Stochastic analysis and control methods for molecular cell biology

Submitted for the degree of Doctor of Philosophy

by

Eszter Lakatos

Imperial College London

Centre for Integrative Systems Biology and Bioinformatics

Department of Life Sciences

Typeset in LaTeX.

This report is the result of my own work.

No part of this dissertation has already been, or is currently being submitted by the author for any other degree or diploma or other qualification.

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

This dissertation does not exceed 100,000 words, including footnotes and excluding appendices, bibliography, tables and equations.

This work is supported by the Schrödinger Scholarship and completed in the Department of Life Sciences at Imperial College, London.

Abstract

Stochastic fluctuations in cellular processes often have a dominant role in evoking heterogeneous behaviour in otherwise identical cells. From a modelling point of view, stochastic systems require specialised techniques to analyse and provide predictions about the underlying processes. Here we approach this problem by presenting several methods for the analysis of stochastic networks in molecular biology.

First, we address the problem of obtaining simulated trajectories of such systems. This task can be computationally prohibitively expensive and a range of approximation schemes have been proposed to lighten the burden. Here, we develop a set of multivariate moment closures that allow us to describe the stochastic dynamics of nonlinear systems. We use multivariate Gaussian, lognormal, and gamma closure to capture how correlations between different molecular species, induced by the reaction dynamics, interact with stochastic effects. In addition to the approximation method we also introduce our Python-based package that provides numerous tools to help non-expert users in stochastic analysis.

Next, we present an approach to quickly evaluate the levels of noise in both naturally evolved and engineered biological systems. Controlling the behaviour of cells by rationally guiding processes is an overarching aim of recent research, but stochasticity proves a formidable obstacle in obtaining reliable models and designs. Here, we compute the states reachable by a stochastic system under some biologically relevant control input in cases with uncertainty about parameter values. Our algorithm provides a set of bounds on molecular dynamics at the level of population

statistics that can be used for evaluating competing models and control strategies.

In the third part, we introduce a formalism to study cellular decision making influenced by noise. Understanding the heterogeneity of phenotypes within a cell population is a long-time challenge in modelling, and Waddington's notion of a developmental landscape has become a powerful conceptual framework for this problem. Here, we derive quantitative potential landscapes that allow us to explore the global dynamics of the system. Using this method we study how different sources of noise and modifications in experimental conditions can induce significant changes in the landscape and the assumed cell fates.

Finally, we propose a framework for model selection taking into account the inherent stochasticity of biochemistry. One of the main objectives of mathematical biology is identifying the most appropriate models and parameter values to describe known experimental results and provide predictions. Here, we develop and apply a Bayesian computational framework to identify dominant mechanisms controlling the variation of protein abundance between cell types. Using single cell expression of p53 in two cell lines, we compare the effect of transcription and protein degradation on the steady state distribution.

Acknowledgements

First of all, I would like to thank my supervisor, Michael Stumpf, without whose guidance this work would have never been completed. Thank you for all the academic and everyday advice and the delightful conversations throughout these past years. I am fortunate to have someone so supportive, approachable and relaxed to help me in my research.

I also want to thank all past and present members of the Theoretical Systems Biology group I had the pleasure to share an office with – you managed to create exactly the motivating environment I dreamt of doing my PhD in. Juliane, thank you for being such a brilliant and strong example, and for all those hour-long discussions on science and life. I hope life will bring you all the success and recognition you deserve, and I get to work next to you again. Anderson, Rav, Delphine, Adam, Patrick, Thalia, Ann, Joe, Quentin, thank you for being rather friends than workmates. I will miss our lunch breaks and spontaneous afternoon pints and picnics.

A huge thanks to all my friends back home and here who helped me to settle and feel at home in London. Mara, I cannot express in words how much it means having you right next door in all times of personal or professional emergencies. I dread the day when our tea-breaks stop, but I know we will be there for each other, and will find a sauna to meet up for sure. Miki, Dia, Dénes, Nóri, Anna, Arnold, Heni and everyone from the Meggysör group, you have all been a tremendous support at points of hardship in these three years. I have been more than lucky to be surrounded by such amazing people. All my friends from LIGC, thanks for the immense fun of

singing with you. I have learnt and gained so much musically and personally just by being part of this community. Finally, Marco, thank you for being my personal goofball and making sure I never stop smiling. I hereby promise, even though adult life is getting inevitably close, we will never grow up.

Last but not least, I am eternally grateful for my family for raising me with so much support and love. All I am and all I achieved so far was only possible due to you having my back in any situation. I can only hope that one day I might be able return a fraction of what you gave to me. Anyu, köszönök mindent és nagyon szeretlek!

Contents

List of Figures **x**

List of Tables **xi**

Acronyms and notations **xii**

1 Introduction **1**

2 Theoretical background **7**

2.1 Example systems 7

2.1.1 Linear gene expression 7

2.1.2 Dimerisation system 9

2.1.3 Oscillatory p53 system 12

2.2 Stochastic formulation 14

2.3 Stochastic simulation techniques 18

2.3.1 Gillespie's Stochastic Simulation Algorithm 20

2.3.2 Stochastic Differential Equations 22

2.3.3 Linear Noise Approximation 27

2.3.4 Moment Expansion Approximation 29

3 Multivariate moment closure techniques **35**

3.1 Introduction 35

3.2 Moment closure formulae 39

3.2.1 Multivariate normal closure 40

3.2.2 Multivariate lognormal closure 41

3.2.3 Multivariate gamma distribution . . . 42
3.2.4 Multivariate gamma closure . . . 43
3.3 Results . . . 48
3.3.1 P53 system . . . 48
3.3.2 Hes1 system . . . 52
3.3.3 Erk/Mek system . . . 55
3.4 MEANS . . . 60
3.5 Discussion . . . 64

4 Reachability analysis of control in stochastic systems 68

4.1 Introduction . . . 68
4.2 Zonotopes . . . 72
4.3 Reachable set computation . . . 73
4.3.1 Problem formulation . . . 73
4.3.2 Reachable set propagation . . . 75
4.3.3 Input for biological systems . . . 79
4.3.4 Parameter uncertainty . . . 80
4.3.5 Nonlinear systems . . . 82
4.3.6 Linearisation error . . . 83
4.4 Applications . . . 89
4.4.1 Controlled gene expression . . . 89
4.4.2 Model validation . . . 91
4.4.3 Bistable system . . . 94
4.5 Discussion . . . 97

5 Potential landscapes of stochastic cellular decision making 100

5.1 Introduction . . . 100
5.2 Landscape computation . . . 104
5.2.1 Exact stochastic computation . . . 104

5.2.2 Approximation through moment equations 106
5.3 Results 109
5.3.1 Schematic transcription factor model 109
5.3.2 Stem cell differentiation model 113
5.3.3 Time-scales of differentiation 116
5.3.4 Dynamical transition between states 118
5.4 Discussion 122

6 Analysis of p53 baseline regulation using model selection 125
6.1 Introduction 125
6.2 Methods 128
6.2.1 Data collection 128
6.2.2 Model and simulation 129
6.2.3 Model selection and inference 131
6.2.4 Distance measure 134
6.2.5 Synthetic experiments for validation 137
6.3 Results 139
6.3.1 Qualitative analysis 139
6.3.2 Model selection and inference results 141
6.3.3 Incorporating mRNA information 144
6.3.4 Absolute quantification experiments 146
6.4 Discussion 148

7 Summary and conclusions 151

Appendix A 154

Appendix B 158

Appendix C 160

Bibliography 164

List of Figures

2.1 Schematic illustration of example systems 11
2.2 Overview of stochastic simulation algorithms 19
2.3 Comparison of the Stochastic Simulation Algorithm and Stochastic Differential Equations 26
2.4 Comparison of Linear Noise Approximation and Moment Expansion Approximation 34

3.1 Two bivariate examples of the new multivariate gamma distribution 44
3.2 Approximation of the p53 oscillatory system 50
3.3 Approximation of the Hes1 oscillatory model 54
3.4 Evaluation of multivariate closures on the Hes1 model 55
3.5 Approximation of the simplified Erk/Mek model 58
3.6 Default graphical output of MEANS 63

4.1 Computation and propagation of zonotopes 73
4.2 Examples of reachable set computation 88
4.3 Reachable states of the controlled stochastic gene expression system 91
4.4 Model evaluation by comparing reachable sets and single measurements 93
4.5 Investigating bistability via reachable sets 96

5.1 Potential landscape and trajectories of the transcription factor model 111
5.2 Potential landscape of the stem cell differentiation model 115
5.3 Statistics of differentiation dynamics 118

5.4 Transition paths during differentiation 121

6.1 Schematic representation describing p53 expression in two cell lines . 130
6.2 Schematics of the ABC SMC algorithm 133
6.3 Distance measure used in the ABC algorithm 135
6.4 Analysis of the proposed models and validation experiments 140
6.5 Results of computational analysis 143
6.6 Model selection incorporating qPCR results 145
6.7 Microfluidic chip measurements of p53 expression level. 147

A1 Computational costs of moment closure methods 157
B1 Extrinsic noise in the transcription factor model 159
C1 Comparison of simulation techniques 161
C2 Evaluation of further distance measures 162
C3 Results of computational analysis on Data set II 163

List of Tables

3.1 Variables and parameter values used in the ERK/MEK system . . . 57

6.1 Parameter values used for synthetic data generation 138

6.2 Prior distributions used in the ABC algorithm 141

A1 Parameters explored for the Michaelis-Menten system 155

A2 Parameters explored for the p53 system 156

Acronyms and notations

Acronyms

ABC : Approximate Bayesian Computation

a. u. : arbitrary unit

CME : Chemical Master Equation

DNA : deoxyribonucleic acid

LNA : Linear Noise Approximation

mRNA : messenger ribonucleic acid

MVG : Multivariate Gamma

ODE : Ordinary Differential Equation

qPCR : quantitative Polymerase Chain Reaction

SDE : Stochastic Differential Equation

SSA : Stochastic Simulation Algorithm

Common notations

$\mathbf{x}$ **:** a multidimensional array, $[x_1, \ldots, x_N]$

$\dot{x}$ **:** time-derivative of x, $\frac{dx}{dt}$

S **:** stoichiometry matrix, where S_{ij} stands for the number of i molecules added in reaction j

$\mathbf{x} \in \mathbb{R}^n$ (or $M \in \mathbb{R}^{nxm}$) **:** a vector (matrix) with n elements (n rows and m columns) all of which are real numbers

$\mathbf{x}^T$ (or M^T) **:** transpose of the vector (matrix)

$x \in \{a, b\}$ **:** $x = a$ or $x = b$

$x \in [a, b]$ **:** x is in the interval between a and b, i.e. $a \leq x \leq b$

A^v **:** a vector containing elements of A above and in its diagonal, i.e the vector $[A_{11}, A_{12}, \ldots, A_{1N}, A_{22}, \ldots, A_{2N}, A_{33}, \ldots, A_{NN}]$

One

Introduction

Nature thrives on diversity. Survival of a population in a changing environment is secured through the diverse characteristics of its members. The ability to generate a variety of unique individuals under identical conditions provides a huge advantage. Therefore understanding how nature regulates and utilises variability is one of the major challenges in current biological research, especially in the mathematical modelling of biology.

Unfortunately, traditional, *deterministic* tools are often not sufficient to model and analyse the diversity observed in biology. They assume that the outcome of the same experiment is always the same and rely on modelling the average behaviour. However, as Darren Huff points out in terms of presenting information [1], knowing only the average value of a quantity can be practically uninformative or even misleading. To illustrate the challenges, let us consider a real-life example.

Stochasticity in an everyday situation

Assume you are at Marble Arch in London and want to take a bus to Kings Cross. Any travel planner can tell you that number 30 goes between the two places and the journey takes about 20 minutes. This is probably an accurate estimate of

the average travel time, but there are certainly quite a few buses that make this trip in more than 30 minutes – all it takes is to have one accident on the road. From your point of view the exact journey time has some amount of randomness, *stochasticity*, attached to it. So if you have a train to catch at Kings Cross, suddenly the average is not enough: you need to know more about the variability in journey lengths to guess how early you have to head off in order to make it in time. For example leaving 60 minutes for an allegedly 20 minute trip is definitely an overreaction. However, whether 30 or 35 is the more reasonable guess you can only tell if you travelled – and recorded the length – enough times on this particular route.

Now consider that bus is not the only way of transport from Marble Arch. You can also use the underground, but in that case you need to change between lines and therefore the average journey is rather 25 than 20 minutes. Nevertheless, the underground is not affected by the general traffic and you expect travel times to have less variability. So, depending on your priorities (e.g. how crucial is it to make your trip under 30 minutes) and how much information is known, you might make the counter-intuitive choice of going for a mode of transport that is on average slower.

Next, take the opposite case, when your arrival time is less important but you cannot leave earlier than 5pm. Due to the differences on bus route 30, you might catch a train at Kings Cross at 5.28 or have to take the next one 20 minutes later, which has more stops. Suddenly a three-minute delay with the bus gets amplified and becomes 40 minutes or more at your destination. If you make this trip every week, you find that the average journey takes a total of 55 minutes, but your actual travel times show substantial variability and a few particularly unlucky instances.

Now suppose that your only concern every time is whether you make it to your final destination before 6.15pm. The result of your 'travel experiment' is now binary. Despite no external differences in your journey (you always leave at 5pm, no bus/train breaks down, you never experience serious weather conditions) a major difference is observed in the outcome due to stochasticity. Clearly, in all the above

questions, simply considering the average journey is not an adequate description of your experience and there are many aspects of randomness to consider.

Stochasticity in biology

Biological processes at the cellular level are in many ways similar to our transport analogy. The average behaviour is frequently not a good descriptor of one particular cell. Often the range of values and the chance of going over some threshold is more crucial. Due to many interactions contributing to a process, variability can become amplified along the chain of reactions. The underlying networks, however, contain many more connections than one might encounter commuting through London, which increases the sources of variation even more. And quite often our observations are limited to some simple read-out of the state (e.g. did the cell become cancerous), like whether you got home before 6.15pm or not.

At the macroscopic scale chemical reaction systems are typically described using *mass-action kinetics*. These capture how the concentrations of chemical species evolve over time, the equivalent to characterising a system through the average behaviour. However, hundreds of experiments in biochemical systems report a strong cell-to-cell-variability that cannot be explained with such a description.

Stochasticity, referring to the random nature of some dynamical systems, stems from the fact that biological processes are driven by random collisions between macromolecules that can be only found in a few dozen or hundred copies in a single cell. Only at a scale where all chemical species are available in sufficiently high abundances, it becomes possible to use the assumption that the behaviour is governed by the speed of reactions, expressed by reaction rates. On the other hand, when mere handfuls of these molecules are distributed over relatively large volumes, it is not guaranteed that reactants of a specific step can be found in sufficient proximity and appropriate conformational state for the reaction to occur [2–4]. The movement of a macromolecule is described through a random process, hence the description of a

single cell becomes governed by probability [5].

Furthermore, there are several other mechanisms contributing to the non-deterministic nature of cellular dynamics. Thermal noise (e.g. fluctuations in the temperature inside the cell) increases inherent stochasticity of production and degradation reactions and results in further variation in the copy numbers of reactants. Random changes in the accessibility of DNA lead to stochastic bursts in the production of many other molecules that can result in the same effect [6]. From a modeller's point of view, randomness can also arise from processes that are impossible to model on the scale of complex biological system: for example influences from upstream components (such as morphogens or the extracellular matrix), which are often highly localised and are subject to noise themselves [7, 8]. All of these factors profoundly affect and shape cell-to-cell variability and heterogeneity in isogenic populations.

For these reasons we often require the explicitly stochastic formulation of chemical reaction systems, and cannot rely on purely deterministic approaches to model (bio-)molecular reaction systems: stochasticity is often an inalienable aspect of such systems [9–13].

The popularity of stochastic biological models has risen in the last few decades with the emergence of new measurement techniques allowing us to observe processes on the single cell level. However, following the dynamics of an entire stochastic system is still a very challenging task. There are several numerical techniques to overcome this problem, which vary in exactness and computational demand, and different aims will require different stochastic approaches. Fortunately, there are approximation techniques that allow stochastic kinetics to be transformed into a deterministic-like model where ordinary differential equations are used to describe how the system changes over time. This makes it possible to apply and modify popular deterministic approaches – developed for inference, system validation and control theory – as well as develop new methodologies to investigate and control cell-to-cell variability [14].

Outline of thesis

In this work we develop and apply stochastic modelling and analysis techniques to understand the origins and influence of noise in cellular processes. We focus on the population-level analysis of elementary macromolecular networks that show characteristics or hold functions very often seen in more complex biological systems. We are particularly interested in how the combination of the average behaviour and the intra-population variability can contribute to a final cellular decision-making process. Through the methods demonstrated in this work we hope to create tools that will lead to a better understanding of biological noise and the design of sophisticated control strategies to regulate crucial cellular processes.

The thesis is organised in the following way. In the next chapter we give an introduction to the mathematical formalism used throughout the thesis. We start with the description of elementary biochemical models. Using these models as examples, we present the basics of stochastic modelling. We also give an overview of stochastic simulation techniques with their strengths and weaknesses.

In chapter three, we present a moment closure technique that complements stochastic approximation methods. The algorithm provides a computationally efficient way to convert a stochastic problem into a deterministic one through a closed set of moment equations. Our technique is applicable to highly nonlinear systems with intermediate numbers of molecules. We demonstrate the advantages of our approximation over other closure methods on three examples. Last, we give a summary of the publicly available Python package implementing the method. Substantial parts of this chapter have been published in [15] and in [16].

The fourth chapter outlines an algorithm for analysing the achievable states of biochemical networks. The method explores the variability and possible limits of the system, based on moment equations generated by approximation techniques.

We develop the algorithm to take into account uncertainties commonly found in biochemistry. It considers unknown or controllable input of molecules, imprecisely measured initial conditions and parameter values and nonlinearity in the underlying dynamics. We apply reachability analysis to investigate noise in a controlled synthetic network, confirm bistability and evaluate a set of plausible models based on limited observations. Parts of this chapter are published and available online in [17].

In the fifth chapter we introduce the notation of potential landscapes to explore phenotypes arising in cellular decision making. We obtain the landscape both through exact stochastic simulations and moment-based approximation. We present a minimalistic model of differentiation and a less simplified model from literature. The landscape structure is explored in detail, together with the path of individual cells along distinguished features in the potential.

Chapter six describes a computational study on the mechanism underlying the differences observed between cell lines. We consider single cell measurements of p53 expression from two tumour cell types and formalise two competing models to explain the data. A Bayesian framework is used for model selection and inference of rate parameters in the models. We validate the method on a synthetic data set and present findings on the experimental observations. The results of this work have been published in [18].

Finally, in chapter seven we turn to the everyday life metaphor introduced in the first chapter and summarise the methods and findings presented in this work.

Two

Theoretical background

2.1 Example systems

First, we present three simple systems that illustrate basic building blocks and principles in biochemical modelling. Each embodies some property common amongst the systems we aim to investigate, but requires a mathematical treatment different from the others. Through these systems we show how a biological phenomenon is translated into a reaction network model. We also provide a 'traditional', ordinary differential equation (ODE) formulation of the systems as a basis of comparison. A stochastic mathematical description of these reaction system models is derived in the next section.

2.1.1 Linear gene expression

The first and most elementary example is the stochastic gene expression system, which is a general description of the basic processes behind the intracellular production of an arbitrary protein. The reactions captured here are present in all biochemical systems, although sometimes not modelled in detail. The model captures the central dogma of biology in the following aspect. A gene encoded in our

DNA is transcribed to form mRNA molecules in the nucleus that are then transported into the *cytosol* (the 'living room' of the cell) and subsequently translated into protein molecules in one step. We assume that the amount or state of DNA is unchanged, as well as that molecules for translation (e.g. ribosomes) are in excess, hence these molecules are not modelled explicitly. The steps of protein production can be summarised by a reaction system of four processes:

$$\begin{aligned} \emptyset &\xrightarrow{k_1} \text{mRNA} \quad \text{(transcription)}; \\ \text{mRNA} &\xrightarrow{k_2} \emptyset \quad \text{(mRNA degradation)}; \\ \text{mRNA} &\xrightarrow{k_3} \text{mRNA} + \text{protein} \quad \text{(translation)}; \\ \text{protein} &\xrightarrow{k_4} \emptyset \quad \text{(protein degradation)}. \end{aligned} \tag{2.1}$$

Constants above each arrow are *reaction rates*, numerical attributes describing the inherent speed of the process. The symbol $\emptyset$ represents that a molecule is produced from (or degraded into) 'nothing', a substance not modelled in the representation: for example the individual molecules used to assemble an mRNA macromolecule. Figure 2.1(a) shows a schematic depiction of the modelled processes.

As the left-hand side of the above reactions always contains only a single (type of) molecule, the terms describing how the system progresses over time are *linear*. For example the relationship between protein molecules produced at a certain time is proportional to the mRNA molecules in the system. Any reaction system that has only zeroth-order (a molecule is created from 'nothing') and first-order (a molecule is transformed into some other molecule(s)) reactions, will be linear. Linear systems have many advantageous properties in terms of analysis, but many biological phenomena cannot be formulated this way.

A deterministic depiction is applied for traditional scenarios when there is a high number of the above molecules in close proximity of each other. Then it can be assumed that the amounts of molecules is well captured by their concentration in a unit volume (denoted by [mRNA] and [protein]) and reactions are continu-

ously happening. We describe the model by writing up the rate of change for each species in a set of *ODE*s that incorporates all reactions decreasing or increasing the concentration. Given these rules and some starting values, one can compute the concentrations at any time, hence this formalism is widely used to describe a model. As an ODE summarises all reactions affecting the abundance of a species, the number of equations equals the number of species, which is typically lower than the number of reactions. For example the set of two differential equations,

$$\begin{aligned} \frac{d[\text{mRNA}]}{dt} &= k_1 - k_2 \cdot [\text{mRNA}], \\ \frac{d[\text{protein}]}{dt} &= k_3 \cdot [\text{mRNA}] - k_4 \cdot [\text{protein}], \end{aligned}$$

gives a deterministic form of the reaction system in Eq. (2.1).

2.1.2 Dimerisation system

Our second system, similarly to the one considered above, is a general model. It describes the reversible interaction of two molecules, *monomers*, forming a new macromolecule. Such systems are very common in the cell and many of the crucial responses are achieved by formation of a dimer that can carry out a different function than its monomers. Further interactions with identical or different monomers lead to *polymers*, each step of which can be captured as a dimerisation event. Some examples of polymers are: antibodies that contain several immunoglobulin units; alcohol dehydrogenase is a dimer; fibrinogen, the main player in blood clotting, is a hexamer. Here we consider the case of a dimer produced from identical monomers.

The system in some sense is simpler than the linear gene expression one: there are only two types of molecules (monomer and dimer) and two reactions (dimer formation and dissociation) needed to capture the dynamics (Fig. 2.1(b)). Formally

the reaction system is

$$\begin{aligned} \text{monomer} + \text{monomer} &\xrightarrow{k_1} \text{dimer (association)}, \\ \text{dimer} &\xrightarrow{k_2} \text{monomer} + \text{monomer (dissociaton)}. \end{aligned} \tag{2.2}$$

And the corresponding ODE model is

$$\begin{aligned} \frac{d}{dt}[\text{monomer}] &= -2k_1 \cdot [\text{monomer}]^2 + 2k_2 \cdot [\text{dimer}] \\ \frac{d}{dt}[\text{dimer}] &= k_1 \cdot [\text{monomer}]^2 - k_2 \cdot [\text{dimer}]. \end{aligned}$$

As the first reaction requires *two* molecules of the monomer to meet and interact, the rate of dimer production will not be a function of monomer concentration, but rather of the square of monomer amount, as apparent from the deterministic formalism. For any system that has a second-order (some molecules are produced through the interaction of two individual molecules) reaction, there is an inevitable *nonlinearity* in the system dynamics. Nonlinearity is an intrinsic feature of many biological systems, however, mathematical analysis of nonlinear systems might lead to complications and requires special techniques to tackle phenomena (for example oscillatory or chaotic behaviour) arising through such quadratic dependence. A commonly used technique therefore is the linearisation of the system: establishing a linear counterpart of the system that can be more easily analysed for characteristics such as stability and parameter dependence. This provides a good estimate of the original system in the proximity of a certain point but might not be used for the entire range of values we are interested in.

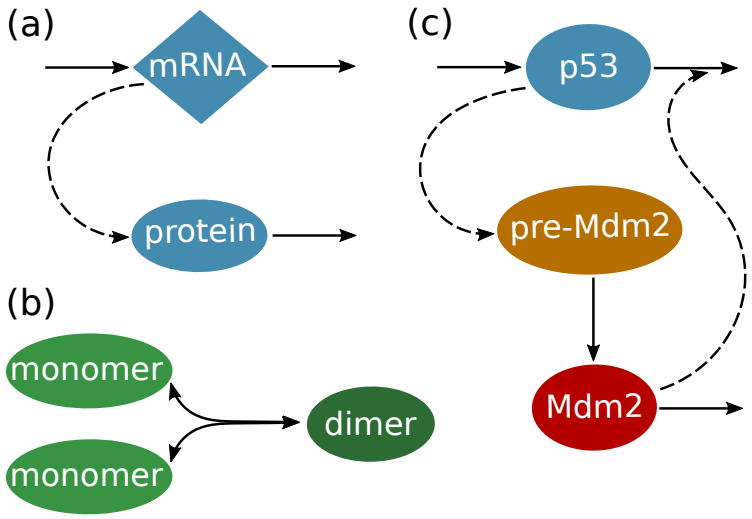


Figure 2.1: Schematic illustration of example systems. **(a)** Linear gene expression system. **(b)** Dimerisation system. **(c)** Oscillatory p53-Mdm2 system. Filled geometric shapes represent macromolecules in all subfigures. Arrows denote biochemical reactions: the arrow pointing towards or away from a macromolecule means production or degradation of the species, respectively. Dashed arrows indicate that the molecule on the starting end of the arrow is not affected by the reaction. For example in (c) the number of p53 molecules is the same after production of pre-Mdm2, but one molecule of pre-Mdm2 is 'used' in the production of each molecule of Mdm2.

2.1.3 Oscillatory p53 system

In our last example we consider a concrete biochemical network that shows a behaviour not covered by the previous examples: oscillations. Oscillations govern many time-sensitive aspects of biology, such as the cell cycle and circadian rhythms. We study the tumour suppressor protein p53, whose many functions include reaction to DNA damage by activating repair mechanisms or initiating cell death. Its malfunction can lead to cells with unhealthy genetic material proliferating and dividing. As a consequence more than 50% of cancer cases are associated with non-healthy p53 function or expression [19].

An oscillatory, pulsing production of p53 is essential for the functionality of the system. Several models have been developed that can show such behaviour, operating on different mathematical principles. Here we use model IV from [20], which focuses on the interaction between p53 and its inhibitor, Mdm2. The production of this protein is through an intermediate molecule, a precursor protein whose production is initiated by p53. Mdm2, in turn, speeds up p53 degradation, reducing its own production indirectly. It is an example of a *negative feedback loop*, as an increase of p53 ultimately counteracts itself, as illustrated in Figure 2.1(c). The model can be summarised with six reactions,

$$
\begin{aligned}
\emptyset &\xrightarrow{k_1} \text{p53} && \text{(p53 production)}\\
\text{p53} &\xrightarrow{k_2} \emptyset && \text{(Mdm2-independent p53 degradation)}\\
\text{p53} &\xrightarrow{k_3,\ k_7,\ \text{Mdm2}} \emptyset && \text{(Mdm2-dependent p53 degradation)}\\
\text{p53} &\xrightarrow{k_4} \text{p53} + \text{pre-Mdm2} && \text{(p53-activated precursor production)}\\
\text{pre-Mdm2} &\xrightarrow{k_5} \text{Mdm2} && \text{(Mdm2 maturation)}\\
\text{Mdm2} &\xrightarrow{k_6} \emptyset && \text{(Mdm2 degradation)},
\end{aligned}
$$

where a species included next to a reaction rate represents that a particular species affects the speed of the reaction without its own amount changing in the process.

For example, in the third reaction, the rate of Mdm2-dependent degradation is

$$k_3 \frac{\text{Mdm2} \cdot \text{p53}}{\text{p53} + k_7},$$

so the decrease of p53 follows a sigmoidal curve and the more molecules of Mdm2 are present, the faster the decay becomes. The same form appears in the deterministic depiction of the system

$$\begin{aligned} \frac{d[\text{p53}]}{dt} &= k_1 - k_2 \cdot [\text{p53}] - k_3 \frac{[\text{Mdm2}] \cdot [\text{p53}]}{[\text{p53}] + k_7} \\ \frac{d[\text{pre-Mdm2}]}{dt} &= k_4[\text{p53}] - k_5[\text{pre-Mdm2}] \\ \frac{d[\text{Mdm2}]}{dt} &= k_5[\text{pre-Mdm2}] - k_6[\text{Mdm2}]. \end{aligned}$$

Similar reactions, following so-called *Hill-type* kinetics [21], are often used as a shorthand instead of dividing the process into several reactions between intermediary complexes. For example if the association/dissociation of these complexes is considerably faster than the reaction they assist or it is at quasi-equilibrium, it is sufficient to only include them implicitly through a sigmoidal function. Note that even though at first glance there are no bimolecular reactions in the above system, the Hill-type reaction introduces a strong nonlinearity – hence the capability to show oscillations – and the model has to be treated accordingly.

2.2 Stochastic formulation

In this section we derive a stochastic model for systems described as a biochemical reactions system. Note that the major point of such a model is *randomness*: at any time we can only make statements about the *probability* of molecular abundances and reactions, not the exact values.

Let us consider a system of N chemical species $(X_1, \ldots, X_N)$ and R reactions, where the lth reaction corresponds to the process

$$\underline{s}_{l,1}X_1 + \cdots + \underline{s}_{l,N}X_N \quad \xrightarrow{k_l} \quad \overline{s}_{l,1}X_1 + \cdots + \overline{s}_{l,N}X_N. \tag{2.3}$$

Here the rate k_l (the stochastic counterpart of a macroscopic reaction speed) relates to the probability of the lth reaction, during which the number of X_i molecules changes from $\underline{s}_{l,i}$ to $\overline{s}_{l,i}$. In a certain time the system is described by its probability $P(\mathbf{x})$ to contain $x_1, \ldots, x_N$ copies of molecules of each species. The actual probability of reaction l occurring is given by its reaction propensity, $a_l(\mathbf{x})$, which consists of the rate parameter k_l and a – typically low-degree polynomial – function of the amount of molecules required for the reaction to happen. Using this quantity, the time evolution of the system is formulated by the Chemical Master Equation (CME) [22, 23],

$$\frac{dP(\mathbf{x})}{dt} = \sum_{l=1}^{R} P(\mathbf{x} - \mathbf{s}_l)a_l(\mathbf{x} - \mathbf{s}_l) - P(\mathbf{x})a_l(\mathbf{x}), \tag{2.4}$$

where $\mathbf{s}_l = \overline{\mathbf{s}}_l - \underline{\mathbf{s}}_l$ is the vector of stoichiometric coefficients of the species, describing the total number of molecules produced or consumed in reaction l. The first term in each summand corresponds to the event that the system is in a different state and, through reaction l, arrives to $\mathbf{x}$, while the second term is the probability of the system moving away from state $\mathbf{x}$ through reaction l. As Eq. (2.4) shows, a system can be completely characterised by a set of reaction propensities and a stoichiometry

matrix collecting the above $\mathbf{s}_l$ vectors,

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \dots & a_r \end{bmatrix},$$
$$S = \begin{bmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \dots & \mathbf{s}_r \end{bmatrix}.$$

Similarly to an ODE depiction, the CME – or a stoichiometry matrix and reaction propensities – and some starting distribution provide a complete description of the system, which in theory makes it possible for the probability distribution to be computed for any time.

In many cases there are already existing deterministic models for a given system, with all parameters known in terms of concentrations, but the situation requires a stochastic treatment. Therefore, values of reaction rates and initial states have to be converted to absolute molecular abundances. The transition is provided by the system size, Ω, calculated as the number of molecules 1 M concentration represents in a system. Hence it is closely related to the physical volume of the system, e.g. a cell or organelle. Initial numbers of species (x) from concentrations ($[X]$) are obtained as $x = [X] \times \Omega$. The conversion of reaction rates (c), depending on their order, is done as follows:

$$\begin{aligned} \text{Zeroth-order: } k &= c \times \Omega; \\ \text{First-order: } k &= c; \\ \text{Second-order: } k &= c/\Omega. \end{aligned}$$

Although in theory higher order reactions should be broken down to a system of maximum second-order reactions, one can use simplifications, such as for example the quasi-steady-state assumption [24], and obtain the appropriate multiplying factor in terms of Ω for more complex reactions, based on that the unit of k is always $1/s$.

Example - Linear gene expression

First, take the linear model of protein expression (Eq. (2.1)) and derive the corresponding stochastic model. In this case the system has $N = 2$ species ($\mathbf{x} = [m, p]^T$, representing mRNA and protein numbers) and $r = 4$ reactions. The first reaction, transcription, will be formalised as

$$0 \cdot m + 0 \cdot p \xrightarrow{k_1} 1 \cdot m + 0 \cdot p,$$

with k_1 either taken directly from a stochastic rate value or derived from the macroscopic transcription rate (c_1), as $c_1 \cdot \Omega$. In an arbitrary human cell which has a volume of 2.4 pl, Ω and the equivalent of 0.1 μM concentration become,

$$\Omega = 6.023 \cdot 10^{23} \cdot 2.4 \cdot 10^{-12} = 1.44 \cdot 10^{11} \text{ [molecules/M]}$$

$$0.1\ \mu M = 10^{-7} \cdot \Omega = 1.44 \cdot 10^{4} = 14400 \text{ [molecules]}.$$

With k_i representing the stochastic rates of each reaction, the reaction propensities and the corresponding stoichiometry matrix are:

$$a_1 = k_1\ ;\ a_2 = k_2 m\ ;\ a_3 = k_3 m\ ;\ a_4 = k_4 p; \tag{2.5a}$$

$$S = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}; \tag{2.5b}$$

from which the CME is easily formed as

$$\begin{aligned} \frac{dP(\mathbf{x})}{dt} = \frac{dP\left(\begin{bmatrix} m \\ p \end{bmatrix}\right)}{dt} &= P\left(\begin{bmatrix} m-1 \\ p-0 \end{bmatrix}\right) k_1 - P\left(\begin{bmatrix} m \\ p \end{bmatrix}\right) k_1 + \\ &+ P\left(\begin{bmatrix} m+1 \\ p \end{bmatrix}\right) k_2(m+1) - P\left(\begin{bmatrix} m \\ p \end{bmatrix}\right) k_2 m + \\ &+ P\left(\begin{bmatrix} m \\ p-1 \end{bmatrix}\right) k_3 m - P\left(\begin{bmatrix} m \\ p \end{bmatrix}\right) k_3 m + \\ &+ P\left(\begin{bmatrix} m \\ p+1 \end{bmatrix}\right) k_4(p+1) - P\left(\begin{bmatrix} m \\ p \end{bmatrix}\right) k_4 p. \end{aligned}$$

Example - Dimerisation

To consider a nonlinear example system as well, let us derive the above descriptive terms for the dimerisation system. The system is consisted of two species (m and d, representing number of monomers and dimers, respectively) and two reactions. The deterministic rate for the association (dimer formation) reaction would be $c_1 \cdot m^2$. However, as we assume we have a strongly limited pool of molecules, we have to account for the selection of two molecules, which we do by using *sampling without replacement*. Therefore the reaction propensitiy of the stochastic reaction becomes $c_1 \cdot m(m-1)/2$. For simplicity let us assume that the stochastic rate parameter, k_1. accounts for both Ω and the $1/2$ multiplier (i.e. $k_1 = c_1 \cdot \Omega/2$), hence the stochastic dimerisation system is described by the following stoichiometry matrix and reaction propensities,

$$S = \begin{bmatrix} -2 & 2 \\ 1 & -1 \end{bmatrix}; \quad a_1 = k_1 m(m-1); \; a_2 = k_2 d. \tag{2.6}$$

From which the CME takes a relatively simple form, as there are only two reactions that might happen at a time-point,

$$\begin{aligned} \frac{dP\left(\begin{bmatrix} m \\ d \end{bmatrix}\right)}{dt} &= P\left(\begin{bmatrix} m+2 \\ d-1 \end{bmatrix}\right) k_1(m+2)(m+1) + P\left(\begin{bmatrix} m-2 \\ d+1 \end{bmatrix}\right) k_2(d+1) + \\ &- P\left(\begin{bmatrix} m \\ d \end{bmatrix}\right) \left(k_1 m(m-1) + k_2 d\right). \end{aligned}$$

2.3 Stochastic simulation techniques

The most general description of stochastic dynamics is typically provided by the CME. However, albeit the CME offers an exact description, it can only be solved for a few, highly idealised cases; therefore other approaches are needed to simulate and predict the system's behaviour. A wide range of methods have been developed that cover the spectrum from exact but slow simulations to fast ODE systems that ignore randomness (see Fig. 2.2) [25].

On one extreme of the range is Gillespie's Stochastic Simulation Algorithm (SSA) [26], which offers a direct numerical simulation of the system to obtain a single exact realisation of the stochastic process modelled. But because of its computational cost it is impossible to model the behaviour of a large population of cells, or of even moderately complex reaction systems. To lighten the burden, several slight variations of the SSA are available that offer improvement in simulation speed especially in dense systems, such as tau-leaping [26,27] and the conditional difference method [28]. Other techniques divide the system dynamics into a deterministic part and a stochastic part to avoid direct simulation of fast reactions. Hybrid methods, for example, divide the species into low and high abundance groups, and model only slow reactions stochastically [29]. Stochastic Differential Equations (SDE), treat changes in mean concentration deterministically and add a noise term to model fluctuations corresponding to each reaction.

The next level of approximation is represented by the Linear Noise Approximation (LNA) [30–32], which similarly separates the trend (drift) and a noise term, and simulates these two as an ODE and an SDE, respectively, in a system with 100-1000 molecules [33]. For smaller but complex reaction systems, which are still practically intractable with the SSA, expansion methods have been developed that capture the dynamics throughout the time evolution of moments of the probability distribution. Most approaches focus on moments of a specific order, for example the so-called

Mass Fluctuation Kinetics (MFK) [34] and the 2MA and 3MA [35] approaches include expressions for the first, second and third central moments. Similarly, the LNA can provide a computationally efficient method for tracking the mean and the variance in systems with non-oscillatory behaviour. Ale et al. [36] developed a moment expansion approximation (MEA) method to automatically generate equations for arbitrary order central moments in a computationally efficient manner. This method is suitable to capture the general behaviour of stochastic systems with complex dynamics and 10-100 molecules, which conventional MFK approaches typically fail to reproduce adequately. The major benefit of this method is nonetheless its ability to approximate highly nonlinear problems. For systems involving only zeroth- and first-order reactions, the evolution of moments can be directly obtained by taking the expected value of the CME – resulting in the same equations can derive through any approximation approach.

In the next sections we give a quick overview of the various techniques available to simulate a stochastic biochemical network. To explore a range of exactness and computational efficiency, we present SSA, SDE, LNA and MEA and demonstrate how are they used as a basis for further analysis.

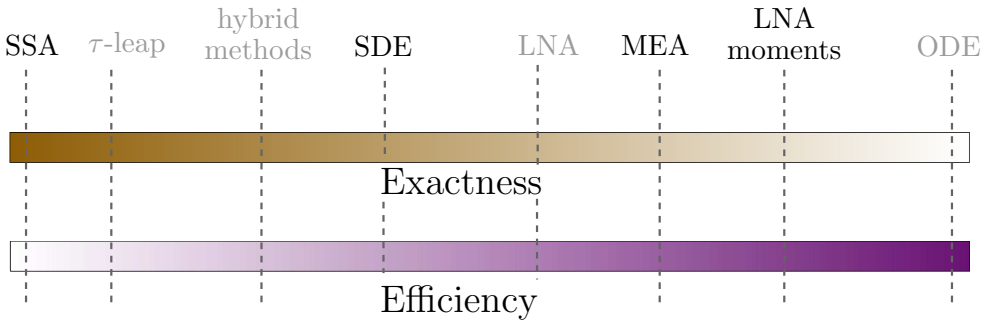


Figure 2.2: Schematic spectrum of stochastic simulation methods, ranging from Gillespie's exact SSA to concentration-level depiction by ODEs. Keep in mind that the exactness/efficiency scale is only for illustration – actual accuracy and computational cost always depend on the problem. Methods described in the thesis are highlighted in black.

2.3.1 Gillespie's Stochastic Simulation Algorithm

Daniel Gillespie developed an algorithm for direct simulation of processes like the CME that accounts for the probabilistic nature of reaction resulting from low molecular numbers and thermal noise by simulating every *jump* or reaction of the process [26]. There are two different formulations for this exact simulation of the CME: the so-called *direct method* and the *first reaction method*. They are mathematically equivalent, but their computational emphasis is different, hence in certain situations one might prove to be less computationally demanding than the other. We summarise the steps of the direct method below.

The heart of the algorithm lies in the fact that a sample path of the CME is technically a Poisson process, and as such, the time between successive reactions will be given by an exponential distribution. The simulation operates by sampling the times between events to map every reaction that happens between a certain starting and final time. The sampling is implemented by the generation of one uniformly (r_l) and one exponentially distributed random number (r_t) in every step to determine which reaction occurs next and the waiting time for the event, respectively. The probability of a given reaction to be chosen is described by its reaction propensity (a_l) compared to the sum of all reaction propensities, while the time for the reaction to occur is obtained from the total summed reaction propensity, i.e. the probability of *any* event to happen. Therefore both samples are directly influenced by the total amount of molecules. Once the next reaction is selected, we increase the time with the generated waiting time and update the molecule count according to the reaction that occurred. The computational steps of the algorithm are as follows:

1. Initialise system, $t = 0$, $\mathbf{x}(0) = x_0$
2. Repeat while $t \leq T_{final}$
   (a) Compute reaction propensities $a_l(\mathbf{x}(t))$ and total propensity $a_{tot} = \sum_l a_l(\mathbf{x}(t))$

(b) Generate random numbers $r_l \sim U(0, a_{tot})$ and $r_t \sim Exp(a_{tot})$

(c) Choose reaction i to occur:

reaction 1 if $r_l \leq a_1$,

reaction 2 if $a_1 < r_l \leq a_1 + a_2$,

$\vdots$

reaction R if $a_{tot} - a_R < r_l \leq a_{tot}$

(d) Update number of molecules with stoichiometry vector, $\mathbf{x}(t) \rightarrow \mathbf{x}(t) + \mathbf{s}_i$

(e) Update time $t \rightarrow t + r_t$

This algorithm provides a *single realisation* of the process that might be substantially different from a second single simulation. Therefore, for most applications a whole population of realisations is needed to obtain a good depiction of the underlying system. With many repeats of the above algorithm required, the computational time of the algorithm becomes a crucial issue.

In an exponential distribution, the higher the parameter value of the distribution is, the smaller samples drawn from this distribution become. Therefore, high molecular abundances increasing a_{tot} directly result in smaller steps, r_t, and a lot of iterations are required to simulate the system until T_{final}. In other words, as every chance encounter of appropriate molecules is sampled, a high number of molecules will place a high load on the sampler and the SSA becomes infeasible.

Although the SSA provides an *exact* solution to the CME, it is important to note that there are certain assumptions made when using this method to represent a physical system. As there is no space considered in the propagation of the system, the simulation is physically accurate only for dilute and well-mixed (i.e. the reactants have a uniform spatial distribution) environments. Throughout the thesis we assume that the systems in question exhibit these properties to the extent that spatial effects are negligible. Nevertheless, methods combining the SSA with spatial partitioning and reaction-diffusion equations can be used to relax these assumptions [37].

