Approaches for the exact reduction of large-scale biochemical models

PhD in Institutions, Markets and Technologies

Curriculum in Computer Science and Systems Engineering

XXXII Cycle

By

Isabel Cristina Pérez Verona

2020

The dissertation of Isabel Cristina Pérez Verona is approved.

Program Coordinator: Prof. Mirco Tribastone, IMT School for Advanced Studies Lucca

Supervisor: Prof. Mirco Tribastone, IMT School for Advanced Studies Lucca

Co-Supervisors: Prof. Andrea Vandin, DTU Lyngby, Denmark., Prof. Max Tschaikowski, TU Wien, Austria

Tutor: Prof. Mirco Tribastone, IMT School for Advanced Studies Lucca

The dissertation of Isabel Cristina Pérez Verona has been reviewed by:

Prof. Adelinde M. Uhrmacher, University of Rostock, Germany

Prof. Tatjana Petrov, University of Konstanz, Germany

Contents

List of Figures

List of Tables

List of Abbreviations

B

BDE	*Backward differential equivalence*
BE	*Backward equivalence*

C

CaM	Calmodulin
CaMKII	Calmodulin-dependent kinase II
CL	Constrained improper lumping
CRN	Chemical Reaction Network
CTMC	Continuous-time Markov chain

E

EGF	Epidermal growth factor
EGFR	Epidermal growth factor receptor
ERODE	Tool for the Evaluation and Reduction of ODEs

F

FDE	*Forward differential equivalence*
FE	*Forward equivalence*
FOXO	Forkhead box type O transcription factor

G

GO	Gene Ontology

M

MAPK	Mitogen-activated protein kinase
MTS	Multi-transition system

N

NGF	Nerve growth factor
NGFR	Nerve growth factor receptor
O	
ODE	Ordinal differential equations
Q	
QSSA	Quasi steady-state approximation
R	
RN	Regulatory network
RTK	Receptor tyrosine kinases
S	
SBML	Systems Biology Markup Language
SBO	Systems Biology Ontology
SE	*Species equivalence*
SIS	Susceptible infected susceptible network
SMB	*Syntactic Markovian bisimulation*
SPOC	Spindle position checkpoint
T	
TS	Transition system

Acknowledgements

When it comes to be grateful the words come naturally in one's mother tongue. Sometimes, words are of no use when it comes to express feelings...but, I will try my best.

I want to start by thanking my family because without them I would no be who I am today. My family has been, and always be my rock, my safe place. Three years ago, the decision to pursuit this PhD was a leap of faith. If had not been for their constant support this thesis would have not been possible. Thank you for listening, for the advises, for the anecdotes, for the love. Thank you for believing in my capabilities and encouraging me to aim for higher goals. The distance is bitter, but I'm grateful that despite being physically far away we are closer than ever in our hearts. Mamá, Papá, Tonitín... los quiero siempre.

These where three years of hard work, of commitment and effort. I'm grateful to Lucca for helping me to see the world with new eyes, and I'm grateful that here I meet you, Jan. Your continuous energy was what pushed me forward when I doubted myself. Today would have not been possible with your support. I have the feeling that in these three years we have grown together, pushing each other to be more efficient, more professional, to be what we really want to be. Thank you for being there for me in the long nights of data analysis, for the debates, for the proofreading... and thank you for reminding me that there are other things in life besides work. We may be leaving Lucca for now, but our memories will not be left being. To my German family, I'm thankful for your support and the confidence you showed in me.

This PhD gave me the opportunity of leaving my comfort zone and immerse myself in a new research field. I was lucky to find enthusiastic colleagues and a constructive work environment. I want to thank my advisor Prof. Mirco Tribastone for all his support and for facing with patience my million questions. The opportunity of brainstorming new projects and ideas kept my mind curious and eager to find new ways of doing research. Thank you for the valuable discussions, for the advises and for keeping always a good mood. It has been a pleasure to work with you.

I want to thank to my colleagues Prof. Andrea Vandin and Prof. Max Tschaikowski. Without their valuable feedback, patience and good energy these three years would have not been the same.

I cannot thank enough to everyone who accompanied me in this journey full of emotions, of new ideas, inspiration. To my professors, to my colleagues, to the friends that I made here, thank you for making me feel a little bit home. To Giada Lettieri and her family, my "famiglia dell'altro oceano", I will keep you always close in my heart.

Last but not least, I would like to acknowledge the financial support that I received from Soroptimist International of Europe and the S.I Club of Lucca through the SIE Scholarship Grants 2019-2020.

Declaration

Chapter 3 is based on Cardelli, Luca., Pérez-Verona, I.C., Tribastone, M., Tschaikowski, Max., Vandin, A., Waizmann, T., 2019. "Aggregation of species with equivalent stochastic dynamics in chemical reaction networks by partition refinement". Submitted to Nature Scientific Reports.

Chapter 4 is based on Pérez-Verona, I.C., Tribastone, M. and Vandin, A., 2019, September. "A Large-Scale Assessment of Exact Model Reduction in the BioModels Repository". Proceedings of the 17th International Conference on Computational Methods in Systems Biology (pp. 248-265). Springer, Cham.

Chapter 5 is based on Pérez-Verona, I.C., Tribastone, M. and Tschaikowski, M., 2019, April. "Fixed-Point Computation of Equilibria in Biochemical Regulatory Networks". Proceedings of the 6th International Workshop on Hybrid Systems Biology (pp. 45-62). Springer, Cham.

Vita

November 13, 1990	Born in Ciego de Ávila, Cuba
Education	
11/2016 - current	Ph.D. student in Computer Science and Systems Engineering (XXXII Cycle) IMT School for Advanced Studies Lucca
10/2013 - 10/2016	M.Sc in Computer Science Final mark: 5.0/5.0 cum laude Central University "Marta Abreu" of Las Villas, Cuba.
09/2009 - 07/2013	B.Sc in Computer Engineering Final mark: 5.33 summa cum laude University "Máximo Gómez Báez" (UNICA), Ciego de Avila.
Working Experience	
09/2013 - 09/2016	Lecturer of Digital Image Processing. University "Máximo Gómez Báez" (UNICA), Ciego de Avila.
09/2013 - 09/2016	Head of software development group and front-end programmer. University "Máximo Gómez Báez" (UNICA), Ciego de Avila.
09/2009 - 07/2013	Recurrent Internship as software developer and analyst Desoft. SA, Cuba

Projects involved

2017-2019	Assessment of exact model reduction in biological models.
2013-2016	Design of a model for the identification of consumption profiles in electrical data. Collaboration with the Center of Studies on Informatics of the Faculty of Mathematics, Physics and Computing, University Martha Abreu of Las Villas, Cuba
2009-2013	Creation of a methodology for the selection of Emerging Patterns. Collaboration with the Pattern Recognition and Artificial Intelligence Research Group of Bioplantas and the Artificial Intelligence Laboratory of University "Máximo Gómez Báez" (UNICA), Ciego de Avila.

Scholarships and Awards

2019-2020	Soroptimist International of Europe Scholarship Grant
2013	Award for best laureate, teaching assistant and young researcher. University "Máximo Gómez Báez" (UNICA), Ciego de Ávila.

Publications

1. Pérez-Verona, I. C., Tribastone, M., & Vandin, A. (2019, September). A Large-Scale Assessment of Exact Model Reduction in the BioModels Repository. In International Conference on Computational Methods in Systems Biology (pp. 248-265). Springer, Cham.
2. Pérez-Verona, I. C., Tribastone, M., & Tschaikowski, M. (2019, April). Fixed-Point Computation of Equilibria in Biochemical Regulatory Networks. In International Workshop on Hybrid Systems Biology (pp. 45-62). Springer, Cham.
3. Pérez Verona, I. C., & Arco García, L. (2016, December). Una revisión sobre aprendizaje no supervisado de métricas de distancia. Revista Cubana de Ciencias Informáticas, 10(4), 43-67.
4. Pérez Verona, I. C. (2016, May). Modelo para el agrupamiento de series de consumo de energía eléctrica basado en el aprendizaje no supervisado de distancias (Master dissertation, Universidad Central Marta Abreu de Las Villas).
5. Millo-Sánchez, R., & Verona, I. C. P., & Vargas, J. A., & Hernández, T. R. (2015). Matrices dispersas en Java para el procesamiento de grandes volúmenes de datos. In Congreso Internacional COMPUMAT (Vol. 2015).
6. Pérez Verona, I. C., & Millo-Sanchez, R., (2015). Adaptación del método de reducción no lineal LLE para la selección de atributos en WEKA. III International Conference yayabociencias 2015. ISBN: 978-959-312-102-6
7. Millo-Sánchez, R., & Verona, I. C. P., & Antón, J. (2015). HTVisualizer: Herramienta para la visualizacion de H-Trees. Acta Informática, 1, 290-306.

Abstract

Mathematical models are a fundamental tool used in many branches of science and engineering to gain insights into the dynamics of systems. In Systems Biology [122], these models facilitate the analysis of complex biochemical networks that describe molecular interactions at different level in living organisms. In the last years, several efforts have been dedicated to the characterization of biological systems [98, 127, 128, 165, 209], specific interaction mechanisms [10, 15, 44, 45] and biological phenomena such as oscillations and bistability [80, 119, 147, 175]. As result, the use of mathematical models of biological processes have shown to be an useful asset in the development of several medical applications including drug design and target therapy [55, 110, 179, 197].

There are two main problems to solve when working with these types of models. The first one is the typically large computational cost required for their analysis, which is giving in part to the large number of configurations in which components such as proteins, or genes are present in the living cell. The second problem is related to the difficulty in calibrating large models, since they are generally associated with many parameters whose experimental estimation in living cells is notoriously difficult to make.

In this dissertation we explore several approaches for the reduction of biological systems, paying special attention to the interpretability of the aggregated system. In general, we are interested in techniques that produce an exact reduction, i.e., algorithms that given an input chemical network, produce a smaller network (consisting of fewer species and reactions)

that preserves the output dynamics of interest to the modeler, e.g. [137, 201].

We present a framework for the automatic analysis of large-scale quantitative repositories of biological models, with special support for models written in the well-known SBML [103] specifications. For networks with stochastic dynamics, we provide equivalences at the level of the Markov chain that can be applied to several models including biological networks and epidemic processes on complex networks. In addition, we approach the simplification of biochemical models by focusing on their steady-state behaviour [2], a stable condition attained by the biological system when the influence of the initial condition can be disregarded. Here, we discuss a method for the computation of equilibrium points, with the guarantee of yielding the unique equilibrium of the ODE under defined graph-theoretical conditions.

Overall, this dissertation provides a detailed analysis of techniques for the reduction of quantitative models of biochemical reaction networks, together with the interpretation of the biological and functional characteristics of the reductions obtained over a wide collection of case studies from the literature. Interestingly, the inspection of the obtained reductions has revealed *reducible motifs*, i.e., structures embedded in the structure of the network that are often compressed in the models analyzed. These findings provide the intuition that these techniques can abstract beyond the behavior of the system to capture structural/functional segments of the network that can be simplified with no (or little) effect in the dynamics of the model.

Chapter 1

Introduction

Biological processes involve molecular actors such as proteins, genes, DNA fragments and enzymes, resulting in the production of a molecular component, or the regulation of an already existing unit. A classical example are signaling pathways, groups of molecules in a cell that work in a coordinated fashion to control one or more cell functions. Each component in the pathway influences the signaling process; if a component does not behave normally, it is not possible to obtain the expected response. In the worst case, malfunctioning components inside this pathway could lead to a disturbed response, such as exacerbated cell growth or cancer [158]. To understand such level of specificity, it is not enough to observe the behavior of each molecular actor, but is required a system-level approach [72, 88, 122].

The base of a biological process comprehends the coordinated interaction among chemical components. These interaction networks can be modeled as a bipartite graph with two types of nodes to represent either the chemical entities or the reactions. The nodes are connected with annotated arcs indicating the stoichiometric information. This representation allows to analyze the network using graph techniques. For instance, [65] checks the network capability to exhibit multiple equilibria by exploring the graph of network complexes and the reaction diagram using deficient theory. Yet, emerging properties such as bifurcations or oscilla-

tions are not easily perceived from the network topology.

Chemical reaction networks (CRNs) are a well-known model to represent the behavior of biological systems. Deterministic and stochastic models can be used to equip such networks with kinetic information, leading to a dynamical representation that describes the time-course change of molecular concentrations of the species of the system. This dynamics can interpreted in two different ways: through ordinary differential equations (ODEs) or continuous-time Markov chains (CTMCs).

In the ODE interpretation each equation tracks the time-course evolution of the concentration of a species in time [199, 210]. Thus, the ODE system can generally be interpreted as an approximation of the behavior of the system. The CTMC representation instead describes the evolution of the system through a transition system. Here, a state of the Markov chain is a vector that contains the population count of the species in that state, and each reaction in the CRN represents a potential transition from a source state to a *target* state. The population count of the species in each target state is updated by observing the species which are consumed and produced in the reaction that generated such transition. Then, the complete state space for a given initial population of species is discovered by evaluating exhaustively the set of reactions in each state.

Both ODE and CTMC representations can be formally related to each other under appropriate conditions, with the ODEs being the thermodynamical limit when the populations of the species are large enough [126]. It may be often useful to consider both interpretations of a CRN; one would take the CTMC semantics as the ground-truth representation and the ODE as a suitable approximation that estimates the first-order moments.

Mechanistic models of biological processes oftentimes yield complex networks involving many state-variables and reactions. The number of species represented also depends on the level of abstraction used by the modeler. For example, the canonical MAPK pathway is oftentimes depicted as a cascade where an input signal is received and transported downstream in a three-level scheme [147], but the number of interactions between human MAPK-related proteins and other mechanisms can

reach the thousands [9].

The high dimensionality of these networks is also given by the number of possible combinations in which molecular actors can be observed. There are occasions when the modeler is interested in observing the behavior of different states of the same molecular species, e.g., active and inactive variants of a protein kinases. Protein kinases and phosphatases are fundamental in cell regulatory processes, acting as bistable switches that can be in either one of two different stable states [147]. The on/off behaviour itself is already an abstraction of the state of these proteins. Many signaling proteins have (several) binding sites specific to be phosphorylated or dephosphorylated by different kinases or phosphatases, respectively. In these circumstances, the computational model needs to account for: the inactive protein (when none of the residues is phosphorylated), the active protein (when all the residues are phosphorylated), and all the possible states of partial-phosphorylation in between; this yields large networks that grow exponentially with each additional molecular state [189].

These characteristics induce two problems: the first is the typically large computational cost for the analysis of these models, which is exacerbated when the modeler needs to make predictions at different conditions, e.g., to understand how a gene expression in a pathway depends on increasing levels of concentration of an input signal [181]. Indeed, the modeler is typically left with computationally intense approaches such as the numerical integration of ODEs (e.g., [6]) or stochastic simulation [85]. The second problem is related to the difficulty in calibrating large models, since they are generally associated with many parameters whose experimental estimation in living cells is notoriously difficult to carry out.

Both problems can be tackled using mathematically founded model abstraction techniques that translate a large model into a smaller one while preserving the dynamics of interest in a controllable fashion, simplifying simulation and validation tasks [62, 141, 174, 180, 196, 203, 204].

Simplification approaches such as sensitivity and conservation analysis can also be used to reduce model complexity. These techniques verify

the robustness of the reaction network to a subset of parameters or minimal steady state modules. For example, in [204] the authors exploit the presence of dependencies (moities) in the stoichiometry matrix to reduce by 10-15% the number of state variables, and write a simplified system in terms of the independent variables only.

Sensitivity analysis can be used in combination with time-scale techniques to identify the lesser influent variables in a model. Time-scale techniques rely on the asymptotic assumption that the dynamics of the system can be separated into *slow* and *fast* processes. Classical examples here are the Briggs-Haldane reduction of the enzyme equation [24] and the quasi steady-state approximation [182]. Recent work in [173] allows to describe the effect of the fast dynamics on the slow time scale directly in terms of the original species by partitioning of the reaction network into *slow* and *fast* modules.

Another approach for model reduction are *exact* model reduction algorithms. This category comprehends a series of reduction techniques that guarantee an (exact) relationship between the original and the reduced model. For instance, the work in [51] applies linear transformations to an ODE system to obtain a reduced system that exactly preserves the output with respect to the original. The reduction framework observes two properties of the system, namely *observability* and *controllability*. The *observability* property characterizes a state space which contains the state variables which can be observed in the system's output. The *controllability* characterizes the influence of the input over the state variables. If the state variables are not influenced by the input, this means that the chosen input has no control (or influence) over the system's output. The reduced system is obtained by simplifying all the unobserved states and focusing instead in the observable ones. Hence, the system can be simplified by removing all equations which do not influence the dynamics of the output variables [51]. In the case of observable states which are also uncontrollable and therefore not influenced by the initial input, their ODE drifts in the reduced system are simply set to the steady state values. This method is used on [52] to exactly reduce a model of EGF and insulin receptor crosstalk consisting of 5,182 ODEs to only 87 ODEs.

Starting from the original network, the idea behind *exact* methods is to produce a smaller network (i.e., consisting of fewer species and reactions) that preserves the output dynamics of interest to the modeler (e.g., [138, 201]). This leads to a coarse-grained reaction network where it is possible to observe the time evolution of a subset of the original species (e.g., the phosphorylated forms of downstream molecular complexes in a signaling pathway) whilst the behavior of some species can be simplified or collapsed into macrovariables. The conformation of the macrovariables is determined by a similarity criterion, i.e., species which are "similar" are grouped together. Thus, the selection of the grouping criteria is dependent on the aims of the modeler.

In [20], the authors apply a domain-oriented approach to collapse the different *macro-states* of scaffold proteins in a signaling circuit of the epidermal growth factor receptor (EGFR). Scaffold proteins are a special type of molecule that has more than one binding site. The state of each binding site encodes specific information such as phosphorylation activity, protein binding or phosphorylation of the bound partner. Then, a state of the scaffold is a vector where each element encodes the state of a specific site. Assuming independence among all sites, it is possible to partition the space of all the states of the scaffold such that for each *macro-state* we can write a macrovariable which is the sum of all proteins displaying the same state in a specific binding site. Thus, the network structure can be replaced by a macro-description where the dynamic is described only in terms of the variation of the concentration of the *macro-states*. Importantly, the macro-description still allows to recover the dynamics of the original microstates, owing to the fact that the micro-state concentrations are exactly expressed as the product of fractional concentrations of *macro-states* [20].

In [36], the grouping criteria captures similarities in the species behavior. Specifically, this framework considers mass-action CRNs, a special case of CRN where the propensity with which a given reaction occurs is proportional to the product of the population levels of the species involved. Given a mass-action CRN where the number of reagent partners in each reaction is at most two; the encoded ODE system is reduced by

establishing *equivalence* relations among the species in the CRN. These relations describe two distinct ways of observable behavior and are based in the notions of probabilistic bisimulation from [130] where, roughly speaking, equivalent states have the same behavior towards any equivalence class [38].

The two behavioral equivalences discussed here, *forward* and *backward* bisimulations (FB and BB respectively), induce a partition of species over the underlying ODE system. FB looks at the reaction and production rates of each species to induce a partition of species such that the sum of the ODEs of each partition block can be written as an explicit function of the sum of variables. On the other hand, BB partitions the species by grouping those that have the same ODE solutions at each point of time, starting from the same initial conditions. As result, these equivalences produce a reduced network that preserves the original dynamics in an exact way. FB induces a partition where each equivalence class represents the exact sum of the concentrations of the species belonging to that class, whereas BB characterizes exact fluid lumpability in terms of properties of the ODE vector field of a CRN [36]. Importantly, being a CRN-to-CRN transformation implies that the reduced network can still be subject to other techniques to further reduce its complexity, e.g., approximate model reduction [37] or sensitivity analysis [60].

