UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

INFORMATION ENGINEERING DEPARTMENT

MASTER DEGREE COURSE IN CONTROL SYSTEMS ENGINEERING

MODELING STOCHASTIC DYNAMICS FOR RATIO CONTROL OF BACTERIA POPULATION

SUPERVISOR

Prof. Luca Schenato

CO-SUPERVISOR

Prof. Jean-Charles Delvenne

GRADUAND

Sara Petrelli

ACADEMIC YEAR

2023-2024

GRADUATION DAY

2 December 2024

Abstract

In recent years, synthetic biology has seen significant advancements. It represents a novel research area which combines engineering and biology proposing to create systems that do not exist in nature, by either synthesizing novel organisms or inserting artificial genetic circuits into living cells. Designing novel artificial genetic systems requires a suitable mathematical model enabling the study, analysis and simulation of the biological system before implementing *in vivo* experiments. Therefore, the first part of this thesis explores the wide range of possible modeling and simulation techniques, highlighting their differences through well-known examples. It focuses especially on realizing stochastic simulations of the genetic toggle switch, starting from its simplified deterministic description.

Among the current key challenges in synthetic biology emerges the co-existence of multiple populations, in order to distribute the workload and the different functionalities. While several strategies exist to regulate the population sizes, different growth and division rates pose significant difficulties. Therefore, an innovative alternative approach considers only one population made of two phenotypically distinct subgroups, where each cell can change its phenotype under the control of an external input. In the literature two controllers have been proposed able to balance the numbers of these two subgroups, whose effectiveness has been validated using an Advanced Agent-Based Cell Simulator. Our work evaluates their performance on a stochastic model, yielding novel and different, yet at the same time promising and realistic, results. Further study is essential to refine the model and better understand the two approaches; however, we believe this thesis already provides a significant basis for the research in this area.

Sommario

Negli ultimi anni, la biologia sintetica ha visto importanti progressi. Questa rappresenta una nuova area di ricerca che combina l'ingegneria e la biologia, proponendo di creare sistemi che non esistono in natura, sia sintetizzando nuovi organismi, sia inserendo circuiti genetici costruiti artificialmente in cellule già esistenti. La realizzazione di nuovi sistemi genetici artificiali richiede però un modello matematico adeguato che ne permetta lo studio, l'analisi e la simulazione ancor prima di implementare l'esperimento *in vivo*. Dunque, la prima parte di questa tesi esplora la vasta gamma di possibili tecniche di modellazione e simulazione, evidenziandone le differenze attraverso esempi comunemente noti. In particolare, la tesi si concentra sulla realizzazione di simulazioni stocastiche dell'interruttore (toggle switch) genetico, a partire dalla sua descrizione deterministica semplificata.

Tra le principali attuali sfide nella biologia sintetica emerge quella riguardo alla coesistenza di più popolazioni, che ha l'obiettivo di distribuire il carico di lavoro e le diverse funzionalità. Sebbene esistano diverse strategie per controllare le dimensioni delle popolazioni, le velocità di crescita e divisione diverse introducono ulteriori importanti difficoltà. Pertanto, un approccio alternativo e innovativo è di considerare un'unica popolazione composta da due sottogruppi fenotipicamente distinti, dove ogni cellula può cambiare il suo fenotipo sotto il controllo di un input esterno. Nella letteratura sono stati proposti due controllori in grado di bilanciare i numeri di questi due sottogruppi, la cui efficacia è stata validata utilizzando un Advanced Agent-Based Cell Simulator. Il nostro lavoro valuta le performance su un modello stocastico, ottenendo risultati nuovi e differenti, ma allo stesso tempo promettenti e realistici. Ulteriori studi sono necessari per perfezionare il modello e comprendere meglio i due approcci; tuttavia, riteniamo che questa tesi fornisca già una base significativa per la ricerca in questo campo.

Acknowledgements

A special and sincere thank you goes to the professors who followed and supported me throughout this thesis. Your knowledge and expertise were incredibly helpful in guiding me to understand an almost completely new topic. I'm deeply grateful for your help, which was not only professional, but your humanity and kindness made this work easier.

To Professor Luca Schenato, my supervisor, I wish to sincerely thank you for your constant patience, availability, advice and empathy shown in this long and challenging project, especially in the last month when time has shortened. Your dedication, interest in the topic, the time you spent discussing and reasoning about it, have really motivated me.

To Professor Jean-Charles Delvenne, who guided me during the initial stages of this thesis when I was feeling lost, I am sincerely grateful for your passion, infinite patience and kindness in always answering my questions and helping me clarifying all my doubts. Thank you for support and guidance which always made me feel reassured.

Contents

List of Figures

Chapter 1

Introduction

Synthetic biology is a recent fast developing field at the border between engineering and biology. From engineering, it requires a suitable model and efficient simulations of biological systems, while from biology, it demands a deep understanding of the system at hand and of the reactions happening inside it. By combining these approaches, synthetic biology aims to engineer biological systems with new or enhanced functionalities. In particular, this is achieved by embedding artificial genetic circuits into living cells (such as bacteria, yeast or fungi) to modify the natural behavior of the biological system, controlling either basic functionalities of the cell, either desired additional productions of proteins or chemical compounds. This is achieved by acting on gene expression, transcription and translation processes, or on molecular interactions [1, 2].

Synthetic biology is a broad interdisciplinary field involving not only engineering and biology, but also chemistry, physics, medicine, philosophy, and many other subjects. Indeed, it can have various applications, ranging from health treatments [3–5] to bioremediation [6], or from information processing [7, 8] to production of biofuels and drugs [9]. Therefore, it has the potential for future significant developments in crucial topics such as environmental sustainability, waste material conversion in new forms of energy, pharmaceutical applications (for instance by providing ways to restore antibiotic sensitivity in bacteria or treating a particular disease), and innovative industrial solutions. As a result, synthetic biology will become an increasingly important and requested field in the future.

1.1 Motivation and State of the Art

The novelty of synthetic biology lies on its buttom-up and modular approach. Indeed, it proposes to start from elementary components such as genes, promoters, ribosomes, and proteins, designed them as needed and finally assembly them to build genetic circuits realizing the desired functionality. Moreover, when addressing a problem, synthetic biology proposes to follow the so called

Design-Build-Test-Learn (DBTL) cycle [10], in which biologists design and build the genetic circuits or the necessary biological components, then test them and analyze the results to learn from them and design enhanced versions of the system. In this process the importance of having an accurate model and simulation becomes clear, especially before building the biological system. Indeed, it provides a proper analysis of the system and of the outcomes of proposed solutions before starting the laboratories experiments. This approach not only allows to save time and money but provides also a detailed knowledge of the system.

Therefore, the first part of the thesis focuses on the modeling, analysis and simulation techniques present in the literature, including both complex but accurate models, and simple but deterministic frameworks. The model allows to study the effects of inserting a new genetic circuit into a cell, ensuring its successful implementation. Indeed, it helps to determine whether the cell exhibits metabolic burden, the circuit is truly effective, and if and how it is interacting with vital cell processes. At the core of biological systems, determining its functionalities, there is gene expression, which is governed by chemical reactions. Thus, the goal becomes to identify a proper mathematical way to model them; the choice of the model will depend on the context and application, obtaining models less or more suitable than others. The main difference between the various approaches regard the consideration or disregard of the noise. This is fundamental to account for potential different behaviors of identical cells, inevitable randomness of molecular collisions, or insufficient and imprecise knowledge of parameters describing chemical reactions.

The most accurate description is the one taking into account all the species and reactions involved in a certain process, together with their inherent randomness. This model is provided by the Chemical Master Equation [11], describing the time evolution of the probability of having a certain number of molecules at every time t. Although this is the most accurate description, it can become heavy and complex, and thus very difficult to handle. Therefore, researchers have looked for alternative approximated models. Among these, the Chemical Langevin Equation [12, 13] (or equivalently the Fokker-Planck Equation [14, 15]) represents a middle ground between stochastic and deterministic models, simplifying the Chemical Master Equation, but still accounting for the noise as gaussian white noise. On the other hand, Chemical Kinetics or Reaction Rate Equations deterministically describe the evolution of the molecules concentration.

The model is necessary to perform a detailed analysis and realize the simulations, which show outcomes supposed to happen in a real scenario. This is then useful to prepare and correctly set up *in vivo* experiments. Moreover, simulations can be very useful to ease the analysis especially when the model is complex or not trivial to interpret, which often happens with the Chemical Master Equation. Therefore, researchers have focused on this in particular, providing different techniques to perform the simulations. The standard approach is the one proposed by D.T. Gillespie in [16],

which applies the Monte Carlo theory to generate realizations of the Chemical Master Equation. Different versions and new approaches have been proposed later by Gillespie himself and other authors. Some of them proposed alternatives to determine the timing and type of reactions, others optimized the algorithm to reduce its computational time [17, 18]. On the other hand, new techniques have been proposed, among which there is the τ-leaping approximation [19], which still implements the Chemical Master Equation, but is valid under specific hypotheses that allow some reactions to be grouped together within a defined time interval, resulting in a simplified algorithm. However, this arises new challenges, hence also in this case different versions are possible; some of them account for a varying time interval [19–21], others for a realistic outcome (avoiding negative numbers of molecules) [22–24]. Finally, for the Chemical Langevin Equation easier simulation methods have been implemented. The most common one is the Euler-Maruyama [25], implementing exactly a discretized version of the Chemical Langevin Equation. The table below summarizes the possible modeling and corresponding simulation techniques.

Mathematical Model	**Simulation tool**	**Main Characteristics**
Chemical Master Equation	Gillespie's SSA - - - - - - - - - - - - - - - - τ-leaping approximation	Stochastic Exact description
Chemical Langevin Equation	Euler-Maruyama Algorithm	Stochastic (gaussian white noise)
Fokker-Planck Equation	Finite volume method	Stochastic (gaussian white noise)
Reaction Rate Equation	ODE's solvers	Deterministic Average description

Table 1.1: Summary table on modeling and simulation methods.

When engineering new circuits it is important to study and analyze them, but also to control their effects on the host cells, in particular when circuits become more and more complex they can lead to metabolic burden, or competition of common resources. To overcome these limitations a common solution is to distribute the workload among multiple populations, so that each population has a specific role and function. An explicative image is the reported below. Two enzymatic reactions are shown, the first one transforms the substrate S in an intermediate I and the second one transforms this into a product P. If both of them happen in the same cell, then they will compete for common resources and if for example the first enzyme has preference over the second one, then there would be an accumulation of the intermediate. Whereas if the reactions are distribute in two different cells, these issues are solved.

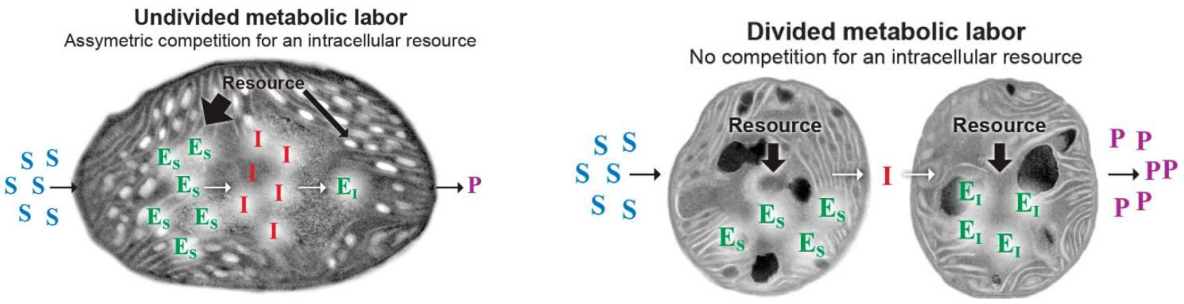
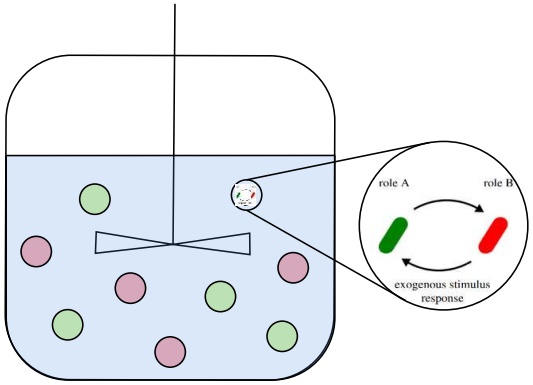


Figure 1.1: Divided and undivided metabolic labor in two steps enzymatic reaction [26].

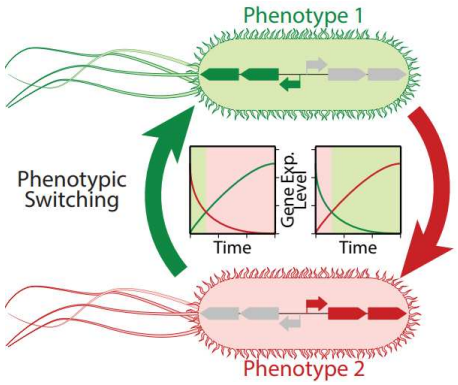
This simple example shows the importance of employing more than one cell population, and hence motivated biologists to consider co-cultures or small consortia where individual populations work together to accomplish a desired output. Microbial consortia represent a rapidly advancing strategy able to overcome limitations in efficiency and robustness that are often seen in monoculture systems. There exists a wide range of applications involving these consortia, which spans from industrial processes to mathematical analysis. For example, in [27] a synthetic ecosystem resembling canonical predator–prey systems has been constructed using two Escherichia coli populations which regulate each other's gene expression and survival. Using this they were able to reproduce extinction, coexistence and oscillatory dynamics typical of the predator and prey populations. Similarly, in [28] oscillations in population levels are generated cultivating together two distinct cell types. Or in [29] multiple logical functions are implemented using combinations of yeast strains. Another particularly innovative application is in the field of Control Systems, it implements multicellular feedback control, realizing typical electronic-based control devices using cell populations and quorum sensing molecules to close the feedback loop. In particular, in [30] they made use of two controller populations to activate or inactivate a third population endowed with a toggle switch; while in [31] and [32] the different populations were used to create a modular P, PI, PD controller in order to regulate the gene expression of a third population. On a different level, there are studies employing microbial consortia to perform bioprocesses [26], or to enhance the production of complex products and optimize metabolic processes across the system [33, 34].

Microbial consortia can be widely used in many other applications due to their ability to mitigate competition for cellular resources, reduce metabolic burdens, perform efficient and stable task division, and promote compartmentalization. However, guaranteeing the stable co-existence of multiple populations requires to regulate their relative sizes, preventing faster growing species from eliminating the slower ones, as the Competitive Exclusion Principle predicts [35]. This problem has been defined in [36] as the Ratiometric Control Problem, having the goal of achieving and maintaining at a certain desired ratio the sizes of the populations in the consortium, despite differences in their growth rates, noise and perturbations. In the literature various solutions have been proposed to regulate the numbers of competitive species controlling the growth and death

rates [37–40], or in general to achieve a desired ration between the concentrations of two microbial populations, while guaranteeing their survival and fast convergence dynamics [41–43]. However, most of these control solutions require to insert additional genetic circuits into the cells, both for establishing communication between the different populations, and both for embedding the entire feedback into every cell to make it able to sense the relative size of all the populations and respond appropriately. Therefore, an innovative and recent solution [44] has been introduced to solve these problems, and additionally to avoid different growth rates typical of different cell populations. This solution proposes indeed to consider a unique population of cells, dividing the labor among two subgroups within this population. Each group will hence have a specific role, identified by a phenotype that can be changed in response to an external stimulus, such as light or injection of inducer molecules (Fig.1.2). This is made possible by means of a bistable memory element, whose states determine the group to which each cell belongs, determining the role assigned to it. Considering one single population eliminates the risk of extinction of one of the subgroups, because cells can switch phenotype to restore the lost group; moreover, this approach allows for online changing of the desired steady-state ratio. Therefore, this represents a highly promising strategy, even though it undoubtedly brings new challenges to accomplish, in particular an accurate model of the system is necessary.



(a) One single population, where each cell can carry out different roles(red or green), in response to exogenous stimuli.



(b) Phenotypic switching inside a cell [45].

Figure 1.2: Ratiometric control with single population.

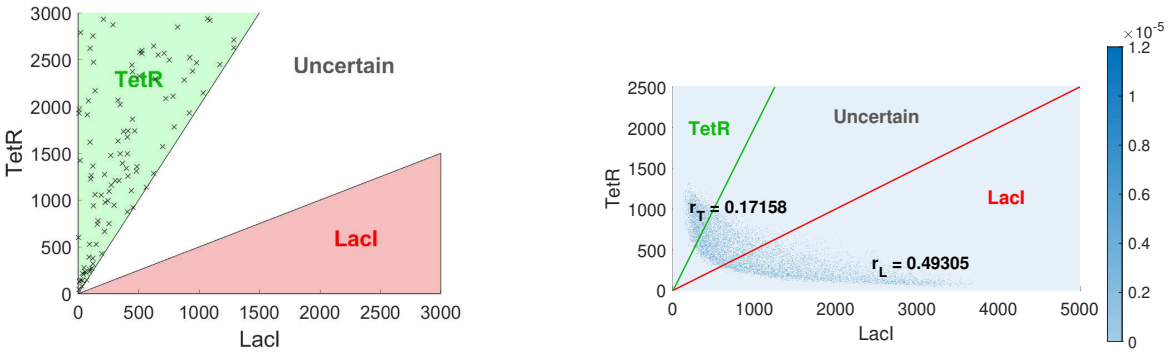
1.2 Objectives and Main Results

Synthetic biology highly relies on the presence of a suitable model and appropriate simulation techniques. Therefore the first objective of this thesis is to provide a comprehensive guide of the main modeling and simulation methods presented in the literature, in order to state their major differ-

ences and validity assumptions. In particular, we focus on the importance of accounting for the noise and on finding an efficient way to simulate the resulting stochastic model. Then we propose to apply this study to the interesting case of ratiometric control problem using a single population as discussed in [44]. In the paper, a deterministic model of the toggle switch is used, though we believe that, especially in this case, considering the noise is essential. Therefore, the second purpose of this work is to develop an efficient and reliable method for stochastically simulating and analyzing the genetic toggle switch, and finally to test the proposed controllers on this new model.

Traditional deterministic approaches describe the system's average behavior, which is valid when there is a large number of molecules. However, even in such cases, these approaches fail to capture the full complexity and richness of the system. Stochastic models can therefore be necessary also in simple cases to describe the variability of the system, as we found for birth-death or compound formation processes. The thesis focuses on three simulation techniques: Gillespie's Stochastic Simulation Algorithm and τ-leaping approximation for the Chemical Master Equation, and Euler-Maruyama method for the Chemical Langevin Equation. We show how the latter can be more efficient than the others, though relying on stronger assumptions that may not always be true. Indeed, to show its application we had to properly tune the parameters in order to satisfy the mentioned hypothesis, however, this would not be feasible in real-world scenarios. For instance, when we considered the genetic toggle switch, we found that this method was not applicable due to low molecular numbers. Whereas, both Gillespie's Stochastic Simulation Algorithm and τ-leaping approximation have been used to simulate the system. However, they require more computational time and τ-leaping can be challenging to implement due to some particular scenarios that one has to take into account. Finally, we wish to emphasize that we developed and tested a method, usually used with Reaction Rate Equations, to stochastically simulate a reduced model of the toggle switch. This represents an innovative and powerful method, which could significantly change the approach to biological stochastic simulations. Indeed, it allows to consider a simplified system bypassing the problem of knowing precisely all the reactions and the corresponding parameters.

The difference between deterministic and stochastic frameworks is made evident by the stability analysis we conducted on the toggle switch. In particular, when using the deterministic model it is possible to perform a bifurcation analysis, that is a study of the quality, quantity and position of the equilibria depending on a parameter, a virtual input in our case. We found that there exists three possible cases, two in which the system is monostable with equilibrium in one or the other stable state of the toggle switch, and one in which it is bistable. Chosen an input and the initial conditions, then the deterministic model predicts that the system will end up in a precise point of the phase space. Whereas, considering the noise, we found that this does not hold anymore, and it is necessary to talk about areas in which it is probable to find the state, rather than exact equilibria. The image below well shows it:

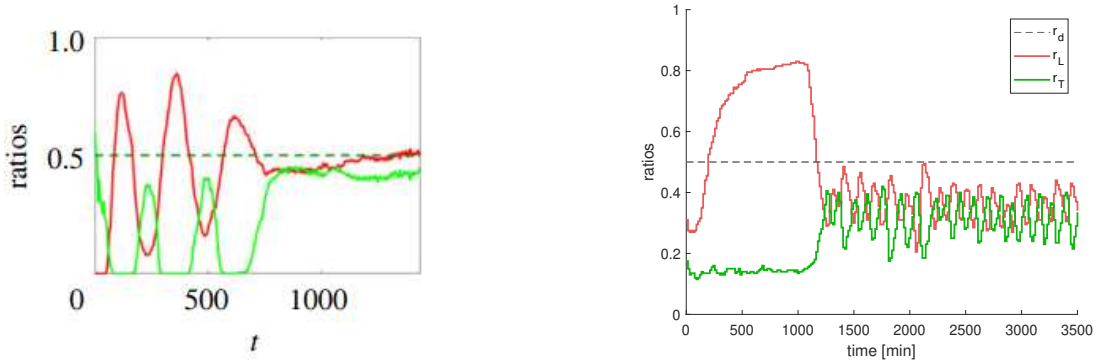


(a) (b)

Figure 1.3: Probabilistic analysis of toggle switch, for inputs providing bistability.(a) Initial conditions.(b) Heat map of probability density function.

On the left, the initial conditions of the toggle switch are shown, all within the green region, for which the deterministic solutions are ensured to remain there, converging to the equilibrium inside it. Though, the image on the right shows the time that the simulations spend in each state; from it we see that a significant part of the trajectories go out of the green region. This proves the importance of considering a stochastic model, but shows also its increased complexity. It is important to highlight that the heat map on the right has been obtained in a precise way we developed in order to return a easily readable interpretation of the simulations, which otherwise can be difficult to comprehend in some cases.

Finally, we have used the model described to validate the relay controller proposed in [44] to solve ratiometric control problem. We realized the simulations in a significantly different way, taking into account the intrinsic noise of chemical reactions, but not a population and chemostat dynamics, which was considered in the paper. However, this should only add noise to the results, without significantly changing them. The images below show the comparison between the two approaches.



(a) (b)

Figure 1.4: Comparison of relay controller outcomes. (a) Result by BSim [44]. (b) Result on stochastic model.

These results show the difference between our approach and the one in [44], supporting the work of our thesis. We found that the relay controller is able to equally balance the two population groups with a similar trend to the one obtained in the paper, though, with a significant non zero steady state error in our case. Based on the study conducted in this thesis, our result seems actually more realistic because the relay controller applies alternating inputs making the system oscillate between the two monostability regions. This reasonably causes cells to spend a significant time in the so-called uncertain set, that is a set of states for which it is not possible to determine which group the cells are in.

1.3 Thesis Structure

The rest of the thesis is organized as follows:

- Chapter 2 explores the main modeling and simulation techniques presented in the literature, performing a systematic analysis with the objective of providing a way to choose the most suitable method depending on the application. Additionally it presents three application examples, showing how to practically obtain the model and perform the simulations, emphasizing the differences between the various methods.
- Chapter 3 introduces a novel method to perform stochastic simulations on a simplified model. It focuses on the example of the genetic toggle switch, proving this approach by comparing the simulation results with experimental data. This chapter also presents a detailed stability analysis of the deterministic model of the toggle switch, providing the bifurcation diagram, the nullclines analysis, and the vector field.
- Chapter 4 discusses Ratiometric Control Problem, an emerging challenge in synthetic biology. It focuses on the solution proposed by [44], validating it on the stochastic model obtained in Chapter 3. Furthermore, this chapter provides a framework to stochastically understand the system, based on the probability distribution.

Chapter 2

Modeling and Simulating Chemical Reactions

Modeling is fundamental to study and understand the system's behavior, but also to design a suitable controller using the simulations' responses, before testing it on the real system.
Mathematical models are clearly approximations of a real system, they are based on assumptions that can be more or less appropriate depending on the application, citing George E. P. Box: *"All models are wrong, but some are useful"* [46].
In particular, in the context of Systems Biology, the goal is to model chemical reactions happening inside cells, that are intrinsically noisy biochemical reactors; indeed, identical cells exposed to the same environmental conditions can show significantly different behaviors [47]. This inherent stochasticity can be due to various factors, such as: small number of molecules involved in fundamental cells' processes (hence fluctuations in molecule numbers can have significant impacts), random molecular collisions (that govern which reactions occur and in what order), variability in cell division (because molecules are randomly partitioned between daughter cells), or multiple steps complex processes (possibly producing consecutive layers of noise). Nevertheless, fluctuations are not always undesirable, cells can exploit them for instance to introduce diversity into a population, to guarantee adaptability in changing environment, or to successfully respond to sudden stresses [47, 48]. However, the noise not only arises from the cell system, it is intrinsic to molecular interactions since their occurrence depends on the random chance of two molecules being in the same position and colliding with enough energy.

One of the goals of this thesis is to provide a guide of the main modeling techniques presented in the literature, with the objective of identifying the most suitable approach depending on the applications and assumptions. It is then essential to understand when fluctuations can be neglected or when they play a decisive role, finding the best way to include them in the model when necessary. Furthermore, including noise in the model can also be seen as a way to account for its inevitable

imprecision since kinetic parameters can only be estimated from experimental data through statistical methods [49].

We chose to implement all the simulations in Matlab since it is widely accessible, can be used very easily and allows a deep understanding of the algorithms, as everything is directly implemented within it; whereas, when algorithms are already pre-implemented it is not really possible to customize them.
On the other hand, using available software could ease the simulation realizations, and in some cases allow to simulate more complete systems, accounting also for additional aspects like environmental factors, daughter cells dynamics or spatial geometry. This is true for example for BSim [50], although it does not implement stochastic models. Whereas, if one wants to include stochasticity, then BNSim [51] or COPASI could be a valid option, even if they probably lack some realistic environmental features [50].

The following tables summarize the main modeling and simulation methods, which will be described in the next sections.

Mathematical Model	**Dependent variable**	**Stochastic process**
Chemical Master Equation	Probability $P(X(t) = x)$	Discrete
Chemical Langevin Equation	Number of molecules $X(t)$	Continuous
Fokker-Planck Equation	Probability $P(Y(t) = x)$	Continuous
Reaction Rate Equation	Concentration of molecules $y(t)$	Not stochastic

Table 2.1: Summary table on modeling methods.

Simulation tool	**Simulation speed**	**Available software**
Gillespie's SSA	Slow if number of molecules or reactions is high	BNSim [51], COPASI [52–54]
τ-leaping approximation	Medium	
Euler-Maruyama algorithm	Medium	BNSim [51]
Finite volume method	Medium	/
ODE's solvers	Fast	BSim [50], COPASI [52–54]

Table 2.2: Summary table on simulation methods.

2.1 Theoretical background

The following section presents the fundamental background and the main modeling techniques discussed in the literature. As general reference works, the reader can consult [11], [55] and [13], additional references will be given throughout the text.

Consider N different species or types of molecules S_i in a well-stirred system and M types of chemical reactions $\boldsymbol{R_j}$ that can occur between them. The assumption of a well-stirred environment implies that the molecules are uniformly spread in the space, hence one can look only at the number of molecules, and not at their single dynamics, ignoring spatial information. Let $\boldsymbol{X(t)}$ be the state variable whose components keep track of the number of molecules for each species, $\boldsymbol{\nu_j}$ the stoichiometric vector representing the state update caused by reaction R_j, and $\boldsymbol{a_j(X(t))}$ the propensity function associated to each reaction such that the probability of R_j taking place in $[t, t+dt)$ is given by $a_j(X(t))\,dt$.

The expression of the propensity functions should make intuitive sense. In fact, they can be justified rigorously from first principles, in particular, the following rules hold:

- Zero order: $\emptyset \xrightarrow{c_j} S_m \quad \longrightarrow \quad a_j(X(t)) = c_j.$
- First order: $S_m \xrightarrow{c_j} S_t \text{ or } \emptyset \quad \longrightarrow \quad a_j(X(t)) = c_j X_m(t).$
- Second order: $S_m + S_n \xrightarrow{c_j} S_t, \text{ with } m \neq n \quad \longrightarrow \quad a_j(X(t)) = c_j X_m(t) X_n(t).$
- Dimerization: $S_m + S_m \xrightarrow{c_j} S_t \quad \longrightarrow \quad a_j(X(t)) = c_j \frac{1}{2} X_m(t)\,(X_m(t) - 1).$

2.1.1 Chemical Master Equation

Collisions between molecules occur in a random manner, to take this stochasticity into account the most accurate way is to consider the probability distribution at time t of the system being in a certain state, i.e. having a certain number of molecules for each species.
The Chemical Master Equation (CME) describes the time evolution of this probability through differential linear equations. In general, it represents a typical way to describe systems that can be modeled as being in a probabilistic combination of states at any given time, and where switching between states is determined by a transition rate matrix. It is indeed very common, for instance, in quantum mechanics or thermodynamical systems because it can capture their probabilistic nature. In a biological context, the CME plays a key role in modeling chemical reactions, especially when the number of molecules is relatively low and the interest is at a microscopic level.

Denote the probability of $X(t)$ being in a particular state x at time t as $\boldsymbol{P(x,t)} = P(X(t) = x)$. To write this probability at time $t + dt$ with $dt \to 0$, it is reasonable to assume that in dt at most one reaction can take place. Observe that to be in state x at time $t + dt$, or the system is already there at time t and no reaction occurs, or the system is in $x - \nu_j$ at time t and the $j-th$ reaction occurs during the time interval. Based on this, recalling the law of total probability and using the definition of the propensity function, one can write the expression of the probability as

$$P(x,t+dt) = \underbrace{\left(1 - \sum_{j=1}^{M} a_j(x)\,dt\right)}_{\substack{\text{Probability of remaining in } x,\\ \text{i.e., no reaction occurs}}} P(x,t) + \sum_{j=1}^{M} \underbrace{a_j(x-\nu_j)\,dt}_{\substack{\text{Probability of going}\\ \text{in } x, \text{ being in } x-\nu_j}} P(x-\nu_j,t).$$

Rearranging it, and taking the limit for $dt \to 0$, we finally obtain the **Chemical Master Equation**:

$$\frac{dP(x,t)}{dt} = \lim_{dt\to 0} \frac{P(x,t+dt) - P(x,t)}{dt} = \sum_{j=1}^{M} \Big(a_j(x-\nu_j)P(x-\nu_j,t) - a_j(x)P(x,t) \Big). \tag{2.1}$$

The derivative $\frac{dP(x,t)}{dt}$ can be seen as the sum of the elements related to the reactions bringing the state from $x - \nu_j$ to x, minus the sum of the ones representing the state moving away from x. Following this procedure, if one knows the reactions and their propensity functions, it is then easy to write the CME.

Considering the vector $P(t) = [P(0,t), P(1,t), ..., P(x,t), ...]^T$ and writing the previous expression for all the possible states in which X can be, one obtains the following linear system representing the CME in a compact way:

$$\dot{P}(t) = A\,P(t),$$

where the entries of A are the reaction rates as appearing in 2.1.A can be very huge since its dimension corresponds to the number of possible states of X, that is potentially infinite, and usually it is very sparse since not all reactions involve all the species.