2.3.2 Stochastic Differential Equations

Next we present the first method to accelerate the simulation of a stochastic system. The computational cost is reduced by approximating the evolution of the system through numerical solution of an Itô process [38]. This approximation relies on modelling the increments of each species over an infinitesimally small time-step (dt) as samples from an N-dimensional normal distribution. The general formula of a Stochastic Differential Equation (SDE) is

$$\mathrm{d}\mathbf{x}_t = \mu(\mathbf{x}_t)dt + \Lambda(\mathbf{x}_t)\mathrm{d}\mathbf{W}_t, \quad \mathbf{x}(t=0) = \mathbf{x}_0, \tag{2.7}$$

where $\mathbf{W}_t$ is an N-dimensional Wiener process – or in other words, a Brownian motion [5, 39]. μ and Λ are arbitrary functions in this formalisation which can be chosen to ensure the above equation truly represents the stochastic process. An interpretation of this form is that the change of the system's state over a small time interval is normally distributed with mean $\mu(\mathbf{x}_t)dt$ and variance $\Lambda(\mathbf{x}_t)^2dt$. Therefore we assume that, in the limit of high molecular abundances, the expected value and variance of Eq. (2.7) should match those derived from the CME solution through the central limit theorem. Hence the SDE form of a stochastic system with stoichiometry matrix S and reaction propensities a_i becomes

$$\mathrm{d}\mathbf{x}_t = S \cdot \mathbf{a}(\mathbf{x}_t) \cdot dt + S \cdot \mathrm{diag}\{\sqrt{\mathbf{a}(\mathbf{x}_t)}\}\mathrm{d}\mathbf{W}_t, \tag{2.8}$$

where $\mathrm{diag}\{v\}$ stands for the $\mathbb{R}^{N\times N}$ matrix with the elements of vector v in its diagonal and zeros elsewhere.

Setting dt to a fixed, small, but finite value, Δt, and iterating from $\mathbf{x}_0$ leads to the *Euler-Maruyama* approximation, a numerical integration technique that can be efficiently evaluated for even considerably complex systems. The algorithm consists of iteratively computing the value of Eq. (2.8) and updating the system's state until a final time is reached. The updates correspond to a regular grid of time intervals

and usually more than one reaction event – however, for a dilute system it is possible to have an update of just a fraction of a reaction's stoichiometry.

As with all numerical methods, this approximation also requires a trade-off between accuracy and efficiency, which can be regulated through the choice of time-step, Δt. Another important issue that has to be accounted for is that – due to the treatment of stochasticity in this formulation – there is no automatic constraint to stop $\mathbf{x}_t$ from assuming negative values. There are several boundary condition strategies to address jumps that would throw a species' amount to negative: (i) by reversing the sign of the Brownian jump; (ii) by setting the species count to zero instead; (iii) by rerunning the step until a valid jump is generated. Depending on the application, one might prefer a particular method, for example because (iii) can lead to an increase in the number of steps required for the overall situation. In the followings when using the Euler-Maruyama method, we always apply the second limit condition, as we interpret the temporary negative values as a 'shortage' of the species, to could be represented by zero copies.

Despite these trade-offs, through simulations on a variety of models we find that in practice, with an appropriately small Δt value, the difference between SDE and SSA simulations is negligible even when the species in consideration are on the magnitude of tens of molecules/cell [40].

Examples – Linear gene expression

SSA

First consider the model of stochastic gene expression in Eqs. (2.1) and (2.5). We demonstrate Gillespie's SSA by going through one cycle of the procedure described in 2.3.1. Let us assume the system currently at is $t = 0.5$ and has five mRNA and 20 protein molecules ($\mathbf{x}(0.5)^T = [5, 20]$), and rate values of reactions are $\mathbf{k} = [10, 2, 5, 1]$. The time-units are normalised according to the protein's average turnover time, so that we gain a general model that could be scaled to reflect a specific macromolecule with known production and degradation rates. Therefore the time is given in arbitrary units (a.u.) and k_1 has units of molecules/a.u., and all other k_i are in 1/a.u.. We will use this scaling technique frequently in later chapters to generalise example systems.

Given the above characteristics, first the propensities (dependent on the current state of the system) are calculated

$$a_1 = k_1 = 10;\ a_2 = k_2 x_1 = 2 \cdot 5 = 10;\ a_3 = 25;\ a_4 = 20;$$

$$a_{tot} = 10 + 10 + 25 + 20 = 65.$$

Then r_l and r_t are generated, which in practice is done via the transformation of two $U(0, 1)$ random variates, in this particular case $u_l = 0.28$ and $u_t = 0.55$:

$$r_l = a_{tot} \cdot u_l = 18.2;\ r_t = -\ln u_t \cdot 1/a_{tot} = 0.0092 \text{ [a.u]}.$$

As $10 < r_l \leq 20$, the second reaction is selected, i.e. the next event is one of the mRNA molecules getting degraded; so we update the state and time of the system accordingly

$$t = 0.5 + 0.0092 = 0.5092 \text{ [a.u.]};\ \mathbf{x}(0.5092)^T = [4, 20].$$

Repeating the above steps until we reach T_{final} or a certain condition is met (e.g. some species completely disappear) provides a single observation of the system –

and through repeated paths the whole distribution as it changes over time can be accurately approximated.

It is clear from the transformation $r_t = -\ln u_t \cdot 1/a_{tot}$ that high molecular abundances or highly probable reactions lead to a small time-step, such as the 0.0092 above. For example with the above rate values and abundances, on average 3800 more steps have to be taken to reach a T_{final} of 60. If the rate parameters are doubled ($[20, 4, 10, 2]$), even though the general behaviour of the system is unchanged, the reactions happen faster and hence an average of 7600 steps are needed to simulate the system until 60. This means a population of few hundred trajectories might take *several minutes* to simulate using a naive implementation. Naturally, this figure can be significantly lowered by informed choice of implementation and simulation details and several solutions are available that execute thousands of steps in less than a second.

SDE

In the next step we obtain the SDE form of the same system and obtain its numerical simulation using the Euler-Maruyama method,

$$dx_1 = (k_1 - k_2 x_1) \cdot dt + \sqrt{k_1 dt} \cdot w_1 - \sqrt{k_2 x_2 dt} \cdot w_2$$
$$dx_2 = (k_3 x_1 - k_4 x_2) \cdot dt + \sqrt{k_3 x_1 dt} \cdot w_3 - \sqrt{k_4 x_2 dt} \cdot w_4,$$

where the $w_i, i = 1 \ldots 4$, are four samples drawn from a *standard normal distribution.* Let us compute the increase of variables if the system is at $\mathbf{x}(0.5)^T = [5, 20]$, with previously mentioned rate values $\mathbf{k} = [10, 2, 5, 1]$. A choice of time-step $dt = 0.02$ and normally distributed samples $w_{1,2,3,4} = [0.72, -1.21, 1.63, 0.49]$ lead to,

$$dx_1 = (10 - 2 \cdot 5) \cdot 0.02 + \sqrt{0.2} \cdot 0.72 + \sqrt{0.2} \cdot 1.21 = 0.8631$$
$$dx_2 = 5 \cdot 0.02 + 1.1526 - 0.31 = 0.9426,$$

and therefore after one time-step the system will be in the state $[5.8631, 20.9426]$.

Figure 2.3(a) shows an example trajectory for each method generated until $2s$, with the above parameter values and starting copy numbers. Note how the SDE takes constant time-steps, while the SSA trajectory has both more dense or more sparse intervals. Histograms showing expression in a large population of cells (Fig. 2.3(b)) demonstrate that on this level the two algorithms result in distributions very similar in shape and summary statistics but statistically different ($p < 0.05$).

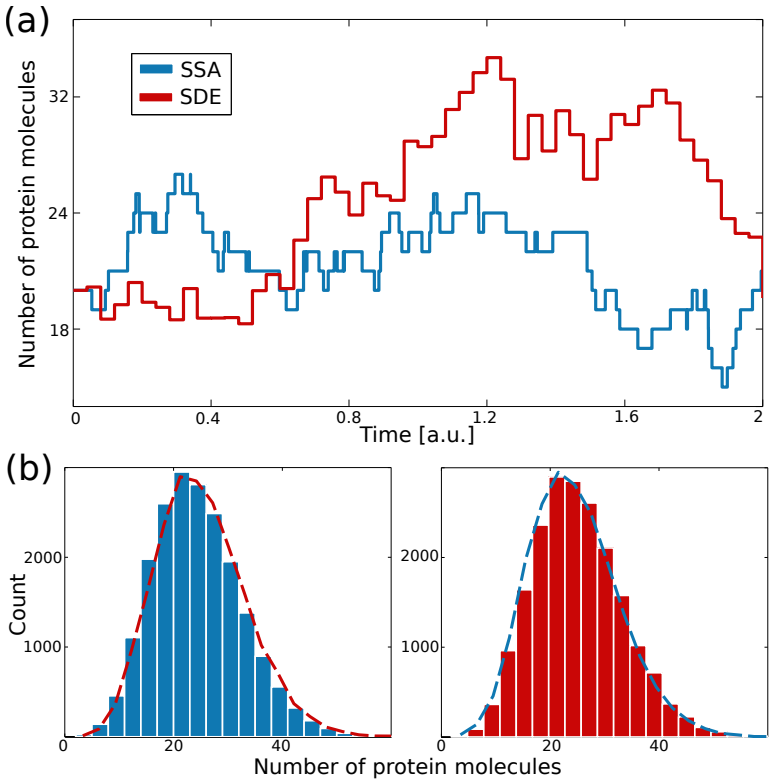


Figure 2.3: Comparison of Gillespie's Stochastic Simulation Algorithm and Stochastic Differential Equation-based simulations. (a) A single trajectory of protein copy number in the stochastic gene expression system, generated by the SSA (blue line) or Euler-Maruyama integration of the SDE form with $\Delta t = 0.04$ (red line). Initial values are $x_1 = 5$, $x_2 = 20$, parameters are $\mathbf{k} = [10, 2, 5, 1]$. **(b)** Distribution of protein abundance at time 10 in a population of 20,000 cells, simulated using SSA (blue) or SDE (red). The distribution obtained using the opposing method is shown with dashes for correspondence.

2.3.3 Linear Noise Approximation

The Linear Noise Approximation (LNA) [30], similarly to SDEs, provides a way to follow a stochastic system through equations describing two parts of the system's state: a deterministic drift term described by ODEs and a noise term captured with SDEs,

$$\mathbf{x}(t) = \boldsymbol{\phi}(t) + \boldsymbol{\xi}(t). \tag{2.9}$$

Given a stochastic system defined by a stoichiometry matrix (S) and a collection of reaction propensities ($\mathbf{a} = [a_1,\ a_2, \ldots, a_r]^T$), the corresponding equations of time-evolution are:

$$\frac{d\boldsymbol{\phi}(t)}{dt} = S \cdot \mathbf{a}(\boldsymbol{\phi}(t)) \tag{2.10a}$$

$$\frac{d\boldsymbol{\xi}(t)}{dt} = S \cdot D(\boldsymbol{\phi}(t))\boldsymbol{\xi}(t) + \sum_{j=1}^{r} S_{:,j} W_j(a_j(\boldsymbol{\phi}(t))), \tag{2.10b}$$

where $W_j(d)$ is a Wiener process and

$$\{D(\boldsymbol{\phi})\}_{ik} = \frac{\partial a_j(\boldsymbol{\phi})}{\partial \phi_k}$$

hence $D = \frac{\partial \mathbf{a}}{\partial \boldsymbol{\phi}}$ is the *Jacobian* of the system. Therefore, the variables $\boldsymbol{\phi}(t)$ and $\boldsymbol{\xi}(t)$ can be formally derived for any t as

$$\boldsymbol{\phi}(t) = \boldsymbol{\phi}(0) + \int_0^t S \cdot \mathbf{a}(\boldsymbol{\phi}(s))ds, \tag{2.11a}$$

$$\boldsymbol{\xi}(t) = \boldsymbol{\xi}(0) + \int_0^t SD(\boldsymbol{\phi}(s))\boldsymbol{\xi}(s)ds + \sum_{j=1}^{r} S_{:,j} W_j \left(\int_0^t \sqrt{a_j(\boldsymbol{\phi}(s))ds} \right), \tag{2.11b}$$

and one can use numerical methods to simulate a trajectory of the variables.

However, in many applications we are interested in population-level behaviour and therefore we use Eq. (2.10) to obtain equations that describe how the mean and variance of a large population of single $\mathbf{x}(t)$ instances changes. The decomposition of the state variable makes it straightforward to follow the change in the mean of the system, as it is already given by Eq. (2.10a), since the deterministic term

corresponds to the mean of the stochastic variables. The evolution of the variance – covariance matrix in multidimensional cases – derived from the formal definition of $\Sigma(t)$ and Eq. (2.10b), is calculated as

$$\frac{d\Sigma}{dt} = SD\Sigma + \Sigma(SD)^T + S \cdot \mathrm{diag}\{\mathbf{a}(\boldsymbol{\phi})\} \cdot S^T.$$

Therefore, by combining the equations for the time-dependent change of $\boldsymbol{\phi}(t)$ and $\Sigma(t)$, we can derive an ODE system

$$\frac{d\boldsymbol{\phi}}{dt} = S \cdot \mathbf{a}(\boldsymbol{\phi}) \tag{2.12a}$$

$$\frac{d\Sigma}{dt} = SD\Sigma + \Sigma(SD)^T + S \cdot \mathrm{diag}\{\mathbf{a}(\boldsymbol{\phi})\} \cdot S^T. \tag{2.12b}$$

This, once solved using standard numerical techniques, provides the time-course of the mean behaviour and variance as one would observe in a large population of single realisation. The major advantage of Eq. (2.12) over other stochastic depictions is that it does not include any probabilistic terms anymore, therefore its simulation and analysis can be carried out through already established *deterministic* techniques. In the next chapters we make excessive use of the above logic to obtain a deterministic representation that still captures the variability of the system, and use it as a starting point for further investigations.

2.3.4 Moment Expansion Approximation

The moment expansion approximation (MEA) technique [36] also tackles the complex problem of following the evolution of the whole system through the moments of the probability distribution. Similarly to the LNA, solution of the CME is translated into solving a set of ODEs tracking the evolution of means, variances, skewness, up to any required moment. Its main advantage over the LNA is that it can provide reliable estimate for system with few dozens of molecules, for which the former method breaks down.

Let us consider the generic description of a stochastic system in terms of stoichiometry matrix, S, and reaction propensities, a_i. The moments of the distribution, $P(\mathbf{x})$, are computed based on the time dependent moment generating function derived from the CME

$$\frac{dm}{dt} = \sum_l \left[(e^{\boldsymbol{\theta}\mathbf{s}_l} - 1) \sum_{\mathbf{x}} e^{\boldsymbol{\theta}\mathbf{x}} P(\mathbf{x}) a_l(\mathbf{x}) \right]. \tag{2.13}$$

The $\mathbf{n}$th raw moment of the system, $\mathbb{E}(\mathbf{x}^{\mathbf{n}}) = \mathbb{E}(x_1^{n_1} \dots x_N^{n_N})$, can be found as the $\mathbf{n}$th order derivative of Eq. (2.13) with respect to the auxiliary variable $\boldsymbol{\theta}$. The expressions obtained this way are evaluated by expanding the $a_l(\mathbf{x})$ in a Taylor series about the mean (μ). For example the equations for the means (i.e. an array containing the mean of each species) are given as,

$$\frac{d\mu}{dt} = S \cdot \left[\sum_l \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} \frac{1}{\mathbf{n}!} \frac{\partial^{\mathbf{n}} a_l(\mathbf{x})}{\partial \mathbf{x}^{\mathbf{n}}} \bigg|_{\mathbf{x}=\mu} \mathbb{E}(\hat{\mathbf{x}}^{\mathbf{n}}) \right],$$

where

$$\begin{aligned} \partial^{\mathbf{n}} &= \partial^{n_1+\dots+n_N} \\ \partial \mathbf{x}^{\mathbf{n}} &= \partial x_1^{n_1} \dots \partial x_N^{n_N} \\ \mathbf{n}! &= n_1! \dots n_N!. \end{aligned}$$

This procedure gives rise to an ODE system describing the evolution of raw moments and containing terms of central moments, $\mathbb{E}(\hat{\mathbf{x}}^{\mathbf{n}}) = \mathbb{E}((x_1 - \mu_1)^{n_1} \dots (x_N - \mu_N)^{n_N})$.

The relationship between raw and central moments is formulated as

$$\mathbb{E}(\hat{\mathbf{x}}^{\mathbf{n}}) = \sum_{k_1=0}^{n_1} \cdots \sum_{k_N=0}^{n_N} \binom{\mathbf{n}}{\mathbf{k}} (-1)^{(\mathbf{n}-\mathbf{k})} \mu^{(\mathbf{n}-\mathbf{k})} \mathbb{E}(\mathbf{x}^{\mathbf{k}}), \tag{2.14}$$

where

$$\begin{aligned} (-1)^{(\mathbf{n}-\mathbf{k})} &= (-1)^{(n_1-k_1)} \dots (-1)^{(n_N-k_N)}, \\ \binom{\mathbf{n}}{\mathbf{k}} &= \binom{n_1}{k_1} \dots \binom{n_N}{k_N}. \end{aligned}$$

We obtain a system of ODEs for the time evolution of central moments by applying Eq. (2.14) to substitute raw moments with central moments, except for the mean concentrations, μ_i.

In systems involving nonlinear connections, each central moment depends on a higher order moment, therefore the set of differential equations generated by moment expansion is in principle infinite and unsolvable. In order to close the equation set we can apply moment closure techniques to derive expressions for higher order moments, based on properties of an assumed underlying distribution. The choice of moment closure can have a substantial effect on the accuracy of the approximation. In the next chapter we will summarise existing solutions for closure and present new formulae that can be applied for a wide range of biological problems.

Example – Linear gene expression

Once again, we take the linear gene expression model of two variables, mRNA and protein abundance (m and p), characterised by four reaction propensities, and a 2×4 stoichiometry matrix (see Eq. (2.5)). Our aim is to obtain deterministic expressions for the evolution of the mean vector and covariance matrix of the system, as these are observables usually obtained in population-level biological experiments. This can be achieved through both the LNA and MEA techniques, yielding the same results, so we now conduct the computation using the LNA. The matrices needed for the calculation are

$$\mathbf{a} = \begin{bmatrix} k_1 \\ k_2 m \\ k_3 m \\ k_4 p \end{bmatrix} \; ; \quad D = \frac{\partial \mathbf{a}}{\partial \mathbf{x}} = \begin{bmatrix} 0 & 0 \\ k_2 & 0 \\ k_3 & 0 \\ 0 & k_4 \end{bmatrix} \; ; \quad \mathrm{diag}\{\mathbf{a}\} = \begin{bmatrix} k_1 & 0 & 0 & 0 \\ 0 & k_2 m & 0 & 0 \\ 0 & 0 & k_3 m & 0 \\ 0 & 0 & 0 & k_4 p \end{bmatrix},$$

and henceforth the equations for the vector of means (here denoted with $\mathbf{x}$) and variances (Σ) become

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= \begin{bmatrix} k_1 - k_2 m \\ k_3 m - k_4 p \end{bmatrix} \\ \frac{d\Sigma}{dt} &= \frac{d}{dt}\begin{bmatrix} \sigma_m & \sigma_{mp} \\ \sigma_{mp} & \sigma_p \end{bmatrix} = \begin{bmatrix} -2k_2\sigma_m & k_3\sigma_m - (k_2+k_4)\sigma_{mp} \\ k_3\sigma_m - (k_2+k_4)\sigma_{mp} & 2k_3\sigma_{mp} - 2k_4\sigma_p \end{bmatrix} + \begin{bmatrix} k_1 + k_2 m & 0 \\ 0 & k_3 m + k_4 p \end{bmatrix}. \end{aligned}$$

These equations are equivalent to those derived through taking the expectation of the CME form of the gene expression system [41].

Example – Dimerisation

Next we consider a nonlinear case, the dimerisation system, in which equations for summary statistics of the system will differ depending on the method used. Taking the stoichiometry matrix and reaction propensities listed in Eq. (2.6), we first apply the LNA to gain an ODE system for the means and variances of the

species:

$$\begin{aligned}
\dot{m} &= 2k_1 m(1-m) + 2k_2 d \\
\dot{d} &= k_1 m(m-1) - k_2 d \\
\dot{\sigma}_m &= 4k_1 m(-2\sigma_m + m - 1) + 4k_1\sigma_m + 4k_2(d + \sigma_{md}) \\
\dot{\sigma}_{md} &= 2k_1 m(\sigma_m - 2\sigma_{md} - m + 1) + k_2(2\sigma_d - \sigma_{md} - 2d) + k_1(2\sigma_{md} - \sigma_m) \\
\dot{\sigma}_d &= k_1 m(4\sigma_{md} + m - 1) + k_2(d - \sigma_d) - 2k_1\sigma_{md},
\end{aligned}$$

where σ_{md} stands for the covariance of monomer and dimer counts. The equations for the mean copy numbers (m and d) only contain terms of the mean values, and in fact are the same equations one would acquire in a deterministic description of the dimerisation system.

In the next step, we use the MEA to obtain the time evolution of the same quantities. As the MEA can provide an arbitrary order of moments, we could obtain equations for not only the means and (co)variances, but further statistics, such as skewness and kurtosis. However, to keep the comparison clear, let us consider a second-order expansion leading to the ODE system

$$\begin{aligned}
\dot{m} &= 2k_1 m(1-m) + 2k_2 d - 2k_1\sigma_m \\
\dot{d} &= k_1 m(m-1) - k_2 d - k_1\sigma_m \\
\dot{\sigma}_m &= 4k_1 m(-2\sigma_m + m - 1) + 4k_1\sigma_m + 4k_2(d + \sigma_{md}) + 4k_1\sigma_m \\
\dot{\sigma}_{md} &= 2k_1 m(\sigma_m - 2\sigma_{md} - m + 1) + k_2(2\sigma_d - \sigma_{md} - 2d) + k_1(2\sigma_{md} - \sigma_m) - 2k_1\sigma_m \\
\dot{\sigma}_d &= k_1 m(4\sigma_{md} + m - 1) + k_2(d - 2\sigma_d) - 2k_1\sigma_{md} + k_1\sigma_m.
\end{aligned}$$

The above equations are arranged to emphasise the similarities and differences with the LNA equivalent. Most terms are identical, but the equations for the means depend on σ_m, due to the nonlinearity of the first reaction. Consequently additional terms appear in the second-order moment equations as well. In fact, the (co)variances have a dependency on third order moments, that we now eliminated, by assuming that all those terms equate to zero.

Next, let us consider a concrete example when $\mathbf{k} = [1.66 \cdot 10^{-3}, 0.2]$ and the current values of the variables are $\mathbf{x}^T = [m, d, \sigma_m, \sigma_{md}, \sigma_d] = [291, 10, 40, -3, 3]$.

Once again, the time-units used in the system are arbitrary, and the units of k_1 and k_2 are 1/molecules×a.u. and 1/a.u., respectively. To quantify the difference between the two equation sets, we compute the change for all species, and calculate the difference as the percentage of change derived through LNA

$$\dot{\mathbf{x}}_{LNA} = \begin{bmatrix} -276.1748 \\ 138.0874 \\ 411.636 \\ -238.0096 \\ 135.1006 \end{bmatrix}, \quad \dot{\mathbf{x}}_{MEA} = \begin{bmatrix} -276.3076 \\ 138.021 \\ 411.902 \\ -238.1424 \\ 135.1671 \end{bmatrix}, \quad \frac{|\dot{\mathbf{x}}_{LNA} - \dot{\mathbf{x}}_{MEA}|}{\dot{\mathbf{x}}_{LNA}} \cdot 100\% = \begin{bmatrix} 0.0481\% \\ 0.0481\% \\ 0.0645\% \\ 0.0558\% \\ 0.0491\% \end{bmatrix}.$$

During numerical solution of the ODE systems these numbers become even smaller and the difference in the obtained trajectories is negligible (see Figure 2.4). The same holds for many biochemical models and parameter sets. However, as the equations are more closely coupled in the MEA formalism (due to the mean evolution's dependence on variance terms), it is possible that the additional terms have an amplified effect over the time evolution. In cases when several reactions involving numerous species are nonlinear, the difference is even more profound. The next chapter focuses on improving the MEA technique for such examples for which the LNA has been shown to be inadequate [36].

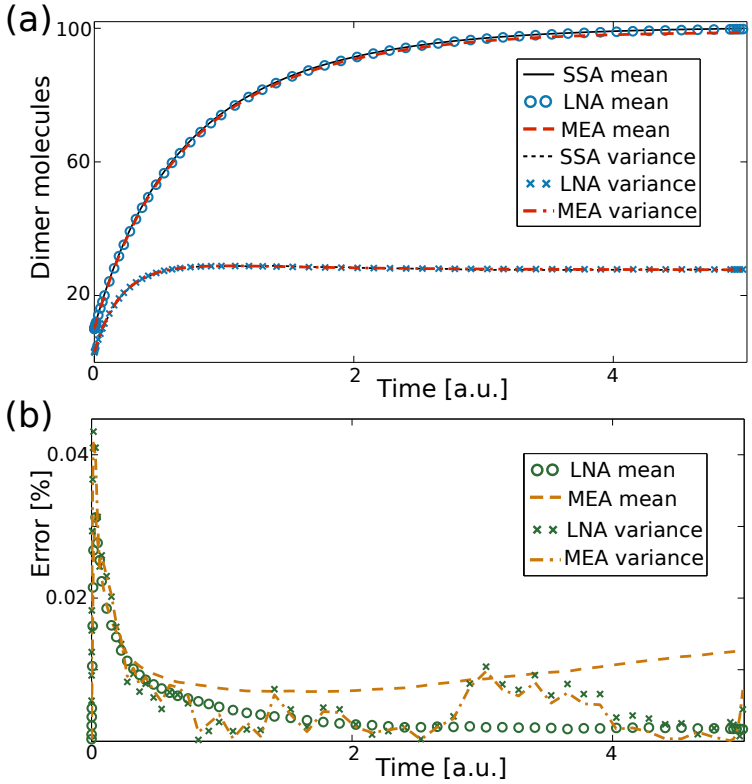


Figure 2.4: Comparison of Linear Noise Approximation and Moment Expansion Approximation on the dimerisation example. **(a)** Trajectory of mean number and variance of dimer molecules computed by the LNA and the MEA, shown together with the summary statistics obtained from 1,000,000 SSA simulations. LNA and MEA results are in blue or red, as indicated in legend. **(b)** Error between the approximated and SSA-derived trajectories, expressed as the % of the SSA-derived trajectory. LNA and MEA results are in green or orange, as indicated. Initial values for the simulations are $\mathbf{x} = [291, 10, 40, -3, 3]$ (used as the mean and covariance matrix parameters of a multivariate normal distribution for SSA simulations), parameters are $\mathbf{k} = [1.66 \cdot 10^{-3}, 0.2]$.

Three

Multivariate moment closure techniques

In the previous section we discussed how the simulation of stochastic systems can be carried out via numerous routes depending on the nature of the application. In this chapter we focus on moment-based approximations and present an extension to these methods.

3.1 Introduction

In the case when stochastic modelling aims to recreate population-level experimental results obtained from thousands of cells, approximation techniques are favoured, as exact simulations often become unfeasible in this volume. Approximations can lighten the computational burden by focusing on the moments of the distribution of cells to track relevant statistics, primarily the average behaviour. Moment expansion methods [36] offer an efficient solution that is applicable even in cases when the molecules investigated are only present in a handful of copies inside the cell. However, in systems where some species are connected through a nonlinear relation (i.e. there is at least one bimolecular reaction), we face a further complication. The evolution of each moment generated by the expansion method depends

on some higher-order moments, resulting in an infinite set of equations that needs to be truncated.

Moment closure approximation techniques generally combine an expansion method with a moment closure *ansatz* to obtain a closed set of differential equations solvable either analytically (this is only possible very rarely) or numerically. The closure of the equations is usually based on the assumption that the variables approximately follow a known distribution for which we already know some relation(s) between the moments, and hence can derive an expression of the higher order moments in terms of lower ones. Goodman [42] and Whittle [43] set higher order terms to zero to close second order equations, effectively applying normal closure at the order of three. Gillespie and colleagues [44, 45] proposed a method to obtain moment equations up to any order, listed expressions as the basis of univariate normal, lognormal and Poisson closure, and demonstrated their algorithm by closing equations of second moments assuming a normal distribution. Krishnarajah et al. [46] directly addressed the problem that arises in the case of highly skewed population distributions when most closure schemes fail as the normal assumption does not hold. They developed second- and third-order beta-binomial, lognormal and mixture closure techniques for two stochastic epidemic models following a single variable. Later they also described multivariate methods for an extended model [47].

In parallel, several approaches have been developed which do not proceed from a particular distribution. Keeling [48] derived moment closure in a one-dimensional logistic system for third order moments by formulating the system in terms of multiplicative instead of additive moments. Singh and Hespanha [49, 50] derived a systematic construction of moment closure functions based on derivative matching, first in one and then in arbitrary dimensions. In both cases, the derived moment closure functions are consistent with the species being (jointly) lognormally distributed. Hausken and Moxnes [51] set higher mixed moments to zero and assumed a Dirac distribution, while Smadbeck and Kaznessis [52] applied an information theoretic

point of view to close the equations based on the assumption that a final set of moments holds all information about the system.

Although it has been shown that expansions up to three moments tend to provide a more precise approximation than only second order ones [53], general closure formulae for closing third or higher order equation systems have not been widely used. Furthermore, many of the existing algorithms have been adapted manually to specific systems and appear not to be concerned with (or capable to provide) general applicability without manual intervention. Crucially, there are no automatic approaches capable of generating and closing moment equations of arbitrary order that could be applied to different systems without substantial modifications to the method itself.

In this work we introduce a new multivariate moment closure approximation technique that can be applied to any stochastic kinetic system with no constraints on the dimension, order or (lack of) linearity of the system and its interactions. Our approximation is based on the MEA method presented in the previous chapter and we extend this method by developing closure schemes following either a multivariate normal, lognormal or a gamma distribution ansatz. We use properties of these three distributions to derive moment closure formulae to express arbitrary order higher moments in terms of lower order ones, replacing truncations used in the original expansion. Our algorithm takes into account the different levels of the potentially intricate interplay between species captured by higher order mixed moments. Since the definition of a multivariate gamma distribution is ambiguous and the existing formulations are not suitable to model arbitrary systems, we also describe a new multivariate gamma distribution generalising existing definitions.

The derived formulae are incorporated into the MEA framework to obtain improved and less costly approximations of nonlinear dynamics. We demonstrate the power of our method on two small oscillatory reaction systems, the p53-Mdm2 negative feedback loop in tumour suppression and the Hes1 system in embryogenesis;

and on a complex two-compartment model of ERK1/2 phosphorylation dynamics. We also present a publicly available Python package implementing the combined framework, called MEANS. The package is completely automated and developed in a user-friendly manner, to make it suitable for the analysis of any stochastic model without thorough understanding of the algorithm. Therefore it can serve as a starting point for parameter inference [54, 55] and model selection [56, 57], distribution reconstruction [58], sensitivity analysis [59], experimental design [60, 61], and design of dynamical systems with desired qualitative [62] and quantitative [63] behaviour.

3.2 Moment closure formulae

Most well-known distributions are completely described by a small set of parameters, meaning that given the first few moments, these parameters can be calculated and used to determine the whole probability distribution as well as its higher order moments. We use this property to substitute expressions of lower order moments in place of $(m+i)$th (where $i \geq 1$) moments appearing in the set of ODEs obtained from a moment-based approximation method. Our final aim is to obtain a closed set of equations

$$\frac{d}{dt}\mathbb{E}(\hat{\mathbf{X}}^{\mathbf{n}}) = f(\boldsymbol{\mu}, \ldots, \mathbb{E}(\hat{\mathbf{X}}^{\mathbf{m}}), mc^k(\mathbf{X})), \tag{3.1}$$

where $\boldsymbol{\mu}$ represents the mean vector of the system and $mc^k(\mathbf{X})$ denotes the moment closure expression based on up to kth order central moments of $\mathbf{X}$. In the followings we derive the expressions for $mc^k(\mathbf{X})$ based on three distributions.

We investigate normal, lognormal and gamma distributions. The first of these is, of course, the most popular distribution due to its prominent role in the central limit theorem. The other two distributions are considered for their asymmetry and restriction to the positive real numbers, two important features of systems describing molecular reactions. Normal and lognormal distributions have been used successfully to model distributions in biochemistry [44,45,49,51]. On the other hand, single cell protein expression levels follow a gamma distribution [64,65] if considered at the population level. Our closure approaches involve *two-parameter distributions*, i.e. the first two moments provide a sufficient description of the distribution. As biological systems usually contain several species which have strong influence over each other, we use multivariate definitions of the distributions, in which the different interaction levels between variables are captured in the different mixed order moments. In this case the two parameters become a vector and a matrix, typically in the form of, respectively, a mean vector and covariance matrix.

For each case, we start the derivation by providing a definition for the corresponding multivariate distribution and its characteristic parameters. Then we derive formulae for all (both non-mixed and mixed) moments of an arbitrary order m (necessarily, $m \geq 3$) in terms of the aforementioned parameters. Expressions of raw moments are translated into central moment terms using methods developed in the MEA algorithm, whereas for the normal distribution we can obtain the central moments directly. Finally, we define a way to calculate the parameters of the distributions from means, variances and covariances of the species, considered known, as their time evolution has already been described in the moment expansion step.

The formulae are compared to the truncation method applied in [36] that sets all higher moments to zero – hence we refer to this method as *zero closure*. As a comparison to the multivariate techniques we also generate univariate closures where only expressions for the marginals are generated, by setting all mixed moments to zero.

3.2.1 Multivariate normal closure

A random vector is said to be N-variate normally distributed if every linear combination of its components has a univariate normal distribution. It is described by the following density function,

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right), \tag{3.2}$$

which reduces to the pdf of the univariate normal distribution, if $\Sigma \in \mathbb{R}^{1\times 1}$.

All higher order central moments of a normal distribution can be calculated using Isserlis's theorem [66],

$$\mathbb{E}(\hat{X}_1^{m_1} \dots \hat{X}_N^{m_N}) = 0, \text{ if } \sum_{i=1}^{N} m_i = \text{odd}, \tag{3.3a}$$

$$\mathbb{E}(\hat{X}_1^{m_1} \dots \hat{X}_N^{m_N}) = \sum_{\mathbb{I}} (\Sigma_{ij}\Sigma_{lk} \dots \Sigma_{az}), \tag{3.3b}$$

where $\hat{X}_i$ denotes the central moments (i.e. $\hat{X}_i = X_i - \mu_i$) and the sum is over all possible partitionings of the set $\mathbb{I} = \{\underbrace{1, 1, \dots 1}_{m_1}, \dots, \underbrace{N, N, \dots N}_{m_N}\}$ into unordered pairs.

In this case, the necessary parameters $\boldsymbol{\mu}$ and Σ are simply described as,

$$\begin{aligned} \mu_i &= \mathbb{E}(X_i), \\ \Sigma_{ii} &= \mathrm{Var}(X_i), \\ \Sigma_{ij} &= \mathrm{Cov}(X_i, X_j), \end{aligned} \tag{3.4}$$

since by definition they correspond to central moments of the species in the system.

3.2.2 Multivariate lognormal closure

For the univariate case, X has a lognormal distribution if its logarithm is normally distributed, thus an N-variate lognormally distributed random vector ($\mathbf{X}$) is characterised by its normal counterpart ($\mathbf{Y}$),

$$\mathbf{X} = \exp(\mathbf{Y}), \tag{3.5}$$

and $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

All higher order (raw) moments can be simply expressed in terms of $\boldsymbol{\mu}$ and Σ using the formula [67]

$$\mathbb{E}(\mathbf{X}^{\mathbf{m}}) = \exp(\mathbf{m}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{m}^T\Sigma\mathbf{m}), \tag{3.6}$$

where $\mathbf{m} = [m_1, m_2, \dots, m_N]^T$ and $m_i \in \mathbb{N}$.

Given our observation of up to second order moments of the components, we can derive expressions in terms of summary statistics of $\mathbf{X}$,

$$\begin{aligned} \Sigma_{ii} &= \ln\left(1 + \frac{\mathrm{Var}(X_i)}{\mathbb{E}(X_i)^2}\right), \\ \mu_i &= \ln(\mathbb{E}(X_i)) - \frac{1}{2}\Sigma_{ii}, \\ \Sigma_{ij} &= \ln\left(\frac{\mathrm{Cov}(X_i, X_j)}{\exp(\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj}))} + 1\right), \end{aligned} \tag{3.7}$$

thus obtaining the parameters $\boldsymbol{\mu}$ and Σ of $\mathbf{Y}$.

3.2.3 Multivariate gamma distribution

The univariate gamma probability density distribution can be parametrised by

$$f_X(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1}\exp(-x/\beta), \tag{3.8}$$

where Γ is the gamma function, β is a scale parameter and α is a shape parameter. Any (raw) moment of a variable $X \sim \text{Gamma}(\alpha,\beta)$ is given by a recurrence relation

$$\mathbb{E}(X^m) = (\alpha)_m \beta^m, \tag{3.9}$$

where $(a)_m = (a+m-1)!/(a-1)! = a(a+1)\dots(a+m-1)$.

There are several ways to define a multivariate gamma (MVG) distribution fulfilling the requirement that individual marginals follow a univariate gamma distribution. Since the sum of two gamma variables with the same scale parameter will again be gamma distributed, it is straightforward to obtain a MVG distribution from linear combinations of independent gamma variables. A thus retrieved multivariate distribution will have gamma distributed marginals and shared components will have a direct influence over the correlation of each dimensional component. Two examples for a MVG distribution can be found in [68] and [69]. However, there are some possible scenarios that cannot be captured well by either of these definitions. For example the three-dimensional case, when two variables (X_1, X_2) are independent while another pair of them (X_2, X_3) is highly correlated. Therefore, we require a new definition that allows for any combination of (positive) variance and covariance values between the coordinates of the N-dimensional vector. In the following paragraph we define a new MVG distribution in terms of linear combinations of gamma random variates that satisfies this condition.

Let Y_{kl} $(k = 1\dots N,\ l = k\dots N)$ be independent gamma variables with shape and scale parameters α_{kl}, β_{kl}. For convenience we also introduce extra – but not

any more independent – gamma variates, Y_{rq} $(r = 2 \dots N,\ q = 1 \dots (r-1))$, such that they equal Y_{kl} if $k = q,\ l = r$. We use these variables in the following to ease the indexing of Y_{kl} – note that in all general formulae Y_{12} and Y_{21} refer to the same variable. Then we define $\mathbf{X} = [X_1, X_2 \dots X_N]$ as,

$$X_i := \sum_{k=1}^{N} \frac{\beta_{ii}}{\beta_{ik}} Y_{ik}, \tag{3.10}$$

which has an N-dimensional multivariate gamma distribution. All of the coordinates are gamma variates, namely $X_i \sim \text{Gamma}(\alpha_i, \beta_i)$, where

$$\begin{aligned} \alpha_i &= \sum_{k=1}^{N} \alpha_{ik}, \\ \beta_i &= \beta_{ii}. \end{aligned}$$

Furthermore, variables X_i and X_j have exactly one component $(Y_{ij} \equiv Y_{ji})$ in common to account for their correlation, whose parameters are chosen independently of any other components contributing to X_i and X_j. Figure 3.1 shows two two-dimensional examples of the new MVG. As seen in the figure, both highly correlated and independent variables can give rise to the same marginals, however, the multivariate density clearly distinguishes between the two cases.

3.2.4 Multivariate gamma closure

Similarly to closure formulae based on the other distributions, our aim is to obtain a closed formulation of mth order moments in terms of the parameter values defining a multivariate gamma distribution. As the individual coordinates, $X_1, \dots, X_N$ follow univariate gamma distributions, all non-mixed higher order moments of the multivariate distribution can be calculated as in Eq. (3.9). Generally,

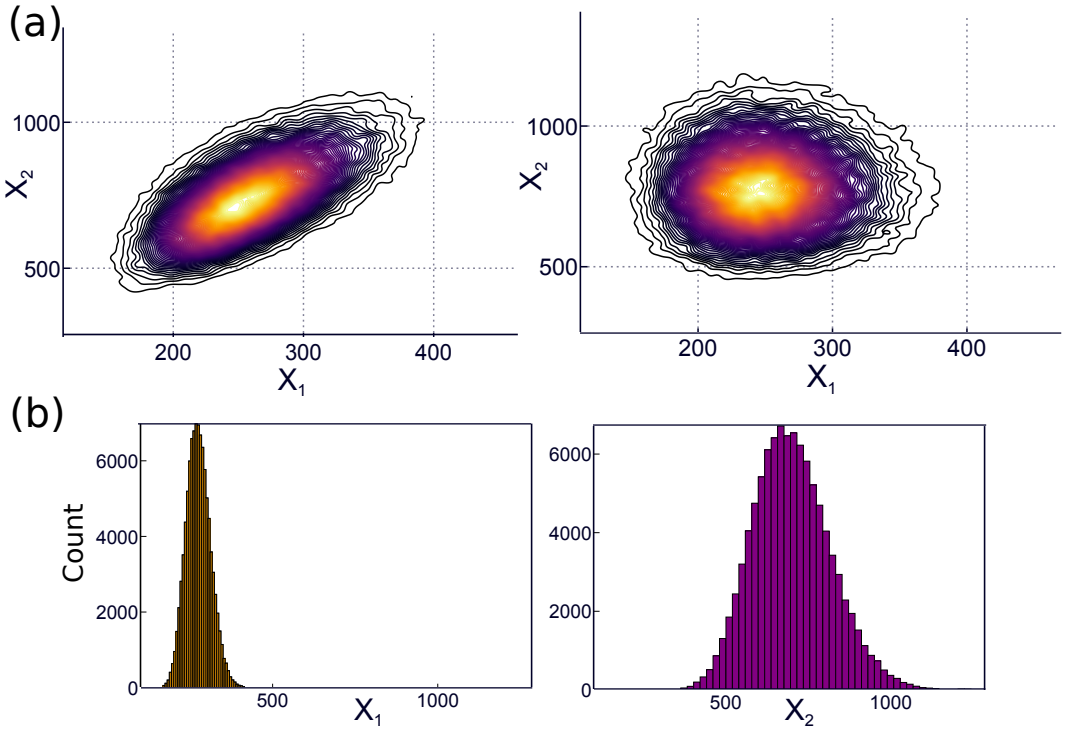


Figure 3.1: Two bivariate examples of the new multivariate gamma distribution. Parameters of the two distributions are: (left) $\alpha_{11} = 25,\ \alpha_{12} = 30,\ \alpha_{22} = 5$; and (right) $\alpha_{11} = 55,\ \alpha_{12} = 0,\ \alpha_{22} = 35$. Scales are $\beta_{11} = 5,\ \beta_{12} = 2,\ \beta_{22} = 20$ in both cases. **(a)** Contour plots presenting the two-dimensional density estimate obtained from 10^5 samples of the two distributions. **(b)** One-dimensional marginal distributions, identical for both bivariate distributions presented above.

the mth order moment of X_i can be described by the formula

$$\begin{aligned} \mathbb{E}(X_i^m) &= \mathbb{E}\left(\left(\sum_{k=1}^{N} \frac{\beta_{ii}}{\beta_{ik}} Y_{ik}\right)^m\right) \\ &= \mathbb{E}\left(\sum_{\mathbf{k}} \mathbb{C}(m,\mathbf{k}) \prod_{r=1}^{N} \left(\frac{\beta_{ii}}{\beta_{ir}} Y_{ir}\right)^{k_r}\right) \\ &= \beta_{ii}^m \left(\sum_{\mathbf{k}} \mathbb{C}(m,\mathbf{k}) \prod_{r} (\alpha_{ir})_{k_r}\right), \end{aligned} \tag{3.11}$$

where the sum is taken over all combinations of nonnegative integers k_r such that they sum up to m; $\mathbb{C}(m, \mathbf{k})$ is the multinomial coefficient,

$$\mathbb{C}(m, \mathbf{k}) = \binom{m}{k_1, \ldots, k_N} = \frac{m!}{k_1! \ldots k_N!}; \tag{3.12}$$

and we made use of the independence of Y_{kl} variables. Based on this formalisation, we derive the closed expression for any mixed moment by adapting the vectorised notation of exponents $\mathbf{m}$ used in Eq. (3.6) and obtaining

$$\mathbb{E}(\mathbf{X}^{\mathbf{m}}) = \boldsymbol{\beta}^{\mathbf{m}} \left(\prod_i \sum_{\mathbf{k}} \mathbb{C}(m_i, \mathbf{k}) (\boldsymbol{\alpha}_i)_{\mathbf{k}} \right), \tag{3.13}$$

where

$$\begin{aligned} \boldsymbol{\beta}^{\mathbf{m}} &= \beta_{11}^{m_1} \ldots \beta_{NN}^{m_N}, \\ (\boldsymbol{\alpha}_i)_{\mathbf{k}} &= (\alpha_{i1})_{k_1} \ldots (\alpha_{iN})_{k_N}, \end{aligned}$$

and the combined product and sum correspond to taking the product of the expanded $X_1^{m_1}, \ldots, X_N^{m_N}$ terms, with $(\alpha_i)_{\mathbf{k}}$ calculated after the $\alpha_{kl} = \alpha_{lk}$ symmetry is taken into account.