The approach in [39] recasts such behavioral equivalences to an Intermediate Drift Oriented Language (IDOL), a syntax drift for general non linear ODEs. This allows to extend the notion of equivalence over species to relations over IDOL variables representing linear systems, CRNs or quantitative models of computing systems such as process algebra [108]. *Forward Differential Equivalence* (FDE) is presented as a notion of equivalence between IDOL variables that allows to write a specification with one variable per equivalence class, representing the sum of the trajectory solutions of its members. Whereas *Backward Differential Equivalence* (BDE) is given as an equivalence relation among variables that have the same semantics when starting from the same initial conditions.

When applied to CRNs, these aggregations fully characterize the behavioral equivalences from [36]. In addition, the notions in [39] extend

the scope of these relations to CRNs that encode Hill kinetics. In this context, FDE captures linear transformations of the original state variables such that in the aggregate ODE each macrovariable represents the sum of the original variables. The FDE partition requires that the evaluation of the polynomial that represents the quotient derivative for an equivalence class be invariant with respect to a redistribution of the values of any pair variables within the equivalence class. On the other hand, BDE associates ODE variables that have identical solutions at all time points.

In [38], the authors introduce two equivalence relations that fully characterize *forward* and *backward* equivalences over mass-action chemical reaction networks with arbitrary degree. Given polynomial ODEs, one can build a reaction network where each ODE variable is represented by a species, and each monomial in the derivatives of such variable is a reaction. Then, a *Forward Equivalence* (FE) is an equivalence relation that partitions the set of species such that each equivalence class describes the time course evolution of the sum of the ODE variables in the original model. A *Backward Equivalence* (BE) instead, induces a partition in the species such that each equivalence class contains ODE variables that have the same solutions at all time points. The reduction procedure can preserve observables of interest and yields a physically intelligible reduced model, since each aggregate corresponds to the exact sum of original variables [38].

The equivalence approaches mentioned above consider the ODE interpretation of the CRN. However, in some cases the modeler is interested in the stochastic behavior of the system. For instance, in cell regulatory processes governed by low-abundance biochemical species [95], the presence of noise caused by the inherent fluctuations in the bio-molecular processes involved [195] may introduce significant variability in gene expression [70]. Stochastic simulation can reduce the combinatorial complexity of a system by restricting the number of elements to be considered to the total of number of protein copies while the number of feasible multiprotein complexes can easily grow to millions [52].

Stochastic reaction networks form a fundamental model across many branches of science to describe populations of species interacting through

reaction channels [89]. The dynamics of these networks is governed by the master equation, a system of linear differential equations which provides the time course of the probability distribution of the underlying continuous-time Markov chain that tracks every configuration of the population levels for each species [207]. Unfortunately, to track every configuration of the species population in a stochastic network implies a combinatorial growth in the number of states of its underlying CTMC, which makes exact analysis impractical in general.

One simplification approach to maintain the stochastic behavior consists in reducing the Markov chain by means of lumpability [115]. The notion of lumpability in the context of a Markov chain characterizes a partition of the original state space such that it is possible to derive a reduced Markov chain with one macrostate per partition block [27]. Complementary methods such as finite state projection [156] and sliding window abstraction [99] truncate the state space by discarding states which accumulate low probability mass. However, these techniques work at the Markov chain level, requiring the computationally expensive enumeration of the state space.

In [40], the authors present *Syntactic Markovian Bisimulation* (SMB), an aggregation method that extends the condition of ordinary lumpability to elementary CRNs, i.e., networks where each reaction is represented as a single transition from the reactant to the product set. SMB can be applied on uni and bi-molecular elementary reactions, i.e., reactions where the set of reactants is at most two. In this context, the SMB equivalence relates two species when the cumulative kinetic parameters of the reactions, where they are involved as reagents, are the same for every equivalence class of products. This yields a structure-preserving coarse-grained network where the dynamics of each macro-species is stochastically equivalent to the exact sum of the original species in each equivalence class.

In this dissertation we explore the biological relevance of the aggregations by several approaches for the reduction of large-scale biological systems. We discuss the reduction capability of each approach, paying special attention to the interpretability of the aggregated systems. We

put special attention to exact aggregation techniques for two main reasons. First, the exact approaches that we consider produce a CRN-to-CRN transformation where the coarse-grained CRN can still be subject to all the other techniques, e.g., [37, 137, 219], to reduce the complexity of the analysis. Second, the collapse of several species into one may carry a physical interpretation that increases our understanding of the system under study. We note that the latter point is of scientific relevance regardless of the extent with which the reduction algorithm compresses the CRN.

We introduce an equivalence-based technique for stochastic mass-action reaction networks that extends the notion of aggregation over the states of the CTMC in [40] to networks with n-ary reactions. In this sense, we provide an equivalence at the level of the Markov chain which can be used to process biological networks with such semantics and more importantly, large complex networks depicting epidemic processes.

Next, we explore the equivalence methods from [38–40]. To extend the scope of these techniques, we propose a framework for the large-scale aggregation of biochemical models that comprehends the encoding of system markup biological language (SBML) [103], a standard language to model biological models, into the syntax of chemical reaction networks [169]. This allows us to tap into the BioModels database, and possibly other repositories that use this encoding format.

In addition, we analyze model reduction abstracting from the dynamics of the system to focus instead in the steady-state behavior. In certain occasions, modelers are interested in the behavior of a system when it is sufficiently away from a transient regime that depends on the conditions with which the ODEs are initialized. This leads to the study of equilibrium points, i.e., states at which the ODE solutions will settle in time.

Generally, the analysis of equilibria could be approached by means of iterative fixed-point algorithms such as Newton's method. However, as discussed in [170], there are two notable issues related to the use of such methods. The first issue is of computational nature and is due to the fact that these methods require the computation of the Jacobian at each iteration, which may be demanding when the number of ODE state vari-

ables is large. The second issue is related to the convergence properties. Indeed, Newton's method does not guarantee convergence in general. Also, it may converge to different fixed points depending on the initial guess in the case of an ODE system with multiple equilibria. Finally, while it can be used to find equilibria, it cannot be used to prove that a given ODE systems has only one equilibrium point.

Here we discuss [170], a method for the computation of equilibrium points with the guarantee of yielding an unique equilibrium of the ODE under defined conditions. Following the standard biochemical representation from [202] and [31], we decompose the behavior of the system such that the local dynamics describes the evolution between the different forms of a given component, and the network dynamics describes the influence that one form of each component exerts on the other components in the system. This leads to a more abstract representation of the reaction network where each locality in the model is given by a single vertex (most active element) and each vertex is connected to the rest by activation or inhibition arcs. We open the loops of the network by suppressing one or several edges of the original network; which allows us to write the local equilibria of some species as a function of the concentration of the other species. Then, we combine those functional expressions into a fixed-point equation that fully characterizes all the fixed-points of the system.

The structure of the thesis is organized as follows:

- **Chapter 2**: Presents a compact review of model reduction approaches used in computational biology. Then, we set the notation that will be used during the thesis and relate the problem of model reduction to the equivalence techniques in [38–40].

- **Chapter 3**: Discusses SE, an equivalence-based technique for stochastic mass-action reaction networks that extends the notion of aggregation over the states of the CTMC in SMB to mass-action networks with n-ary reactions.

- **Chapter 4**: Proposes a framework for the large-scale assessment of exact model reduction on quantitative models of biological processes. Here we provide a detailed report of the reduction results obtained over a snapshot from BioModels Database, where we look at the performance of several equivalence-based reduction techniques.

- **Chapter 5**: Shows a method for the efficient computation of equilibrium points, with the guarantee of yielding the unique equilibrium of the ODE under defined conditions [170].

- **Chapter 6:** Summarizes the main findings of the considered streams for the reduction of quantitative systems of biochemical networks and exposes the main contributions of this dissertation.

Chapter 2

Background

In this chapter we fix the thesis notation and explore diverse variants for the representation of chemical processes. Then we review popular approaches for model reduction paying special attention to lumping techniques. To conclude, we review the definitions of the semantics of the species equivalence techniques used in this thesis.

2.1 Chemical reaction networks

A **Chemical Reaction Network** (CRN) is a pair $(\mathcal{S}, \mathcal{R})$ where $\mathcal{S}$ and $\mathcal{R}$ are sets of species and reactions, respectively. The interaction among the species is described through reactions in the form:

$$\rho \xrightarrow{f} \pi \tag{2.1}$$

where ρ is the multiset of *reactant* species, and π is the multiset of *products*. We denote $\mathcal{MS}(S)$ as the set of all multisets of species in $\mathcal{S}$. The stoichiometric coefficient of a species S is denoted as $\rho(S)$ and $\pi(S)$ to indicate the species's multiplicity in the multisets of reactants and the products, respectively. The term f represents a *propensity function* $f : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}_{\geq 0}$. Which gives the rate at which the reaction occurs. Once the reaction fires,

the state update is set by the net *stoichiometry* $\pi - \rho$.[1]

2.2 ODE semantics

According to the deterministic semantics of CRNs [210], a CRN $(\mathcal{S}, \mathcal{R})$ is associated with an ODE system which tracks the time course evolution of the vector of concentrations of the species at time t. The rate of change of a component $[X](t) = ([X_S](t))_{X \in \mathcal{S}}$ is given by:

$$\frac{d[X_S](t)}{dt} = \sum_{(\rho \xrightarrow{f} \pi) \in \mathcal{R}} (\pi_S - \rho_S) \cdot f([X](t)).$$

which essentially states that variation on X's amount is the difference between the concentration of produced and consumed X over all reactions in $\mathcal{R}$.

A CRN with mass-action kinetics is a special case of CRN, where the interaction between the chemical species is described by the law of mass-action [211]. In a deterministic mass-action CRN each reaction is associated with a kinetic parameter $\lambda > 0$ and the propensity function f_λ, is given by:

$$f_\lambda(x) = \lambda \cdot \prod_{S \in \mathcal{S}} x_S^{\rho_S} \tag{2.2}$$

If the number of reactants in each reaction is restricted to at most two interacting species, i.e. $\rho \xrightarrow{f_\lambda} \pi$ where $|\rho| \leq 2$, then we are in presence of an *elementary* mass-action CRN.

Example 1 (Mass-action CRN)**.** *Let $(\mathcal{S}_E, \mathcal{R}_E)$ be a CRN with species $\mathcal{S}_E = \{S_1, S_2, S_3, S_4, S_5\}$, and reactions*

$$\begin{aligned}\mathcal{R}_E = \{&S_1 \xrightarrow{2} S_5,\ S_1 \xrightarrow{1} 2S_3,\ S_3 + S_5 \xrightarrow{3} S_3,\\ &S_2 \xrightarrow{2} S_3,\ S_2 \xrightarrow{1} 2S_5,\ S_4 + S_5 \xrightarrow{3} S_3\}\end{aligned}$$

The network in Example 1 is an *elementary* mass-action network that will be used as running example through this chapter. Following the

[1] As usual, the + and − operators denote multiset union and difference, respectively, while the multiplicity of a species denotes its stoichiometric coefficient.

law of mass action, we can express the dynamics of the network in the following ODE system:

$$\begin{aligned}
\frac{d[X_{S_1}](t)}{dt} &= -3 \cdot [X_{S_1}](t) \\
\frac{d[X_{S_2}](t)}{dt} &= -3 \cdot [X_{S_2}](t) \\
\frac{d[X_{S_3}](t)}{dt} &= 2 \cdot ([X_{S_1}](t) + [X_{S_2}](t)) + 3 \cdot ([X_{S_4}](t) \cdot [X_{S_5}](t)) \\
\frac{d[X_{S_4}](t)}{dt} &= -3 \cdot [X_{S_4}](t) \cdot [X_{S_5}](t) \\
\frac{d[X_{S_5}](t)}{dt} &= 2 \cdot [X_{S_2}](t) - 3 \cdot [X_{S_5}](t) \cdot \big([X_{S_3}](t) + [X_{S_4}](t)\big)
\end{aligned} \tag{2.3}$$

2.3 CTMC semantics

A state of the underlying Markov chain is a multiset of species, i.e., an element of $\mathcal{MS}(S)$, where σ_i denotes the population count of species S_i in that state. For instance, the state $\sigma_e = (1, 0, 0, 2)$ encodes the initial population $S_1 + 2S_4$ in the CTMC of $(\mathcal{S}_E, \mathcal{R}_E)$. For the analysis of the stochastic semantics we consider only CRNs with mass-action kinetics, which is the case covered by the techniques herein analyzed. In this context, a reaction R_k in the CRN induces a transition from state σ to a state $\sigma + \pi_k - \rho_k$ in the underlying Markov chain, with mass-action propensity given by:

$$\alpha_k \prod_{i=1}^{n} \binom{\sigma_i}{\rho_{k,i}}$$

The CTMC specification is mediated by a multi-transition system (MTS) that records transition multiplicities. This is needed to account for all the reactions contributing to the same transition. For instance, consider reactions $A + B \xrightarrow{\alpha_1} B + C$ and $A \xrightarrow{\alpha_2} C$. In both reactions one unit of A is consumed to produce one unit of C, however the production of C occurs at different speed in each reaction, i.e, α_1 and α_2.

Definition 1 (Multi-transition system)**.** *The multiset of outgoing transitions* $out(\sigma)$ *from state* $\sigma \in \mathcal{MS}(S)$ *is obtained as*

$$out(\sigma) = \left\{ \left| \sigma \xrightarrow{\lambda} \sigma - \rho + \pi \mid (\rho \xrightarrow{\alpha} \pi) \in R, \lambda = \alpha \cdot \prod_{X \in \rho} \binom{\sigma(X)}{\rho(X)} \right| \right\}$$

The set of reachable states from σ*, denoted by* $reach(\sigma)$*, is the smallest set such that:*

1. $\sigma \in reach(\sigma)$*;*
2. *if* $\sigma' \in reach(\sigma)$*, then the target states of each transition in* $out(\sigma')$ *belong to* $reach(\sigma)$*.*

Finally, for an initial state $\sigma_0 \in \mathcal{MS}(S)$*, the MTS for* (S, R) *and* σ_0 *is the union of the multisets of transitions outgoing from any reachable state, i.e.,* $MTS(\sigma_0) = \biguplus_{\theta \in reach(\sigma_0)} out(\theta)$*.*

Consider the state $\sigma_{0e} = S_1 + S_4$ as an initial population for $(\mathcal{S}_E, \mathcal{R}_E)$ (Example 1). Then, the possible outgoing transitions for σ_{0e} are:

$$out(\sigma_{0e}) = \{\!|\sigma_{0e} \longrightarrow S_5 + S_4, \sigma_{0e} \longrightarrow S_4 + 2S_3|\!\}$$

due to the first and second reactions in $\mathcal{R}_E$, respectively.

The CTMC underlying a CRN for an initial state consists of all states and transitions generated by applying exhaustively the reactions on all generated states, starting from the initial one. Given an MTS, the CTMC is obtained by collapsing all transitions between the same source and target into a single CTMC transition and summing their rates. Let (S, R) be a reaction network and σ_0 an initial population. This means that the underlying *CTMC of* (S, R) *for* σ_0 has states $reach(\sigma_0)$ and transitions given by

Definition 2 (CTMC)**.**

$$MC(\sigma_0) = \left\{ \sigma \xrightarrow{r} \theta \mid \sigma, \theta \in reach(\sigma_0) \ \wedge \ \sigma \neq \theta \ \wedge \ r = \sum_{(\sigma \xrightarrow{r'} \theta) \in MTS(\sigma_0)} r' \right\}$$

For any two states $\sigma, \theta \in \mathcal{MS}(S)$ the element of the generator matrix *of $MC(\sigma_0)$ from σ to θ is defined as:*

$$q(\sigma,\theta) = \begin{cases} r & \text{if } \sigma \neq \theta \ \wedge \ (\sigma \xrightarrow{r} \theta) \in MC(\nu_0) \\ - \sum_{\theta' \in \mathcal{MS}(S) \text{ s.t. } \theta' \neq \sigma} q(\sigma,\theta') & \text{if } \sigma = \theta \\ 0 & otherwise \end{cases}$$

For any set of states $\mathcal{M} \subseteq \mathcal{MS}(S)$, we define $q[\sigma, \mathcal{M}] = \sum_{\theta \in \mathcal{M}} q(\sigma, \theta)$.

For any two distinct states σ and σ', we denote $q(\sigma, \sigma')$ as the sum of the transition rates from σ to σ' across all reactions. These values form the generator matrix Q, which characterizes the dynamical evolution of the Markov chain by means of the master equation $\dot{p} = Qp^\top$. Each component of its solution, $p_\sigma(t)$, is the probability of being in state σ at time t [207], starting from an initial condition where the probability mass is concentrated at the initial state $\hat{\sigma}$, i.e., $p_{\hat{\sigma}}(0) = 1$.

2.4 Model reduction

Methods of model reduction project the set of reactants (and the set of reactions) to some space where a subset of the original dynamic behavior can be approximated [190]. This can be done by finding relationships among the reactants and thus reducing the amount of species which needs to be observed independently. For instance, time-scale exploitation methods separate the components in the system into groups of *slow* or *fast* components in comparison with the dynamics of the variables of interest. This is possible thanks to the different speeds on which chemical processes occur inside a biological system [216]. Some examples of time-scale techniques are the steady-state (SSA) and equilibrium approximations (EA) [79] and the work of [145, 187] to characterize the slow manifold.

A well known example in this context is the work of Briggs-Haldane for the simplification of the enzyme-substrate equation [24]. The equation is a two step mechanism that ends with the production of new species. The first step characterizes the reversible binding of a substrate with a

mediator enzyme forming a complex. The second step is given by the decomposition of the enzyme-substrate complex, releasing the enzyme units and transforming the substrate in a new species, namely *product*. Intermediate species, such as the enzyme-substract complex, tend to exist locally in the environment of the reaction mixture. In this mechanism, the concentration of the molecular complex depends only on the concentration of the reacting enzyme and substrate.

Briggs quasi steady-state assumption (QSSA) is that total concentration of the enzyme, meaning the free enzyme and the substrate-bound enzyme, is much lower than the concentration of the substrate and the *product*. Assuming that the mechanism is mass-preserving, the concentration of the complex is negligible in comparison with both the concentration of the substrate and the *product* [24]. Thus, the rate change of the complex is also negligible in comparison with the rest of the mechanism and a slow-time scale model can be written by setting the net formation of such complex to zero. This comprises the two step mechanism into a single equation, which is the Michaelis-Menten formula.

A major problem of time-scale approaches in general is the need of a deep knowledge of behavior of the species to identify a clear separation between the different timescales. This can be particularly difficult, especially when analyzing large models. In such cases, sensitivity analysis techniques can be used to split the set of species to recognize components with little or no influence in the dynamics of the system, e.g., [60, 219].

There are many streams for the reduction of biochemical models; a detailed tutorial covering of several approaches for model reduction is provided in [161]. For the interested reader, we also recommend consulting [178, 190]. The work in [178] discusses a compilation of algorithms for the reduction of dynamical models in computational biology, whereas [190] reviews general approaches of model reduction for large-scale biological systems. In this section we focus on lumping techniques applied to the context of model reduction. This provides the mathematical background required in the next sections to discuss ordinary and exact lumpability as equivalence relations for states of the Markov chain.

Lumping techniques addresses the model reduction problem by finding a mapping between the variables in the original system and a system of lower dimension. In this sense, at least one state variable is removed from the original model and replaced with a new *lumped* variable that represents a direct mapping from the original.