The CME equation represents the most accurate model to describe biological systems, taking into account the probability of each reaction happening, but its dimension is a big limitation, making it often difficult to handle, even if it is linear. Indeed, it is possible to find its explicit solution only in very easy cases, such as death processes or reversible reactions [56]. Sometimes it is possible to recognize a known probability distribution when the system is at steady state, as happens for instance with birth-death processes [11]. However, in general it is not possible to get the explicit expression of the probability distribution, one usually has to simulate its realizations, as will be presented in the simulations part.

Reference Example: Degradation process

Let us introduce a simple example to which we will refer throughout this section to show how to obtain the different models in practice and how to simulate them. Later, other examples will also be analyzed.

Consider the degradation process $X \xrightarrow{k} \emptyset$, the propensity function of this reaction depends on k and on the number n of molecules of X, hence, it can be written as $a(n) = kn$. In this case the stoichiometric vector is just equal to -1, since there is only one species and one reaction causing n to decrease. Following what has been said earlier, the probability $P(n,t)$ of having n molecules will increase if we have $n+1$ molecules and a death reaction occurs, and will decrease if we have n molecules and one dies. Let n_0 be the initial number of molecules X, then the CME is given by:

$$\frac{dP(n,t)}{dt} = k(n+1)P(n+1,t) - knP(n,t), \qquad \text{for } n = 0,1,...,n_0.$$

By defining $P(t) := \begin{bmatrix} P(0,t) & P(1,t) & \dots & P(n_0,t) \end{bmatrix}^T$, the CME can be written in a compact form:

$$\frac{dP(t)}{dt} = \underbrace{\begin{bmatrix} 0 & k & 0 & 0 & \cdots & 0 \\ 0 & -k & 2k & 0 & \cdots & 0 \\ 0 & 0 & -2k & 3k & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & & 0 & -(n_0-1)k & n_0 k \\ 0 & \cdots & & & 0 & -n_0 k \end{bmatrix}}_{A} P(t), \tag{2.2}$$

that is a linear finite system. As we can see, the matrix A is very sparse and its dimension increases with the number of molecules.

2.1.2 Chemical Langevin Equation

Let us introduce another modeling method, which will be simpler and lighter than the Chemical Master Equation, though, not always suitable. This is based on the Langevin Equations, stochastic differential equations composed by two parts, a deterministic one, and a stochastic one accounting for the intrinsic fluctuations. It has been introduced to describe the Brownian motion [57] (random motion of a small particle immersed in a fluid), and it can be used in general to approximate Markov jump processes [12]. In the case of chemical reactions, the fluctuations in the number of a species

can be seen as the irregular motion of a Brownian particle [57], hence, the Langevin Equations can be employed to stochastically describe chemical reactions, approximating the CME. This will lead to an easier and faster simulation method, as it will be shown later.

The assumptions under which the Chemical Langevin Equation (CLE) is valid are:

- There exists a time interval τ small enough such that the propensity functions do not significantly change in it, i.e., relatively few reactions take place.

- The same τ is big enough so that $a_j(X(t))\,\tau \gg 1$, i.e., every reaction fires many more times than once during τ.

The first assumption allows to freeze the system over each time interval τ, updating the state based on the total number of each reaction fired during τ. If the propensity functions were exactly constant, we could describe the number of the $j-th$ reaction happening in τ with a Poisson distribution of parameter $a_j(X(t))\,\tau$ since this represents the probability of R_j firing in τ.
If also the second assumption holds, then, by the Central Limit Theorem, the Poisson distribution can be approximated with a Gaussian one with mean and variance equal to $a_j(X(t))\,\tau$. In particular, we can consider it to be true if the Poisson parameter is greater than 10-20 [58]. Therefore, in practice the CLE becomes reliable when there is an high number of molecules and when τ is appropriately chosen (by doing a trade-off). Having said this, one can proceed with writing the equations, so that the state $X(t)$ is updated every τ units of time according to the number of reactions occurring, that is given by the normal distribution $\mathcal{N}\left(a_j(X(t))\,\tau, a_j(X(t))\,\tau\right)$ for each reaction R_j:

$$Y(t+\tau) = Y(t) + \sum_{j=1}^{M} \nu_j\, \mathcal{N}\left(a_j(X(t))\,\tau, a_j(X(t))\,\tau\right),$$

where we have used $Y(t)$ to emphasize that now the state is a (random) real-valued vector; whereas, with $X(t)$ we were denoting a (random) integer-valued vector.
Recalling that a Gaussian distribution with mean μ and variance σ^2 can be written as $\mu + \sigma Z_j$, with $Z_j \sim \mathcal{N}(0,1)$, the previous expression becomes:

$$Y(t+\tau) = Y(t) + \tau \sum_{j=1}^{M} \nu_j\, a_j(Y(t)) + \sqrt{\tau} \sum_{j=1}^{M} \nu_j \sqrt{a_j(Y(t))} Z_j, \tag{2.3}$$

Note that actually the sign of ν_j in the last term of the equation does not matter because of the property of linear combination of independent Gaussian variables:

$$W = \sum_i \alpha_i Z_i \sim \mathcal{N}\left(\sum_i \alpha_i \mu_i, \sum_i \alpha_i^2 \sigma_i^2\right) \iff Z_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right) \text{ i.i.d.}$$

Finally, if we make $\tau \to 0$, the last expression can be seen as the discretization of

$$\frac{dY(t)}{dt} = \sum_{j=1}^{M} \nu_j a_j(Y(t)) + \sum_{j=1}^{M} \nu_j \sqrt{a_j(Y(t))}\, dZ_j(t), \tag{2.4}$$

that is what is properly called **Chemical Langevin Equation** in the continuous random variable $Y(t)$, with dZ_j representing M independent temporally uncorrelated scalar Brownian motions.
In general, the Langevin Equations are composed by a drift term (the first one on the right side) and a diffusion one (second on the right side), it is noteworthy that the presented CLE inferred the forms of both the drift and diffusion terms from the premises underlying the CME [59].

The so-obtained model is now based on a nonlinear finite set of equations with dimension equal to the number of species, not the number of possible states. This will ease the simulation part, though we will have to be careful in satisfying the given assumptions.

Reference Example: Death/Degradation process

Consider again the reference example of the degradation process, recalling that $a(n) = kn$ and $\nu = -1$, we can write the CLE as follows:

$$\frac{dY(t)}{dt} = -kY(t) - \sqrt{kY(t)}\, dZ, \qquad dZ \sim \mathcal{N}(0,1). \tag{2.5}$$

2.1.3 Fokker-Planck Equation

The Fokker-Planck Equation (FPE) [14, 15] can be used to describe the evolution of the probability density function of having a certain number of molecules at time t, when Brownian motion is included. In particular, it is an exact description if the noise really acts as a Gaussian white noise.
In the context of biological systems, when the assumptions of the Chemical Langevin Equation hold, the Fokker-Planck Equation can be derived from an approximation of the CME keeping the first two terms of the Kramers-Moyal expansion [60], or from the CLE using the drift and diffusion functions [61].
However, it is completely equivalent to the Chemical Langevin Equation and we have not encountered many papers using it to describe biological systems. Moreover, as it happened for the CME, it would be difficult to interpret the Fokker-Planck Equation since the variable of interest is the probability distribution and hence we should find a method to simulate it. For these reasons we have decided to not delve into it, focusing on the other models presented in this thesis.

2.1.4 Reaction Rate Equation

When the amount of molecules of each species is very large the fluctuating terms of the CLE can be neglected, being left only with deterministic equations. To be more precise, one should talk about both the number of molecules and the volume, formally we refer to it as the thermodynamic limit, i.e., the limit for which the molecules and the volume tend to infinity, keeping the concentration constant so that the reaction stays the same. Otherwise, if only the number of molecules increases in a fixed volume, then one should also change the reaction rate coefficients because reactions become more likely to happen.
Therefore, in the termodynamic limit we can look at the concentration and not at the number of molecules. Usually the concentration is expressed as the number of moles or molecules over the total volume Ω, hence, we can refer to it as $y(t) = \frac{Y(t)}{N_A \Omega}$ or $y(t) = \frac{Y(t)}{\Omega}$ and study its evolution over time. N_A is the Avogadro number, used to obtain the number of moles corresponding to a certain number of molecules; the choice of using the molar or molecular concentration is arbitrary. We have seen that the former is often used, in general when taking the values from a paper or a document, one should check which concentration definition has been used and then make the correct scaling. Substituting $Y(t)$ with the expression of the concentration in the CLE, and making the number of molecules and volume tend to infinity, then one can observe that the deterministic part of the CLE will grow as the volume, and the stochastic one as the square root of the volume, hence, we can claim that the ODE part dominates [55].

To better understand the scaling of the coefficients we recall that the propensity functions are proportional to the number of species $Y(t)$, that can be expressed as $Y(t) = y(t)\,\Omega$ or $Y(t) = y(t)\,\Omega N_A$. Thus, the propensity functions can be written in terms of the concentration $y(t)$, remaining with a factor Ω^m or $(\Omega N_A)^m$, where m corresponds to the order of the reaction [11, 16]. If now we divide the CLE by the volume or by the product of the Volume and the Avogadro number, and we neglect the stochastic part, then, on the left side we remain with the derivative of the (molar or molecular) concentration of molecules, and on the right side with the same drift term of the CLE, but with the coefficients of the propensity functions multiplied by Ω^{m-1} or $(\Omega N_A)^{m-1}$. If we call c_j the stochastic reaction constants of the CME or CLE and k_j the reaction rate constants of the RRE, then their relationship can be expressed as $c_j = \frac{k_j}{\Omega^{m-1}}$ or $c_j = \frac{k_j}{(\Omega N_A)^{m-1}}$. From a theoretical point of view, actually the difference between them is more complicated and relates to the conceptual differences that exist between the stochastic and deterministic approaches, but from a practical point of view it is sufficient to know that the RRE can be obtained from the CLE neglecting the noise and possibly scaling the reaction coefficients.

Following this procedure, we finally obtain the **Reaction Rate Equation** (RRE):

$$\frac{dy(t)}{dt} = \sum_{j=1}^{M} \nu_j a_j(y(t)), \tag{2.6}$$

where $y(t)$ is a continuous real-valued vector representing the concentration of each molecule. We have therefore obtained a set of nonlinear deterministic equations of dimension equal to the number of species in the system. Normally, given the reaction rates, it is straightforward to write the RRE, considering that each reaction in the system affects the rate of change of the species involved (the derivative of their concentration) proportionally to the reaction rate and the concentration of the reacting species. This is what is generally called Law of mass action.

Reference Example: Death/Degradation process

The RRE equation for the degradation process considered earlier can be easily obtained as:

$$\frac{dy(t)}{dt} = -k\,y(t),$$

where the coefficient k is the same of the CME and CLE because the degradation process is a 1^{st} order reaction. Indeed, doing the procedure previously described, if one starts from the CLE in (2.5) neglecting the noisy part, then substitutes $Y(t)$ with $y(t)\,\Omega$, and then divides the equation by Ω, one obtains exactly the given model.

In this case, it is also possible to find the explicit solution $y(t) = e^{-kt}\,y(0)$, hence one clearly understands why RRE are normally preferred: they are easy to derive, compact, deterministic, and often explicitly solvable, though, this approximation is not always true. Indeed, with systems that can be highly affected by noise, stochasticity is necessary to correctly describe them, giving the possibility of focusing on single realizations, rather than on their average concentration.
Therefore, the RRE can be very useful and powerful, but one has to use it carefully. In particular, it can be very interesting to compare their solution with the average solution of CME or CLE (when RRE are valid of course) to see if the deterministic model can capture the trend of the real one. In the following chapter the simulations of the models presented here will be shown to better compare them and understand when one is more suitable than the other.

2.2 Simulation Methods

As anticipated, to study the system and its behavior, the model is an interesting and useful tool. In some cases we can directly use it to gain some properties of the system, such as stability, steady state, equilibria, ... Though, when dealing, for instance, with the Chemical Master Equation, that can be very huge and that describes the probability evolution, it often becomes difficult to directly get information on the state, save for exceptional cases. Hence, the usual and easiest way

to use the CME is actually to simulate it. A similar reasoning can be done also for the Chemical Langevin Equation, therefore, in this chapter we will provide some methods to simulate the models previously presented.

2.2.1 *Stochastic Simulation Algorithm* for CME

There exist different ways to simulate the Chemical Master Equation, the standard method has been proposed by D.T. Gillespie in *Exact Stochastic Simulation of Coupled Chemical Reactions, 1977* [16]. It is an exact method that numerically simulates the time evolution of a well-stirred chemically reacting system, fully accounting for the inherent randomness of such systems.
Later, Gillespie himself and other authors have proposed modified versions of it [17, 18], or new simulation methods, among which the main ones are the τ-leaping [19] and the Hybrid [62, 63] methods.

Gillespie's Stochastic Simulation Algorithm (SSA) is fully equivalent to the Chemical Master Equation, even though it never uses it explicitly and does not aim to numerically solve it. Instead, it is a systematic, computer-oriented procedure in which Monte Carlo techniques are employed to simulate the same Markov process that the master equation describes analytically. The goal is to simulate the time evolution of the N species as a random-walk process, knowing only their initial values $X_i(0)$ and the shape and parameters' values of the M propensity functions. In particular, at every time t it computes, based on the current propensity functions, the time at which the next reaction will occur and its index; then it updates the state according to the reaction fired.

Therefore, Gillespie's SSA looks at the single reactions and updates the state every time a reaction occurs. Although this represents the most accurate simulation method, it can become extremely slow especially when reactions are fast and frequent or when the molecular populations increases, and hence the time step for the next reaction becomes very small. We can indeed say that the SSA has a computational cost that scales with the number of reaction occurrences, so, systems with one or more "fast" reactions become costly and inefficient to simulate in this way. For this reason, other techniques have been proposed, making small approximations, but getting faster algorithms; they will be presented later in the chapter. However, one of the great advantages of the SSA, that makes it so reliable, is that in this procedure the infinitesimal time increments dt are never approximated by small but finite time steps, which can be a common source of computational inaccuracies and instability in standard numerical methods for solving the deterministic reaction rate equations. This will be especially advantageous when dealing with systems in which the molecular population levels can change suddenly and sharply with time.

Returning to the algorithm, as reference we have mainly considered the paper of Gillespie [16] where he explains it. Consider the so-called Reaction Probability Density Function $P(\tau, j)$, that is

such that $P(\tau, j)d\tau$ represents the probability that given the state at time t, the next reaction will occur in the infinitesimal time interval $(t+\tau, t+\tau+d\tau)$, and will be an R_j reaction. This can be computed as the product of $P_0(\tau) :=$ Probability that no reaction occurs in the interval $(t, t+\tau)$ given the state at time t and $a_j(X(t))\,d\tau$, that we know represents the probability of the $j-th$ reaction happening in the time interval $(t+\tau, t+\tau+d\tau)$.
To find the analytical expression of $P_0(\tau)$, we can consider $P_0(\tau+d\tau)$ as follows.

$$\begin{aligned} P_0(\tau+d\tau) &= \text{Probability of no reaction happening neither in } [t, t+\tau), \text{ neither in } [t+\tau, t+\tau+d\tau) = \\ &= P_0(\tau)\left(1 - \text{Sum of probabilities of each reaction happening in } [t+\tau, t+\tau+d\tau)\right) = \\ &= P_0(\tau)\left(1 - \sum_{j=1}^{M} a_j(X(t))\,d\tau\right) \end{aligned}$$

Bringing $P_0(\tau)$ at the left of the expression and dividing everything by $d\tau \to 0$, one gets:

$$\frac{P_0(\tau+d\tau) - P_0(\tau)}{d\tau} = \frac{dP_0(\tau)}{d\tau} = -\underbrace{\left(\sum_{j=1}^{M} a_j(X(t))\right)}_{a_0(X(t)) \geq 0} P_0(\tau)$$

This is a linear scalar stable ODE that has solution $P_0(\tau) = e^{-a_0(X(t))\,\tau}$. Therefore, we can write the Reaction Probability Density Function $P(\tau, j)$ as

$$P(\tau, j)\,d\tau = P_0(\tau)\,a_j(X(t))\,d\tau \;\longrightarrow\; P(\tau, j) = \begin{cases} a_j(X(t))\,e^{-a_0(X(t))\,\tau} & \text{if } 0 \leq \tau < \infty \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

Notice that this probability is not the one described by the Chemical Master Equation, and differently from that one, it depends on all the reaction constants and the current numbers of molecules of all reactant species. However, its expression is derived from the same hypothesis of the CME, hence it can be used equivalently to the CME. The following figure, taken from [18], shows the shape of this probability:

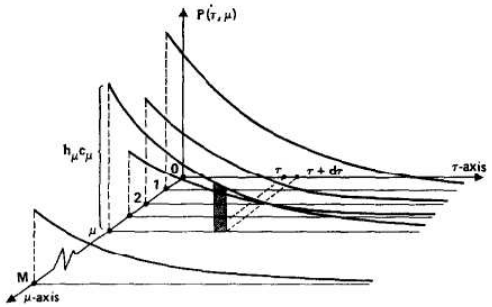


Figure 2.1: Schematic plot of the reaction probability density function $P(\tau, j)$. The shaded area is by definition equal to $P(\tau, j)d\tau$, and the sum of the areas under all the M curves is equal to one[18].

To derive the proper algorithm, let us write (2.7) as $P(\tau, j) = \frac{a_j(X(t))}{a_0(X(t))}\, a_0(X(t))\, e^{-a(X(t))\,\tau}$, then, from here τ can be seen as an exponential random variable with mean (and standard deviation) $\frac{1}{a_0(X(t))}$, and j as a statistically independent integer random variable with point probabilities $\frac{a_j(X(t))}{a_0(X(t))}$. Hence, two random variables, j and τ, should be generated accordingly, respectively, to the probabilities just mentioned. To do so, we can generate two random numbers r_1 and r_2 from a uniform distribution in the unit interval, and compute:

$$\tau = \frac{1}{a_0(X(t))} \ln\left(\frac{1}{r_1}\right) \qquad \text{and} \qquad j \;\text{ such that } \sum_{k=1}^{j-1} a_k < r_2\, a_0(X(t)) \leq \sum_{k=1}^{j} a_k. \tag{2.8}$$

This gives life to what is called *Direct Method*. Using this, we obtain the original version of the SSA, that can be schematized by the following pseudocode:

Algorithm 1 SSA for CME

$t_f \leftarrow$ simulation length, $V \leftarrow$ Stoichiometric matrix with ν_j as columns
for all simulations **do**
  Clear t, x
  $h \leftarrow 1, t \leftarrow 0, x \leftarrow X(0)$
  while current time $t(h) < t_f$ **do**
    $a \leftarrow \begin{bmatrix} \dots & a_j & \dots \end{bmatrix}$, propensity functions, using current state $x(h)$
    $a_0 \leftarrow$ sum of the propensity functions
    $a_{cum} \leftarrow$ cumulative sum of a
    $r_1, r_2 \leftarrow$ random numbers from uniform distribution
    $\tau \leftarrow \frac{1}{a_0} \ln\left(\frac{1}{r_1}\right)$
    $j \leftarrow \min(\text{find}(a_{cum} \geq r_2\, a_0))$
    $t(h+1) \leftarrow t(h) + \tau$
    $x(h+1) \leftarrow x(h) + V(:, j)$
    $h \leftarrow h + 1$
  end while
end for

The SSA so derived is a rigorous and exact method to simulate the time evolution of the system described through the CME, thus including the intrinsic fluctuations. The algorithm provides single realizations of the probability distribution of the CME, hence it can be useful to seek for some particular behavior. However, to get a statistically complete picture of the temporal evolution of the system, we must actually carry out several independent realizations, each starting with the same initial set of molecules and proceeding for the same time. This approach can become very inefficient

when the number of reactions or molecules is very high since each simulation will then require a large amount of time to be run. It is also possible to change the ending condition of the *while* loop to, for instance, exiting if $a(X(t)) = 0$, $X_i(t) \lesseqgtr$ specified bound, or $h <$ prescribed number of reactions allowed, or as the user needs.

Finally, we mention that in the literature there have been different proposals to modify the SSA algorithm, between which one uses the *First-reactions Method*, or the *Next-reactions Method*, or another one a generalization of both the *Direct* and *First-reactions* methods in place of the *Direct Method*. Other solutions are based, for example, on reordering the propensity functions to reduce the time spent by the algorithm; if the reader is interested, all these versions are mentioned in [17]. They can provide some interesting insights on the algorithm, though, they do not provide a significant improvement in the computational time with respect to the original SSA, hence we do not refer to them in this thesis.

2.2.2 *τ-leaping approximation* for CME

We have seen how the SSA can become slow when reactions occur very close to each other, the τ-leaping method is an alternative algorithm providing a faster way to simulate the CME, at the cost of making some approximations. In particular, it divides the time interval into time steps of size τ, groups all the reactions firing in each time step, and updates the state based on the number of reactions that occurred in τ. This method is valid only under the leap condition, which means that there exists τ sufficiently small so that the propensity functions are almost constant over each time step. However, one can observe that when the SSA is slow, there will be probably many reactions happening and the reactant population would be big, hence, in this situation we can group together some reactions knowing that to have the propensity functions significantly changing we need a very high number of reactions since the large dimension of the population makes small changes not really visible.

The correct choice of τ is what determines the accuracy of the algorithm, and can represent an important difficulty in its implementation. It is possible to choose it by trial and error, or by imposing some desired behaviors or tolerance thresholds on the variability of the propensity functions, or it is also possible to change it as the simulation proceeds. The last approach is the one we have implemented. However, from the *Direct method* of the SSA we know that there will be a reaction that will occur within $\frac{1}{a_0(X(t))} \ln\left(\frac{1}{r_1}\right) \geq \frac{1}{a_0(X(t))}$, hence, in general it is true that the closer τ is to $1/a_0(X(t))$, the more precise the algorithm is, although choosing $\tau \sim 1/a_0(X(t))$ would be inefficient from a velocity point of view. Therefore, normally a trade-off between speed and accuracy is necessary. For example, in [19] they propose as general rule to use the τ-leaping method if

$\tau > 2/a_0(X(t))$, this is what we will refer to in our implementation.

Let us recall that $a_j(X(t))\,\tau$ represents the probability of the $j-th$ reaction happening in τ, if this is constant in the time interval, then we can claim that the number of reactions happening for each j is given by a Poisson random variable $\mathscr{P}_j\left(a_j(X(t))\,\tau\right)$. Of course $a_j(X(t))\,\tau$ in reality is not exactly constant, hence, using the Poisson random variable, we get an approximation of the system evolution, in which the state components are updated at each time step as

$$X_i(t+\tau) = X_i(t) + \sum_{j=1}^{M} \mathscr{P}_j\left(a_j(X(t))\,\tau\right)\,\nu_j(i), \qquad \text{for } i = 1,...,N. \tag{2.9}$$

This can also be written in a compact form for the whole state $X(t)$ as

$$X(t+\tau) = X(t) + \sum_{j=1}^{M} \mathscr{P}_j\left(a_j(X(t))\,\tau\right)\,\nu_j,$$

or, defining $V = \begin{bmatrix} \dots & \nu_j & \dots \end{bmatrix}$ and $P = \begin{bmatrix} \dots & \mathscr{P}_j\left(a_j(X(t))\,\tau\right) & \dots \end{bmatrix}$, one obtains:

$$X(t+\tau) = X(t) + V \cdot P^T,$$

which exploits the matrix product between V and P to perform the sum.

The algorithm can be schematized in the following steps:

Algorithm 2 τ-leaping for CME

$t_f \leftarrow$ simulation length, $\tau \leftarrow$ time step satisfying leap condition
$t \leftarrow \begin{bmatrix} 0 & \dots & t_f \end{bmatrix}$, time vector with step amplitude τ
$x \leftarrow X(0)$
$V \leftarrow$ Stoichiometric matrix with ν_j as columns
for all simulations **do**
  for all time steps **do** ▷ *A different exiting condition can be added depending on the application*
    $a \leftarrow \begin{bmatrix} \dots & a_j & \dots \end{bmatrix}$, propensity functions, using current state
    $P \leftarrow \begin{bmatrix} \dots & \mathscr{P}_j\left(a_j\,\tau\right) & \dots \end{bmatrix}$, Poisson samples with mean $a_j\,\tau$
    State at next step $\leftarrow$ current state $+ V \cdot P^T$
  end for
end for

The described algorithm is the easiest and probably most naive way to implement the τ-leaping approximation, indeed, one should check the leap condition at every time step and decide how to handle the possible negative values for the state. Moreover, since the value of the state is changing, it would be more efficient to update τ as the algorithm proceeds.

Regarding the possibility of receiving a negative number of molecules, this could happen when a degradation reaction fires too many times in τ. Actually, if this is the case, it means that the leaping condition is not truly satisfied, because the propensity functions are evidently changing of a significant amount. Hence, the best approach would be to change τ, or to switch to the SSA when the number of molecules is particularly low and also small changes are significant. It is also possible to set the number of molecules to zero instead of the negative value, or to generate a new Poisson sample, even though, the former solution could return a wrong algorithm if for instance it inhibits a reaction that otherwise would happen, and the latter one could slow down the algorithm if a high number of Poisson samples trials becomes necessary.
In the literature, there exist other proposed methods [22–24] to avoid getting negative number of molecules. One idea [22, 23] is based on substituting the Poisson random variables with binomial random variables. The other approach [24], instead, introduces what are called Critical Reactions, i.e., reactions that could bring the state to a negative value, and uses them to modify the original τ-leaping algorithm so that no more than one firing of a critical reaction can occur in a single τ, which makes it impossible for any critical reaction to produce a negative species population count. This represents a very powerful, robust and potentially more accurate implementation than the original one since it can be reduced to the SSA if all reactions are considered critical, or to the standard τ-leaping procedure if no reaction is treated as critical. If the reader is interested, all the steps are well described on page 8 of [24], we do not report the detailed procedure here because for simplicity we have decided to use Gillespie's SSA if a negative number of molecules is returned.

As we were mentioning, it is also possible to change τ adaptively during the simulation. This causes of course a complication in the algorithm, but on the other hand it allows to speed it up by choosing every time the minimum τ satisfying the leap condition, when possible. Indeed, it may happen that τ is equal to infinity if, for example, the number of molecules is too low; hence, it would be necessary to switch to the SSA because the τ-leaping approximation cannot be applied. However, the big advantage updating τ during the simulations is that only the real necessary steps are performed, though, the step sizes and numbers will be different in every simulation, which does not happen if τ is fixed. In this direction there have been many proposals; the first one was made by Gillespie in [19], then he redefined it together with Petzold in [20], and finally Cao, Gillespie and Petzold further improved it in [21]. However, the general idea was to introduce a control problem to optimally choose the different values of τ during the algorithm based on a control parameter ε that imposes a maximum variation in the propensity functions [19, 20]:

$$|a_j(X(t+\tau)) - a_j(X(t))| \leq \varepsilon\, a_0(X(t)).$$

The method presented in [20] has been proven to be very efficient in selecting τ, although requiring additional computational effort to compute it. Hence, in [21] they proposed a new τ selection procedure that approximates the condition in the previous method looking at the state changes rather than at the propensity functions. This allows to still have a reliable algorithm, that will be easier to implement and faster to execute, especially when there are many reactions and species. For these reasons, in our work we have referred to this method.

Therefore, the inequality becomes:

$$\Delta_\tau X_i = X_i(t+\tau) - X_i(t) = \sum_{j=1}^{M} \mathscr{P}_j\left(a_j(X(t))\,\tau\right)\, \nu_j \leq \max\{\varepsilon_i X_i(t), 1\},$$

where ε_i is chosen differently for every species depending on the kind of reactions involving it, with the objective of guaranteeing that the relative changes in the propensity functions are all bounded, at least approximately, by ε. We ask the reader to refer to the paper for the detailed expression and reasoning behind this method. Practically, to ensure this condition it is sufficient to choose τ at every iteration such that:

$$\tau = \min_{i=0,\ldots,N} \left\{ \frac{\max\{\varepsilon_i x_i, 1\}}{|\hat{\mu}_i(x)|}, \frac{\max\{\varepsilon_i x_i, 1\}^2}{\hat{\sigma}_i^2(x)} \right\},$$

with

$$\hat{\mu}_i(x) \triangleq \sum_{j=0,\ldots,M} \nu_{ij} a_j(x) \qquad \text{and} \qquad \hat{\sigma}_i^2(x) \triangleq \sum_{j=0,\ldots,M} \nu_{ij}^2 a_j(x)$$

representing the mean and standard deviation of $\Delta_\tau X_i$. Moreover, our design choice has been to switch to Gillespie's SSA whenever τ-leaping is not applicable, i.e. $\tau = \infty$, the number of molecules becomes negative, or $\tau < \frac{2}{a_0(X(t))}$ since it would not be more efficient than the SSA.

As we have seen, although the τ-leaping method seems appealing due to its ability to accelerate the simulations, one should be careful in particular with selecting appropriate values for τ or ε, verifying the validity of the underlying assumptions, handling negative values, and determining if and when to switch to the SSA. These design choices introduce complexity, making the algorithm non-trivial to implement and often dependent on specific cases, whereas the SSA could be applied in the same way regardless of the system involved.

The following pseudocode shows our final implementation of the τ-leaping approximation, including all the considerations previously discussed:

Algorithm 3 τ-leaping for CME, Version 2 with adaptive τ

$t_f \leftarrow$ simulation length
$V \leftarrow$ Stoichiometric matrix with ν_j as columns
$\varepsilon \leftarrow \left[\dots \varepsilon_i \dots\right]$, as defined in the paper [21]
for all simulations **do**
  Clear t, x
  $h \leftarrow 1, t \leftarrow 0, x \leftarrow X(0)$
  while current time $t(h) < t_f$ **do** ▷ *A different exiting condition can be added depending on the application*
    $a \leftarrow \left[\dots \quad a_j \quad \dots\right]$, propensity functions, using current state $x(h)$
    $a_0 \leftarrow$ sum of the propensity functions
    $\hat{\mu} \leftarrow V \cdot a^T$
    $\hat{\sigma^2} \leftarrow V.^2 \cdot a$ ▷ *$V.^2$ denotes the element-wise square operation of V*
    $\tau_\mu \leftarrow \left[\dots \frac{\max\{\varepsilon_i \cdot x_i(h), 1\}}{\hat{\mu}_i} \dots\right]$
    $\tau_\sigma \leftarrow \left[\dots \frac{\max\{\varepsilon_i \cdot x_i(h), 1\}^2}{\hat{\sigma}_i} \dots\right]$
    $\tau \leftarrow \min\{[\tau_\mu, \tau_\sigma]\}$ ▷ *$[\,\cdot\,,\,\cdot\,]$ denotes array concatenation*
    $P \leftarrow \left[\dots \quad \mathscr{P}_j\left(a_j \tau\right) \quad \dots\right]$, Poisson samples with mean $a_j \tau$
    if $\tau = \infty$ or $\tau < \frac{2}{a_0}$ or $x(h) + V \cdot P^T < 0$ **then**
      Use Gillespie's SSA
    else
      $x(h+1) \leftarrow x(h) + V \cdot P^T$
      $t(h+1) \leftarrow t(h) + \tau$
    end if
    $h \leftarrow h + 1$
  end while
end for

To conclude, we finally mention also an alternative way of applying the τ-leaping method, proposed in [19]: the Estimated-Midpoint τ-leap method, taking inspiration from the Estimated-Midpoint procedure used with the Euler method. In a few words, the difference is in the parameter of the Poisson distribution, which in this case is computed using the propensity functions calculated in $X(t) + \bar{\lambda}/2$, where $\bar{\lambda} = \tau \sum_j a_j(X(t))\, \nu_j$ is the expected state change. Therefore, instead of computing them in $X(t)$, they propose doing it in the middle point between $X(t)$ and $X(t+\tau)$. This

method has been proven to work in the easiest example of the degradation process, but it required further adjustments for more complex systems, hence, we do not see the necessity to implement it. Finally, there exists a different method, the k_α-leap method [19], where the leaping is based on a predetermined number of firings of a specified reaction channel rather than on a predetermined time when the firings of each reaction are added. This solution is basically equivalent to the τ-leaping method, and actually it could be worse in very specific cases.
For these reasons, our simulations have been obtained by implementing the pseudoce described in Algorithm 3.