To give an example, the first mixed raw moment – and based on it, the covariance – of two variables will be,

$$\begin{aligned} \mathbb{E}(X_i X_j) &= \beta_{ii}\beta_{jj} \left(\sum_{(k,l) \neq (i,j)} \alpha_{ik}\alpha_{jl} + (\alpha_{ij})_2 \right), \\ \mathrm{Cov}(X_i, X_j) &= \mathbb{E}(X_i X_j) - \beta_{ii}\beta_{jj} \left(\sum_{k,l} \alpha_{ik}\alpha_{jl} \right) \\ &= \alpha_{ij}\beta_{ii}\beta_{jj}. \end{aligned} \tag{3.14}$$

One can see that, to calculate the final formulae, we need all α_{kl} values, while only the 'diagonal' β_{kk} elements of the scale parameters are used. To emphasise the similarity with the previous closure techniques, we summarise the descriptive parameters in a vector of scale parameters, $\boldsymbol{\beta}$, and a matrix of shape parameters,

$\mathbf{A}:\ \mathbf{A}_{kl} = \alpha_{kl}$. By reformulating Eqs. (3.9) and (3.14),

$$\begin{aligned} \beta_i &= \mathrm{Var}(X_i)/\mathbb{E}(X_i), \\ \mathbf{A}_{ij} &= \mathrm{Cov}(X_i, X_j)/(\beta_i\beta_j), \\ \mathbf{A}_{ii} &= \mathbb{E}(X_i)/\beta_i - \sum_{k\neq i} \alpha_{ik}, \end{aligned} \tag{3.15}$$

we can obtain all desired parameters in terms of maximum second order moments.

Examples

Isserlis's theorem

To illustrate how Isserlis's theorem provides a way to express any even ordered moment of a normally distributed N-dimensional variable, consider the fourth order central moment $\mathbb{E}(\hat{X}_1^2\hat{X}_2\hat{X}_3)$. First, we need every possible partitioning of the set $\{1,1,2,3\}$ into unordered pairs, which in this case will give us three sets composed of two pairs,

$$\{(1,1);(2,3)\};\ \{(1,2);(1,3)\};\ \{(1,3);(1,2)\}.$$

Then in Eq. (3.3b), for each set we generate the product $\Sigma_{ij}\cdot\Sigma_{lk}$, where the indices are determined by the pairs in the set, and take the sum of these products,

$$\mathbb{E}(\hat{X}_1^2\hat{X}_2\hat{X}_3) = \Sigma_{11}\Sigma_{23} + \Sigma_{12}\Sigma_{13} + \Sigma_{13}\Sigma_{12},$$

to obtain the desired central moment.

Closure formulae

For the closure formulae derived above we take an example when fourth order moments are approximated based on the assumed distribution. Let us consider a system with at least two variables and investigate how $\mathbb{E}(\hat{X}_1^2\hat{X}_2^2)$ can be expressed in terms of lower order moments. This expression does not depend on any other

variables, hence, regardless of the total number of species in the system, the same formula can be substituted into Eq. (3.1) and applied in all considered systems.

Truncation (zero-closure) means setting all higher order moments to zero, while in all univariate closures mixed moments become zero similarly. Therefore in both cases

$$\mathbb{E}(\hat{X}_1^2\hat{X}_2^2) = 0.$$

Assuming a normal distribution and applying the expressions derived in Eqs. (3.3) and (3.5) we obtain

$$\mathbb{E}(\hat{X}_1^2\hat{X}_2^2) = 2\sigma_{12} + \sigma_1\sigma_2,$$

where

$$\begin{aligned} \sigma_{12} &= \mathrm{Cov}(X_1, X_2) \\ \sigma_1 &= \mathrm{Var}(X_1) \\ \sigma_2 &= \mathrm{Var}(X_2). \end{aligned}$$

If, on the other hand, we assume a lognormal distribution and apply our formulae accordingly, the above moment becomes,

$$\begin{aligned} \mathbb{E}(\hat{X}_1^2\hat{X}_2^2) &= \mu_1^2\mu_2^2(\sigma_1 + \sigma_2 - 1) + \\ &+ 4\mu_1\mu_2(\sigma_{12} + \mu_1\mu_2) - \\ &- 2\left(\frac{\sigma_1}{\mu_1^2} + \frac{\sigma_2}{\mu_2^2} + 2\right)(\sigma_{12} + \mu_1\mu_2)^2 + \\ &+ \frac{(\sigma_{12} + \mu_1\mu_2)^4(\mu_1^2 + \sigma_1)(\mu_2^2 + \sigma_2)}{\mu_1^4\mu_2^4}, \end{aligned}$$

where $\sigma_1, \sigma_2, \sigma_{12}$ are as before and μ_1, μ_2 denote the mean of X_1 and X_2.

Finally, using the above notation for means, variances and covariance, in gamma closure

$$\mathbb{E}(\hat{X}_1^2\hat{X}_2^2) = 2\sigma_{12}^2 + \sigma_1\sigma_2 + \frac{6\sigma_1\sigma_2\sigma_{12}}{\mu_1\mu_2}$$

is used to substitute the expected value of $\hat{X}_1^2\hat{X}_2^2$.

3.3 Results

In the following examples we first investigate the performance of moment closure techniques on two systems, both of which exhibit nonlinear dynamics and show oscillatory behaviour. Sometimes even a deterministic model suffices to capture the mean behaviour, but this depends on the parameter values. There are regimes in the parameter space, where deterministic and low-order stochastic approximation methods qualitatively deviate from the true behaviour. In the first two models we focus on parameter values taken from this regime so that we can explore the improvement our method can provide in the most difficult cases. In our last example we consider a system exhibiting less complex – non-oscillatory – dynamics and demonstrate how our closure techniques can be applied on systems of many interacting species.

Throughout this section we will use the following notation: an mth-order moment closure means the ODE system consists of equations for up to mth order moments and a closure applied in order to develop substituting terms for the $(m+1)$th order moments. As expected, normal closure is equivalent to zero closure in all even order closures, as odd-order terms are always substituted by zeroes. Therefore in most of the analysis we will focus on moment closure approximations based on moments up to an odd order.

3.3.1 P53 system

We first illustrate the properties of our method in the oscillatory p53-Mdm2 system [20]. We denote the three variables, p53, Mdm2 precursor and Mdm2, by x_1, x_2, x_3, respectively. As it has been illustrated in Figure 2.1(c), the variables are connected through a feedback loop with nonlinear rate laws. The reaction network

is described by the stoichiometry matrix

$$S = \begin{bmatrix} 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \tag{3.16}$$

and six reaction propensities

$$\begin{aligned} a_1 &= k_1, & a_2 &= k_2 x_1, \\ a_3 &= k_3 \frac{x_1 x_3}{x_1 + k_7}, & a_4 &= k_4 x_1, \\ a_5 &= k_5 x_2, & a_6 &= k_6 x_3, \end{aligned} \tag{3.17}$$

where k_1 is the p53 production rate, k_2 is the constitutive p53 degradation rate, k_3 and k_7 are the parameters of p53 degradation through Mdm2, k_4 and k_5 are the Mdm2 production and maturation rate, and k_6 is the Mdm2 degradation rate.

Despite the sustained oscillations at the single-cell level, as cells go out of phase a dampened oscillation can be observed in the mean population dynamics. (See Figure 3.2(a)-(b).) Ale et al. [36] has shown that the deterministic approach and LNA are not able to capture this mean behaviour, and in a truncation-based moment approximation at least six moments need to be included. This shows that the variance and the skewness have little effect on the mean, while expanding the moment equations significantly changes the behaviour of the system. We use the parameters and initial values from Ale et al. [36], $\mathbf{x}(0) = [70, 30, 60]$, $\mathbf{k} = [90, 0.002, 1.7, 1.1, 0.93, 0.96, 0.01]$, and compare our results to 10,000 stochastic simulations representing an assembly of individual cells.

In Fig. 3.2(c)-(d) we compare third-order moment closure approximations (including the corresponding zero closure) and the true behaviour obtained from averaging 10,000 SSA runs for state x_1. For the first quarter of the trajectory all closure methods perform similarly, but after that zero closure method is not able to capture the dampening of oscillations while all the other closures show a decrease in

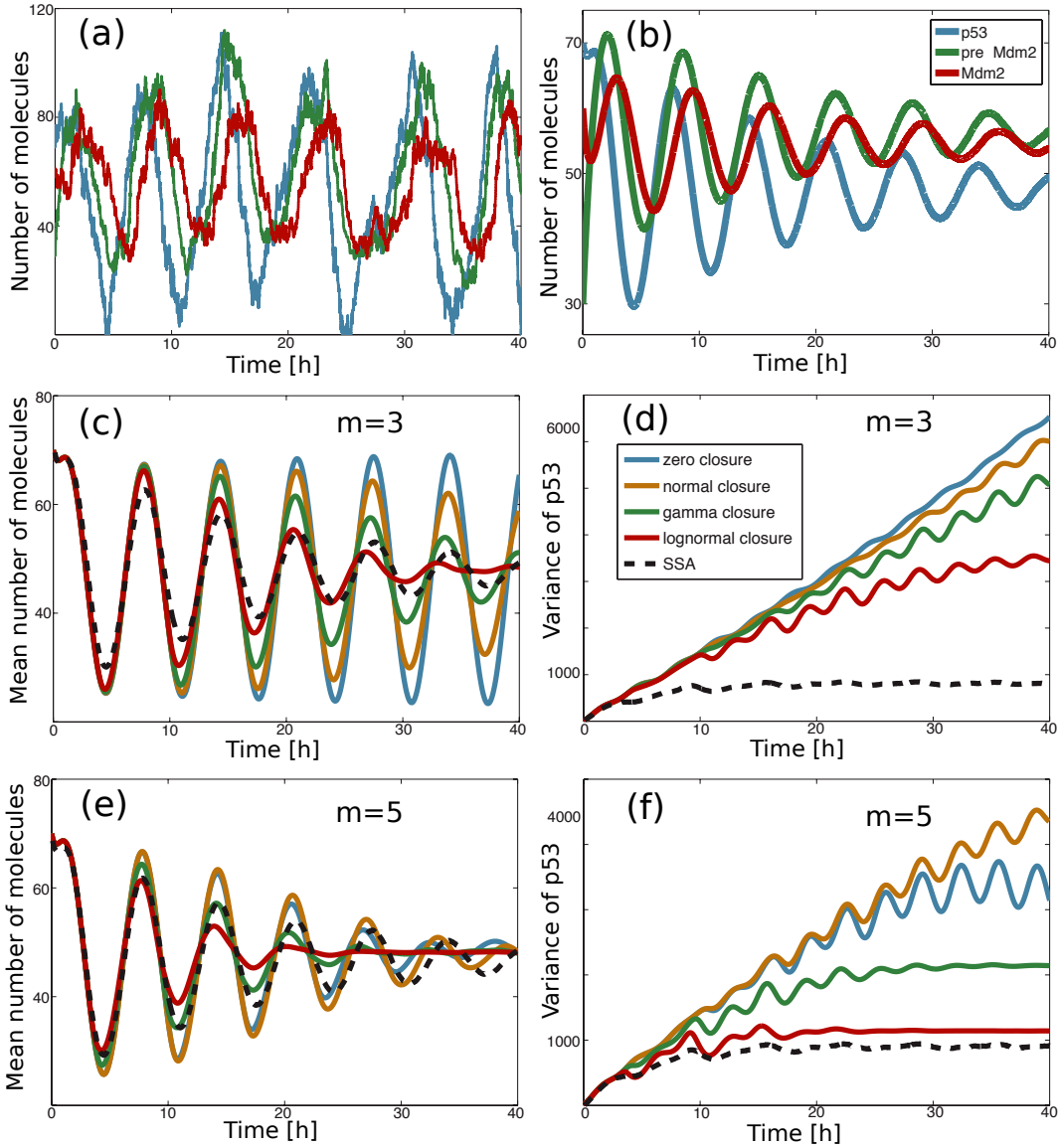


Figure 3.2: Approximation of the p53 oscillatory system with parameters $\mathbf{k} = [90, 0.002, 1.7, 1.1, 0.93, 0.96, 0.01]$ and initial conditions $\mathbf{x} = [70, 30, 60]$. **(a)** A single exact realisation by SSA; colours are as in the legend in part (b). **(b)** Mean behaviour obtained as the average of 10,000 SSA simulations. **(c)** and **(d)** Trajectory of the mean and variance for p53 protein number (x_1) calculated by applying third order zero (blue), normal (orange), gamma (green) and lognormal (red) closure. The dashed black line shows the result of 10,000 SSA runs in correspondence. **(e)** and **(f)** Trajectory of the mean and variance of p53 protein number calculated through fifth order moment closures.

amplitude. As expected, normal closure shows the least pronounced improvement, as the real distribution for p53 molecules is highly asymmetric, which suggests that a Gaussian will not provide a sufficient description. Gamma and lognormal closure both perform well, the latter being the closest but slightly overestimating the dampening. The quality of the approximation is quantified as the cumulative squared difference between the respective methods and the SSA trajectory. By this measure the lognormal closure approximates the true behaviour best, but the error of gamma closure also appears to level off.

As was shown previously for zero closure, the variance is not correctly estimated for fewer than six moments. Using multivariate lognormal closure, the qualitative behaviour of the variance of x_1 is estimated from three moments just as well. On the other hand, assuming a gamma distribution does only slightly improve the approximation qualitatively, as no levelling-off is observed over the time-course considered here. We also extend the simulated time interval to 160 hours and find that the performance of the various closure methods does not change; distribution-based closures all reach the steady state but show varying dampening in the trajectory, while the zero closure method fails to converge to a steady state and the variance approximated by this method grows exponentially instead of levelling off.

We also investigate fifth order distribution-based closure methods, which are compared to sixth order zero closure (Figure 3.2(e) and (f)). All approximation methods overestimate the dampening, and zero closure starts a slight increase around 30 minutes, while lognormal and gamma closure converge to the correct final value. On the other hand the variance of p53 is approximated very well, especially by the lognormal closure method. We also evaluate the difference between multivariate and univariate fifth-order closures and find that the time evolution of variances and covariances diverge more than for third-order closures. This is in agreement with previous findings which showed that including sixth order moments significantly improves the estimation of the variance [36].

We also find that even order moment closures (except for zero closure) result in an ODE system that is numerically too unstable to be evaluated for the entire time course. We use several ODE solvers to obtain trajectories, all of which fail because the system is too *stiff*. That is, the derivatives for some moment equations become too large and hence the time step of the solver needs to be reduced below the minimal allowed step size. Zero closure produces stable and reasonable trajectories, suggesting that some subsequent moments are so strongly coupled that if we include only an approximation of odd-order terms, relevant information is lost providing the stability of the system. We note that problems with ODE solvers occur frequently with moment expansion approaches, especially if vast parameter spaces are explored as is the case in inference; we suspect that ODE's resulting from MEA (or related approaches) may require specific solvers.

3.3.2 Hes1 system

Our second example is the feedback loop on transcription factor Hes1, involved in the regulation of embryonic segmentation, which is also depicted in a 3-variables, 6-reactions model [62]. This system is very similar to the oscillatory p53 model detailed previously. The observables are the level of Hes1 mRNA (x_1) and number of Hes1 protein molecules in the cytosol (x_2) and in the nucleus (x_3) — not all of these can, however, be simultaneously measured *in vivo*. The system is formulated in terms of stoichiometries and propensities as

$$S = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \tag{3.18}$$

$$\begin{aligned} a_1 &= k_1 x_1, & a_2 &= \tfrac{1}{1+(x_3/k_2)^{k_3}}, \\ a_3 &= k_1 x_2, & a_4 &= k_4 x_1, \\ a_5 &= k_5 x_2, & a_6 &= k_1 x_3, \end{aligned} \tag{3.19}$$

where k_1 is the Hes1 degradation rate assumed to be the same for each species, k_2 is the amount of Hes1 protein in the nucleus at half-maximal transcription of the mRNA, k_3 is the Hill coefficient, k_4 is the Hes1 mRNA translation rate, and k_5 is the rate of nuclear transport of Hes1 protein.

The parameter values used (in the above order) are $\mathbf{k} = [0.03, 15, 7, 0.1, 0.008]$, and we start the simulations from $\mathbf{x}(0) = [15, 10, 6]$. As in the case of the p53 system, the results are compared to the average result of 10,000 SSA runs. Similarly to the previous example, sustained oscillations are present in single realisations, but only transiently appear in the mean behaviour and the system soon converges to a steady-state value (Figure 3.3(a) and (b)).

The very low copy numbers of species (on the order of 1-10 molecules) make it harder to follow the evolution of the probability distribution. Due to the stiffness of equations and numerical error introduces through the solution of the ODE system, mean values can take negative values resulting in instability and unrealistic behaviour. We again find that including information of odd order moments without continuing the expansion to the subsequent even order makes the system unstable, hence odd (even) order zero (distribution-based) closures fail to produce a realistic system, although the resulting equations are still numerically tractable.

On the other hand, Fig. 3.3(c) shows that, similarly to the p53 system, third order moment closure methods are able to capture the characteristic dampening. Although even zero closure converges to a steady state (four moments are shown in comparison, as the third order closure is unstable), only lognormal closure estimates the steady state value correctly. As for the variances, values are underestimated by the approximation methods, but the shape of the curve is well captured in all cases and gamma and lognormal closures perform the best.

We also investigate the difference between applying univariate and multivariate closure techniques for this system. In contrast to the p53 system, there is a significant difference between mean values obtained with or without the inclusion of mixed

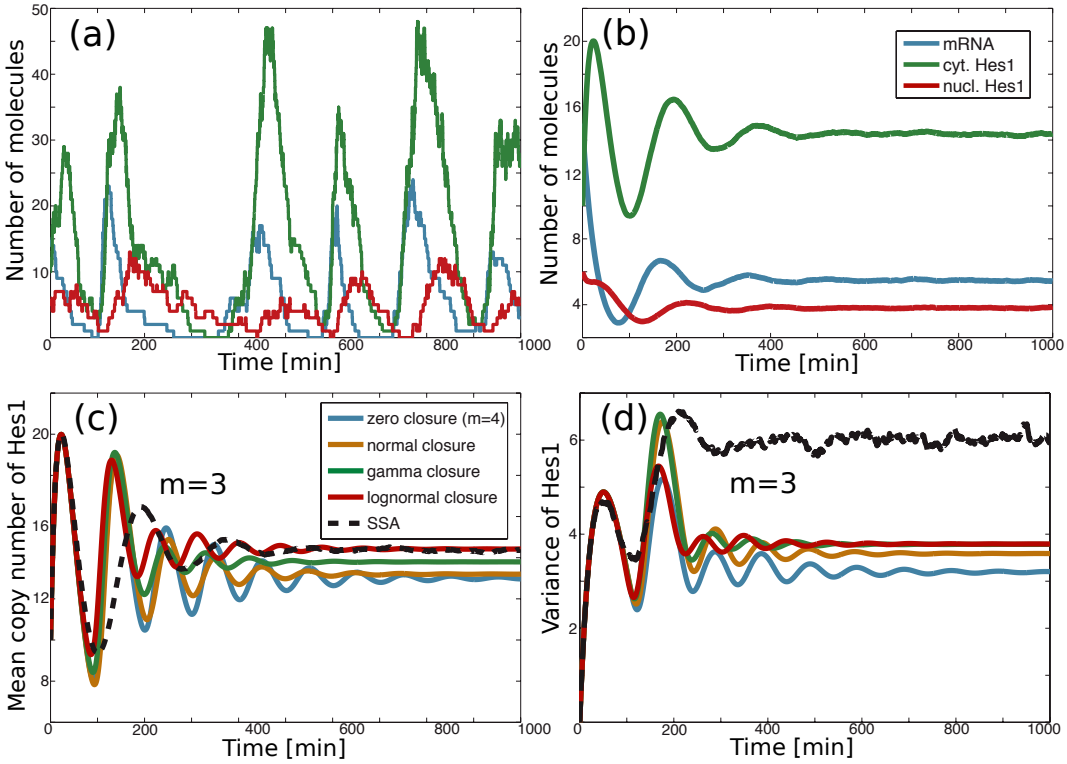


Figure 3.3: Approximation of the Hes1 oscillatory model with parameters $\mathbf{k} = [0.03, 1.5, 7, 0.1, 0.008]$ and initial conditions $\mathbf{x} = [15, 10, 6]$. **(a)** Single SSA simulation of the system; colours as in part (b). **(b)** Average of 10,000 SSA realisations. **(c)** Mean number of Hes1 molecules in cytosol (x_2) obtained from SSA simulations (dashed black line), fourth order zero closure (blue line) and third order normal, gamma and lognormal moment closures (orange, green and red lines, respectively). **(d)** Variance of nuclear Hes1 molecular number calculated as above.

moments, as shown for lognormal closure in Fig. 3.4(a). In univariate closure cases, the amplitude of oscillations reduces over time but they remain throughout the time course, while multivariate closures converge to the steady state much faster. Interestingly, closure without mixed moment formulae performs about the same as fourth order zero closure, although the latter contains more information of fourth order moments.

Furthermore, we compare closures based on the MVG distribution definition of

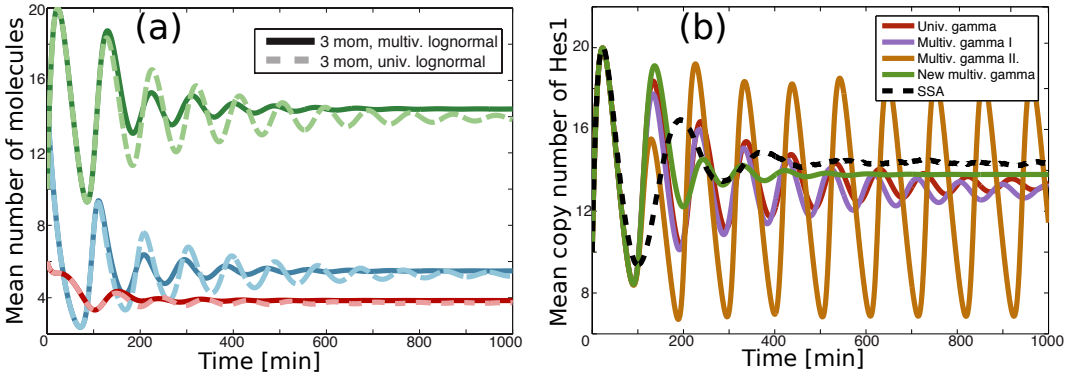


Figure 3.4: Evaluation of multivariate closures on the Hes1 model. (a) Trajectory of the mean for all variables, obtained by univariate (dashed lines) and multivariate (solid lines) lognormal closure. Blue, green and red lines represent mRNA, cytosolic and nuclear protein levels, respectively, as indicated in Figure 3.3(b). **(b)** Mean number of cytosolic Hes1 molecules obtained with an ensemble of 10,000 SSA realisations (black dashed lines) and with univariate (in red) and multivariate closures based on the definition of [68] (purple), [69] (orange) and this work (green).

Mathai and Moschopoulos [68], Furman [69] and the one presented in Eq. (3.10). Figure 3.4(b) shows approximations of the cytosolic Hes1 copy number as compared to the average of SSA runs. Surprisingly, only our new MVG closure offers a superior estimate compared to univariate closure, which suggests that other, more restricted MVG definitions cannot properly model the intricate relationship between higher order mixed moments. This, in agreement with our findings in Fig. 3.4(a), confirms that in this regime the dynamics of the system highly depend on the whole ensemble of interactions between the species.

3.3.3 Erk/Mek system

Up to now we have considered systems with 'challenging' dynamics (see [62]). Here we demonstrate the advantages of our method on a large complex system, modelling the widely studied mitogen-activated kinase, Erk, and its cognate kinase, Mek. Mitogen-activated protein kinases (MAPK) are involved in regulating cellular

fates such as proliferation, differentiation and apoptosis, and therefore many deterministic and stochastic models are available including various details of regulatory and regulated interactions. We choose a very simplified minimal model by Harrington et al. [70], but even this model contains seven species and eight reactions. We reformulate the model in stochastic terms with stoichiometry matrix

$$S = \begin{bmatrix} -1 & 1 & 1 & -1 & 1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}, \tag{3.20}$$

and reaction propensities

$$\begin{aligned} a_1 &= k_1 x_1 x_2, \qquad a_2 = k_2 x_3, \quad a_3 = k_3 x_3, \\ a_4 &= k_4 x_1 x_5, \qquad a_5 = k_5 x_6, \quad a_6 = k_6 x_6, \\ a_7 &= k_7 x_4, \qquad\qquad a_8 = k_8 x_7, \end{aligned} \tag{3.21}$$

where the variables ($x_i,\ i = 1 \dots 7$) denote the absolute numbers of Mek, Erk1, Mek-Erk1 complex, Erk1p, Erk2, Mek-Erk2 complex and Erk2p molecules, respectively. We adjusted initial conditions and parameters to fit the above stochastic model assuming an average system size of $9.24 \cdot 10^5$ (based on the 1540 μm^3 average volume of HeLa cells) and also lowered molecular concentrations into a regime where stochastic effects are prevalent; for a list of the values used in the simulations, see Table 3.1. In our analysis, we focus on the levels of Erk1p and Erk2p, phosphorylated Erk1 and Erk2. In this regime exact simulation of the stochastic dynamics is typically computationally prohibitive and approximate methods are required if we wish to allow for stochastic dynamics.

Variable	Initial value	Parameter	Value
x_1	9240	k_1	$9.524 \cdot 10^{-8}$
x_2	1848	k_2	$8.8 \cdot 10^{-2}$
x_3	0	k_3	$3 \cdot 10^{-1}$
x_4	0	k_4	$9.524 \cdot 10^{-8}$
x_5	7392	k_5	$8.8 \cdot 10^{-2}$
x_6	0	k_6	$2 \cdot 10^{-1}$
x_7	0	k_7	$1.4 \cdot 10^{-2}$
		k_8	$1.4 \cdot 10^{-2}$

Table 3.1: Variables and parameter values used in the ERK/MEK system. Variable values represent the number of molecules initially in the system, while parameter values are given in 1/sec (molecules/sec).

We study second order moment closures, compared to an average of 10,000 SSA runs. We find that in this case of relatively simple, at most second order rate laws, the mean behaviour is completely captured by the approximation methods. Hence, in contrast to the previous two examples, higher order moments have negligible influence on the evolution of the mean. As expected, there is no difference between zero and normal closure, as both methods set all third order terms to zero, but also the results from gamma closure are the same. Fig. 3.5(a) shows the mean $\pm$ two standard deviations obtained by gamma closure (zero and normal closure would result in the same figure) as well as a *single* SSA realisation for both outputs Erk1p and Erk2p. The average trajectory of the 10,000 SSA realisations completely agrees with the approximated means, hence for the sake of transparency it is not shown in Fig. 3.5(a). The same holds even when we reduce the system size 100 times, having only a couple of molecules from some species in the system.

On the other hand, the equation set produced by lognormal closure is numerically

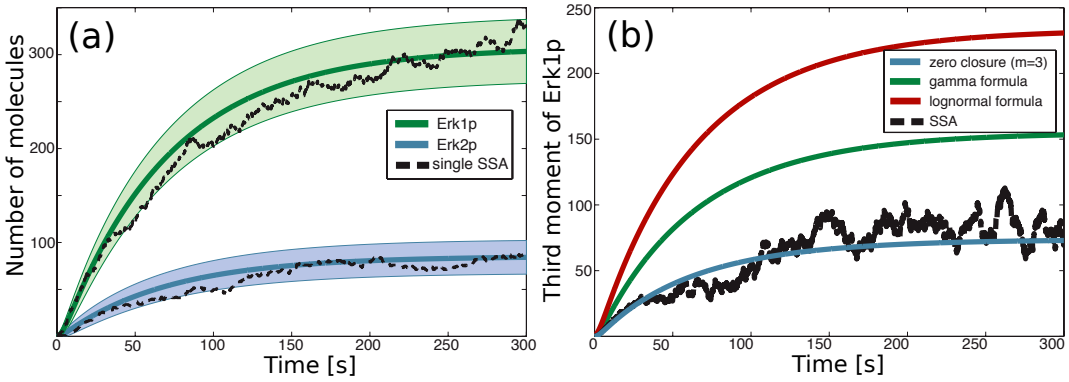


Figure 3.5: Approximation of the simplified Erk/Mek model. Parameters and initial conditions as in Table 1. **(a)** Mean trajectory (solid line) of the two output variables (Erk1p and Erk2p) ± two standard deviations (shaded areas) calculated by second order gamma closure and a single SSA realisation (black dashed lines). **(b)** Approximations of the third central moment of Erk1p obtained by full third order zero closure (blue) and applying gamma (green) and lognormal (red) closure formulae based on the full second order moment closure. The black dashed line shows the evolution of the third moment obtained from a population of 10,000 single realisations. Normal closure is not shown as it is zero for the whole time-course.

unstable and cannot be evaluated by the ODE solvers. This is probably due to the distribution of the molecules having a low skewness compared to its variance, hence a lognormal assumption would force moments to take unrealistic values. This, together with the fact that most variables, especially x_3 and x_6, change very rapidly in the first few seconds, introduces instability into the system. Similarly, variances produced by closure approximations and calculated from the SSA runs completely match, although the SSA-based variance already shows some uncertainty due to limited sample size.

A second order approximation of the system consists of 35 equations, each with between five and 20 terms. Third order approximations would produce an equation set of 118 variables with right-hand side expressions of around 30 terms, that is considerably slower to simulate (not to mention the original costs of determining the

equations — which is here done automatically using computer algebra). Additionally, the risk of introducing numerical errors is also higher. However, deriving third order moments from SSA simulations is not feasible as at least a million simulations would be required to be able to follow the evolution of such moments reliably. In such situations the formulae used during approximation to close the moment equations can be explicitly applied to estimate moments of higher orders based on available moment values. In Figure 3.5(b) we plot the third moment of Erk1p calculated from Eqs. (3.6) and (3.13) by using the numerically obtained second order moment values. For comparison we also show the trajectory obtained by the *third order* zero closure approximation. Note that zero closure – unsurprisingly – provides the best estimate as it is an order higher than the other approximations. Lognormal closure overestimates the actual moment by a factor of three, which supports our previous observation that the instability of lognormal closure arises because this closure ansatz assumes a more highly skewed distribution given the mean and (co)variance values. Gamma closure also overestimates the moment, but qualitatively follows the dynamics, hence provides more information on the actual time evolution than normal closure, which is equivalent to assuming constant zero third order moments.

Some results suggest that if moment closures can be evaluated by their performance – or stability – on estimating the mean, an improved prediction can be obtained by taking a consensus of the 'accepted' third order predictors. For example in Fig 3.5(a) both gamma and normal closure perfectly match the mean, and indeed the average of trajectories derived from gamma and normal closure formulae would give a more precise approximation of the actual third order moment.

3.4 MEANS

A python package for Moment Expansion Approximation, iNference and Simulation

The formulae presented above are implemented together with the automated moment computation in a Python package for ***M**oment **E**xpansion **A**pproximation, i**N**ference and **S**imulation*. The package was developed as a group project by Quentin Geissmann, Saulius Lukauskas and Sisi Fan. I supervised and tested the mathematical aspect of the project, implemented the newly defined gamma distribution and closure, drafted the corresponding manuscript and contributed to the maintenance of the github repository [16].

With the package we aim to address the issue that many algorithms and their corresponding packages with a similar function restrict the calculation to polynomial reaction rates. In addition, an analysis based on them might require expensive softwares or computational skills not generally available to all potential users ([71, 72]; see also http://www.ece.ucsb.edu/∼hespanha). Our package, on the other hand, provides an automated moment equation generation method, suitable for arbitrary rate laws; together with the choice of distribution-based closure formulae, including the multivariate gamma, not available in other packages. Furthermore, the package is complemented with an implementation of an exact stochastic simulation algorithm, as well as support for different ODE solvers and sensitivity analysis, a parameter estimation tool, and specific functions for the visualisation of results. In the following section we provide a brief introduction to the package available from ***https://github.com/theosysbio/means***.

MEANS is a Python package, primarily developed to be used interactively in the IPython/Jupyter notebook framework. Pre-requisites and installation instructions are listed in the accompanying documentation. MEANS enables the definition and exact simulation of stochastic models; moment-based approximations for arbitrary

stochastic system and closure order; numerical solution, sensitivity analysis and parameter inference of ODE systems generated by the approximation or defined by the user; and customisable visualisation for the output. Detailed tutorial notebooks demonstrating the functionalities of the package are available from the `tutorial` directory.

Input: The stochastic system can be supplied by manual input of a symbolic representation of the variables and reactions, all supported by the SymPy library [73]; this is then converted into a specific model object. There is also a parser enabling model definition from SBML files. Furthermore, four predefined models are available from the `examples` directory. Our package also offers a set of routines for serialisation and deserialisation of the MEANS-specific objects to and from files, therefore previously defined models are easily saved and loaded.

Approximation and closure: The MEA method can be called through the `mea_approximation` function, which requires two arguments: a MEANS model object; and the closure order. The default closure method is *scalar* (truncation); distribution-based methods can be used by setting the `closure` argument of the function. MEANS supports normal, lognormal and gamma closures, the latter based on our new definition of the multivariate gamma distribution designed to fit general biochemical networks. When univariate distributions are desired, the parameter `multivariate=False` can be passed to the closure function. We can also generate LNA expressions by calling the function `lna_approximation` where the model is the single argument.

Simulation: MEANS uses Assimulo [74] as the ODE solver back-end, and supports all solvers and settings implemented in that package, which can be accessed via the optional argument `solver`. The default value is *ode15s*, which corresponds to the MATLAB function with the same name. The simulation returns a collection of trajectory objects, each containing a list of time-points, values and a description of the moment. Stochastic systems can be also directly simulated in MEANS by

providing the model and the number of simulations as input.

Furthermore, parameter sensitivities can be computed while performing the ODE simulations using the built-in support of CVODE solvers [75]. Sensitivity analysis returns a collection of trajectory objects with the additional `.sensitivity_data` attribute.

Inference: Parameter estimation is also available in MEANS. For this function the observed data can be entered manually or read from a file; and then converted into a trajectory object with appropriate descriptors so that means and higher moments can be matched. Variable parameters are defined by a dictionary of the symbolic names and corresponding ranges of allowed values; range *None* indicates unbounded parameters, and all unspecified parameters are assumed constant. Starting values for the parameters can be user-defined or generated automatically by optimising runs from random starting positions sampled from a defined range. We advise the latter to lower the chance of finding local minima.

Visualisation: Trajectory objects have automatic built-in visualisation in the notebook, as shown in Fig. 3.6(a). A customisable `.plot()` method allows the use of all options of the standard `matplotlib` plot function. Sensitivity behaviour can be illustrated using `.plot_perturbations()`, which visualises the result of perturbing a single parameter with a specified `delta` value (see in Figure 3.6(b)). Inferred trajectories can also be visualised using the result object's `.plot()` method, cf. Fig. 3.6(c). Besides plotting primary results, MEANS also offers functions to plot the distance landscape and intermediate solutions in order to aid the analysis of inference results.

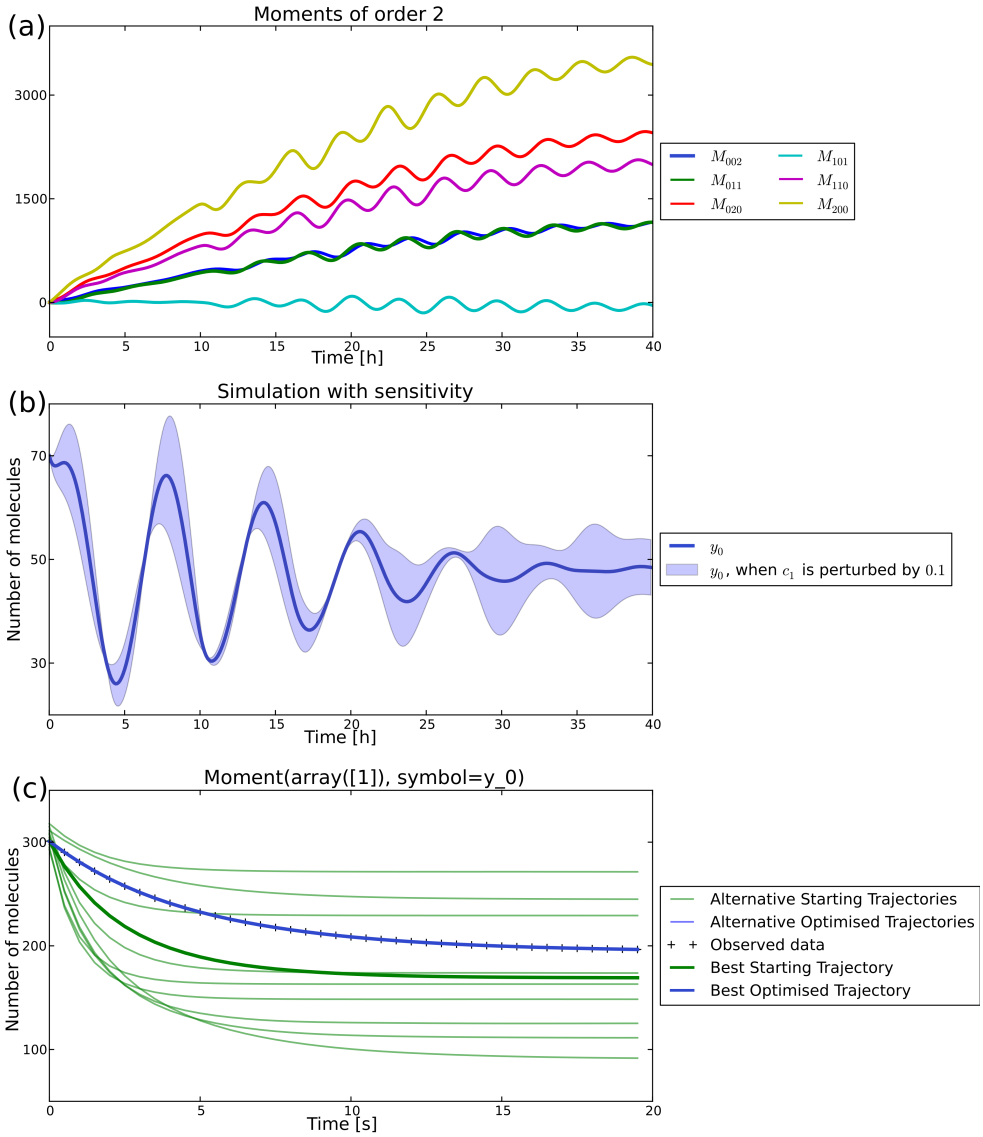


Figure 3.6: Default graphical output of MEANS. Labels are generated automatically. **(a)** Second order moments of the oscillatory p53 system. Initial values and parameters are as in Fig. 3.2. **(b)** Outcome of sensitivity analysis on mean number of p53 molecules in the oscillatory model: original trajectory with shaded region indicating values taken if a single parameter is perturbed. **(c)** Inference with restarts using the dimerisation system and synthetic data: light lines show starting trajectories considered, dark lines show the best starting position and the final fit. Observed data points are obtained by simulating the dimerisation model with parameters as in Fig. 2.4.

3.5 Discussion

In this chapter we have introduced a general moment closure approximation technique based on N-dimensional multivariate forms of normal, lognormal and gamma distributions. We have derived expressions that automatically determine higher order moments in terms of means, variances and covariances, which can be used in any application concerning summary statistics of dependent random variables. Here we have combined the formulae with a general moment expansion approach to follow the time evolution of complex nonlinear systems and shown that moment closure methods are able to approximate the mean behaviour and higher moments in the regime of 5-100 molecules, where many other methods tend to fail [31]. Furthermore, we have also presented MEANS, a Python package, implementing the entire process of moment closure approximation based on multivariate normal, lognormal and newly defined gamma distributions. Besides the derivation of moment equations, MEANS also offers several methods to simulate stochastic models, as well as tools for sensitivity analysis and parameter inference.

With recent developments in single cell recording techniques our approach offers many benefits and allows us to make better use of the information in such experimental data [76, 77]. Since our closure-based approximation method transforms the stochastic problem into a set of ODEs, where moments of the probability distribution are treated as new variables, the approach can be directly used in many applications designed for deterministic systems. For example standard inference schemes [55] can be used to estimate parameters of stochastic systems via conditioning not only on the mean but also variance, skewness – however this is only possible if sufficient amount of experimental measurements is available. We have shown that the use of moment closure formulae instead of truncation considerably reduces the number of moments required to approximate the correct mean behaviour, while going to the same order improves the estimation of higher moments. The reduced system size can be bene-

ficial in most inference tasks, as these algorithms involve evaluating the system for thousands of parameter sets. Furthermore, combined with moment-based distribution reconstruction techniques [78, 79], our method can provide a way to access the whole distribution of the system without analytically solving the CME or costly SSA simulations. Since parametric sensitivities are easily obtained from our analysis (see e.g. [36] and [59]), our framework can also be used to study parameter dependence of the system evolution, e.g. through deriving Waddington landscapes [80] from the probability distribution [81]. However, note that sometimes high order moments are needed for the unique reconstruction of the distribution, and the overall cost of our algorithm and the reconstruction technique might exceed that of the generation of SSA trajectories. On the other hand, in cases when limited number of moments is sufficient, but a range of parameter values has to be investigated (such as in the case of inference), the speed-up provided by our method can be substantial, as illustrated in Appendix A.

Our closure methods rely on the assumption that the observed species follow a specific joint distribution, therefore choosing the best closure technique can be problematic and one has to be aware of the limitations arising from the chosen method. For some systems prior knowledge might help in determining the best distribution to model the species [65], but usually we are not so fortunate and can only base our decision on very general properties. For example, normal closure can lead to errors or failure in systems where the skewness of the probability distribution over states is significantly different from zero. Both the p53 and Hes1 systems exhibit such dynamics and normal closure performs much worse than the two asymmetric distributions. In light of our analyses it appears that lognormal closure should be more successful in cases with highly skewed distributions. In cases where the lognormal ansatz fails, gamma closure might still be applied, but due to positiveness of the univariate gamma distribution higher order mixed moments should all be positive, and hence negatively correlated variables cannot be precisely modelled

with this ansatz. However, our implementation of a new MVG distribution allows the system to assign negative values to these moments based on the corresponding covariances.

In addition, the current algorithm can be refined by including more moments in the calculation of closure expressions. In cases where the actual distribution cannot be described by any of the distributions considered here, an improved approximation might be obtained by deriving the parameters from different pairs of lower order moments – e.g. by expressing $\mathbf{m}$ and $\mathbf{\Sigma}$ of a lognormal distribution in terms of mean and skewness as well as mean and variance – and using a weighted average of the results. However, this method would either require tighter restrictions on the closure order, m, or considerable computational time.

In case we have no prior information about the distribution of species, a possible way of finding the optimal closure is to calculate all three second order approximations and compare their stability, as incorrect assumptions can lead to unstable systems, such as in the case of the Erk/Mek network. Moreover, if experimental data or data from a couple of SSA simulations is available, the reliability of each closure can be estimated, similar to the estimation of necessary order moments in Ale et al. [36]. If the system size allows, comparing full third order moments (obtained from third order closure) to the ones calculated by the respective assumptions can also help in deciding which distribution fits the data best. After the distribution assumptions are evaluated with one of the above approaches, the optimal or a mixture of reasonable closures can be used to model the distribution of the system. For example closure formulae can be used as an estimator of higher order moments even when further expanding the approximation is not feasible or efficient. However, although the three included cases can fit a very broad range of possible molecular distributions, note that there are still some examples which cannot be well approximated, e.g. bistable systems and other multimodal distributions. Investigating further distributions, such as Poisson or (negative) binomial distribution, can also

improve the approximation as they have been shown to have relevance in some biochemical models, but they lack a general multivariate formulation for which closure formulae could be derived automatically for an arbitrary system.

For some such systems general criteria have recently been developed [82] that allow us to determine whether moment closure approximations can be usefully applied. Whenever nonlinear dynamics and noise interact it poses a challenge for computationally convenient approximations [83]; approaches such as the one introduced here can easily improve the analysis of such systems compared e.g. to the LNA. However, many biologically important processes will defy these attempts, and for these there may be no alternative to costly SSA simulations. Moreover, we would like to emphasise that even in cases when SSA simulations are not feasible, moment closure methods are not the only technique that can provide insight into the system – quite often there are more specialised and better suited tools. For example in oscillatory systems, such as the p53 and Hes1 examples, studying the average power spectrum of the model can provide more useful information [84], especially compared to a low-order moment estimate that has significant error. Where approximations probably have a role to play is in statistical inference: here even approximations to the true pdf may greatly help in determining the likelihoods of observed data [32, 36, 85, 86].