Let us recall X, the vector containing the concentration of all species. We can express the change in time of this vector as:

$$\frac{dX}{dt} = \mathbf{f}(X; k) \qquad \text{s.t} \qquad X_i(t_0) = [S_i](t_0) \tag{2.4}$$

where $k = (k_1, \ldots, k_n)$ is the vector of kinetic constants which characterizes the speed at which the reactions occur; and $\mathbf{f}$ is a function that calculates the variation of the concentration values depending on k. The objective is to produce a reduced system $d\hat{X}/dt = \hat{\mathbf{f}}(\hat{X}; \hat{k})$ where $\hat{X}$ can be a subset of the original species so the dimension of $\hat{X}$ is lower than X [125].

Linear lumping transforms the vector of concentrations to a lower dimension vector so fewer species are tracked independently, meaning:

$$\hat{X} = \sum_{j=0}^{n} m_{ij} X_j \qquad\qquad i = \{1. \ldots, l\} \tag{2.5}$$

where m_{ij} is an element of the lumping matrix M. Li and Rabitz [139] apply this scheme to systems with arbitrary non linearities by introducing invariant subspaces of the reaction system. Here, the reverse calculation of equation 2.5 is given by:

$$X = \bar{M}\hat{X} \tag{2.6}$$

where $\bar{M}$ is the right inverse of M, such that the multiplication of $M \cdot \bar{M}$ is equal to the $\hat{n}$-identity matrix $I_{\hat{n}}$. Evaluating this in equation 2.4 we get

$$\frac{d\bar{M}\hat{X}}{dt} = \mathbf{f}(\bar{M}\hat{X}; k) \tag{2.7}$$

Assuming $\bar{M}$ constant, it is then possible to extract it out of the derivative so $\bar{M}(d\hat{X}/dt) = \mathbf{f}(\bar{M}\hat{X}; k)$. Reorganizing we obtain

$$\frac{d\hat{X}}{dt} = M\mathbf{f}(\bar{M}\hat{X}; k) \tag{2.8}$$

If the system is lumped in this way, then the number of governing equations for the dynamics of the considered system is reduced from n in equation 2.4 to $\hat{n}$ in equation 2.8.

As pointed out in [125], given a reactivity matrix K the lumping error in unimolecular systems can be calculated as:

$$E \equiv MK - \hat{K}M \tag{2.9}$$

where $\hat{K}$ is a diagonal matrix with the corresponding values of the the eigenvalues of K. If the system is exactly lumpable (with respect to the lumping matrix M) then equation 2.9 yields a unique null matrix. Whereas if the system is not exactly lumpable, E depends on the diagonal matrix $\hat{K}$.

The necessary and sufficient condition for exact lumping of nonlinear systems is that the transpose of the Jacobian matrix $J^R(X)$ of $f(X)$ has non trivial fixed invariant subspaces $\mathcal{M}$ [139]. To protect the components of interest (*observables*), i.e., variables which will not be lumped, M can be written as

$$M = \begin{pmatrix} M_{obs} \\ M_{det} \end{pmatrix} \tag{2.10}$$

where M_{obs} are the *observables* and M_{det} are the components determined from a search of the invariant subspaces of the system.

Finding such mapping becomes a difficult task, especially in very large models. A good alternative in such cases is to treat the species as a continuum, which means that the reduced model consists of a single lumped macrovariable that can be tracked by monitoring the distribution of the data [47, 160]. Here, we focus on discrete lumping, i.e., approaches that take into account each individual variable in the original system.

The term lumping is used to describe a wide range of methods for the reduction of dynamical systems [190]. According to the contribution of the original species in the aggregated system, lumping can fall in two categories: proper and improper. Proper lumping refers to a scheme where each of the original variables appears in only one lumped variable of the reduced model [161, 190]. Following the principle of invariant response, the reaction rate of the lump depends then only on the sum of the lumped species. In the improper lumping scheme, instead, each variable from the original model can contribute to more than one lump [161].

In the recent years, lumping techniques have been used in the literature to reduce a number of biochemical systems [34, 36, 38, 40, 63, 78, 194, 200]. In [102] the authors used a combined approach of sensitivity analysis together with PCA to remove redundant reactions over the simulation conditions followed by a proper lumping scheme for further reduction. The calculation of $\bar{M}$ is given in terms of α, a matrix that represents the share of each individual species within its lump, i.e., α_{ij} represents the share of species j within the lump i. The lumping scheme was validated by reducing a model describing oxidation of fuel-rich methane mixtures from a size of 155 species to 53. Another combined approach of lumping and optimization is proposed in [56]; where the authors reduce effectively a 20D model of yeast glycolytic oscillations to 8 dimensions using a lumping and subsequent optimization approach (LASCO).

In [63] the authors propose a proper lumping algorithm for the automatic order reduction of system biology models and tested it reducing a 26-state NF-κB signaling model to 12 final states. The authors use the Penrose inverse, a unique matrix M^+ which is the pseudo-inverse matrix of M as candidate for $\bar{M}$. The lumping algorithm searches through all the possible lumping schemes and tries to lump them; i.e., in each step the lump is verified by comparing the result of the simulation of the original to the reduced model. This procedure is repeated for all possible lumpable configurations.

Similar to the approach in [21], the authors in [124] use domain-oriented lumping [21, 51] over a layer-based formalism for the reduction of signaling pathways. First, the model is separated in three different layers

according to a biological process: ligand binding, binding site phosphorylation and effector binding, and the reaction rates for within each process is assumed to be equal. Then, the macrospecies are built by lumping species in each layer, e.g., species which are phosphorylated can be lumped together regardless of their binding status. This allows to write a reduced system with equations given in terms of the macro-species and a gross rate.

The work of [194] also divides the system in modules, but in this case based on properties of the reaction graph. The lumpability criterion groups states so that the interaction among the states within a lump occur at a much faster time scale than the rest of the reactions in the model. Groups of states that are connected by such fast reactions are called DGFR. Each DGFR is checked for strong components (SC), i.e. components in the group which are interconnected with fast reactions. If there are no fast reactions to states outside the strong component, then this is a *sink* component. Finally, the algorithm checks if it is possible to lump strong components with any of the *sink* SC in their DGFR.

In [78] the authors use a rule-based description to facilitate the analysis of the reaction mixture. The authors work with a model of protein interactions in the form of a site-graph, a graph generalization where the nodes represent proteins and their sites (i.e., binding domains, residues, etc.) and the edges indicate bonds between proteins. The dynamics of the protein interactions is modeled by constructing a Markov chain on a space of site-graphs. Unlike the species representation of the reaction network where thousands of entities are required to model the dynamics of the system, rule-based modeling [14, 58] allows to generate the complete state space using a smaller set of rules that describe the interaction patterns between the molecular entities. The framework in [78] identifies patterns, namely abstract species or stochastic fragments, on the rule-based language *Kappa* (κ) [58] that represent a syntactic criteria that yield a sufficient condition of weak lumpability on the Markov chain. The detection of fragments involves characterizing the states of the CTMC that can be lumped while preserving lumpability [171]. This is achieved by equating pairs of sides which are related directly or indi-

rectly within the left-hand or right-hand side of a rule [81, 171], enforcing a strong independence between the uncorrelated sites.

2.5 Equivalences over species in a CRN

Two of the chapters of this thesis assess the equivalence-based approaches from [36, 38–40] for the reduction of quantitative biological models. Importantly, the equivalence over species provided by these methods can be seen as a generalization of equivalence relations for Markov chains.

Markov chain lumpability *Ordinary* and *exact* lumpability are two equivalence relations that can be established over the states of a Markov chain. Given a generator matrix Q, ordinary lumpability is an equivalence relation that partitions the state space into N blocks $P_1, \ldots, P_N$, such that all states in a block have equal aggregate rate toward any block [27, 115], that is,

$$\sum_{\vartheta \in P_j} q(\sigma, \vartheta) = \sum_{\vartheta \in P_j} q(\sigma', \vartheta) \quad \text{for } P_i, P_j \text{ and any } \sigma, \sigma' \in P_i.$$

These rates form the generator matrix of the lumped Markov chain, where each macro-state corresponds to a partition block. Ordinary lumpability preserves the stochastic behavior in the sense that the probability of being in one block in the lumped Markov chain is equal to the sum of the probabilities of being in states of that block in the original Markov chain [27].

On the other hand, exact lumpability partitions the state space into N blocks $P_1, \ldots, P_N$, if and only if it holds that for any two blocks P_i and P_j and any states σ, σ' in P_i

$$\sum_{\vartheta \in P_j} q(\vartheta, \sigma) = \sum_{\vartheta \in P_j} q(\vartheta, \sigma')$$

Both lumpability criteria observe the transitions within states of the Markov chain. Ordinary lumpability looks at the outgoing transitions from each state, whereas exact lumpability looks at the incoming transitions.

Overview of the equivalence-based techniques The equivalence techniques hereby mentioned identify a criterion which allows to induce an *equivalence* relation on the species of a CRN. This allows to partition the set of species using the *equivalence* relation as splitting condition, such that equivalent species are lumped in the same block and one of them is chosen as block representative.

To build the reduced network we update the set of reactions by replacing each species by its representative. This ensures that the resulting network is written having a macro-species per partition block. Then, the reaction rates of the updated reactions are scaled down by the product of sizes of the block that contains the reagent species. Finally, we reduce the set of reactions by merging all those that have the same reactants and products by summing their kinetic parameters [38]. The final outcome is a reduced network where each representative can be interpreted as a macro-species that tracks the sum of the populations of the distinct species in the original network that belong to the same *equivalence* class.

These techniques use a partition refinement algorithm to compute the largest equivalence, which is the coarsest reduction. Given an equivalence $\mathcal{E}$ and an initial partition $\mathcal{H}_{\mathcal{E}}$, which is the trivial partition consisting in all species, the algorithm refines such partition to find the largest equivalence. The algorithm maintains a reference to the current candidate partition and a set of blocks that will be used to check the candidate partition ("splitters"). Both, structures are initialized to $\mathcal{H}_{\mathcal{E}}$. Then, a fixed-point iteration algorithm splits a block of the current candidate partition whenever it falsifies $\mathcal{E}$ with respect to a splitter.

The algorithm refines the falsifying block into sub-blocks that have equal values for the quantities in the partitioning criteria, thus creating a new potential splitter to be used in future iterations. Importantly, the largest sub-block may be ignored [164], and therefore not included in the set of splitters. The refinement process stops when no further block is found that breaks the equivalence condition, hence the current partition must be the largest refinement of $\mathcal{H}_{\mathcal{E}}$. The computational cost of such calculations is polynomial in time and space for FE, BE and SMB.

To make this thesis self-contained, in the next sessions we review the definitions of the semantics of such equivalence techniques. A brief overview of these techniques can be found in [169]. For a detailed tutorial, consult [199]. In addition, all the equivalence techniques discussed here are available in ERODE [35], a public software tool for the reduction and simulation of differential equations[2].

2.5.1 *Forward* and *Backward* equivalences

Forward and *Backward* equivalences (FE and BE) are the polynomial dynamical systems analogues of *ordinary* and *exact* lumpability in Markov chains, respectively. Both reduction techniques can be applied over mass-action CRNs as equivalence relations on species. Furthermore, both equivalences can be efficiently checked relying only in structural conditions of the reactions [38].

For completeness, we restate the FE and BE definitions from [38]:

Definition 3 (FE-BE deffinition)**.** *Given a CRN $(\mathcal{S},\mathcal{R})$, a partition $\mathcal{H}$ of species is χ ($\chi \in \{FE, BE\}$) iif. for any two blocks $H, H' \in \mathcal{H}$ and any two species $S_i, S_j \in H$ it holds:*

$$\mathbf{c}_\chi(S_i,\eta,H') = \mathbf{c}_\chi(S_j,\eta,H') \quad \forall \eta.\ \exists (S_k + \eta \xrightarrow{\lambda} \pi) \in \mathcal{R} \text{ for } S_k \in \{S_i, S_j\}$$

where $\mathbf{c}_\chi$ maps a species (S_i,S_j), a multiset of reagent partners (η) and a block (H') into a real number computed by inspecting only the reactions.

Before defining the specific mapping functions for FE and BE, let us recall the notions of net stochiometry provided in [38]. Let $S_i \in \mathcal{S}$ be a species, and $G \subseteq \mathcal{S}$. The net stochiometry of S_i and H'due to reagents η is defined in equations 2.11 and 2.12, respectively.

$$\phi(\eta, S_i) = \sum_{\eta \xrightarrow{\alpha} \pi} (\pi_i - \eta_i) \cdot \alpha \quad (2.11) \qquad \phi(\rho, H') = \sum_{S_1 \in H'} \phi(\eta, S_i) \quad (2.12)$$

Then, to verify that χ is an FE we compute:

$$\mathbf{c}_{FE}(S_i,\eta,H') = \frac{\phi(S_i+\eta, H')}{[S_i+\eta]!}$$

[2]Available for download from `https://sysma.imtlucca.it/tools/erode/`

$$\mathcal{H}_f = \{\{S_1, S_2\}, \{S_3, S_4\}, \{S_5\}\}$$
$$\mathcal{S}_f = \{S_{1,2}, S_{3,4}, S_5\}$$
$$\mathcal{R}_f = \{S_{1,2} \xrightarrow{1} S_5,\ S_{1,2} \xrightarrow{0.5} 2S_{3,4}, S_{3,4} + S_5 \xrightarrow{3} S_{3,4},$$
$$S_{1,2} \xrightarrow{1} S_{3,4},\ S_{1,2} \xrightarrow{0.5} 2S_5\}$$

(a) FE-reduction

Figure 1: FE-reduction of $(\mathcal{S}_E, \mathcal{R}_E)$ from Example 1.

where the operator $[\cdot]!$ denotes the multinomial coefficient induced by a multiset of species η:

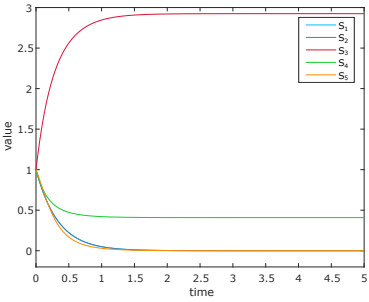
$$\eta! = \binom{|\eta|}{\{\eta(S_i) \mid S_i \in S\}} = \frac{|\eta|!}{\prod\limits_{S_i \in S} (\eta(S_i)!)}$$

The equivalence established by FE relates species such that it is possible to rewrite the ODEs underlying the CRN in terms of sums of the variables in each block. This is easily seen in Figure 1 where the FE-condition identifies the lumpable species $\{S_1, S_2\}$ and $\{S_3, S_4\}$ in $\mathcal{S}_E, \mathcal{R}_E$ (Example 1). The FE partition $\mathcal{H}_f$ divides the set of species such that the lumpable species are treated as a macrovariable in the FE-reduced CRN $(\mathcal{S}_f, \mathcal{R}_f)$. In our representation, each macrovariable is denoted with a subscript that refers to the indices of the species that it represents.

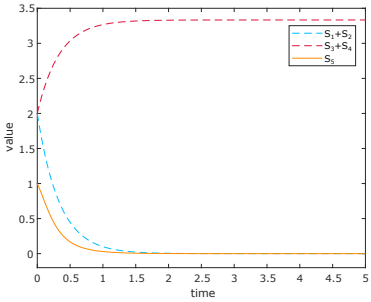
The FE-reduced network is represented in the following ODE system:

$$\frac{d[X_{S_{1,2}}](t)}{dt} = -3 \cdot [X_{S_{1,2}}](t)$$
$$\frac{d[X_{S_{3,4}}](t)}{dt} = \frac{3}{2} \cdot [X_{S_{1,2}}](t)$$
$$\frac{d[X_{S_5}](t)}{dt} = \frac{3}{2} \cdot [X_{S_{1,2}}] - 3 \cdot [X_{S_{3,4}}](t) \cdot [X_{S_5}](t)$$

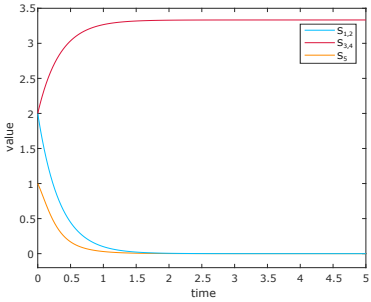
The macrovariables in CRN $(\mathcal{S}_f, \mathcal{R}_f)$ can be used to study the concentration of the sum of the original lumped variables. For instance, the traces of $S_{3,4}$ in the FE-reduced ODE system (Fig 2 c) are used to track $S_3 + S_4$ in the original model (Fig 2 a). Instead, the un-lumped species,



(a) Solutions of the original model



(b) Original model with sums of solutions



(c) Solutions of the FE-reduced model

Figure 2: ODE simulation of the FE-reduced model from Example 1.

such as S_5, represent the same chemical entity in both the original and the FE-reduced model. Assuming all species start from the same initial conditions, i.e., $[S_i] = 1$ $(i = 1, \ldots, 5)$, Figure 2 shows the solutions of the original and the FE-reduced model.

The equivalence established by BE relates two species S_i and S_k if they start from the same initial conditions and have the same behavior in time, i.e., equal consumption and production rates. To verify that an equivalence χ (Definition 3) is a BE we compute:

$$\mathbf{c}_{BE}(S_i, \eta, H') = \sum_{S_k \in H'} \sum_{\mathcal{M} \in \eta} \frac{\phi(S_k + \mathcal{M}_i, S_i)}{|S_k + \mathcal{M}|_{H'}}$$

$$\mathcal{H}_b = \{\{S_1, S_2\}, \{S_3\}, \{S_4\}, \{S_5\}\}$$
$$\mathcal{S}_b = \{S_{1,2}, S_3, S_4, S_5\}$$
$$\mathcal{R}_b = \{S_{1,2} \xrightarrow{1} S_5,\ S_{1,2} \xrightarrow{0.5} 2S_3,\ S_3 + S_5 \xrightarrow{3} S_3,$$
$$S_{1,2} \xrightarrow{1} S_3,\ S_{1,2} \xrightarrow{0.5} 2S_5,\ S_4 + S_5 \xrightarrow{3} S_3\}$$

(a) BE-reduction

Figure 3: BE reduction of $(\mathcal{S}_E, \mathcal{R}_E)$ from Example 1.

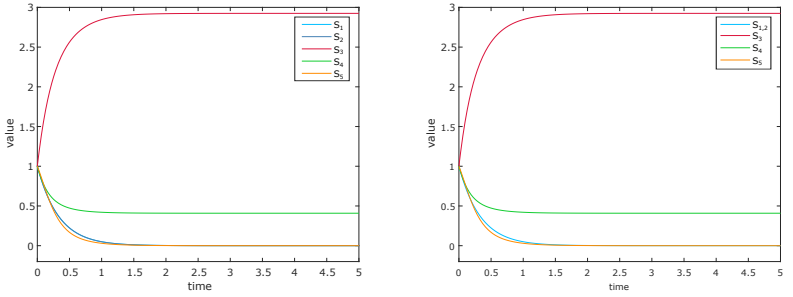
where $|\pi|_{H'} = |\{S_l \in H' | \pi_l \leq 0\}|$ counts the number of different species in H' occurring in π.

Inspecting our running example we can observe that species S_1 and S_2 are consumed at the same rate in the first, second, fourth and fifth reactions of $\mathcal{R}_E$. Furthermore, those are the only reactions where these species are involved. Hence, the species S_1 and S_2 are lumpable up to BE. Figure 3 shows $\mathcal{H}_b$, the partition that BE induces on $(\mathcal{S}_E, \mathcal{R}_E)$, and $(\mathcal{S}_b, \mathcal{R}_b)$, the BE-reduced network.