2.2.3 *Euler-Maruyama method* for CLE

The Euler-Maruyama(E-M) [25] represents exactly the implementation of the discretized CLE in (2.3) with time step τ. It is interesting to note that, following the procedure used to obtain the CLE, the E-M method can be regarded as the τ-leaping one when the Poisson distribution is approximated with a Gaussian one. We recall that this is possible if τ is large enough so that $a_j(X(t))\,\tau \gg 1$, while still satisfying the leaping condition. This implies that τ must be chosen such that in each time step all the reaction channels fire many more times than once yet none of the propensity functions changes appreciably. Then, the jump Markov process $X(t)$ can be approximated by the continuous Markov process $Y(t)$ defined by the standard form of the chemical Langevin equation (2.4).

The algorithm implementing the E-M method can be obtained directly from the expression (2.3), we report the pseudocode below.

Algorithm 4 Euler-Maruyama for CLE

$t_f \leftarrow$ simulation length, $\tau \leftarrow$ time step satisfying leap condition and $a_j(X(t))\,\tau \gg 1$
$t \leftarrow \begin{bmatrix} 0 & \dots & t_f \end{bmatrix}$ with step amplitude τ
$x \leftarrow X(0)$
$V \leftarrow$ Stoichiometric matrix with ν_j as columns
for all simulations **do**
  for all time steps **do** ▷ *A different exiting condition can be added depending on the application*
    $a \leftarrow \begin{bmatrix} \dots & a_j & \dots \end{bmatrix}$, propensity functions, using current state
    $Z \leftarrow \begin{bmatrix} \dots & Z_j & \dots \end{bmatrix}^T$, $Z_j \leftarrow \mathcal{N}(0,1)$
    State at next step $\leftarrow$ current state $+\tau \cdot V \cdot a^T + \sqrt{\tau} \cdot V \cdot \begin{bmatrix} \dots & \sqrt{a_j} \cdot Z_j \dots \end{bmatrix}^T$
  end for
end for

The so-described algorithm is very easy to implement, and can execute really fast, depending of course on the choice of τ, that is probably the most demanding part, as we have seen for the τ-leaping method. We recall that the Langevin equation (and hence the simulations obtained with the E-M method) is valid when τ is small enough so that the propensity functions undergo a relatively small change over $[t, t+\tau)$ and big enough so that the reaction channels fire many more times than once in the same time interval; both assumptions are normally satisfied when there are abundant molecular populations. In any case, finding the correct τ could not be trivial, we have seen that there are various ways of selecting τ, some more naive and others more accurate adjusting it during the run of the simulation. Unfortunately, in the case of the E-M method we have not found an implementation with an adaptive τ, therefore, our choice has been to find the value of τ doing a trade-off between all the requirements, and to switch to Gillespie's SSA if one of the necessary assumptions does not hold, if a negative number of molecules is returned, or if τ is so small that it is not convenient to apply the E-M method.

To conclude we would like to mention that there exist also some other methods [64] that are not simulating the discretized Langevin equations, but that are numerically solving them, between which, the Milstein scheme and the derivative-free Milstein scheme [65]. Moreover, in cases where there are reactions particularly faster than others, it is possible to use a continuous approximation for the fast reactions and a discrete Markov process for the slow ones. This gives life to a hybrid framework [62, 63] which could be very powerful, even though we did not refer to it because the E-M method represents the standard way to simulate the CLE. Actually, the Euler method could also be performed by directly using the corresponding Matlab command, but this would not give the possibility of controlling if the assumptions are satisfied or of handling negative results. An example with this implementation will be shown later, but in general in this work we have used our implementation of E-M, as described above. Finally, in the following page the reader can find some examples where E-M gives reliable results, and hence is convenient to use, and others in which one of the assumptions does not hold, thus it is better to refer to the τ-leaping method or the SSA.

2.2.4 *Finite Volume Method* for FPE

One of the methods to simulate the Fokker-Planck Equation is the so-called finite volume method [60]. It is based on discretizing in space the FPE to obtain an approximation of it taking into account the boundary conditions, then what is left can be solved by time integration. The solution obtained in this way differs from the one of the CME mainly because of the derivation of the FPE that consists in applying the Taylor expansion, but ignoring the terms of order three and higher.
There exist also other numerical methods [66, 67] to solve the FPE that are based on Monte Carlo simulations, they can be referred to as finite difference and finite element methods. Though, as we

mentioned before, we did not aim to delve into them because it is not a common used model from what we could see. The CLE is instead usually preferred and used to simulate the FPE, since it is equivalent to it. Following this idea, in this thesis we will only consider the CLE, and not the FPE.

2.2.5 *ODEs solvers* for RRE

The simplest way to model biological systems is through Reaction Rate Equations (2.6) using the Law of mass action. Even though they can be very easy to derive and study, they obviously represent an approximation of the real system that is valid if the molecular numbers are particularly high and the volume tends to infinity. Indeed, under these assumptions we can consider just the concentration and neglect the stochasticity of the system. In any case, it is in principle impossible to predict the exact molecular population levels unless we take into account the precise positions and velocities of all the molecules in the system, which is unfeasible. Hence, every model will be an approximation and an attempt to simulate the real system, becoming acceptable depending on the applications and assumptions.

However, analytical solutions to the reaction-rate equations can be found only for rather simple systems, so it is usually necessary to solve these equations numerically. This can be done by exploiting, for example, the Matlab solver command *ode45*.
Once we obtain the deterministic behavior of the system, we can compare it with the average of the stochastic simulations. We will see that RRE are an accurate model with very simple systems, whereas, in other cases (for instance with bistability) the average molecular population levels will not exactly satisfy any closed system of ODEs.

Reference Example: Degradation process

Let us consider the example of the degradation process that has already been used to derive the different models; here we will show how to implement the simulation methods presented, and we will compare their results. Along with only the simulating purpose, we have chosen the coefficient k arbitrarily, not representing any real system. Below, the pseudo-reaction and the models we have already derived are reported.

Degradation process: $X \xrightarrow{k} \emptyset, \quad a(n) = kn, \quad \nu = -1, \quad k = 0.1\,s^{-1}.$

CME: $$\frac{dP(n,t)}{dt} = k(n+1)P(n+1,t) - knP(n,t), \quad \text{for } n = 0,1,...,n_0 = X(0).$$

CLE: $$\frac{dY(t)}{dt} = -kY(t) - \sqrt{kY(t)}\,dZ, \quad dZ \sim \mathcal{N}(0,1).$$

RRE: $$\frac{dy(t)}{dt} = -ky(t), \quad \text{with solution } y(t) = e^{-kt}y(0).$$

We recall that the CME and the CLE look at the number of molecules, whereas the RRE at the their concentration. When passing from one to the other, there can be a scaling in the coefficient if the reaction order is greater than 1, hence in this case the coefficient k stays the same.

The following graphs show the results obtained with the different simulation methods presented in the previous paragraph, starting always with the same initial condition $X(0) = 100$:

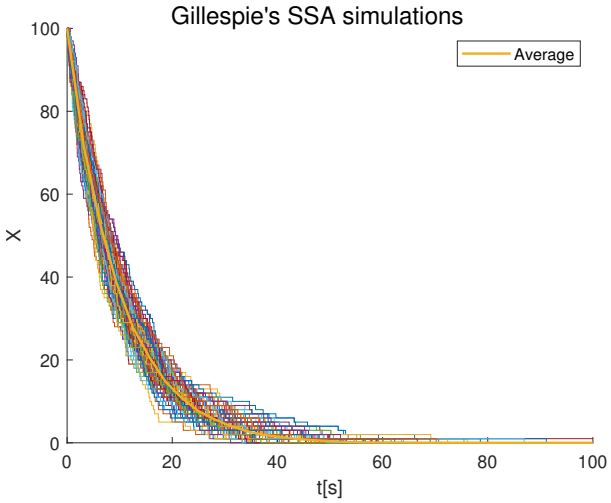


Figure 2.2: 100 simulations of degradation process by SSA, with $k = 0.1\,s^{-1}$.

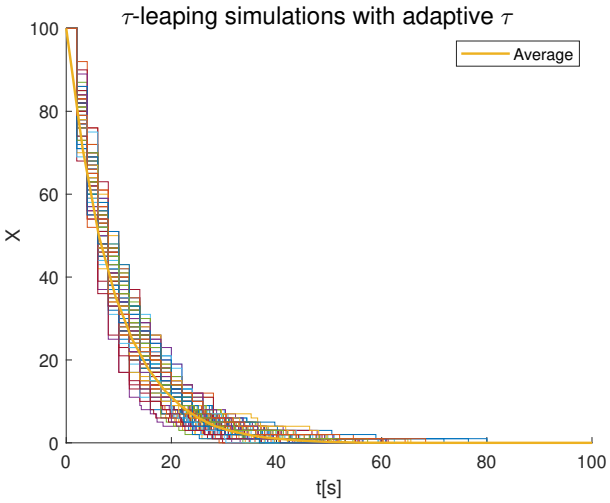


Figure 2.3: 100 simulations of degradation process by adaptive τ-leaping method, with $k = 0.1\,s^{-1}$.

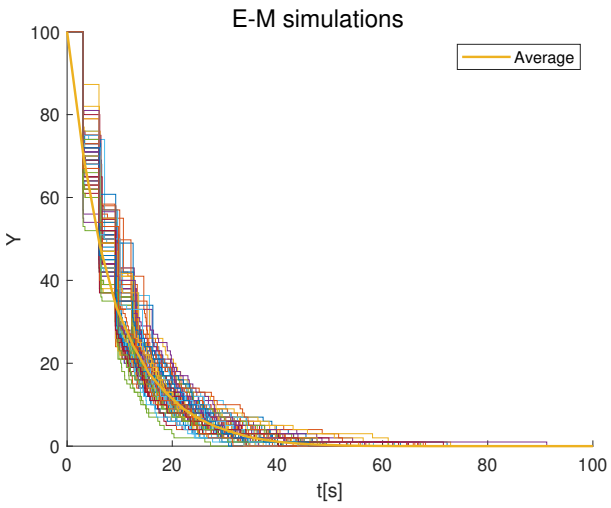


Figure 2.4: 100 simulations of degradation process by E-M method, with $k = 0.1\,s^{-1}$ and $\tau = 3$.

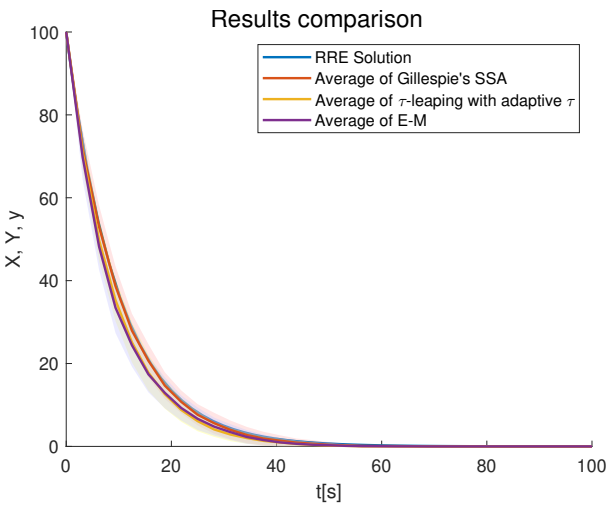


Figure 2.5: Comparison of RRE solution and averages of the simulations obtained from SSA, τ-leaping and E-M for the degradation process, with $k = 0.1\,s^{-1}$. Standard deviation from the averages is shown with light colored bands.

Fig. 2.2 has been obtained using Gillespie's SSA in the standard way (by Direct method), Fig. 2.3 using τ-leaping approximation letting τ change adaptively according to the method described in Algorithm 3, and Fig. 2.4 using E-M method with fixed value, chosen to satisfy the following requirements:

- $\tau|$ the propensity functions do not change too much (leap condition).
  In this case, their change is directly proportional to the one of the state. Qualitatively, looking at the simulations, we assumed that a good value of τ could be $2-5$, it should be smaller at the beginning and larger in the end.

- $a_j(X(t))\,\tau \gg 1\ \forall j$ to make the approximation of the Poisson r.v with the Gaussian r.v. valid. Since the maximum value of $a_j(X(t))$ is 10, for $\tau = 3$ it holds: $a(X(t))\,\tau = 30$ at the beginning, but as time passes, the approximation is no longer valid. Indeed, in the end the population is small and we know that the CLE is valid when the population is big.

- $\tau > 2/a_0(X(t))$ so that it is worth to use E-M method.
  In the example the maximum value of $a_0(X(t))$ is $100k = 10$, when $t = 0$; hence, we have to choose $\tau > 0.2$ at the beginning, and larger as the simulation proceeds (since $a_0(X(t))$ decreases).

We have chosen $\tau = 3$ by trial and error doing a trade-off between all the constraints, knowing that after a while one of the conditions will probably not be satisfied. We recall that this was necessary because for the E-M method we could not find a version with adaptive τ, hence we had to choose a fix value of τ, knowing that the assumptions will not always be satisfied. When this happens, both in τ-leaping and E-M, we call Gillespie's SSA. This is indeed clearly visible from the previous plots in the parts where the steps are very small. Moreover, the SSA is called to avoid negative numbers of molecules, as it has been explained in the previous section.

Finally, in Fig. 2.5 we have compared the RRE solution with the averages of the simulations, including also their standard deviation with a light colored band around the mean. In terms of the standard deviation and average, all the results are very similar; hence, the RRE is reliable to describe the degradation process on average. Indeed, the RRE is equivalent (except for the scaling of the volume) to the evolution of the expected value $< n(t) >= \sum_{n=0}^{n_0} nP(n,t)$ computed using the CME, that in this particular case is possible to solve explicitly. On the other hand, τ-leaping and E-M give slightly different results, probably because the population is not large enough to have the assumptions satisfied. In any case, for the number of simulations considered, the SSA still represents an efficient method to include stochasticity exactly and study single realizations. If, instead, the number of simulations become very large one should probably consider τ-leaping or E-M method to save some computational time, meanwhile allowing for some losses in accuracy.

We have run 100 simulations and the time spent by the different algorithms is the following:

SSA: 1.1487 s, τ-leaping: 1.4348 s, E-M: 1.2621 s, RRE: 0.0628 s.

As we were mentioning, from this small number of simulations we cannot appreciate the velocity of the τ-leaping or E-M algorithm, this should become significant when the number of simulations is very high ($\sim$ 10000 using adaptive τ [20]) or when the population is large. Whereas , it is already evident why RRE is the most preferred, when applicable.

Moreover, it is worth noting that all algorithms return time and state vectors of different dimension and scale, because of the presence of SSA or adaptive τ, hence, one has to be careful and standardize them if some operations need to be performed. For example, it was necessary to obtain the average and standard deviation of the simulations; in particular, we have used Matlab *interp1* to have vectors "equally spaced" and of the same dimension.

Finally, we have also tried to use Matlab to simulate the CLE, through the command *simByEuler*, the resulting simulations are reported below.

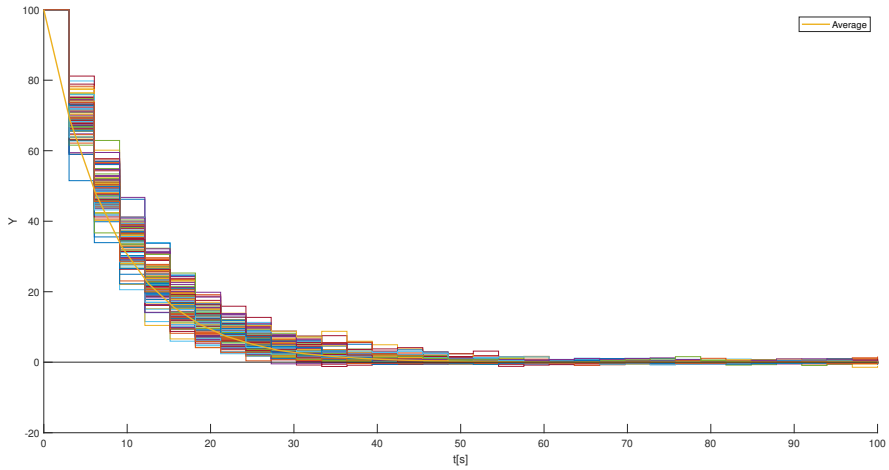


Figure 2.6: 100 simulations of degradation process by Matlab *simByEuler*, with $k = 0.1\,s^{-1}$ and $\tau = 3$.

As one can see from the figure above, the number of molecules is sometime negative, and probably there are moments in which the assumptions of the CLE are not satisfied, indeed in our implementation was necessary to use Gillespie's SSA. Using Matlab to simulate the CLE represents a faster and easier way, though, it does not allow to "control" if we are satisfying all the requirements, or if the number of molecules is negative. This is why throughout this thesis we will use our implementation of the E-M method.

Comparison of different modeling and simulation methods

The following table summarizes the key features of each model and its corresponding simulations, as discussed so far.

Table 2.3: Comparison between different model and simulation methods

<table>
<tr><th>MATHEMATICAL MODEL</th><th>SIMULATION TOOL</th></tr>
<tr><td>Chemical Master Equation</td><td>Gillespie SSA</td></tr>
<tr><td rowspan="3">Advantages:
Exact stochastic linear model.
Useful with low number of molecules.
Limitations:
Describes the evolution of the probability distribution.
Dimension equal to total number of possible states.
Usually difficult to handle.</td><td>Advantages:
Equivalent to CME.
Advantageous when molecular population levels can change suddenly.
Limitations:
Can be very slow when the number of frequent reactions is high.
Updates the state every time a reaction occurs.</td></tr>
<tr><td>τ-leaping approximation</td></tr>
<tr><td>Advantages:
Groups together more reactions.
Can be fast if propensity functions do not change significantly.
Limitations:
Not always valid.
Not trivial implementation.
Necessity to handle negative numbers of molecules.</td></tr>
<tr><td colspan="2">If there exists τ such that the propensity functions do not significantly change in it and $a_j(X(t))\tau \gg 1$, then:</td></tr>
</table>

<table>
<tr><th>MATHEMATICAL MODEL</th><th>SIMULATION TOOL</th></tr>
<tr><td>Chemical Langevin Equation</td><td>E-M method</td></tr>
<tr><td>Advantages:
SDE with a drift and a diffusion part.
Dimension of the model equal to number of species.
Describes evolution of the number of molecules.
There exist numerical methods to solve it.
Limitations:
Not always valid.
The state is a real continuous r.v. (not integer).</td><td>Advantages:
Very fast if the number of simulations or molecules is high.
Limitations:
Difficult to choose the correct τ.
Not versatile τ.
Necessity to handle negative numbers of molecules.</td></tr>
<tr><td colspan="2">If the population and the volume tend to infinity, then:</td></tr>
<tr><td>Reaction Rate Equation</td><td>ODEs solver tools</td></tr>
<tr><td>Advantages:
Deterministic model.
Dimension of the model equal to the number of species.
Gives an average analysis.
Limitations:
Often not reliable.
Describes evolution of the concentration of molecules.</td><td>Advantages:
There exist available solvers.
Fast.
Limitations:
Not possible to see individual simulations.</td></tr>
</table>

2.3 Application examples

In this section we present two main examples within the biological context. The objective is to illustrate the different modeling and simulation methods, highlighting the differences among them, and identifying which is the most suitable depending on the application. For this reason, the coefficients' values have been chosen arbitrarily, or to be able to apply all the simulation methods, and do not correspond to real-world scenarios.

For each example we will derive the Chemical Master Equation, the Chemical Langevin Equation and the Reaction Rate Equation, and we will simulate all of them with the methods that have been presented earlier:

- Gillespie's SSA to simulate the CME.
- Adaptive τ-leaping approximation to simulate the CME as described in [21], with variability threshold of the propensity functions equal to $\varepsilon = 0.2$. The algorithm switches to Gillespie's SSA when assumptions are not satisfied or a negative result would be returned.
- E-M method with fixed τ to simulate the CLE and variability threshold of the propensity functions equal to $\varepsilon = 0.2$. The algorithm switches to Gillespie's SSA when assumptions are not satisfied or a negative result would be returned.
- Matlab *ode45* to solve the RRE. This will be compared with the averages of the simulations obtained by the stochastic methods.

For simplicity and clarity, we will often omit the dependence on time t in the number or concentration of molecules, assuming that they always depend on time. Moreover, we will distinguish the number of molecules X from their concentration denoting the last one as $[X]$.

Birth-Death process

$$\emptyset \xrightarrow{k_1} X \xrightarrow{k_2} \emptyset, \qquad \text{with } k_1 = 5, k_2 = 0.1s^{-1}.$$

From the pseudo-reactions we can write the propensity functions as $a_1 = k_1$, $a_2 = k_2 X$, and the stoichiometric vectors as $\nu_1 = 1$ and $\nu_2 = -1$. The choice of k_1 and k_2 is not casual in this case, it is what allows us to apply the τ-leaping approximation and the E-M method. Indeed, as we will show, the system will tend to k_1/k_2 and if this is too small, then it would be more reasonable to use Gillespie's SSA.

From (2.1), the infinite dimensional CME for the birth-death process can be written as:

$$\frac{dP(n,t)}{dt} = -(k_1 + k_2\, n)\, P(n,t) + k_1\, P(n-1,t) + k_2\,(n+1)\, P(n+1,t), \qquad \text{for } n = 0,1,...,\infty.$$

It is also straightforward to write the CLE as in (2.4):

$$\frac{dX}{dt} = k_1 - k_2 X + \sqrt{k_1}\, dW_1 - \sqrt{k_2 X}\, dW_2, \qquad \text{with } dW_j \sim \mathcal{N}(0,1).$$

Notice that $\sqrt{k_1}\, dW_1$ can be seen as a normal distribution with zero mean and standard deviation equal to $\sqrt{k_1}$, and similarly $\sqrt{k_2 X}\, dW_2$ as a normal distribution with zero mean and standard deviation equal to $\sqrt{k_2 X}$. Then, exploiting the property of the linear combination of two gaussian variables, we can claim that their sum is a normal distribution with zero mean and standard deviation equal to $\sqrt{k_1} + \sqrt{k_2 X}$. Hence, the CLE can be written also as:

$$\frac{dX}{dt} = k_1 - k_2 X + \left(\sqrt{k_1} + \sqrt{k_2 X}\right) dW, \qquad \text{with } dW \sim \mathcal{N}(0,1).$$

To write the RRE, we recall that we can ignore the noisy part of the CLE, and write X on the right-hand side of the equation in terms of the concentration $[X]$:

$$\frac{d[X]\,\Omega}{dt} = k_1 - k_2\,[X]\,\Omega, \quad \text{where } \Omega \text{ is the Volume of the cell where the reactions take place.}$$

In principle, there should also be the Avogadro number if one is considering the molar concentration, though, for our purposes of only simulating the system, we can ignore it.
The following Reaction Rate Equation is then obtained:

$$\frac{d[X]}{dt} = \frac{k_1}{\Omega} - k_2\,[X]$$

Looking at the RRE we can easily find also its solution, given by

$$[X](t) = \frac{k_1}{k_2\,\Omega} + \left(X(0) - \frac{k_1}{k_2\,\Omega}\right) e^{-k_2 t} \qquad \xrightarrow{t\to\infty} \qquad \mu = \frac{k_1}{k_2\,\Omega}.$$

We can now proceed to the simulation phase, we have performed 100 simulations using Gillespie's SSA, τ-leaping and E-M method as previously described. Specifically, the value of $\tau = 3$ for the

E-M algorithm has been selected, so that $a_j\tau > 10$ most of the time:

$$\tau > \frac{10}{a_1} = \frac{10}{k_1} = 2 \qquad \text{and} \qquad \tau > \frac{10}{a_2} = \frac{10}{k_2 X} = \frac{100}{X}, \quad \text{for } X \in [0, \sim 80].$$

At the same time, in the majority of the cases, $\tau = 3$ is also satisfying the condition $\tau > \frac{2}{a_0} = \frac{2}{k_1 + k_2 X}$ and is such that the propensity functions are almost constant. The results obtained are shown in the figure below with $\Omega = 1$ for simplicity, since in any case we are not considering a real scenario.

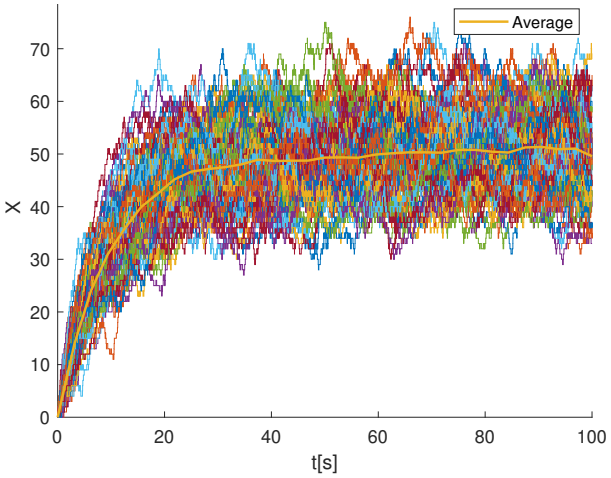


Figure 2.7: 100 simulations of birth-death process by SSA, with $k_1 = 5, k_2 = 0.1\,s^{-1}$.

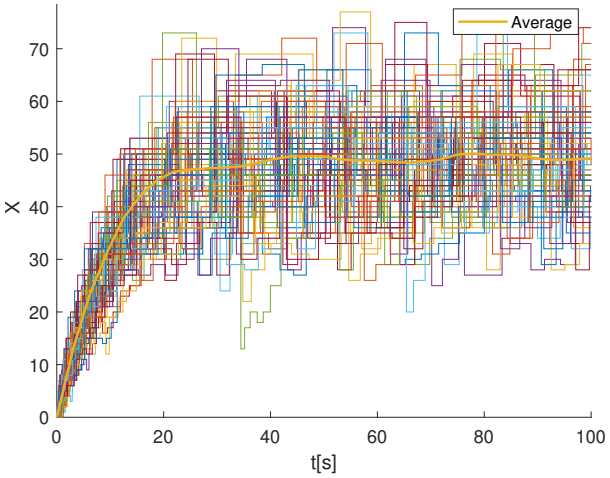


Figure 2.8: 100 simulations of birth-death process by adaptive τ-leaping method, with $k_1 = 5, k_2 = 0.1\,s^{-1}$ and $\varepsilon = 0.2$

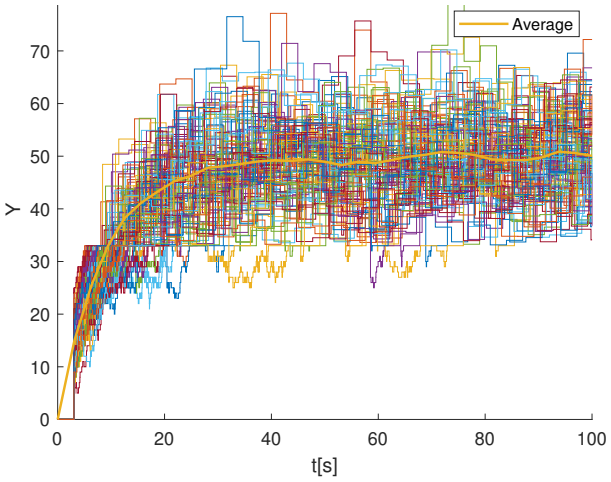


Figure 2.9: 100 simulations of birth-death process by E-M method, with $k_1 = 5, k_2 = 0.1\,s^{-1}$, $\tau = 3$ and $\varepsilon = 0.2$.

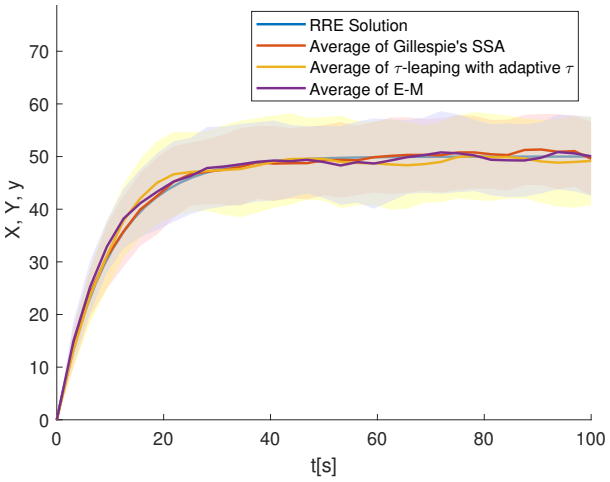


Figure 2.10: Comparison of RRE solution and averages of the simulations obtained from SSA, τ-leaping and E-M for the birth-death process, with $k_1 = 5, k_2 = 0.1\,s^{-1}$. Standard deviation from the averages is shown with light colored bands.

Let us first consider Fig. 2.10 from which it is possible to observe that all the methods yield similar

average results that show the same trend as the one obtained by the RRE. Normally, in a real case scenario, there will be a different scaling in the RRE because of the volume factor, but the trend would still be the same. Indeed, it is possible to show that the evolution of the expected value computed using the CME matches that described by the RRE, aside from the volume scaling factor. Hence, we can claim that all the methods are in this case valid, and if someone is interested in the average behavior, RRE provides reliable and fast results. Though, using the deterministic model one loses information on the real system, which in this case presents high variability between the trajectories, i.e. different cells could end up in a very different state.

We recall that both τ-leaping and E-M methods have been implemented so that Gillespie's SSA is used when one of the assumptions does not hold, the number of molecules is negative, or τ is quite small that it is more convenient to use the SSA, since it is exact. We would like to highlight that in the simulations reported here τ-leaping uses the SSA only at the very beginning for $t < 1,2,3\ s$ since the resulting τ value would be too small; similarly, the E-M method uses the SSA at the first iteration since the state is equal to zero, and then rarely calls it again. This means that in the given example, both methods are applicable, although, it is worth noticing that using different reaction coefficients, it is possible that this may no longer be true. Using for instance $k_1 = 1$, we have noticed that Gillespie's SSA was almost always used when calling τ-leaping or E-M, probably because the number of molecules was too low, indeed the steady-state equilibrium was 10, instead of 50.

Finally, we report the time required by each algorithm:

SSA: 1.6769 s, τ-leaping: 1.7354 s, E-M: 1.9102 s, RRE: 0.053003 s.