A related further application of our framework is in the field of experimental and control design [60, 87, 88]. This field offers numerous mathematical and engineering tools developed for differential equations systems, which applied to a stochastic model followed with our algorithm could lead to automatic, *in silico* experimental design for biochemical systems. In the next chapter we introduce such a method for evaluation of control and modelling.

Four

Reachability analysis of control in stochastic systems

The previous chapters introduced approximation techniques that convert a stochastic problem into a set of deterministic equations. In this chapter we combine these results with a deterministic method from control theory to investigate regulatory models in terms of variability.

4.1 Introduction

Much of the research in synthetic and systems biology in the last decade has focused on the study of elementary biological systems. This has often been with the aim of controlling and modifying them in order to achieve new functional modules exhibiting novel and useful behaviour [89], such as sustained oscillations [90] and bistability [91]. It is now becoming possible to use engineered biological systems made of characterised components to solve specific problems, such as information processing, energy production or the production of chemicals.

However, biological systems are inherently noisy and probabilistic in nature, which can pose significant difficulty for anyone aiming for a reliable, well-characterised

module. Although several external control techniques have been developed, which are able to avoid some of the variability in a population [92, 93], noise at the level of molecular processes is often unavoidable and can, for example, affect quite profoundly how information is transmitted along the molecular networks underlying cell function [94, 95]. Furthermore, due to the unreliability of measured quantities, our understanding of the underlying mechanisms might be mistaken leading to suboptimal analysis and design. Therefore we have to assess the practical limits on the amount of noise in a general biological module, in order to evaluate the efficiency of a control design or the reliability of a mechanistic model.

Reachability analysis has been widely used in control design and engineering for applications such as verification of electrical or mechanical networks and hybrid automata [96–100]. The analysis focuses on the computation of the subset of the state space that can be reached within a certain time-limit, given some starting position of the system and external inputs. The technique can also be used to verify that a certain undesired state is not reached under realistic operating conditions, or that the behaviour of the system is robust and qualitatively/quantitatively holds for different conditions, including different realistic inputs to the system.

To this end reachability is generally calculated under varying levels of uncertainty regarding details of the system – such as initial state, input signal and rate parameters — usually formalised by assuming that these parameters come from a set of plausible values. Therefore, unless analytical solutions are derived for some abstraction of the system, the applicability of the computation heavily relies on the choice of the set representation [98,101,102]. Some representations might prove computationally expensive and hence impractical for high dimensional systems, while a simple shape representation can lead to crude over-approximation of the reachable set. Methods have been proposed using several techniques such as polygonal projections [103], oriented hyper-rectangles, special polyhedra [104], ellipsoids [102], or level sets [98]. In this work we use zonotopes [96], a centrally symmetric type of

polytopes that can be conveniently represented by a list of vectors.

Although there is already an enormous body of work on reachable set computation for problems in engineering, including highly nonlinear cases [105], hybrid automata [106], and differential-algebraic equations [107], the complexity, frequent nonlinear behaviour, and strict constraints on many of the model parameters of biochemical systems require the development of specialised analysis methods. There are already a few applications of reachability techniques to biological examples with special emphasis on the treatment of the nonlinearity of the system, either through direct computation [108] or through hybridisation-based methods using either static or dynamic partitioning of the state space [109–111]. The work by Dang et al. [111] has also been expanded to take into account the possible lack of knowledge of parameter values [112]. The stochasticity in biological systems has been tackled in even more diverse ways: through computing bounds on the probability function [113], using stochastic hybrid systems [114], or analytically deriving invariant sets for linear equations obtained from the stochastic model [115].

In this work we propose a computationally efficient and flexible method to compute the reachable set (in a zonotopic representation) of stochastic biochemical systems; besides stochasticity we also consider possibly nonlinear rate laws, controlled or uncertain input signals, and uncertainty about model parameters that might also represent control over these values. The main steps behind our derivation are the following: (i) we first obtain an ODE representation of the system's mean and (co-)variances; (ii) then use an iterative procedure to obtain a conservative approximation of consecutive reachable sets up to a final time of interest. Here we primarily use the LNA for the first step, but also consider the MEA [15, 16, 36] to generate equations for the second step. We derive a new method to tightly approximate realistic biological input signals, and a piece-wise temporal linearisation method is applied to deal with many common nonlinearities. We also give conservative approximation formulae for the reachable set if rate parameters of the system are not

precisely known — which is very often the case in biochemical systems.

The method is demonstrated on two elementary modules fundamental to mathematical models and regulatory designs of biochemical networks. The first system presents the use of reachability analysis for the study of noisy biochemical reactions and evaluating a control on the levels of cell heterogeneity; the second example considers the task of model (in)validation for cases when high cell-to-cell variability poses a challenge to estimating the system's true behaviour. Finally, we explore the limitations of our method and how it can be utilised as a quick indicator of 'challenging' dynamics.

4.2 Zonotopes

A crucial aspect of reachable set computation is the choice of representation for the sets considered. We have to find a compromise between precision – i.e. giving a very close approximation of the real reachable set – and efficiency – i.e. using a representation method which does not require particular computational expenses. A great advantage of zonotopes, our chosen set representation for reachability analysis, is that they can provide a compact encoding for quite complex polytopes that cannot be captured with basic geometric forms.

A zonotope [116] is described by the position of its centre and a set of generator vectors as

$$\mathcal{Z} = (\mathbf{c}; < \mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_p >) := \left\{ \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} = \mathbf{c} + \sum_{i=1}^{p} a_i \mathbf{g}_i,\ -1 \leq a_i \leq 1 \right\}. \quad (4.1)$$

Figure 4.1(a) shows an example of a zonotope with two and three generators. In the following we use the shortened notation $(c\ ; G)$ to represent $\mathcal{Z}$, where $G \in \mathbb{R}^{n \times p}$ is a generator matrix formed from the generator column vectors. Zonotopes are a convenient representation as they are closed under *Minkowski-addition* (also known as *dilation*) and *affine transformations*, the two key operations in reachability analysis [96]. Furthermore, the above can be calculated through simple matrix-vector operations: given zonotopes $\mathcal{Z} = (c\ ; G)$ and $\mathcal{W} = (d\ ; H)$, $(H \in \mathbb{R}^{n \times q})$ and the affine transformation $\mathbb{T}(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$,

$$\mathcal{Z} \oplus \mathcal{W} = \{\mathbf{x} + \mathbf{y} \mid \mathbf{x} \in \mathcal{Z},\ \mathbf{y} \in \mathcal{W}\} = (c + d\ ; [G, H]), \quad (4.2a)$$

$$\mathbb{T}(\mathcal{Z}) = \{A\mathbf{x} + \mathbf{b} \mid \mathbf{x} \in \mathcal{Z}\} = (Ac + \mathbf{b}\ ; AG), \quad (4.2b)$$

where $[G, H]$ is the concatenation of the 'vector lists' in G and H, i.e. an $n \times (p+q)$ matrix whose first p columns consist of the matrix G, and its last q columns are of matrix H.

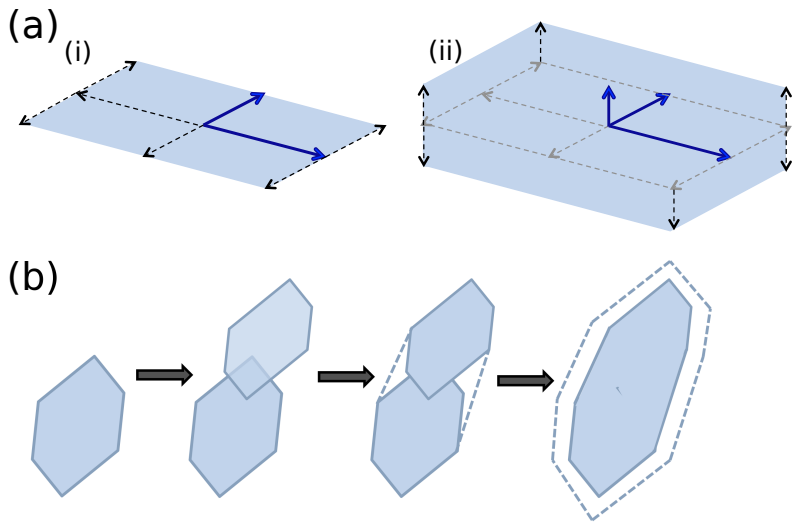


Figure 4.1: Computation and propagation of zonotopes. **(a)** Example of a zonotope (light blue shaded area) with two generators (blue arrows) **(i)**. **(ii)** The same zonotope after the addition of a third generator. **(b)** General method for the computation of reachable set for a time interval. Given the reachable set at time t (shaded area in leftmost panel), it is propagated into the reachable set at $t + \tau$ (second from the left), then the convex hull of the two sets are computed (third panel) and finally this set is enlarged to account for trajectories of non-linear dynamics (area encircled by dashes in rightmost panel).

4.3 Reachable set computation

4.3.1 Problem formulation

First take a state-space formalisation [117] of the general deterministic system described by the evolution of the n-dimensional state variable, $\mathbf{x}(t)$,

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(0) = \mathbf{x}_0,$$

where $\mathbf{u}(t) \in \mathbb{R}^m$ is an input signal, and $f(\mathbf{x}, \mathbf{u})$ is a generally nonlinear but time-independent transition function. Such deterministic systems can be obtained in several ways through approximation of a stochastic biochemical representation. In the following work we primarily use the LNA [30], to derive equations that describe

the population average and variance as the variables in ODEs.

Consider a stochastic system with N-dimensional state variable $\chi(t)$ describing the abundance of modelled species at time t. The system is described by a stoichiometry matrix, S, and reaction propensities, $F = [a_1,\ a_2, \ldots, a_r]'$. The LNA separates the state variable into a macroscopic part ($\phi(t)$) and random fluctuations ($\xi(t)$). This form yields an ODE system describing the rate of change of the means ($\mathbb{E}$) and covariance matrix (Σ). This set of ODEs (Eq. (2.12)) can be summarised by the $\mathbf{x}(t) \in \mathbb{R}^n$ vector, where $n = N(N+3)/2$,

$$\dot{\mathbf{x}}(t) = \frac{d}{dt}\begin{bmatrix} \mathbb{E}(t) \\ (\Sigma(t))^v \end{bmatrix} = \begin{bmatrix} S \cdot F(\mathbb{E}(t)) \\ (SD\Sigma(t))^v + (\Sigma(t)(SD)^T)^v + (S\mathrm{diag}\{F(\mathbb{E}(t))\}S^T)^v \end{bmatrix} = f(\mathbf{x}(t)),$$

where the upper triangle vectorisation operator stands for row-continuous listing of all elements in and above the diagonal: for example $(\Sigma(t))^v = [\Sigma_{11}, \Sigma_{12}, \ldots, \Sigma_{1N}, \Sigma_{22}, \ldots, \Sigma_{2N}, \Sigma_{33}, \ldots, \Sigma_{NN}]$. We can thus obtain a generic state-space form of a stochastic system.

Using this formalism we focus on the case where there is a single input ($\mathbf{u}(t) \in \mathbb{R}$) and the input-dependence in the above equation can be separated as

$$f(\mathbf{x}, \mathbf{u}) = g(\mathbf{x}) + Bu, \tag{4.3}$$

where $B \in \mathbb{R}^{n\times 1}$ is the input matrix, which specifies which states are affected by the input.

Now assume that, instead of starting the system from a known $\mathbf{x}_0$ value, we have a set in the state-space, $\mathcal{I}$, containing all possible initial states – for example because we observe a range of expression values in a population of cells. Similarly, the input signal also comes from a set, $\mathcal{U}$, which is bounded by some value $\mu \in \mathbb{R}$ so that $||u(t)|| \leq \mu\ \forall t$. Then Eq. (4.3) is presented as a differential inclusion [118] of the form

$$\dot{\mathbf{x}}(t) \in g(\mathbf{x}(t)) \oplus B\mathcal{U};\quad \mathbf{x}(0) \in \mathcal{I}. \tag{4.4}$$

The reachable set of this system at any time t is defined as

$$\mathcal{R}_t := \{\mathbf{y} \in \mathbb{R}^n \mid \exists(\mathbf{x},\ u);\ \mathbf{x}(0) \in \mathcal{I},\ \mathbf{u}(s) \in \mathcal{U},\ \mathbf{x}(s) \text{ follows (4.4) } \forall s \leq t,\ \mathbf{y} = \mathbf{x}(t)\}, \tag{4.5}$$

i.e. all the states that can be achieved having the above described system start from a possible initial state and affected by plausible input signals. Similarly, we can define $\mathcal{R}_{[0,t]}$ as all the states reachable within the time interval $[0, t]$, by computing all individual reach sets: $\mathcal{R}_{[0,t]} = \bigcup_{s\in[0,t]} \mathcal{R}_s$.

4.3.2 Reachable set propagation

We start our derivation considering the case when function f is a linear function of $\mathbf{x}$; under such condition the original system (Eq. (4.3)) can be represented in the linear time-invariant (LTI) form,

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + Bu(t);\quad \mathbf{x}(0) = \mathbf{x}_0, \tag{4.6}$$

where $A \in \mathbb{R}^{n\times n}$ is the state (transition) matrix. As noted before, we focus on single input systems, but all results can be easily generalised to other values of input dimension, $m \neq 1$. The solution of this system can be derived analytically as

$$\mathbf{x}(t) = e^{At}\mathbf{x}_0 + \int_0^t e^{A(s)}Bu(t-s)ds, \tag{4.7}$$

which can be further simplified if we assume a constant input signal, $u(t) = u\ \forall t$. We will discuss later why such an assumption is valid in our application. Further assuming an invertible transition matrix leads to the expression

$$\int_0^t e^{A(s)}ds \cdot Bu = A^{-1}\left(e^{At} - I\right)Bu, \tag{4.8}$$

where I is the $n \times n$ identity matrix. In the following for convenience we use the notation $\psi(A, t) = A^{-1}\left(e^{At} - I\right)$ and compute the state of the system at time t as

$$\mathbf{x}(t) = e^{At}\mathbf{x}_0 + \psi(A, t)Bu. \tag{4.9}$$

We consider an equidistant partitioning of the time horizon into N intervals, with time-step $\tau = T/N$. We start the derivation of reachable sets from $\mathcal{I}$ to $\mathcal{R}_T$ by handling the case of the autonomous system, i.e. when $u \equiv 0$.

Given a reachable set at an arbitrary time-point, $\mathcal{R}_t = (c\,; G)$, the new set reached under zero input can be calculated according to Eq. (4.9), as

$$\mathcal{R}_{t+\tau} = e^{A\tau}\mathcal{R}_t, \tag{4.10}$$

since the second term becomes zero for autonomous systems. This operation is in fact an affine transformation $x \to Mx$ (with $M = e^{A\tau}$) and hence the resulting set will also be a zonotope, as given by Eq. (4.2),

$$\mathcal{R}_{t+\tau} = (e^{A\tau}c\,; e^{A\tau}G).$$

As the same equation applies for all values of t, the transition matrix can be applied iteratively to propagate $\mathcal{I}$ up to the final reach set, $\mathcal{R}_T$. In case the aim is to map the entire space the system explores between times 0 and T, the reachable sets of the time intervals, $\mathcal{R}_{[i\tau,(i+1)\tau]}$ has to be derived from the two end-points. Generally, this can be done by computing the convex hull of $\mathcal{R}_{i\tau}$ and $\mathcal{R}_{(i+1)\tau}$ and then extending this set to contain all affine solutions for one time-step (see [101] and Figure S1(b) for an illustration). However, zonotopes are not closed under this operation and it is desirable to avoid the exact computation of a convex hull. We use the conservative ("over–")approximation presented in [96] to derive a set enclosing all points reachable between 0 and τ. All following intervals can be propagated from this set, as for all $t \in [0, \tau]$, $\mathcal{R}_t \in \mathcal{R}_{[0,\tau]}$ and hence $e^{A\tau}\mathcal{R}_t \in e^{A\tau}\mathcal{R}_{[0,\tau]}$.

The above computation is easily extended to account for a constant input signal, u. According to Eqs. 4.8 and 4.10, $\mathcal{R}_{t+\tau}$ is given by

$$\mathcal{R}_{t+\tau} = e^{A\tau}\mathcal{R}_t + A^{-1}\left(e^{A\tau} - I\right)Bu,$$

which is conveniently also a zonotope, since $A^{-1}\left(e^{A\tau} - I\right)Bu = \psi(A, \tau)$ is a constant vector. However, in systems with uncertainty (so, for example, where the input

signal is unknown) we need to transform and enlarge, or *bloat*, the set $\mathcal{R}_{i\tau}$ in each iteration, to account for the system dynamics under all admissible input/parameter values. This corresponds to taking all combinations of an unknown initial value (contained in $\mathcal{I}$) and unknown parameter (input signal or reaction rate value) and obtaining every state the system can thus possible reach. This is generally done by taking the Minkowski sum of the reachable set propagated from $\mathcal{I}$, $\mathcal{R}_{t+\tau}$; and the bloating sets, β_μ and β_δ, accounting for the uncertainty introduced in one time-step by the input or rate parameter set, respectively. In the following sections we derive these bloating sets in a zonotope formalism for convex, one-dimensional sets of input and parameter values.

To summarise, using the short form $\mathcal{R}_i$ to denote either $\mathcal{R}_{i\tau}$ or $\mathcal{R}_{[(i-1)\tau,i\tau]}$ given in a zonotope form, the next reachable set is calculated as

$$\mathcal{R}_{i+1} = \left(e^{A\tau}c_i \; ; e^{A\tau}G_i\right) \oplus \beta_\mu \oplus \beta_\delta. \tag{4.11}$$

where the terms β_μ and β_δ are computed in each iteration. We describe how to obtain these bloating factors in the next sections.

Computation of $\psi(A, t)$

The reachable set calculation based on Eq. (4.9) requires the inverse of A which is only possible if A is regular. However, biological models often lead to transition matrices that are singular regardless of the exact parameter values chosen. One can overcome this problem by either solving the expression symbolically or numerically computing the integral of the matrix exponential [119]. The prior is too computationally demanding to be used in a nonlinear or parameter-varying system where this calculation has to be executed in each time-step, while the latter might not be applicable to arbitrary matrices [120]. We use the generalist approach of deriving an approximation of Eq. (4.8), based on the integral of the Taylor-series of e^{As} and

computing a bloating factor β_ν to correct for the small error thus introduced [107],

$$\psi(A,t) = \sum_{i=0}^{\nu} \frac{A^i t^{i+1}}{(i+1)!} \oplus \underbrace{\left(0 \; ; e^{|A|t} - \sum_{i=0}^{\nu} \frac{|A|^i t^i}{i!}\right)}_{\beta_\nu}. \tag{4.12}$$

The case when ν approaches infinity leads to the exact Taylor expansion of $\psi(A,t)$, but large values of ν make the above calculation expensive. In practice, however, values over 5 already lead to an almost negligible error, while both parts of the above expression are still easily computed. We find, through exploration of different scenarios that $\nu = 8$ provides the best compromise between computational efficiency and accuracy. Furthermore, whenever A is time-independent, the computation can be executed in the initialisation of the algorithm, hence non-singular matrices do not present an additional load for our algorithm.

On the other hand, we want to emphasise that the inevitable singularity of the transition matrix informs us that the variables of our model are not independent, and one or more could be expressed by the others. For example in the dimerisation system mentioned in Section 2, the total number of monomers and dimers is not independent, as the molecules can only either associate or dissociate and change between being free or involved in a dimer. In such cases one might write down conservation equalities to eliminate some variables and reactions, until the resulting set is independent, whence A becomes invertible [121]. Furthermore, due to the elimination step, the system size is reduced and hence both computational time and numerical error decrease. This approach, however, is not easily automated and usually requires manual modification of the system before the analysis can begin. Nevertheless, if the conservation equations are available together with the original system description, it is more advantageous than the approximation of $\psi(A,\tau)$.

4.3.3 Input for biological systems

In the algorithm proposed for systems in electronics [96], the input set, $\mathcal{U}$, is taken to be an n-dimensional hypercube enclosing all points between $[-\mu, \mu]$ in all dimensions – equivalent to the radius μ ball in the infinity norm. As $\mathcal{U}$ is symmetric with no distinction between the dimensions, the bloating set representing the effect of the input, β_μ, is a hypercube similar to $\mathcal{U}$ with radius given by

$$\frac{e^{||A||\tau} - 1}{||A||}\mu,$$

where $||A||$ is the maximal row-sum of the elements of A (i.e. the infinity norm). Such a generalised approximation of the input set is sometimes necessary, as the analysis is motivated by a target or an avoidable set in the state-space and the input set is meant to capture all variations induced by noise in a physical/electrical system.

In the case of biological inputs, however, we usually have a more detailed information about $\mathcal{U}$, meaning that the aforementioned general input set would contain lots of implausible signals and hence the over-approximation would be too loose to provide useful information on the actual reachable sets. For example, in the typical case the control input is implemented as a certain type of molecular species added to the system; this cannot take negative values, and it is also reasonable to assume an upper limit on the number of molecules injected at any time point, which can be viewed as the bound μ. Furthermore, setting appropriate values of the input matrix, B, to zero is a good representation of the fact that only certain variables (i.e. the one representing the type of input molecule) are directly affected by the input signal.

As $\mathcal{U}$ is typically not centred around zero, we divide the input effect into two parts. A drift term, corresponding to the effect associated with the centre of $\mathcal{U}$ (u_c), which defines the change in the system's state under a constant average input signal; and an uncertainty term representing our lack of knowledge about the exact

input value, based on the maximal difference, u_d, between the centre and possible values of $\mathcal{U}$. Thus the whole input set is taken into account as the interval $\mathcal{U} = [u_c - u_d; u_c + u_d]$; for instance, in the example above, $u_c = \mu/2$ and $u_d = \mu/2$. The drift and uncertainty terms are calculated using Eq. (4.8), and the bloating set is summarised in zonotope form,

$$\beta_\mu = (\beta_{\mu\text{–drift}} \, ; \beta_{\mu\text{–uncertainty}}) = (\psi(A,\tau)Bu_c \, ; \psi(A,\tau)Bu_d) \, . \tag{4.13}$$

The major advantage of this representation lies in the ease of performing the Minkowski addition in Eq. (4.11) to derive the reachable set under uncertain input.

4.3.4 Parameter uncertainty

In the previous section we summarised techniques for treating control and/or uncertainty in a simple, additive input term. It is straightforward to extend this method to account for more subtle inputs, even those characterised by a chemical reaction, as long as the transition matrix of the system, A, is unaffected. Such a case can arise for example if the rate of a zeroth-order reaction is under our control. We will discuss this case in our first example.

However, often we have to make predictions but only have approximate values of the parameters contributing to matrix A, either due to imperfect knowledge of reaction rates [122], or because we can control a reaction through some interaction that is not or cannot be modelled explicitly. Therefore, we derive a way to account for uncertainty in case the matrix A is also drawn from a set or ensemble of matrices. For example, let us consider the case where a single reaction rate, k, possibly affecting more than one element of A, is not known precisely or controlled. We assume that k has some plausible upper and lower bounds and hence it can be considered as coming from an interval centred at the nominal value $\hat{k}$, i.e. $k \in [\hat{k} - \delta, \hat{k} + \delta]$. We approach the problem by following the nominal dynamics using the previously defined matrix, A, and defining a bloating term to enclose all solutions arising from

admissible parameter values, so that

$$\mathcal{R}_{i+1}(A(k)) \subseteq \mathcal{R}_{i+1}(A(\hat{k})) \oplus \beta_\delta,$$

where $A(k)$ denotes the matrix formed using the parameter value k. For a positive $\mathbf{x}$, we can derive a bound for $A(k)\mathbf{x}$ with $A(\hat{k})\mathbf{x} + D\mathbf{x}$ where D is the $n \times n$ matrix computed as $|A(\hat{k}+\delta) - A(\hat{k})|$.

In order to derive a general formula that can be obtained efficiently, we make $D\mathbf{x}$ independent of the state variable by using a conservative estimate of $\mathbf{x}$. We approximate $D\mathbf{x} \leq D|\mathbf{x}|_{max}$, where $|\mathbf{x}|_{max}$ can be approximated generally for the whole algorithm, or, more practically, in each time step based on the states in the current reachable set. $|\mathbf{x}|_{max}$ is derived from the zonotope definition in Eq. (4.1): given a reachable set in the form (c, G), the ith coordinate of the maximal vector is computed by the formula

$$\{|\mathbf{x}|_{max}\}_i = \{c\}_i + \sum_{k=1}^{p} |\{g_k\}_i|. \tag{4.14}$$

Note that $|\mathbf{x}|_{max}$ might not be in the reachable set, but there is at least one point for each coordinate for which $\{\mathbf{x}\}_i = \{|\mathbf{x}|_{max}\}_i$ holds. $|\mathbf{x}|_{max}$ is derived as $|c| + \sum |G|$ with all summations carried out *by rows*. We also take into account a rough estimate of the next reachable set, so that in non-converging cases $|\mathbf{x}|_{max}$ is not underestimated: $|\mathbf{x}|_{max} = \max\{|c| + \sum |G|, e^{A\tau}(|c| + \sum |G|)\}$.

For each coordinate of the system the difference between dynamics under any k and the centre, $\hat{k}$, is enclosed in the interval defined by the most extreme value of possible dynamics,

$$\forall d_i = \{A(k)\mathbf{x}\}_i - \{A(\hat{k})\mathbf{x}\}_i : \quad d_i \in [-\{D|\mathbf{x}|_{max}\}_i; \{D|\mathbf{x}|_{max}\}_i], \tag{4.15}$$

from which a zonotope representation can be obtained as $(\mathbf{0}\,; \mathrm{diag}\{D|\mathbf{x}|_{max}\})$. From here we proceed considering this conservative approximation of the deviation from the nominal trajectory as an input set centred at zero; in agreement with the fact that

trajectories of the nominal value, $\hat{k}$, are already calculated through the transition matrix. As in the previous section, the input set can be represented in zonotope form, so

$$\beta_\delta = (\mathbf{0}\ ; \psi(A,\tau)\mathrm{diag}\{D|\mathbf{x}|_{max}\})\,, \tag{4.16}$$

provides the bloating set for parameter uncertainty to be used in Eq. (4.11).

4.3.5 Nonlinear systems

Biological systems are often nonlinear, as even the simplest dimer-formation requires second-order rate laws that cannot be eliminated from the system. Nonlinear systems can lead to complex and unpredictable behaviour and their analysis can lead to unexpected complications that have to be tackled by specialised methods [123–125]. The most popular way to overcome the challenges of nonlinearity is by performing a (piece-wise) linearisation of the system, in each segment of an either temporal or spatial partitioning of the time-horizon and state-space studied. In this work we apply the former approach and in each step linearise the system around its current centre using first-order Taylor expansion, as described in [126]. This results in a system which is piece-wise linear, and has LTI properties between two time-steps, but overall time-varying and is therefore only an approximation of the real underlying dynamics. The linearised system calculated at time $i\tau$ is

$$\dot{\mathbf{x}}(t) = \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}=c_i} \times (\mathbf{x}(t) - c_i) + f(c_i) = A_i\mathbf{x}(t) - A_i c_i + f(c_i). \tag{4.17}$$

As this system is now linear, we can use results on LTI systems to approximate the reachable set at time $(i+1)\tau$ given the reach set $\mathcal{R}_{i\tau}$. The propagation of a single trajectory, $\mathbf{x}(i\tau)$ to $\mathbf{x}((i+1)\tau)$, is computed as

$$\begin{aligned}\mathbf{x}((i+1)\tau) &= e^{A_i\tau}\mathbf{x}(i\tau) - A_i\int_0^\tau e^{A_i\tau - s}ds\cdot c_i + \int_0^\tau e^{A_i\tau-s}ds\cdot f(c_i)\\ &= e^{A_i\tau}\mathbf{x}(i\tau) - e^{A_i\tau}c_i + c_i + \psi(A_i,\tau)f(c_i),\end{aligned}$$

which corresponds to the affine transformation $\mathbf{x} \to M\mathbf{x} + m$, where now $M = e^{A_i\tau}$ and $m = -e^{A_i\tau}c_i + c_i + \psi(A_i, \tau)f(c_i)$. Hence, given a zonotope $(c_i \ ; \ G_i)$, the next reachable set of the linearised system is obtained as

$$\mathcal{R}^l_{i+1} = (Mc_i + m \ ; \ MG_i).$$

However, this set is only the set of states reachable through the linearised system, Eq. (4.17). To enclose points reached through the original nonlinear dynamics, an additional bloating factor is added that accounts for the difference between the original and the linearised dynamics

$$\mathcal{R}_{i+1} = (c_i + \psi(A_i, \tau)f(c_i) \ ; e^{A_i\tau}G_i) \oplus \beta_\epsilon. \tag{4.18}$$

The bloating term, β_ϵ, is computed iteratively based on the same procedure proposed in [126]. However, we use the zonotopic set representation to obtain a tight estimate of the effects of nonlinearity.

4.3.6 Linearisation error

Let $\epsilon(\mathbf{x})$ denote the difference of the nonlinear function f and its corresponding linear system for a given time step,

$$\epsilon(\mathbf{x}) = |f(\mathbf{x}) - [A_i(\mathbf{x} - c_i) + f(c_i)]|\,, \tag{4.19}$$

i.e. the *local error* of the linearisation algorithm. In a similar fashion to how parameter uncertainty is tackled, we aim for a conservative approximation of this function that can be evaluated independently of $\mathbf{x}$. The local error is zero at the centre, c_i. Using the substitution $\mathbf{x} = c_i + \Delta\mathbf{x}$, we obtain ϵ as a function of $\Delta\mathbf{x}$, the distance between a point and the centre of the current set of states. Next, we define an upper bound for $\Delta\mathbf{x}$, based on the same principle as Eq. (4.14),

$$\{|\Delta\mathbf{x}|_{max}\}_i = \sum_{k=1}^{p} |\{g_k\}_i|, \tag{4.20}$$

and approximate $\epsilon(\Delta x)$ with a function monotonously increasing with Δx. Hence we obtain an estimation for which $\epsilon(\mathbf{x}) \leq \hat{\epsilon}(|\Delta \mathbf{x}|_{max})\ \forall \mathbf{x} \in \mathcal{R}$. Similarly to Eq. (4.16), this expression enables us to define a generator set,

$$G_\epsilon = \mathrm{diag}\{\hat{\epsilon}(|\Delta x|_{max})\}, \tag{4.21}$$

which represents a bound on the nonlinearity of the system.

To obtain the reach set of the time-interval $[i\tau, (i+1)\tau]$, further approximations have to be applied to enclose all trajectories between the two time-points. As the piece-wise linear system is time-varying, the estimation of the time-interval $[0, \tau]$ is not sufficient. Instead, as mentioned before (Fig. 4.1(b)), the convex hull of two delimiting reach sets is calculated and bloated, the bloating factor for which can be derived in several ways as described in [96] and [126].

Although the computation in Eq. (4.18) assumes a constant time-step, it can be modified to incorporate variable steps, τ_i, without loss of efficiency, as in the nonlinear case the matrix exponential, $e^{A_i\tau}$, should anyway be computed in each iteration. Consequently, the first order linearisation used above can be combined with a higher-order, adaptive method by computing the central trajectory (c_i) and each time-step (τ_i) with the higher-order numerical technique, followed by reachable set computation based on these quantities. This way stiff systems with rapidly changing dynamics can be tackled without significant additional computing time, and slowly changing systems can gain a speed-up compared to fixed time-step calculations. Alternatively, we could replace the linearisation with a more flexible description of the stochasticity than the LNA, such as moment closures [15] or finite–state projection methods [127].

Example – Dimerisation

To demonstrate how a stochastic – and in particular, nonlinear – system is transformed into state-space representation, let us consider the dimerisation system, described in Section 2.1.2. Representative quantities of the system are listed in Eq. (2.6), and as derived in Eq. (2.15), the LNA leads to a differential equation system of five variables, which we will denote by the state vector, $\mathbf{x} \in \mathbb{R}^5$. The state-space form of the evolution is a summary of the above equations as $\frac{d}{dt}\mathbf{x} = f(\mathbf{x})$.

$$\frac{d}{dt}\underbrace{\begin{bmatrix} m \\ d \\ \sigma_m \\ \sigma_{md} \\ \sigma_d \end{bmatrix}}_{\mathbf{x}} = \begin{bmatrix} 2k_1 m(1-m) + 2k_2 d \\ k_1 m(m-1) - k_2 d \\ 4k_1 m(-2\sigma_m + m - 1) + 4k_1\sigma_m + 4k_2(\sigma_{md} + d) \\ 2k_1 m(\sigma_m - 2\sigma_{md} - m + 1) + k_2(2\sigma_d - \sigma_{md} - 2d) + k_1(2\sigma_{md} - \sigma_m) \\ k_1 m(4\sigma_{md} + m - 1) + k_2(d - 2\sigma_d) - 2k_1\sigma_{md} \end{bmatrix}.$$

To obtain a linearised system for time-point $t_i = i\tau$, we take the Jacobian of the above transition function,

$$\frac{df}{d\mathbf{x}} = \begin{bmatrix} -4k_1 m + 2k_1 & 2k_2 & 0 & 0 & 0 \\ 2k_1 m - k_1 & -k_2 & 0 & 0 & 0 \\ -8k_1\sigma_m + 8k_1 m - 4k_1 & 4k_2 & -8k_1 m + 4k_1 & 4k_2 & 0 \\ -4k_1\sigma_{md} + 2k_1\sigma_m - 4k_1 m + 2k_1 & -2k_2 & 2k_1 m - k_1 & -4k_1 m + 2k_1 - k_2 & 2k_2 \\ 4k_1\sigma_{md} + 2k_1 m - k_1 & k_2 & 0 & 4k_1 m - 2k_1 & -2k_2 \end{bmatrix}.$$

This translates to a different transition matrix at each time step. Hence, for example, if at t_i the centre is $c_i = [150, 75, 90, -45, 25]'$, the evolution of the system for the next time-step according to Eq. (4.17) will take the form

$$\frac{d}{dt}\begin{bmatrix} m \\ d \\ \sigma_m \\ \sigma_{md} \\ \sigma_d \end{bmatrix} = \underbrace{\begin{bmatrix} -588k_1 & 2k_2 & 0 & 0 & 0 \\ 299k_1 & -k_2 & 0 & 0 & 0 \\ -1916k_1 & 4k_2 & -716k_1 & 4k_2 & 0 \\ -238k_1 & -2k_2 & 299k_1 & -358k_1 - k_2 & 2k_2 \\ 119k_1 & k_2 & 0 & 588k_1 & -2k_2 \end{bmatrix}}_{\frac{df}{d\mathbf{x}}\big|_{\mathbf{x}=c_i} = A_i} \begin{bmatrix} m - 150 \\ d - 75 \\ \sigma_m - 90 \\ \sigma_{md} + 45 \\ \sigma_d - 25 \end{bmatrix} + \underbrace{\begin{bmatrix} -44700k_1 + 150k_2 \\ 22350k_1 - 75k_2 \\ 475072k_1 \end{bmatrix}}_{f(c_i)},$$

with some k_1 and k_2 parameter values. It is important to note that depending on the values of c_i, the above A_i matrix often becomes singular or near-singular (almost zero determinant). This is expected, since the dimerisation system (Fig. 2.1(b)) is in fact a closed system where the total number of monomers (either free or assembled into a dimer) is constant – therefore monomer and dimer numbers hold redundant information. We can either use Eq. (4.12) to approximate the inverse or reformulate the system by incorporating the knowledge that the amount $d + 2m$ is constant.

The calculation of the linearisation error is based on the difference function (Eq. (4.19)) between the true value of $f(\mathbf{x})$ and the linearly computed value $A_i(\mathbf{x} - c_i) + f(c_i)$. This function is vector-valued, hence we can explore the local error for each coordinate. In the dimerisation system, with centre c, the first-step estimation of the linearisation error of the first coordinate (i.e. mean monomer number) at state x is expressed in terms of deviation from the centre ($\Delta x_1 = x_1 - c_1$) as

$$\begin{aligned}\{f(c+\Delta\mathbf{x})\}_1 - \left\{ \frac{df}{d\mathbf{x}}\bigg|_{\mathbf{x}=c} \Delta\mathbf{x} + f(c) \right\}_1 &= -2k_1(c_1+\Delta x_1)(c_1+\Delta x_1-1) + 2k_2(c_2+\Delta x_2) - \\ &- [-(4k_1c_1+2k_1)\Delta x_1 + 2k_2\Delta x_2 - 2k_1c_1(c_1-1) + 2k_2c_2] \\ &= -2k_1\Delta x_1^2.\end{aligned}$$

The entire vector of $\epsilon(\mathbf{x})$ can be derived similarly. As the error function is the absolute value of the above and we aim to derive a monotonous function of $\Delta\mathbf{x}$, we use the inequality $|a+b| \leq |a| + |b|$ to obtain $\hat{\epsilon}(\mathbf{x})$. In practice this vector is stored with symbolic variables and evaluated at each time point with estimations of the maximal deviation from Eq. (4.20). For example, for an arbitrary zonotope yielding the vector $|\Delta\mathbf{x}|_{max} = [20, 10, 5, 3, 4]$:

$$\hat{\epsilon} = \begin{bmatrix} 2k_1dx_1^2 \\ k_1dx_1^2 \\ 4k_1dx_1^2 + 8k_1dx_1dx_3 \\ 2k_1dx_1dx_3 + 2k_1dx_1^2 + 4k_1dx_1dx_4 \\ k_1dx_1^2 + 4k_1dx_1dx_4 \end{bmatrix} = k_1 \begin{bmatrix} 800 \\ 400 \\ 2400 \\ 720 \\ 640 \end{bmatrix}.$$

In other nonlinear examples, it is possible for some variables to have no linearisation error if they have linear dynamics in the first place, i.e. they are not involved in any multi-molecular reactions. However, over the time course from $i\tau$ to $(i+1)\tau$,

the linearisation error in the other variables will have an influence on them, which is reflected in the vectors building up the linearisation error bloating set, β_ϵ.

Example – Reachable set computation

Figure 4.2 provides a visual demonstration of the major computational steps introduced in this chapter. The toy model used for illustration is the *deterministic* system with transition matrix $A = \begin{bmatrix} -3 & 0 \\ 2 & -1 \end{bmatrix}$ that can be interpreted as a biological system with the following reactions:

$$x_1 \xrightarrow{3} \emptyset, \quad x_1 \xrightarrow{2} x_1 + x_2, \quad x_2 \xrightarrow{1} \emptyset.$$

In addition, there is a fourth reaction $\emptyset \xrightarrow{u} x_1$ under our control, providing controllable input to the system. We assume our knowledge of the starting state is inaccurate and given with error bounds: $x_1(0) = 20 \pm 5$ and $x_2(0) = 35 \pm 10$. In our zonotope formalism the initial set is captured as $\mathcal{I} = \left(\begin{bmatrix} 20 \\ 35 \end{bmatrix} ; \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix} \right)$. Figure 4.2(a) shows the first few reachable sets computed as this initial set is propagated through the reachability algorithm, without any external input. Note that, as the transition matrix defines a stable, contracting system, the reachable sets get smaller and show convergence towards the origin, the stable steady state of the system.

Figure 4.2(b) illustrates the computation of quantities $|\mathbf{x}|_{max}$ and $|\Delta x|_{max}$ – used in the correction terms for parameter uncertainty and non-linearity – on a two-dimensional zonotope. The illustration also highlights that the actual state corresponding to the coordinates $|\mathbf{x}|_{max}$ or distance $|\Delta x|_{max}$ is usually not part of the zonotope.

Finally, the bloating used in the propagation of systems under input is demonstrated in Figure 4.2(c). Given the reach set, $\mathcal{R}_1$, and an input signal that might assume any values between 0 and 100, we first compute the reachable set with the fixed value $u = 50$. In the next step the set is bloated with a set of generators such

that the two extrema (0 and 100) are both accounted for; giving rise to a reachable set, $\mathcal{R}_2$, containing all states achievable.

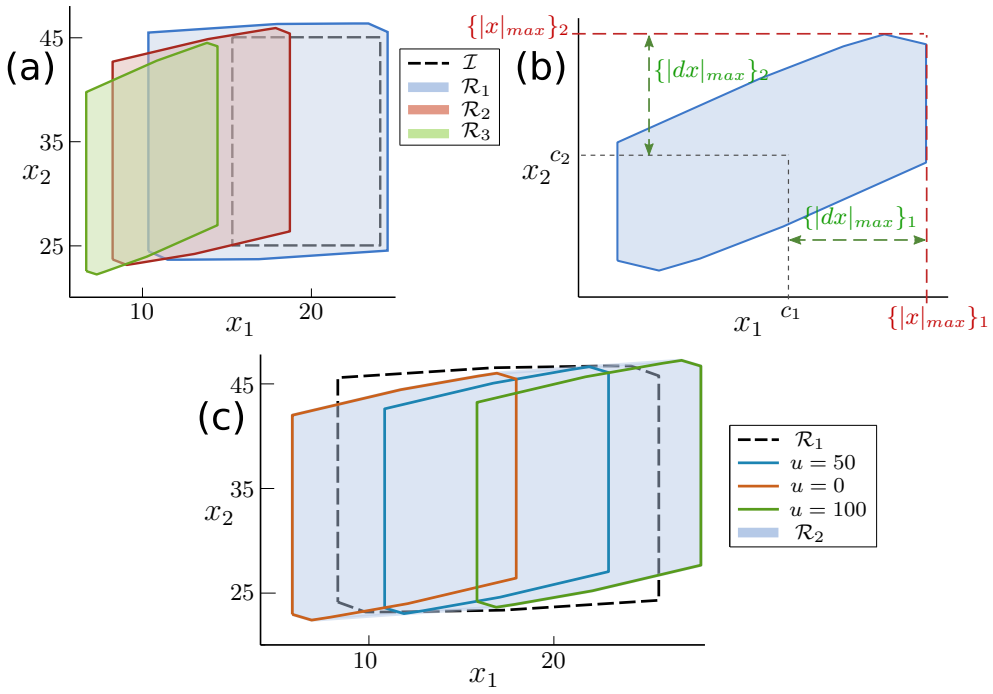


Figure 4.2: Examples of reachable set computation. **(a)** Propagation of the initial set, $\mathcal{I}$ (rectangle outlined by black dashes), following the system dynamics by transition matrix A, with zero input and time-step $\tau = 0.08$. Blue, red and green shaded areas show the set of states reachable in the time-intervals $[0, \tau]$, $[\tau, 2\tau]$, and $[2\tau, 3\tau]$, respectively. **(b)** Two-dimensional example of the computation of $|\mathbf{x}|_{max}$ and $|\Delta x|_{max}$. The centre of the zonotope and its respective coordinates are indicated with black dotted line, maximal coordinates with red dashes. Green dashes indicate the distance defining $|\Delta x|_{max}$. **(c)** Computation of bloating set accounting for a range of input signal values. Black dashes show the original reachable set, blue solid line the reachable set obtained with the middle of input range, 50. Blue shading represents the bloated reachable set. Dark orange and green shapes outline reachable sets with 0 and 100 input, respectively, which are both contained in the bloated set.

4.4 Applications

4.4.1 Controlled gene expression

The first example considered is the controlled stochastic gene expression system [115, 128]. In spite of its simplicity, this model is one of the most important examples as protein production is a necessary and elementary building module in both naturally evolved and synthetic biological systems. Many applications might rely on the steady operation of such a unit, hence it is crucial to study what the achievable noise levels and possible states of operation are.

The model of stochastic gene expression has two variables and four reactions, characterised by the stoichiometry matrix and reaction propensities described in Eq. (2.5). The *controlled* system we investigate is technically the same, except that transcription is not constitutive, but regulated by a signal. The corresponding reaction propensities are

$$a_1 = k_1\mu; \quad a_2 = k_2 m; \quad a_3 = k_3 m; \quad a_4 = k_4 p,$$

where the multiplier in the first reaction, μ, represents our control over the transcription rate. In this formalisation $0 \leq \mu \leq 1$, and we may assume discrete ($\mu = 0$ or $\mu = 1$) or continuous ($\mu \in [0, 1]$) values. A typical control signal is delivered through a light-sensitive setup [129], where transcription can be switched on or off in every time unit, represented by μ being a sequence of zeros and ones.

After performing the LNA, just like in the dimerisation example above, we obtain a set of five ODEs, determining the time evolution of the means and (co)variances of the studied species. The system is started from $\mathcal{I} = \mathbf{0}$ and we investigate the reachable sets up to time $T = 10$ units with time-step $\tau = 0.01$ and reaction rates $\mathbf{k} = [100, 5, 100, 1]$ Just as in Section 2.3, the parameters are scaled to represent a general system where the time is in arbitrary units.