The macrovariable $S_{1,2}$ in the BE-reduced CRN $(\mathcal{S}_b, \mathcal{R}_b)$ represents the sum of original species $S_1 + S_2$. However, thanks to the BE condition, for each block we can recover the individual solutions of the original species by dividing the value of the macrovariable by the number of species in the block. For instance, we can recover the values of the concentrations of S_1 and S_2 by halving the value the concentration of $S_{1,2}$.

Figure 4 shows the solutions of the original and the BE-reduced model. Where the BE-reduced system is given by:

$$\frac{d[X_{S_{1,2}}](t)}{dt} = -3 \cdot [X_{S_{1,2}}](t)$$
$$\frac{d[X_{S_3}](t)}{dt} = 2 \cdot [X_{S_{1,2}}](t),$$
$$\frac{d[X_{S_4}](t)}{dt} = -\frac{1}{2} \cdot [X_{S_{1,2}}](t),$$
$$\frac{d[X_{S_5}](t)}{dt} = \frac{3}{2} \cdot [X_{S_{1,2}}](t) - 3 \cdot [X_{S_5}](t) \cdot \big([X_{S_3}](t) + [X_{S_4}](t)\big)$$



(a) Solutions of the original model (b) Solutions of the BE-reduced model

Figure 4: ODE simulation of the BE-reduced model from Example 1.

2.5.2 *Forward* and *Backward* differential equivalences

FDE and BDE, introduced in [39], are notions of equivalences established between IDOL variables, a syntax language that represents the ODE drifts. The aggregations provided by FDE and BDE bring to mind those of FE and BE. The equivalence established by FDE allows to write a quotient program where each equivalence class is represented using a single variable, or class representative that preserves the sum of the original trajectories. On the other hand, the equivalence established by BDE relates variables that have the same semantics whenever they start from the same conditions, i.e., variables whose drifts have the same value.

For completeness, here we restate the original definitions from [39]:

Definition 4 (Definition FDE)**.** *Let p be an IDOL program and $\mathcal{V}_p = \{x_1, \ldots, x_n\}$ the set of variables in p, where f_i denotes the drift of variable x_i. Let $\mathcal{H}$ be a partition of ODE variables, such that $H = \{x_{H,1}, \ldots, x_{H,|H|}\}$ is a block of such partition. Then, $\mathcal{H}$ is an FDE iff.*

$$\theta(p) \longrightarrow \bigwedge_{H \in \mathcal{H}} \left(\sum_{x_i \in H} f_i = \sum_{x_i \in H} f_i \left[x_j / \frac{\sum_{x_k \in H} x_k}{|H'|} : H' \in \mathcal{H}, x_j \in H' \right] \right) (\mathbf{\Phi}^{\mathcal{H}})$$

Definition 5 (Definition BDE)**.** *Let p be an IDOL program and $\mathcal{H}$ a partition of ODE variables. Then, $\mathcal{H}$ is an BDE iff.*

$$\theta(p) \longrightarrow \bigwedge_{H \in \mathcal{H}} \left(x_{H,1} = \cdots = x_{H,|H|}\right) \longrightarrow \bigwedge_{H \in \mathcal{H}} \left(f_{H,1} = \cdots = f_{H,|H|}\right) \left(\mathbf{\Psi}^{\mathcal{H}}\right)$$

When applied to ODEs encoding CRNs, FDE and BDE provide a full characterization of FE and BE (Section 2.5.1). However, the greater generality of these methods comes at the cost of a more computationally expensive implementation based on encodings in satisfiability modulo theory (SMT) formulas. Observe the following example taken from [169]. The formula $\psi^{\mathcal{H}_b}$ encodes the check whether partition $\mathcal{H}_b$ is a BDE:

$$\psi^{\mathcal{H}_b} := \left(X_{S_1} = X_{S_2}\right) \implies \left(-3 \cdot X_{S_1} = -3 \cdot X_{S_2}\right)$$

which checks that if all variables in the same block are equal (the premise) then they must evolve in the same way, i.e., their derivative should evaluate to the same value (the conclusion). The formula has two free real variables, X_{S_1} and X_{S_2}, corresponding to S_1 and S_2. By using an SMT solver, e.g., Z3 [59], we can check if $\mathcal{H}_b$ is a BDE by checking for the satisfiability of $\neg\psi^{\mathcal{H}_b}$. If there exists an assignment for X_{S_1} and X_{S_2} that makes $\neg\psi^{\mathcal{H}_b}$ true, then $\mathcal{H}_b$ is not a BDE. From the running example this is not the case, and hence it is a BDE (as expected from it being a BE).

2.5.3 Syntactic Markovian bisimulation

Syntactic Markovian bisimulation (SMB)[40] is an equivalence technique for the reduction of stochastic elementary mass-action CRNs that can be seen as an instantiation of FE to the stochastic semantics of CRNs. In the SMB-reduced network, a macrovariable tracks the sum of the populations of the distinct species in the original network that belong to the same SMB-equivalence class.

Therefore, for any given initial condition of the original network, it is possible to directly generate its lumped Markov chain from the reduced network by fixing a matching initial condition up to sums of populations. For completeness, here we restate the original definitions from [40].

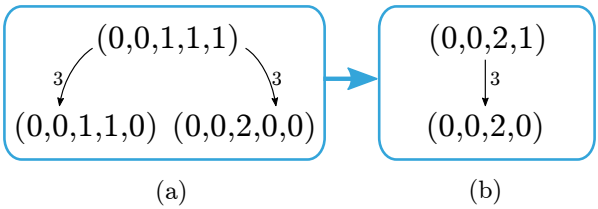


Figure 5: Example of a SMB reduction. (a) Graphical representation of the CTMC underlying $(\mathcal{S}_E, \mathcal{R}_E)$ (Example 1) starting from an initial state $\sigma_e = (0, 0, 1, 1, 1)$ (b) The lumped CTMC.

SMB computes the cumulative rate that characterizes the transformation of a certain reagent ρ into a product π by checking the syntax of the reactions using the following notion of reaction rate:

Definition 6 (SMB-Reaction rate)**.** *Let (S, R) be an reaction network, and $\rho, \pi \in \mathcal{MS}(S)$. The reaction rate from ρ to π is defined as*

$$\mathbf{rr}_{SMB}(\rho, \pi) = \sum_{(\rho \xrightarrow{\alpha} \pi) \in R} \alpha$$

For any $\mathcal{M} \subseteq \mathcal{MS}(S)$, we define $\mathbf{rr}_{SMB}[\rho, \mathcal{M}] = \sum_{\pi \in \mathcal{M}} \mathbf{rr}_{SMB}(\rho, \pi)$.

We can now define SMB

Definition 7 (SMB)**.** *Let (S, R) be an reaction network, $\mathcal{R}$ an equivalence relation over S, $\mathcal{R}^{\uparrow}$ the multiset lifting of $\mathcal{R}$ and $\mathcal{H}^{\uparrow} = \mathcal{MS}(S)/\mathcal{R}^{\uparrow}$. We say that $\mathcal{R}$ is a syntactic Markovian bisimulation (SMB) for (S, R) if and only if*

$$\mathbf{rr}_{SMB}[X + \rho, \mathcal{M}] = \mathbf{rr}_{SMB}[Y + \rho, \mathcal{M}]$$

$$\textit{for all } (X, Y) \in \mathcal{R}, \rho \in \mathcal{MS}(S), \textit{and } \mathcal{M} \in \mathcal{H}^{\uparrow}.$$

We define the syntactic Markovian bisimilarity of (S, R) as the union of all SMBs of (S, R).

Figure 5 a shows the CTMC underlying $(\mathcal{S}_E, \mathcal{R}_E)$ for an initial state $(0, 0, 1, 1, 1)$, which encodes the initial species population $S_3 + S_4 + S_5$. From this initial state we can reach states: $(0, 0, 1, 1, 0)$ and $(0, 0, 2, 0, 0)$ due to the third and sixth reactions in $\mathcal{R}_E$, respectively. Both states are reached with the same transition rate. It can be shown that the partition

of species $\mathcal{H}_s = \{S_1\}, \{S_2\}, \{S_3, S_4\}, \{S_5\}$ induces an ordinary lumpable partition of the Markov chain, which means that we can write a reduced CTMC having one state per equivalence class. The states $(0, 0, 1, 1, 0)$ and $(0, 0, 2, 0, 0)$ in Figure 5a form an ordinary lumpable partition that is represented in the lumped CTMC (Figure 5b) by the state $(0, 0, 2, 0)$.

$$
\begin{aligned}
\mathcal{H}_s &= \{\{S_1\}, \{S_2\}, \{S_3, S_4\}, \{S_5\}\} \\
\mathcal{S}_s &= \{S_1, S_2, S_{3,4}, S_5\} \\
\mathcal{R}_s &= \{S_1 \xrightarrow{2} S_5, S_1 \xrightarrow{1} 2S_{3,4}, S_{3,4} + S_5 \xrightarrow{3} S_{3,4}, \\
&\qquad S_2 \xrightarrow{2} S_{3,4}, S_2 \xrightarrow{1} 2S_5\}
\end{aligned}
$$

(a) SMB-reduction

Figure 6: Coarsest SMB and SMB reduction of $(\mathcal{S}_E, \mathcal{R}_E)$ from Example 1.

Figure 6 shows the SMB partition $\mathcal{H}_s$ and reduced CRNs, for the running example. The CTMCs of $(\mathcal{S}_s, \mathcal{R}_s)$ are reductions in terms of CTMC ordinary lumpability [27] of the ones obtained from $(\mathcal{S}_E, \mathcal{R}_E)$. Therefore, similarly to FE, SMB produces a coarse-grained version of the original CRN which allows to reason in terms of sums of variables [169]. We remark that the partition $\mathcal{H}_s$ is a refinement of $\mathcal{H}_f$. Indeed, it has been shown that SMB implies FE, but not vice versa [40]. This will be confirmed in Section 4.2.

Chapter 3

Abstraction of stochastic behaviour by species aggregation

In this chapter we discuss *Species Equivalence* (SE), a reduction technique that can be seen as the stochastic analogue of an aggregation method developed for mass-action networks in [38]. First, we define the notions behind SE, followed by a technical comparison with SMB (discussed in Section 2.5.3). We conclude by showing the applicability of SE to the reduction of biological networks and epidemic processes on complex networks.

3.1 Abstraction of stochastic behavior

As discussed in Section 2.5, ordinary lumpability preserves stochastic equivalence in the sense that the probability of each block/macro-state is equal to the sum of the probabilities in each original state belonging to that block [27, 115]. However, to verify the conditions for ordinary lumpability requires the full enumeration of the CTMC state space, which grows combinatorially with the multiplicities of initial state and the number of reactions [27, 115].

SE detects ordinary lumpability directly at the finitary level of the reaction network by identifying an equivalence relation (i.e., a partition) of the species which induces an ordinary lumpable partition over the multisets representing the CTMC states. Similar to [40], we consider a natural lifting of a partition $\mathcal{H}$ of species to multisets of species $\mathcal{H}^{\uparrow}$ through the notion of *multiset lifting*. This allows to relate multisets that have the same cumulative multiplicity form each partition block. That is, two multisets/states σ_1, σ_2 belong to the same block $B \in \mathcal{H}$ if it holds that $\sum_{S \in H} \sigma_1(S) = \sum_{S \in H} \sigma_2(S)$ for all blocks of species $H \in \mathcal{H}$.

At the basis of SE is the notion of the reaction rate $\mathbf{rr}(\rho, \pi)$ from reagents ρ to products π, as the analogue to the entries of the CTMC generator matrix:

$$\mathbf{rr}(\rho,\pi) = \begin{cases} \sum\limits_{\substack{\rho_k \xrightarrow{\alpha_k} \pi_k, \\ \rho=\rho_k, \pi=\pi_k}} \alpha_k & \text{if } \rho \neq \pi, \\ -\sum\limits_{\pi_k \in \boldsymbol{\pi}, \rho \neq \pi_k} \mathbf{rr}(\rho, \pi_k) & \text{if } \rho = \pi. \end{cases}$$

An SE-partition $\mathcal{H}$ splits the network by relating species that have equal sums of values across all members within each SE-equivalence class. Thus, two multisets of species σ, σ' are equal up to the partition $\mathcal{H}$ *iff*

$$\sum_{S_i \in H} \sigma_i = \sum_{S_i \in H} \sigma'_i, \qquad \text{for all blocks of species } H \in \mathcal{H}.$$

Let us denote $\boldsymbol{\pi}/\mathcal{H}$ as the partition induced on the set of products $\boldsymbol{\pi}$ of the network up to $\mathcal{H}$. Then, we say that the partition of species $\mathcal{H}$ is an SE if for any two species S_i and S_j in a block of $\mathcal{H}$, and for any block of products $\mathcal{M} \in \boldsymbol{\pi}/\mathcal{H}$ we get

$$\sum_{\pi \in \mathcal{M}} \mathbf{rr}(S_i + \rho, \pi) = \sum_{\pi \in \mathcal{M}} \mathbf{rr}(S_j + \rho, \pi) \tag{3.1}$$

for all ρ such that $S_i + \rho$ or $S_j + \rho$ are in the set of reagents $\boldsymbol{\rho}$.

3.1.1 Computation of the SE-reduced network

The partition that SE induces on the products carries over to the states of the underlying Markov chain, thereby yielding an ordinarily lumpable

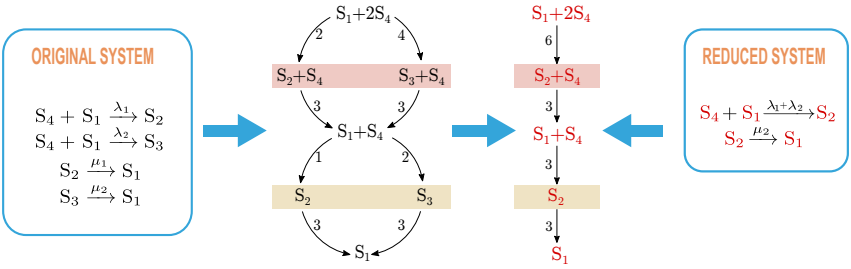


Figure 7: Illustration of SE on a simple network with four species. From left to right: original system and its corresponding Markov chain generated from an initial population S_1+2S_4, reduced Markov chain and reduced system. The colored boxes represent lumpable blocks in the ordinary lumpable partition and the red labels indicate variables in the reduced systems.

partition that aggregates all states that are equal up to the SE. Given a partition $\mathcal{H}$, it is possible to construct a reduced network using an algorithm similarly to [36], where it was developed for network reductions with deterministic rate interpretation.

Each representative in the reduced network can be interpreted as a macro-species that tracks the sum of the populations of the distinct species in the original network that belong to the same SE equivalence class. Therefore, for any given initial condition $\hat{\sigma}$ of the original network, it is possible to directly generate its lumped Markov chain from the reduced network by fixing a matching initial condition up to sums of populations.

Figure 7 shows an example of SE reduction. The original system is a CRN with species S_1, S_2, S_3, S_4. Next, we observe the underlying Markov chain derived from the initial state $\hat{\sigma}_0 = S_1 + 2S_4$. The Markov chain is shown in a customary graph representation where each node is a state and each arc is labeled with the transition rate according to mass-action kinectics, i.e., $q(\sigma, \sigma')$ is the transition rate from state σ to state σ'. The colored boxes represent the two lumpable blocks in an ordinary lumpable partition of the Markov chain (here it suffices to check that the outgoing transitions are equal for states in blocks of size two).

The partition of species $\mathcal{H} = \{\{S_1\}, \{S_2, S_3\}, \{S_4\}\}$ can be shown to be an SE, hence, states that are equal up to the sum of the second and third coordinate form a lumpable partition block. This SE gives rise to a reduced network by choosing the representatives S_1, S_2 and S_4 for each block (colored in red in the figure to distinguish them from original species names). The reduced network has fewer reactions due to the fact that reactions in the original network are merged into a single after renaming. Similar to the equivalence techniques described in the background, the maximal SE can be calculated using a partition refinement algorithm that refines the initial partition, i.e., a singleton block containing all species. Following the blue arrow we observe the underlying Markov chain of the reduced network derived from the matching initial state $\hat{\sigma}_0 = S_1 + 2S_4$. The Markov chain of the reduced network essentially corresponds to the lumped Markov chain of the original network (as indicated by the matching colors of the nodes).

3.2 Technical Results

Now we can formally define Species Equivalence[1].

Definition 8 (Multiset lifting)**.** *Let (S, R) be an reaction network, $\mathcal{R} \subseteq S \times S$ be an equivalence relation over S, and $\mathcal{H}$ be the partition induced by $\mathcal{R}$ over S. The* multiset lifting *of $\mathcal{R}$ on $\mathcal{MS}(S)$, denoted by $\mathcal{R}^{\uparrow} \subseteq \mathcal{MS}(S) \times \mathcal{MS}(S)$, is:*

$$\mathcal{R}^{\uparrow} \triangleq \{(\sigma_1, \sigma_2) \mid \ \sigma_1, \sigma_2 \in \mathcal{MS}(S) \ \wedge \ \forall H \in \mathcal{H} : \sigma_1(H) = \sigma_2(H)\}$$

The multiset lifting $\mathcal{R}^{\uparrow}$ of $\mathcal{R}$ can be readily seen to be an equivalence relation over $\mathcal{MS}(S)$. Let $\mathcal{H}^{\uparrow} = \mathcal{MS}(S)/\mathcal{R}^{\uparrow}$. For any equivalence class of multisets $\mathcal{M} \in \mathcal{MS}(S)/\mathcal{R}^{\uparrow}$ we have

$$\rho(H) = \rho'(H) \quad \textit{for any } \rho, \rho' \in \mathcal{M} \textit{ and any } H \in \mathcal{H}.$$

Therefore we might use $\mathcal{M}(H)$ to denote $\sigma(H)$, with σ being any element of $\mathcal{M}$.

[1] The mathematical framework of this chapter is as in [32]. For matters of completeness, we also included the corresponding proofs, which are the result of the work of my coauthors A. Vandin, T. Waizmann and M. Tschaikowski.