As we can see, neither E-M, neither τ-leaping are faster than the SSA; in this case it is actually more convenient to use Gillespie's SSA. We expect τ-leaping and E–M to be more efficient when the number of molecules or reactions is higher. Indeed, increasing the steady-state equilibrium, τ-leaping and E-M become more efficient, for instance, using $k_1 = 100$, we got the following computational times:

SSA: 2.9039 s, τ-leaping: 1.675 s, E-M: 1.661 s, RRE: 0.064063 s,

from which one sees that E-M is slightly faster than τ-leaping, that is faster than the SSA.

Therefore, in general, one has to pay attention to weather τ-leaping and E-M are truly applicable and more efficient than Gillespie's SSA; if so, they can offer significant improvements in speed and computational efficiency. The user should choose the simulation method carefully, depending on the application.

Two Species Compound process

$$X + Y \underset{k_2}{\overset{k_1}{\rightleftarrows}} X:Y, \qquad \text{with } k_1 = 0.01\,s^{-1},\ k_2 = 1\,s^{-1}.$$

Let us call $x = [X \quad Y \quad X:Y]^T$ the state, then, the propensity functions a_j and the stoichiometric vectors ν_j for both reactions are the following:

- First reaction: $a_1 = k_1 X Y$ and $\nu_1 = \begin{bmatrix} -1 & -1 & +1 \end{bmatrix}^T$.
- Second reaction: $a_2 = k_2 X:Y$ and $\nu_2 = \begin{bmatrix} +1 & +1 & -1 \end{bmatrix}^T$.

Again, the choice of the reaction coefficients has been made so that it was feasible to apply τ-leaping and E-M methods, indeed, if the coefficients are such that the number of molecules is pretty low, then the algorithms will almost only apply Gillespie. This happens for instance for $k_1 = 0.01\,s^{-1}, k_2 = 0.1\,s^{-1}$, because in this case the number of free molecules X is smaller.

To write the CME, it is easier to consider the state changes due to only one reaction, as shown below, since we supposed that in dt only one reaction will take place.

$$\begin{bmatrix} n_X + 1 \\ n_Y + 1 \\ n_{X:Y} - 1 \end{bmatrix} \underset{a_2}{\overset{a_1}{\rightleftarrows}} \begin{bmatrix} n_X \\ n_Y \\ n_{X:Y} \end{bmatrix} \underset{a_1}{\overset{a_2}{\rightleftarrows}} \begin{bmatrix} n_X - 1 \\ n_Y - 1 \\ n_{X:Y} + 1 \end{bmatrix},$$

where $n_X =$ number of free molecules X, $n_Y =$ number of free molecules Y, and $n_{X:Y} =$ number of molecules $X:Y$. Let us call N_X the total number of molecules X (both free and in $X:Y$) and N_Y the total number of molecules Y (both free and in $X:Y$), with $N_Y \geq N_X$, w.l.o.g.. Then, it is possible to write the following finite dimensional Chemical Master Equation:

$$\begin{aligned} \frac{dP(n_X, n_Y, n_{X:Y}, t)}{dt} = & k_1 (n_X + 1)(n_Y + 1) P(n_X + 1, n_Y + 1, n_{X:Y} - 1, t) + \\ & + k_2 (n_{X:Y} + 1) P(n_X - 1, n_Y - 1, n_{X:Y} + 1, t) - [k_1 n_X n_Y + k_2 n_{X:Y}] P(n_X, n_Y, n_{X:Y}, t), \end{aligned}$$

with $n_X = 0, \ldots, N_X$, $n_Y = 0, \ldots, N_Y$ and $n_{X:Y} = 0, \ldots, N_X$.

Actually, it is possible to simplify this expression and obtain the state in only one dimension simply by observing that there are no birth or death processes, hence the total number of molecules will be

constant over time. This allows to write Y and $X:Y$ in terms of X as follows:

$$\begin{aligned} n_X + n_{X:Y} &= N_X \\ n_Y + n_{X:Y} &= N_Y \end{aligned} \quad \Longrightarrow \quad \begin{aligned} n_{X:Y} &= N_X - n_X \\ n_Y &= N_Y - N_X + n_X, \text{ with } N_Y \geq N_X. \end{aligned}$$

Therefore, the evolution of the state is fully described by only the evolution of $n_X(t) =: n(t)$, this brings a big advantage to the model moving the system from a three dimensional one to a scalar one. We can now write the CME in $P(n,t)$ as:

$$\begin{aligned} \frac{dP(n,t)}{dt} =& k_1 (n+1)(N_Y - N_X + n + 1) P(n+1,t) + k_2 (N_X - n + 1) P(n-1,t) + \\ & - [k_1 n (N_Y - N_X + n) + k_2 (N_X - n)] P(n,t), \qquad n = 0, ..., N_X. \end{aligned}$$

Similarly, we can write again the propensity functions and the stoichiometric vectors considering only the number of free molecules X:

- First reaction: $a_1 = k_1 X (N_Y - N_X + X)$ and $\nu_1 = -1$.
- Second reaction: $a_2 = k_2 (N_X - X)$ and $\nu_2 = +1$.

From here we can easily obtain the CLE:

$$\frac{dX}{dt} = -a_1 + a_2 - \sqrt{a_1}\, dW_1 + \sqrt{a_2}\, dW_2, \qquad \text{with } dW_j \sim \mathcal{N}(0,1).$$

Exploiting the properties of linear combination of independent Gaussian variables as done in the Birth-Death process, we can write it also as:

$$\frac{dX}{dt} = -a_1 + a_2 + (-\sqrt{a_1} + \sqrt{a_2})\, dW, \qquad \text{with } dW \sim \mathcal{N}(0,1).$$

Now we can substitute the expressions of a_j and ν_j, and write:

$$\frac{dX}{dt} = -k_1 X (N_Y - N_X + X) + k_2 (N_X - X) + \left(-\sqrt{k_1 X (N_Y - N_X + X)} + \sqrt{k_2 (N_X - X)} \right) dW.$$

From here it is easy to get the RRE, though, we have to be careful about the coefficients' values since we have a biomolecular reaction. In particular, if we neglect the noisy terms and we write $X = [X]\,\Omega$, with $\Omega =$ Volume of the cell we are considering (not including the Avogadro number always because it will only scale the coefficients), then we obtain:

$$\frac{d[X]\,\Omega}{dt} = -k_1\,[X]\,\Omega\,(N_Y - N_X + [X]\,\Omega) + k_2\,(N_X - [X]\,\Omega),$$

which, calling $\Delta N = N_Y - N_X$, can be written as

$$\frac{d[X]}{dt} = -k_1\,[X]\,\Omega\left(\frac{N_Y}{\Omega} - \frac{N_X}{\Omega} + [X]\right) + k_2\left(\frac{N_X}{\Omega} - [X]\right) = k_2\,\frac{N_X}{\Omega} - (k_1\,\Delta N + k_2)\,[X] - k_1\,\Omega\,[X]^2.$$

For simulations' purposes here we have made the easiest choice of $\Omega = 1$, knowing that it should change depending on the application. In general, different values of the volume can cause the RRE to behave differently than the other models; in this case we have tested it changing Ω, and we have verified that it acts as the mean of the simulations, just scaled by the volume.

We have performed 100 simulations of the different algorithms that implement the models derived above in the variable X since the evolution of the other molecules can be derived from it, as we have seen. We would like to highlight that with E-M it was necessary to relax the inequality on the variability of the propensity functions by increasing the value of ε. This can cause losses in accuracy, though, it is a price to pay to implement E-M, otherwise it was almost always using Gillespie's SSA because the propensity functions were changing very quickly, which implies in any case that the time step must be quite small. We have chosen $\tau = 0.7$, trying to satisfy the usual conditions. Having made this choice, we were able to perform the $E-M$ method, even though the time step was small (not be particularly efficient), and ε was quite big (larger changes in the propensity functions). The results of the simulations with $N_X = 50$, $N_Y = 60$ and $X(0) = N_X$ are visible in the figure below.

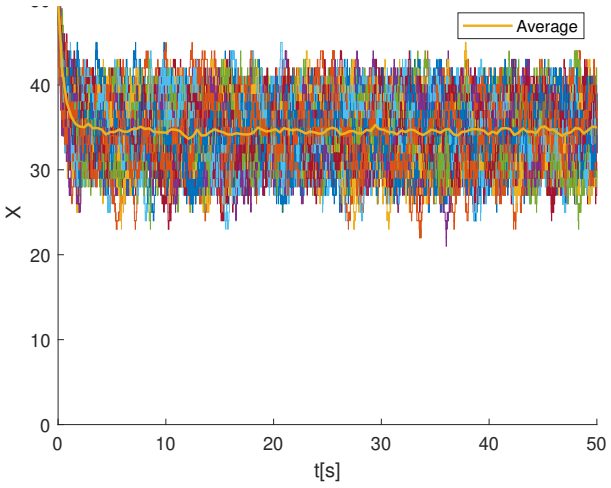


Figure 2.11: 100 simulations of compound formation process by SSA, with $k_1 = 0.01\,s^{-1}$, $k_2 = 1\,s^{-1}$.

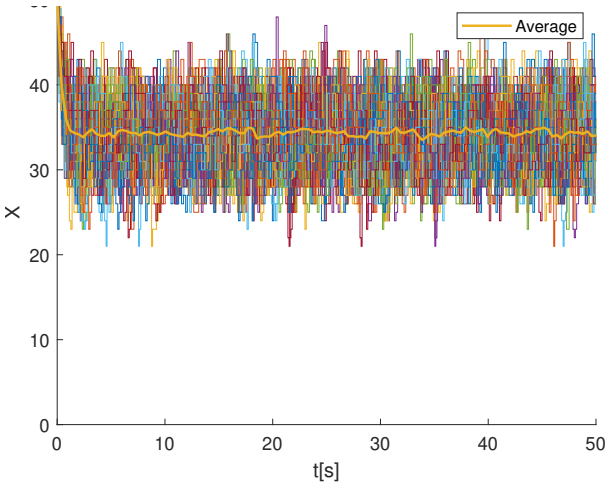


Figure 2.12: 100 simulations of compound formation process by adaptive τ-leaping method, with $k_1 = 0.01\,s^{-1}$, $k_2 = 1\,s^{-1}$ and $\varepsilon = 0.2$.

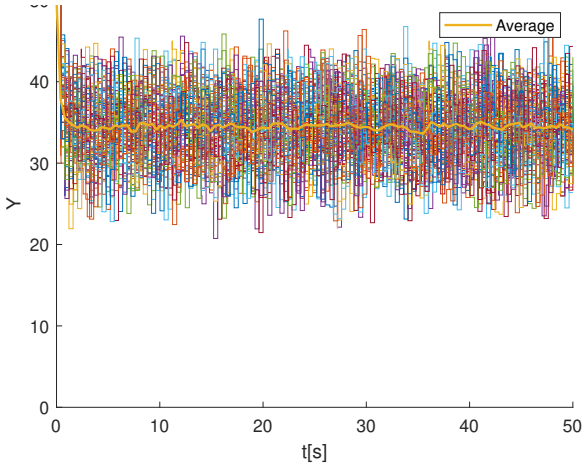


Figure 2.13: 100 simulations of compound formation process by E-M method, with $k_1 = 0.01\,s^{-1}, k_2 = 1\,s^{-1}, \tau = 0.7$ and $\varepsilon = 0.4$.

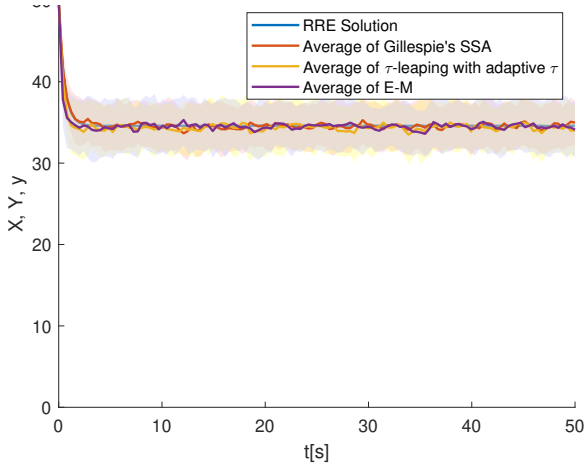


Figure 2.14: Comparison of RRE solution and averages of the simulations obtained from SSA, τ-leaping and E-M for the compound formation process, with $k_1 = 0.01\,s^{-1}, k_2 = 1\,s^{-1}$. Standard deviation from the averages is shown with light colored bands.

Also in this case, as we could expect, the results of the different models are pretty similar, and the considerations previously done still hold true. Moreover, it is noteworthy that in the E-M algorithm it was necessary to call Gillespie's SSA also when the state update was causing $X > N_X$, which is of course not feasible in reality. This additional case was similarly added also in the implementation of the τ-leaping approximation, even if this was performing particularly well with the chosen parameters and the SSA was actually never utilized. In any case, it is necessary to add it if for example we increase ε causing higher variability that can lead to $X > N_X$. In fact, the most correct way to handle these cases would probably be to introduce critical reactions, as has been proposed in the literature [24] when a negative number of molecules is returned.

However, we can claim that the simulations obtained in the described way are all reliable. The execution times are reported below:

SSA: 1.9338 s, τ-leaping: 2.1343 s, E-M: 1.5504 s, RRE: 0.050579 s.

We can easily see that in this case the τ-leaping approximation is not efficient, probably because the time steps must be very small since the propensity functions change very quickly. The E-M method gives a more satisfactory result, but we have to recall that it was necessary to choose a larger ε. Therefore, if one is interested in the average trend, the RRE are without any doubt the best choice, whereas, if one wants to look at the single realizations, in this case probably Gillespie's SSA still represents the best option since it does not do any approximation and it takes almost the same time as the other methods.

Chapter 3

From RRE to Simplified Stochastic Models, via Quasi-Steady-State-Assumption

Stochastic modeling provides a very rich description of biological systems, especially when dealing with complex systems where fluctuations can significantly influence reaction outcomes. The high variability of stochastic trajectories is already evident in the Birth-Death process studied in the previous chapter, even if the average of the simulations followed the trend of the deterministic solution. In many cases, accounting for stochasticity is not only interesting but also essential, especially when fluctuations have a profound effect on the physiology of the cell and statistical averages do not accurately describe the chemical dynamics inside it. This happens, for example, in bistable systems where stochasticity introduces the possibility of switching between the two stable states, even if starting from identical initial conditions or from the opposite basin of attraction. However, performing stochastic simulations can be computationally demanding, especially when using Gillespie's Stochastic Simulation Algorithm, as we have seen. Therefore, finding an efficient and reliable approach is crucial to save time and computations.

The objective of this chapter is to propose a novel and powerful method to perform stochastic simulations by reducing system's dimension, thereby decreasing computational complexity. This approach can be particularly effective with complex and huge systems involving numerous reactions, as can happen when transcription factors are employed in gene expression or protein production. The method we propose lies on time scale separation, i.e. the idea that there exist fast and slow dynamics and that they can be treated differently. This is often observed in systems where reactions involve more than two molecules, actually hiding underlying fast intermediate elementary reactions. When treating these cases with RRE, there already exists a way to reduce the system [11] based on the fact that fast dynamics can be considered at steady state after an initial fast transient (relative to the timescale of interest for the slower variables). Hence, by imposing $\frac{d(\cdot)}{dt} = 0$ for the fast dynamics, one derives the corresponding steady-state value and can substitute it in the slow

dynamics, usually obtaining Hill-type functions. The time scale separation in this case reduces to the so-called quasi-steady-state assumption (QSSA), which can be formally derived also from singular perturbation theory [68].

This approach is a typical approximation technique used with RRE, which, due to its great potential, has motivated researchers to explore ways to apply it also in a stochastic context. Notable examples include its application to Chemical Master Equations [69–72], Gillespie's SSA [73], and τ-leaping approximation [74]. Taking inspiration from these papers, here we propose a method (that should be formally proven, but this goes beyond the possibilities of this thesis) to reduce stochastic models and hence ease their simulations under the same hypothesis of the QSSA. The idea is to start from RRE, apply the QSSA, then substitute the resulting steady-state expressions in the slow dynamics, identify some kind of new propensity functions, and finally use them to write the new stochastic models and to perform the corresponding simulations. It is worth to make a small note on this passage from RRE to stochastic models: the former indeed describes the evolution of the concentration, whereas the latter of the number of molecules. Thus, once one obtains the reduced RRE, before deriving the new propensity functions, it is necessary to use the definition of the concentration to obtain the correct values of the parameters, which could be scaled by the volume with respect to the ones appearing in the RRE, as explained in the previous chapter.

This method offers a promising way to ease the simulation process by reducing the number of species and reactions, and at the same time to bypass the problem of estimating a large quantity of parameters about which we usually do not possess enough information. By adopting this approach it is no longer necessary to make use of detailed first-principles to develop stochastic models; they can now be derived from existing robust ODE models that may encapsulate detailed chemical kinetics by various Hill functions and quasi-steady-state assumptions [74].

A common example, that we will soon examine, is the genetic toggle switch. Since transcription factors are involved in the reactions and they bound very rapidly to the DNA promoter sites, the described method can be applied by considering their equations to be at steady state and using them inside the dynamics of the mRNA. In this way, the mRNA equations can be written via Hill functions of the transcription factors. Similarly, since the dynamics of the mRNA are faster than those of proteins, it is possible to further simplify the system by reducing it to the protein dynamics alone, with the mRNA and other species incorporated into the proteins activation functions. The detailed method and results are provided in the following pages.

3.1 The Genetic Toggle Switch

The genetic toggle switch is a synthetic bistable gene-regulatory network which has analogous functions to its electronic counterpart. In particular, it acts as a binary memory element exhibiting two different stable states, each corresponding to either the production or inhibition of a specific protein. This bistable nature requires the use of stochastic modeling because deterministic ones would not be sufficient to describe all the possible behaviors and switchings the toggle switch can encounter. Multistability, in general, plays a significant role in some of the basic processes of biological life and evolution. Indeed, it might account for the maintenance of phenotypic differences in the absence of genetic or environmental variations, or it can be used to explain cell differentiation [75]. Therefore, being able to properly model multistable systems can be of great importance. Stochastic modeling represents a reliable way to do it, even though it often requires detailed information about chemical kinetics and computationally intense simulations, especially when there is a large number of reactions, as in the case of the toggle switch. For these reasons, model reduction through QSSA, including the noise as previously described, becomes particularly valuable.

Implementation

The genetic toggle switch can be engineered by inserting into a living cell a new piece of DNA describing the production of two repressing proteins creating a double negative feedback. This is such that if there is abundance of one protein then the other one is almost not produced and we consider the system to be in *State 1*; otherwise, if there is abundance of the second protein, the production of the first one is repressed and we can say we are in *State 2*. Such behavior can be obtained by making use of inducible promoters, which are such that the transcription of the corresponding genes can be controlled by changing the external concentration of an inducer. Their general functioning is shown in the image below. Observe that inducible promoters provide a simple way of switching genes on and off, or in general of controlling their expression level; hence, the genetic toggle switch exploits exactly this principle.

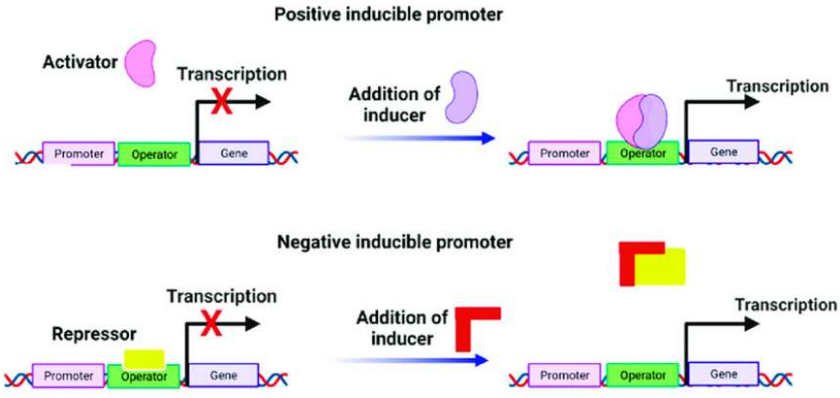
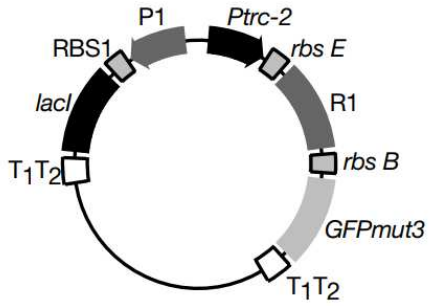
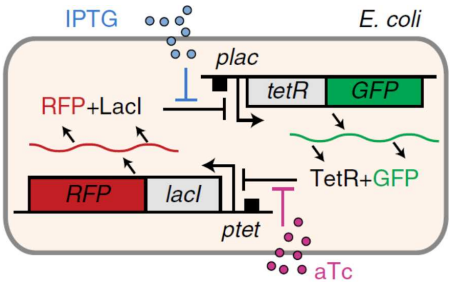


Figure 3.1: Schematic regulation with positive and negative inducible promoters.

Gardner and Collins implemented the genetic toggle switch *in vivo* for the first time in the Escherichia coli [76], constructing the plasmid in Fig.3.2a through typical biological techniques. In particular, the circuit includes the lac and tet promoter, and the genes encoding the LacI and TetR repressor proteins, which can bound to the isopropyl β-D-1-thiogalactopyranoside (IPTG) and anhydrotetracycline (aTc) inducers. The advantage of using IPTG and aTc is that they have virtually no effect on the cell other than changing the binding affinity of the respective repressor, and thus they can be used to control the functioning of the promoters. Moreover, the lac and tet promoters also control the expression of a reporter gene (e.g. a fluorescent protein) that is produced proportionally to the LacI or TetR protein; hence, it can be used in the experiments to analyze the quantity of each protein through a fluorescence microscope. In this case, the reporter genes are the green fluorescent protein (GFP), which will be proportional to TeR, and the red fluorescent protein (RFP), proportional to LacI. The plasmid implementing this circuit and the relative scheme of the main reactions are shown in the following figure.



(a) The toggle switch plasmid. Promoters are marked by solid rectangles with arrowheads. Genes are denoted with solid rectangles. Ribosome binding sites and terminators are denoted by outlined boxes [76].



(b) Genetic Toggle Switch Scheme: Proteins LacI and TetR mutually repress each other's promoter, unless they bind with inducers aTc or IPTG. The quantity of LacI and TetR is proportional to reporter genes RFP and RFP rispectively.

The toggle switch so implemented constitutes a bistable system, where in one state LacI is mainly produced, and in the other one the production of TetR is dominating. The inducers can be added to perform switching between these two states by inhibiting the repressing action of the protein to which they can bind. As anticipated, the toggle switch is of fundamental importance in biological systems because it plays a key role in processes like cell differentiation and decision making. Indeed, it allows cells to have memory of some previous stimulus by maintaining a high expression level of a specific repressor protein [77]. Additionally, it is at the base of an emerging challenge in synthetic biology: Ratiometric Control Problem using a Single Population. This will be discussed in the next chapter, here we focus, instead, on the toggle switch model and analysis.

3.1.1 Deterministic Model using QSSA and Bistability Analysis

A schematic representation of the toggle switch was shown in Fig. 3.2b, though, in reality many other reactions are present. The total set of reactions is shown below. On the left and right sides we have reported the reactions involved with the production of LacI and TetR respectively. We have denoted by DNA_L and DNA_T the operator region of the DNA where LacI and TetR can bind and hence act as repressors. The remaining notations should be self-explanatory.

Transcription

$$DNA_L + RNAp \underset{k_L^2}{\overset{k_L^1}{\rightleftarrows}} DNA_L : RNAp \qquad DNA_T + RNAp \underset{k_T^2}{\overset{k_T^1}{\rightleftarrows}} DNA_T : RNAp$$

$$DNA_L : RNAp \xrightarrow{k_L^3} mRNA_L + DNA_L + RNAp \qquad DNA_T : RNAp \xrightarrow{k_T^3} mRNA_T + DNA_T + RNAp$$

Translation

$$mRNA_L + Ribo \underset{k_L^5}{\overset{k_L^4}{\rightleftarrows}} mRNA_L : Ribo \qquad mRNA_T + Ribo \underset{k_T^5}{\overset{k_T^4}{\rightleftarrows}} mRNA_T : Ribo$$

$$mRNA_L : Ribo \xrightarrow{k_L^p} mRNA_L + Ribo + LacI \qquad mRNA_T \xrightarrow{k_T^p} mRNA_T + TetR$$

Repressing action

$$TetR + DNA_L \underset{k_L^7}{\overset{k_L^6}{\rightleftarrows}} TetR : DNA_L \qquad LacI + DNA_T \underset{k_T^7}{\overset{k_T^6}{\rightleftarrows}} LacI : DNA_T$$

Proteins-Inducers interaction

$$TetR + aTc \underset{k_L^9}{\overset{k_L^8}{\rightleftarrows}} TetR : aTc \qquad LacI + IPTG \underset{k_T^9}{\overset{k_T^8}{\rightleftarrows}} LacI : IPTG$$

Inducers diffusion

$$u_{aTc} \xrightarrow{k_{aTc}} aTc \qquad u_{IPTG} \xrightarrow{k_{IPTG}} IPTG$$

Degradation

$$mRNA_L \xrightarrow{g_L^m} \emptyset \qquad mRNA_T \xrightarrow{g_T^m} \emptyset$$

$$mRNA_L : Ribo \xrightarrow{g_L^m} Ribo \qquad mRNA_T : Ribo \xrightarrow{g_T^m} Ribo$$

$$LacI \xrightarrow{g_L^p} \emptyset \qquad TetR \xrightarrow{g_T^p} \emptyset$$

$$aTc \xrightarrow{k_{aTc}} \emptyset \qquad IPTG \xrightarrow{k_{IPTG}} \emptyset$$

The number of reactions and parameters involved is very high, this is why we were anticipating that model reduction can be fundamental in this case. The reduced deterministic model below can be obtained by re-arrenging the equations to only leave the independent ones, and by applying the QSSA since reactions can reasonably be distinguished between fast and slow ones because of the presence of transcription factors.

$$\begin{cases} \frac{d}{dt}[mRNA_{LacI}] = k_L^{m0} + \frac{k_L^m}{1+\left(\frac{[TetR]}{\theta_{TetR}}\frac{1}{1+\left(\frac{[aTc]}{\theta_{aTc}}\right)^{\eta_{aTc}}}\right)^{\eta_{TetR}}} - g_L^m\,[mRNA_{LacI}] \\ \frac{d}{dt}[mRNA_{TetR}] = k_T^{m0} + \frac{k_T^m}{1+\left(\frac{[LacI]}{\theta_{LacI}}\frac{1}{1+\left(\frac{[IPTG]}{\theta_{IPTG}}\right)^{\eta_{IPTG}}}\right)^{\eta_{LacI}}} - g_T^m\,[mRNA_{TetR}] \\ \frac{d}{dt}[LacI] = k_L^p\,[mRNA_{LacI}] - g_L^p\,[LacI] \\ \frac{d}{dt}[TetR] = k_T^p\,[mRNA_{TetR}] - g_T^p\,[TetR] \\ \frac{d}{dt}[aTc] = k_{aTc}\,(u_{aTc} - [aTc]) \\ \frac{d}{dt}[IPTG] = k_{IPTG}\,(u_{IPTG} - [IPTG]) \end{cases} \tag{3.1}$$

The model describes the evolution of 6 species interacting through 12 reactions, a significant simplification from the original system which involved 18 species and 30 reactions, thus one can see why this method could be essential and pivotal. From the model, it is possible to see that the effects of the repressor proteins and of the inducers are encapsulated in the mRNA equations through Hill functions, as anticipated. This is the typical way these interactions can be described; and hence one can directly assume this Hill-type shape for the activation functions and fit the model in (3.1) to some calibration data, using a global optimization tool, as done in [78]. The values found in this way are the ones we will refer to throughout the thesis, and are summarized in Table 3.1 below.

The main advantage of applying the QSSA is that only the necessary parameters' values must be estimated, without the need of knowing in advance all the coefficients appearing in the pseudo-reactions. This efficiency, along with the simplicity of the resulting model, has led to the use of (3.1) in numerous studies. Examples include [78] and [79], where the objective was to control the system at its unstable equilibrium, similarly to the classic inverted pendulum problem; [44], which focuses on ratio control of a population of cells endowed with the toggle switch; [77] that investigates the dynamics of the system when subject to pulse-width modulated inputs; and [30], where a population endowed with the toggle switch is controlled by a multicellular controller.

Parameters	Values	Parameters	Values
k_L^{m0}	0.3045 mRNA min^{-1}	k_T^{m0}	0.3313 mRNA min^{-1}
k_L^m	13.01 mRNA min^{-1}	k_T^m	5.055 mRNA min^{-1}
k_L^p	0.6606 a.u. mRNA min^{-1}	k_T^p	0.5098 a.u. mRNA^{-1}min^{-1}
g_L^m	0.1386 min^{-1}	g_T^m	0.1386 min^{-1}
g_L^p	0.0165 min^{-1}	g_T^p	0.0165 min^{-1}
θ_{LacI}	124.9	θ_{TetR}	76.40
η_{LacI}	2.00	η_{TetR}	2.152
θ_{aTc}	35.98	η_{aTc}	2.00
θ_{IPTG}	0.2926	η_{IPTG}	2.00
k_{aTc}	0.04 min^{-1}	k_{IPTG}	0.04 min^{-1}

Table 3.1: Parameters and corresponding values for the genetic toggle switch [78].

The model in (3.1) can be further reduced by applying again the QSSA to the equations regarding the mRNAs and the inducers, which is possible because they degrade faster than the proteins. In this way, the following approximated reduced model in only the repressor proteins is obtained.

$$\begin{cases} \frac{d}{dt}[LacI] = \frac{k_L^p k_L^{m0}}{g_L^m} + \frac{\frac{k_L^p k_L^m}{g_L^m}}{1+\left(\frac{[TetR]}{\theta_{TetR}} \frac{1}{1+\left(\frac{u_{aTc}}{\theta_{aTc}}\right)^{\eta_{aTc}}}\right)^{\eta_{TetR}}} - g_L^p [LacI] \\ \frac{d}{dt}[TetR] = \frac{k_T^p k_T^{m0}}{g_T^m} + \frac{\frac{k_T^p k_T^m}{g_T^m}}{1+\left(\frac{[LacI]}{\theta_{LacI}} \frac{1}{1+\left(\frac{u_{IPTG}}{\theta_{IPTG}}\right)^{\eta_{IPTG}}}\right)^{\eta_{LacI}}} - g_T^p [TetR] \end{cases} \tag{3.2}$$

Albeit still deterministic, this system plays a key role in understanding the bistable nature of the toggle switch. Indeed, since it is a two-dimensional system we can easily perform the nullcline analysis and plot the 2D vector field and phase portrait. We recall that the nullclines are the curves obtained by setting to zero the derivative of each variable, thus they represent the points on which the corresponding variable is not changing. Setting to zero both $\frac{d}{dt}[LacI]$ and $\frac{d}{dt}[TetR]$ of (3.2), one obtains the following two nullcline curves, which have a sigmoidal shape visible in Fig.3.3.

$$
\begin{aligned}
[LacI] &= \frac{k_L^p k_L^{m0}}{g_L^m g_L^p} + \frac{\frac{k_L^p k_L^m}{g_L^m g_L^p}}{1 + \left(\frac{[TetR]}{\theta_{TetR}} \frac{1}{1+\left(\frac{u_{aTc}}{\theta_{aTc}}\right)^{\eta_{aTc}}} \right)^{\eta_{TetR}}} \\
[TetR] &= \frac{k_T^p k_T^{m0}}{g_T^m g_T^p} + \frac{\frac{k_T^p k_T^m}{g_T^m g_T^p}}{1 + \left(\frac{[LacI]}{\theta_{LacI}} \frac{1}{1+\left(\frac{u_{IPTG}}{\theta_{IPTG}}\right)^{\eta_{IPTG}}} \right)^{\eta_{LacI}}}
\end{aligned}
\tag{3.3}
$$

The crossing points of these curves represent the equilibria of the system because here both protein concentrations do not vary [80]. The vector field assigns a vector to each point in the phase space, representing the derivative in that point. This is a typical way to analyze dynamical systems, indeed from this it is also possible to see the trajectory corresponding to any initial condition, obtaining the phase portrait. The vector field can be obtained, for example, with the Matlab command *quiver*, the resulting plot for the reduced toggle switch with zero inputs $u_{aTc} = 0$ and $u_{IPTG} = 0$ is shown in the image below.