As the reachable sets enclose any possible sequence of μ, including $\mu(t) \equiv 0$ and $\mu(t) \equiv 1$, any reachable set will have to contain the trajectories implied by a sequence starting with n zeros, which equal to the ones evoked by the signal lacking the zeros, just shifted in time. Hence $\mathcal{R}_{[0,\tau]} \subseteq \mathcal{R}_{[\tau,2\tau]} \subseteq \cdots \subseteq \mathcal{R}_{[T-\tau,T]}$, and it is sufficient to study the last reachable set of the iteration.

Figure 4.3 shows projections of the final reachable set together with two sample trajectories. The samples are computed as the mean and variance of the population consisting of 10,000 direct realisations of the original stochastic system. The input signal $\mu(t)$ is defined as a piece-wise constant function with values randomly drawn from $\{0, 1\}$ and switching once or twice every time-unit (u_1 and u_2, respectively); the signal is kept the same for all realisations contributing to a particular trajectory in Fig. 4.3. We also take into account some uncertainty regarding the protein degradation rate, i.e. that the value of k_4 is unknown or can be externally changed to values within 5% of the nominal value, 1. The light blue regions in Figure 4.3 show the estimate of the reachable set under such uncertainty. Interestingly, while a substantial part of the mRNA–protein mean space can be covered, the reachable states on the protein mean–variance plane are limited to a very narrow band. Therefore, if relying on the production of a protein with this module we will have to make a compromise: either have a low number of proteins produced, or a high number but with great variability (as may be expected given the Poisson nature of this process).

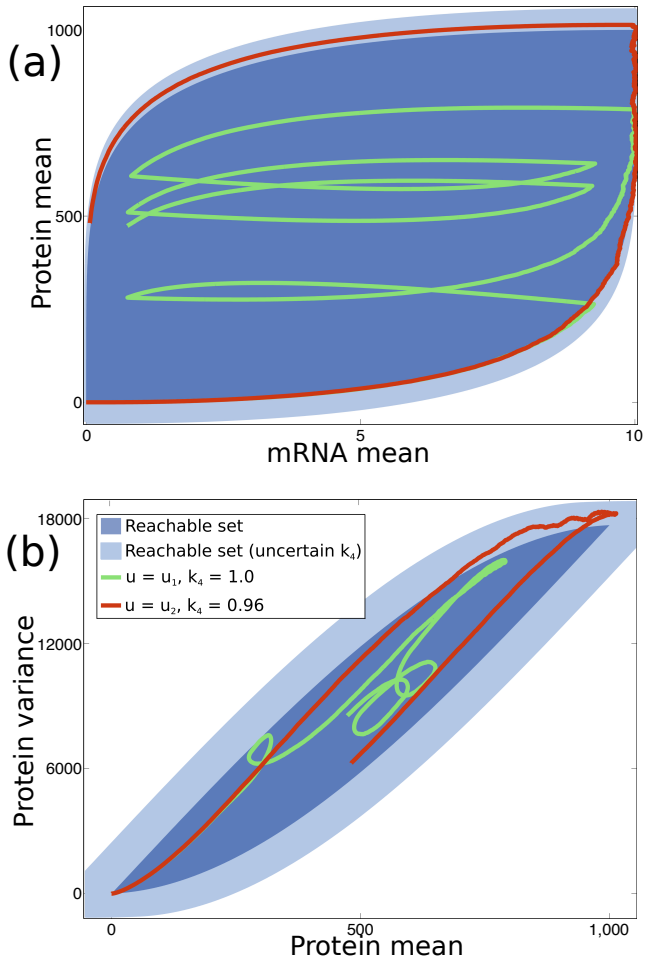


Figure 4.3: Reachable states of the controlled stochastic gene expression system with additional uncertainty. Blue shaded regions show projection of the final reachable set to **(a)** the mRNA mean–protein mean plane and **(b)** the protein mean–protein variance plane. Dark and light blue shades indicate reachable sets without and with 5% uncertainty in parameter k_4. Red and green example trajectories are calculated from 10,000 exact simulations, under the input sequences $u_1 = [1,0,1,1,1,0,1,0,1,0]$ and $u_2 = [1,1,1,1,1,1,1,1,0,0,0,0,0,1,0,1,1,0,1,1]$, respectively, with protein degradation values as indicated in legend.

4.4.2 Model validation

In our second application we consider a chain of three molecules affecting each other's production and aim to demonstrate how our reachability analysis can con-

tribute to model validation for stochastic systems. The output of our method in this application is similar to the results of Anderson and Papachristodoulou [130], but since our focus is on stochastic systems and measurements affected by intrinsic noise, we derive reachable intervals of a given model structure in terms of mean $\pm$ standard deviations.

Let us consider the network consisted of molecules A, B and C in Figure 4.4 that can represent any chain of interacting species with similar reaction networks; for example a simple model of transcription factors, where only protein levels are modelled explicitly. The regulatory effect of these molecules is through the production rate of their target species. Three different wiring schemes are considered: (i) molecule A induces molecule B that in turn activates the production of C; (ii) molecule A also activates molecule C, such that this effect is more profound than the activation via B; (iii) molecule C feeds back onto, and induces the production of, A. In each model we simplify the mathematical description by assuming equal degradation rates for all species; hence the models can be summarised by five parameters: k_{AB}, k_{BC}, k_{AC}, k_{CA} and k_{deg}, where the subscript XY refers to the activation of Y by X. In all models $k_{AB} = 1$ and $k_{deg} = 0.8$ and the other parameters are chosen to reflect the connections of the model and produce similar maximal values in the output, C.

In order to model measurements, we generate five individual stochastic trajectories from Model (i), with parameter values $k_{BC} = 1$, $k_{AC} = k_{CA} = 0$ and an initial A value randomly chosen from the range $[80, 120]$. The parameter values are in [1/a.u.]. We then sample the amount of molecule C, associated with the experimentally measurable output, from these simulations at six time-points; and evaluate the distance of each measurement point from the reachable set of mean values at that specific time. The distance is normalised by the maximum reachable value of the standard deviation and zero for all points within the reachable set. As Figure 4.4(b) and (c) show, it is very unlikely that the observations arose from Model (ii). Although the true model shows the best correspondence to the data, Model (iii)

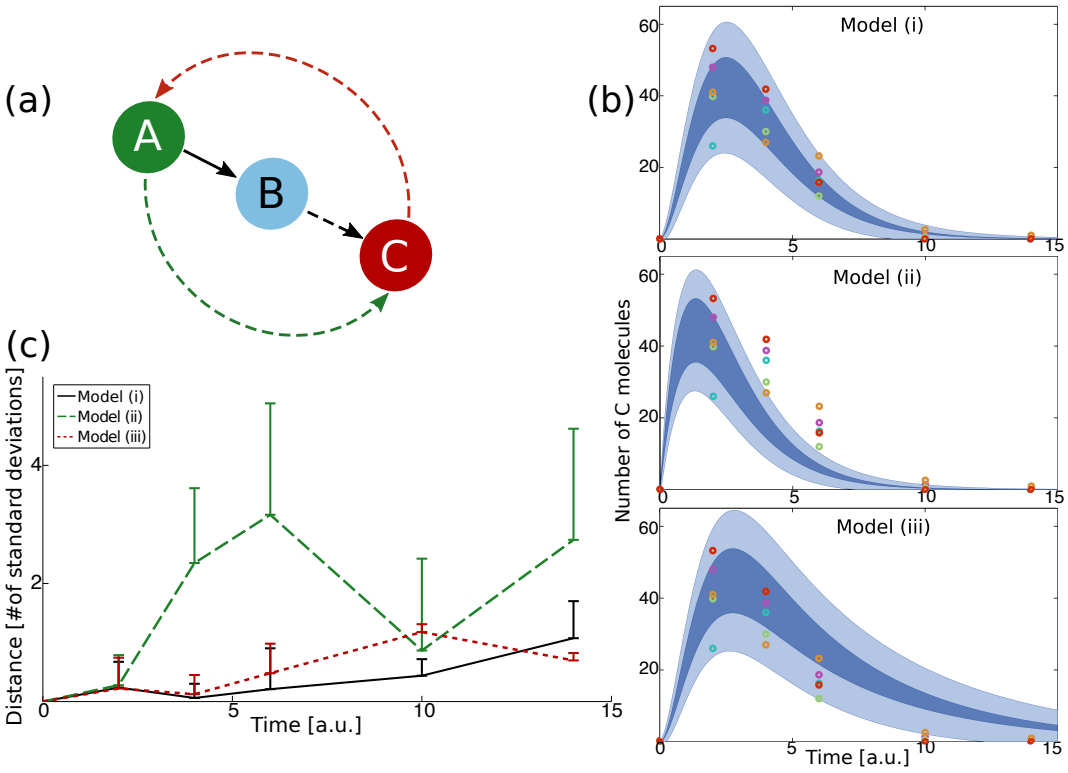


Figure 4.4: Model evaluation by comparing reachable sets and single measurements. **(a)** Schematics of the three reaction chain models. Models differ in rates corresponding to dashed arrows. (i): $k_{BC} = 1,\ k_{AC} = k_{CA} = 0$. (ii): $k_{BC} = 0.1,\ k_{AC} = 0.9,\ k_{CA} = 0$. (iii): $k_{BC} = 1,\ k_{AC} = 0,\ k_{CA} = 0.2$. **(b)** Reachable region over time of the output (molecule C) starting from the initial set $80 \leq A_0 \leq 120$, $B_0 = C_0 = 0$. Dark areas represent reachable values of the mean, light blue shades are the $\pm$ 1 standard deviation region computed with the maximal reachable value of the variance. Coloured circles are sample points taken from single exact simulations of Model (i). **(c)** Distance of observation points from the reachable set of mean values. Lines show the average distance of observed data at each time point for each model (colours as indicated in legend), the top of error bars depict the maximal distance at the evaluation points.

— representing a feedback system — cannot be discarded due to the wide range of values molecule C can reach in that specific wiring scheme. If the relative noise level is reduced by raising the average initial abundance (to $\sim$ 300 molecules), or mea-

surements from another species (e.g. molecule A) are also available, the distinction between different models becomes clear with Model (i) unambiguously fitting the data best.

4.4.3 Bistable system

Our last example is a model exhibiting bistability, on which we demonstrate limitations of zonotope-based reach set computation, but also how reachability analysis can be used for *qualitative* exploration. As bistability plays a crucial role in cellular decision making [131], we choose a concrete stem cell differentiation model to examine this phenomenon. The model describes the interaction of the Oct4-Sox2 complex with stem cell marker Nanog [132]. Although it is a simplified representation of the underlying network that summarises numerous reactions into rates used in modelling, it still produces bistable behaviour. The model has two variables (N and OS) and five reactions,

$$a_1 = k_1 \cdot \left(\frac{k_2 OS^{\gamma_{os}}}{(K + OS^{\gamma_{os}})k_{deg}} \right)^2; \qquad a_2 = k_{deg} OS;$$
$$a_3 = \frac{k_3 OS^{\gamma_{os}}}{K + OS^{\gamma_{os}}}; \qquad a_4 = \frac{k_4 N^{\gamma_n}}{K + N^{\gamma_n}}; \qquad a_5 = k_{deg} N,$$

where the reactions stand for auto-activated production of the Oct4-Sox2 complex, degradation of the complex, Oct4-Sox2 activated production of Nanog, auto-activated production of Nanog, and degradation of Nanog, respectively. Accordingly, the stoichiometry matrix of the system is,

$$S = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 \end{bmatrix}.$$

The exponents γ_{os} and γ_n indicate co-operativity, and take values 1 and 2 respectively, as Nanog is known to form dimers. The term in a_1 is squared to represent the combination of Oct4 and Sox2-related kinetics that, in our example, have the same reaction rates. We take parameter values from [132] that make the model bistable: $[k_1, k_2, k_3, k_4, k_{deg}, K] = [0.03, 50, 0.1, 14, 1, 10]$.

Here we focus on how bistability affects our prediction of the average behaviour. Therefore, in practice, our analysis is equivalent to computing the reachable sets of the corresponding deterministic system. We fix rate parameters to the above values, and introduce uncertainty through allowing a range of initial values. As noise in the initial Oct4-Sox2 concentration, OS_0, has no significant effect on the dynamics, we fix its value to 60. Similarly, variance and covariance values do not influence mean Nanog level, hence their value is also without variability, $\sigma_{os} = 10$, $\sigma_{os-n} = 1$, $\sigma_n = 0.1$. On the other hand, as the system is in a bistable regime, Nanog initial values determine which of the two fixed points – corresponding to high-Nanog, stem cell phenotype or low-Nanog, differentiated phenotype – is reached by the system. Therefore we explore different initial values of Nanog, N_0, with a 25% uncertainty in each case.

The value $N_0 = 0.6$ is close to the boundary between the two basins of attraction (c.f. Figure 2A in [132]). Hence the initial set $\mathcal{I} = [0.45, 0.75]$ inevitably contains a sample of states from both basins. We first compute the reachable set up until 15 arbitrary time-units. As zonotopes are limited to describe centrally symmetric and convex reach sets, the set enclosing both high- and low-Nanog steady states also includes a set of implausible, negative values that arise as the symmetric counterparts of the highly positive (> 10) Nanog values. We exclude this range of values as it is clearly an artefact of set representation and concentrate on the strictly positive subset of reachable states. Figure 4.5 shows Nanog mean values in blue on a logarithmic scale as a function of time.

As a comparison, we also conduct reachability analysis restricted to the stem cell or differentiated cell regime. The former is started from the initial set $\mathcal{I} = [1.5, 2.5]$ and the latter from Nanog values $N_0 \in [0.3, 0.5]$. The collective evolution of trajectories is shown with red and green reach sets in Figure 4.5.

It is clear from the comparison of the shaded regions in Figure 4.5 that the final reachable set obtained with $N_0 = 0.6$ indeed contains both steady states.

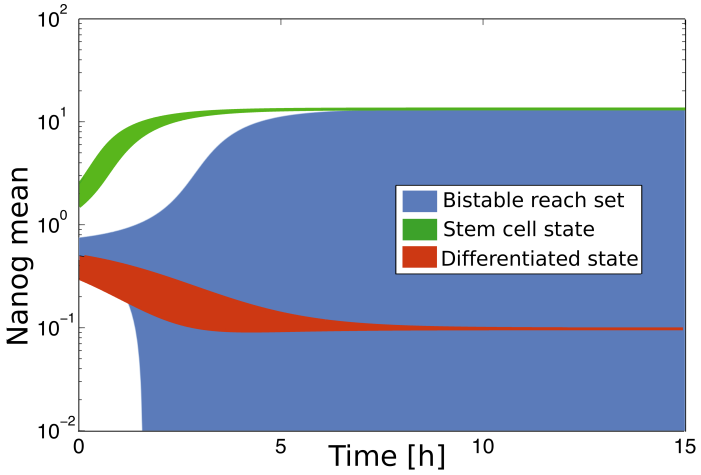


Figure 4.5: Investigating bistability via reachable sets of a stem cell differentiation model. Shaded regions show a conservative estimate of the achievable mean Nanog levels over time. The dark blue region is an over-approximation of the reachable set of the bistable system, when the analysis is started from an initial set of bistability, i.e. a set enclosing states leading to both fixed points. $N_0 = 0.6 \pm 0.15$. Red and green shades show reachable sets computed from 'monostable' initial sets, from which all trajectories converge to the differentiation fixed point (low Nanog level, in red) or to the stem cell fixed point (high Nanog level, in green). Initial Nanog values are 0.4 ± 0.1 and 2 ± 0.5, respectively.

Furthermore, the upper limit of the reach set is not only conservative, but also tight. Therefore the only crude over-approximation is due to symmetry and the high-Nanog state is well outlined even in this analysis. In addition, obtaining negative reachable sets for variables that should not assume negative values can be used as an indicator of discontinuous behaviour within the analysis. After such an observation, restricting the set of initial values – similarly to how the second and third reachable sets are obtained in Fig. 4.5 – can shed light on the underlying dynamics.

4.5 Discussion

In this work we have introduced a method to compute the states that a stochastic biochemical network can access under a control input signal. The method is particularly applicable to determine reachable mean-variance values of the investigated species, and hence estimate noise levels of a system. Our aim is to provide a computationally efficient tool that could be used in the search for the best modelling description or regulatory design of stochastic networks.

Here we used the LNA to extract deterministic equations of the mean and covariance matrix that served as a basis for our reachable set computation. An identical technique can be applied for equations obtained via the MEA algorithm [15]. Both can be accessed on github as parts of our Python package, MEANS [16]. Then we presented a general algorithm for linear and nonlinear reachability analysis based on zonotopes, together with formulae for handling relevant input types: addition of molecules and control of a reaction rate. All steps are implemented in Matlab to semi-automatically generate the reachable sets for a problem defined in terms of a transition function, input matrix and bounds, and parameter uncertainty. The algorithm can be executed on any given set of ODEs describing some characteristics of the system; hence it is applicable regardless of the approximation method used for the generation of equations and also to deterministic biological models.

We demonstrate the method on three schematic models of biological macromolecules: a controlled gene expression system; a cascade of modifying molecules, e.g. transcription factors; and a bistable model of stem cell differentiation. In all examples the approximation is conservative, hence trajectories randomly generated from a set of admissible signals and initial values are confined within the set our method predicts. In linear cases, if we have precise knowledge of the parameter values and input bounds, the predicted reachable set will be exact as well, i.e. each point in it can be actually accessed by an appropriate input sequence. This can-

not be ensured for nonlinear systems (or uncertain rate values) as the conservative estimates used for the nonlinearity (or reaction rate effect) are obtained from approximations the system might not take. However, our estimation reflects the general characteristics of the system – e.g. if some variables of a generally nonlinear system have linear equations – and hence unnecessary over-approximations are avoided to provide a tight estimate. This is crucial in the evaluation of various control designs on the basis of mean-variance values reachable through them.

Our method is based on a zonotope representation of reachable, initial and parameter sets. Set operations can be efficiently computed using this representation method, which allows us to apply our reachability analysis to many-variable, complicated biological networks. Since zonotopes can be stored as a set of column vectors summarised in a matrix, all above computational steps are typically executed very efficiently (in a few seconds), especially in languages dedicated for matrix-vector operations, such as Matlab and Python. In addition, in the case of linear systems with a fixed time-step, matrix functions needed for the computation of reachable sets (i.e. $e^{A\tau}$) can be evaluated in advance to offer additional speed-up. On the other hand, a limitation of the algorithm is space: the number of generator vectors in the reachable set increases in every iteration, and for a long time horizon or small time step the algorithm can become expensive in terms of required memory space. Furthermore, the number of zonotopes added in each step is proportional to the system size. In smaller systems, such as our examples, this effect is still negligible, but for high-dimensional cases (from about 20-variable systems) the increase is more significant. For such cases, one can turn to the zonotope-reduction technique presented by Girard [96] to limit the size of zonotopes to an adjustable value.

Note also that zonotopes, and hence the sets in our analysis, are by definition convex and centrally symmetric. Therefore systems where the actual reachable set is concave or consists of multiple sets — especially networks with bi- and multistability — will be largely over-approximated. In our last example, we have tested

the algorithm on a bistable model of stem cell differentiation. As the two expression level states in such cases are typically of different magnitudes, our method is forced to enclose a high proportion of negative values besides also incorporating a non-accessible region between the two states. However, we also found that this drop in approximation quality is a good indicator of the system entering a bistable regime of initial conditions/input signal. A similar over-approximation case can arise for imprecise parameter values of the transition matrix: although admissible parameter sets are symmetric sets given as a nominal value with error bounds, the influence of parameters is usually not symmetric on the reachable set. Therefore zonotopes lead to a rough approximation when some rates show high uncertainty. Just like in the case of multi-stability, a preliminary analysis using our general method can reveal these issues, and point one towards a more exhaustive study using a series of reachable sets for fixed values (or small intervals) of initial values or rate parameters to cover the range of interest and extrapolate for tighter approximation.

As often the case in the analysis of biochemical systems, incomplete information about the model structure and crucial constants of the system makes any detailed and informative analysis impossible. Our method is not applicable for the exploration of differences in model structures; this problem can be treated implicitly by either doing independent analysis of all candidate models, or by choosing rate parameter uncertainty such that zero (i.e. no interaction) is amongst the admissible values. The latter is likely to give rise to an over-generalised approximation, as described above; the former approach is only advisable for a small number of possible models, like in our second example. Furthermore, limited measurements and high levels of noise can also influence the method's power for validation, as Figure 4.4(c) demonstrates: in such cases flexible models might be favoured even over the true model. Therefore we advise the use of other tools, such as Topological Sensitivity Analysis [133] coupled with our method for a more thorough investigation of model space.

Five

Potential landscapes of stochastic cellular decision making

We have so far presented methods for simulation or evaluation of nonlinear biochemical systems. However, these methods can give only limited insight into systems with multistability, such as those describing development. In this chapter we explore stem cell differentiation on a potential landscape of multiple stable states.

5.1 Introduction

In his book in 1957, Conrad Waddington presented the metaphor for cell development in which cells are captured as marbles rolling through a landscape of hills and valleys aiming towards a local point of minimum elevation [80]. An individual ball starts its journey in a valley at the back of the landscape and as it progresses forward (the caption of time in the original formalism) and automatically downward, it might face branching points along the path, representing the series of binary fate choices made by a developing cell. Every point a ball travels through represents a physical condition of the cell, for example a specific level of a measured gene product or protein. Although the number of possible states is theoretically infinite, the

number of phenotypes observed in actual cells are often very limited. In this view, the final basins of low elevation where a high proportion of cells ends up correspond to these experimentally observable cell types.

The landscape notation has been widely used as a qualitative way of understanding and illustrating the dynamics driving development ever since [134–136]. Moreover, even though Waddington did not intend his landscape to be any more than a visualisation tool, landscapes can be quantitatively realised through exploration of the behaviour of the underlying network in terms of a probabilistic landscape framework. Recent studies have proposed a "pseudo-potential" to quantify the epigenetic landscape for a cell fate regulatory network, in which the elevation of the surface is derived from the inverse of the probability of observing a particular state in phase space [81, 137–141]. States that have the highest probability locally will have lower potential and hence will act as the valley-bottoms on the landscape, surrounded by a basin of attraction which in this picture would correspond to cells with slightly different states but exhibiting the same phenotype.

Even though the landscape depiction might be adopted to many decision making processes over the lifetime of a cell, its major uses are still in describing development and stem cell differentiation processes [142, 143]. In this context the final attractors of the landscape are differentiated cell types with well-defined patterns of robust gene expression, and differentiation occurs through transitions along the surface, while some cells might reside or return to the original basin of the pluripotent phenotype.

The potential landscape of differentiation can be derived in two major ways depending on the type of information available. In case a great number of measurements showing the expression states of cells is available, an empirical landscape might be obtained by using the observed probability of each state. Furthermore, the evolution of the landscape and its major features can be derived in a model-free manner if the distribution is measurable at several time-points [144]. On the other hand, when the system is characterised by a biochemical network model, the potential is

computed based on the model and its possible behaviours. Such model-based probability distribution or landscape might be compared to experimental observations, but can also be used for qualitative investigation such as stability or bifurcation analysis [131]. Here we adopt a model-based approach for mapping the potential landscape of cellular decision making.

The models proposed to shed light on the fundamental principles of differentiation are extremely versatile [145]. However, there are some key features most mathematical depictions share, such as a core network of genes/proteins associated with the maintenance and loss of pluripotency. It has been shown that Oct4 – acting in close cooperation with Sox2 – is essential for establishing stem cell state, but can promote differentiation when expressed at raised level [146, 147]. The above complex is supported by a third player, Nanog, which unlike the Oct4-Sox2 duo, is consistently associated with the maintenance of stable pluripotent state [148, 149]. These three all act as transcription factors while being target genes themselves and are connected through several positive feedback loops [150, 151]. Furthermore, they are also connected to a plethora of downstream and upstream regulatory genes and factors with similar network characteristics [152–154]. Many of these can be categorised as *stem cell* or *differentiation* genes/factors depending on their relationship with the core circuit and whether they are associated with raised or lowered levels upon differentiation.

The majority of models of this Nanog-Oct4-Sox2 core structure use Hill-type kinetics to investigate how the complex feedback loops between the three players lead to the experimentally observed potential for assuming substantially different cellular states under similar conditions [155, 156]. Numerous studies focused on the heterogeneity and (inducible) reversibility observed in Nanog levels, and consequently in phenotypic cell fate, considering stochastic effects in modelling [132, 157, 158]. Several different mechanisms have been proposed to mechanistically explain the bimodality of measured Nanog distributions, such as extrinsic and transcriptional

noise and cyclic behaviour [132,159–162]. Although models hypothesise different underlying mechanisms, they all consistently describe a stable but dynamic bistability in the system.

In this work, we derive potential landscapes of differentiation based on two schematic models of the network underlying pluripotency. First, we consider two methods for the computation of the probability distribution of the system and the landscape, given an arbitrary model described in stochastic terms. We use either an ensemble of exact SSAs or – to be used in cases when former is not feasible – a local approximation through a moment-based estimate of the dynamics. In addition, simulated paths in the phase space of pluripotency and differentiation are studied to provide information on the fluxes along the landscape.

The technique is first applied for a minimalistic model that captures dynamical interactions between transcription factors and target markers; and achieves bistability through external excitation. We show how essential features of stem cell regulatory networks can be reproduced and examined with this framework. Next, we consider a stem cell model from the literature with intrinsically arising fluctuations [161]. We obtain the potential landscape of the system to confirm bistability and spontaneous transitions between states. Since the level of heterogeneity in the population depends on externally set quantities (such as reaction rates and constants), we define summary characteristics to quantify the underlying structure and map changes induced by these quantities. Finally, we study the relationship between steady-state structural information and dynamical paths explored during the transition between stable phenotypes, and identify a distinguished transition state that seems essential in differentiation [163, 164].

5.2 Landscape computation

Let us consider a stochastic biochemical system consisting of N interacting species and R reactions, as described in Eq. (2.3). The system is characterised by the probability of each reaction (typically dependent on the copy number of some species) happening and the change in abundance each species experiences would the reaction undergo. Given these quantities collected in the reaction propensity vector, $F(\mathbf{X}) = [a_1(\mathbf{X}), \ldots, a_R(\mathbf{X})]$, and the stoichiometry matrix, $S = [\mathbf{s}_1, \ldots, \mathbf{s}_R]$, we want to describe the probability of all possible combinations of molecular abundance levels, namely,

$$P\left(\mathbf{X}(t) = \begin{bmatrix} \psi_1 \\ \vdots \\ \psi_N \end{bmatrix} \middle| \, F(\mathbf{X}), S\right), \quad \forall \psi_i \in \mathbb{N}. \tag{5.1}$$

In our investigation, we obtain the probability distribution for either a specific timepoint (denoted $\mathcal{P}_t$) as in the above equation, or for the steady-state of the system ($\mathcal{P}_{ss}$). Note that $\mathcal{P}_t$ is conditioned on the model structure and parameters provided (S and F), but not on the initial conditions, $\mathbf{X}(0)$, as we seek to explore the *global* distribution.

The landscape is then obtained from $\mathcal{P}_t$ as its negative logarithm [131, 138, 139],

$$\mathcal{U}(\boldsymbol{\psi}(t)) = -\ln(\mathcal{P}_t), \tag{5.2}$$

providing a dimensionless quantification of the quasi-potential.

5.2.1 Exact stochastic computation

The most straightforward and precise way to obtain the above probability distribution is through direct simulation of the system. Using Gillespie's SSA, we can observe trajectories of individual cells, the collection of which depicts the probability density as more and more exact simulations are considered. Such an approach is

limited by how well the model depicts the system, but gives an exact account of the potential landscape within these limits.

Let $\mathcal{X}^i_{\mathbf{x}_0}$ denote the ith realisation started from $\mathbf{x}_0$ initial state, obtained using the SSA described previously. At any given time-point, t, we obtain the probability distribution from n individual samples; either as an N-dimensional histogram or using a *kernel density estimator* [165, 166],

$$\mathcal{P}_t = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}\left(\boldsymbol{\psi} - \mathcal{X}(t)\right),$$

with a multivariate normal kernel, $\mathcal{K}_{\mathbf{H}} \sim \mathcal{N}(\mathbf{0}, \mathbf{H})$. Even though $\mathbf{H}$ can be chosen to be any arbitrary symmetric, positive definite matrix, in practice, it is usually a diagonal matrix, with either identical or varying values in its nonzero elements. To ease computation and avoid introducing any bias, we choose $\mathbf{H}$ as a multiple of the identity matrix.

Due to the stochastic nature of the system, the collection of trajectories outlines not only the stable states and most likely paths, but also the basins and valleys surrounding them that one would also observe in real measurements. In the collection $\mathcal{X}_{\mathbf{x}_0}$, more than one stable state/path can be present. However, some regions of the state space surrounded by a low probability region – a mountain range in the landscape metaphor – might be observable only from a very specific set of starting positions. Therefore we obtain trajectories from a range of initial conditions, both through systematically scanning a grid and randomising the starting state.

The true biological processes underlying cellular decisions involve a complex network of molecules and chemical reactions with substantially different availability or speed that makes modelling very demanding computationally. As our main focus is on the phenotype level captured by specific expression patterns, we apply simplifications to avoid the complexity arising from multi-scale models of DNA, mRNA and protein interactions. We rely on the assumption that, due to the complexity of the underlying chemical processes, protein translation is much slower than DNA-binding

and transcription [167, 168]. Therefore we assume that processes modifying DNA occupancies and transcription rates are at quasi-equilibrium, and the expression dynamics of a given molecule can be described with a single nonlinear (Hill-type) term depending on its known regulatory partners, for example transcription factors. In particular, the models only include species of the same type – either mRNA or protein molecules, depending on application – and a few (usually one) nonlinear creation and a linear degradation reaction for each species.

5.2.2 Approximation through moment equations

As the above method requires a huge collection of SSA realisations, it might not be feasible or practical in some applications. In order to estimate Eq. (5.1) in a computationally efficient way, we use a deterministic approximation of the original stochastic system. Let us assume that even though the system might have several stable states, each high-probability state (i.e. the attractors of the corresponding potential field) can be well approximated by a known distribution and their weighted sum describes the probability over the entire state space. Then a few low-order moments (typically the mean and the variance) provide a good characterisation of the distribution; and we can use moment-based approximation techniques to obtain a set of ODEs describing the evolution of these moments. In the following the LNA [30] and a normal approximation is used; however, a similar approach can be derived for other distributions, such as log-normal or gamma, with the help of the MEA algorithm [15, 36].

First, we derive a coupled, nonlinear ODE system to obtain the first two moments of the stochastic system in terms of stoichiometry matrix and propensities, as described in equations (2.12),

$$\frac{d}{dt}\begin{bmatrix}\mathbb{E}(t)\\ \Sigma^v(t)\end{bmatrix} = \begin{bmatrix}S\cdot F(\mathbb{E}(t))\\ (SD\Sigma)^v + (\Sigma(SD)^T)^v + (S\cdot \mathrm{diag}\{F(\mathbb{E}(t))\}\cdot S^T)^v + (\eta I)^v\end{bmatrix}, \quad (5.3)$$

where $\mathbb{E}(t)$ and $\Sigma(t)$ stand for the vector of mean values and the covariance matrix, respectively. Σ^v represents the vectorised form of the symmetric matrix, as introduced in the previous chapter. The last term, ηI, is a diagonal matrix representing an additional diffusion term that allows us to include some level of external noise in the system. However, for the scope of this study we focus on cases with only intrinsic noise, when $\eta = 0$. $\mathcal{P}_t$ is defined as an N-variate normal distribution with the above computed parameters,

$$\mathcal{P}_t = \frac{1}{\sqrt{(2\pi)^N |\Sigma(t)|}} \exp\left(-\frac{1}{2}(\boldsymbol{\psi} - \mathbb{E}(t))'\Sigma(t)^{-1}(\boldsymbol{\psi} - \mathbb{E}(t))\right). \tag{5.4}$$

The ODE set to be solved has dimension $N(N+3)/2$, so in case of large multidimensional models a further approximation might be needed to ease computational simulations. By assuming independence of variables, we can split the probability into a product as in [142], $P(\mathbf{X})_t = P(X_1)_t \times P(X_2)_t \times \cdots \times P(X_N)_t$, and thus reduce the system to a $2N$-dimensional one,

$$\frac{d}{dt}\begin{bmatrix}\mathbb{E}(t)\\ \boldsymbol{\sigma}(t)\end{bmatrix} = \begin{bmatrix}S \cdot F(\mathbb{E}(t))\\ SD\boldsymbol{\sigma} + \boldsymbol{\sigma}(SD)^T + S \cdot F(\mathbb{E}(t)) \cdot S^T\end{bmatrix},$$

where $\boldsymbol{\sigma}(t)$ is an array formed of the variances of $X_1, \ldots, X_N$. Besides significant reduction in the dimensionality of the problem, the mean and variance for each species can be evaluated independently, in parallel with the others [137, 142]. However, to ensure no further information is lost, we advise the use of Eq. (5.3) as long as computations are feasible.

In systems with multistability, the probability obtained in Eq. (5.4) corresponds to one basin of attraction in the system (or a snapshot in time on the way to a particular basin of attraction), and there are several other distributions contributing to the landscape. As our approximation method is deterministic, it captures one of these components at a time, depending on the initial conditions used to simulate Eq. (5.3). The total probability distribution at a given time-point is computed as the weighted sum of output distributions obtained from simulations on a fine grid of

initial values, where weight are derived according to the representation percentage of each distribution in the total number of simulations.

The steady-state distribution, $\mathcal{P}_{ss}$, can be obtained analytically by identifying all-positive solutions of the equation system

$$\frac{d}{dt}\begin{bmatrix}\mathbb{E}\\ \Sigma^v\end{bmatrix} = \mathbf{0},$$

however, when the reaction propensities are described by a rational function of many terms, this approach might prove inefficient. Instead, we base evaluation of the steady state on the relation $\mathcal{P}_{ss} \approx \mathcal{P}_{ss+\tau}$. We assume the steady state is reached when

$$|\mathbb{E}_i(t+\tau) - \mathbb{E}_i(t)| < \epsilon \quad \text{and} \quad |\Sigma_i(t+\tau) - \Sigma_i(t)| < \epsilon \quad \forall i,$$

so the change in any variable over time is negligible. Again, a wide range of initial conditions is considered to capture multistability, and obtained values are further checked to test whether they satisfy the above equation.

5.3 Results

5.3.1 Schematic transcription factor model

In our first analysis, we aim to capture some of the typical behaviour observed in decision making networks with a system as minimalistic as possible. Transcription factors have a major role in such scenarios – therefore our proposed basic system is formed of a transcription factor (x_1) and its downstream component (x_2). The expression level of the second species corresponds to a specific phenotype, and provides an observable output of the system. For this example, let x_2 denote the abundance of a *stem cell marker*: then high x_2 means the cell is in pluripotent state, and low x_2 indicates a differentiated cell type.

The central interaction is between the two modelled molecules: here we assume that the transcription factor *inhibits* transcription and consequently production of the output marker. The Hill-type formalism is used to simplify the model; the system has four reactions, a production and a degradation reaction for each species, respectively. All other molecules involved in these processes are implicitly captured in the propensities of the reactions. Furthermore, we include an external source that has the ability to introduce excitation to the system [155, 159]. This is represented by an additional zeroth-order reaction producing x_1 with rate k_S that depicts an external input (e.g. differentiation signal) altering the expression of the transcription factor. The resulting stochastic system is fully specified by the reaction propensities and the corresponding stoichiometry matrix,

$$a_1 = k_1; \quad a_2 = k_2 x_1; \quad a_3 = \frac{k_3}{1 + x_1/k_4}; \quad a_4 = k_2 x_2; \quad a_5 = k_S,$$

$$S = \begin{bmatrix} 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 \end{bmatrix}.$$

This system is monostable, but the choice of parameters and external stimulus defines whether the only fixed point is a stem cell or differentiated state. We initialise

the system with stem cell conditions ($\mathbf{x_0} = [1\ \ 8]^T$) and parameter values that, in the absence of differentiation signal, keep the stem cell state ($\mathbf{k} = [2, 2, 40, 1],\ k_S = 0$). Figure 5.1(a) shows two examples of the stem cell marker abundance over time (obtained using SSA) with and without a $k_S = 10$ signal introduced at time $t = 50$.

First, we create a combined landscape of the two phenotypes by computing the probability distribution for both states (corresponding to signal on/off) independently, and then taking their sum (Figure 5.1(b)). As the system is simple, both exact and approximated simulation is feasible, and they lead to results with identical summary statistics. Note that, unlike the landscapes presented in [81] or [142], the two states shown here would not appear at the same time intrinsically. However, in a scenario when not all cells receive the same signal (due to spatial heterogeneities or noise in an upstream pathway) both valleys can get populated simultaneously. We explore this phenomenon in Appendix B.

Next, we investigate the details of transition between stem cell and differentiated state. We first simulate the system until it settles in the stem cell steady state (until $t = 100$), then introduce a sudden change in signal $k_S : 0 \rightarrow 10$ to push cells towards differentiation and take frequent observations. LNA simulations are used to map the mean path taken by cells between the states (Figure 5.1(c)). We also generate 10,000 independent exact trajectories with SSA for each time-point observed along the transition, to depict biological observations that often require a destructive technique. The probability distribution is reconstructed from the LNA-derived moments and the SSA-based cell population – the two methods yield almost identical outputs apart from some noise in the exact simulations inevitable at such population size. Similarly, the mean path computed by averaging individual cells at each time-point corresponds well with the approximated mean path in Figure 5.1(c). Since the Hill-type kinetics effectively introduce a delay into the action of the transcription factor, the path in phase space shows a separation between the two variables; a quick change in x_1 levels is followed by a slower decrease of x_2. Therefore

our system qualitatively reproduces the dynamics we expect to see in more accurate decision making models.

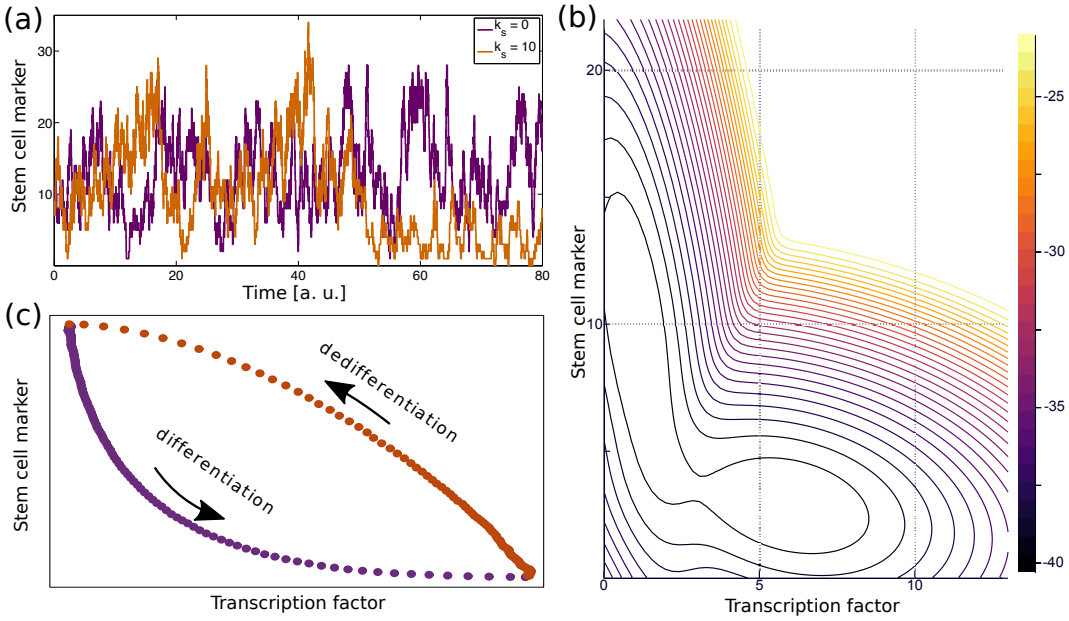


Figure 5.1: Potential landscape and trajectories of the transcription factor model. **(a)** Two example simulations of the model with (orange line) or without (purple line) external signal at time 50. **(b)** Potential landscape obtained from the combined probability distribution of the two steady states. **(c)** Mean path followed by cells during differentiation (purple curve, direction indicated by arrow) and the reverse process, dedifferentiation (orange curve). Each point denotes a time-point between $t = 100$ and $t = 105$ with time-step of 0.03. $\mathbf{k} = [2, 2, 40, 1]$ in all panels, k_S is set to 10 for switched-on signal and zero otherwise.

Furthermore, as the modelled interactions do not impose a strong directionality on the system, we can also investigate the reverse transition, *dedifferentiation* from differentiated to stem cell state. With the above parameter values, removal of the external signal ($k_S : 10 \rightarrow 0$) allows the system to escape back to a high stem cell marker phenotype. This behaviour seemingly contradicts the assumption that there is a point of no return after which cells commit to a chosen fate. However, a reversal of the process can be achieved not only in *in silico* but in

in vitro experiments through appropriate perturbations that help cells gain – so-called induced – pluripotency [154, 169]. We are interested whether our model can recreate a similar phenomenon. We follow the same simulation strategy as in the 'forward' case: first obtain the differentiated steady-state, then record transition following a drop in external input. We find that, as expected, there is again a delay in stem cell marker response compared to change in transition factor levels. An important – and in hindsight mathematically obvious – consequence is that the forward and backward process move through a different path in phase space (purple and orange curves in figure 5.1(c)), similarly to how tumour-evolution and differentiation paths have been suggested to divert from their reverse transition path [139, 142]. We also follow the transition of the entire probability distribution from low to high stem cell marker state, which aligns with the mean observations, as expected. (Animation of the transition between the two states is available at `https://github.com/elakatos/Thesis/blob/master/TF_SSA_transition.gif`.) This shows that the external signal substantially alters the landscape features, such as changing the attractor structure and major valleys of migration [81, 138].

In summary, despite the simplicity of the model, we are able to observe some crucial – and we believe, widely shared – characteristics of cellular decision making processes, such as bistability, different timescales of cellular response, dynamic change of landscape topology and a separation of back and forward transition paths.

5.3.2 Stem cell differentiation model

In our next study, we continue our analysis with a specific stem cell differentiation model that is still a simplified representation of the real biological system, but includes more details of the known interactions behind the process [161]. The model focuses on four key players: Nanog (N), the complex Oct4-Sox2 (O), Fgf4 (F) and an otherwise unspecified differentiation gene that we choose to be represented by Gata6 (G). A schematic representation of their network is shown in Figure 5.2(a); all arrows stand for a non-linear interaction that accounts for implicitly modelled species and processes, like the formation of the Oct4-Sox2 complex or Nanog dimerisation. In addition, two typical cell media used to promote self-renewal and stem cell state, LIF and 2i/3i are also included in the model with their most dominant mechanism of action on the network. These are treated as single – though controllable – constant values, L and I_3. We apply the quasi-equilibrium assumption and treat all modifying interactions as factors in the same birth process, with their associated rates practically weighting their positive or negative effect on production. A first order degradation reaction, assuming the same rate, k_{deg}, is added for each species, therefore in total the system is described by eight reactions,

$$\begin{aligned}
a_1 &= \frac{k_0 O(k_1 + k_2 N^2 + k_0 O + k_3 L)}{1 + k_0 O(k_2 N^2 + k_0 O + k_3 L + k_4 F^2) + k_5 O G^2}, \\
a_2 &= \frac{k_6 + k_7 O}{1 + k_7 O + k_8 G^2}, \\
a_3 &= \frac{k_9 + k_{10} O}{1 + k_{10} O + k_{11} I_3}, \\
a_4 &= \frac{k_{12} + k_{13} G^2 + k_{15} O}{1 + k_{13} G^2 + k_{14} N^2 + k_{15} O}, \\
a_5 &= k_{deg} N, \\
a_6 &= k_{deg} O, \\
a_7 &= k_{deg} F, \\
a_8 &= k_{deg} G.
\end{aligned} \tag{5.5}$$

Thanks to the compact form of propensities, the system can still be efficiently simulated using a collection of SSAs. The parameter values used are taken from [161], with one exception: the baseline production of Gata6, k_{12} (b_0 in [161]) is increased to 1 to speed up differentiation events. Hence the rates used throughout this analysis are,

$$\mathbf{k} = [0.005, 0.01, 0.4, 1, 0.1, 0.00135, 0.01, 0.01, 1, 1, 0.01, 1, 5, 1, 0.005, 1, 1, 0, I_3],$$

with the penultimate value representing LIF amount (kept at 0), and I_3 assuming a range of values depending on the *in silico* experiment. As the model is schematic and not based on exact rates, time is in arbitrary units and rates have units of 1/a.u. ($k_0, k_3, k_7, k_{10}, k_{11}, k_{15}, k_{deg}$), molecules/a.u. ($k_1, k_6, k_9, k_{12}$) or 1/molecules $\times$ a.u.($k_2, k_4, k_5, k_8, k_{13}, k_{14}$).