Definition 9 (Species Equivalence)**.** *Then, we say that $\mathcal{R}$ is a species equivalence (SE) for (S, R) if and only if:*

$$\mathbf{rr}[X+\rho, \mathcal{M}] = \mathbf{rr}[Y+\rho, \mathcal{M}], \quad \textit{for all } (X, Y) \in \mathcal{R}, \rho \in \mathcal{MS}(S), \textit{and } \mathcal{M} \in \mathcal{H}^{\uparrow}.$$

To show that SE is a sufficient condition for ordinary lumpability, the first step is to express $q[\sigma, \tilde{M}]$, the cumulative transition rate from a CTMC state σ to states belonging to a block $\tilde{M}$ of the multiset lifting of an SE, in terms of the reagents and products of the reactions that generate those transitions. Given the source and target blocks M and $\tilde{M}$, respectively, and for a given block of multisets $\bar{M}$, we define the set:

$$\bar{M}_{M \to \tilde{M}} = \{\pi \in \mathcal{MS}(S) | \exists \sigma \in M, \rho \in \bar{M} \text{ s.t } \rho \subset \sigma \text{ and } (\sigma - \rho + \pi) \in \tilde{M}\} \quad (3.2)$$

which collects all products π of reactions which can be executed in a state $\sigma \in M$, such that the reagents belong to $\bar{M}$ and the target state is in block $\tilde{M}$. Importantly, it can be shown that $\bar{M}_{M \longrightarrow \bar{M}}$ is a block of the multiset lifting. Then, for any two distinct blocks of the multiset lifting, $M, \tilde{M} \in \mathcal{H}^{\uparrow}$ and for any $\sigma \in M$ it holds that:

$$q[\sigma, \tilde{M}] = \sum_{\bar{M} \in \mathcal{H}^{\uparrow}} \sum_{\substack{\rho \in \bar{M} \\ \rho \subset \sigma}} \prod_{S \in \rho} \binom{\sigma(S)}{\sigma(S)} \cdot \mathbf{rr}[\rho, \bar{M}_{M \to \bar{M}}] \quad (3.3)$$

thus, expressing the aggregate rate in terms of the source state σ and quantities depending on the multiset lifting. Furthermore, we can express $q[\sigma, \tilde{M}]$ in a way that does not depend on the source state σ, but only on the block of the multiset lifting to which it belongs. For any $M, \tilde{M} \in \mathcal{H}^{\uparrow}$, for any $\sigma \in M$, it holds that:

$$\sum_{\rho \in \tilde{M}} \prod_{S \in \rho} \binom{\sigma(S)}{\rho(S)} = \prod_{H \in \mathcal{H}} \binom{M(H)}{\tilde{M}(H)} \quad (3.4)$$

To formalize this we need to show that for any two blocks of the multiset lifting $M, \tilde{M} \in \mathcal{H}^{\uparrow}$ and for any two states $\sigma, \sigma_0 \in M$, we have:

$$q[\sigma, \tilde{M}] = q[\sigma', \tilde{M}]$$

Indeed, let us assume that $M \neq \tilde{M}$. Due to the properties of SE, we can factor out the reaction-rate quantities in Equation. 3.3. Using $\rho^{\bar{M}}$ to denote any element of $\bar{M}$, we obtain

$$q[\sigma, \tilde{M}] = \sum_{\bar{M} \in \mathcal{H}^{\uparrow}} \mathbf{rr}[\rho^{\bar{M}}, \bar{M}_{M \to \bar{M}}] \sum_{\substack{\rho \in \bar{M} \\ \rho \subset \sigma}} \prod_{S \in \rho} \binom{\sigma(S)}{\rho(S)} \tag{3.5}$$

$$= \sum_{\bar{M} \in \mathcal{H}^{\uparrow}} \mathbf{rr}[\rho^{\bar{M}}, \bar{M}_{M \to \bar{M}}] \sum_{\rho \in \bar{M}} \prod_{S \in \rho} \binom{\sigma(S)}{\rho(S)}$$

$$= \sum_{\bar{M} \in \mathcal{H}^{\uparrow}} \mathbf{rr}[\rho^{\bar{M}}, \bar{M}_{M \to \bar{M}}] \prod_{H \in \mathcal{H}^{\uparrow}} \binom{M(S)}{\bar{B}(S)} \tag{3.6}$$

where the last equality follows by Equation 3.4. This closes the case, because the terms appearing on the right-hand side of Equation 3.5 do not depend on σ or σ_0. Since the case $M = \tilde{M}$ follows from the case $M \neq \tilde{M}$, see [[205], Proposition 1], we infer the sufficiency of SE.

3.2.1 Comparison with SMB

Syntactic Markovian Bisimulation (Section 2.5.3) is an earlier variant of SE . It is easy to see that SMB is strictly finer than SE; that is, an SE is also an SMB but the converse is not true in general. We now provide a simple reaction network which shows that SMB is stricter than SE. The reaction network has species $\{A, B\}$ and only one reaction: $\{A \xrightarrow{1} B\}$. We have that the partition $\{\{A, B\}\}$ is an SE, but it is not an SMB. Indeed, we have only one class of multi-set equivalent products, consisting of $\{A\}$ and $\{B\}$, with

Example 2 (SMB stricter than SE)**.**

$$\mathbf{rr}(A, \{\{A\}, \{B\}\}) = 0 \quad \textit{and} \quad \mathbf{rr}(B, \{\{A\}, \{B\}\}) = 0$$
$$\mathbf{rr}_{SMB}(A, \{\{A\}, \{B\}\}) = 1 \quad \textit{and} \quad \mathbf{rr}_{SMB}(B, \{\{A\}, \{B\}\}) = 0$$

Table 3: Comparison between SMB and SE.

		Number of species			*Number of reactions*		
Model	*Ref.*	*Orig.*	*SMB*	*SE*	*Orig.*	*SMB*	*SE*
e9	[189]	262 146	222	222	3 538 944	990	990
e8	[189]	65 538	167	167	786 432	720	720
e7	[189]	16 386	122	122	17 032	504	504
e6	[189]	4 098	86	86	368 664	336	336
e5	[189]	1 026	58	58	7 680	210	210
NIHMS80246-S1	[19]	708	555	555	7 432	5 416	5 416
NIHMS80246-S4	[19]	213	66	66	2 230	432	432
NIHMS80246-S6	[19]	24	2	2	88	3	3
machine	[193]	14 531	10 855	10 855	194 054	142 165	142 165
pcbi.1000364.s006	[12]	2 562	1 907	1 907	41 233	33 141	33 141
pcbi.1000364.s005	[12]	471	348	348	5 033	4 083	4 083
Tyrosine phosphorylation	[48, 49]	730	217	217	5 832	1 296	1 296
Nag2009	[157]	920	364	364	12 740	4 020	4 020
Tumor supressor protein	[11]	796	503	503	5 797	4 210	4 210
mmc1	[68]	348	72	72	284	142	142
fceri_ji	[74, 189]	354	105	105	3 680	732	732
sx_fceri_ji_sitel	[10]	348	215	215	3 447	1 782	1 782
fceri_gamma2_asym	[74, 189]	10 734	3 744	**351**	187 468	5 708	**3 132**
sx_fceri_ji_sitelll	[74, 189]	2 506	1 281	**154**	32 776	16 841	**1 140**
sx_insulin	[123]	2 768	2 719	2 719	38 320	37 760	37 760

In Table 3[2] we compare SMB against SE in a number of models of models from the literature that were previously analyzed in [38, 40]. For most of the models, including those in Section 3.5 , the maximal SMB and SE aggregations coincide. However, SE yields coarser aggregations for two variants of the FcϵRI signaling pathway model presented in [74, 189].

FcεRI is a tetrameric receptor complex that has high affinity for the Fc region of immunoglobulin (IgE), an antibody involved in defense response to bacterium and inmune response. FcεRI signaling is a process

[2]These aggregations where obtained using ERODE in its version win32.x86_64_1.0.0.201810151158.

connected to cell surface receptor signaling and immune response regulation. The system starts with the binding of IgE to the FcεRI receptor on the surface of a signal-receiving cell. This initiates a series of ordered phosphorylation events within the receptor microenviroment. The β and γ-chains mediate signaling in the intracellular environment through their immunoreceptor tyrosine-based activation motifs (ITAMs). Exhibiting a cooperative behavior, the ITAMs motifs recruit molecules with a SrC-homology-2 (SH_2) domain.

The order of phosphorylation events in signaling pathways can be either sequential or random [66, 86]. The hierarchical organization of phosphorylation reactions is common in molecules with multiple phosphorylation residues, e.g., in [64] the state of one site in GPCR dictates whether or not other phosphosites become efficiently phosphorylated. In signal-transduction, for instance, protein binding events oftentimes require one of the interacting sites to be phosphorylated in advance. This imposes an ordering on the phosphorylation and binding events in the model [19]. In the FcεRI signaling pathway, the phosphorylation of the ITAM sites of β and γ conditions the binding of the SH2 domains of Lyn and Syk.

The models **fceri_gamma_asym** and **fceri_fyn_lig** are prototypical models depicting *early* events in FcεRI signaling [189] consisting of 10 734 species and 187 468 reactions, and 2 506 species and 32 776 reactions, respectively. The general mechanism in both models consists in a FcεRI receptor with a ligand-binding site α and two protein docking sites β and γ that are used to recruit the Src kinases Lyn and Syk, respectively. The crosslinking events in FcεRI lead to tyrosine phosphorylation of the β by an active Lyn molecule. The phosphorylated FcεRI receptor binds Syk molecules using the γ site. Moreover, receptor-bound Lyn can transphosphorylate an adjacent receptor in a dimer on both phosphorylation sites, which causes the recruit of Syk and the activation loop of Syk proteins bound to the same dimer.

In **fceri_gamma_asym** the site γ is modeled as dimeric unit with sites labeled $\gamma 1$ and $\gamma 2$. Figure 8 (A) shows the four molecular components of the model: a bivalent ligand, two proteins: Lyn and Syr, and the FcεRI

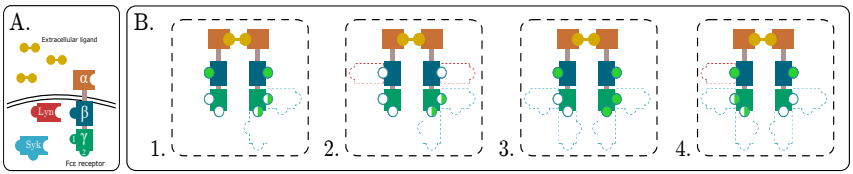


Figure 8: (A) Components in the fceri_gamma_asym model, a model for the early events in FcεRI signaling induced by a bivalent Ligand [19] where the FcεRI receptor has a dimeric γ site. Binding sites β and $\gamma 1, \gamma 2$ can be either phosphorylated (green circles) or unphosphorylated (white circles); green-white circles represent either phosphorylation state. (B) Examples of macrovariables obtained in the reduced model.

receptor. The result of SE evaluation is a reduced model consisting of 351 species and 2532 reactions. In a closer inspection of the aggregated system, we observe that states of the dimerized receptor are aggregated in equivalence classes, e.g., Figure 8(B1-4). For instance, an equivalence between receptor dimmers with active β site, where at least one γ site has bound a Syr, causing the phosphorylation of the bound γ; results in the collapse of eight receptor dimmer units into the macrovariabe shown in Figure 8 (B1). This equivalence relation extends to receptor dimmers which are bound to one or more Lyn units (Figure 8(B1) and (B3), respectively), and to receptor dimmers whose $\gamma 1$ and $\gamma 2$ sites can be bound to Syr (Figure 8(B4)).

The model **fceri_fyn_lig** includes the phosphorylation of one of the sites of the Lyn by another Lyn unit in a transphosphorylation process occuring in the receptor dimmer. This model also includes a second Src protein called Fyn that can bind to the β site of the FcεRI receptor, and whose phosphorylation occurs in the receptor dimmer by Lyn transphosphorylation. Additionally, the bivalent ligand is modeled with two states, which increases the overall combinatorial complexity. The SE-reduced model for **fceri_fyn_lig** consists of 154 species and 900 reactions. The SE-aggregations for the FcεRI states in this model are consistent with those discussed above. In addition, in both **fceri_gam-ma_asym** and **fceri_fyn_lig**, the SE-aggregations also relate four states of the free Syk protein, regardless the phosphorylation state of its binding sites.

3.2.2 Comparison with Forward Equivalence.

Although SE and FE work on different semantics, the fact that they are both equivalences over species that induce analogous aggregations at the semantic level calls for the question of establishing a formal relation between the two equivalences.

Theorem 1. *Let (S, R) be an reaction network and $\mathcal{R}$ an equivalence relation over S. Then, if $\mathcal{R}$ is an SE for (S, R), it also is an FE for (S, R'), with*

$$R' = \{(\rho \xrightarrow{\alpha'} \pi) \mid (\rho \xrightarrow{\alpha} \pi) \in R, \alpha' = \frac{\alpha}{\prod\limits_{Y \in S} (\rho(Y)!)}\} \tag{3.7}$$

***Proof of Theorem* 3.2.2** Let $\mathcal{H}$ be the partition induced by $\mathcal{R}$ over S, and $\mathcal{H}^{\uparrow}$ the partition induced by $\mathcal{R}^{\uparrow}$ over $\mathcal{MS}(S)$. In order to close the proof we show that the condition of SE in (S, R) implies the condition of FE in (S, R').

Note that the factor $\frac{1}{[\rho]!}$ in $\mathbf{fr}$ has the effect of scaling the rates of reaction in $(\rho \xrightarrow{\alpha'} \pi)$ in R' back to its original rate in R, divided by $|\rho|!$. It is hence easy to see that $\mathbf{fr}_{(S,R')}(\rho, H) = \mathbf{fr}'_{(S,R)}(\rho, H)$, with

$$\mathbf{fr}'(\rho, H) := \frac{1}{|\rho|!} \cdot \sum_{Y \in H} \sum_{(\rho \xrightarrow{\alpha} \pi) \in R} (\pi(Y) - \rho(Y)) \cdot \alpha$$

In the rest of the proof we will hence use $\mathbf{fr}'_{(S,R)}$ rather than $\mathbf{fr}_{(S,R')}$. Since all $\mathbf{rr}$ and $\mathbf{fr}'$ are defined over (S, R), we avoid to write the reaction network as prefix.

What we want to show is that $\mathbf{rr}[X + \rho, \mathcal{M}] = \mathbf{rr}[Y + \rho, \mathcal{M}]\ \forall \mathcal{M} \in \mathcal{H}^{\uparrow}$, implies $\mathbf{fr}'[X + \rho, H] = \mathbf{fr}'[Y + \rho, H]\ \forall H \in \mathcal{H}$. All multi-sets in a given $\mathcal{M} \in \mathcal{H}^{\uparrow}$ contain same number of species of each block $H \in \mathcal{H}$. Thus, with a slight abuse of notation we can write $\mathcal{M}(H)$ to denote $\sum_{Z \in H} \pi^{\mathcal{M}}(Z)$, with $\pi^{\mathcal{M}}$ any multi-set in $\mathcal{M}$. Let us assume that $X + \rho$ belongs to an equivalence class $\mathcal{M}' \in \mathcal{H}^{\uparrow}$. We remark that this implies that $Y + \rho$ belongs to $\mathcal{M}'$ as well. Then, a reaction with reagents $X + \rho$, rate α and product any $\pi \in \mathcal{M}$, produces comulatively species in H with rate $(\mathcal{M}(H) - \mathcal{M}'(H)) \cdot \alpha$. Finally, for each $H \in \mathcal{H}$ we easily obtain

$$\begin{aligned}
\mathbf{fr}'[X+\rho, H] &= \frac{1}{|X+\rho|!} \cdot \sum_{\mathcal{M} \in \mathcal{H}^{\uparrow}} (\mathcal{M}(H) - \mathcal{M}'(H)) \cdot \left(\sum_{\pi \in \mathcal{M}} \sum_{(X+\rho \xrightarrow{\alpha} \pi) \in R} \alpha \right) \\
&= \frac{1}{|X+\rho|!} \cdot \sum_{\mathcal{M} \in \mathcal{H}^{\uparrow} s.t. \mathcal{M}' \neq \mathcal{M}} (\mathcal{M}(H) - \mathcal{M}'(H)) \cdot \left(\sum_{\pi \in \mathcal{M}} \sum_{(X+\rho \xrightarrow{\alpha} \pi) \in R} \alpha \right) \\
&= \frac{1}{|X+\rho|!} \cdot \sum_{\mathcal{M} \in \mathcal{H}^{\uparrow} s.t. \mathcal{M}' \neq \mathcal{M}} (\mathcal{M}(H) - \mathcal{M}'(H)) \cdot \mathbf{rr}[X+\rho, \mathcal{M}]
\end{aligned}$$

This follows similarly for Y, where we have as well $(Y + \rho) \in \mathcal{M}'$. Furthermore, we have $|X + \rho| = |Y + \rho|$. This allows us to conclude $\mathbf{fr}'[X + \rho, H] = \mathbf{fr}'[Y + \rho, H]$, closing the proof.

Remarks An important remark to be made regarding this result is that it requires to scale the rates of reactions in Equation (3.7). This is due to an inherent, well-known inconsistency existing between the CTMC and ODE semantics of chemical reaction networks. While, as discussed in Section 2, rates are scaled by binomial coefficients in the CTMC semantics in order to capture the combinatorial nature of the discrete molecular interactions, in the ODE semantics does not make such difference, e.g., [76]. We refer to [30] for a more in depth discussion on this.

It is interesting to note that a *different* ODE semantics would be possible, grounded on a limit result by Kurtz which establishes the ODE solution as the asymptotic behavior of a sequence of infinitely large CTMCs induced by the same chemical reaction network with increasing volumes of a solution having given initial concentrations of species. This interpretation would lead to same scaling as those considered in Theoreom 3.2.2 also in the ODE case [126]. We leave it for future work to understand if, by appropriately adapting FE to this different ODE semantics, Theorem 3.2.2 can be stated so as to relate SE and FE *on the same reaction network*.

3.3 Applications to Systems Biology

Next, we discuss a series of case studies to highlight the physical interpretability of SE aggregations.

3.3.1 Species equivalence in multi-site phosphorylation processes

Mechanistic models of signaling pathways are prone to a rapid growth in the number of species and reactions due to the distinct configurations in which a molecular complex can be found [180]. A prototypical situation is multisite phosphorylation, a fundamental process in eukaryotic cells that is responsible for various mechanisms such as the regulation of switch-like behavior [93, 198].

Let us take for example scaffold proteins; the multiple sites allow the scaffold to bind several signaling proteins at the same time. This is useful for instance when it is necessary to transport proteins to a specific binding partner avoiding unnecessary interactions. In other cases the scaffold recruits several kinases to form a larger multi-kinase complex capable to interact with specific targets. An example of this is the scaffold kinase suppressor of Ras 1 (KSR 1) which can bind all the three kinases of the MAPK pathway [155]. In addition, Ras 1 promotes MEK's phosphorylation by transporting the recruited MEK molecules close Ras in the cellular membrane [151]. The multisite characteristic is also present in several kinases involved in signaling pathways, e.g., the double phosphorylation of MAPK discussed in case study 4.3.1, where MAPK phosphorylation and dephosphorylation is mediated by kinases and phosphatases in each specific residue.

Figure 9 shows a prototypical model of multisite phosphoryation in protein with n docking sites. Each site behaves independently from the others and exhibits the same affinity with kinase K. In the example, K binds the protein in site i which is unphosphorylated (i_u). Upon binding, the kinase transfers a phosphate group to the protein causing phosphorylation (i_p).

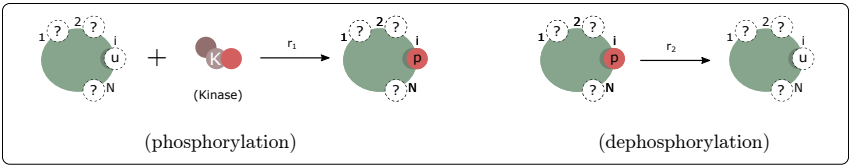


Figure 9: Phosphorylation/dephosphorylation processes in a N-site protein [180]. The state of each binding site can be either phosphorylated (p), unphosphorylated (u) or unknow (?). Phosphorylation occurs according to a random mechanism where the binding of kinase (K) to site i is described through kinetic parameter r_1. Dephosphorylation is modeled as a spontaneous process where kinetic parameter r_2 gives the speed in change of i

The dephosphorylation of the protein occurs as a result of its own enzymatic activity, i.e., the protein looses the phosphate group and the site i_p returns to its original unphosphorylated state. This mechanism is common in signal transduction systems, when self-catalysis of the receiver domain terminates the signal process, e.g., [91].