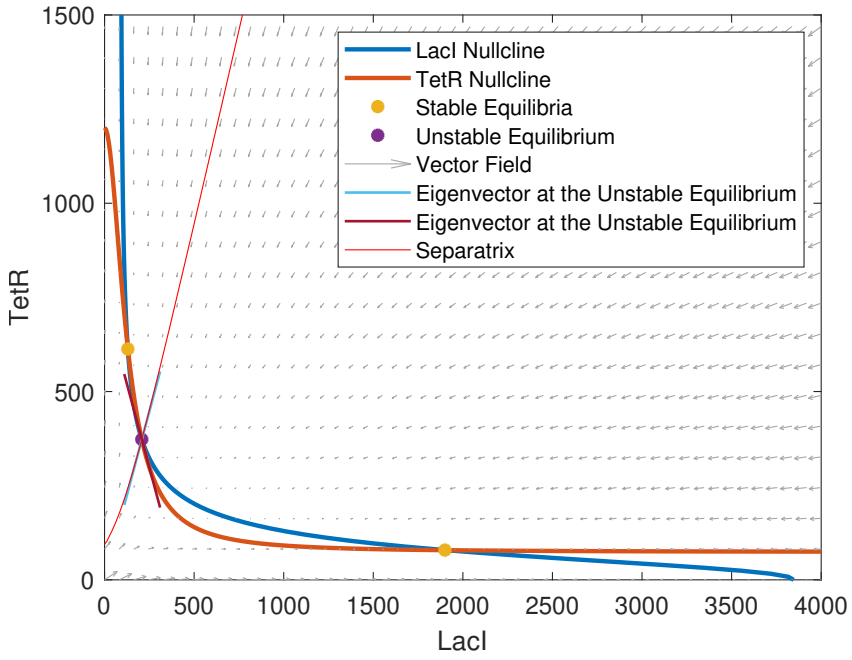


Figure 3.3: Phase portrait, nullclines, equilibria, separatrix of the reduced toggle switch model in (3.2), with $u_{aTc} = 0$ and $u_{IPTG} = 0$.

From the phase portrait it is straightforward to see that without input the system has three equilibria, specifically the (low LacI, high TetR) and the (high LacI, low TetR) are the stable ones, while the third one is unstable. Practically the toggle switch will always be in one of these states or will switch between them, moving away from the unstable equilibrium. The two stable equilibria correspond to the two ON/OFF states of the toggle switch, in one the production of LacI is active and

the one of TetR is suppressed; and in the other equilibrium the opposite holds. The phase space can thus be divided into two regions of attraction, each one leading to one of the stable equilibria. Indeed, when treating the system deterministically, the region of the initial conditions determines the equilibrium to which the system will converge. The separatrix (red curve in the plot) is by definition the curve separating these two regions, obtained by solving the system in (3.2) with opposite sign, $\dot{x} = -f(x)$, with initial condition in the unstable equilibrium. The separatrix can also be linearly approximated by one of the eigenspaces (light blue line) associated with the linearized system at the unstable equilibrium, as visible in Fig.3.3.

Up to this point, we have considered the inputs to be equal to zero. By changing them, one can observe some interesting behaviors. In particular, the system can become monostable if, for instance, one adds sufficient u_{IPTG}, which will bound to LacI, inhibiting its repressing action. As a result, there will be an abundance of TetR, which will further repress LacI production, leading the system to a single stable equilibrium at (low LacI, high TetR). Similarly, the other equilibrium can be obtained by adding a sufficient quantity of u_{aTc}, which will inhibit the repressing action of TetR and cause an increase in the quantity of LacI.

The effect of the inputs on the quality and quantity of equilibria can be studied via the bifurcation diagram. This is a common way to analyze how the equilibria of a system change depending on a parameter. In this case, we have two possible parameters u_{aTc} and u_{IPTG}, but they can be reduced to only one virtual input by imposing the following convex combination, as done in [44]:

$$u_{aTc} \in [0, U_{aTc}] \quad \text{and} \quad u_{IPTG} = \left(1 - \frac{u_{aTc}}{U_{aTc}}\right) U_{IPTG}, \quad \text{with } U_{aTc} = 100\,ng/ml, U_{IPTG} = 1\,mM.$$

This imposes a relationship between the inputs, hence we can consider as bifurcation parameter either one of them individually or a combination of both; for instance, we chose to use the difference of the standardized inputs, defined as follows.

$$u_{virtual} = \frac{u_{aTc}}{U_{aTc}} - \frac{u_{IPTG}}{U_{IPTG}} = 2\frac{u_{aTc}}{U_{aTc}} - 1 \in [-1, 1], \quad \text{for } u_{aTc} \in [0, U_{aTc}]. \tag{3.4}$$

This is a virtual input we can give to the system to study the effects on the position and number of equilibria. From this it is possible to retrieve the values of the original inputs as:

$$u_{aTc} = \frac{U_{aTc}}{2}(u_{virtual} + 1) \quad \text{and} \quad u_{IPTG} = \frac{U_{IPTG}}{2}(1 - u_{virtual}).$$

To draw the bifurcation diagram, we need to reduce to only one variable between LacI and TetR, or to a scalar function of both these proteins. Since they are in a direct relationship, we can focus only on one of them and obtain the value of the other one from the expression in (3.3). We have chosen to look at LacI, and for each value of the virtual input, we have found and plotted its equilibria,

distinguishing between stable and unstable equilibrium points. The bifurcation diagram obtained in the way described is shown in the figure below.

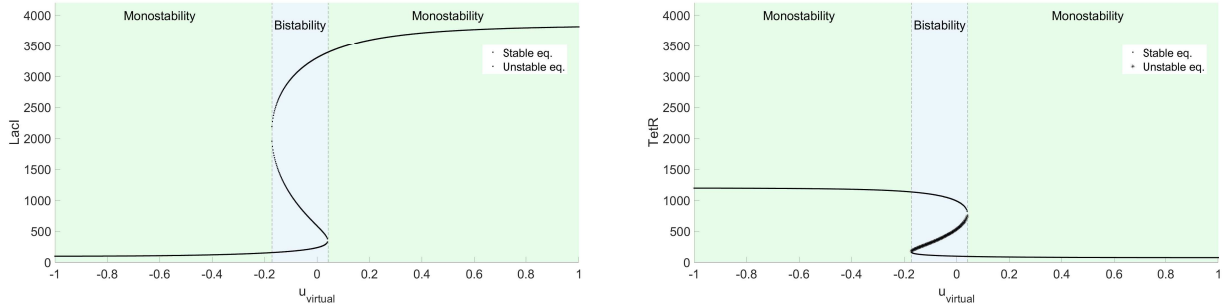
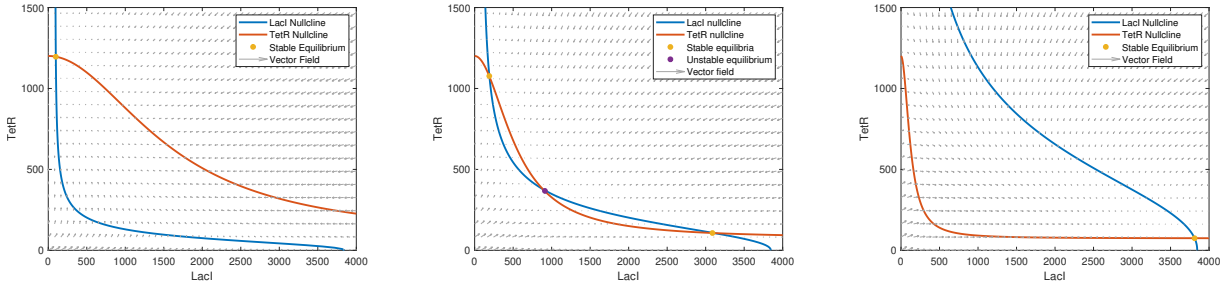


Figure 3.4: Bifurcation diagram for LacI and TetR of reduced toggle switch model (3.2), with parameter the virtual input defined in (3.4).

The figure shows that there exists an interval in which the system exhibits bistability, for $u_{virtual} \in [-0.172, 0.042]$, for which it holds:

$$u_{virtual} \in [-0.172, 0.042] \quad \Longrightarrow \quad u_{aTc} \in [41.4, 52.1]\, ng/ml \text{ and } u_{IPTG} \in [0.586, 0.479]\, mM.$$

For the other values, the system is instead monostable, in particular with an equilibrium in (low LacI, high TetR) when $u_{virtual} < -0.172$, and in (high LacI, low TetR) when $u_{virtual} > 0.042$. Below, the nullclines are shown for three significant cases $u_{virtual} = -1$, $u_{virtual} = -0.07$ and $u_{virtual} = 1$, that implies $u_{aTc} = 100\, ng/ml$, $u_{IPTG} = 0$, $u_{aTc} = 46.75\, ng/ml$ and $u_{IPTG} = 0.5325 mM$ and $u_{IPTG} = 1\, mM$, $u_{aTc} = 0$, for which the system is monostable.



(a) $u_{virtual} = -1$ ($u_{aTc} = 0$ and $u_{IPTG} = 1mM$).

(b) $u_{virtual} = -0.07$ ($u_{aTc} = 46.75\, ng/ml$ and $u_{IPTG} = 0.5325 mM$).

(c) $u_{virtual} = 1$ ($u_{aTc} = 100\, ng/ml$ and $u_{IPTG} = 0$).

Figure 3.5: Phase portrait, nullclines, (low LacI, high TetR) equilibrium of the reduced toggle switch model for different inputs.

3.1.2 Stochastic Model using QSSA

The deterministic model considered has provided valuable insights into the system's behavior, enabling to determine whether the system exhibits two or just one stable equilibrium, or to identify

the corresponding regions of attraction. However, this approach is not sufficient to fully describe a real genetic toggle switch. For instance, even if an initial condition lies in one of the basins of attraction, it is not always true that the real system will end up, or remain for all times, in the corresponding equilibrium, as the RRE instead predicts. Similarly, the bifurcation diagram of Fig. 3.4 identifies the areas of mono- and bi-stability, though in a real case scenario, the boundaries would not be so precise. Especially near the limit values of u_1 and u_2, one of the stable equilibria will be very close to the unstable one and basically "disappear". Furthermore, real biological systems exhibit variability at the cellular level: even under identical experimental conditions, there could be different genetic switching in different cells. Therefore, to perform a reliable and realistic analysis, it is essential to consider a stochastic model, which accounts for the inherent noise characteristic of biological systems.

To obtain stochastic models, we can apply what has been explained previously, namely: consider the system in (3.1) obtained by QSSA, write it in the number of molecules, extrapolate the propensity functions, and use them in the stochastic modeling and simulations as presented in Chapter 2. This approach has been applied in [79] to the toggle switch model to perform stochastic simulations with Gillespie's Algorithm.

Following the procedure described, the first step requires to write the equations in the number of molecules. To do so, let us recall the definition of molecular and molar concentration: $[X] = \frac{X}{\Omega}$ and $[X] = \frac{X}{\Omega N_A}$. The choice of one of them strictly depends on the measurement unit adopted in the specific reference document; in our case, we used the values of [78], reported in Table 3.1, where the measurement unit is the number of molecules, not of moles. Therefore, we can substitute the definition of molecular concentration into the equations in (3.1). To illustrate this step, let us consider only the first equation:

$$\frac{d}{dt} mRNA_{LacI}\,\Omega = k_L^{m0} + \frac{k_L^m}{1 + \left(\frac{TetR\,\Omega}{\theta_{TetR}} \frac{1}{1+\left(\frac{aTc\,\Omega}{\theta_{aTc}}\right)^{\eta_{aTc}}} \right)^{\eta_{TetR}}} - g_L^m\, mRNA_{LacI}\,\Omega.$$

Dividing both sides by Ω, one gets:

$$\frac{d}{dt} mRNA_{LacI} = \frac{k_L^{m0}}{\Omega} + \frac{\frac{k_L^m}{\Omega}}{1 + \left(\frac{TetR\,\Omega}{\theta_{TetR}} \frac{1}{1+\left(\frac{aTc\,\Omega}{\theta_{aTc}}\right)^{\eta_{aTc}}} \right)^{\eta_{TetR}}} - g_L^m\, mRNA_{LacI},$$

and defining the new coefficients scaled by the volume as $k_L^{m0\prime} = \frac{k_L^{m0}}{\Omega}$, $k_L^{m\prime} = \frac{k_L^m}{\Omega}$, $\theta_{TetR}{}' = \frac{\theta_{TetR}}{\Omega}$, $\theta_{aTc}{}' = \frac{\theta_{aTc}}{\Omega}$, then one obtains:

$$\frac{d}{dt}mRNA_{LacI} = k_L^{m0\prime} + \frac{k_L^{m\prime}}{1+\left(\frac{TetR}{\theta_{TetR}{}'}\frac{1}{1+\left(\frac{aTc}{\theta_{aTc}{}'}\right)^{\eta_{aTc}}}\right)^{\eta_{TetR}}} - g_L^m\, mRNA_{LacI}.$$

The same reasoning can be applied to the remaining equations in 3.1, obtaining something analogous for the $mRNA_{TetR}$, after defining the coefficients as $k_T^{m0\prime} = \frac{k_T^{m0}}{\Omega}$, $k_T^{m\prime} = \frac{k_T^m}{\Omega}$, $\theta_{LacI}{}' = \frac{\theta_{LacI}}{\Omega}$, $\theta_{IPTG}{}' = \frac{\theta_{IPTG}}{\Omega}$. The other equations instead stay the same, just written in the number of molecules. In particular, we have supposed that the inputs u_{aTc} and u_{PITG} are given as a concentration, if not, a scaling factor would simply be applied to the coefficients k_{aTc} and k_{IPTG} to account for this difference.

Finally, let us define the state as $X = \begin{bmatrix} mRNA_{LacI} & mRNA_{TetR} & LacI & TetR & aTc & IPTG \end{bmatrix}^T$, and order the reactions as birth and death processes for each species in X. Under this setup, the propensity functions will have the following expressions:

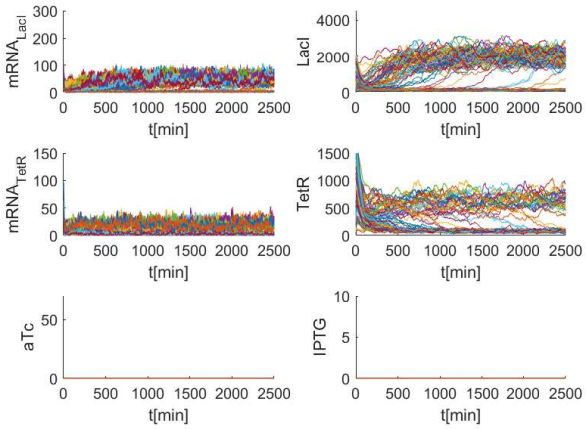
Birth	**Death**
$a_1 = k_L^{m0\prime} + \frac{k_L^{m\prime}}{1+\left(\frac{TetR}{\theta_{TetR}{}'}\frac{1}{1+\left(\frac{aTc}{\theta_{aTc}{}'}\right)^{\eta_{aTc}}}\right)^{\eta_{TetR}}},$	$a_2 = g_L^m\, mRNA_{LacI},$
$a_3 = k_T^{m0\prime} + \frac{k_T^{m\prime}}{1+\left(\frac{LacI}{\theta_{LacI}{}'}\frac{1}{1+\left(\frac{IPTG}{\theta_{IPTG}{}'}\right)^{\eta_{IPTG}}}\right)^{\eta_{LacI}}},$	$a_4 = g_T^m\, mRNA_{TetR},$
$a_5 = k_L^p\, mRNA_{LacI},$	$a_6 = g_L^p\, LacI,$
$a_7 = k_T^p\, mRNA_{TetR},$	$a_8 = g_T^p\, TetR,$
$a_9 = k_{aTc}\, u_{aTc},$	$a_{10} = k_{aTc}\, aTc,$
$a_{11} = k_{IPTG}\, u_{IPTG},$	$a_{12} = k_{IPTG}\, IPTG.$

The stoichiometric vectors can be written just by looking at the equations in (3.1). Since there are only birth and degradation processes, the entries will be equal to ± 1, or 0 when the species is not affected by the corresponding reaction. The vectors can be stacked in the following stoichiometric matrix:

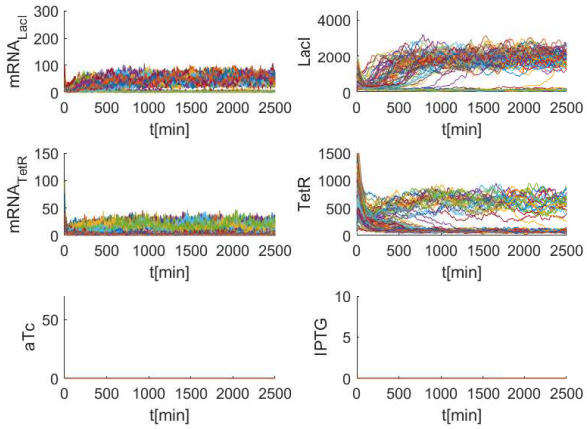
$$V = \begin{bmatrix} \nu_1 & \nu_2 & \dots & \nu_{12} \end{bmatrix} = \begin{bmatrix} +1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & +1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & +1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & +1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & -1 \end{bmatrix}.$$

Finally, knowing the propensity functions and the stoichiometric vectors, we can simply use them to derive any desired model and corresponding simulation method. Comparable approaches have been applied in [74] using the τ-leaping method for a similar case, in [73] using Gillespie Algorithm for the enzymatic reaction, and in [79] specifically for simulating the toggle switch. Hence, we have strong reasons to believe that this method could be both valid and effective. The key advantage is that it is possible to consider existing robust ODE models that may encapsulate detailed chemical kinetics by various Hill functions and quasi-steady-state assumptions, and use these to directly derive reliable stochastic descriptions.

Below we show the plots obtained following this approach for Gillespie and τ-leaping algorithms, using zero inputs. Since the experiments in [76] with the toggle switch were performed in the Escherichia coli, we have considered its volume $\Omega = 1\,\mu m^3$ [81].



(a) 100 simulations of the toggle switch in 6 variables (3.1) by Gillespie's SSA, with zero input.



(b) 100 simulations of the toggle switch in 6 variables (3.1) by adaptive τ-leaping method, with zero input and $\varepsilon = 0.01$.

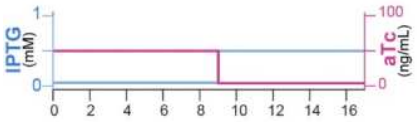
From the simulations, the two equilibria are clearly visible, and we can claim that the toggle switch

is quite robust since only a few number of trajectories is switching equilibrium. Moreover, we can see that qualitatively both algorithms return similar results, hence τ-leaping represents a valid option, even if it is important to mention that sometimes during its run, it was necessary to call Gillespie's SSA because the number of molecules was too low (especially for the $mRNA_{TetR}$) and the algorithm was generating negative numbers. This is why we had to choose a very small $\varepsilon = 0.01$, to ensure that the state change was minimal. The time spent by each algorithm is the following:

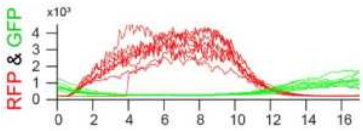
$$\text{SSA: } 424.5333\ s, \qquad \tau\text{-leaping: } 256.3125\ s.$$

Therefore, τ-leaping approximation is reliable and faster than the SSA, and can be used to speed up simulations. We have also tried to implement the E-M method, though, the approximation of the Poisson distribution with a Gaussian one was never true, and E-M was always calling Gillespie. This is probably due to the fact that the number of *mRNA* molecules was quite low, violating the main hypothesis of this method: to have a high number of molecules.

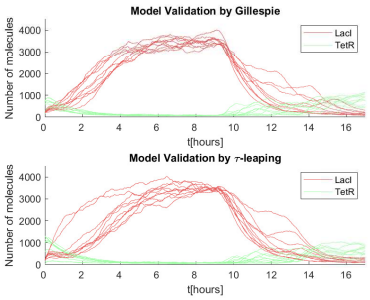
The stochastic models so obtained have been validated by comparing their realizations with experimental data taken from the Supplementary Information of [78]. The figures below show this comparison for various combinations of the inputs. In almost all the cases, both Gillespie's SSA and τ-leaping method return simulations similar to the real data, except for the last two cases (Fig. 3.11 and 3.12) in which, when simulating the system, the quantity of LacI decreases after the reduction of u_{aTc}, whereas, in the real system this reduction is not observed. This discrepancy may arise because, in an experimental setting, decreasing the amount of u_{aTc} is not instantaneous, whereas in the simulations we can assume that the change is immediate. Consequently, in a real experiment, the quantity of LacI may also decrease over time due to the continuous presence of u_{IPTG} and the reduction of u_{aTc}, but probably it was not possible to see it in the time provided. Indeed, IPTG would continue to bind to LacI preventing the repression of TetR, and, as aTc diminishes, TetR will further repress LacI. Together, these two factors will eventually result in a reduction in LacI levels. This hypothesis is supported by the fact that in [78] an asymmetrical exchange of aTc and IPTG is considered, in particular the out-exchange rate of aTc is half of the value reported in Table 3.1. Hence, taking this into account should provide more accurate simulations, resulting in a much slower decrease in the concentration of LacI in the last two tests, comparable with the experimental data. However, we have adopted the values of Table 3.1 because they are the ones used in [44], that is the work we will refer to in the next chapter.



(a)

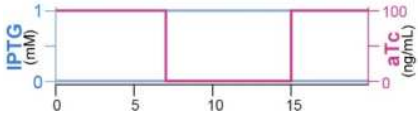


(b)

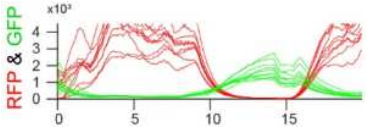


(c)

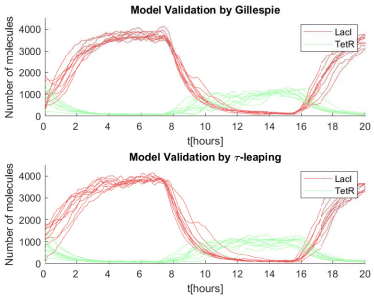
Figure 3.7: (a) Given inputs. (b) Experimental data [78]. (c) 10 model simulations by Gillespie's SSA (top) and τ-leaping (bottom).



(a)

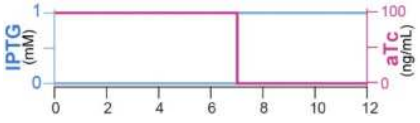


(b)

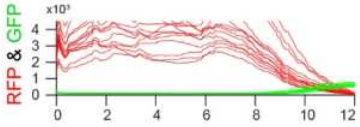


(c)

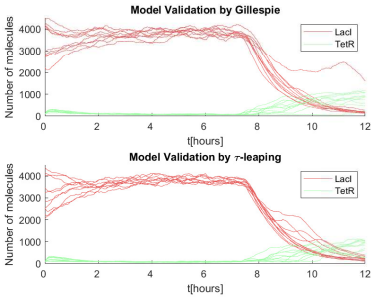
Figure 3.8: (a) Given inputs. (b) Experimental data [78]. (c) 10 model simulations by Gillespie's SSA (top) and τ-leaping (bottom).



(a)

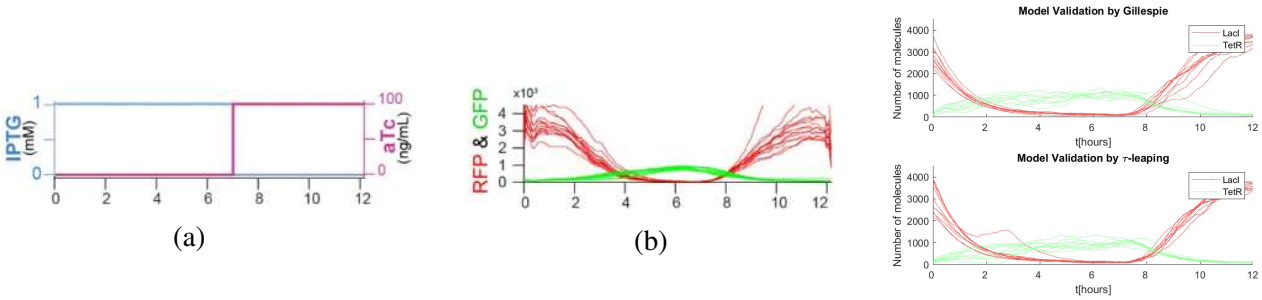


(b)



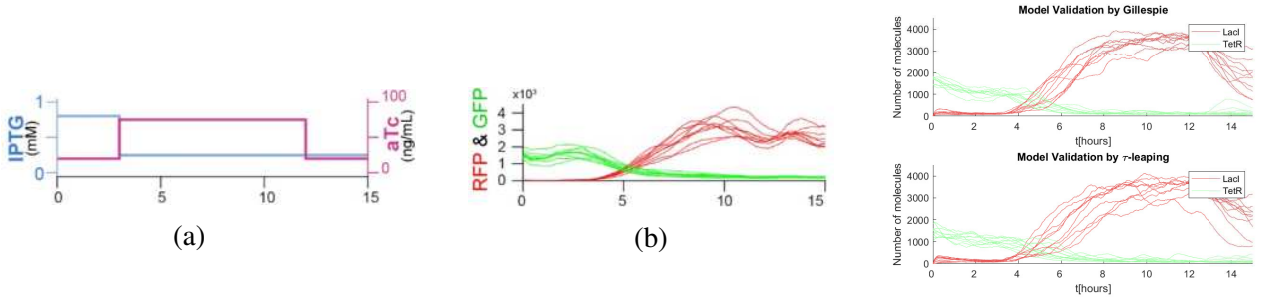
(c)

Figure 3.9: (a) Given inputs. (b) Experimental data [78]. (c) 10 model simulations by Gillespie's SSA (top) and τ-leaping (bottom).



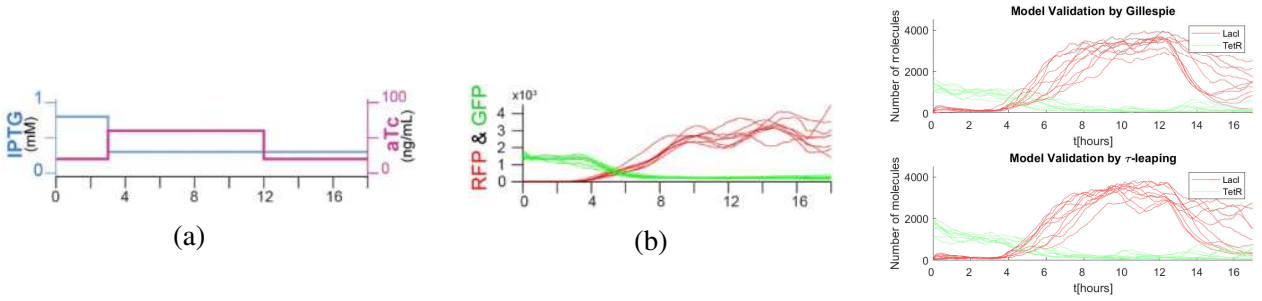
(a) (b) (c)

Figure 3.10: (a) Given inputs. (b) Experimental data [78]. (c) 10 model simulations by Gillespie's SSA (top) and τ-leaping (bottom).



(a) (b) (c)

Figure 3.11: (a) Given inputs. (b) Experimental data [78]. (c) 10 model simulations by Gillespie's SSA (top) and τ-leaping (bottom).

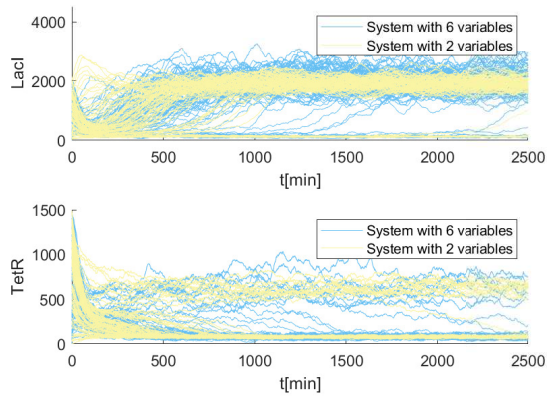


(a) (b) (c)

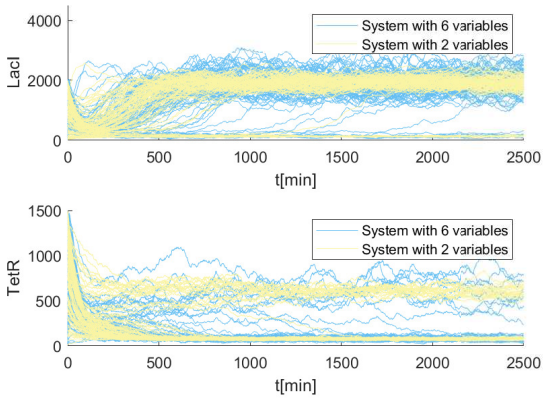
Figure 3.12: (a) Given inputs. (b) Experimental data [78]. (c) 10 model simulations by Gillespie's SSA (top) and τ-leaping (bottom).

Comparison with reduced stochastic model

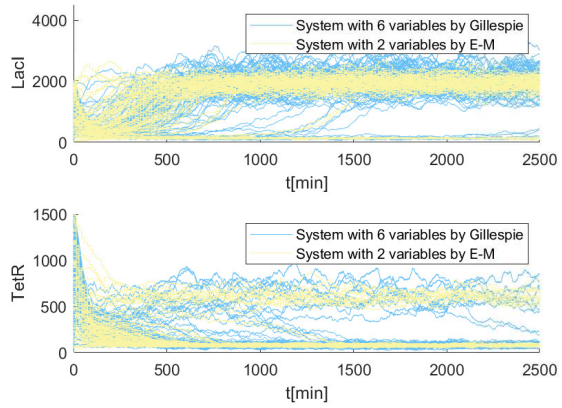
The same approach to derive stochastic simulations can in principle be applied to the system in (3.2), reduced to only two variables. The results are shown in the following figures, where we have compared the trajectories obtained by the system, before and after applying the QSSA to (3.1).



