
9	Interleukin-2 (IL-2) signalling pathway	NetPath_14	NetPath database
10	Interleukin-3 (IL-3) signalling pathway	NetPath_15	NetPath database
11	Interleukin-4 (IL-4) signalling pathway	NetPath_16	NetPath database
12	Interleukin-5 (IL-5) signalling pathway	NetPath_17	NetPath database
13	Interleukin-6 (IL-6) signalling pathway	NetPath_18	NetPath database
14	Interleukin-7 (IL-7) signalling pathway	NetPath_19	NetPath database
15	Interleukin-9 (IL-9) signalling pathway	NetPath_20	NetPath database
16	Interleukin-10 (IL-10) signalling pathway	NetPath_132	NetPath database
17	Interleukin-11 (IL-11) signalling pathway	NetPath_147	NetPath database
18	Kit Receptor signalling pathway	NetPath_6	NetPath database
19	Leptin signalling pathway	NetPath_22	NetPath database
20	Notch signalling pathway	NetPath_3	NetPath database
21	Prolactin signalling pathway	NetPath_56	NetPath database
22	Receptor activator of nuclear factor kappa-B ligand (RANKL) signalling pathway	NetPath_21	NetPath database
23	T Cell Receptor (TCR) signalling pathway	NetPath_11	NetPath database
24	Transforming growth factor beta (TGF-beta) receptor signalling pathway	NetPath_7	NetPath database
25	Tumor necrosis factor (TNF) alpha signalling pathway	NetPath_9	NetPath database
26	Thyroid-stimulating hormone (TSH) signalling pathway	NetPath_23	NetPath database
27	Thymic stromal lymphopoietin (TSLP) signalling pathway	NetPath_24	NetPath database
28	TIE2/TEK signalling pathway	NetPath_138	NetPath database
29	Wnt signalling pathway	NetPath_8	NetPath database
30	Carbon metabolism	BIOMD0000000051	Biomodels database
31	E. coli metabolic adaptation	BIOMD0000000244	Biomodels database
32	Influenza virus replication	BIOMD0000000463	Biomodels database
33	TNF signalling network	BIOMD0000000407	Biomodels database
34	Yeast pheromone pathway	BIOMD0000000032	Biomodels database
35	Degradation of beta-catenin by the destruction complex	R-HSA-195253.1	Reactome database
36	DNA replication	R-HSA-69306	Reactome database
37	Interferon gamma signalling pathway	R-HSA-877300	Reactome database
38	Interferon alpha/beta signalling pathway	R-HSA-909733	Reactome database
39	Cholesterol biosynthesis	R-HSA-191273	Reactome database
40	DAG and IP3 signalling	R-HSA-1489509	Reactome database
41	Growth hormone receptor signalling	R-HSA-982772	Reactome database
42	Inositol phosphate metabolism	R-HSA-1483249	Reactome database
43	Integrin alphaIIb beta3 signalling	R-HSA-354192	Reactome database
44	ISG15 antiviral mechanism	R-HSA-1169408	Reactome database
45	Meiotic recombination	R-HSA-912446	Reactome database
46	Peroxisomal lipid metabolism	R-HSA-390918	Reactome database
47	PIP3 activates AKT signalling	R-HSA-1257604	Reactome database
48	RAF-independent MAPK1/3 activation	R-HSA-112409	Reactome database
49	RAF/MAP kinase cascade	R-HSA-5673001	Reactome database
50	JAK2/STAT5 signalling pathway	-	Bachmann *et al.* (2011)



Supplement Figure 3: **The schematic of the chain of monomolecular reactions**



Supplement Figure 4: **Schematic of the sequence of bimolecular reactions with a hub.**

Supplement Table 2: Parameter values and initial conditions used in the simulation of the chain of monomolecular reactions.

k_0	k_+	k_-	$[x_1](0)$	$[x_2](0)$	$[x_3](0)$	$[x_4](0)$	$[x_5](0)$	$[x_6](0)$	$[x_7](0)$	$[x_8](0)$	$[x_9](0)$	$[x_{10}](0)$
0.5	1	0.2	10	0	0	0	0	0	0	0	0	0

Supplement Table 3: Parameter values and initial conditions used in the simulation of the sequence of bimolecular reactions with a hub.

k_0	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}	k_{11}	k_{12}	k_{13}	k_{14}
100	5	1	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.2	0.21	0.22	0.23

k_{15}	k_{16}	k_{17}	k_{18}	k_{19}	k_{20}	k_1^-	k_2^-	k_-	k_d	$[x_0](0)$	$[x_1](0)$ to $[x_{29}](0)$
0.24	0.25	0.26	0.27	0.28	0.29	0.1	0.01	0.001	0.001	50	0