To keep track of the status of each individual site of the protein in the example, we require 2^n distinct molecular species in a formal network [180]. Let us call our protein A, then each species representing a state of A can be written in the form $A(b_1, ..., b_n)$ where a site b_j can be either phosphorylated ($b_j = 1$) or not ($b_j = 0$). Then, for all $j = 1, ..., n$ and for any combination of the site states $s_1, ..., s_{j-1}$, $s_{j+1}, ..., s_n$, the resulting mass-action network is given by:

$$A(s_1, ..., s_{j-1}, 0, s_{j+1}, ..., s_n) + K \xrightarrow{r_1} A(s_1, ..., s_{j-1}, 1, s_{j+1}, ..., s_n) + K$$
$$A(s_1, ..., s_{j-1}, 1, s_{j+1}, ..., s_n) \xrightarrow{r_2} A(s_1, ..., s_{j-1}, 0, s_{j+1}, ..., s_n)$$

To simplify this mathematical model, it is often assumed that the kinetic parameters are equal at all phosphorylation sites [189]. Under this setting, SE can provide a stochastically equivalent reduction using N equivalence classes, each aggregating the behavior of all distinct protein configurations with the same number of phosphorylated sites. More formally, if we consider a block of species H_i that groups all configurations

that have exactly i phosphorylated sites, $H_i = \{A(s_1, ..., s_n) | s_1 + ... + s_n = i\}$ for $i = 0, ..., n$, then the maximal SE is given by the partition $\{\{K\}, H_0, ..., H_n\}$.

Occasionally, protein dephosphorylation occurs by the influence of an enzyme, i.e., a phosphatase. In an opposite manner to the kinase which induces phosphorylation, the phosphatase binds the protein on its phosphorylated site to sequestrate the phosphate group. This results in a reaction where the site bound to the phosphatase becomes dephosphorylated, e.g., the double dephosphorylation of MAPK by MKP (Sec. 4.3.1). Considering the prototypical CRN above, and a phosphatase F, the equation describing A's dephosphorylation becomes:

$$A(s_1, ..., s_{j-1}, 1, s_{j+1}, ..., s_n) + F \xrightarrow{r_2} A(s_1, ..., s_{j-1}, 0, s_{j+1}, ..., s_n) + F$$

We tested SE using a model describing the enzyme-mediated phosphorylation and dephosphorylation of a 2-site protein [189]. The model consists of 18 species and 24 reactions describing the interactions of a substrate $A(b1_{\{u,p\}}, b2_{\{u,p\}})$, an enzyme $K(k)$ and a phosphatase $F(f)$, where $k, f = \{1, 2\}, k \neq f$ indicate the site to which K or F are bound. Here, SE produces equivalence classes that aggregate components with symmetric site configuration, e.g., the equivalence class $\{A(b1_u, b2_p)$, $A(b1_p, b2_u)\}$ equates the two forms of the partially phosphorylated substrate. This equivalence extends to all the states of the substrate, regardless if they are bound with K, F, or in a complex where the substrate is bound to both E and F at the same time. This yields a SE-reduced model given by 12 species and 24 reactions.

3.3.2 Identification of equivalent molecular complexes in CaMKII mechanism

The assumption of equal kinetic parameters is not necessary to achieve aggregation with SE. We show this on *MODEL100115100*, a kinetic model of the interactions between calcium (Ca^{2+}), calmodulin (CaM), and the calcium-CaM dependent protein kinase II (CaMKII) [167] taken from BioModels Database. CaMKII is a serine/threonine serine kinase that is regulated by the Ca^{2+}|CaM complex and is considered to play a fundamental role in the mechanism of synaptic plasticity [143].

The structure of a molecule of CaMKII consists of: a variable segment, a catalytic, an anti-inhibitory and a self-association domains. The variable and self-association domains determine CaMKII isophorms and the sensitivity to react with calcium and calmodulin; for example, CaMKIIβ shows higher affinity for Ca^{2+}|CaM than CAMKIIα [25, 107]. After binding the Ca^{2+}|CaM complex, the calcium-calmodulin complex CaMKII becomes active. Then, as the amounts of Ca^{2+}|CaM increase, the neighboring sub-units of CaMKII activate via autophosphorylation which could lead to a sustained calcium-dependend persistent activation.

In the model under consideration, Ca^{2+} can bind to two pairs of domains located at the amino (N) and carboxyl (C) termini of CaM, thus from now own we will refer to the complex Ca^{2+}|CaM as a single unit CaM_{nNcC} where n, c=(0,1,2) indicate the amount of Ca^{2+} ions bound to N and C respectively. Figure 10 (A) represents the kinetic interactions of CaMKII and CaM_{nNcC} (CaM). From left to right, the monomeric units of CaM bind to CaMKII and form a complex referred to as KCaM. Then, KCaM experiences a reversible dimerization that may lead to autophosphorylation. The active units of KCaM (KCaM*) can interact with any unphosphorylated KCaM unit (indicated in the figure by the "?" sign) and form a phosphorylated complex. The $KCaM^*$ complex decomposes either in a reverse reaction releasing the two original units, or can lead to autophosphorylation of the associated unphosphorylated KCaM unit. This dynamics is represented in a reaction network consisting of 156 species and 480 reactions.

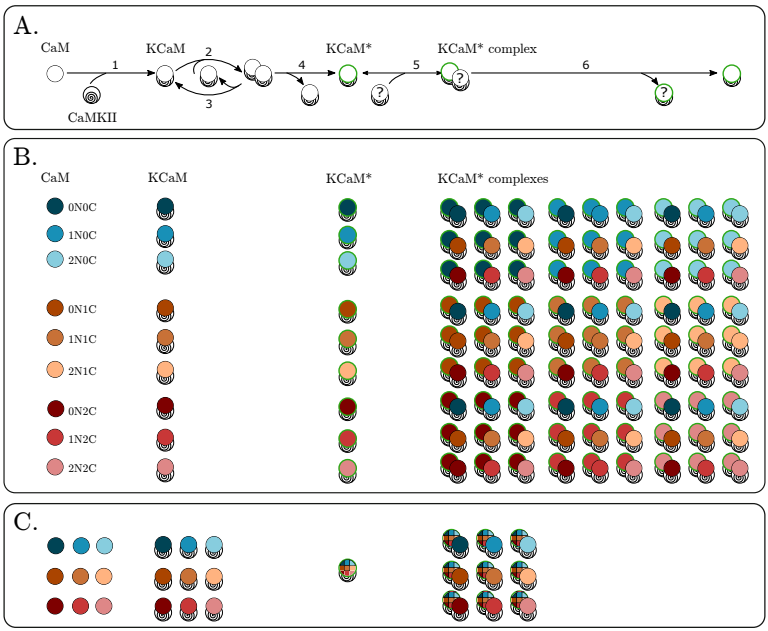


Figure 10: (A). Graphical representation of the dynamics in the kinetic model of Ca^{2+}, CaM and CaMKII adapted from [167]. (B) Units of CaM and CaMKII in the original model; calcium-bound CaM is represented by coulored circles with label $nNcC$, where n and c corresponds to the number of Ca^{2+} ions associated to each termini. (C) Units in the SE reduced model.

The maximal SE finds that all phosphorylated monomers are equivalent (Figure 10 B), although their dynamics are characterized by distinct kinetic parameters to account for phosphorylation rates that depend on the number of bound Ca^{2+}[184]. Furthermore, such equivalences carry over to all complexes where they are involved. This leads to equivalence classes consisting of nine molecular species each, with an overall reduction from 156 species and 480 reactions to 76 species and 264 reactions. Notably, important quantities to observe in this model are the amounts of free and bound CaM[142], both recoverable from the reduced network.

3.3.3 Internalization of the GTPase cycle in a SPOC model

In the examples above, SE can be physically interpreted as a reduction that preserves both the structure of equivalent molecular species as well as their functions. In our next case study we show that SE can also aggregate species that exhibit contrasting functionality. We consider the mechanistic model from [44] of spindle position checkpoint (SPOC).

SPOC intervenes in the process of cell division by verifying the correct alignment of the nucleus between mother and daughter cells [135], an important requisite in the cell cycle. A correct alignment of the mitotic spindle determines the fate of the two daughther cells in the process of cell division. During mitosis, the mother cell develops a mechanism called spindle apparatus whose function is to ensure the proper alignment of the genetic material, so that upon division two identical daughter cells are created. The mitotic spindle is an elipsoide structure composed of proteins and microtubules, all structural components whose function is to bind on specific sites the mother's chromosomes.

A Bfa 1 |Bub 2 GTPase complex in the SPOC inhibits the mitotic exit network (MEN) by regulating its upper stream component: the Ras-like GTPase Tem 1. Under correct alignment the GAP complex is inhibited by kinase Cdc 5 who phosphorylates Bfa 1 [90]; under misalignment, the kinase Kin 4 phosphorylates Bfa 1 preventing the inhibitory phosphorylation by Cdc 5 [168]. Stabilization of the spindle poles means that the nucleus has moved to a daugher cell, which allows Tem 1 to return to its active phase and start the exit from mitosis.

Figure 11 (B) adapts the Tem 1 GTPase cycle from [44]. The most upstream event of the pathway shows Tem 1, which is regulated by the (GAP) complex Bfa 1 |Bub 2 (from now on Bfa 1). Tem 1 binds to the yeast centrosomes (spindle pole bodies, SPBs) via GAP-dependent and GAP-independent sites. The intrinsic GTPase switching cycle of Tem 1 is modeled as a reversible first-order reaction that converts Tem 1^{GTP} into Tem 1^{GDP} and vice versa. When bound directly to the SPB, Tem 1 does not interact with Bfa 1 whereas the units of Tem 1 in the cytosol interact with units of Bfa 1 located in both, cytosol and SPB. These interactions

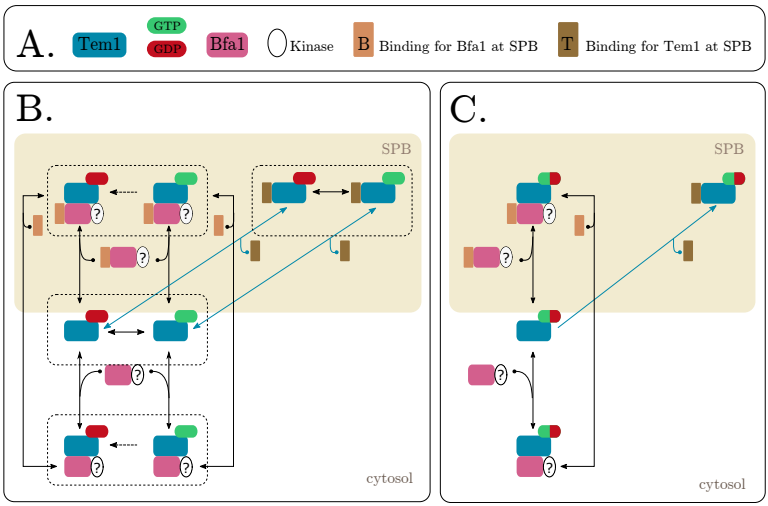


Figure 11: (A) Subunits of the model. (B) Adaptation of the SPOC dynamical model from [44]. Beige boxes indicate the SPB compartment. Reactions crossing the compartment boundary represent the reversible SPB association of the respective species or complexes. Reactions in blue mark the intrinsic Tem1 GTPase-cycle and reversible SPB association. The dashed boxes represent the SE equivalence classes. (C) SE reduced network.

occur for all instances of Bfa 1 regardless its phosphorylation state as indicated by the "?" sign. The dashed arrows indicate GTP hydrolysis by the respective Bfa 1 |Tem 1 $^{\text{GTP}}$ complexes, which is accelerated depending on Bfa 1's GAP activity.

The maximal SE collapses complexes that are equal up to the GTP- or GDP-bound state (Figure 11) (B), yielding eight equivalence classes that contain pairs of molecular species. The original network with 24 species and 71 reactions is reduced to 16 species and 36 reactions, from which one may recover observables of interest such as the total amount of active Bfa 1 [44].

3.4 Epidemic processes in networks

Models of epidemic processes are well established since the celebrated work by Mermack and McKendrick [116]. They have received considerable attention from several disciplines including biology, computer science, mathematics, physics, and sociology due to the generality with which, in addition to the diffusion of pathogens, the epidemic analogy can be applied to a variety of phenomena such as the spreading of rumor [146], opinion [42], as well as computer viruses [212]. The availability of large datasets in a range of socio-technical systems has prompted the study of epidemic processes on complex networks that consider the heterogeneity of real-world processes, which is neglected in simpler variants that assume a well-mixed, uniform environment [166].

Aggregation of epidemic processes on networks has been studied by Simon *et al.* [185], relating symmetries in the graph with lumping on the Markov chain. Symmetry is formalized in terms of nodes belonging to the same orbit, thereby satisfying the property that there exists a graph automorphism relating them. Then, the orbit partition, i.e., the partition of nodes where each block is a distinct orbit, induce a Markov chain lumping that tracks the number of nodes in each block of the orbit partition that are in any given state [185].

Here we show that SE can be seen as a complementary, exact aggregation method for epidemic processes on complex networks. As an example, we study the well-known susceptible-infected-susceptible (SIS) model, where each node in the network in the susceptible state can be infected with a rate proportional to the number of infected neighbors, and recover from infection according to an independent Poissonian process. Let $A = (a_{ij})$, with $A \in \mathbb{R}^{N \times N}$, define the adjacency matrix of a graph with N nodes representing the network topology, with $a_{ij} > 0$ denoting the presence of a possibly weighted edge between node i and j.

The network experiments where conducted by my co-author M. Tribastone. The set-up of the experiment, together with the corresponding analysis of the results is detailed here since it contributes to the understanding of the general applicability of SE.

Table 4: Coarse graining of SIS models on real-world networks.

	Original size		*Reduced size*		
Network	N	E	N	E	*Orbits*
tntp-ChicagoRegional [69]	1 467	2 596	635	932	166
ego-facebook [150]	2 888	5 962	35	104	35
as20000102 [134]	6 474	27 790	3 885	19 437	3 690
arenas-pgp [16]	10 680	48 632	8 673	44 074	7 944
web-webbase-2001 [17]	16 062	51 186	5 253	24 232	3 574
as-caida20071105 [134]	26 475	106 762	13 393	69 184	13 252
ia-email-EU [134]	32 430	108 794	6 262	53 228	6 259
topology [218]	34 761	215 440	19 246	168 782	19 128
douban [217]	154 908	654 324	59 524	462 128	59 493

The SIS epidemic process can be described by the network

$$S_i + I_j \xrightarrow{a_{ij}\lambda} I_i + I_j, \qquad I_i \xrightarrow{\gamma} S_i,\ 1 \leq i, j \leq N,\ j \neq i, \tag{3.8}$$

where the first reaction models infections by neighbors and the second reaction is the spontaneous recovery, with parameters λ and μ respectively. In a similar fashion, different variants of the process, such as SIR, SIRS, and SEIR [166], can be described. Any physically meaningful initial condition $\hat{\sigma}$ for this network must be such that each node i is initially infected ($\hat{\sigma}_{S_i} = 0$, $\hat{\sigma}_{I_i} = 1$) or susceptible ($\hat{\sigma}_{S_i} = 1$, $\hat{\sigma}_{I_i} = 0$). This setting makes stochastic models of epidemics spreading on complex networks difficult to study exactly because the state of each individual node is tracked explicitly [213], leading to a state space size with 2^N distinct configurations [185]. With SE it is possible to discover an ordinary lumpability of the underlying Markov chain, without ever generating it, on the network of Eq. 3.8, which has exponentially smaller size because it has $2N$ species and $E + N$ reactions, where E is the number of nonzero entries in the adjacency matrix of the graph.

It can be shown that, for the SIS model, the maximal SE is the trivial partition where all the species are in a single block. This is an invariant property stating that the total population of individuals in the system is constant [185]. A non-degenerate reduction may be obtained

by considering initial partitions with two blocks, hereafter denoted by $\mathcal{S} = \{S_i \mid 1 \leq i \leq N\}$ and $\mathcal{I} = \{I_i \mid 1 \leq i \leq N\}$, that separates species associated with nodes in the susceptible state from those in the infected state, respectively. Such a setting leads to noticeable SE reductions for SIS models evolving on several real-world networks from the literature (Table 4).

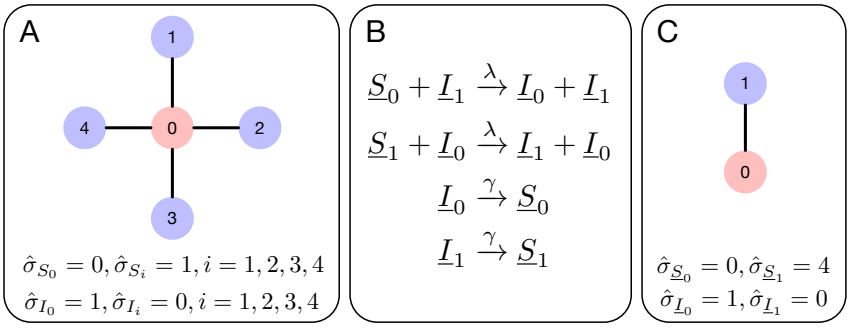


Figure 12: Example of SE reduction of SIS dynamics on a coarse-grained network. (A) Star network over which an SIS process evolves according to Eq. 3.8, starting from an initial condition where the infection starts at node 0. (B) Reduced network from the SE that partitions the species into blocks $\{S_0\}$, $\{I_0\}$, $\{S_1, S_2, S_3, S_4\}$ and $\{I_1, I_2, I_3, I_4\}$ (species representatives are underlined in the figure for clarity). (C) The SE partition induces a partition on the graph with blocks $\{0\}$ and $\{1, 2, 3, 4\}$. The reduced network corresponds to the description of the SIS dynamics on the quotient graph. The lumpability relation holds for an initial condition of the reduced network that is consistent with the initial condition of the original network up to SE, hence the initial number of agents in each macro-node of the quotient graph is equal to the sum of the initial conditions of the equivalence class that it subsumes.

An inspection of the obtained SE equivalence classes reveals that each refinement of the initial block $\mathcal{S}$ matches a refinement of block $\mathcal{I}$ for the same subset of nodes of the graph. As a result, SE naturally induces a partitioning of the graph, and the reduced network can thus be understood as a description of the SIS dynamics on the quotient graph where each macro-node subsumes a partition block of nodes induced by SE (Figure 12). According to this structural interpretation, the reduced

model is still an epidemic process, albeit on a coarse-grained network; as such, it is still amenable to the applicability of a wide range of analysis techniques developed for such systems [166, 213]. These include mean-field and pair approximation [43, 149, 208], whose computational cost for the generation and solution of the resulting nonlinear differential equations may benefit from the availability of a stochastically equivalent reduced model.

Since SE is a sufficient condition for Markov chain lumping, coarser aggregations of the Markov chain state space could be obtained in principle. Indeed the line graph in Figure 12 C admits the orbit partition that collapses nodes 0 and 1, thereby inducing a lumping following [185]. However, this is not detected by SE. The reason is that the lumpability relation induced by SE must hold for all population vectors that are equal up to SE. However, the lumpable partition derived with the approach in [185] violates this property because it does not collapse in the same block the states $S_0 + I_0 + S_1 + I_1$ and $S_0 + S_0 + I_1 + I_1$, which preserve the sums of infected and susceptible individuals. In our tested real-world networks in Table 4, the maximal SE induces a partition on the nodes of the graph which is a refinement of the orbit partition, although in many cases it is not considerably stricter. On the other hand, SE can be applied to models that do not satisfy the conditions required in [185]. For example, the SIS dynamics on the star network of Figure 12 can be lumped also in the case of node-specific kinetic parameters, whilst the results in [185] require equal transmission and recovery rates at every node.