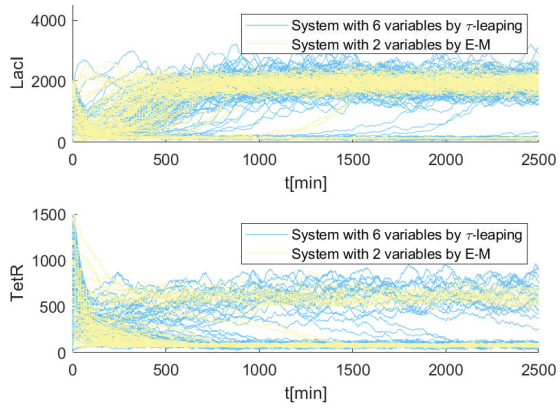
(a) Simulations performed by Gillespie's SSA.



(b) Simulations performed by adaptive τ-leaping, with $\varepsilon = 0.01$.



(c) Simulations performed by E-M method with $\varepsilon = 0.2$ and $\tau = 15$; and by Gillespie's SSA.

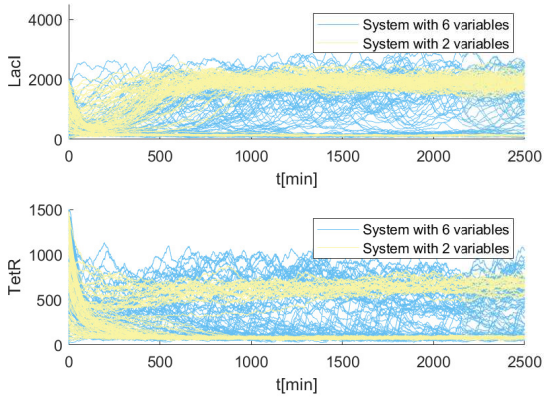


(d) Simulations performed by E-M method with $\varepsilon = 0.2$ and $\tau = 15$; and by τ-leaping approximation with $\varepsilon = 0.01$.

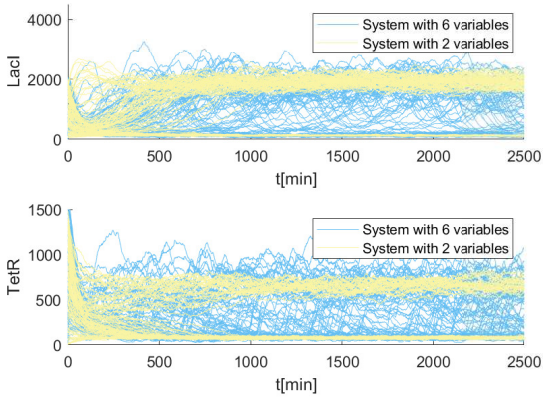
Figure 3.13: Comparison of 100 simulations of the reduced toggle switch model in 2 variables (3.2), with the one in 6 variables (3.1), with $u_{aTc} = 0$ and $u_{IPTG} = 0$.

As the reader can see, in this case it was also possible to perform E-M simulations, using $\varepsilon = 0.2$. This is probably due to the fact that we have thrown away the evolution of the molecules that were present in low numbers, allowing for the hypothesis of the CLE to be true. In particular, Fig. 3.13 compares the results of the reduced system obtained by E-M method with those of the system in (3.1) of 6 variables, obtained by the SSA and τ-leaping respectively, since E-M method was not applicable in that system.

The figures above reveal that the behavior of the stochastic systems in two or six variables is quite different. Indeed, the former is less noisy than the latter, which does not seem to be an issue with zero input, but it actually is for example with $u_{IPTG} = 0.05\,mM$. Indeed, Fig.3.14 shows that for this input the trajectories of the system in 6 variables oscillate and switch between the two stable states, whereas the ones in 2 variables do not move from the equilibria. We have reported only the simulations made by Gillespie's SSA and τ-leaping algorithm, but the same happens also with the E-M method.



(a) Simulations performed by Gillespie's SSA.



(b) Simulations performed by adaptive τ-leaping, with $\varepsilon = 0.01$.

Figure 3.14: Comparison of 100 simulations of the reduced toggle switch model in 2 variables (3.2), with the one in 6 variables (3.1), with $u_{aTc} = 0$ and $u_{IPTG} = 0.05\,mM$.

The different behavior of the two systems can be attributed to the fact that the QSSA has been applied to the deterministic system. As a result, when constructing the stochastic version, the noise associated with the equations at steady state is omitted, leading to a reduced variability in the simplified system. However, we can expect that there exists a way to include this noise into the reduced stochastic model, obtaining more reliable simulations.

3.1.3 Discussion

In this chapter, we have introduced, tested, and discussed an innovative and powerful method for performing stochastic simulations on a reduced system without necessarily accounting for all the reactions, or requiring prior knowledge of all the corresponding parameters. This approach takes inspiration from the well-established technique used to reduce Reaction Rate Equations via time scale separation. Based on previous work, we have shown how to extend it in order to develop a corresponding stochastic framework.

The example considered is the genetic toggle switch, a bistable circuit involving 18 species and 30 reactions. Among these, certain species exhibit rapid dynamics, which can be approximated as being at steady state, which allows to reduce the system to one with 6 species and 12 reactions (3.1), or even further to a two-dimensional system (3.2). The latter has been used to perform a stability and bifurcation analysis, revealing the existence of two monostable regions (one for each stable equilibrium) and one bistable region, depending on the value of a virtual input. However, when the noise is considered, the boundaries between these regions become less defined, as will be shown in detail in the next chapter. Hence, the objective of this chapter was to propose an efficient way to account also for intrinsic noise that, particularly in the case of the toggle switch, could produce unwanted switches or different outcomes across the cells, making its consideration essential.

Therefore, we have presented a method to produce stochastic simulations of the system in 6 variables, and tested its validity by comparing the simulation results with the experimental data provided in [78]. Simulations were conducted using both Gillespie's algorithm and the τ-leaping method, which showed behaviors similar to the real system, with τ-leaping being nearly twice as fast. Instead, in this case, the E-M method was not applicable, probably due to low molecular numbers of certain species.

Finally, we have tried to perform stochastic simulations using the reduced system with only two variables, but unfortunately, the variability observed in this system was significantly lower than the one in the six-dimensional case. This could be explained by the fact that, moving from a system in 6 dimensions to one in 2 dimensions, part of the noise (the one of the fast dynamics) has been neglected, thereby causing less variability. Nevertheless, we are confident that there can be a way to take into account this noise artificially adding it to the 2-dimensional system to obtain reliable simulations also in this case. If such a method does not already exist, it could represent a significant future research direction, which could lead to the development of even more efficient stochastic simulation techniques and models. However, to ensure accurate representations, this thesis relies on the six-variable model in (3.1), incorporating the noise as described.

Chapter 4

Application to Ratiometric Control Problem

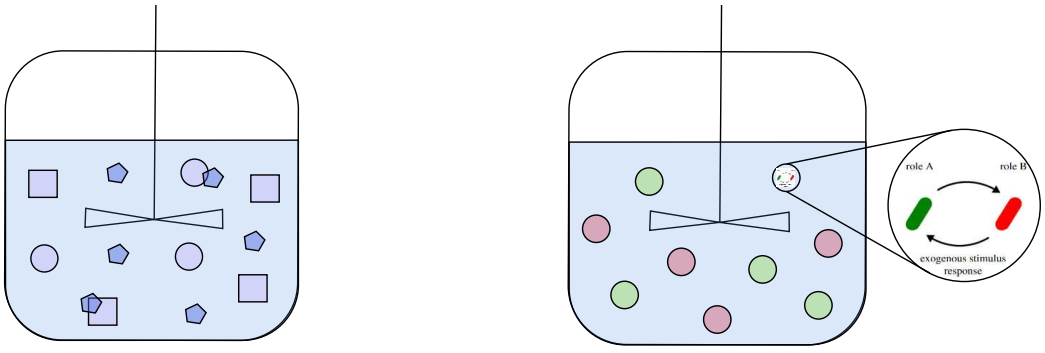
Synthetic biology is a highly promising and innovative field; however, its novelty inevitably brings significant challenges. These are mainly due to the desire of inserting huge and complex circuits into a single living cell, which may lead to unintended side effects, such as metabolic burden, competition of limited resources, retroactivity, or incompatible chemical reactions [44]. To address these problems, the standard solution proposed in the literature consists in distributing the workload, or the different functionalities, among multiple cell populations in a microbial consortium [82]. Therefore, instead of having a population where each cell is producing a set of proteins or is having various roles, there will be more populations with each population having a specific function, for example to produce only one type of protein. This alleviates the burden of single cells and enables compartmentalization and collaboration between diverse microbial consortia. Moreover, it can be significantly advantageous to distribute and optimize complex metabolic pathways or to facilitate the production of high-value compounds.

Although this approach seems very appealing, it arises a new challenge: controlling the ecosystem as a whole rather than only the individual cells and reactions happening inside them. This can be particularly difficult because of different growth and death rates, diverse metabolic burdens, various environmental conditions, or distinct reactions to a stimulus, all of which can cause one population to displace the others. Indeed, according to the competitive exclusion principle, competitors species cannot co-exist [35], even if growth and death rates are only slightly different. Therefore, researchers have started to seek a way to regulate the relative sizes of the different populations in order to counteract the competitive exclusion principle, preventing faster-growing species from eliminating the slower ones. This problem is commonly known as Population Ratio Control [41, 42, 44], and has the formal objective of balancing the populations so that their relative numbers satisfy a desired ratio.

4.1 Single vs Double Population Approach

The goal of Ratiometric Control is to regulate the relative sizes of two populations, in order to guarantee their stable co-existence. As discussed in the Introduction, microbial consortia can have different important applications, among which multicellular feedback controller, implementation of multiple logical functions, and innovative solutions for industrial applications, thanks to task division within populations. Though, to be able to ensure the survival of all populations, controlling their proportions is essential. Usually, this is addressed by embedding additional genetic circuits that enable cells to sense and respond to each other's relative size through quorum sensing molecules (Fig.4.1a). In particular, by sensing the density of the other group, cells can either increase their growth rate by producing some growth regulatory protein, or decrease their number by means of toxin–antitoxin mechanisms.

However, dealing with more populations requires to guarantee their survival and to insert additional genetic circuits enabling communication within the consortium. To overcome this challenge, an innovative solution has been proposed, approaching the problem from a different perspective which focuses on only one population. In this setup, all cells grow at the same rate, and there is no need of communication and sensing molecules anymore. The key idea is to distinguish two groups within the same population using a reversible bistable memory element, so that each group has a specific role (Fig. 4.1b). At the same time, cells can quickly and easily switch between groups in response to exogenous stimuli from the environment, such as the injection of inducer molecules or light. This is particularly useful if for example one group dies, or if a new population ratio is asked. This approach represents a cutting-edge solution that, to the best of our knowledge, has been proposed and studied only by D. Salzano, D. Fiore and M. di Bernardo in [44]. Their detailed proposal is discussed in the following pages.



(a) Two cells populations (circles and squares) in a chemostat, communicating through quorum sensing molecules (pentagons) that control their growth or death rates.

(b) One single population (circles), where each cell can carry out different roles (red or green), in response to exogenous stimuli.

Figure 4.1: Ratiometric Control Solutions.

Before analyzing the single population approach, we would like to mention some works proposing solutions to the ratiometric control problem using two populations. In [43] the authors implemented autonomous regulation of the consortia composition via cell-based signal translator and growth-controller modules. Other studies [37–40] dealt with competitive populations, realizing stable co-cultures using various strategies, between which controlling the toxin production or the dilution rate. This last strategy was also used in [41] to regulate the ratio between the concentrations of two microbial populations in a chemostat while guaranteeing their survival and fast convergence dynamics. Finally, in [42] both the total cell population and the relative ratio between two cell strains have been regulated via dual feedback control modulating either cell growth, or cell death processes, with communication implemented by quorum sensing molecules.

4.2 The Single Population Approach [44]

This chapter aims to present the pioneering work conducted by Salzano et al. in [44], where ratiometric control problem is treated using only one population. The aim is to create a population divided into two groups whose relative numbers satisfy a given ratio. By focusing on a single population, this approach bypasses challenges associated with different growth rates or communication molecules. Moreover, having cells of only one population should allow fast switching and recovery if one of the groups dies. As anticipated, the innovative idea is that cells can transition between these two groups via a reversible bistable memory element, that in the paper has been implemented with a genetic toggle switch. Although introducing a circuit into a cell could affect the growth of the host organism, this impact is likely less significant compared to the more complex circuits required when managing two separate populations; hence, this represents a very promising solution.

The population considered is composed of bacterial cells endowed with a reversible genetic toggle switch. In particular, they considered the toggle switch presented in the previous chapter, which was implemented *in vivo* by Gardner et al. in [76], and can be modeled by the equation (3.2). We recall that the different cells can exhibit different states due to inherent noise and can be controlled individually by an external stimulus. In this context, the same input is common to all cells, and it is precisely the heterogeneity of the responses that makes this method effective. This variability, indeed, allows some cells to adopt one state, others to settle in another state, and still others to switch dynamically between equilibria. To distinguish between the stable states of the toggle switch, to each of them is associated a reporter protein (RFP or GFP) which determines the phenotype (i.e. a set of observable characteristics) of each cell. Specifically, if a cell has the green phenotype, it will mainly produce TetR, and viceversa for the red phenotype.

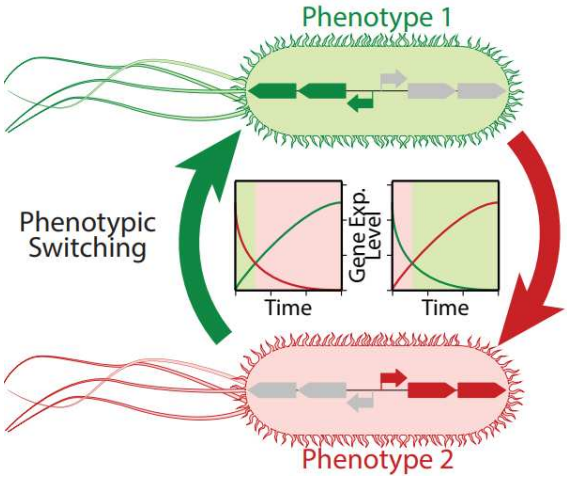


Figure 4.2: Phenotypic switching inside a cell [45].

In [44] they applied ratiometric control to the bioproduction of protein dimers. Therefore, to the production of the repressing proteins LacI and TetR, and of the reporter genes GFP and RFP, it is associated also the production of the corresponding monomer, in order to obtain the desired dimer. Since they assumed that the two monomers have equal transcription and translation rates, the dimer is produced at maximum rate if the consortium is divided into two symmetric groups with a $1:1$ ratio. Depending on the application, there could be different transcription and translation rates which would require different ratios between the two groups. In any case, r_d should be such that it guarantees high efficiency in the production of the desired metabolic end product.

Therefore, we can state the objective as follows:

> Given cells belonging to the same strain, endowed with a reversible bistable memory mechanism, design a feedback law $u(t,x) = \begin{bmatrix} u_{aTc}(t,x) & u_{IPTG}(t,x) \end{bmatrix}^T$ such that at steady state the consortium is divided into two cell groups whose ratios converge to the desired values $r_{d,LacI}(t) = \frac{n_{LacI}(t)}{N(t)}$ and $r_{d,TetR}(t) = \frac{n_{TetR(t)}}{N(t)}$, where $n_{LacI}(t)$ and $n_{TetR}(t)$ are the number of LacI and TetR molecules, $N(t)$ the total number of molecules in the consortium, and x the state of the system denoted by the phenotype of each cell.

This can be achieved by designing a controller for the scheme outlined in Fig.4.3, where a fluorescence microscope measures the expression of the reporter proteins (proportional to LacI and TetR) in each cell. These measurements will be fed into the feedback control algorithm that compares the current ratio $r(t)$ between the two groups with the desired one and computes online the appropriate control inputs. The calculated inputs, together with a mixture of growth medium, will then be delivered to the population using a pair of syringes.

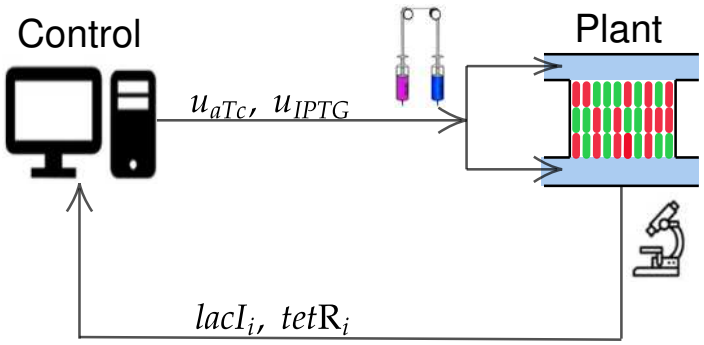
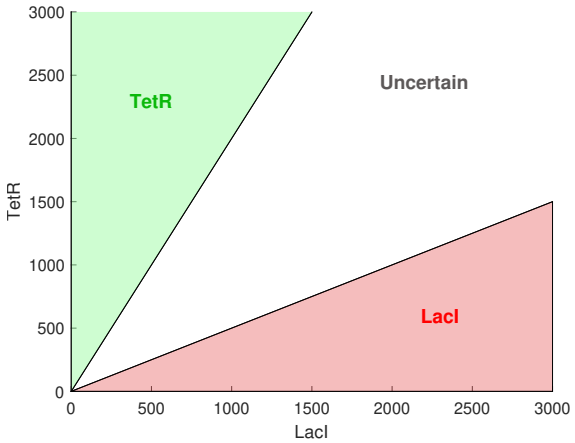


Figure 4.3: Control Scheme for Ratiometric Control Problem as faced in [44].

Before introducing the proposed controllers, it is important to specify the model used in this study. The authors considered a deterministic model of the reduced toggle switch (3.2), which was simulated using the agent-based cell simulator BSim [50]. This simulator generates realistic realizations of a cell population, accounting for environmental geometries, cellular growth, and physical interactions; though, it does not include stochastic intracellular dynamics.

Moreover, it is important to precise how the phenotype of each cell is determined. Fig. 3.3 of the previous chapter, shows the vector field of the deterministic model, together with the separatrix distinguishing the two regions of attraction of the stable equilibria. Though, due to higher dimensional dynamics of the real system and unavoidable uncertainties affecting it, it is not possible to define such precise boundary between the two regions. It is necessary, indeed, to consider an additional region where neither of the two proteins is produced in higher abundance than the other. This region is called the uncertain set because it is not possible to identify if the toggle switch is in TetR or LacI state since their values are close to the separatrix. The areas that the authors have defined in [44] are the ones we refer to, shown in the figure below.



$$TetR_t = \{i \in \mathcal{N}_t \,|\, 2\,LacI_i(t) < TetR_i(t)\}.$$

$$LacI_t = \{i \in \mathcal{N}_t \,|\, LacI_i(t) > 2\,TetR_i(t)\}.$$

$$Uncertain_t = \{i \in \mathcal{N}_t \,|\, i \notin TetR_t, i \notin LacI_t\}.$$

Figure 4.4: Regions of attraction for stable equilibria of reduced toggle switch, as defined in [44].

However, the existence of the uncertain set implies that the total number of cells is not simply the sum of those producing LacI and TetR, but it is given by $N(t) = n_{LacI}(t) + n_{TetR}(t) + n_{Uncertain}(t)$. Thus, when considering the current ratios, we have to distinguish between $r_{LacI}(t) = \frac{n_{LacI}(t)}{N(t)}$ and $r_{TetR}(t) = \frac{n_{TetR}(t)}{N(t)} \neq 1 - r_{LacI}(t)$ and keep track of both of them. Observe that, instead, the desired ratio can be specified for only one protein, as the ideal scenario assumes there are no cells in the uncertain set; hence $r_{d,LacI}(t) + r_{d,TetR}(t) = 1$ holds. For instance, let r_d represent the desired ratio for LacI, then the errors will be defined as $e_{LacI}(t) = r_d - r_{LacI}(t)$ and $e_{TetR}(t) = (1 - r_d) - r_{TetR}(t)$. Therefore, the objective will be achieved when both errors converge to zero, which slightly complicates the control laws. In the paper two possible controllers are implemented: the relay and the PI controller. To each of them corresponds a specific kind of input $u = [u_{aTc}\ u_{IPTG}]^T$ that has been adopted.

4.2.1 Control Design

Relay controller

The inputs satisfy the T-junction condition, meaning that u_{aTc} and u_{IPTG} are mutually exclusive and with fixed amplitude:

$$u = [U_{aTc}\ 0]^T \quad \text{or} \quad u = [0\ U_{IPTG}]^T. \tag{4.1}$$

The relay controller is probably the simplest feedback controller, indeed, based on the output error interval, it generates a piece-wise constant input signal that belongs to a discrete set. In particular, in this case, based on the major error in absolute value, it returns the input able to reduce this error. The control law can be written as:

$$u(t) = \begin{cases} u_1 & \text{if} \quad |e_{LacI}| \geq |e_{TetR}| \\ u_2 & \text{if} \quad |e_{LacI}| < |e_{TetR}| \end{cases}, \quad \text{where:}$$

$$u_1 = \begin{cases} \begin{bmatrix} 0 \\ U_{IPTG} \end{bmatrix} & \text{if} \quad e_{LacI} \leq 0 \\ \begin{bmatrix} U_{aTc} \\ 0 \end{bmatrix} & \text{if} \quad e_{LacI} > 0 \end{cases}, \qquad u_2 = \begin{cases} \begin{bmatrix} U_{aTc} \\ 0 \end{bmatrix} & \text{if} \quad e_{TetR} \leq 0 \\ \begin{bmatrix} 0 \\ U_{IPTG} \end{bmatrix} & \text{if} \quad e_{TetR} > 0 \end{cases}. \tag{4.2}$$

A shutdown condition ($u = [0\,0]^T$) can be included when both errors are equal to zero, to allow finite convergence to zero error. Indeed, without it, the control would continuously alternate between the two possibilities, causing the inputs to oscillate.

PI controller

The inputs satisfy the Dial-A-Wave combination, which implies u_{aTc} and u_{IPTG} being in the following convex combination:

$$u_{aTc} \in [0, U_{aTc}] \quad \text{and} \quad u_{IPTG} = \left(1 - \frac{u_{aTc}}{U_{aTc}}\right) U_{IPTG}. \tag{4.3}$$

Note that actually the T-junction is a special case of the Dial-A-Wave, when one of the inputs is equal to zero.

The PI controller is given by the sum of a proportional and an integral action, which is known to ensure zero regulation error at steady state. In this case the input is smoother and the error converges to zero, even if slower than before. The control law is the following:

$$\begin{cases} u_{aTc} = k_{P,1}\, e_{LacI}(t) + k_{I,1} \int_0^t e_{LacI}(\tau)\, d\tau - \left(k_{P,2}\, e_{TetR}(t) + k_{I,2} \int_0^t e_{TetR}(\tau)\, d\tau\right) \\ u_{IPTG} = \left(1 - \frac{u_{aTc}}{U_{aTc}}\right) U_{IPTG} \end{cases} . \tag{4.4}$$

To improve performances, they also added an anti-windup scheme that sets to zero the internal state of the integrator when the error is 0 or changes sign, and a dynamic saturation defined as:

$$\begin{cases} u_{aTc} \in [0, 50] & |e_{LacI}| < |e_{TetR}| \\ u_{aTc} \in [0, 100] & \text{otherwise} \end{cases} .$$

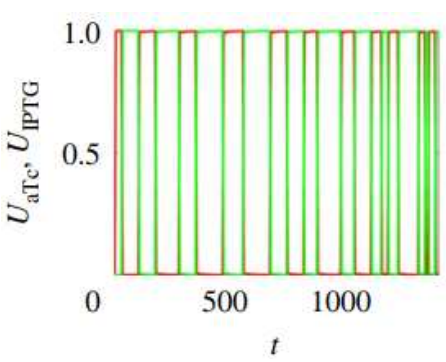
4.2.2 Simulation Results

As anticipated, the simulations have been performed using BSim, in particular they accounted for realistic physical and technological constraints summarized in the following table:

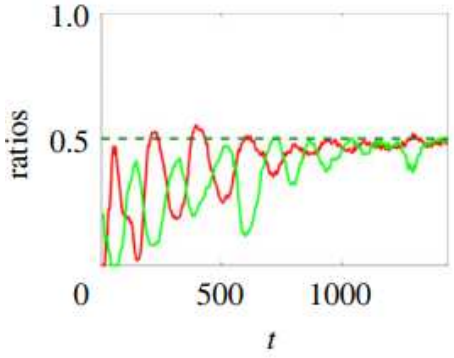
Target Variable	**Constraint**	**Motivation**
Input values	$U_{aTc} \in [0, 100]$ mg/ml, $U_{IPTG} \in [0, 1]$ mM	Avoids excessive stress on cells.
Input switching frequency	$\leq \frac{1}{15 \text{ minutes}}$	Reduces osmotic stress on cells.
Time delay	40 seconds	Accounts for time spent by inducers to flow into cell chambers.
Sampling time	≥ 5 minutes	Prevents excessive photo-toxicity.
Experiment duration	≤ 24 hours	Avoids substantial cell mutations.

Table 4.1: Constraints considered in BSim by [44].

The figures below show the results of the simulations using the controllers described, where the proportional and integral gains, and the values of the maximum concentrations of the inducers, have been empirically tuned by trial-and-error. Their values are reported in the captions of the corresponding figure.

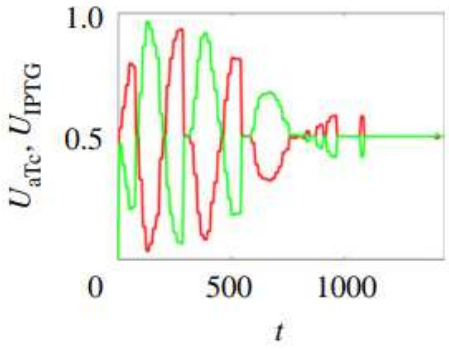


(a) Control inputs.

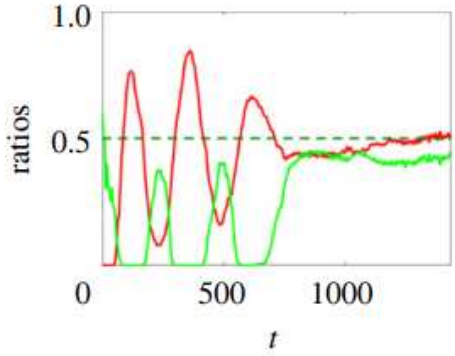


(b) Evolution of the ratios r_{LacI} and r_{TetR}.

Figure 4.5: Relay controller (4.2) for ratiometric control problem with one population simulated using BSim [44]. Parameters $r_d = 0.5$, $U_{aTc} = 60\,ng/ml$ and $U_{IPTG} = 0.5\,mM$.



(a) Control inputs.



(b) Evolution of the ratios r_{LacI} and r_{TetR}.

Figure 4.6: PI controller (4.4) for ratiometric control problem with one population simulated using BSim [44]. Parameters: $r_d = 0.5$, $U_{aTc} = 100\,ng/ml$, $U_{IPTG} = 1\,mM$, $k_{P,1} = 100$, $k_{P,2} = 1.5$, $k_{I,1} = 1.5$ and $k_{I,2} = 0.05$.

The figures illustrate the validity and efficacy of the controllers implemented in the paper, even in the presence of realistic physical and technological constraints. In the case of the relay controller the inputs are switching between the two cases defined in (4.2) causing the output to never really reach the desired value, although getting very close to it. On the other hand, when using the PI controller, the inputs evolutions are smoother and after some time they stabilize to the value of 0.5 for u_{IPTG} and 50 for u_{aTc}, even though the response of the system is a bit slower, but still with higher accuracy at steady state. Moreover, the authors accounted for cell-to-cell variability by changing the model parameters. This can cause some cells to be uncontrollable, meaning that they

are either monostable or unswitchable, and hence there does not exist an input able to bring them from one state to the other. In the paper, they proved that the error goes to zero as long as no uncontrollable cells are present; otherwise, a small, unavoidable residual error will remain. Similarly, the controllers have been tested varying r_d and have always returned satisfactory results, showing the robustness of the controller. Therefore, the ratiometric control problem is solvable for any r_d if and only if all the cells are controllable. Finally, they observed that actually the growth rates of the two phenotypes differ because of different metabolic loads caused by the production of LacI or TetR. However, this did not lead to divergence phenomena; nevertheless, it represents an interesting observation that is worth exploring further in the future, potentially by accounting for the different metabolic loads directly in the model.

4.3 Analysis via Probability Distribution

The simulations performed in BSim demonstrate the efficiency of the controllers proposed in [44] to balance the ratios of two groups of cells within the same population. However, as anticipated, despite BSim can account for factors such as cells morphology, realistic growth and division dynamics, mutual cells interactions, and environmental parameters [50], it models intracellular dynamics through ordinary or delayed differential equations, which do not capture the intrinsic noise of chemical reactions. Therefore, our aim is to employ stochastic models to test the controllers, and hence compare the results with those of the paper.

Before doing so, we would like to introduce a way to easily compare stochastic simulations results using the probability density function at steady state. This is a typical framework for analyzing stochastic scenarios, such as Markov chains, because it gives the possibility of directly seeing the most probable event, which could not be well described by the average behavior. Applying the negative logarithm to the probability density function, one obtains the potential energy function, which is equivalent to the probability, though it could ease the study of global properties showing the basins of attraction and the barriers between equilibria. Energy-like functions are used in many applications from biology to economics or physics. Within biology, very interesting examples regard proteins conformations or species evolution. Specifically, in the former the lowest-energy landscape represents the structure assumed by the protein, and thus can be used to find the correct amino acid sequence; whereas in the latter the potential measures the likelihood of a species to survive in a given external environment. In any case, because of the noise, the system can explore different energy levels, leading to different results.

In the case of the toggle switch, observing the probability distributions $P([LacI, TetR])$, $P([LacI])$ and $P([TetR])$ for different inputs can be very useful to see the basins of attraction of the equilibria, to visualize and count the number of cells producing LacI or TetR, and to study the robustness of the system and the controllers, providing in general a much reacher description than the deterministic one. In a stochastic framework it is common to talk about functional cellular attractors, i.e. areas where it is more likely to find the state, rather than considering exact equilibria. Indeed, observe that if the RRE converges to a steady-state, typically the stochastic solution is instead such that it is the likelihood of a particular steady-state value to converge, rather than the state value itself [73]. For this reason, studying the probability distribution can be very important, though in many cases it is not trivial to obtain its explicit expression. In [83] they actually found a way to solve the CME for a symmetric toggle switch by using the Hartree mean field approximation. Here, we propose an easier approach, which constructs the probability distribution from Gillespie's or τ-leaping realizations of the CME. It is noteworthy that, since observable trajectories cannot capture the complete nature of the system because of the existence of rare events, this method is valid if a sufficiently high number of simulations is present. Other diverse methods have been tried considering stationarity or detailed balance conditions, or the CLE when applicable; however, these approaches did not yield reliable results, likely because they were not appropriate in this context.

The method we used computes $P([LacI, TetR])$ as the sum of the time spent by all the trajectories in each state (LacI, TetR) after a transient period, normalized by the total time of all simulations. The detailed steps we followed are outlined in the pseudocode below.