Unlike the previous example, this model shows stochastic bistability. Cells initiated with the same original expression level might avoid or undergo differentiation. Differentiation is modelled as an irreversible process, therefore theoretically on the infinite time limit the whole population changes into a more specialised cell type. However, in biologically relevant time-scales, a significant proportion of cells stays self-renewing. Furthermore, this model does not include cell duplication events that would ensure the persistence of a stem cell population. Instead, as a simplification, we focus on a fixed time period ($t = 3000$ in arbitrary time-units) when mapping landscapes.

We use the expression of Nanog and Gata6 to indicate stem cell and differentiated phenotypes. Since there is a strong negative correlation between these two molecules, a state when both are high or low simultaneously is extremely unlikely. Formally, we categorise a cell as differentiated if $N < 10$ and $G > 50$. Figure 5.2(b) shows example simulations with $I_3 = 2$ and intermediate initial expression levels, under which conditions many cells stay pluripotent (i), but some stochastically shift into differentiated state (ii).

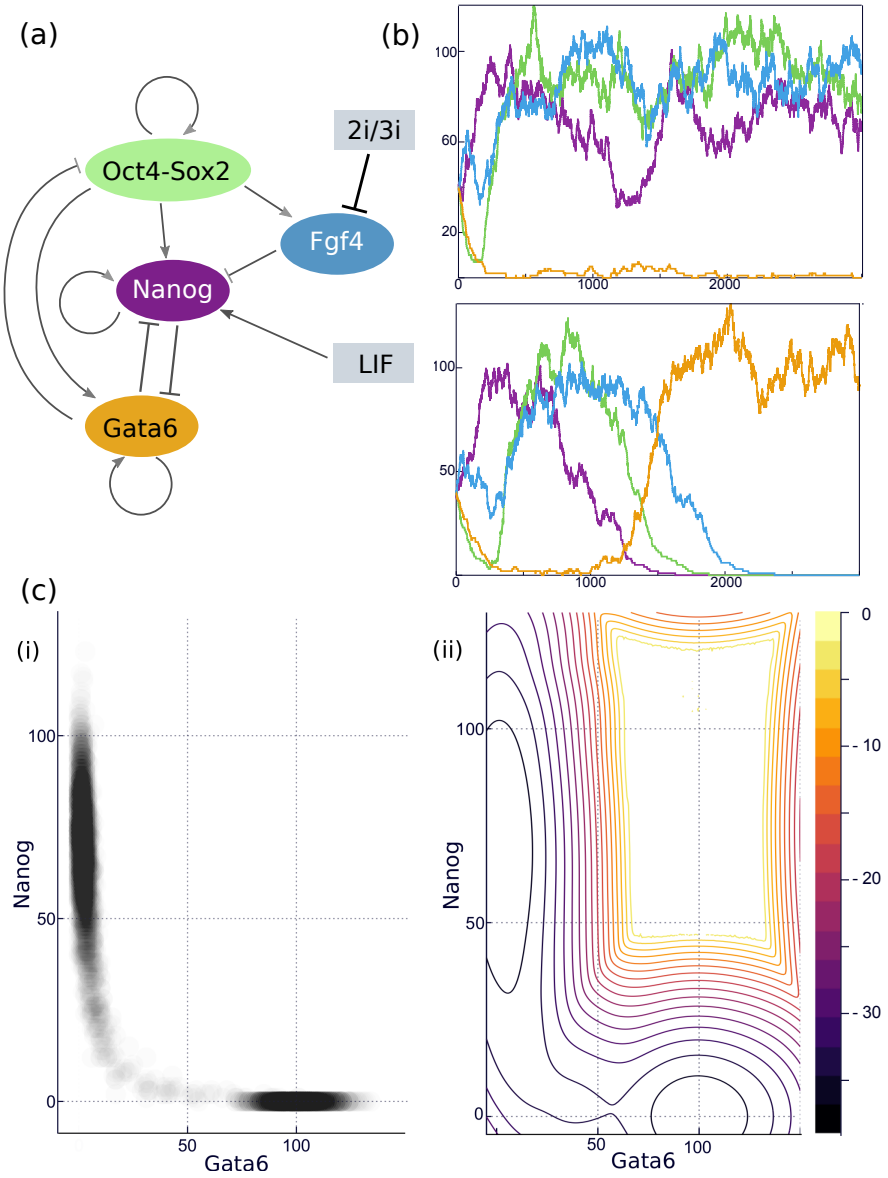


Figure 5.2: Potential landscape of the stem cell differentiation model. **(a)** Schematics of the stem cell differentiation model. **(b)** Simulated example of a cell that stays pluripotent (upper axes) and a cell undergoing differentiation (lower axes). The abundance of each species is indicated by a line coloured according to the filled ovals in (a). **(c)** Probability distribution and landscape of the system with $I_3 = 2$. **(i)** Scatter plot of the whole population on the Gata6-Nanog plane. **(ii)** The landscape computed as the negative logarithm of the probability distribution.

We reconstruct the whole probability distribution for several different concentrations of the stem cell medium and explore changes in the landscape structure due to this external signal. The landscape computed at a moderate level of 2i/3i ($I_3 = 2$) is shown on panel (c) in Figure 5.2. The two major populations are best balanced at this concentration; lower I_3 values push more cells to differentiate, higher I_3 increases the number of the maintained pluripotent cells. The topology of the landscape is the same nevertheless – 2i/3i concentration only affects the relative potential of the two phenotypic basins.

5.3.3 Time-scales of differentiation

Due to the strong stochasticity in the system, some cells might differentiate even when their conditions would intuitively keep them in a stem cell state. The magnitude of this random process holds valuable information on the dynamics of the transition towards the differentiated phenotype. Therefore, we calculate summary statistics in phase space to characterise dynamics, in particular the speed of transition, on the landscape in a given condition. First, we compute the average time differentiation takes for cells starting from a specific point of the landscape, i.e. the *mean first passage time*. We fix the initial values of Fgf4 and Oct4-Sox2 to 40 to avoid priming cells toward one fate or another. We later repeat the analysis with different F and O values and see no difference in the obtained statistics. Starting values of Nanog and Gata6 are explored on an 11×11 grid from 0 to 100 and a population of 40 trajectories is simulated from each point. Differentiation time is defined as

$$t_{\text{diff}} = \min\{t \mid t < 3000,\ N(t) < 10,\ G(t) > 50\}.$$

We choose these criteria as they define a 'point of no return': no trajectories were observed where stem cell state is regained after such conditions are met. Although stochastic effects might push a cell outside this region, it will shortly converge back

to the differentiated phenotype. If no time-point fulfils the conditions, we label the cell as undifferentiated. Then the average, standard deviation and coefficient of variation of t_{diff} is computed from the population at each initial value, as well as the number of cells that stayed pluripotent. Figure 5.3(a) presents the mean time required for differentiation as a heat-map in the Gata6-Nanog phase space. We also map the average, minimum and maximum values of N at time 3000. In all cases there is a clear region that has almost zero transition time, as all trajectories start from close proximity of the differentiation basin so that the time needed to converge there is negligible. The region surrounding the stem cell phenotype is more extended, with very similar summary statistics that show random instead of starting value-dependent inhomogeneity. The two regions are separated by a narrow band of initial conditions that show the biggest fractional variability. This suggests the separating region contains quickly differentiating and prolongedly pluripotent cells in almost equal proportion. Interestingly, a change in 2i/3i concentration has no effect on the boundary, but makes the separation of regions more profound (Figure 5.3, rows (i) and (ii)). With increased I_3, the number of cells not differentiated also increases. However, the dispersion of the cells that avoid differentiation is dominated by intrinsic stochasticity.

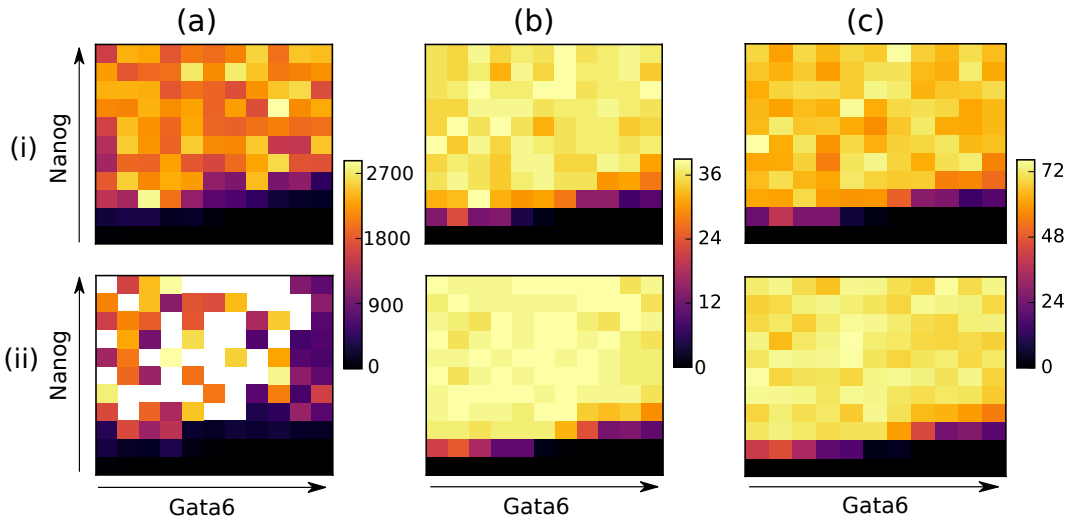


Figure 5.3: Statistics of differentiation dynamics depending on initial expression levels. Rows **(i)** and **(ii)** show heat-maps obtained with a 2i/3i concentration of 2 and 3, respectively. **(a)** Average time (in arbitrary units) required for cells to reach the differentiated state basin. Unfilled, white areas represent that no cells from that area differentiated within the simulation time limit. **(b)** Number of cells out of 40 that did not differentiate during the course of simulation. **(c)** Average abundance of Nanog at final time-point.

5.3.4 Dynamical transition between states

We have so far touched upon the dynamics of differentiation, but have not investigated the transition between the two states in detail. The potential landscape (Fig. 5.2(c)) outlines a narrow valley as the most likely candidate to connect them. To compare the final potential landscape with paths taken by cells, we devise the *in silico* equivalent of an experiment inducing differentiation. First, we 'cultivate' the cell population in a high 2i/3i medium by simulations with initial values from the $N \in [60, 100]$, $G \in [0, 16]$ region and $I_3 = 10$. A total of 1250 output cell readings is produced. Then we change the external conditions, $I_3 : \ 10 \rightarrow 0$, and simulate 20 paths from each point in the population. As demonstrated by randomly selected trajectories in Figure 5.4(a), the dynamical paths show good correspondence with the previously derived static landscape.

Most of the time along the path is spent in either of the two valleys: cells linger in the stem cell valley for varying amount of time, but as their Nanog level decreases below a critical threshold, they quickly transit to the differentiated cell valley. In fact, some trajectories do not exit the self-renewing state (such as cell 2 on panel (a)), accounting for the less populated but still prevailing stem cell population in Figure 5.2(c). We also reconstruct the shift of the whole probability distribution (available at `https://github.com/elakatos/Thesis/blob/master/SC_SSA_transition.gif`) that follows the same channel between basins of attraction. As figure 5.3(a) suggests, the path cells follow is independent of their exact starting state inside the stem cell basin. Comparing a collection of cells initiated from the same subregion of the basin with randomly selected cells demonstrates no relationship between path and starting proximity. We also test the system in a reversed differentiation scenario: cells that already reached the differentiation valley are put into high 2i/3i medium ($I_3 = 10$) and tracked. The trajectories confirm that once differentiation is achieved the cells are not allowed to regain pluripotency, as expected from an irreversible switch-like process (Fig. 5.4(b)).

The majority of the Gata6-Nanog phase space lies on a high-potential plateau that thus remains unexplored by the cells. Therefore, we have to artificially introduce conditions that allow the mapping of cell dynamics on the value range of this plateau. We first uniformly sample cells from an uncovered volume of four-dimensional space, for example $\{N, O, G\} \in [60, 100]$ and $F \in [10, 50]$. Then follow the same experimental setup as before and induce differentiation by setting I_3 to 0. Surprisingly, from this region, trajectories quickly converge into the Nanog-high valley and then follow the dynamics observed on a stem cell starting population. Figure 5.4(c) shows the first 500 time-units of five randomly sampled simulations. On the other hand, initial values sampled from another region of the high-potential plateau ($\{O, F, G\} \in [60, 100];\ N \in [20, 60]$) lead the majority of cells towards the differentiation basin. Nevertheless, there is a significant fraction of cells, solely with

initial Nanog abundance over 40, that go towards and linger in the stem cell state.

Interestingly, as illustrated in Figure 5.4(d), cells which go through differentiation do not follow a direct route. This finding does not contradict the landscape of Fig. 5.2(c) as dynamic fluxes are not captured in the potential. They first visit a *transition state*, characterised by simultaneously low Gata6 and Nanog expression, then follow the connecting valley to a differentiated phenotype. This phenomenon is consistent for a wide range of initial conditions, which suggests that any cell – unless it has already differentiated – has to go through the transition state. Cells that reside in the transition state, on the other hand, have the ability to progress towards either states, as seen in Fig. 5.4(e). This is in agreement with other models reporting a distinguished transition state along the way of cell specialisation that has the potential for any cell fate [81, 163, 164].

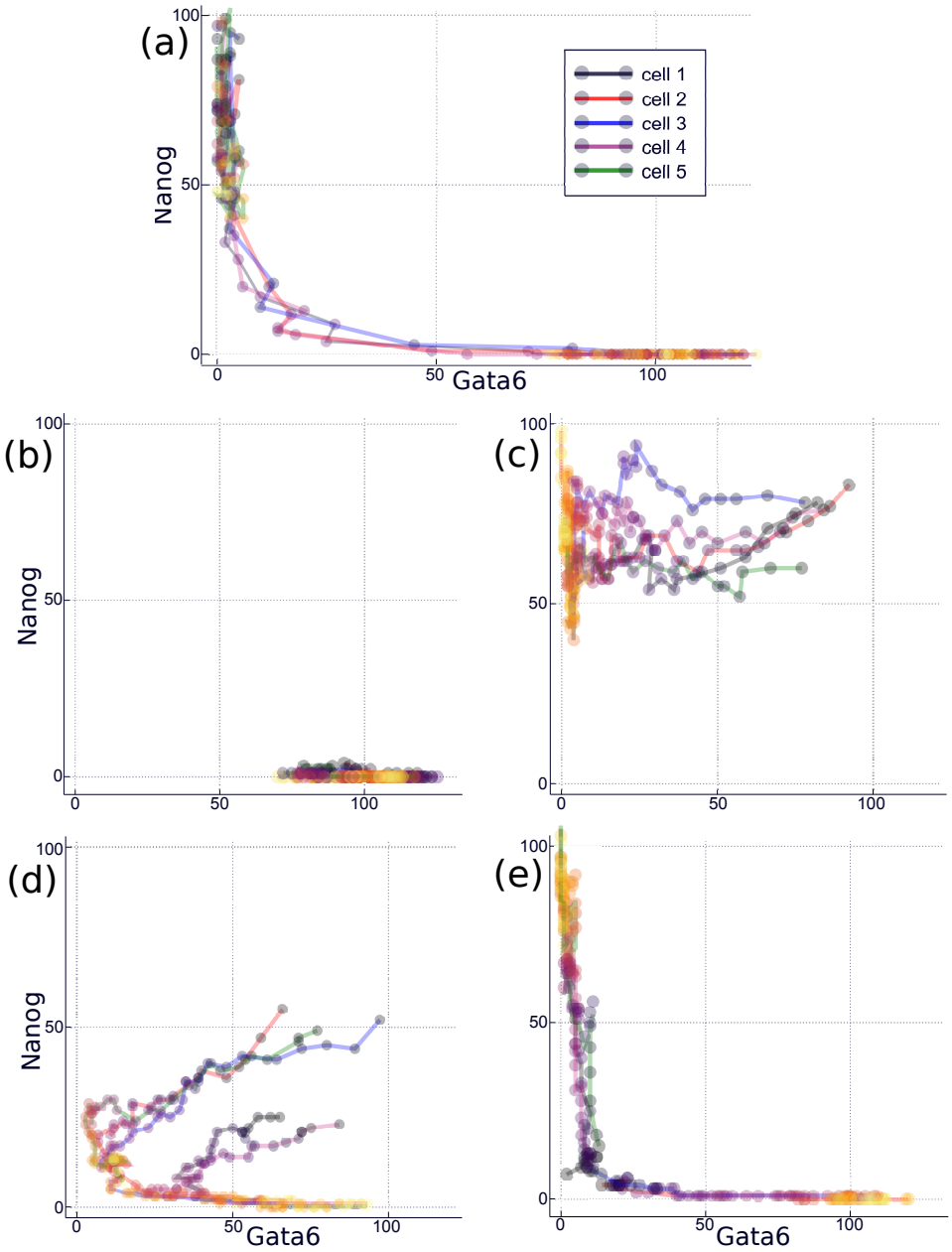


Figure 5.4: Transition paths during differentiation. **(a)** Trajectories of five randomly selected stem cells after removal of 2i/3i medium, plotted on the Gata6-Nanog phase space. **(b)** Cells started from the differentiated state and put into $I_3 = 10$. **(c-d)** Cells started from the unexplored region of high Gata6 and high Nanog (c) or low Nanog (d) levels. **(e)** Trajectories started from the transition state under $I_3 = 5$. Different line colours denote different cells, as indicated in legend. Marker colours represent time-wise order in a trajectory, with grey and yellow corresponding to earliest and latest observations, respectively.

5.4 Discussion

In this chapter we have introduced a framework to visualise the influencing factors and dynamics underlying stem cell differentiation. We obtained quasi–potential landscapes, a quantitative implementation of Waddington's epigenetic landscape metaphor [80], to explore the state topography of multistable stochastic systems. Landscapes offer a powerful tool for exploring the global properties of phenotype-level cell states and the dynamic flows determining the transition between them. They combine the information carried by a huge ensemble of trajectories and are capable of encoding crucial dependencies between variables and parameters, especially if considered as dynamic entities [131]. As our knowledge of gene regulatory networks increases, the methods based on potential landscapes could become a routine part of their analysis to uncover phenotypic behaviours. Therefore we presented two methods for numerically approximating the probability distribution – and subsequently, the potential landscape – for an arbitrary model in multi-dimensional space associated with macromolecule expression levels.

We considered two models of differentiation that represent different mechanisms of achieving a bimodal distribution of a specified marker. Many behaviours observed in *in vitro* and *in silico* experiments can be reproduced even in a representation that is much simpler than the true dynamics or the established core network used for modelling differentiation. However, this model relies on extrinsic variability to drive bimodal behaviour and experimental studies suggest that in the underlying system intrinsic stochasticity has a dominant role [170]. In the second representation considered, we indeed have shown large heterogeneities can be produced solely by random fluctuations under identical conditions. We studied how changing an external constant, representing an experimentally controllable concentration, modifies landscape topography and the mean first passage time between attractor states. We also found that, despite the apparent vastness of available states, transitions

are associated with a small, restricted and ultimately unstable region. Similarly to previous studies establishing the notion of a transition state in differentiation, our simulations show that cells in this state are not yet primed for either fate and lead to a heterogeneous final distribution.

In order to approximate the potential landscape, we used the LNA to derive equations of the means and (co)variances of the system that allow us to reconstruct the probability distribution using a Gaussian kernel. Often, even without normality, the approximation can give a good account of the system's behaviour – however, in many biochemical systems the true distribution might be too far from a normal distribution. For such cases, our approximation can be based on the MEA instead to gain equations of an arbitrary number of moments. The moments then can be used in an ensemble manner to inform the reconstruction with a known multivariate distribution kernel [15] or a maximum-entropy based method [58].

It is important to keep in mind that the landscape is limited by the models and methods used in its creation. The phase space quantified through potentials is far from static and more and more evidence seems to suggest that a huge variety of factors can directly alter the landscape [171]. For example, it has been suggested that cells might regulate their susceptibility to environmental noise [134] – that would correspond to controllable heights of between-attractor barriers on the landscape. Spatial interaction between cells is also crucial in differentiation dynamics [172–174], which adds a further dimension to landscape topology. This means an additional layer has to be considered when devising a model: which entities and processes can influence the underlying structure and noise level.

A possible extension of the potential landscape framework to account for some of the above factors is the inclusion of extrinsic noise through rate parameters. In our study we only considered identical conditions or a deterministic external signal. However, small differences between cells – which in modelling practice would be formalised as a distribution of possible rate parameters instead of a single value –

cannot be avoided in a real scenario, and this extrinsic component has been shown to contribute significantly to the overall heterogeneity in certain cases [95]. The effect of extrinsic noise could be sampled by adding an additional layer of grid-like scanning or randomisation, as in the case of sampling the state space for initial values. We study the effect of this extension on the schematic transcription factor model in Appendix B. However, for more complex systems this approach can easily become computationally infeasible, therefore an approximation of the parameter distribution, for example through the Unscented Transform, is recommended [123,124,175]. In particular, a combination of exact and approximation methods (for example SSA simulations with the Unscented Transform) could offer a good compromise between information content and computational complexity.

Six

Analysis of p53 baseline regulation using model selection

The previous chapters have presented methods for exploring capacities of mechanistic models in biochemical networks. In this final research chapter we present an inference framework that combines stochastic modelling with single cell experiments to investigate regulation of protein expression.

6.1 Introduction

The regulation of gene expression and, specifically, the variation in protein abundance has received increasing attention recently. There are four main steps that contribute to protein abundance – transcription, mRNA degradation, translation and protein degradation [176]. The abundance of a specific protein is then determined by the relative rates of these steps and can vary by many orders of magnitude per cell [177]. If the rates for one of these steps significantly differ across cell lines, it would naturally have a dominant influence on determining the variation in protein abundance. There have been a number of reports, including system-wide studies, which conclude that the variation in protein abundance is predominantly controlled

at either the level of transcription [178–180] or translation [181–184]. In the reports by Schwanhusser et al. [183] and Li et al. [180], protein degradation only accounts for either 5% or 8%, respectively, of the variance in true protein levels [185]. On the other hand, the noise identified in biochemical systems was shown to arise to a large extent from degradation reactions [31]. In spite of increasing interest in the field, there is not much known about individual proteins: a specific product might be under substantially different control mechanisms, depending on its function and the molecular networks it contribute to.

Our study here focuses on identifying the dominant determining factor behind basal p53 protein abundance. Erroneous regulation of the corresponding gene has been shown to have a strong correlation with tumour development, with more than 50% of total cancer cases exhibiting a non-healthy p53 expression pattern [19]. In particular, the tumour suppressor protein is intrinsically involved in some specific cancers with over 96% of ovarian cancer cases involving some p53 mutation [186].

One of the most important and widely investigated aspects about the regulation of p53 is its relationship with Mdm2, an E3 ubiquitin ligase, whose transcription is activated by p53. Upon expression Mdm2 readily binds to and ubiquitinates p53, henceforth targeting it for degradation by the proteasome [187,188]. The outcome of this relationship is that, under non-stressed conditions, p53 levels are automatically kept relatively low; as any expressed p53 will readily promote the transcription of Mdm2 and ultimately its own degradation.

On the other hand, in the case of stress response, this relationship can be interrupted in varying ways depending on the stress involved. For example, under conditions of DNA damage, induced either by ionising radiation or ultraviolet irradiation, phosphorylation of both p53 [189–192] and Mdm2 [193] prevent their interaction. Under excessive proliferative conditions the presence of activated oncogenes induces the Arf protein, which is able to bind and sequester Mdm2 [194]. Similarly, in the case of damage to the ribosome, large ribosomal RNAs leak into the nucleus and

bind to Mdm2, once again sequestering it away from its partner [195, 196]. The outcome of these stresses is the same: without its inhibitory binding partner p53 is able to rapidly accumulate in cells and activate some downstream effector.

In this chapter we study the mechanism of control p53 is under in non-stressed conditions, in particular the process regulated to set baseline expression levels in a specific cell line. The relationship between p53 and Mdm2 might be taken to imply that p53 protein is controlled predominately through degradation rather than repression of the gene itself. In order to investigate this we report, using single cell immunofluorescence measurements, the steady-state distribution of p53 protein abundance in two human cancer cell lines, MCF7 and BE. We demonstrate how the distributions can be studied through stochastic modelling using a two-stage model of gene expression. In our computational analysis, we formalise mathematical models based on the two competing hypotheses that either the rate of mRNA transcription or protein degradation is the predominant determinant of basal p53 protein abundance in these cells. We further provide modifications of an Approximate Bayesian framework to utilise the full information available in the measured distributions. Through subsequent parameter estimation and model selection we conclude that differences in abundance levels within these cell types may be best explained by mechanisms which affect the rate of p53 protein degradation. This conclusion is perhaps expected in the light that that p53 expression in stressed conditions is orchestrated through protein degradation [197]; while on the other hand it provides a clear counterexample to the general finding that most protein levels are regulated by mRNA transcription.

6.2 Methods

6.2.1 Data collection

All experimental data sets presented in this chapter were generated by Ali Salehi-Reyhani and Michael Barclay at the Department of Chemistry, Imperial College London. The contributions are detailed in [18].

We use single-cell data of fluorescently labelled p53 content within two different cancer cell lines: MCF7, a widely used human breast adenocarcinoma cell line; and BE, a line of human colon-cancer cells. The cell lines were chosen as – unlike many cancer cell lines – p53 is active in both. The signal is generated with immunostaining against p53 following the methods described in [198], and measured with widefield fluorescent microscopy (Nikon Ti-E, Nikon, Japan). Images are analysed for all measured fluorescent channels using automated algorithms written for FiJi [199]. To compensate for the epifluorescence illumination profile of the microscope, tile scans of fluorophores in a mixed solution are acquired and used to flatten images. The exact details of cell preparation and image acquisition are given in [18].

The main computational analysis is focused on a set of data obtained in 2015; however, for validation the whole procedure is repeated on measurements from a second experiment performed in 2014 with the same imaging and quantification technique. Both contain a pair of measured distributions for MCF7 and BE cells obtained simultaneously and hence under identical conditions.

Besides fluorescent measurements of p53 protein levels, mRNA levels are also assessed through *quantitative Polymerase Chain Reaction* (qPCR) to provide further experimental verification of modelling results. Total mRNA is obtained from cultured and subsequently lysed cells using a commercially available isolation kit. The total yield is recorded through spectrophotometry, which is used to achieve optimal concentration for the reverse transcription reaction. The quantitative step is per-

formed using a real-time PCR machine (ABI Prism 7700, Applied Biosystems, USA) and for analysis results of TP53 amplification are recorded together with housekeeping gene HPRT1. Relative mRNA levels were calculated using the comparative CT method normalised to the first MCF7 sample [200].

6.2.2 Model and simulation

We model the underlying dynamics with the simple linear model of stochastic gene expression introduced in Equation (2.1). In this formalisation only mRNA and p53 protein levels are modelled explicitly and all other regulatory molecules and processes influencing this system are implicitly modelled in the rate values of the individual reactions.

Given the limited number of measurements, many parameter combinations can give rise to the same target distribution, and we expect that only ratios of reaction rates can be determined instead of absolute values. As our aim is to perform inference and model selection, simplifications can be beneficial by narrowing the parameter space to be searched. Therefore we normalise all parameters by the mRNA degradation rate – assuming, effectively, that this reaction is the least affected by the change of conditions between samples – and hence focus our analysis on a base model with three free parameters: k_1, k_2 and k_3, as indicated in Figure 6.1. All rates estimated in the inference step will be in arbitrary time-units based on the mRNA half-life.

We formulate two models based on our hypotheses that p53 expression is dominantly regulated by (I) transcription, (II) protein degradation. In each model both cell lines follow the simple base model with the same rate parameters, except for a modifying factor influencing the crucial regulatory reaction. Model I corresponds to a transcription-controlled baseline p53 expression, where the difference in protein level between the two cell lines arises from a scaling factor on transcription, s_1. Similarly, Model II attributes the difference to reduced p53-degradation in BE

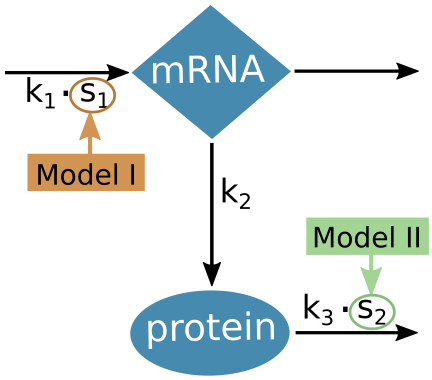


Figure 6.1: Schematic representation describing p53 expression in two cell lines. Blue filled forms denote modelled species, arrows represent reactions with their respective rates shown next to them. The two factors, s_1 and s_2 are only effective in Model I and II, respectively, and assumed to be unity otherwise.

cells, expressed in a second scaling factor, s_2. Computationally, the two models are represented in the form of different reaction propensities corresponding to processes in MCF7 and BE cells. Let $m_1,\ p_1$ and $m_2,\ p_2$ denote the mRNA and protein levels in MCF7 and BE cells, respectively; and $a_i^{(1)}$ and $a_i^{(2)}$ the reaction propensities in the two cell types. Then Model I demonstrated in Figure 6.1 corresponds to the following propensities

$$\text{MCF7: } a_1^{(1)} = k_1,\ a_2^{(1)} = m_1,\ a_3^{(1)} = k_2 m_1,\ a_4^{(1)} = k_3 p_1; \tag{6.1a}$$

$$\text{BE: } a_1^{(2)} = s_1 \cdot k_1,\ a_2^{(2)} = m_2,\ a_3^{(2)} = k_2 m_2,\ a_4^{(2)} = k_3 p_2. \tag{6.1b}$$

Similarly, the model focusing on protein degradation rate can be formulated as,

$$\text{MCF7: } a_1^{(1)} = k_1,\ a_2^{(1)} = m_1,\ a_3^{(1)} = k_2 m_1,\ a_4^{(1)} = k_3 p_1; \tag{6.2a}$$

$$\text{BE: } a_1^{(2)} = k_1,\ a_2^{(2)} = m_2,\ a_3^{(2)} = k_2 m_2,\ a_4^{(2)} = s_2 \cdot k_3 p_2, \tag{6.2b}$$

where the propensities $a_1, \ldots, a_4$ stand for transcription, mRNA degradation, translation, and protein degradation, respectively.

Several of the previously detailed simulation approaches can be used for simulation of the system, each offering different levels of exactness and computational

efficacy. An exact realisation using Gillespie's SSA would offer the highest resolution, however, as the system includes several million units of protein molecules, this simulation method would prove computationally prohibitive, and the simulation of a whole population, let alone thousands of populations is not feasible in our case. Therefore we apply an SDE formulation, which requires only negligible approximation compared to exact simulations, but offers a significantly faster way to simulate a relatively dense but still stochastic system. As a verification we compare SSA and SDE simulations with rate values based on the inferred parameters and do not find a difference in output (see Figure C1 in Appendix C).

For each parameter combination, we simulate n instances of the system for an appropriate time so that the system settles for the steady-state values, then compare the arising steady-state distribution with the target cell population. The population size, n, is chosen as $n = \max\{n_{mcf7}; n_{be}\}$, and the population formed from the first n_{mcf7} (or n_{be}) simulated cells is compared to the experimental MCF7 (BE) protein numbers.

6.2.3 Model selection and inference

For model selection, we use an Approximate Bayesian Computation (ABC) method based on sequential Monte Carlo (SMC) sampling [55, 201, 202]. ABC methods offer various advantages making them an ideal candidate for parameter estimation and model selection. The Bayesian framework [203] offers a thorough analysis of the parameter space – instead of a single optimal value – in the form of a *posterior distribution*, which describes for the range of possible parameter combinations, how likely they would give rise to the observed data [204]. This approach makes it possible to identify crucial dependencies between parameters and thus understand the mechanisms behind the model and the data set. ABC substitutes the calculation of the likelihood function – that formalises the probability of measuring the data under a selected model and parameter set – with a distance calculated

between simulated and target data sets [205–207]. This provides a way to overcome difficulties if an exact likelihood cannot be analytically derived and does not require substantial changes to accommodate for various models.

Sequential Monte Carlo sampling extends the applicability of the ABC method by sampling the parameter space and evaluating the distance function through a series of intermediary distributions, in order to avoid impractically long convergence times that might occur if the correct parameter subset is very tightly distributed [207, 208]. The intermediate distributions are constrained by a series of acceptance thresholds, corresponding to an increasingly better agreement between simulation and experimental data, until the distance goes below a certain target value.

Each iteration of the algorithm consists of the following steps illustrated in Figure 6.2 [55, 201]. First, a batch of parameter combinations (often referred to as *particles*) is sampled from the prior distribution. The initial distribution is defined manually based on our knowledge of realistic parameter ranges, and afterwards each prior distribution is calculated as the posterior of the previous iteration with some additional disturbance [202]. In model selection, when there are several competing models, first a model is chosen randomly according to its proportion in the previous posterior, then we move on to the particle sampling step. The system is simulated for each particle and the output (e.g. trajectory of a variable or steady-state distribution) is compared to the experimental data we aim to fit. Based on the distance between the simulation and data and a threshold value (ϵ_i for the ith iteration), we either accept or discard a particle [205, 209]. These steps are repeated until a pre-defined number of accepted particles have accumulated (typically in the range of 500-2000). An intermediate posterior distribution is generated from these values and the next iteration step is entered targeting a smaller ϵ value [210]. The algorithm is repeated until a predefined final threshold, ϵ_f, is reached so that all simulated data sets fall within a distance of ϵ_f from the experimental measurements.

Model selection is based on the percentage of accepted final particles obtained

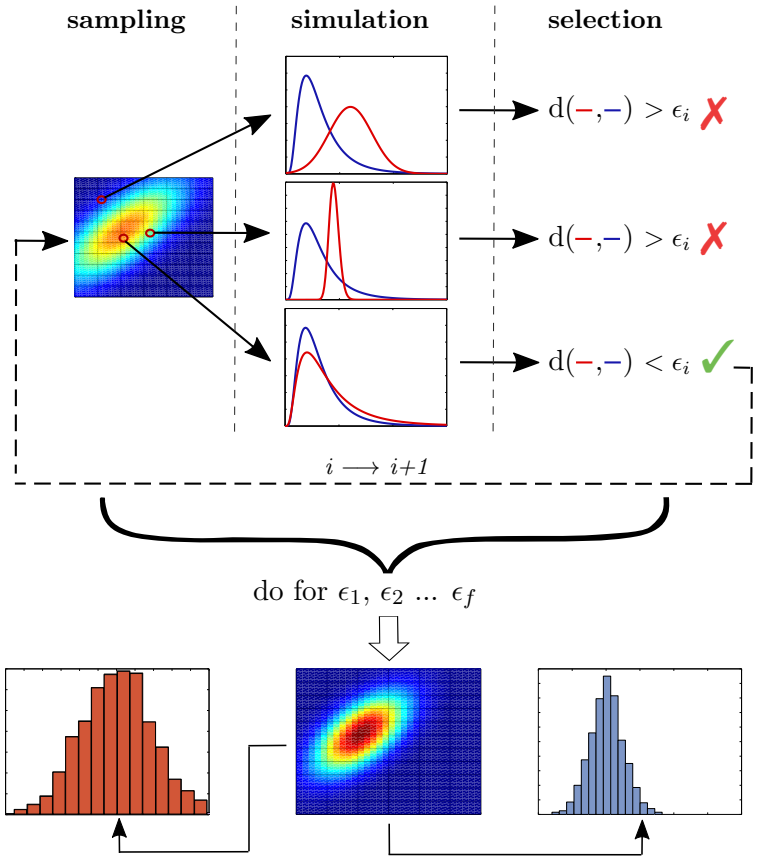


Figure 6.2: Schematics of the ABC SMC algorithm. The three steps on the upper panel are repeated in an iterative manner to estimate parameter values that can fit experimental measurements. Details are given in the main text.

from each model, yielding a model evidence value between 0 and 1, with both models starting from 0.5 evidence. As this value directly derives from the acceptable parameter combinations from each model, the ratio of evidences provides a Bayes factor to evaluate models [211]. A ratio above 4.0 represents substantial evidence in favour of a given model according to the scale first proposed by Jeffreys [212].

6.2.4 Distance measure

The reliability and discriminating power of our estimation procedure is strongly influenced by the properties of the distance function used to compare simulated and target distributions. The ABC algorithm is usually used with some summary statistics (e.g. Euclidean distance between the means of the populations), but in our case the main information lies in the form of the distribution, so we aim to find a distance measure that has the ability to compare whole distributions. We also require this measure to (i) take its minimum at the true optimal value and (ii) show a steep increase as we diverge from the true parameter combination. The third concern about measures is (iii) how sensitive they are to the noise appearing in the distribution due to the stochasticity of the system and the small number of measurements. To select the best distance function, we check their properties on an example population, following the base model for a single cell line. A synthetic dataset is obtained by simulating 100 cells with parameters $[k_1, k_2, k_3] = [15, 50000, 50]$. This is used as the target distribution to compare against populations derived from other parameter combinations. The parameters are scanned for the following ranges: k_1 and k_3 is checked with 21 equidistant values, centred at the 'true' values; k_2 is varied between five values. The results are plotted for all k_2 values as two-dimensional distance landscapes of k_1 and k_3 (Figure 6.3 and Figure C2).

As all landscapes point out that the parameters cannot be unambiguously identified, no distance function will have a clear minimum at the actual rate values. However, some show a rather flat landscape around the true value, while others have more power in selecting the right parameter range. Furthermore, techniques comparing the two empirical *probability* distribution functions are highly influenced by the choice of method or bin size used converting the samples into a distribution. Therefore we conclude that distance functions evaluated according to the *cumulative* distribution function (the two upper examples in Figure 6.3(a)) show the best

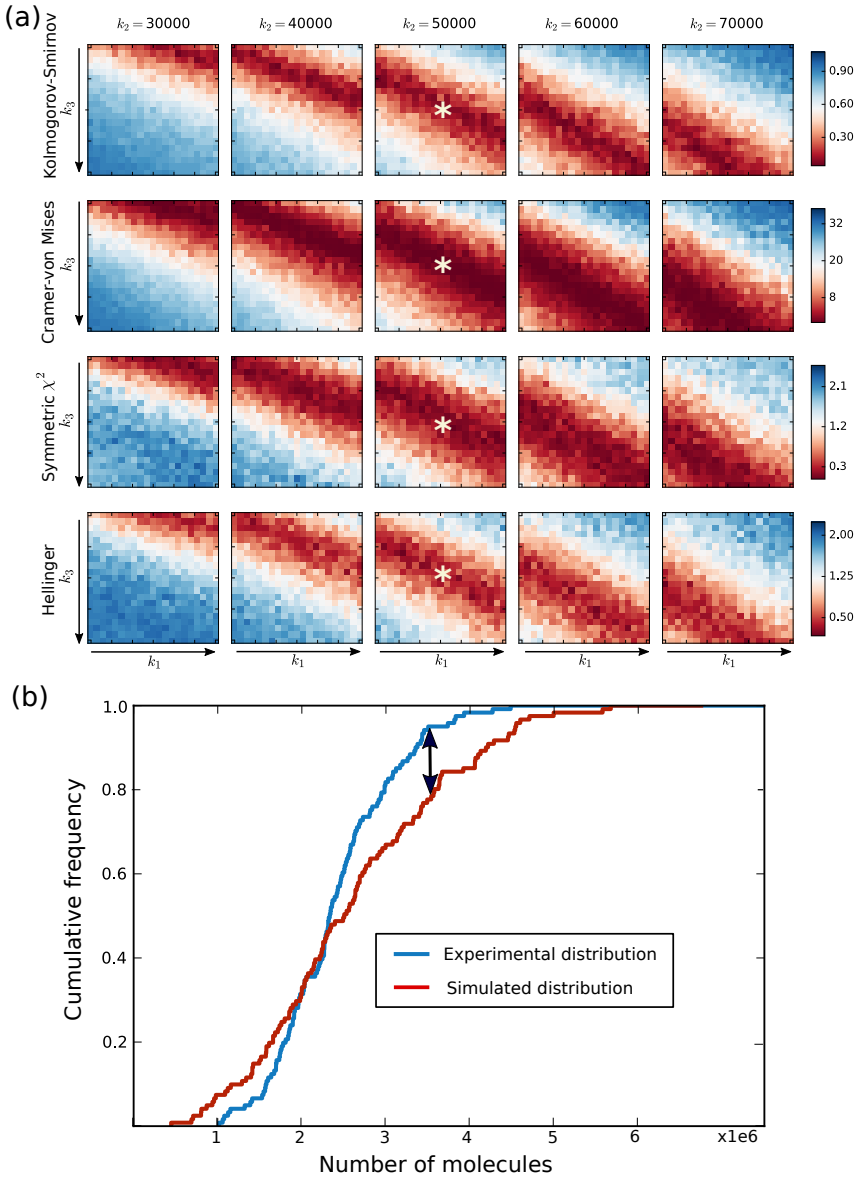


Figure 6.3: Distance measure used in the ABC algorithm. **(a)** Colormap of distance values computed with different measures for a range of k_1, k_3 and k_2 values. All rows are normalised to the same colour bar shown at the right end of the row. The true values used in the simulated dataset are indicated by white asterisk. Evaluation of further distance functions is attached in Figure C2. **(b)** Example of the computation of Kolmogorov-Smirnov distance between an experimentally measured distribution (MCF7 cells, blue line) and a simulated distribution obtained with random parameter values (red line).

properties. We choose the Kolmogorov-Smirnov distance function [213], calculated as

$$\mathrm{d}_{KS} = \sup_{x \in \mathbb{R}} \left| P_{sim}(X \leq x) - P_{targ}(X \leq x) \right|, \tag{6.3}$$

where P_{sim} is the simulated and P_{targ} is the target distribution, and the distance function is evaluated at a large sample of x values as implemented in the Python built-in function, *ks_2samp*. Figure 6.3(b) demonstrates the computation of the distance.

The Kolmogorov-Smirnov distance has a further advantage: as it is widely used in statistics for evaluating if two distributions are significantly different, it comes together with a logical choice of ϵ threshold values in the iterative estimation algorithm. According to the classical two-sample Kolmogorov-Smirnov test our two populations are *rejected* to be coming from the same distribution, at confidence level α, if their distance fulfils

$$d_{KS} > c(\alpha) \sqrt{\frac{n_{sim} + n_{targ}}{n_{sim} \cdot n_{targ}}},$$

where $c(\alpha)$ is the critical value defined by the Kolmogorov-Smirnov function. On a similar basis, we choose to *accept* a parameter combination in iteration i, if the obtained simulated distribution has a normalised distance lower than the constant corresponding to α_i,

$$d = d_{KS} \sqrt{\frac{n_{sim} \cdot n_{targ}}{n_{sim} + n_{targ}}} < c(\alpha_i) = \epsilon_i.$$

With a sequence of thresholds such that $\epsilon_{i+1} < \epsilon_i$, we automatically imply $\alpha_{i+1} > \alpha_i$, which can be interpreted as an increasing confidence in the two distributions being identical. For example at $\epsilon = 1.224 \rightarrow \alpha = 0.1$ we reject any simulated populations that have less than 10% chance of being sampled from the same distribution.

The distance (including corrections for the sample size) for each cell line is evaluated independently and the final distance is computed and assessed as

$$d_{total} = \max\{d_{mcf7}, d_{be}\} < \epsilon_i,$$

ensuring that both simulated populations will have a fit to their corresponding experimental distribution at least as close as the current ϵ_i value requires.

Since we have access to data regarding relative mRNA abundance in the cell lines, this information also gets incorporated in a modified version of the distance function. In order to filter out populations in which the average mRNA levels are not on the same magnitude in the two cell lines, an additional penalising term is added based on the difference between mean mRNA numbers ($\overline{m}_{mcf7}$ and $\overline{m}_{be}$):

$$d_{modified} = d_{total} + d_m = d_{total} + \frac{|\overline{m}_{mcf7} - \overline{m}_{be}|}{\overline{m}_{mcf7} + \overline{m}_{be}}\mu, \tag{6.4}$$

where we control the weight of the penalty with the factor μ. Results without constraints on mRNA levels are obtained by setting $\mu = 0$.