3.4.1 SIS model with heterogeneous rates

Consider now a variant of the network presented in Fig 12. In this variant we assume node-dependent transmission and recovery rates that depend on whether the node is at the center or at the periphery of the star. More specifically, we consider the following mass-action reaction network:

$$
\begin{array}{lll}
I_0 \xrightarrow{\gamma_1} S_0 & S_0 + I_1 \xrightarrow{\beta_1} I_0 + I_1 & S_1 + I_0 \xrightarrow{\beta_2} I_1 + I_0 \\
I_1 \xrightarrow{\gamma_2} S_1 & S_0 + I_2 \xrightarrow{\beta_1} I_0 + I_2 & S_2 + I_0 \xrightarrow{\beta_2} I_2 + I_0 \\
I_2 \xrightarrow{\gamma_2} S_2 & S_0 + I_3 \xrightarrow{\beta_1} I_0 + I_3 & S_3 + I_0 \xrightarrow{\beta_2} I_3 + I_0 \\
I_3 \xrightarrow{\gamma_2} S_3 & S_0 + I_4 \xrightarrow{\beta_1} I_0 + I_4 & S_4 + I_0 \xrightarrow{\beta_2} I_4 + I_0 \\
I_4 \xrightarrow{\gamma_2} S_4 & &
\end{array}
\tag{3.9}
$$

For given $\gamma_1 \neq \gamma_2$ and $\beta_1 \neq \beta_2$, this model admits the same SE as above, namely

$$\mathcal{H} = \big\{\{S_0\}, \{I_0\}, \{S_1, S_2, S_3, S_4\}, \{I_1, I_2, I_3, I_4\}\big\}.$$

In fact, using the SE theory it is possible to prove that such equivalence holds *for any* choice of rates such that $\beta_1 \neq \beta_2$ and $\gamma_1 \neq \gamma_2$. To do so, let us assume without loss of generality that the kinetic rates are expressed as rational numbers $\beta_i = n_{\beta_i}/d_{\beta_i}$, $\gamma_i = n_{\gamma_i}/d_{\gamma_i}$, $i = 1, 2$. Then, we consider a time rescaling of this parameters such that all of them are non-negative integers, e.g., by multiplying each such parameter by the product of all denominators $d_{\beta_1} d_{\beta_2} d_{\gamma_1} d_{\gamma_2}$.

Let us denote rescaled parameters by $\hat{\beta}_i$ and $\hat{\gamma}_i$, $i = 1, 2$. Then, the original reaction network in Eq. 3.9 and the one with the rescaled kinetic parameters are equivalent in the sense that they give rise to the same stochastic behavior (i.e., the solution of the master equation) up to a time rescaling. Importantly, the rescaled parameters $\hat{\beta}_i$ and $\hat{\gamma}_i$ can now be interpreted as auxiliary species P_{γ_1}, P_{γ_2}, P_{β_1} and P_{β_2} in the following extended mass-action network with unitary kinetic rate parameters and auxiliary species P_{β_1}, P_{β_2}, P_{γ_1}, and P_{γ_2}:

$$P_{\gamma_1} + I_0 \xrightarrow{1} S_0 + P_{\gamma_1} \quad P_{\gamma_2} + I_1 \xrightarrow{1} S_1 + P_{\gamma_2} \quad P_{\gamma_2} + I_2 \xrightarrow{1} S_2 + P_{\gamma_2}$$

$$
\begin{array}{cc}
P_{\gamma_2} + I_3 \xrightarrow{1} S_3 + P_{\gamma_2} & P_{\gamma_2} + I_4 \xrightarrow{1} S_4 + P_{\gamma_2} \\
P_{\beta_1} + S_0 + I_1 \xrightarrow{1} I_0 + I_1 + P_{\beta_1} & P_{\beta_1} + S_0 + I_2 \xrightarrow{1} I_0 + I_2 + P_{\beta_1} \\
P_{\beta_1} + S_0 + I_3 \xrightarrow{1} I_0 + I_3 + P_{\beta_1} & P_{\beta_1} + S_0 + I_4 \xrightarrow{1} I_0 + I_4 + P_{\beta_1} \\
P_{\beta_2} + S_1 + I_0 \xrightarrow{1} I_1 + I_0 + P_{\beta_2} & P_{\beta_2} + S_2 + I_0 \xrightarrow{1} I_2 + I_0 + P_{\beta_2} \\
P_{\beta_2} + S_3 + I_0 \xrightarrow{1} I_3 + I_0 + P_{\beta_2} & P_{\beta_2} + S_4 + I_0 \xrightarrow{1} I_4 + I_0 + P_{\beta_2}
\end{array}
$$

Here the auxiliary species act as *catalysts*, i.e., their population remains constant at the occurrence of every reaction. Thus, it is easy to see that if the initial condition is such that it matches the values in the original (time-rescaled) model, i.e.,

$$
\hat{\sigma}_{P_{\beta_1}} = \hat{\beta}_1, \qquad \hat{\sigma}_{P_{\beta_2}} = \hat{\beta}_2, \qquad \hat{\sigma}_{P_{\gamma_1}} = \hat{\gamma}_1 \qquad \hat{\sigma}_{P_{\gamma_2}} = \hat{\gamma}_2,
$$

then the two networks give rise to the same underlying Markov chain. The expanded reaction network, however, can provide a statement of SE that is independent from the actual values of the parameters. Indeed, it can be shown that the partition $\mathcal{H}'$ defined as

$$
\mathcal{H}' = \mathcal{H} \cup \big\{ \{P_{\beta_1}\}, \{P_{\beta_2}\}, \{P_{\gamma_1}\}, \{P_{\gamma_2}\} \big\}
$$

is an SE for the expanded reaction network. This induces a lumpable partition on the underlying Markov chain (hence, on the underlying Markov chain of the original reaction network) for all states independently of the values of the auxiliary species — hence, independently of the choice of transmission and recovery rates in the original reaction network.

3.5 Discussion

Table 5 shows the results obtained by SE over a collection of models taken from BioModels repository [136], JWS Online [162] as well as benchmark networks obtained from BioNetGen [14] rule-based models. The table divisions are set to group models that appear in the same scientific publication, although they describe different biological mechanisms.

Models M1-M2 show reductions in the number of species but not in the number of reactions. A manual inspection of these models and the

resulting SE equivalence classes reveals that SE equates species with no dynamical role in the network as they can be interpreted as distinct auxiliary species that are used to model zero order reactions in the form $I \xrightarrow{\alpha} I + A$, degradation reactions such as $A \xrightarrow{\alpha} SINK$, or purely catalytic species such as $A + C \xrightarrow{\alpha} B + C$, where I and $SINK$, and C are the auxiliary species that are otherwise not involved as reagents in the network.

Models M3-M12 present a small reduction in the number of species and reactions, that can be explained in a similar fashion to the previous models. M20 on the other hand, is a simplified model for prothrombinase action based on [133]. The model describes the transformation of coagulation factor II (Prothrombin) (II) in Meizothrombin (M) and Prethrombin (P2), as an intermediate step for factor II's activation.

Model M13 depicts sugar isomerization, one of the main reaction pathways in the sugar-casein systems, in which aldoses and ketoses can isomerize into each other. The reaction network also considers the Maillard reaction, in which sugars react with the lysine residues [23]. Here, SE aggregates species that are produced and never consumed in a single macro variable which produces a reduced system consisting of six species and ten reactions from the original model (11 species and 11 reactions).

Interestingly, in a number case studies (such as those discussed in Section 3.3) we observe that symmetries carry over to equivalences at the level of the underlying quantitative semantics. This appears often in case studies analyzing protein phosphorylation and signaling pathways. In addition, it is worthy to remark that when applying FE to the models in tables 5 and 3 we observe that in all cases the largest computed FE equivalence corresponds to the discussed largest SE, except for the models *MODEL1511170000-02* (Table 3).

Table 5: SE reductions.

			Species		*Reactions*	
Id	*Name*	*Cit.*	*Orig.*	*Red.*	*Orig.*	*Red.*
1	BIOMD0000000362	[28]	34	30	45	45
2	MODEL4780784080	[96]	14	13	16	16
3	BIOMD0000000077	[15]	8	7	8	7
4	BIOMD0000000188	[175]	19	13	20	19
5	BIOMD0000000189	[175]	17	10	14	13
6	MODEL1511170000	[144]	7	6	8	7
7	MODEL1511170001	[144]	7	6	8	7
8	MODEL1511170002	[144]	8	7	10	9
9	MODEL1608100000	[1]	4	3	4	3
10	MODEL1609100000	[84]	7	2	7	1
11	MODEL8262229752	[140]	22	16	25	17
12	MODEL1411130000	[152]	8	4	8	6
13	BIOMD0000000052	[22]	11	6	11	10
14	MODEL1001150000	[167]	156	76	480	264
15	BIOMD0000000699	[44]	24	16	71	36
16	BIOMD0000000702	[44]	24	16	71	36
17	BIOMD0000000705	[188]	26	22	61	52
18	BIOMD0000000706	[188]	40	36	127	102
19	S6K - Insulin model I	[111]	15	10	16	10
20	S6K - Insulin model II	[111]	15	10	16	10

3.6 Concluding remarks

Stochasticity is a key tool to understand a variety of phenomena regarding the dynamics of reaction networks, but the capability of exactly analyzing complex models escapes us due to the lack of analytical solutions and the high computational cost of numerical simulations in general. Species equivalence enables aggregation in the sense of Markov chain lumping by identifying structural properties on the set of reactions, without the need of the costly enumeration of the state space. Owing to the polynomial space and time complexity of the reduction algorithm, it can be seen as a universal pre-processing step that exactly preserves the stochastic dynamics of species of interest to the modeler. Since species

equivalence gives rise to a network where the reactions preserve the structure (up to a renaming of the species into equivalence classes), the reduction maintains a physical interpretation in terms of coarse-grained interactions between populations of macrospecies.

Another consequence of the availability of a reduced network is that our method is orthogonal to any of the analysis techniques developed for stochastic reaction networks. Numerical simulations may run faster because they traverse fewer reactions at each time step [40]; when feasible, one can generate the underlying Markov chain to be further analyzed or reduced [99, 156, 205]; the reduced network can be subjected to complementary coarse-graining techniques concerned with time-scale separation [29, 87, 114, 186]. More generally, since the reduced network preserves the stochastic dynamics in the sense specified above, it can be used as the basis for various forms of approximate analysis such as linear noise or moment closure approximation [183], where the complexity of the resulting system of equations grows rapidly with the network size.

Chapter 4

Exact reduction of quantitative biological models

In this chapter we present a framework for the large-scale assessment on quantitative models of biological systems, that includes the conversion of SMBL models into a CRN-like syntax. To test our framework we use a snapshot of 1640 biological models extracted from BioModels Database. We evaluate the performance of each equivalence technique in terms of the reduction in the number of species and reactions. To conclude, we discuss a series of case studies that illustrate the physical interpretability of the obtained aggregations.

4.1 Repositories and databases of biological models

The growing interest in computational biology has lead to an increased amount of publicly accessible peer-reviewed biological models. A brief review on the current state of the art of the model databases and resources in systems biology is proposed in [83].

Initiatives such as the Gene Ontology (**GO**) [7] consortium facilitate the availability of annotated biological data. An important part of this inititiave is the GO's Reference Genome Project, which provides a framework to infer annotations of biological components (e.g., genes or protein and RNA molecules produced by a gene) from a broad set of genomes from experimental annotations in a semi-automated manner [82]. The biological data is integrated from different resources and databases such as **TAIR**, **WormBase**, **FlyBase**, the Protein ANalysis THrough Evolutionary Relationships Pathway (**PANTHER**) and the EBI project for human UniProtKB-Gene Ontology Annotation [**UniProtKB-GOA**].

The availability of data also influenced the creation of repositories that store mechanistic models of biological processes. Repositories such as **NFsim** [189] gather synthetic mechanisms that describe processes such as multisite phosphorylation in proteins, polymerization, branching and severing in cells, and cellular behavior in bacteria, among others.

Other database resources instead focus on specific processes. A popular metabolic pathway database is the **Reactome** project [73]. This database lists a collection of curated human pathways with molecular details of signal transduction, transport, metabolism and other cellular processes. **Reactome** gathers curated core pathways and reactions in human biology, with each of its components being cross-referenced in sequence databases such as **NCBI, UniProt**, the Kyoto Encyclopedia on Genes and Genomes (KEGG) [113] (**KEGG**) and **GO** [7]. In addition to curated human events, one can also find inferred orthologous events in 22 non-human species including mouse, rat, chicken, puffer fish, worm, fly, yeast, E.coli and two plants.

While **Reactome** is mostly dedicated to the human organism, there are biological databases with a broader scope such as **MetaCyc** [41], a general reference database on metabolism, or **KEGG** [113], focused on high-level functions of biological systems from molecular-level information. It is not clear where to draw the line for a proper classification of these biological resources, but a suitable cutting point is database specificity. For example, organism-specific resources such as **EcoCyc** [117] and **ECMDB**[94] are largely dedicated to the metabolism and genes of

Escherichia Coli. However, there is a rich variety of resources focusing in several areas such as metabolism (**BiGG** [120], **BioSilico** [100], **BioCarta** [159]); or molecular interactions, i.e enzymes (**BRENDA** [109], **ENZYME**[8]) or lipids (**LIPID MAPS** [53, 192]).

Since many model repositories use their own storage format, one intuitive criterion to categorize these database resources is provided by their underlying data structure. For example, **CellML** [153] uses XML markup language to characterize a wide range of curated electrophysiological and cellular processes. A common standard for model encoding is given through the systems biology markup language (SMBL) [103]. SBML is machine-readable format based on XML used to describe biological systems where the biological entities are involved in, and modified by processes that occur over time, such as chemical reaction networks.

Several database resources employ this encoding, being so popular that even a fragment of the models contained in **KEGG** of metabolic pathways have been annotated in it. Furthermore, **JWS Online** [162] uses a database implementation with a native format that mirrors the SBML specification, and encodes several quantitative models of biochemical processes and a platform for model simulation.

4.1.1 The BioModels repository

The BioModels Database is a repository of computational models of biological processes [136]. It hosts dynamical quantitative models described in peer-reviewed scientific literature as well as models generated automatically from pathway resources such as KEGG [113], BioCarta [159], MetaCyc [41], PID [172] and SABIO-RK [215]. BioModels covers a wide range of models from several biological categories such as biochemical reaction systems, kinetic models, metabolic networks, steady-state models and signaling pathways. Models are available, among other formats, in SBML [103].

The BioModels repository is divided into two sections: the *curated branch* and the *non-curated branch*. The former contains models that have

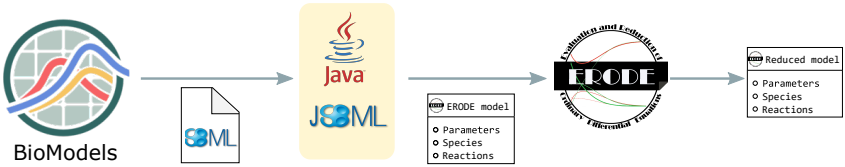


Figure 13: Workflow overview. Models were downloaded from the BioModels repository in the SBML format. We implemented a tool to translate the SBML description into the CRN-like input (`.ode` format) of ERODE. The output of ERODE is a reduced CRN with reactions involving *macro-species*, each representing the sum within an equivalence class of original species. We manually inspected the ERODE output to provide a physical interpretation of the obtained equivalences.

been manually checked and their components annotated using unambiguous identifiers [112] that refer to external biological databases [50, 75, 191] or ontologies (such as Gene Ontology [7], SBO [54] or ChEBI [61]). Models are curated following the Minimum Information Required in the Annotation of Models guidelines (MIRIAM) [131]. Models that are not MIRIAM-compliant are stored in the non-curated branch, which also contains non-kinetic models such as flux balance analysis models, or models which are not supported by SBML (e.g., spatial models). A more detailed description of this database is available at [46].

4.1.2 Encoding SBML into CRN syntax

We developed a tool for the conversion of SBML models to ERODE format, a CRN-like syntax. We build the tool using Java 1.8.0_121 and NetBeans IDE 8.2 (Build 201705191307) runing on Windows 10 v.1709. To access the objects in the SBML description [67, 132] we used the *jsbml* library[1] version 1.2.

An ERODE specification is a file divided in three ordered sections: parameters (kinetic rates), species, and reactions; each section is delimited using the labels **`begin/end`**. The reactions are defined using mass-action kinetics by default unless the keyword **`arbitrary`** is used. Af-

[1]Here we explain each step of the translation process; for specific details of the SBML structure, please consult the official resources.

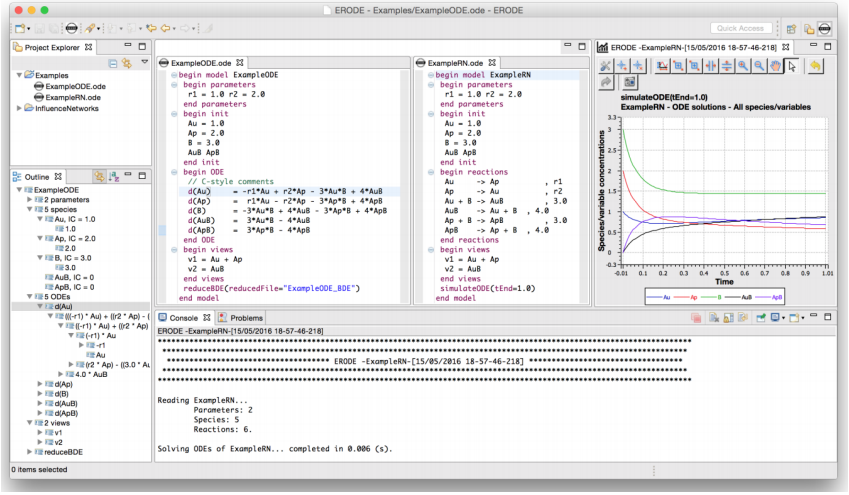


Figure 14: A screenshot of ERODE

ter the components of the model have been declared, the user can write list of commands for analysis, reduction, and export. Figure 14 shows a screenshot of ERODE.

The SBML specification is a more complex file that encodes also information of the species compartment, rules of assignment for the species and parameters, propensity rules constraining reaction rates, etc. The conversion process is regulated using three structures that verify the *correctness* of the conversion of each parameter, species and reactions in the CRN encoding:

- `BannedNamesList` stores a list of the reserved words in the ERODE environment and its main function is to verify that a *term* (species, parameters, etc.) extracted from the SBML specification can be written into ERODE. If the name of the *term* is a banned name, we make a name replacement in the ERODE model and annotate the original name.

- `SBOChecker` stores a dictionary of SBO labels depicting mass-action kinetics. We use this to verify the kinetics of each reaction upon the

translation (See 4.1.2.3).

- `KineticFormulaParser` contains methods that operate over the kinetic law describing the reaction rate.

4.1.2.1 Parameters

An SBML *parameter* associates a symbol with a value, so that it can be useds in mathematical formulas. Parameters can have constant values for the complete duration of a simulation, or their value can be updated according to some specific rule. See for example *Assignment Rules* and *Rate Rules* in [103–106].