Algorithm 5 Probability Distribution

$t_f \leftarrow$ simulation length
$t_s \leftarrow \begin{bmatrix} 0 & \dots & t_f \end{bmatrix}$, standard vector
$t_{ss} \leftarrow$ steady state time
$C \leftarrow 2000x4000$ zero matrix
for all simulations **do**
  $x \leftarrow$ state evolution returned by a stochastic simulation method
  $x_L, x_T \leftarrow$ state components describing the evolution of the proteins *LacI* and *TetR* at steady state ($t \geq t_{ss}$)
  $x_{L,s}, x_{T,s} \leftarrow$ standardized versions of x_L, x_T based on t_s
  for all points (L, T) in the phase space grid **do**
    $C \leftarrow C+$ number of elements such that $(x_{L,s}, x_{T,s}) = (L, T)$
  end for
  $P \leftarrow \frac{C}{\text{total times*number of simulations}}$
end for

As visible in the pseudocode, we have represented the phase space via a grid of points, assigning to each point the sum of the elements in the state vector that were equal to the corresponding value on the grid. This process has been repeated for all simulations after removing a transient interval, which we defined as 1000 minutes. However, it is important to emphasize that, to do this correctly, we needed to standardize the state vector before calculating the time spent, because both Gillespie's SSA and τ-leaping method return vectors not equally spaced, as the state is updated every time a reaction occurs or for each adapting τ, rather than at fixed time intervals. In particular, we considered as standard vector the time vector divided in intervals of 15 minutes and used the *interp1* command to obtain the standard version of the state.

Once the probability distribution $P(LacI, TetR)$ was obtained, this has been plotted as a 2D histogram, together with the 1D marginal probability density functions of $LacI$ and $TetR$, as visible in Fig.4.7a and 4.8a. Note that in the figures it was necessary to multiply $P(LacI, TetR)$ by a factor of 130 to make it visible in the same plot of the marginal probabilities. Finally, we recall that these can be computed by integrating $P(LacI, TetR)$ along the direction of each protein:

$$P(LacI) = \int_0^\infty P(LacI, TetR)\, dTetR \quad \text{and} \quad P(TetR) = \int_0^\infty P(LacI, TetR)\, dLacI. \tag{4.5}$$

Fig.4.7b and 4.8b show the corresponding heat maps, where to each grid point is assigned a color based on the magnitude of the value at that point. To better visualize the quantity of cells producing LacI or TetR, the regions of attraction defined in [44] are also plotted. To determine the precise number of cells in each region is therefore sufficient to integrate the probability density over the respective LacI or TetR region, obtaining the two ratios as follows:

$$r_{LacI} = \int_{\text{LacI region}} P(LacI, TetR)\, dLacI\, dTetR, \tag{4.6}$$

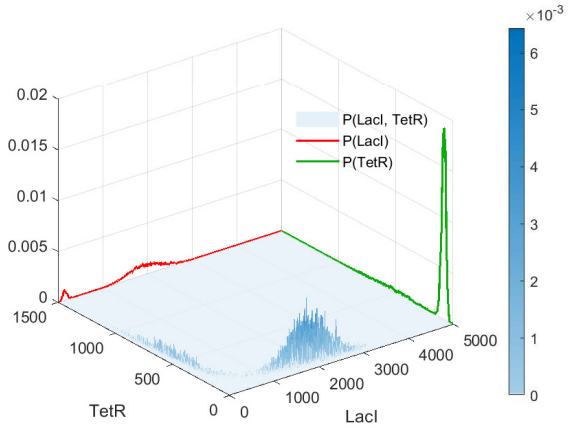
$$r_{TetR} = \int_{\text{TetR region}} P(LacI, TetR)\, dLacI\, dTetR, \tag{4.7}$$

recalling that $\int_{\mathbb{R}^2} P(LacI, TetR)\, dLacI\, dTetR = 1$.

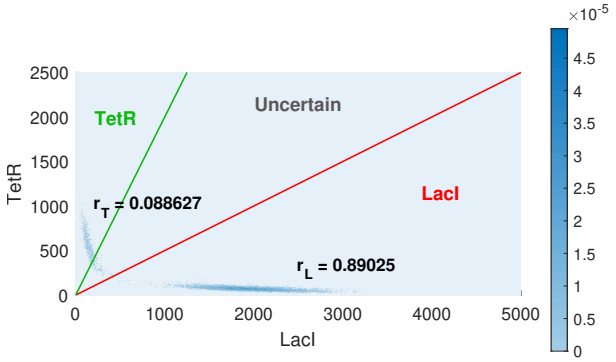
Note that the definition of the regions of attraction is arbitrary and will influence the resulting ratio values. The choice made in [44] is reasonable, as it approximates the shape of the nullclines of Fig.3.3, 3.5a, 3.5c. Therefore, we have adopted this definition in our analysis.

The procedure described has been applied to 100 simulations of the toggle switch system defined by equation (3.1), involving six variables. The simulations have been obtained as described in the previous chapter, both by Gillespie's SSA and τ-leaping method, each starting from random initial conditions. The results of both methods are very similar, proving the validity of the τ-leaping approximation, making it the preferred choice due to its significantly faster computation.

The figures below show the probability density function and the corresponding heat map derived from the simulations using Gillespie's SSA and τ- leaping method with zero input. Consistently with the nullcline configuration in Fig.3.3, we see that most of the cells are in the LacI region.

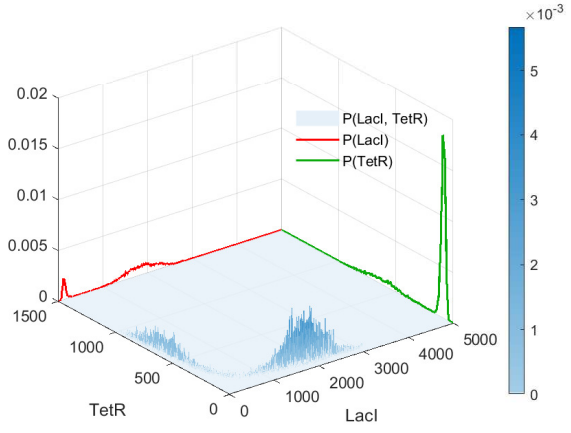


(a) 2D Probability density function of LacI and TetR, scaled by a factor of 130. 1D Marginal probabilities of LacI and TetR individually.

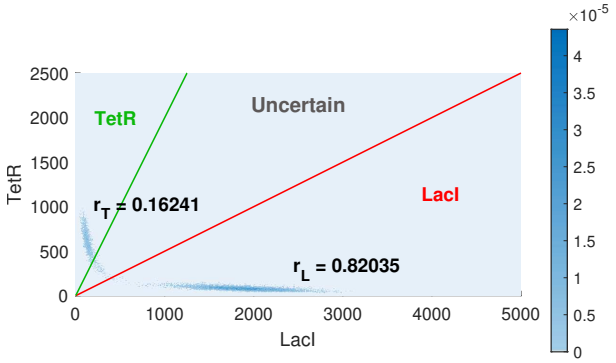


(b) Heat map of 2D Probability density function of LacI and TetR, with regions of attraction.

Figure 4.7: Probabilistic analysis via Gillespie's SSA with $u = [0\ 0]^T$.



(a) 2D Probability density function of LacI and TetR, scaled by a factor of 130. 1D Marginal probabilities of LacI and TetR individually.



(b) Heat map of 2D Probability density function of LacI and TetR, with regions of attraction.

Figure 4.8: Probabilistic analysis via τ-leaping approximation with $u = [0\ 0]^T$.

Using these graphs, as anticipated, we are able, for example, to compare the different models or to analyze the effects of the inputs. Observe that this method is much more powerful than the deterministic models, as it returns the values of the ratios that are most probable. Furthermore, it is more effective than looking solely at the simulations, especially if unwanted switchings are present.

In the previous chapter, the bifurcation analysis was performed using the deterministic model, while here we aim to analyze the effects of the inputs on the probability distribution through stochastic simulations. Depending on the input values, we can expect the probability density function to exhibit either one or two peaks, corresponding to whether the system is monostable or bistable. However, close to the border of these regions, probably frequent transitions will occur, leading to less defined peaks in the probability density function. As in the bifurcation analysis, let us consider the convex combination of the inputs defined in (4.3). We will study four cases of interest, corresponding to the monostability, bistability, or critical regions in the bifurcation diagram:

- $u_{virtual} = 1 \longrightarrow u = [U_{aTc}\ 0]^T$, for which there is a unique equilibrium in the LacI region.
- $u_{virtual} = -1 \longrightarrow u = [0\ U_{IPTG}]^T$, for which there is a unique equilibrium in the TetR region.
- $u_{virtual} = -0.07 \longrightarrow u = [46.75\,ng/mL\ 0.5325\,mM]^T$, for which we expect two equilibria, one in each region.
- $u_{virtual} = -0.24 \longrightarrow u = [38\,ng/mL\ 0.62\,mM]^T$, for which we expect to have basically one equilibrium in the TetR region, but with a lot of noise.

These input values are highlighted in the figures below by vertical red lines to allow a visual fast interpretation.

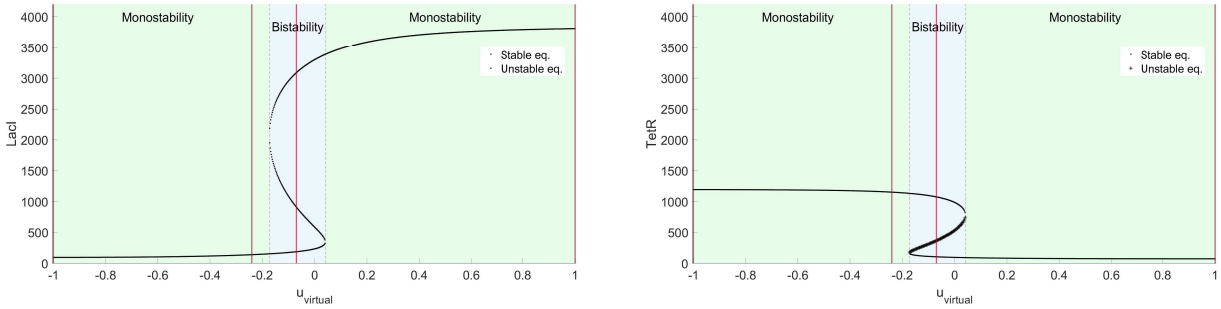


Figure 4.9: Bifurcation diagram for LacI and TetR of reduced toggle switch model (3.2), with parameter the virtual input defined in (3.4). Highlighted in red four cases of interest.

For each of these cases, the simulations have been generated using the τ-leaping method, with the initial conditions belonging to three different sets: all phase space, LacI region, or TetR region. In this way, it is possible to study the impact on the probability density function of both the inputs and the initial conditions. We have decided to compute the probability distribution after $t = 1000$ minutes since before the system is still adjusting and the probability would not be reliable. Moreover, we computed it during two different time intervals, $t \in [1000, 2500]$ minutes and $t \in [3500, 4000]$ minutes, to see the time evolution of the probability distribution. The plots obtained for all the cases described are shown in the figures below.

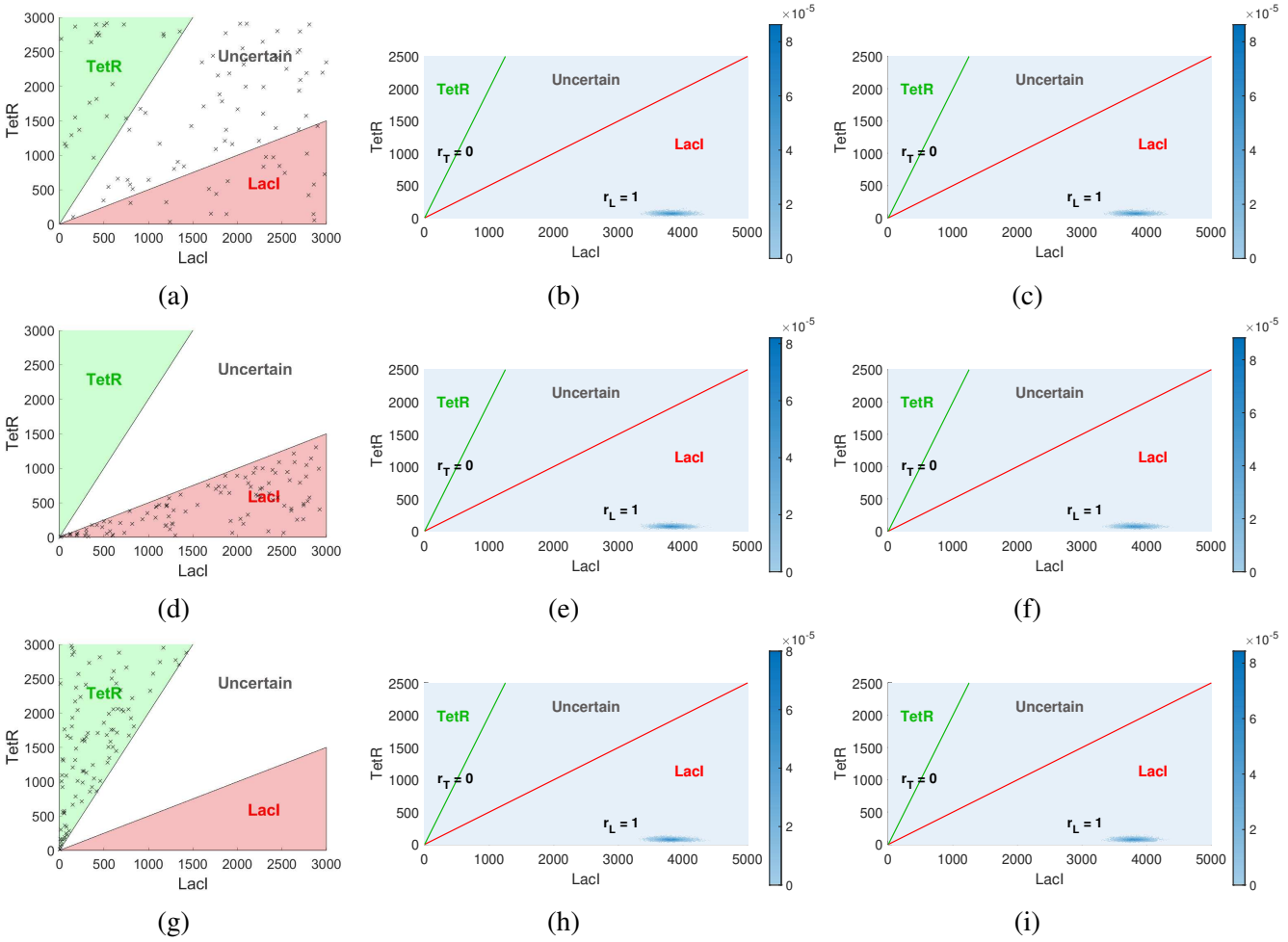
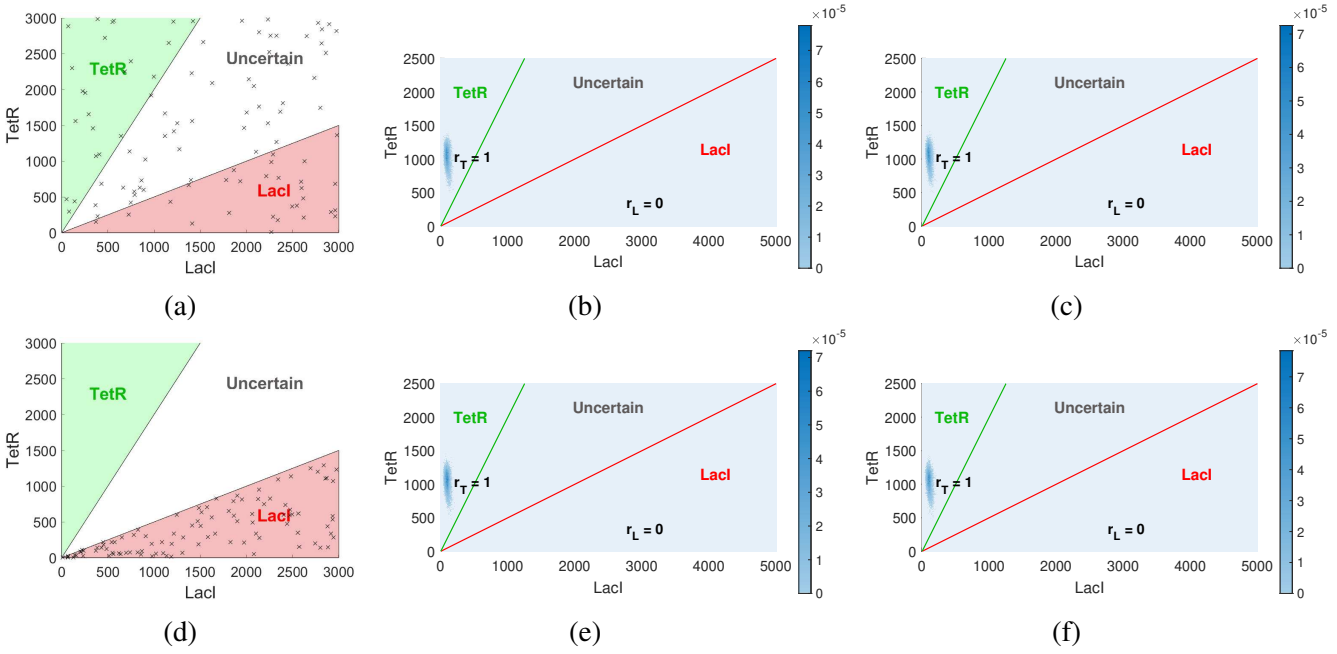
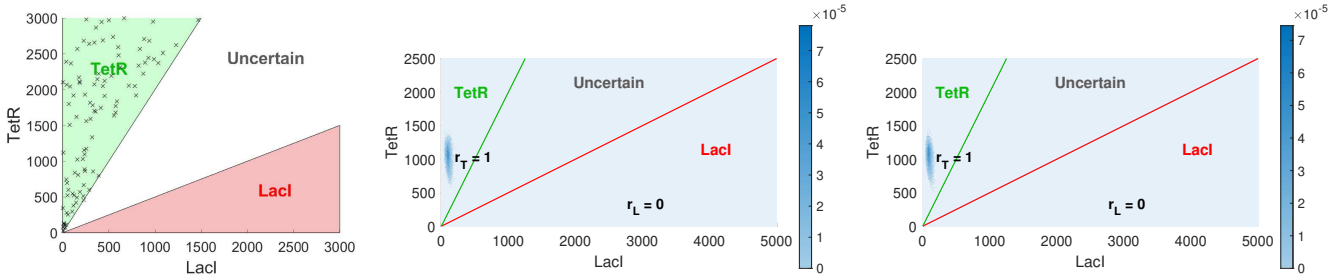


Figure 4.10: Probabilistic analysis, case 1: $u = [U_{aTc}\ 0]^T$, $u_{virtual} = 1$. (a),(d),(g) Initial conditions. (b),(e),(h) Heat map of probability density function for $t \in [1000, 2500]$ minutes. (c),(f),(i) Heat map of probability density function for $t \in [3500, 4000]$ minutes.



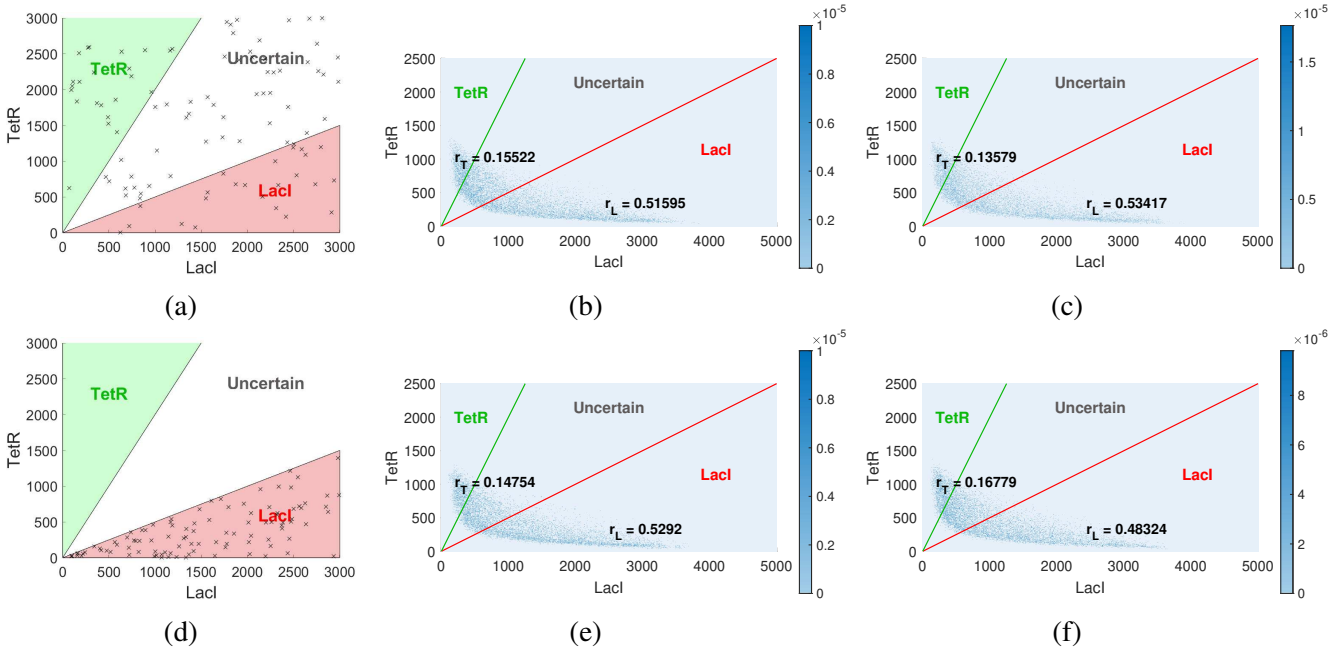


(g) (h) (i)

Figure 4.11: Probabilistic analysis, case 2: $u = [0\ U_{IPTG}]^T$, $u_{virtual} = -1$. (a),(d),(g) Initial conditions. (b),(e),(h) Heat map of probability density function for $t \in [1000, 2500]$ minutes. (c),(f),(i) Heat map of probability density function for $t \in [3500, 4000]$ minutes.

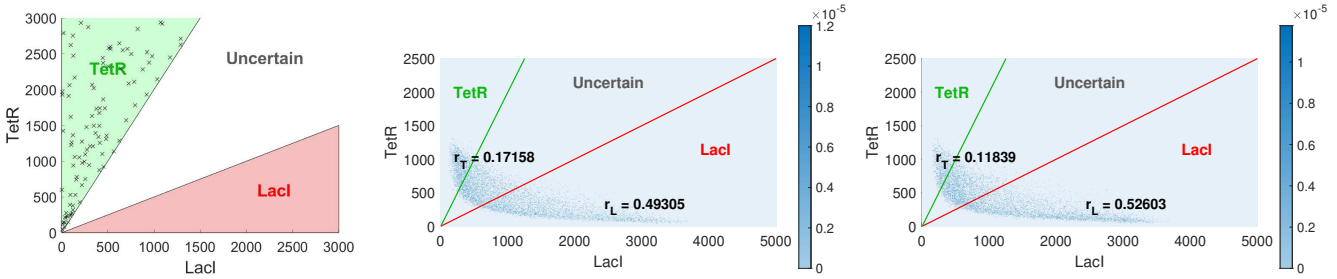
From these two cases, it is already visible that, given a certain input, the simulations converge all to the same stationary distribution, regardless of the initial conditions. Indeed, these only influence the system during the early stages, before the cells have stabilized. Moreover, as we could expect, the single equilibria are clearly visible in the heat maps, with the ratios stabilizing at either 0 or 1 for all initial conditions. This means that the inputs are strong enough not only to make the system monostable, but also to keep the trajectories very close to the deterministic equilibrium, despite the presence of the noise.

When the system is in the bistability region, trajectories can converge to either of the two equilibria. However, many transitions occur because of noise, and we cannot talk about true bistability. As visible in the figures below, this causes almost half of the cells to be in the uncertain set, which implies that the potential barrier between the equilibria is very low, making it easy for the cells to switch between the stable states.



(a) (b) (c)

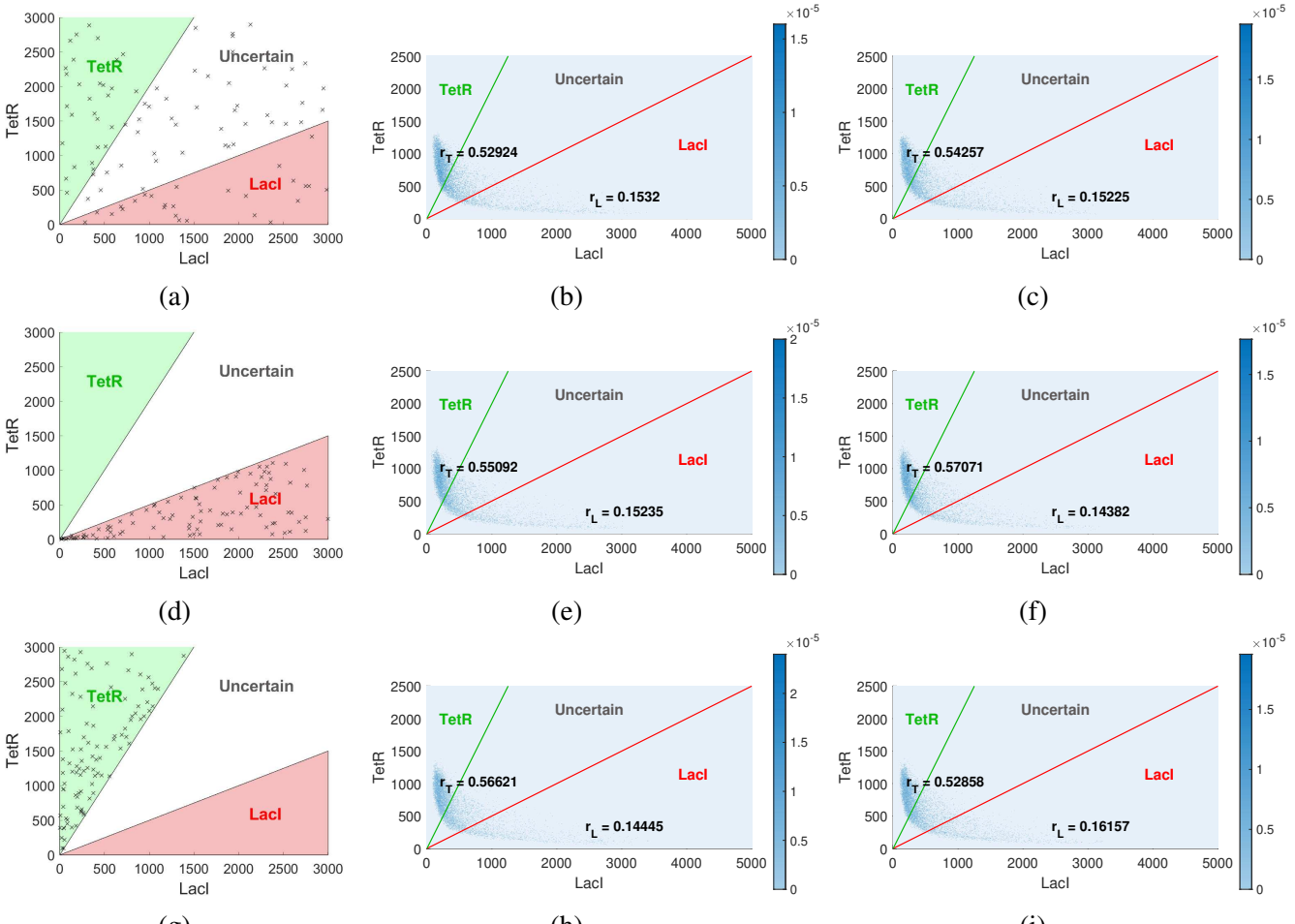
(d) (e) (f)



(g) (h) (i)

Figure 4.12: Probabilistic analysis, case 2: $u = [46.75\,\frac{ng}{mL}\ 0.5325\,mM]^T$, $u_{virtual} = -0.07$. (a),(d),(g) Initial conditions. (b),(e),(h) Heat map of probability density function for $t \in [1000, 2500]$ minutes. (c),(f),(i) Heat map of probability density function for $t \in [3500, 4000]$ minutes.

Finally, let us consider the last case, in which the inputs are such that the deterministic system is monostable with equilibrium in the TetR region, though at the same time it is very close to the bistability region.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 4.13: Probabilistic analysis, case 2: $u = [38\,ng/mL\ 0.62\,mM]^T$, $u_{virtual} = -0.24$. (a),(d),(g) Initial conditions. (b),(e),(h) Heat map of probability density function for $t \in [1000, 2500]$ minutes. (c),(f),(i) Heat map of probability density function for $t \in [3500, 4000]$ minutes.

On one hand, we know that the solution of the RREs would have behaved exactly as in the second case with $u = [0\ U_{IPTG}]^T$, for which there was a unique equilibrium in (low LacI, high TetR). On the other hand, the stochastic simulations reveal that only slightly more than half of the trajectories stays in the TetR region, while the others are either in the uncertain or the LacI region, resulting in a completely different outcome compared to Fig. 4.11.

Observing the outcomes of the cases presented, we can conclude that unless the input is strong enough to lead all simulations to the same equilibrium, as in the first two cases, there will be many uncertain cells. Indeed, around the bistability region, even if the system starts from a certain region of attraction, it will not necessarily remain there, but it will rather transit between the different regions. Moreover, we highlight that changing the inputs close to this region does not really influence the stability analysis, but rather the presence and impact of the noise. This shows the difference between deterministic and stochastic stability analysis, proving the importance of including the noise into the models. However, despite oscillations in the system, it still produces LacI and TetR proteins, even though at a slower rate. It is worth noting that realizations alone cannot lead to this conclusion, whereas the probability density function offers a clearer way to interpret the results, even in the presence of oscillating trajectories. Finally, we emphasize that, as anticipated, the probability distribution converges to a stationary one after a fast transient.

In conclusion, the most significant result we have shown is that there does not exist an input so that there are simultaneously two equilibria and no uncertain cells. This will be particularly valuable when analyzing the controller's response. Moreover, we have observed that when the system is supposed to have a single equilibrium, actually also the other equilibrium slightly appears, showing bimodality, i.e. the presence of two distinct peaks in the probability distribution even if the underlying system is monostable. Similarly, when the system is supposed to be bistable, in reality it is difficult to distinguish between the two equilibria. Therefore, we can claim that in the stochastic framework it is not appropriate to refer to the concept of bistability; and to fully understand the system's behavior, it is essential to examine the probability density function.