6.2.5 Synthetic experiments for validation

To validate the discriminative ability of the applied methodology in its final form, we also perform model selection on synthetic datasets, generated to reflect the properties of the actual measurements and our hypotheses. We use the base model (Eq. (6.1a)) to simulate a population of MCF7 cells, consisting of $n_{mcf7} = 150$ observations. In order to show that our normalisation (i.e. expressing independent parameters by their ratios to the mRNA degradation rate) has no influence on the computation, the before-normalisation model is used, with parameter values as indicated in Table 6.1. A BE cell data set of 300 cells is generated independently, with three different instances corresponding to cases when transcription, translation or protein degradation are regulated differently in this cell line to result in a higher baseline p53 expression level. Accordingly, the rate of transcription, translation or protein degradation, respectively, is chosen to be markedly different from that of MCF7 cells; while other values are perturbed slightly as we cannot expect the exact same values in a realistic system.

Data set	$\mathbf{k_1}$	$\mathbf{k_2}$	$\mathbf{k_3}$	$\mathbf{k_{mrna-deg}}$
MCF7	30	$5 \cdot 10^6$	30	5
BE – transcr.	248	$5.3 \cdot 10^6$	28	4.8
BE – transl.	31	$1.6 \cdot 10^7$	28	4.8
BE – prot.deg.	31	$5.3 \cdot 10^6$	5.6	4.8

Table 6.1: Parameter values used for synthetic data generation. All target data sets consist of the MCF7 data combined with one of the three synthetic BE measurements.

The thus generated data sets are used as target distributions in the model selection algorithm between Model I (transcription-based regulation, Eq. (6.1)) and Model II (degradation-based control, Eq. (6.2)). In every other detail, the same procedure is used as described above for real observations.

6.3 Results

6.3.1 Qualitative analysis

We first qualitatively compare the experimental distributions to the two proposed mechanisms of regulation. In order to understand the general shape of distributions the two models would give rise to, we generate simulated populations with parameters $[k_1, k_2, k_3, s_1, s_2] = [5, 1, 5 \cdot 10^6, 10, 5, 0.2]$. To make them comparable the distributions from the two cell lines are both normalised by their respective mean p53 expression level. Figure 6.4(a) shows the normalised distributions on the same axis, obtained using the transcription and the protein degradation based model.

Due to the amplification effect throughout protein production, one expects low average mRNA levels to result in a high-variance protein distribution, as small fluctuations in mRNA abundance get overly represented during translation. On the other hand, relatively high number of mRNA molecules would prove more stable against such inevitable stochasticity and could thus lead to a narrower protein distribution. Therefore, regardless of other downstream influences, we expect to see an direct effect of relative mRNA abundance on the coefficient of variation. As demonstrated in the data simulated from Model I (Fig. 6.4(a) (uppermost graph)), the highly expressing cell line (here, red dashed line) should show considerably smaller relative variation. While in the case of similar mRNA regulation and levels, both cell lines have about the same normalised variance of p53 amount (middle panel in Fig. 6.4(a)).

We also visualise the experimentally measured expression levels in the same way. The shape of the two distribution of the cell lines are qualitatively similar, and the normalised experimental distributions are close to each other, as highlighted in the lowermost panel of Figure 6.4(a). Since their coefficients of variation does not show the difference expected from Model I (in either data sets), we propose that mRNA

levels are on the same magnitude in both cell lines and regulation is protein-based.

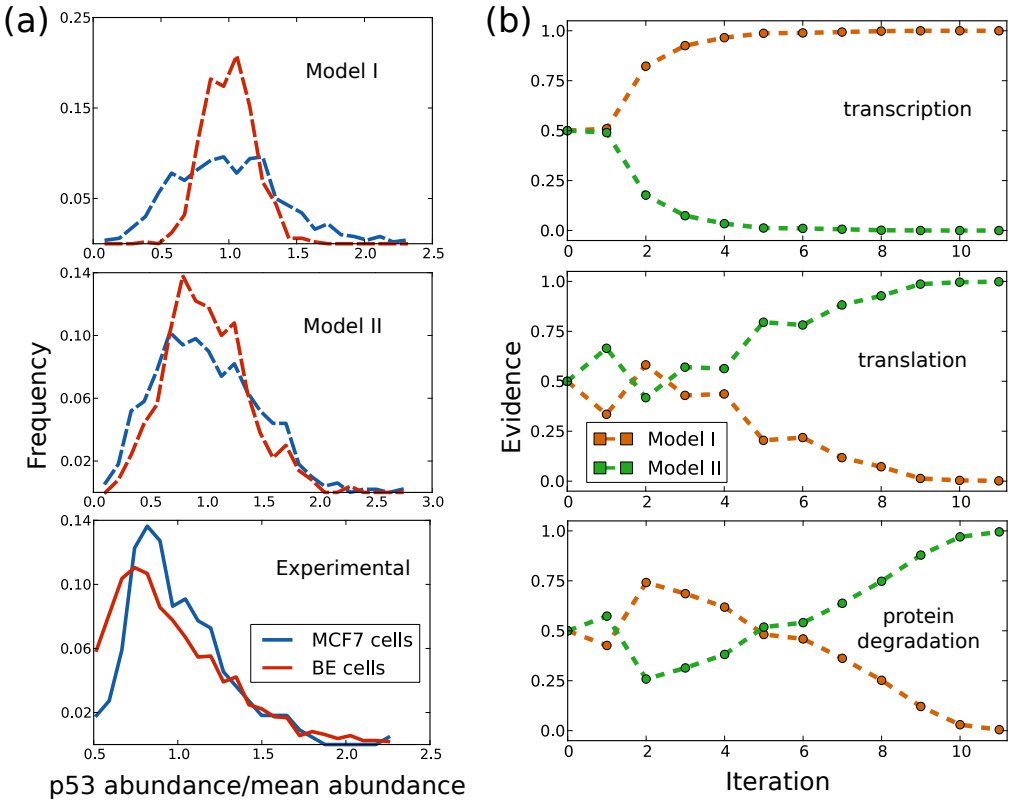


Figure 6.4: Analysis of the proposed models and the model selection process through simulated data. (a) Distributions of p53 protein level in the two cell types as obtained from simulations of Model I (uppermost panel) and Model II (middle) with a toy parameter set and from experimental measurements (lowermost panel). All expression levels are normalised to their respective mean value to enable comparison of differently expressing cell lines. Blue and red lines show the frequency distribution of MCF7 and BE p53 levels, respectively. **(b)** Evidence representing the ratio of accepted fits obtained from Model I (in dark orange) and Model II (in green) with model selection performed on simulated validation data sets. Synthetic data are generated with rate parameters corresponding to (uppermost panel) transcription-level, (middle) translation-level and (lowermost) protein degradation-level regulation of baseline p53 level, as detailed in Table (6.1).

6.3.2 Model selection and inference results

We first present the results of computational analysis on the synthetic validation data set. Figure 6.4(b) shows the evidence supporting the two competing models in all three validation experiments. As expected, in the case of the transcription-based control data set, Model I is quickly picked up as the one underlying the difference between cell lines, even though no information on mRNA levels is used for discrimination. In the other two cases, the degradation-based model is selected with decisive evidence. Since in both cases a protein-level process is governing the level of p53, even in the second experiment we expect protein degradation control to provide a much better explanation of the observed data than transcription regulation. These results confirm that we can confidently discriminate between a regulatory effect on the mRNA level versus on the protein level and, even though our prior expectations point towards the second case, our method has no bias against Model I.

After the discriminative power of our algorithm has been confirmed, we carry out the same analysis on experimentally measured distributions in MCF7 and BE cells.

	$\mathbf{k_1}$	$\mathbf{k_2}$	$\mathbf{k_3}$	$\mathbf{s_1}$	$\mathbf{s_2}$
Data I – repeat 1	[1, 20]	[$5{\cdot}10^5$, $1.2{\cdot}10^7$]	[0, 22]	[0, 6]	[0, 2]
Data I – repeat 2	[1, 20]	[$5{\cdot}10^5$, $1.2{\cdot}10^7$]	[0, 35]	[0, 6]	[0, 2]
Data I – repeat 3	[1, 10]	[$2{\cdot}10^6$, $1.2{\cdot}10^7$]	[0, 28]	[0, 6]	[0, 2]
Data II – repeat 1	[0, 18]	[$5{\cdot}10^5$, 10^7]	[0, 15]	[0, 4]	[0, 2]
Data II – repeat 2	[0, 15]	[10^6, 10^7]	[0, 12]	[0, 4]	[0, 2]
Data II – repeat 3	[0, 10]	[$2{\cdot}10^6$, 10^7]	[0, 15]	[0, 4]	[0, 2]

Table 6.2: Prior distributions used in the ABC algorithm. All priors are uniform with upper and lower bounds as indicated in their respective rubric. Results of Data I are presented in the main text, the analysis of Data II can be found in Appendix C.

The algorithm is initialised with fixed initial conditions $p_{init} = 4{\cdot}10^6$; $m_{init} = 100$ and uniform initial priors for all parameters. The ranges of uniform distributions are listed in Table 6.2. After the first few iterations, the evidence for the degradation-based model quickly exceeds that of the transcription regulation model, with all the evidence in the final parameter set supporting Model II. This result is consistent throughout a range of prior distributions for the model parameters (Fig. 6.5(a)). The superiority of Model I in the first section of iterations is due to the high acceptance rate that favours the more flexible transcription-based model, probably finding a local minimum in the parameter space. However, as the required quality of fit increases, this model is not able to reproduce the experimental data well enough.

The coupled parameter inference concludes that p53 degradation is about five times *slower* in BE cells, consistent with our expectations given the general properties of the BE cell line. An example population simulated with an instance of the final, best fitting group of parameter sets is shown in Figure 6.5(b) together with the experimentally measured data. The two distributions show no significant difference even at a level of $\alpha = 0.3$. The main distinction between simulated and experimental populations is in their tail: the simulated dataset fails to reproduce the significantly different cells lying far off from the main body of the distribution. However, it has been shown that even non-stressed cells can show oscillatory patterns in their p53 expression [214]. Therefore we conclude that the high-expressing measurement points are cells measured in the high point of a stochastic p53 burst and they form a separate population of outliers, whose dynamics are beyond the scope of our model. Analysing such cells together with the main distribution can lead to erroneous results, so less weight should be placed on these cells when accessing goodness of fit and we filter them for a better representation of the resulting distributions.

The analysis is repeated on a second data set (obtained under identical conditions) for validation. Prior distributions used in the ABC algorithm are listed

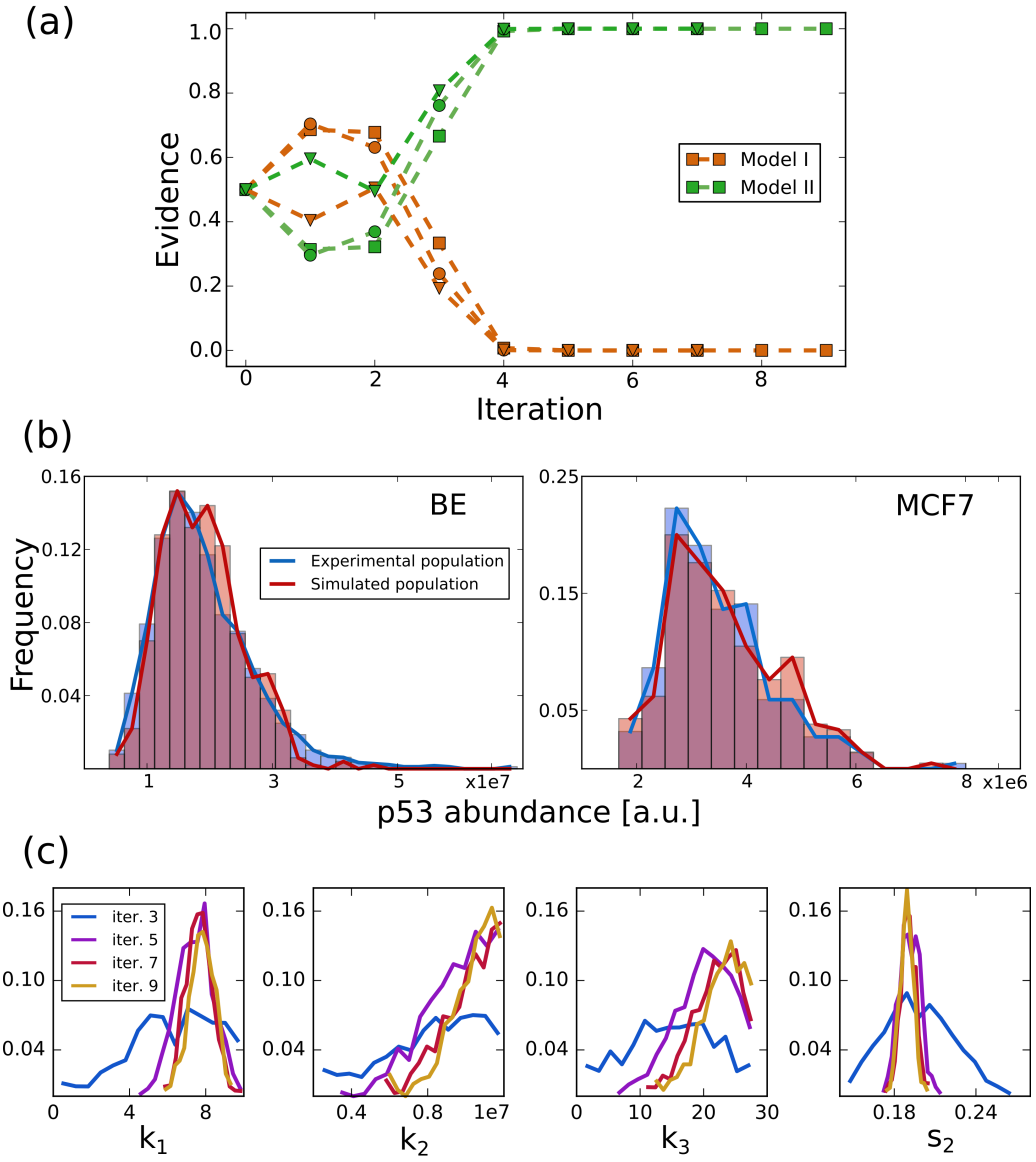


Figure 6.5: Results of computational analysis. **(a)** Dark orange and green curves show evidence for the transcription- and the degradation-based model, different markers denote replicates with different prior parameter distributions used at initialising the algorithm. **(b)** Histograms of the distribution of p53 abundance (measured as arbitrary fluorescence units) in (left) BE and (right) MCF7 cells. Blue distributions are measured experimentally, red distributions are simulated during inference using the best fitting parameter combination. α values of Kolmogorov-Smirnov distance are $\alpha = 0.34$ (BE) and $\alpha = 0.59$ (MCF7). **(c)** Posterior parameter distributions after iterations 3, 5, 7 and 9, as indicated in legend.

in Table 6.2. The results are included in the Appendix (Figure C3). Once again, Model II shows a higher level of evidence after the first few iterations, with a final Bayes factor of 6.35 supporting it over Model I. Furthermore, the experimental distributions are reproduced by the simulated model up to a level of $\alpha = 0.5$.

In addition, we confirm the robustness of our result by repeating the analysis on slightly modified versions of our basal expression models. It is known that the dynamics of p53 are regulated through a negative feedback loop that introduces nonlinearity into the system; so we modify our model structure to contain p53-dependent, nonlinear rates of protein degradation. The rate k_3 is substituted by $k_3 p$ (or $s_2 k_3 p$) and we perform the same analysis on these models. The results are in agreement with our previous findings, as Model I is discarded rapidly in the model selection process; however, the target distribution cannot be fitted to the same quality as with the entirely linear model. This is probably due to the nonlinear model's sensitivity to the protein degradation rate: only a small range of values can approximate the experimental population, and hence the inference algorithm converges very slowly, while the original model provides a more robust steady-state distribution.

6.3.3 Incorporating mRNA information

The relative abundance levels of p53 mRNA, as obtained through real-time quantitative PCR measurements, are shown on Figure 6.6(a). The measurements confirm that there is no significant difference between mRNA levels in the two cell lines, in agreement with our expectations based on Figure 6.4(a) and with the computational analysis unambiguously selecting protein-level regulation.

We also repeat our analysis using the extended distance function incorporating a constraint on relative mRNA expression levels. The validation data set (Data II) is used since in this case the transcription based model was not excluded completely, and hence we can gain the most by incorporating new data in this analysis. Figure

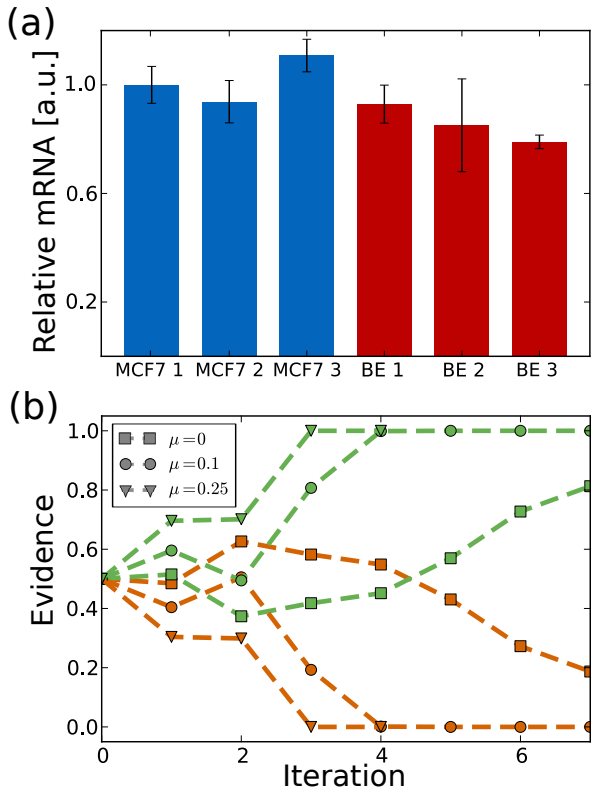


Figure 6.6: Model selection incorporating qPCR results. **(a)** Results of real-time PCR measurement quantifying relative mRNA abundance in MCF7 (red bars) and BE (blue bars) cells. All levels are normalised to the first MCF7 sample (leftmost data bar). **(b)** Model selection results using the distance function extended with a term penalising difference in mRNA abundance. Dark orange and green curves show evidence for Model I and II, respectively, for three values of the penalty weight, μ, as indicated in legend.

6.6(b) shows the evidence supporting Model I (protein degradation based) and Model II (transcription rate based) in each consecutive iteration for three different values of the penalty weight, μ, in Eq. (6.4). Even without taking into account the mRNA measurements ($\mu = 0$), the analysis confirms a Bayes factor of substantial evidence for the protein degradation based model, while for higher μ values the other model is completely unsupported in the final population. Unsurprisingly, increasing the weight of the penalty term also leads to a faster elimination of the transcription-

based model, as cellular data simulated with this model typically have a 10-20-fold difference in mRNA levels.

6.3.4 Absolute quantification experiments

Besides single-cell measurements, providing the basis of the above analysis, we also aim to quantify the exact copy number of p53 in single cells. For this aim, single-cell images are generated on a Microfluidic Antibody Capture (MAC) chip using an 'all optical' cell manipulation method in which all operations on cells (positioning, lysis and analysis) are performed using lasers acting as optical tweezers [215]. Figure 6.7(a) shows a schematic drawing of the procedure. First, cells are placed in solution in a microfluidic device which have 25-50 individual lysis chambers to which single cells are transported using dual-beam optical tweezers. There cells are lysed by an infrared pulse and the thus freed p53 protein molecules are bound by a primary antibody coating the surface of the chamber. The proteins are detected through a fluorescently labelled secondary antibody with Total Internal Reflection Fluorescence microscopy [216, 217] that ensures a high signal-to-noise ratio. This way an absolute quantification of p53 can be achieved with read-outs in exact copy number of protein molecules up to the precision of a hundred molecules [218].

We compare p53 expression in MCF7 cells obtained with the above method and immunofluorescence, and find that fluorescence data provides a good match to chip experiments (Fig. 6.7(b)). Note that p53 abundance in both experiments is normalised to its respective mean to ease comparison. The coefficients of variation – that have proved to be an important characteristic of p53 distributions – confirm the fit: 0.3639 and 0.3945 for the chip and microscopy data, respectively. In the case when rightmost outliers are omitted, for previously discussed reasons, the two coefficients of variation are 0.3048 and 0.2892. We also confirm using a Kolmogorov-Smirnov test that the two distributions are not significantly different ($\alpha = 0.9$).

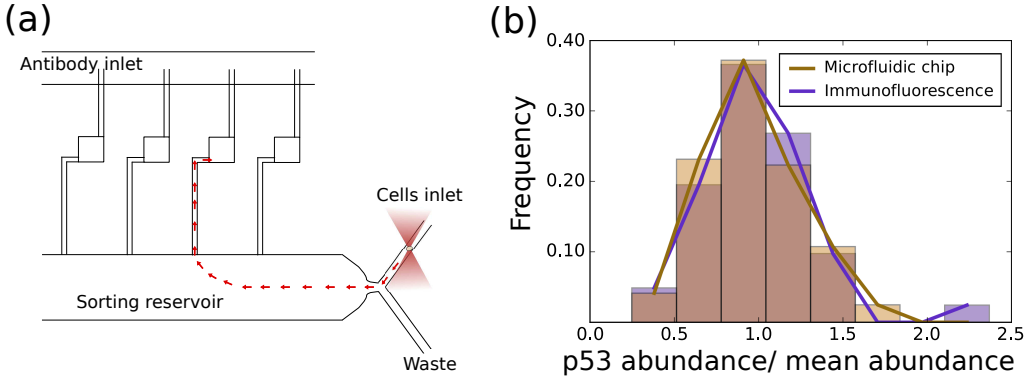


Figure 6.7: Microfluidic chip measurements of p53 expression level. **(a)** Schematic structure of the microfluidic chip and cell manipulation used in the single-molecule resolution experiments. The red arrows illustrate a typical path travelled by a cell, from when it is captured by the optical tweezers (represented by the beams) to the cell lysis chamber. **(b)** Histogram of normalised p53 copy number in MCF7 cells, measured on a microfluidic chip (gold bars and line) and by single-cell immunofluorescence (purple bars and line). Kolmogorov-Smirnov distance: 0.0954, $\alpha = 0.931$.

Given these findings we hope that, with further advancements in experimental calibration, measurements obtained with high-throughput and cost-effective methods, like fluorescent microscopy, can be converted into exact copy number data, and our computational framework could provide conclusions and predictions not only on the single-cell but the single-molecule level.

6.4 Discussion

In this work we have shown that the combination of steady-state single cell measurements and computational techniques can be a powerful tool in understanding the mechanisms which help establish the baseline state of cells. We have measured the steady-state basal p53 mRNA and protein abundances in two human cancer cell lines. In particular, we found that, despite the high number of proteins under transcriptional control [185], the regulation of basal p53 protein level in these human cancer cell lines is achieved through the rate of p53 protein degradation.

Understanding the contributions of transcriptional and translational control is important in ultimately understanding the underlying mechanism of cancer and how drugs may sometimes fail. For example, Zhang et al. reported that the variation in protein expression in colorectal tumours was poorly explained by the abundance in corresponding mRNA [184]. Certainly, a lack of correlation between mRNA and protein abundances has been repeatedly reported as single cell analysis becomes increasingly quantitative [64, 219, 220]. So far, system-wide studies have challenged one another in the conclusion that either transcriptional or translational control/variation contributes most to the variance in protein abundance [180, 183]. As far as we are aware, this is the first report that concludes protein degradation is the predominant contributor to the variance in protein abundance of, specifically, basal p53 protein.

This makes sense in terms of the role p53 plays within the cell. Sometimes referred to as the *guardian of the genome* [221], p53 acts as an integrator of cellular stresses and must accumulate and respond rapidly. With this goal in mind, it stands to reason that it is perhaps faster to constitutively express and degrade p53 rather than activate the slow process of transcription and translation in times of stress. The data shown here does not imply that transcription rate and other processes has no effect on p53 abundance and dynamics. p53 mRNA itself might play an important role: indeed, it is known that p53 mRNA can directly bind to Mdm and

interrupt the protein-protein interaction that leads to p53 degradation [222]. An increase in Mdm2 and p53 mRNA immuno-precipitation has been observed after treatment with the DNA damaging drug doxorubicin. Importantly there seems to be little correlation between transcript levels and corresponding protein abundance.

We would like to emphasise that our aim is not to create an accurate model of p53 dynamics, but to identify a minimalistic model that can reproduce the steady-state distribution and can be used to distinguish between the two regulatory mechanisms in question. We choose to use a simplified model to reflect experimental limitations, namely that we have no ability to alter individual reactions with sufficient accuracy and selectivity. Drugs or other treatments usually have an effect on the underlying network at several points, especially in the case of p53 which has a particularly complex upstream and downstream interaction network [223]. Even if we include a more detailed picture in modelling, experimental interventions would influence numerous of these detailed processes, possibly without our knowledge. Therefore our model is based on the appropriate abstraction level where we can reliably discriminate between modelled processes. Even though the mechanism of action behind cell differences cannot be pinpointed with this model, our aim is to avoid any erroneous conclusions and imprecise identification. As an example, from our current measurements we would not have been able to distinguish if the smaller degradation rate is due to differences in Mdm2 abundance or phosphorylation status. However, the current level of measured detail can be sufficient for certain applications, such as for identification of the main point of action of a drug (e.g. whether it alters p53 expression through mRNA or protein level changes). We are also aware that the simple gene expression model used in this work cannot explain many characteristics of p53 dynamics that arise from nonlinear feedback loops. However, as only steady-state measurements of basal p53 distribution are available, using a simplified model can provide an adequate computational description of our observations.

Nevertheless, the ABC model selection framework makes it easy to repeat the

analysis with extended models or different hypotheses. Our plans for possible extensions include: explicit modelling of Mdm2 as a species and the reactions leading to degradation by Mdm2, as well as chemicals that modify this interaction [224]; modelling relevant post-translational modifications and their respective components, such as phosphorylation [192, 225] and ubiquitination [187]; including compartmentalisation, i.e. separate modelling of nuclear and cytosol-bound p53 [226–228]. Extension of the computational framework is tied to further experiments such as simultaneous observations of p53 and Mdm2, as well as time-course data at different locations in the cell. In addition, the main framework could be extended or modified for the analysis of other independent datasets, as long as the simple two-step expression model can be considered appropriate – or an alternative model is known – and single-cell resolution data is available. Throughout the course of these extensions we hope to create a specialised but flexible framework that may be broadly applied to study other proteins of interest to help researchers further elucidate the mechanisms controlling protein abundance and how these may be altered in disease or drug therapy.

Seven

Summary and conclusions

Stochastic fluctuations are crucial in biology for ensuring adaptation and evolution in changing environments. On the other hand, inherent stochasticity on the level of the individual (molecule, cell or organism) is ultimately smoothened out on the macroscopic population level (cell fate, tissue identity or animal populations). Shedding light on the intricate interplay between these two levels is key in understanding how biological processes operate and how could we regulate them in order to achieve goals in synthetic biology or personalised medicine.

The application of mathematical modelling for questions in biology has been a powerful tool in understanding nature. It has enabled researchers over the recent decades to uncover mechanisms never suspected before and progress towards gaining control over biological processes. However, many of the traditional, deterministic, techniques cannot describe stochastic systems appropriately. In this thesis we have presented several methods for the analysis of systems highly affected by randomness.

Let us revisit the everyday example introduced in the first chapter: using public transport between Marble Arch and Kings Cross. As we established there, random fluctuations in the time your trip takes can have a non-negligible effect, and many

aspects of this randomness has to be considered to truly understand the process.

As a first step in planning your trip, a more complete information on the distribution of journey times with bus 30 is needed. However, it is understandable if you do not wish to carry out thousands of individual observations – i.e. bus journeys – just to be able to draw this distribution. Instead assume that there is a website listing a few features: the average journey length, how much variability is in it and what is the probability of having an unusually long travel. As these are the most crucial characteristics you need to know, these three values provide sufficient information on bus 30 to plan your travels. In a similar way, the method introduced in Chapter 3 helps quick simulation and evaluation of biochemical systems, by providing access to descriptive statistics on distributions of molecular abundances in the cell. We have shown that our method estimates the features of systems with complex interactions better than previous techniques, and can be generally applied to many scenarios.

Next you want to evaluate if a certain strategy, for example favouring the underground over bus 30, is more desirable for catching a train from Kings Cross. You have to take into account the combination of average journey time and variability, but there might be additional unknowns to consider: for example that heavy raining elongates bus travel times, while the underground can get suspended due to flooding. You need to obtain an upper limit of journey time or variability taking into account the mode of travel and some uncertainty about weather condition. Our reachability analysis presented in Chapter 4 deals with this problem: we obtain a conservative estimate of plausible average and variance values, and use this to evaluate or validate control models in the context of biological networks. We applied our method for an elementary controlled building block of synthetic biology and found inevitable levels of variability, as well as validated models using only a small number of observations.

The next issue raised in the Introduction was making specific observations of your journey: whether or not you arrive to your destination by 6.15pm. Alternatively, assume the scenario that there is a significant difference between travel times

on Tuesdays and Wednesdays. Despite the limitations of your observations, you want to understand what factor determine the final output. Through incorporating prior knowledge on the transport system you might be able to identify some crucial interactions: for example that the time you leave from home has little relevance as long as you are inside a certain time window; on the other hand, arriving to Kings Cross before 4.40pm almost certainly ensures your success. In Chapter 5 we took a visual framework to explore cellular decision making, which is typically characterised by two (or a few) significant states despite the complex underlying processes governed by fluctuations. We successfully reproduced some key phenomena of stem cell dynamics and identified a distinguished transition state along the course of differentiation. Next, in Chapter 6 we presented an algorithm for selecting between two hypotheses explaining the difference observed between two cell lines. We found that the protein degradation rate of the tumour suppressor p53 has a dominant role in determining the baseline expression level in single cells.

None of the models and methods presented here are optimal in the absolute sense. Given the level of complexity and our limited understanding of molecular cell biology, we cannot hope for a description that is ideal for all questions we might want to investigate. This thesis developed and combined approaches from fields of science that traditionally have little interaction between them, and created *simple* albeit *general* techniques to help stochastic mathematical modelling. We hope that the research done here, together with similar endeavours, will lead researchers to the quantitative understanding of biological processes at the single cell level.

Appendix A

In this section we include benchmarking results demonstrating the computational cost of the multivariate moment closure technique presented in Chapter 3. We focus on the case when moment approximation offers the most advantages: when the same model has to be simulated for a high number of parameters, e.g. in statistical inference problems.

In order to quantify the computational time needed for moment expansion, we use MEANS, run in the Jupyter notebook environment. We explore two example systems implemented amongst the pre-defined models of MEANS, the Michaelis-Menten system (see [36] IIIA) and the oscillatory p53 system (Eq. (3.17)). We time the algorithm at different closure degrees by simulating each model with $20, 100, 500$ and 2500 different parameter combinations. In comparison a naïve SSA implementation in the same environment is used to simulate 500 individual trajectories for each parameter in order to estimate the mean of the population.

In the Michaelis-Menten example the three rate parameters are varied, while the initial enzyme and substrate amount are kept constant. For each set of parameter combinations, we explore a range of k_1, k_2 and k_3 values between a defined upper and lower bound (for example using the *linspace* command). The specifics are listed in Table A1. Similarly, in the p53 system k_1, k_3 and k_5 are varied, while all other parameters are kept constant. The range of rate values and number of explored rates in each scenario are listed in Table A2.

Comb.	$\mathbf{k_1}$	$\mathbf{k_2}$	$\mathbf{k_3}$
20	5: $[1.66, 2.2] \cdot 10^{-3}$	2: $[1, 1.1] \cdot 10^{-3}$	2: $[1, 1.1] \cdot 10^{-1}$
100	5: $[1.66, 2.2] \cdot 10^{-3}$	10: $[0.05, 20] \cdot 10^{-3}$	2: $[1, 1.1] \cdot 10^{-1}$
500	5: $[1.66, 2.2] \cdot 10^{-3}$	10: $[0.05, 20] \cdot 10^{-3}$	10: $[0.05, 20] \cdot 10^{-1}$
2500	25: $[0.96, 2.36] \cdot 10^{-3}$	10: $[0.05, 20] \cdot 10^{-3}$	10: $[0.05, 20] \cdot 10^{-1}$

Table A1: Parameters explored for the Michaelis-Menten system. For each set of parameter combinations, the possible values of a rate parameter are represented by a vector with equidistantly spaced values in the interval given in square brackets. The number of elements in the vector is given before the interval. From the vectors, all possible combinations are generated (total number is given in the first column) and used to simulate the model. Initial conditions in all combinations are: $S(0) = 301$, $P(0) = 0$ and $E(0) = 120$.

Comb.	$\mathbf{k_1}$	$\mathbf{k_3}$	$\mathbf{k_5}$
20	5: $[75, 105]$	2: $[1.4, 1.7]$	2: $[0.93, 1.2]$
100	5: $[75, 105]$	10: $[1, 1.9]$	2: $[0.93, 1.2]$
500	5: $[75, 105]$	10: $[1, 1.9]$	10: $[0.8, 1.4]$
2500	25: $[40, 140]$	10: $[1, 1.9]$	10: $[0.8, 1.4]$

Table A2: Parameters explored for the p53 system. For each set of parameter combinations, the possible values of a rate parameter are represented by a vector with equidistantly spaced values in the interval given in square brackets. The number of elements in the vector is given before the interval. The final parameter vector and initial conditions for a given combination are: $\mathbf{k} = [k_1, 0.002, k_3, 1.1, k_5, 0.96, 0.01]$, $\mathbf{x}(0) = [70, 30, 60]$.

Parameters to be scanned are set deterministically to be the same in all cases so that variations of running time arising from different reaction rates are accounted for. For both models, first moment equations of closure order $m = \{1, 2, 3, 4\}$ are generated, using normal closure for odd orders and gamma closure for even orders (to ensure nonzero terms arising from the closure method). Then these equation sets are simulated using the default ODE-solver (*ode15s*) of MEANS. In addition, 500 SSA trajectories are generated using Gillespie's direct method.

Figure A1 shows the computational times on a logarithmic scale. All moment closure data points are made up of two parts: the time required to perform the moment closure to generate the underlying ODE system; and the simulation of the ODE system for each parameter. In both examples the computational overhead of generating the equations makes the fourth order closure non-economic for a small number of parameters – a disadvantage that becomes less profound as the number of combinations increases. It is also clear that for the p53 system, which has six reactions and highly nonlinear dynamics, the complexity of the equations system makes higher order closures more expensive to simulate. This difference, on the

other hand, is more negligible for the 'simpler' Michaelis-Menten system. Also, note that, while the use of moment closures already shows advantages over the SSA, in practice an ensemble of more than 500 realisations should be used to estimate summary statistics, especially for predicting higher order moments.

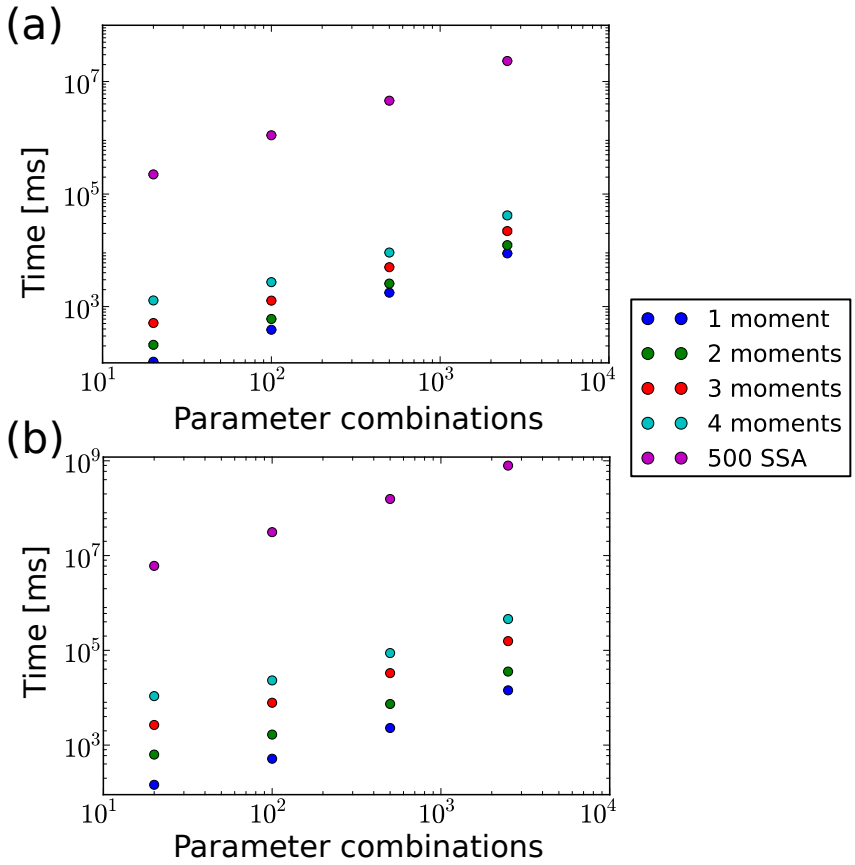


Figure A1: Computational costs of moment closure methods. The time required to simulate the **(a)** Michaelis-Menten **(b)** p53-Mdm2 oscillatory system in a parameter scan task, using MEANS with different closure orders (as indicated in the legend) and trajectories averaged from 500 runs of Gillespie's direct method. Parameters are as detailed in Tables A1 and A2.

Appendix B

In this section we include further analysis of the schematic transcription factor system we described in Chapter 5. In particular, we investigate how an inhomogeneous external signal – representing an extrinsic source of variability – alters the distribution of the system.

In order to simulate the effect of extrinsic noise, we use the same simulation procedure as in the main analysis in Section 5.3.1. However, the external signal, denoted by k_S, is not set to a fixed value for all cells. Instead, for each cell, a value is sampled from a bimodal distribution with two modes at low and high signal level,

$$k_s \sim (\mathcal{N}(1.5, 1) + \mathcal{N}(11, 2)) .$$

The signal is kept at the same level throughout the simulation, and a total of 10,000 cells are simulated. Figure B1 shows the distribution after 80 time-units. It is clear that, unlike in the noise-free case, a mixture of the two steady-states is present, and hence the population exhibits heterogeneous cell fate (i.e. both pluripotent and differentiated cells if evaluated according to stem cell marker level).

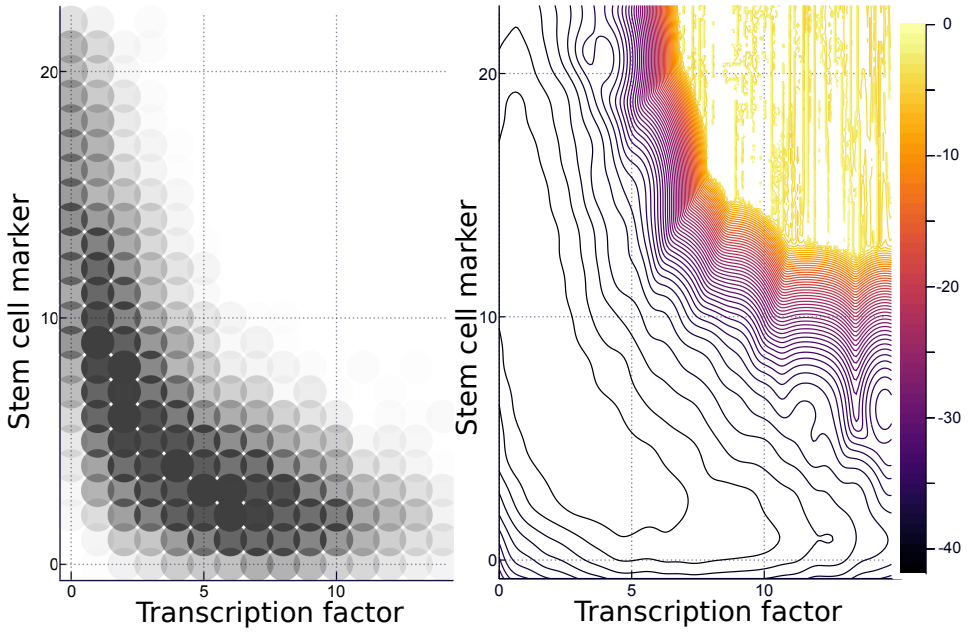


Figure B1: Extrinsic noise in the transcription factor model. Scatter plot of the probability distribution (left) and the corresponding landscape (right) of final expression in a population of cells being activated by inhomogeneous signal. The outlined distribution on the left is discrete due to the low number of molecules considered.

Appendix C

In this section we include further figures supporting our analysis of baseline p53 regulation using single cell measurements, detailed in Chapter 6. We present control simulations comparing stochastic simulation techniques, further distance measures not detailed in the main text and results of the model selection and inference algorithm on a secondary data set.

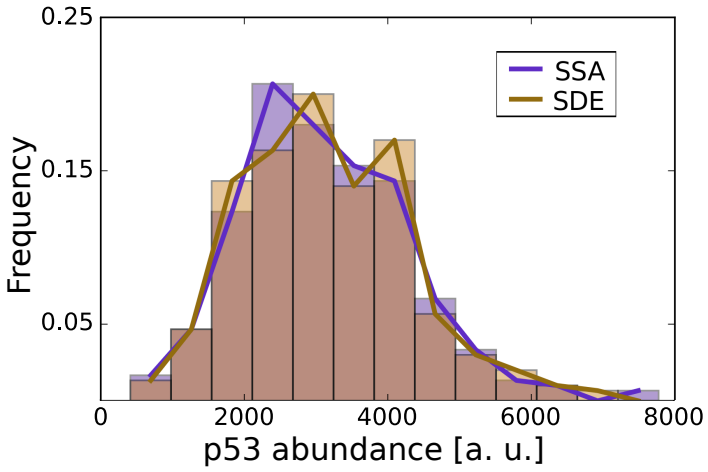


Figure C1: Comparison of simulation techniques. Steady-state populations of p53 expression levels in MCF7 cells simulated with the SSA (purple) or using a SDE formalism (gold). The rate of translation is reduced by three orders of magnitude to aid the exact stochastic simulations, which otherwise require infeasible time and memory.

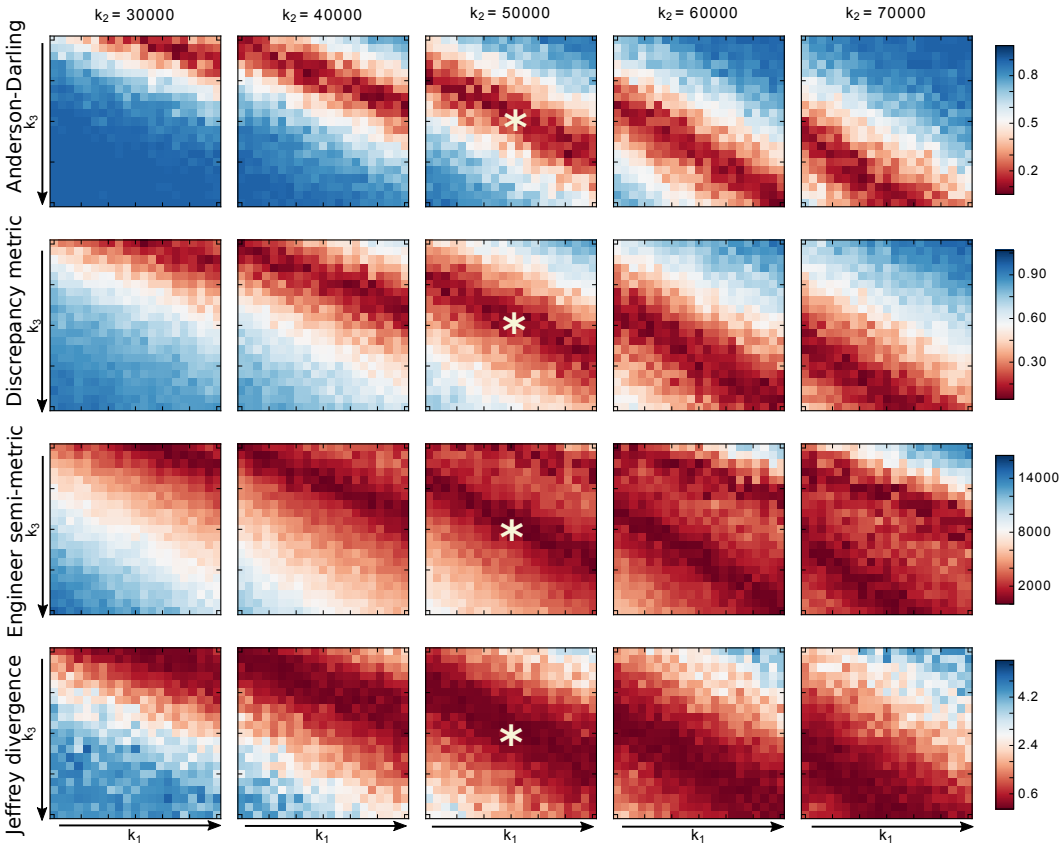


Figure C2: Evaluation of further distance measures. Colormap of distance values computed with different measures for a range of k_1, k_3 and k_2 values. All rows are normalised to the same colour bar shown at the right end of the row. The true values used in the simulated dataset are indicated by white asterisk.