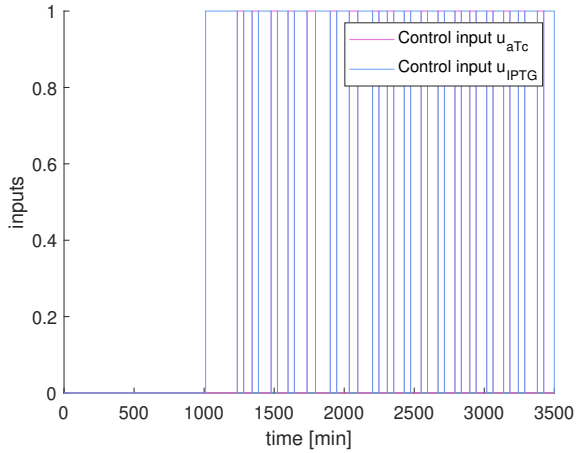
4.4 Relay Controller Validation via Stochastic Models

A detailed analysis of the system's properties has been performed, including stochasticity directly into the model, differently from the approach in [44]. In this chapter, the stochastic model will be utilized to test the controllers proposed in the paper and determine whether they yield comparable results. Indeed, we recall that the controllers were previously tested using BSim, showing their ability to reach the desired ratio (Fig. 4.5, 4.6). However, the toggle switch was modeled by

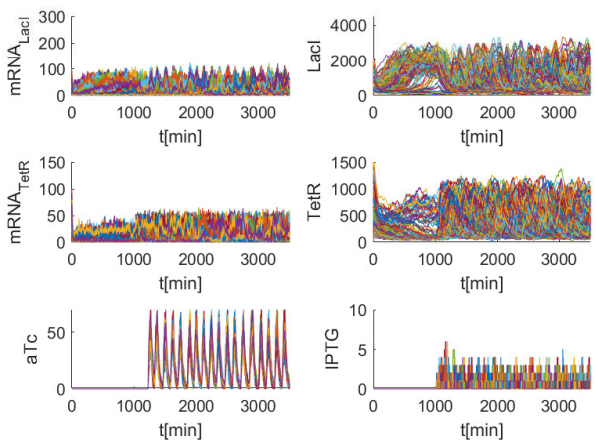
deterministic equations, and, as we have observed, this simplification reduces the richness of the real system and can lead to discrepancies when compared with stochastic simulations. Another key difference with the study carried out in the paper is that we have not considered a realistic population and chemostat dynamics, that they instead implemented using BSim. Nevertheless, our study focuses on the foundational model, offering better understanding of the system and more appropriate controllers designs. Finally, we anticipate that we have focused on the relay controller given in (4.2), as it already returned interesting and novel results compared to those presented in the paper. Therefore, for the purposes of this thesis, it is not necessary to explore more complex controllers.

In [44] a microfluidic chamber with about 200 cells was considered. Since the cells can be seen as independent stochastic processes (indeed the probability of each cell being in a specific state at a given time is a time-dependent random variable [84]), we can consider 200 simulations, which were generated using the τ-leaping method because of its reliability and computational efficiency. However, performing the simulations as before requires to assume ergodicity, which in this case implies that every time an input is given, the cells will have enough time to stabilize. If this condition holds, then simulations can be performed one after the other, with the state sampled at the end of each run. Though, this would require an impractically long amount of time, which is not feasible in practice. Indeed, in the paper they supposed possible substantial cells mutations after 24 hours. Therefore, it is necessary to consider an alternative method to realistically perform the desired simulations.

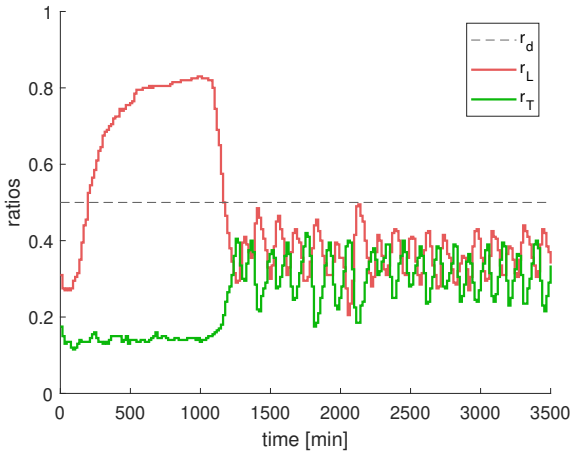
A reasonable way to generate the simulations taking into account the previous considerations, is to discretize the simulation time length and, at each time step, perform all the realizations simultaneously. In particular, the time steps we considered are of 15 minutes because it is the minimum input switching frequency considered in the paper to not overstress the cell (Table 4.1). Although this will result in longer computational time, it ensures reliable outcomes. To implement this algorithm, during each time interval, the control law generates the inputs for the next 15 minutes, and then the corresponding simulations are performed. The algorithm implements the relay controller in (4.2) using the error at the end of the previous time step. Specifically, it calculates the error by counting the number of cells producing LacI or TetR and taking the difference between these and the desired values. Observe that the determination of which protein each cell produces has been made using the regions of attraction of the paper, depicted in Fig. 4.4. Finally, it is important to note that the controller was applied only after the first 1000 minutes, allowing the cells to first stabilize in the most probable configuration, which should yield greater realism in our results. Below we report the plot of 200 simulations generated as described, all starting from random initial conditions. Note that the inputs are normalized to their maximum value, as done in the paper.



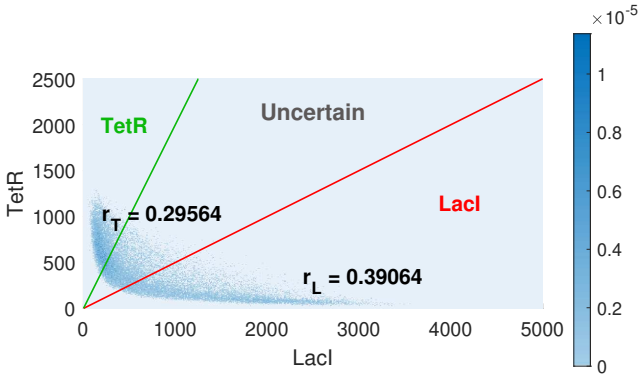
(a) Control inputs.



(b) τ-leaping simulations.



(c) Ratios evolution.



(d) Heat map of probability distribution.

Figure 4.14: Relay controller (4.2) on a population of 200 cells, using the stochastic model of the toggle switch. Parameters $r_d = 0.5$, $U_{aTc} = 60\,ng/ml$ and $U_{IPTG} = 0.5\,mM$.

From the figures, it is evident that the trajectories oscillate continuously, resulting in a significant number of uncertain cells (approximately 30% of the whole population), making it impossible to reach the desired ratio. Despite this, the controller is still able to maintain a balance between the number of cells producing LacI and TetR, even though at levels below 0.5. This result differs from the one presented in [44], which, as previously explained, used a significantly different simulation approach. However, we can claim that our findings appear more realistic because the oscillations in the input cause the cells to continuously move from one equilibrium to the other, which reasonably leads to a large proportion of cells being in the uncertain state, as observed in the plots. This is confirmed also by the fact that even when starting from the desired ratio, the controller produces an error similar to before. This is visible in the figure below.

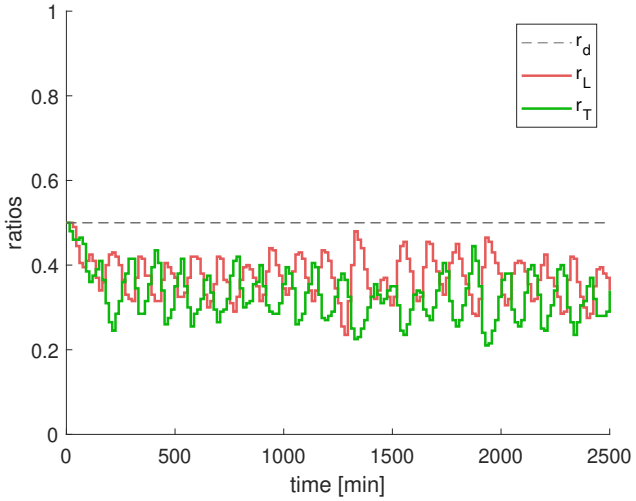


Figure 4.15: Relay controller (4.2) on a population of 200 cells, using the stochastic model of the toggle switch, starting at r_d. Parameters $r_d = 0.5$, $U_{aTc} = 60\,ng/ml$ and $U_{IPTG} = 0.5\,mM$.

4.5 Discussion

In this chapter the importance of ratiometric control has been discussed, considering both the case of multiple and single population. Regulating the relative number of two populations, or of two subgroups within a single population, is indeed fundamental to prevent one group from displacing the other, and can serve for diverse purposes ranging from multicellular controllers to innovative industrial applications. In this work, we focus on the promising single population approach proposed by Salzano et al. in [44]. This is expected to perform better than the case with two populations by avoiding problems related to different growth rates or communication molecules. Additionally, it will enable fast switching between the two groups of cells, which can be particularly useful if one of the groups dies or if the desired ratio significantly changes. The proposed method consists in equipping each cell with a bistable toggle switch, whose state determines the role of the cell, which can be modified by means of an external input. Therefore, ratiometric control problem is solved using two feedback controllers, the relay and the proportional integral controllers. These have been validated *in silico* via BSim simulations, which account for cell morphology, growth and division dynamics, cell-to-cell variability, mutual cell interactions, and environmental parameters. Although BSim represents a useful tool commonly employed to simulate cell populations, it models the toggle switch via delayed Reaction Rate Equations, which do not account for the intrinsic noise typical of chemical reactions.

Therefore, the goal of this chapter was to firstly study the stochastic behavior of the toggle switch under different inputs, and then to evaluate the controller performances on this model. To achieve this, we started by introducing a method to compute the probability distribution from the stochastic realizations. This allows to directly determine the value of the ratios and to better understand the

system's response. In particular, by looking at the heat maps of the probability distribution, we observed that using only one of the two inputs allows to drive all simulations to one unique equilibrium; whereas, when using a combination of the inputs, especially close to the bistability region, a significant amount of cells stay in an uncertain state, where it is not possible to determine whether they are producing LacI or TetR. This still holds true even when starting from a deterministically stable equilibrium, as trajectories continue to oscillate regardless of the initial conditions. Therefore, no pair of input values satisfying the given convex combination (except for the exclusive input solution) can yield zero uncertain cells. The presence of a high number of uncertain cells makes it more difficult to identify the different equilibria, emphasizing the differences between the deterministic and stochastic models, and supporting the importance of our proposal.

Finally, we have tested the proposed relay controller and obtained inconsistent results with those of the paper, sufficiently supporting the reasons behind our thesis. However, in the future also the PI controller should be tested in order to further understand the system.
The relay controller has been tested by implementing an algorithm which simulates simultaneously all the cells of a population. This accounts for inherent randomness of chemical reaction, though, it does not incorporate the chemostat dynamics that are likely considered in BSim. Including it would allow to account also for the possibility that some cells can be flushed out of the chamber and that new ones can appear, introducing additional noise at steady state. Hence, we can expect that developing a method to include also this factor into our model would produce even more oscillations in the simulations. However, even without considering it, our approach already yields different results from those of the paper. In particular, using the relay controller, we have found that this is able to balance the number of cells producing LacI and TetR, though leaving approximately 30% of the population in the uncertain state. It is important to note that, as mentioned earlier, even the cells producing one of the proteins do not remain all the time in the corresponding region; they likely stay there only for a small period of time and then move out. Since this applies to all cells, there still is an overall production of the two proteins, but this is not maximized because part of the population remains in the uncertain state. Therefore, an interesting direction for the future would be to seek inputs able to create an high barrier between the equilibria, limiting unwanted switching. If this is possible, then the barrier could be lowered when it is necessary to let the cells transit, and then raised again to ensure robusteness.

Conclusions

The thesis begins with a comprehensive study and detailed analysis of the current main modeling and simulation methodologies, to then extend a typically deterministic technique to the stochastic framework, proposing a possibly cutting-edge method to ease stochastic simulations. This has been used to prove the importance of an accurate model which includes the noise, showing the fundamental differences with the deterministic framework. In particular, we considered the controller proposed in the literature to achieve regulation of cell groups within a single population, which was validated by BSim simulations. This has been tested on the stochastic model bringing novel, but reasonable, and hence promising, important results. These can therefore be used to design more appropriate and innovative controllers.

"Assumptions of Chemical Langevin Equation breaks down with genetic toggle switch."

We provided a systematic evaluation of deterministic and stochastic modeling techniques for chemical reactions, with different methods depending on the applications. In particular, we have described the hypothesis underlying each of these methods, and presented a pseudo code to simulate the stochastic models. The Chemical Master Equation represents an exact description that is very useful when the molecular numbers are low and fluctuations are significant, though its dimension equals the total number of possible states. It can be simulated exactly by the Gillespie's Algorithm, which can be particularly slow if the number of frequent reactions is high. This led to the introduction of a new method, the τ-leaping approximation. Although this represents a valid alternative, satisfying the leaping condition can be challenging, and potential negative numbers of molecules must be handled. We presented an algorithm implementing a recent version of the τ-leaping method which allows to easily compute online an appropriate value of τ, and which switches to the Gillespie's SSA when necessary. The Chemical Langevin Equation is applicable if additionally the Poisson distribution describing the number of reactions in the τ-leaping approximation can be considered to be Gaussian. This provides an important simplified model which can significantly ease the study and the simulations when valid, indeed, to it is possible to apply sophisticated tools commonly used with Brownian motion. Though, one has to be careful because for example we surprisingly found that the assumptions broke down when trying to use the Chemical Langevin Equation with the toggle switch case. Finally, the Reaction Rates Equation represents a commonly used model, although it is valid only when the population and the volume tend to

infinity, which justifies the use of the average concentration. This can still be highly useful when studying the model, as in the case of the toggle switch to perform the bifurcation analysis, but, as already said, it can often be unreliable or insufficient.

"Limitation of separation of time-scales in stochastic dynamics."

One of the main results of this thesis is the validation in the case of the toggle switch of a method to obtain simplified stochastic models starting from Reaction Rate Equation. This consists in applying the time scale separation, a technique developed for ordinary differential equations that allows to reduce the system dimension. We have presented a way to add the noise to this reduced system and obtain stochastic simulations comparable with experimental data for various inputs. Though, this holds for a first time reduced model, indeed, when we attempted to further reduce the system, the resulting simulations showed less variability. This is likely due to the fact that, by applying the quasi-steady-state-assumption to the deterministic equations, we neglected some significant noisy terms. Therefore, it would be interesting to study a way to artificially include them again in the reduced stochastic model in order to obtain realistic results also in this case. However, in our work we considered the validated model, reduced to six variables. This is the one used to show that even if the deterministic system predicts a certain output, actually in the stochastic framework not only we have oscillations, but a completely different result can be obtained when compared with the deterministic one.

"Stochastic dynamics show limitation of population ratio control."

The final part of the thesis focuses on ratiometric control problem and in particular on the single-population approach. This represents a recent and pivotal solution to balance the relative sizes of two cell groups that are part of the same population. Because of its novelty, this proposal requires further detailed study before it is possible to transform it a real experiment. Our contribution is significant in this direction, because we found a novel result, by accounting for intrinsic randomness of chemical reactions. This differs from the one present in the literature because it exhibits non-zero steady state error due to the presence of an high number of uncertain cells. Indeed, we showed that there not exists a pair of inputs satisfying the proposed convex combination such that there are two equilibria and no uncertain cells. Therefore, for example to have an equal size of the two groups, the relay controller provided continuously oscillating inputs to make some cells move to one region, and when this becomes too populated, to let others return to the previous region, and so on. Though, this inevitably causes a big portion of uncertain cells, as shown in our validation results. Despite this, it is exactly this heterogeneity of the responses that enables ratiometric control.

4.6 Potential Future Directions

On the model...

This thesis opens up several questions for further research aiming to understand stochastic dynamics and achieve ratiometric control in a single population. As we have seen, at the base there is the necessity of finding even more efficient modeling and simulation techniques. In particular, some of the mentioned alternative methods can be investigated, focusing especially in finding a way to generally apply the quasi-steady-state assumption in the stochastic framework. On the other hand, a population and chemostat dynamics should be included in the stochastic model to account also for cell interaction and spatial factors.

... and on the controller design.

The other key future direction consists in firstly testing the proposed PI controller, and consequently develop more sophisticated controllers, including approaches that exploit stochastic dynamics. For instance, it could be interesting to look for an input pair able to ”control” the noise, lowering the barrier between the equilibria when it is necessary to let the cells transit, and then raising it again to ensure robusteness of the system. Such approach could lead to the design of non-standard controllers, specific for the study case considered. Another interesting question is to understand whether it is possible to find a controller able to decrease the number of uncertain cells. We believe that to answer this question, future work will have also to investigate if precisely maintaining the desired ratio is truly reasonable. Indeed, if the system slightly move away from it, then a necessary transient time interval with possibly more uncertain cells is necessary to let cells go in the desired regions and hence reach again the desired ratio.

To conclude...

We provided an insightful study of a possible solution to the ratiometric control problem. While promising, this approach still requires further refinement and validation via *in vivo* experiments. These create a fundamental bridge between theory and practice, albeit raising new significant challenges related to real-world experiments.

Bibliography

[1] Peter Gray et al. ≪Synthetic biology in australia: an outlook to 2030≫. In: (2018).

[2] A Michael Sismour and Steven A Benner. ≪Synthetic biology≫. In: *Expert Opinion on Biological Therapy* 5.11 (2005). PMID: 16255644, pp. 1409–1414. DOI: 10.1517/14712598.5.11.1409. eprint: https://doi.org/10.1517/14712598.5.11.1409. URL: https://doi.org/10.1517/14712598.5.11.1409.

[3] Malathy Krishnamurthy et al. ≪Bacterial genome engineering and synthetic biology: combating pathogens≫. In: *BMC microbiology* 16 (2016), pp. 1–11.

[4] Heinz Neumann and Petra Neumann-Staubitz. ≪Synthetic biology approaches in drug discovery and pharmaceutical biotechnology≫. In: *Applied microbiology and biotechnology* 87 (2010), pp. 75–86.

[5] Jan Claesen and Michael A Fischbach. ≪Synthetic microbes as drug delivery systems≫. In: *ACS synthetic biology* 4.4 (2015), pp. 358–364.

[6] Jan Roelof van der Meer and Shimshon Belkin. ≪Where microbiology meets microengineering: design and applications of reporter bacteria≫. In: *Nature reviews. Microbiology* 8.7 (July 2010), pp. 511–522. ISSN: 1740-1526. DOI: 10.1038/nrmicro2392. URL: https://doi.org/10.1038/nrmicro2392.

[7] Tae Seok Moon et al. ≪Genetic Programs Constructed from Layered Logic Gates in Single Cells≫. In: *Nature* 491 (Oct. 2012). DOI: 10.1038/nature11516.

[8] Ahmad S Khalil and James J Collins. ≪Synthetic biology: applications come of age≫. In: *Nature Reviews Genetics* 11.5 (2010), pp. 367–379.

[9] Marco Mauri et al. ≪Enhanced production of heterologous proteins by a synthetic microbial community: Conditions and trade-offs≫. In: *PLOS Computational Biology* 16.4 (2020), e1007795.

[10] Evan Appleton. ≪A design-build-test-learn tool for synthetic biology≫. PhD thesis. Boston University, 2016.

[11] Domitilla Del Vecchio and Richard M Murray. *Biomolecular feedback systems*. Princeton University Press Princeton, NJ, 2015.

[12] Desmond J Higham and Raya Khanin. «Chemical master versus chemical Langevin for first-order reaction networks». In: *The Open Applied Mathematics Journal* 2.1 (2008).

[13] Daniel T Gillespie. «The chemical Langevin equation». In: *The Journal of Chemical Physics* 113.1 (2000), pp. 297–306.

[14] *Brownian Motion: Fokker-Planck Equation*. URL: https://gu-statphys.org/media/mydocs/LennartSjogren/kap7.pdf.

[15] W. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*. Springer, 1985.

[16] Daniel T Gillespie. «Exact stochastic simulation of coupled chemical reactions». In: *The journal of physical chemistry* 81.25 (1977), pp. 2340–2361.

[17] Daniel T Gillespie. «Stochastic simulation of chemical kinetics». In: *Annu. Rev. Phys. Chem.* 58.1 (2007), pp. 35–55.

[18] Daniel T Gillespie. «A general method for numerically simulating the stochastic time evolution of coupled chemical reactions». In: *Journal of computational physics* 22.4 (1976), pp. 403–434.

[19] Daniel T Gillespie. «Approximate accelerated stochastic simulation of chemically reacting systems». In: *The Journal of chemical physics* 115.4 (2001), pp. 1716–1733.

[20] Daniel T Gillespie and Linda R Petzold. «Improved leap-size selection for accelerated stochastic simulation». In: *The journal of chemical physics* 119.16 (2003), pp. 8229–8234.

[21] Yang Cao, Daniel T Gillespie, and Linda R Petzold. «Efficient step size selection for the tau-leaping simulation method». In: *The Journal of chemical physics* 124.4 (2006).

[22] Tianhai Tian and Kevin Burrage. «Binomial leap methods for simulating stochastic chemical kinetics». In: *The Journal of chemical physics* 121.21 (2004), pp. 10356–10364.

[23] Abhijit Chatterjee, Dionisios G Vlachos, and Markos A Katsoulakis. «Binomial distribution based τ-leap accelerated stochastic simulation». In: *The Journal of chemical physics* 122.2 (2005).

[24] Yang Cao, Daniel T Gillespie, and Linda R Petzold. «Avoiding negative populations in explicit Poisson tau-leaping». In: *The Journal of chemical physics* 123.5 (2005).

[25] Lucy Ham, Megan A Coomer, and Michael PH Stumpf. «The chemical Langevin equation for biochemical systems in dynamic environments». In: *The Journal of Chemical Physics* 157.9 (2022).

[26] Stephen R Lindemann et al. «Engineering microbial consortia for controllable outputs». In: *The ISME journal* 10.9 (2016), pp. 2077–2084.

[27] Frederick K Balagaddé et al. «A synthetic Escherichia coli predator–prey ecosystem». In: *Molecular systems biology* 4.1 (2008), p. 187.

[28] Ye Chen et al. «Emergent genetic oscillations in a synthetic microbial consortium». In: *Science* 349.6251 (2015), pp. 986–989.

[29] Sergi Regot et al. «Distributed biological computation with multicellular engineered networks». In: *Nature* 469.7329 (2011), pp. 207–211.

[30] Davide Fiore et al. «Multicellular feedback control of a genetic toggle-switch in microbial consortia». In: *IEEE Control Systems Letters* 5.1 (2020), pp. 151–156.

[31] Vittoria Martinelli et al. «Multicellular PI control for gene regulation in microbial consortia». In: *IEEE Control Systems Letters* 6 (2022), pp. 3373–3378.

[32] Vittoria Martinelli et al. «Multicellular pd control in microbial consortia». In: *IEEE Control Systems Letters* 7 (2023), pp. 2641–2646.

[33] Kang Zhou et al. «Distributing a metabolic pathway among a microbial consortium enhances production of natural products». In: *Nature biotechnology* 33.4 (2015), pp. 377–383.

[34] J Andrew Jones et al. «Experimental and computational optimization of an Escherichia coli co-culture for the efficient production of flavonoids». In: *Metabolic engineering* 35 (2016), pp. 55–63.

[35] Garrett Hardin. «The competitive exclusion principle: an idea that took a century to be born has implications in ecology, economics, and genetics.» In: *science* 131.3409 (1960), pp. 1292–1297.

[36] Davide Salzano, Davide Fiore, and Mario di Bernardo. «Ratiometric control for differentiation of cell populations endowed with synthetic toggle switches». In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. 2019, pp. 927–932. DOI: 10.1109/CDC40024.2019.9029592.

[37] Spencer R Scott et al. «A stabilized microbial ecosystem of self-limiting bacteria using synthetic quorum-regulated lysis». In: *Nature microbiology* 2.8 (2017), pp. 1–9.

[38] Razan N Alnahhas et al. «Majority sensing in synthetic microbial consortia». In: *Nature Communications* 11.1 (2020), p. 3659.

[39] Alex JH Fedorec et al. «Single strain control of microbial consortia». In: *Nature communications* 12.1 (2021), p. 1977.

[40] Patrick De Leenheer and Hal Smith. «Feedback control for chemostat models». In: *Journal of Mathematical Biology* 46.1 (2003), pp. 48–70.

[41] Davide Fiore et al. «Feedback ratiometric control of two microbial populations in a single chemostat». In: *IEEE Control Systems Letters* 6 (2021), pp. 800–805.

[42] Xinying Ren et al. ≪Population regulation in microbial consortia using dual feedback control≫. In: *2017 IEEE 56th annual conference on decision and control (CDC)*. IEEE. 2017, pp. 5341–5347.

[43] Kristina Stephens et al. ≪Bacterial co-culture with cell signaling translator and growth controller modules for autonomously regulated culture composition≫. In: *Nature communications* 10.1 (2019), p. 4129.

[44] Davide Salzano, Davide Fiore, and Mario di Bernardo. ≪Ratiometric control of cell phenotypes in monostrain microbial consortia≫. In: *Journal of the Royal Society Interface* 19.192 (2022), p. 20220335.

[45] Kaushik Raj et al. ≪Fundamental trade-off between speed of switching and robustness of genetic switches limits dynamic control of metabolism≫. In: *bioRxiv* (2022), pp. 2022–03.

[46] George E. P. Box. ≪Science and Statistics≫. In: *Journal of the American Statistical Association* 71.356 (1976), pp. 791–799.

[47] Mukund Thattai and Alexander Van Oudenaarden. ≪Intrinsic noise in gene regulatory networks≫. In: *Proceedings of the National Academy of Sciences* 98.15 (2001), pp. 8614–8619.

[48] Mads Kaern et al. ≪Stochasticity in gene expression: from theories to phenotypes≫. In: *Nature Reviews Genetics* 6.6 (2005), pp. 451–464.

[49] Ankur Gupta and James B Rawlings. ≪Comparison of parameter estimation methods in stochastic chemical kinetic models: examples in systems biology≫. In: *AIChE Journal* 60.4 (2014), pp. 1253–1268.

[50] A Matyjaszkiewicz et al. *BSim 2.0: An Advanced Agent-Based Cell Simulator. ACS Synth Biol. 2017; 6: 1969–1972*.

[51] Guopeng Wei, Paul Bogdan, and Radu Marculescu. ≪Efficient modeling and simulation of bacteria-based nanonetworks with BNSim≫. In: *IEEE Journal on Selected Areas in Communications* 31.12 (2013), pp. 868–878.

[52] COPASI team. *COPASI: Biochemical System Simulator*. URL: https://copasi.org/.

[53] Stefan Hoops et al. ≪COPASI—a complex pathway simulator≫. In: *Bioinformatics* 22.24 (2006), pp. 3067–3074.

[54] Pedro Mendes et al. ≪Computational modeling of biochemical networks using COPASI≫. In: *Systems Biology* (2009), pp. 17–59.

[55] Desmond J Higham. ≪Modeling and simulating chemical reactions≫. In: *SIAM review* 50.2 (2008), pp. 347–368.

[56] Raúl Toral and Pere Colet. *Stochastic numerical methods: an introduction for students and scientists*. John Wiley & Sons, 2014.

[57] Robert Zwanzig. ≪Brownian Motion And Langevin Equations≫. In: *Nonequilibrium Statistical Mechanics*. Oxford University Press, Apr. 2001. ISBN: 9780195140187. DOI: 10.1093/oso/9780195140187.003.0001. eprint: https://academic.oup.com/book/0/chapter/421856268/chapter-pdf/52320848/isbn-9780195140187-book-part-1.pdf. URL: https://doi.org/10.1093/oso/9780195140187.003.0001.

[58] Giulio D'Agostini. *Probability and measurement uncertanty*. https://www.roma1.infn.it/~dagos/PRO/node262.html. 2001.

[59] Hana El Samad et al. ≪Stochastic modelling of gene regulatory networks≫. In: *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal* 15.15 (2005), pp. 691–711.

[60] Paul Sjöberg, Per Lötstedt, and Johan Elf. ≪Fokker-Planck approximation of the master equation in molecular biology≫. In: *Computing and Visualization in Science* 12 (2009), pp. 37–50.

[61] Parham Radpay. ≪Langevin Equation and Fokker-Planck Equation≫. In: (2020).

[62] Andrew Duncan, Radek Erban, and Konstantinos Zygalakis. ≪Hybrid framework for the simulation of stochastic chemical kinetics≫. In: *Journal of Computational Physics* 326 (2016), pp. 398–419.

[63] Howard Salis and Yiannis Kaznessis. ≪Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions≫. In: *The Journal of chemical physics* 122.5 (2005).

[64] Peter E Kloeden et al. *Stochastic differential equations*. Springer, 1992.

[65] Silvana Ilie and Monjur Morshed. ≪Automatic simulation of the chemical Langevin equation≫. In: (2013).

[66] William Zheng. ≪Methods for Solving Fokker Planck≫. In: (2019).

[67] Lukas Pichler, Arif Masud, and Lawrence A Bergman. ≪Numerical solution of the Fokker–Planck equation by finite difference and finite element methods—a comparative study≫. In: *Computational Methods in Stochastic Dynamics: Volume 2* (2013), pp. 69–85.

[68] Lee A Segel and Marshall Slemrod. ≪The quasi-steady-state assumption: a case study in perturbation≫. In: *SIAM review* 31.3 (1989), pp. 446–477.

[69] CW Gardiner. ≪Adiabatic elimination in stochastic systems. I. Formulation of methods and application to few-variable systems≫. In: *Physical Review A* 29.5 (1984), p. 2814.

[70] JAM Janssen. ≪The elimination of fast variables in complex chemical reactions. I. Macroscopic level≫. In: *Journal of statistical physics* 57 (1989), pp. 157–169.

[71] JAM Janssen. «The elimination of fast variables in complex chemical reactions. II. Mesoscopic level (reducible case)». In: *Journal of statistical physics* 57 (1989), pp. 171–185.

[72] Shev MacNamara et al. «Stochastic chemical kinetics and the total quasi-steady-state assumption: application to the stochastic simulation algorithm and chemical master equation». In: *The Journal of chemical physics* 129.9 (2008).

[73] Christopher V Rao and Adam P Arkin. «Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm». In: *The Journal of chemical physics* 118.11 (2003), pp. 4999–5010.

[74] Tianhai Tian and Kevin Burrage. «Stochastic models for regulatory networks of the genetic toggle switch». In: *Proceedings of the national Academy of Sciences* 103.22 (2006), pp. 8372–8377.

[75] Michel Laurent and Nicolas Kellershohn. «Multistability: a major means of differentiation and evolution in biological systems». In: *Trends in biochemical sciences* 24.11 (1999), pp. 418–422.

[76] TS Gardner, CR Cantor, and JJ Collins. *Construction of a genetic toggle switch inEscherichia coli. Nature403: 339–342*. 2000.

[77] Davide Fiore, Agostino Guarino, and Mario Di Bernardo. «Analysis and control of genetic toggle switches subject to periodic multi-input stimulation». In: *IEEE control systems letters* 3.2 (2018), pp. 278–283.

[78] Jean-Baptiste Lugagne et al. «Balancing a genetic toggle switch by real-time feedback control and periodic forcing». In: *Nature communications* 8.1 (2017), pp. 1–8.

[79] Agnes Köhler. «Controlling a Toggle Switch–Stochastic Modeling of Cell Populations». MA thesis. Technische Universität München, 2015.

[80] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. Chapman and Hall/CRC, 2019.

[81] Ron Milo and Rob Phillips. *Cell biology by the numbers*. Garland Science, 2015.

[82] Jasmine Shong, Manuel Rafael Jimenez Diaz, and Cynthia H Collins. «Towards synthetic microbial consortia for bioprocessing». In: *Current Opinion in Biotechnology* 23.5 (2012), pp. 798–802.

[83] Keun-Young Kim and Jin Wang. «Potential energy landscape and robustness of a gene regulatory network: toggle switch». In: *PLoS computational biology* 3.3 (2007), e60.

[84] *Stochastic process*. URL: https://en.wikipedia.org/wiki/Stochastic_process.