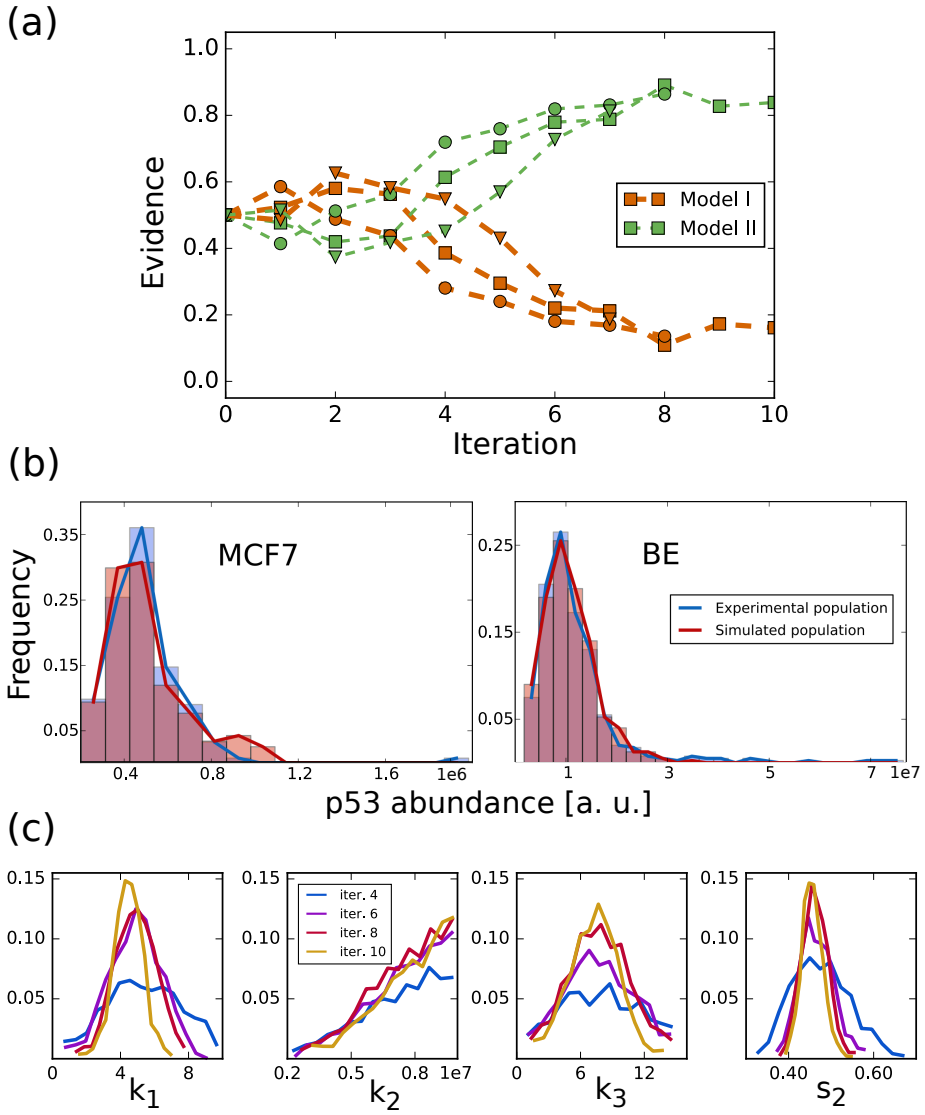


Figure C3: Results of computational analysis on Data set II. (a) Dark orange and green curves show evidence for the transcription- and the degradation-based model, different markers denote replicates with different prior parameter distributions used at initialising the algorithm. **(b)** Histograms of the distribution of p53 abundance (measured as arbitrary fluorescence units) in (left) MCF7 and (right) BE cells. Blue distributions are measured experimentally, red distributions are simulated during inference using the best fitting parameter combination. α values of Kolmogorov-Smirnov distance are $\alpha = 0.575$ (MCF7) and $\alpha = 0.51$ (BE). **(c)** Posterior parameter distributions after iterations 4, 6, 8 and 10, as indicated in legend.

Bibliography

[1] Huff, D.: How to lie with statistics. Norton, New York (1954).

[2] Delbrück, M.: Statistical fluctuations in autocatalytic reactions. *The Journal of Chemical Physics*, 8(1):120 (1940).

[3] Fedoroff, N. and Fontana, W.: Genetic networks. Small numbers of big molecules. *Science*, 297(5584):1129 (2002).

[4] Wilkinson, D. J.: Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews: Genetics*, 10(2):122 (2009).

[5] Brown, R.: A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. (1828).

[6] Paulsson, J.: Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157 (2005).

[7] Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S., and Elowitz, M. B.: Gene regulation at the single-cell level. *Science*, 307(5717):1962 (2005).

[8] Eldar, A. and Elowitz, M. B.: Functional roles for noise in genetic circuits. *Nature*, 467(7312):167 (2010).

[9] Thattai, M. and van Oudenaarden, A.: Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614 (2001).

[10] Swain, P. S., Elowitz, M. B., and Siggia, E. D.: Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795 (2002).

[11] Rao, C. V., Wolf, D. M., and Arkin, A. P.: Control, exploitation and tolerance of intracellular noise. *Nature*, 420(6912):231 (2002).

[12] Kaern, M., Elston, T. C., Blake, W. J., and Collins, J. J.: Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews: Genetics*, 6(6):451 (2005).

[13] Vardi, N., Levy, S., Assaf, M., Carmi, M., and Barkai, N.: Budding yeast escape commitment to the phosphate starvation program using gene expression noise. *Current Biology*, 23(20):2051 (2013).

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

[15] Lakatos, E., Ale, A., Kirk, P. D. W., and Stumpf, M. P. H.: Multivariate moment closure techniques for stochastic kinetic models. *The Journal of Chemical Physics*, 143(9):094107 (2015).

[16] Fan, S., Geissmann, Q., Lakatos, E., Lukauskas, S., Ale, A., Babtie, A. C., Kirk, P. D., and Stumpf, M. P. H.: MEANS: python package for Moment Expansion Approximation, iNference and Simulation. *Bioinformatics* (2016).

[17] Lakatos, E. and Stumpf, M. P. H.: Control mechanisms for stochastic biochemical systems via computation of reachable sets. *Royal Society Open Science*, 4(8) (2017).

[18] Lakatos, E., Salehi-Reyhani, A., Barclay, M., Stumpf, M. P. H., and Klug, D. R.: Protein degradation rate is the dominant mechanism accounting for the

differences in protein abundance of basal p53 in a human breast and colorectal cancer cell line. *PLoS One*, 12(5):e0177336 (2017).

[19] Levine, A. J.: p53, the cellular gatekeeper for growth and division. *Cell*, 88(3):323 (1997).

[20] Geva-Zatorsky, N., Rosenfeld, N., Itzkovitz, S., Milo, R., Sigal, A., Dekel, E., Yarnitzky, T., Liron, Y., Polak, P., Lahav, G., and Alon, U.: Oscillations and variability in the p53 system. *Molecular Systems Biology*, 2:2006.0033 (2006).

[21] Hill, A. V.: The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *Journal of Physiology*, 40:iv (1910).

[22] McQuarrie, D. A.: Stochastic approach to chemical kinetics. *Journal of Applied Probability*, 4(3):413 (1967).

[23] Gillespie, D. T.: A rigorous derivation of the Chemical Master Equation. *Physica A*, 188(1-3):404 (1992).

[24] Rao, C. V. and Arkin, A. P.: Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *The Journal of Chemical Physics*, 118(11):1 (2003).

[25] Schnoerr, D., Sanguinetti, G., and Grima, R.: Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. *Journal of Physics A: Mathematical and Theoretical*, 50(9):093001 (2017).

[26] Gillespie, D. T.: Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340 (1977).

[27] Gillespie, D. T.: Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716 (2001).

[28] Spill, F., Maini, P. K., and Byrne, H. M.: Optimisation of simulations of stochastic processes by removal of opposing reactions. *The Journal of Chemical Physics*, 144(8):084105 (2016).

[29] Salis, H. and Kaznessis, Y.: Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. *The Journal of Chemical Physics*, 122(5):054103 (2005).

[30] Van Kampen, N.: Stochastic processes in physics and chemistry. North-Holland Personal Library. Elsevier Science (2011).

[31] Komorowski, M., Mikisz, J., and Stumpf, M. P. H.: Decomposing noise in biochemical signaling systems highlights the role of protein degradation. *Biophysical Journal*, 104(8):1783 (2013).

[32] Komorowski, M., Finkenstadt, B., Harper, C., and Rand, D.: Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics*, 10(1):343 (2009).

[33] Wallace, E. W. J., Gillespie, D. T., Sanft, K. R., and Petzold, L. R.: Linear noise approximation is valid over limited times for any chemical system that is sufficiently large. *IET Systems Biology*, 6(4):102 (2012).

[34] Gómez-Uribe, C. A. and Verghese, G. C.: Mass fluctuation kinetics: Capturing stochastic effects in systems of chemical reactions through coupled mean-variance computations. *The Journal of Chemical Physics*, 126(2):024109 (2007).

[35] Ullah, M. and Wolkenhauer, O.: Investigating the two-moment characterisation of subcellular biochemical networks. *Journal of Theoretical Biology*, 260(3):340 (2009).

[36] Ale, A., Kirk, P., and Stumpf, M. P. H.: A general moment expansion method for stochastic kinetic models. *The Journal of Chemical Physics*, 138(17):174101 (2013).

[37] Andrews, S. S. and Bray, D.: Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. *Physical Biology*, 1(3-4):137 (2004).

[38] Itô, K.: On stochastic differential equations. Memoirs of the American Mathematical Society. American Mathematical Society, New York, NY (1951).

[39] Papoulis, A.: Probability, random variables and stochastic processes. McGraw-Hill Companies, 3rd edn. (1991).

[40] Manninen, T., Linne, M.-L., and Ruohonen, K.: Developing Itô stochastic differential equation models for neuronal signal transduction pathways. *Computational Biology and Chemistry*, 30(4):280 (2006).

[41] Wilkinson, D. J.: Stochastic modelling for systems biology. Mathematical and Computational Biology. Chapman & Hall / CRC, London, UK (2006).

[42] Goodman, L. A.: Population growth of the sexes. In Mathematical Demography, vol. 6 of *Biomathematics*, 469–480. Springer Berlin Heidelberg (1977).

[43] Whittle, P.: On the use of the normal approximation in the treatment of stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 19(2):pp. 268 (1957).

[44] Gillespie, C. S.: Moment-closure approximations for mass-action models. *IET Systems Biology*, 3(1):52 (2009).

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

[46] Krishnarajah, I., Cook, A., Marion, G., and Gibson, G.: Novel moment closure approximations in stochastic epidemics. *Bulletin of Mathematical Biology*, 67(4):855 (2005).

[47] Krishnarajah, I., Marion, G., and Gibson, G.: Novel bivariate moment-closure approximations. *Mathematical Biosciences*, 208(2):621 (2007).

[48] Keeling, M. J.: Metapopulation moments: coupling, stochasticity and persistence. *Journal of Animal Ecology*, 69(5):725 (2000).

[49] Singh, A. and Hespanha, J. P.: Lognormal moment closures for biochemical reactions. In IEEE Conference on Decision and Control, 2063–2068 (2006).

[50] Singh, A. and Hespanha, J. P.: A derivative matching approach to moment closure for the stochastic logistic model. *Bulletin of Mathematical Biology*, 69(6):1909 (2007).

[51] Hausken, K. and Moxnes, J. F.: A closure approximation technique for epidemic models. *Mathematical and Computer Modelling of Dynamical Systems*, 16:555 (2010).

[52] Smadbeck, P. and Kaznessis, Y. N.: A closure scheme for chemical master equations. *Proceedings of the National Academy of Sciences*, 110(35):14261 (2013).

[53] Grima, R.: A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *The Journal of Chemical Physics*, 136(15):154105 (2012).

[54] Milner, P., Gillespie, C. S., and Wilkinson, D. J.: Moment closure based parameter inference of stochastic kinetic models. *Statistics and Computing*, 23(2):287 (2013).

[55] Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H.: Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187 (2009).

[56] Kirk, P., Thorne, T. W., and Stumpf, M. P. H.: Model selection in systems and synthetic biology. *Current Opinion in Biotechnology*, 24(4):767 (2013).

[57] Fearnhead, P., Prangle, D., Cox, M. P., Biggs, P. J., and French, N. P.: Semi-automatic selection of summary statistics for ABC model choice. *Statistical Applications in Genetics and Molecular Biology*, 13(1):67 (2014).

[58] Mead, L. and Papanicolaou, N.: Maximum-entropy in the problem of moments. *Journal of Mathematical Physics*, 25:2404 (1984).

[59] Komorowski, M., Costa, M. J., Rand, D. A., and Stumpf, M. P. H.: Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proceedings of the National Academy of Sciences*, 108(21):8645 (2011).

[60] Liepe, J., Filippi, S., Komorowski, M., and Stumpf, M. P. H.: Maximizing the information content of experiments in systems biology. *PLoS Computational Biology*, 9(1):e1002888 (2013).

[61] Silk, D., Kirk, P., Barnes, C. P., Toni, T., and Stumpf, M. P. H.: Model selection in systems biology depends on experimental design. *PLoS Computational Biology*, 10(6):e1003650 (2014).

[62] Silk, D., Kirk, P. D. W., Barnes, C. P., Toni, T., Rose, A., Moon, S., Dallman, M. J., and Stumpf, M. P. H.: Designing attractive models via automated identification of chaotic and oscillatory dynamical regimes. *Nature Communications*, 2:489 (2011).

[63] Barnes, C. P., Silk, D., Sheng, X., and Stumpf, M. P. H.: Bayesian design of synthetic biological systems. *Proceedings of the National Academy of Sciences*, 108(37):15190 (2011).

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

[65] Friedman, N., Cai, L., and Xie, X. S.: Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Physical Review Letters*, 97(16):168302 (2006).

[66] Isserlis, L.: On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):pp. 134 (1918).

[67] Crow, E. L. and Shimizu, K.: Lognormal distributions: theory and applications, vol. 88 of *Statistics, textbooks and monographs*. M. Dekker, New York (1988).

[68] Mathai, A. M. and Moschopoulos, P. G.: On a multivariate-gamma. *Journal of Multivariate Analysis*, 39:135 (1991).

[69] Furman, E.: On a multivariate gamma distribution. *Statistics & Probability Letters*, 78:2353 (2008).

[70] Harrington, H. A., Komorowski, M., Beguerisse-Díaz, M., Ratto, G. M., and Stumpf, M. P. H.: Mathematical modeling reveals the functional implications of the different nuclear shuttling rates of Erk1 and Erk2. *Physical Biology*, 9(3):036001 (2012).

[71] Schnoerr, D., Sanguinetti, G., and Grima, R.: Comparison of different moment-closure approximations for stochastic chemical kinetics. *The Journal of Chemical Physics*, 143(18):185101 (2015).

[72] Azunre, P., Gómez-Uribe, C., and Verghese, G.: Mass fluctuation kinetics: analysis and computation of equilibria and local dynamics. *IET Systems Biology*, 5:325 (2011).

[73] SymPy Development Team: SymPy: Python library for symbolic mathematics (2016).

[74] Andersson, C., Fuhrer, C., and Akesson, J.: Assimulo: A unified framework for ODE solvers. *Mathematics and Computers in Simulation*, 116:26 (2015).

[75] Hindmarsh, A. C., Brown, P. N., Grant, K. E., Lee, S. L., Serban, R., Shumaker, D. E., and Woodward, C. S.: Sundials: Suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software*, 31(3):363 (2005).

[76] Kügler, P.: Moment fitting for parameter inference in repeatedly and partially observed stochastic biological models. *PLoS One*, 7(8):e43001 (2012).

[77] Munsky, B., Trinh, B., and Khammash, M.: Listening to the noise: random fluctuations reveal gene network parameters. *Molecular Systems Biology*, 5(1):318 (2009).

[78] John, V., Angelov, I., Oencuel, A. A., and Thevenin, D.: Techniques for the reconstruction of a distribution from a finite number of its moments. *Chemical Engineering Science*, 62:2890 (2007).

[79] Hasenauer, J., Wolf, V., Kazeroonian, A., and Theis, F. J.: Method of conditional moments (MCM) for the chemical master equation : A unified frame-

work for the method of moments and hybrid stochastic-deterministic models. *Journal of Mathematical Biology* (2013).

[80] Waddington, C. H.: The strategy of the genes: A discussion of some aspects of theoretical biology. Allen & Unwin, London (1957).

[81] Li, C. and Wang, J.: Quantifying Waddington landscapes and paths of non-adiabatic cell fate decisions for differentiation, reprogramming and transdifferentiation. *Journal of The Royal Society Interface*, 10(89):20130787 (2013).

[82] Schnoerr, D., Sanguinetti, G., and Grima, R.: Validity conditions for moment closure approximations in stochastic chemical kinetics. *The Journal of Chemical Physics*, 141(8):084103 (2014).

[83] Erguler, K. and Stumpf, M. P. H.: Statistical interpretation of the interplay between noise and chaos in the stochastic logistic map. *Mathematical Biosciences*, 216:90 (2008).

[84] Percival, D. B. and Walden, A. T.: Spectral analysis for physical applications. Cambridge University Press, Cambridge, UK (1993).

[85] Golightly, A. and Wilkinson, D. J.: Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781 (2005).

[86] Golightly, A. and Wilkinson, D. J.: Bayesian parameter inference for stochastic biochemical network models using particle Markov Chain Monte Carlo. *Interface Focus*, 1(6):807 (2011).

[87] Gilbert, E. G.: The decoupling of multivariable systems by state feedback. *SIAM Journal on Control*, 7(1):50 (1969).

[88] Apgar, J. F., Toettcher, J. E., Endy, D., White, F. M., and Tidor, B.: Stimulus design for model selection and validation in cell signaling. *PLoS Computational Biology*, 4(2):e30 (2008).

[89] Khalil, A. S. and Collins, J. J.: Synthetic biology: Applications come of age. *Nature Reviews: Genetics*, 11(5):367 (2010).

[90] Elowitz, M. B. and Leibler, S.: A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335 (2000).

[91] Gardner, T. S., Cantor, C. R., and Collins, J. J.: Construction of a genetic toggle switch in Escherichia coli. *Nature*, 403(6767):339 (2000).

[92] Milias-Argeitis, A., Summers, S., Stewart-Ornstein, J., Zuleta, I., Pincus, D., El-Samad, H., Khammash, M., and Lygeros, J.: In silico feedback for in vivo regulation of a gene expression circuit. *Nature Biotechnology*, 29(12):1114 (2011).

[93] Menolascina, F., Fiore, G., Orabona, E., De Stefano, L., Ferry, M., Hasty, J., di Bernardo, M., and di Bernardo, D.: *In-Vivo* real-time control of protein expression from endogenous and synthetic gene networks. *PLoS Computational Biology*, 10(5):e1003625 (2014).

[94] Mc Mahon, S. S., Lenive, O., Filippi, S., and Stumpf, M. P. H.: Information processing by simple molecular motifs and susceptibility to noise. *Journal of The Royal Society Interface*, 12(110):0597 (2015).

[95] Filippi, S., Barnes, C. P., Kirk, P. D. W., Kudo, T., Kunida, K., McMahon, S. S., Tsuchiya, T., Wada, T., Kuroda, S., and Stumpf, M. P. H.: Robustness of MEK-ERK dynamics and origins of cell-to-cell variability in MAPK signaling. *Cell Reports*, 15(11):2524 (2016).

[96] Girard, A.: Reachability of uncertain linear systems using zonotopes. In Hybrid Systems : Computation and Control, 291–305. Springer (2005).

[97] Audemard, G., Bozzano, M., Cimatti, A., and Sebastiani, R.: Verifying industrial hybrid systems with MathSAT. *Electronic Notes in Theoretical Computer Science*, 119(2):17 (2005).

[98] Cross, E. A. and Mitchell, I. M.: Level set methods for computing reachable sets of systems with differential algebraic equation dynamics. In American Control Conference, 2008, 2260–2265 (2008).

[99] Prandini, M. and Hu, J.: Application of reachability analysis for stochastic hybrid systems to aircraft conflict prediction. *IEEE Transactions on Automatic Control*, 54(4):913 (2009).

[100] Scacchioli, A., Bayen, A. M., and Stojadinovic, B.: Assessment of uncertainty propagation in the dynamic response of single-degree-of-freedom structures using reachability analysis. *Journal of Engineering Mechanics*, 140(6) (2014).

[101] Althoff, M., Le Guernic, C., and Krogh, B. H.: Reachable set computation for uncertain time-varying linear systems. In Proceedings of the 14th International Conference on Hybrid Systems: Computation and Control, HSCC '11, 93–102. ACM, New York, NY, USA (2011).

[102] Kurzhanski, A. B. and Vályi, I.: Ellipsoidal calculus for estimation and control. Systems & control. IIASA Boston, Laxenburg, Austria (1997).

[103] Greenstreet, M. R. and Mitchell, I.: Reachability analysis using polygonal projections, vol. 1569, 103–116. Springer Berlin Heidelberg (1999).

[104] Han, Z. and Krogh, B. H.: Reachability analysis of large-scale affine systems using low-dimensional polytopes. In Hybrid Systems: Computation and Control, 287–301. Springer-Verlag, Berlin, Heidelberg (2006).

[105] Asarin, E., Dang, T., and Girard, A.: Reachability analysis of nonlinear systems using conservative approximation. In Hybrid Systems: Computation and Control, LNCS 2623, 20–35. Springer-Verlag (2003).

[106] Mitchell, I. M. and Templeton, J. A.: A toolbox of Hamilton-Jacobi solvers for analysis of nondeterministic continuous and hybrid systems. In Hybrid Systems : Computation and Control, 480–494. Springer-Verlag (2005).

[107] Althoff, M. and Krogh, B. H.: Reachability analysis of nonlinear differential-algebraic systems. *IEEE Transactions on Automatic Control*, 59(2):371 (2014).

[108] Szederkényi, G., Kovács, M., and Hangos, K. M.: Reachability of nonlinear fed-batch fermentation processes. *International Journal of Robust and Nonlinear Control*, 12(12):1109 (2002).

[109] Brim, L., Fabriková, J., Drazan, S., and Safránek, D.: Reachability in biochemical dynamical systems by quantitative discrete approximation. *Computing Research Repository*, abs/1107.5924 (2011).

[110] Halász, Á. M., Kumar, V., Imieliński, M., Belta, C., Sokolsky, O., Pathak, S., and Rubin, H.: Analysis of lactose metabolism in E. coli using reachability analysis of hybrid systems. In IEEE Proceedings – Systems Biology, 130–148 (2007).

[111] Dang, T., Le Guernic, C., and Maler, O.: Computing reachable states for nonlinear biological models. *Theoretical Computer Science*, 412(21):2095 (2011).

[112] Testylier, R. and Dang, T.: Analysis of parametric biological models with non-linear dynamics. In Proceedings of the First International Workshop on Hybrid Systems and Biology, vol. 92 of *Electronic Proceedings in Theoretical Computer Science*, 16–29. Open Publishing Association (2012).

[113] El-Samad, H., Fazel, M., Liu, X., Papachristodoulou, A., and Prajna, S.: Stochastic reachability analysis in complex biological networks. In American Control Conference, 2006, 6 – 12 (2006).

[114] Riley, D., Koutsoukos, X., and Riley, K.: Verification of biochemical processes using stochastic hybrid systems. In IEEE International Symposium on Intelligent Control, 100–105 (2007).

[115] Parise, F., Valcher, M. E., and Lygeros, J.: On the reachable set of the controlled gene expression system. In IEEE Conference on Decision and Control. Los Angeles, California, USA (2014).

[116] Fukuda, K.: From the zonotope construction to the Minkowski addition of convex polytopes. *Journal of Symbolic Computation*, 38:1261 (2004).

[117] Szederkényi, G., Lakner, R., and Gerzson, M.: Intelligent control systems: An introduction with examples. Springer Science & Business Media (2006).

[118] Aubin, J.-P. and Cellina, A.: Differential inclusions: Set-valued maps and viability theory, vol. 264 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, Heidelberg (1984).

[119] Van Loan, C.: Computing integrals involving the matrix exponential. *IEEE Transactions on Automatic Control*, 23(3):395 (1978).

[120] Lu, Y. Y.: Computing a matrix function for exponential integrators. *Journal of Computational and Applied Mathematics*, 161(1):203 (2003).

[121] Feliu, E. and Wiuf, C.: Variable elimination in chemical reaction networks with mass-action kinetics. *SIAM Journal of Applied Mathematics*, 72(4):959 (2012).

[122] Erguler, K. and Stumpf, M. P. H.: Practical limits for reverse engineering of dynamical systems: A statistical analysis of sensitivity and parameter inferability in systems biology models. *Molecular BioSystems*, 7:1593 (2011).

[123] Barnes, C. P., Silk, D., and Stumpf, M. P. H.: Bayesian design strategies for synthetic biology. *Interface Focus*, 1(6):895 (2011).

[124] Julier, S., Uhlmann, J., and Durrant-Whyte, H. F.: A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Transactions on Automatic Control*, 45(3):477 (2000).

[125] Mosconi, F., Julou, T., Desprat, N., Sinha, D. K., Allemand, J.-F., Croquette, V., and Bensimon, D.: Some nonlinear challenges in biology. *Nonlinearity*, 21(8):T131 (2008).

[126] Han, Z. and Krogh, B.: Reachability analysis of nonlinear systems using trajectory piecewise linearized models. In American Control Conference, 1505–1510 (2006).

[127] Munsky, B. and Khammash, M.: The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics*, 124(4) (2006).

[128] Ingram, P. J., Stumpf, M. P. H., and Stark, J.: Nonidentifiability of the source of intrinsic noise in gene expression from single-burst data. *PLoS Computational Biology*, 4(10):e1000192 (2008).

[129] Tabor, J. J., Levskaya, A., and Voigt, C. A.: Multichromatic control of gene expression in Escherichia coli. *Journal of Molecular Biology*, 405(2):315 (2011).

[130] Anderson, J. and Papachristodoulou, A.: On validation and invalidation of biological models. *BMC Bioinformatics*, 10(132) (2009).

[131] Ferrell, J. E., Jr: Bistability, bifurcations, and Waddington's epigenetic landscape. *Current Biology*, 22(11):R458 (2012).

[132] Glauche, I., Herberg, M., and Roeder, I.: Nanog variability and pluripotency regulation of embryonic stem cells–insights from a mathematical model analysis. *PLoS One*, 5(6):e11238 (2010).

[133] Babtie, A. C., Kirk, P., and Stumpf, M. P. H.: Topological sensitivity analysis for systems biology. *Proceedings of the National Academy of Sciences*, 111(52):18507 (2014).

[134] Pujadas, E. and Feinberg, A. P.: Regulated noise in the epigenetic landscape of development and disease. *Cell*, 148(6):1123 (2012).

[135] Feinberg, A. P., Koldobskiy, M. A., and Gondor, A.: Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Reviews: Genetics*, 17(5):284 (2016).

[136] Teramoto, Y., Takahashi, D. Y., Holmes, P., and Ghazanfar, A. A.: Vocal development in a Waddington landscape. *eLife*, 6:e20782 (2017).

[137] Sasai, M. and Wolynes, P. G.: Stochastic gene expression as a many-body problem. *Proceedings of the National Academy of Sciences*, 100(5):2374 (2003).

[138] Wang, J., Xu, L., Wang, E., and Huang, S.: The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. *Biophysical Journal*, 99(1):29 (2010).

[139] Wang, J., Zhang, K., Xu, L., and Wang, E.: Quantifying the Waddington landscape and biological paths for development and differentiation. *Proceedings of the National Academy of Sciences*, 108(20):8257 (2011).

[140] Zhou, J. X., Aliyu, M. D. S., Aurell, E., and Huang, S.: Quasi-potential landscape in complex multi-stable systems. *Journal of The Royal Society Interface*, 9(77):3539 (2012).

[141] Zhang, B. and Wolynes, P. G.: Stem cell differentiation as a many-body problem. *Proceedings of the National Academy of Sciences*, 111(28):10185 (2014).

[142] Li, C. and Wang, J.: Quantifying the landscape for development and cancer from a core cancer stem cell circuit. *Cancer Research*, 75(13):2607 (2015).

[143] Yamanaka, S.: Elite and stochastic models for induced pluripotent stem cell generation. *Nature*, 460(7251):49 (2009).

[144] Luo, Y., Lim, C. L., Nichols, J., Martinez-Arias, A., and Wernisch, L.: Cell signalling regulates dynamics of Nanog distribution in embryonic stem cell populations. *Journal of The Royal Society Interface*, 10(78):20120525 (2013).

[145] Herberg, M. and Roeder, I.: Computational modelling of embryonic stem-cell fate control. *Development*, 142(13):2250 (2015).

[146] Niwa, H., Miyazaki, J., and Smith, A. G.: Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nature Genetics*, 24(4):372 (2000).

[147] Chew, J.-L., Loh, Y.-H., Zhang, W., Chen, X., Tam, W.-L., Yeap, L.-S., Li, P., Ang, Y.-S., Lim, B., Robson, P., and Ng, H.-H.: Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Molecular and Cellular Biology*, 25(14):6031 (2005).

[148] Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S.: The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell*, 113(5):631 (2003).

[149] Chambers, I. and Smith, A.: Self-renewal of teratocarcinoma and embryonic stem cells. *Oncogene*, 23(43):7150 (2004).

[150] Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R., and Young, R. A.: Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947 (2005).

[151] Rodda, D. J., Chew, J.-L., Lim, L.-H., Loh, Y.-H., Wang, B., Ng, H.-H., and Robson, P.: Transcriptional regulation of Nanog by Oct4 and Sox2. *Journal of Biological Chemistry*, 280(26):24731 (2005).

[152] Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I. R.: Dissecting self-renewal in stem cells with RNA interference. *Nature*, 442(7102):533 (2006).

[153] Lin, T., Chao, C., Saito, S., Mazur, S. J., Murphy, M. E., Appella, E., and Xu, Y.: p53 induces differentiation of mouse embryonic stem cells by suppressing Nanog expression. *Nature Cell Biology*, 7(2):165 (2005).

[154] Takahashi, K. and Yamanaka, S.: Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663 (2006).

[155] Chickarmane, V., Troein, C., Nuber, U. A., Sauro, H. M., and Peterson, C.: Transcriptional dynamics of the embryonic stem cell switch. *PLoS Computational Biology*, 2(9):1080 (2006).

[156] Lakatos, D., Travis, E. D., Pierson, K. E., Vivian, J. L., and Czirok, A.: Autocrine FGF feedback can establish distinct states of Nanog expression in pluripotent stem cells: a computational analysis. *BMC Systems Biology*, 8:112 (2014).

[157] MacArthur, B. D., Please, C. P., and Oreffo, R. O. C.: Stochasticity and the molecular mechanisms of induced pluripotency. *PLoS One*, 3(8):e3086 (2008).

[158] MacArthur, B. D. and Lemischka, I. R.: Statistical mechanics of pluripotency. *Cell*, 154(3):484 (2013).

[159] Kalmar, T., Lim, C., Hayward, P., Muñoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J., and Martinez Arias, A.: Regulated fluctuations in Nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biology*, 7(7):e1000149 (2009).

[160] Chambers, I. and Tomlinson, S. R.: The transcriptional foundation of pluripotency. *Development*, 136(14):2311 (2009).

[161] Chickarmane, V., Olariu, V., and Peterson, C.: Probing the role of stochasticity in a model of the embryonic stem cell - heterogeneous gene expression and reprogramming efficiency. *BMC Systems Biology*, 6(98) (2012).

[162] Bessonnard, S., De Mot, L., Gonze, D., Barriol, M., Dennis, C., Goldbeter, A., Dupont, G., and Chazaud, C.: Gata6, Nanog and Erk signaling control cell fate in the inner cell mass through a tristable regulatory network. *Development*, 141(19):3637 (2014).

[163] Arias, A. M. and Hayward, P.: Filtering transcriptional noise during development: concepts and mechanisms. *Nature Reviews: Genetics*, 7(1):34 (2006).

[164] Muñoz Descalzo, S. and Martinez Arias, A.: The structure of Wntch signalling and the resolution of transition states in development. *Seminars in Cell & Developmental Biology*, 23(4):443 (2012).

[165] Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 832–837 (1956).

[166] Parzen, E.: On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 1065–1076 (1962).

[167] Segal, E. and Widom, J.: From DNA sequence to transcriptional behaviour: a quantitative approach. *Nature Reviews: Genetics*, 10(7):443 (2009).

[168] Li, G.-W., Burkhardt, D., Gross, C., and Weissman, J. S.: Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624 (2014).

[169] Okita, K., Ichisaka, T., and Yamanaka, S.: Generation of germline-competent induced pluripotent stem cells. *Nature*, 448(7151):313 (2007).

[170] Singer, Z. S., Yong, J., Tischler, J., Hackett, J. A., Altinok, A., Surani, M. A., Cai, L., and Elowitz, M. B.: Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Molecular Cell*, 55(2):319 (2014).

[171] Nicol-Benoit, F., le Goff, P., and Michel, D.: Drawing a Waddington landscape to capture dynamic epigenetics. *Biology of the Cell*, 105(12):576 (2013).

[172] Cao, Y., Liang, C., Naveed, H., Li, Y., Chen, M., and Nie, Q.: Modeling spatial population dynamics of stem cell lineage in tissue growth. *Proceedings of the Conference of the IEEE Engineering in Medicine and Biology Society*, 2012:5502 (2012).

[173] Smith, Q., Stukalin, E., Kusuma, S., Gerecht, S., and Sun, S. X.: Stochasticity and spatial interaction govern stem cell differentiation dynamics. *Scientific Reports*, 5:12617 (2015).

[174] MacLean, A. L., Kirk, P. D. W., and Stumpf, M. P. H.: Cellular population dynamics control the robustness of the stem cell niche. *Biology Open*, 4(11):1420 (2015).

[175] Toni, T. and Tidor, B.: Combined model of intrinsic and extrinsic variability for computational network design with application to synthetic biology. *PLoS Computational Biology*, 9(3):e1002960 (2013).

[176] Shahrezaei, V. and Swain, P. S.: Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256 (2008).

[177] Lenive, O., W Kirk, P. D., and H Stumpf, M. P.: Inferring extrinsic noise from single-cell gene expression data using approximate Bayesian computation. *BMC Systems Biology*, 10(1):81 (2016).

[178] Battle, A., Khan, Z., Wang, S. H., Mitrano, A., Ford, M. J., Pritchard, J. K., and Gilad, Y.: Genomic variation. Impact of regulatory variation from RNA to protein. *Science*, 347(6222):664 (2015).

[179] Jovanovic, M., Rooney, M. S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E. H., Fields, A. P., Schwartz, S., Raychowdhury, R., Mumbach, M. R., Eisenhaure, T., Rabani, M., Gennert, D., Lu, D., Delorey, T., Weissman, J. S., Carr, S. A., Hacohen, N., and Regev, A.: Dynamic profiling of the protein life cycle in response to pathogens. *Science*, 347(6226):1259038 (2015).

[180] Li, J. J., Bickel, P. J., and Biggin, M. D.: System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*, 2:e270 (2014).

[181] Ghazalpour, A., Bennett, B., Petyuk, V. A., Orozco, L., Hagopian, R., Mungrue, I. N., Farber, C. R., Sinsheimer, J., Kang, H. M., Furlotte, N., Park, C. C., Wen, P.-Z., Brewer, H., Weitz, K., Camp, D. G., II, Pan, C., Yordanova, R., Neuhaus, I., Tilford, C., Siemers, N., Gargalovic, P., Eskin, E.,

Kirchgessner, T., Smith, D. J., Smith, R. D., and Lusis, A. J.: Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genetics*, 7(6):e1001393 (2011).

[182] Kristensen, A. R., Gsponer, J., and Foster, L. J.: Protein synthesis rate is the predominant regulator of protein expression during differentiation. *Molecular Systems Biology*, 9:689 (2013).

[183] Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M.: Global quantification of mammalian gene expression control. *Nature*, 473(7347):337 (2011).

[184] Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J. C., Carr, S. A., Tabb, D. L., Coffey, R. J., Slebos, R. J. C., Liebler, D. C., and NCI CPTAC: Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382 (2014).

[185] Li, J. J. and Biggin, M. D.: Gene expression. Statistics requantitates the central dogma. *Science*, 347(6226):1066 (2015).

[186] Cancer Genome Atlas Research Network: Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609 (2011).

[187] Honda, R., Tanaka, H., and Yasuda, H.: Oncoprotein MDM2 is a ubiquitin ligase E3 for tumor suppressor p53. *FEBS Letters*, 420(1):25 (1997).

[188] Moll, U. M. and Petrenko, O.: The MDM2-p53 interaction. *Molecular Cancer Research*, 1(14):1001 (2003).

[189] Banin, S., Moyal, L., Shieh, S.-Y., Taya, Y., Anderson, C. W., Chessa, L., Smorodinsky, N. I., Prives, C., Reiss, Y., Shiloh, Y., and Ziv, Y.: En-

hanced phosphorylation of p53 by ATM in response to DNA damage. *Science*, 281(5383):1674 (1998).

[190] Siliciano, J. D., Canman, C. E., Taya, Y., Sakaguchi, K., Appella, E., and Kastan, M. B.: DNA damage induces phosphorylation of the amino terminus of p53. *Genes & Development*, 11(24):3471 (1997).

[191] Shieh, S. Y., Ikeda, M., Taya, Y., and Prives, C.: DNA damage-induced phosphorylation of p53 alleviates inhibition by MDM2. *Cell*, 91(3):325 (1997).

[192] Sakaguchi, K., Herrera, J. E., Saito, S., Miki, T., Bustin, M., Vassilev, A., Anderson, C. W., and Appella, E.: DNA damage activates p53 through a phosphorylation-acetylation cascade. *Genes & Development*, 12(18):2831 (1998).

[193] Khosravi, R., Maya, R., Gottlieb, T., Oren, M., Shiloh, Y., and Shkedy, D.: Rapid ATM-dependent phosphorylation of MDM2 precedes p53 accumulation in response to DNA damage. *Proceedings of the National Academy of Sciences*, 96(26):14973 (1999).

[194] Stott, F. J., Bates, S., James, M. C., McConnell, B. B., Starborg, M., Brookes, S., Palmero, I., Ryan, K., Hara, E., Vousden, K. H., and Peters, G.: The alternative product from the human CDKN2A locus, p14(ARF), participates in a regulatory feedback loop with p53 and MDM2. *The EMBO Journal*, 17(17):5001 (1998).

[195] Gilkes, D. M., Chen, L., and Chen, J.: MDMX regulation of p53 response to ribosomal stress. *The EMBO Journal*, 25(23):5614 (2006).

[196] Pestov, D. G., Strezoska, Z., and Lau, L. F.: Evidence of p53-dependent cross-talk between ribosome biogenesis and the cell cycle: effects of nucleolar protein Bop1 on G(1)/S transition. *Molecular Cell Biology*, 21(13):4246 (2001).

[197] Lavin, M. F. and Gueven, N.: The complexity of p53 stabilization and activation. *Cell Death & Differentiation*, 13(6):941 (2006).

[198] Stadler, C., Skogs, M., Brismar, H., Uhlén, M., and Lundberg, E.: A single fixation protocol for proteome-wide immunofluorescence localization studies. *Journal of Proteomics*, 73(6):1067 (2010).

[199] Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., and Cardona, A.: Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676 (2012).

[200] Schmittgen, T. D. and Livak, K. J.: Analyzing real-time pcr data by the comparative c(t) method. *Nature Protocols*, 3(6):1101 (2008).

[201] Liepe, J., Barnes, C., Cule, E., Erguler, K., Kirk, P., Toni, T., and Stumpf, M. P. H.: ABC-SysBio–approximate Bayesian computation in Python with GPU support. *Bioinformatics*, 26(14):1797 (2010).

[202] Liepe, J., Kirk, P., Filippi, S., Toni, T., Barnes, C. P., and Stumpf, M. P. H.: A framework for parameter estimation and model selection from experimental data in systems biology using Approximate Bayesian Computation. *Nature Protocols*, 9(2):439 (2014).

[203] Bayes, M. and Price, M.: An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions*, 53:370 (1763).

[204] Golightly, A. and Wilkinson, D. J.: Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781 (2005).

[205] Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W.: Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791 (1999).

[206] Beaumont, M. A., Zhang, W., and Balding, D. J.: Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025 (2002).

[207] Marjoram, P., Molitor, J., Plagnol, V., and Tavare, S.: Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324 (2003).

[208] Sisson, S. A., Fan, Y., and Tanaka, M. M.: Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760 (2007).

[209] Toni, T. and Stumpf, M. P. H.: Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104 (2010).

[210] Del Moral, P., Doucet, A., and Jasra, A.: Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411 (2006).

[211] Kass, R. E. and Raftery, A. E.: Bayes factors. *Journal of the American Statistical Association*, 90(430):773 (1995).

[212] Jeffreys, H.: Theory of Probability. Oxford University Press (1961).

[213] Massey, F. J.: The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68 (1951).

[214] Loewer, A., Batchelor, E., Gaglia, G., and Lahav, G.: Basal dynamics of p53 reveal transcriptionally attenuated pulses in cycling cells. *Cell*, 142(1):89 (2010).

[215] Ashkin, A., Dziedzic, J. M., Bjorkholm, J. E., and Chu, S.: Observation of a single-beam gradient force optical trap for dielectric particles. *Optics Letters*, 11(5):288 (1986).

[216] Ambrose, E. J.: A surface contact microscope for the study of cell movements. *Nature*, 178(4543):1194 (1956).

[217] Axelrod, D.: Cell-substrate contacts illuminated by total internal reflection fluorescence. *The Journal of Cell Biology*, 89(1):141 (1981).

[218] Salehi-Reyhani, A., Kaplinsky, J., Burgin, E., Novakova, M., deMello, A. J., Templer, R. H., Parker, P., Neil, M. A. A., Ces, O., French, P., Willison, K. R., and Klug, D.: A first step towards practical single cell proteomics: a microfluidic antibody capture chip with TIRF detection. *Lab on a Chip*, 11(7):1256 (2011).

[219] Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M.: Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology*, 4(9):117 (2003).

[220] Gedeon, T. and Bokes, P.: Delayed protein synthesis reduces the correlation between mRNA and protein fluctuations. *Biophysical Journal*, 103(3):377 (2012).

[221] Lane, D. P.: p53, guardian of the genome. *Nature*, 358(6381):15 (1992).

[222] Candeias, M. M., Malbert-Colas, L., Powell, D. J., Daskalogianni, C., Maslon, M. M., Naski, N., Bourougaa, K., Calvo, F., and Fåhraeus, R.: P53 mRNA controls p53 activity by managing Mdm2 functions. *Nature Cell Biology*, 10(9):1098 (2008).

[223] Kohn, K. W.: Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Molecular Biology of the Cell*, 10(8):2703 (1999).

[224] Choong, M. L., Yang, H., Lee, M. A., and Lane, D. P.: Specific activation of the p53 pathway by low dose actinomycin D: a new route to p53 based cyclotherapy. *Cell Cycle*, 8(17):2810 (2009).

[225] Ashcroft, M., Kubbutat, M. H. G., and Vousden, K. H.: Regulation of p53 function and stability by phosphorylation. *Molecular and Cellular Biology*, 19(3):1751 (1999).

[226] Rotter, V., Abutbul, H., and Ben-Ze'ev, A.: P53 transformation-related protein accumulates in the nucleus of transformed fibroblasts in association with the chromatin and is found in the cytoplasm of non-transformed fibroblasts. *The EMBO Journal*, 2(7):1041 (1983).

[227] Liang, S.-H. and Clarke, M. F.: Regulation of p53 localization. *European Journal of Biochemistry*, 268(10):2779 (2001).

[228] Inoue, T., Wu, L., Stuart, J., and Maki, C. G.: Control of p53 nuclear accumulation in stressed cells. *FEBS Letters*, 579(22):4978 (2005).