Global parameters are defined at the top of the SBML model and can be used by any of the elements of the model, although there are also parameters which are defined *locally* for the use of a specific reaction. The following SBML snippet specifies a parameter from *BIOMD0000000030*, a model from BioModels Database that we will use as running example to explain the conversion process.

```
1 <parameter id="k1" metaid="metaid_0000019" name="k1" value="0.02"/>
```

We translate this into ERODE using the parameter's *id* and *value* and write it in the ERODE parameters list, i.e., `k1 = 0.02` (delimited by **`begin parameters/end parameters`**). In the case of non-constant parameters we replace the value of the parameter by the rule assignment directly in the reaction rates (Section 4.1.2.3)

4.1.2.2 Species

An SBML *species* can represent a chemical species, or groups of reactant entities that take part in a chemical reaction. Species are associated in compartments, which indicate the section of the cell on which the species is present. The attributes `initialAmount` or `initialConcentration` are used to denote the species initial quantity. These attributes are mutually exclusive, which means that only one can have a value in the species instance. In general, the quantity of the species varies during simulation, except in the case of *constant* species or species which belong to the

boundary of the system, i.e., species which can intervene in the reactions as reactants but their values are not determined by a kinetic law.

If the species is *constant*, this means that the concentration of the species is not altered during the reaction process (Section 4.1.2.3). This is also true for the *boundary* species, however, the concentration in time of these species can be updated via specific rules (SBML manual [104–106]). The next SBML snippet defines a species:

```
1 <species compartment="cell" id="M" initialConcentration="800"
2         metaid="metaid_0000005" name="MAPK"/>
```

For each species ERODE only requires an identifier and the species' initial concentration. Thus, we write the species in the example as `M=800` within ERODE 's species declaration section (delimited by **`begin init/end init`**). In the case of species whose concentration is determined by rules, e.g., boundary species, we replace the species by its corresponding rule assignment directly in the reaction rate (Section 4.1.2.3).

4.1.2.3 Reactions and other structures

An SBML *reaction* represents a transformation in the concentration of one or more species. The reversibility of a reaction process is given by the *optional* attribute `reversible`, by default set to `true`. Two lists, called `listofReactants` and `listOfProducts`, store the species which are being consumed and produced during the reaction, respectively. An *optional* list, called `listOfModifiers`, contains species that mediate in the reaction such that their concentration is not affected. In our encoding, we treated as modifiers all species labeled as *constant* or part of *boundary* system.

The speed of the reaction is dictated by a `KineticLaw` object, which inherits an optional attribute `sboTerm` of type `SBOTerm` that defines the type of reaction rate encoded by the `KineticLaw` instance. The SBO identifier contained in `sboTerm` refers to a term from the vocabulary defined in the System Biology Ontology repository. The inspection of this attribute is what allows to identify the type of kinetics of the reaction by capturing the label assigned to the term in the ontology, e.g. the la-

```
<reaction id="reaction_0000001" metaid="metaid_0000046"
        name="binding MAPKK on Tyr site of MAPK">
  <listOfReactants>
    <speciesReference metaid="_063184" species="M"/>
    <speciesReference metaid="_063196" species="MAPKK"/>
   </listOfReactants>
   <listOfProducts>
     <speciesReference metaid="_063208" species="M_MAPKK_Y"/>
   </listOfProducts>
   <kineticLaw metaid="_063220">
     <math xmlns="http://www.w3.org/1998/Math/MathML">
      <apply>
        <times/>
          <ci>cell</ci>
            <apply>
              <minus/>
                <apply>
                  <times/><ci>k1</ci><ci>M</ci><ci>MAPKK</ci>
                </apply>
                <apply>
                  <times/><ci>k_1</ci><ci>M_MAPKK_Y</ci>
                </apply>
            </apply>
       </apply>
      </math>
    </kineticLaw>
</reaction>
```

Figure 15: Sample SBML reaction adapted from *BIOMD0000000030*

bel "SBO:0000562" corresponds to "mass action like rate law for second order irreversible reactions, one reactant, one essential stimulator".

The `math` element inside the `KineticLaw` holds a MathML formula defining the rate of the reaction. Each formula can involve species contained in any of the three internal lists of the reaction, and both local and global parameters. Figure 15 shows a snippet of an SBML reaction, where we omitted the `annotation` tag for the sake of readability.

ERODE represents the reactions as $\rho \longrightarrow \pi, f$, remembering the notation in equation 2.1. In each reaction, ρ is the sum of the elements (with their corresponding stoichiometry) contained in the two list of species `listofReactants` and `listOfModifiers`. Then, π is the sum of the elements in `listOfModifiers` and `listOfProducts`. However, if the reaction does not have modifiers, then we need to account for two particular cases. First, if `listofReactants` is empty, i.e., $\rho = \emptyset$, we

are in the presence of the spontaneous creation of a product. We encode this reaction setting as reactant the *dummy* species `Source`, which has no dynamics, e.g., `Source -> S` encodes the creation of species `S`. Second, if the reaction does not have defined products, i.e., $\pi = \emptyset$, then this is a reaction which encodes the spontaneous degradation of a species. We solve this by creating a dummy `SINK` species which will be set as product of the reaction, e.g., `S -> SINK`. The SBML reaction in Figure 15 is then translated into the following ERODE reaction:

```
M+MAPKK -> M_MAPKK_Y,arbitrary cell*(k1*M*MAPKK - k_1*M_MAPKK_Y)
```

Here, the left- and right-hand sides of the reactions are taken from the SMBL lists as described above, the term `cell` is the compartment where the reaction occurs and the **`arbitrary`** keyword denotes a reaction with a generic non-mass-action propensity function.

The conversion of the `KineticLaw` to a reaction rate f is not straightforward. We pay particular attention to the recognition of mass-action models to which FE, BE, and SE can be applied. As discussed, SBML allows the direct specification of the *type* of kinetics by means of appropriate SBO labels in the `KineticLaw` object. However, we encountered cases where the `sboTerm` attribute is omitted, and upon manual inspection of the reactions these are effectively mass-action kinetics. An example is given in the reaction in Figure 15. It is easily noticeable that the reaction is encoding a reversible mass-action reaction with kinetic parameters `cell * k1` and `cell * k_1`, describing the forward and backward reactions respectively.

We manually detected such occurrences of non-tagged mass-action reactions and translated them into ERODE mass-action ones. We inferred the forward and reverse rate functions as the left and right operand, respectively, of the topmost `minus` MathML tag (Line 16). This leads to the two following ERODE irreversible reactions (as ERODE does not support reversible reactions):

```
M + MAPKK -> M_MAPKK_Y,   cell *  k1
M_MAPKK_Y -> M + MAPKK,   cell * k_1
```

Testing ERODE can export the ODEs underlying a model as a MATLAB function. Likewise, in BioModels all models come with an encoding as MATLAB functions. We tested our converter over a large random selection of BioModels files by checking that their MATLAB functions and those exported by ERODE corresponded.

4.1.3 Data processing

In our experiments we used the BioModels repository's snapshot of the 26th of July 2017. The data consists of 640 models in the curated branch *BIOMD0000000001-640* and 1000 models in the non-curated branch (with ids ranging from *MODEL0072364382* to *MODEL9811206584*). We performed a preprocessing step to filter out models that could not be used for the analysis. In the non-curated branch only 491 models are kinetic models described as ODE systems, while the others are described in formalisms, e.g., logical or flux balance analysis models outside the scope of applicability of species equivalences. Overall, we could process 448 models from the curated branch and 219 from the non-curated one, for a sanitized dataset of 667 models. Of these, 43 were identified as mass-action CRNs (as detailed in Section 4.1.2).

The most frequent reasons for discarding a model were (within parenthesis we give the frequency in the curated branch, which we assume to be more stable):

- syntactic limitations in our converter prototype, including the lack of support for models without explicit reactions where the dynamics is given by rate rules over a set of parameters,
  e.g., *BIOMD0000000020* (114);

- models with unsupported propensity functions such as `tanh` and `exp` (31);

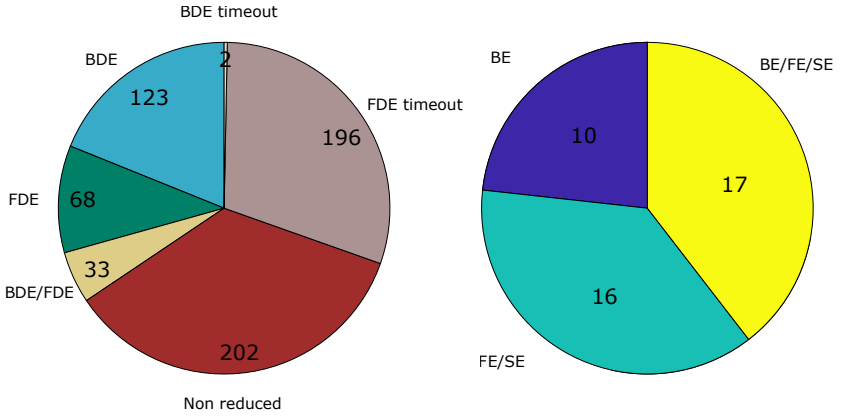
- models with species with *Assignment Rules*, used to model features such as delayed equations and hybrid systems, not supported by ERODE (47).

4.2 Reduction results

Non mass-action models were analyzed using FDE and BDE, whilst for mass-action ones we used FE, BE, and SE. In a preliminary analysis we considered the maximal equivalences for all cases, computed by starting the partition-refinement algorithms with the initial singleton partition with a single block containing all species in the CRN. However, in 32 cases we found that the maximal FDE/FE collapsed *all* species and reactions. This is because these CRNs are *closed* and *mass-preserving*, meaning that the concentrations (represented by the ODE solutions for each species) just flow among the species, but the *total* cumulative concentration is constant. Therefore these systems can be self-consistently written as a single-equation ODE with zero derivative (and initial concentration equal to that total cumulative concentration).[2] In these 32 cases, we built more meaningful ad-hoc initial partitions to be used in the partition-refinement algorithm by isolating variables of interest to the modeler, in accordance to the model's output as evinced from the scientific publication associated with each model.

For each equivalence we computed the reduced CRN, recording the resulting number of species and reactions as a measure of the effectiveness of the exact reduction techniques. Figure 16 counts the models that could be reduced by at least one technique, regardless of the reduction ratio. For the non mass-action models (Figure 16a), 224 models (36%) could be reduced. In particular, only 33 models could be reduced by both FDE and BDE, suggesting that they are not comparable. Several models (196, 31%) could not be analyzed due to the excessive computational cost of FDE, whereas only 2 due to BDE (we used a time-out of 8 hours). This is consistent with more (and more complex) SMT checks required by FDE with respect to BDE [39].

[2]We remark that this situation is analogous to ordinary lumpability in Markov chains, where the coarsest ordinarily lumpable partition contains all states because the sum must always preserve the probability mass at all time points.



(a) Non mass-action (b) Mass-action

Figure 16: Reduction results.

All the mass-action models (Figure 16b) could be reduced by at least one equivalence relation. Ten models (23%) could be reduced with BE only, seventeen models (40%) could be reduced with the three methods and sixteen models (37%) could be reduced with SE and FE. The fact that the models that are reducible with SE can also be reduced by FE derives from the aforementioned property that SE implies FE.

Figure 17 shows a scatter plot to summarize the reduction ratio for each model using the species equivalence that yielded the best reduction. The average reduction ratio in the number of species and reactions for each method is: BDE (23%, 8%), FDE (50%, 48%), BE (15%, 10%), FE (37%, 30%) and SE (36%, 29%). Figure 18 illustrates the reductions obtained. For each species equivalence, we group the models in 6 histogram bins (0%, (0%-20%], ..., (80%-100%]) in two series showing the reduction ratio of the species (red) and the reactions (blue). There are models which reduce in the number of species but not the reactions. This can be due to an equivalence among species with dynamical role in the network as they can be interpreted as distinct auxiliary species that are used to model zero order reactions, such as `I -> I + A`, degradation

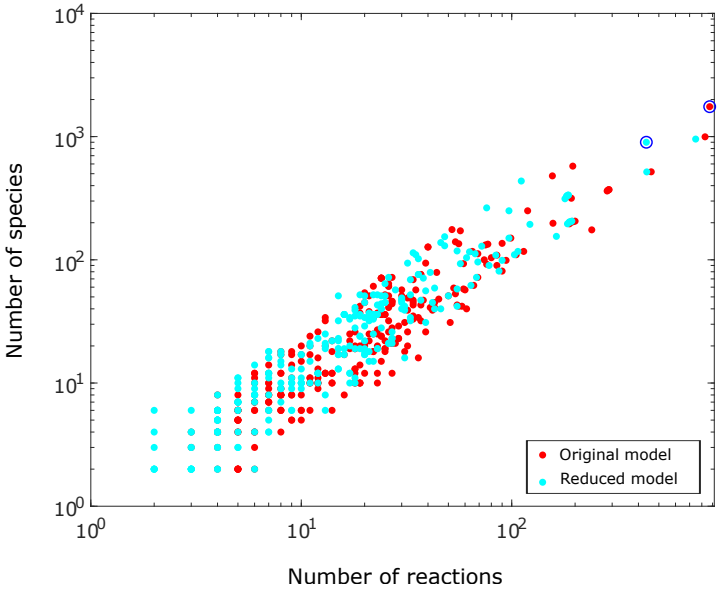
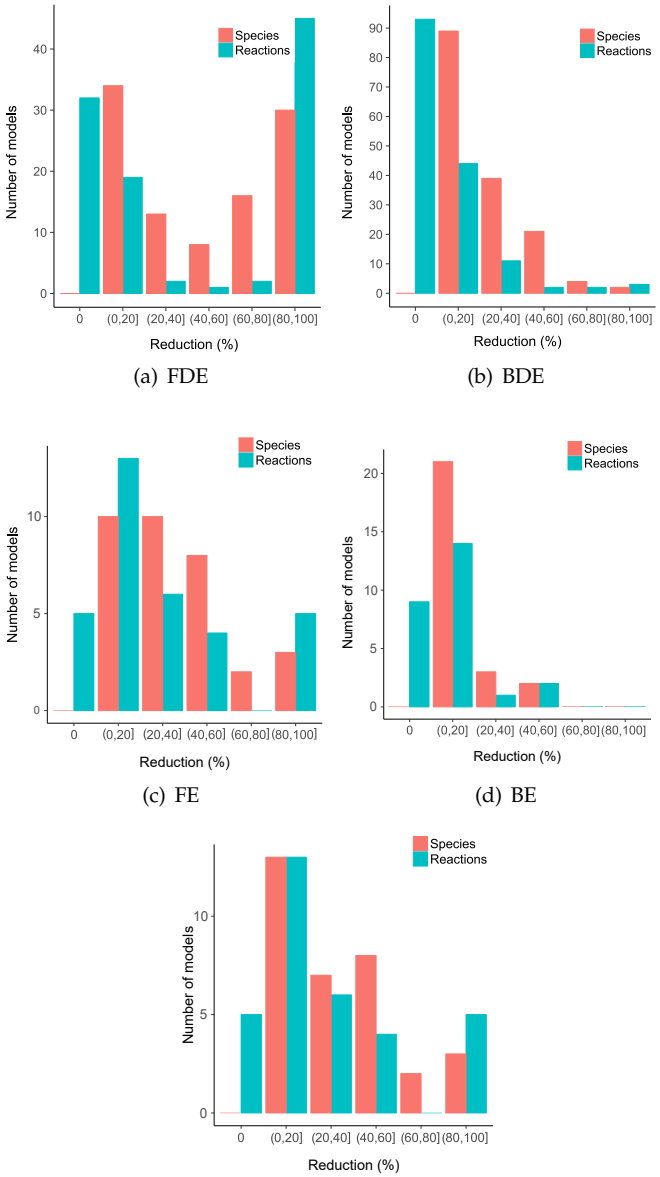


Figure 17: Comparison among original and reduced species and reactions (log scale).

reactions such as `A -> SINK`, or purely catalytic species such as `C` in the reaction `A + C -> B + C`, with `I, SINK` and `C` species that are otherwise not involved as reagents in the network.



(a) FDE (b) BDE

(c) FE (d) BE

(e) SE

Figure 18: Reduction ratios for the species and reactions for each species equivalence.

4.3 Case Studies

We selected three case studies to highlight the physical interpretability of the reductions obtained.

4.3.1 Aggregation of symmetric molecular complexes in MAPK phosphorylation scheme

The Mitogen-activated protein kinase, popularly known as MAPK is a fundamental system in cell signaling processes, and consists of a large family of enzymes that exhibit a switch-like behavior in a phosphorylation cascade. There are several families in the MAPK signaling pathway, and each of them is initialized by specific extracellular agents which leads to the activation of a particular MAPK. The JNK and ERK 1/ 2 pathways can both be activated by RTK receptors via a Ras|Raf complex, although other activation paths are possible for both kinases, e.g., ERK 1/ 2 can also be activated by the Integrin receptor through a Ras|Raf interference. In addition, the p_{38} MAPK family can be activated through both the IL -1 R and TNFR receptors, but also by multistep activation from Cdc 42.

In general, the mechanism starts with an external stimulus that activates precursor enzymes or a target receptor. This leads to the eventual activation of a MAPK module and starts a phosphorylation cascade with the activation of the MAPK kinase kinase (MAPKKK). MAPK's are a three tier cascade where the active kinase in each level mediates the phosphorylation of the kinase downstream. The active MAPKKK phosphorylates and activates MAPKK, which in turn activates MAPK, a serine/threonine-specific protein kinase that requires phosphorylation on conserved threonine (T) and tyrosine (Y) residues.

Here, we analyze *BIOMOD0000000030*, a model of the double phosphorylation/dephosphorylation of MAPK. The kinetic mechanism shows four distinct forms for MAPK: inactive (M), active (Mpp), and two partially phosphorylated forms MpY and MpT which represent the phosphorylation in the tyrosine and threonine residues, respectively. This dynamics is represented in a model with 18 species and 32 reactions.

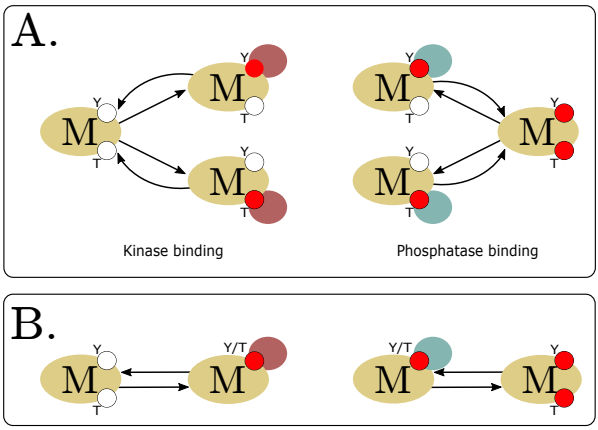


Figure 19: (A) Mechanisms for the phosphorylation and dephosphorylation of MAPK (M) [147]. Red and white circles indicate respectively a phosphorylated (on) or unphosphorylated (off). Initially both residues are *off* and turn *on* upon kinase binding. Reversely, when M is activated, the residues becomes *off* upon phosphatase binding. (B) BE yields compact representations of the mechanism in (A) by establishing an equivalence over the two symmetric complexes in each binding mechanism.

The model shows distributive kinetic mechanisms in both the phosphorylation and dephosphorylation processes (Figure 19). The MAPK molecule is phosphorylated by interaction with a kinase that upon binding transfers a phosphate group, resulting in the phosphorylation of the MAPK's site. Then, the kinase releases the intermediate monophosphorylated molecule and a new collision is required to phosphorylate MAPK's remanent residue. The same mechanism but in reverse is observed between the phosphorylated form of MAPK and a phosphatase.

BE produces a reduced model that equates symmetric molecular complexes in both the phosphorylation and dephosphorylation mechanism. The molecular complexes of the MAPK and the kinase are set in the same equivalence class, regardless the site where the kinase is bound. In the same fashion we group the molecular complexes of the fully phosphorylated MAPK and the phosphatase. This yields a reduced model of 16 species and 28 reactions.

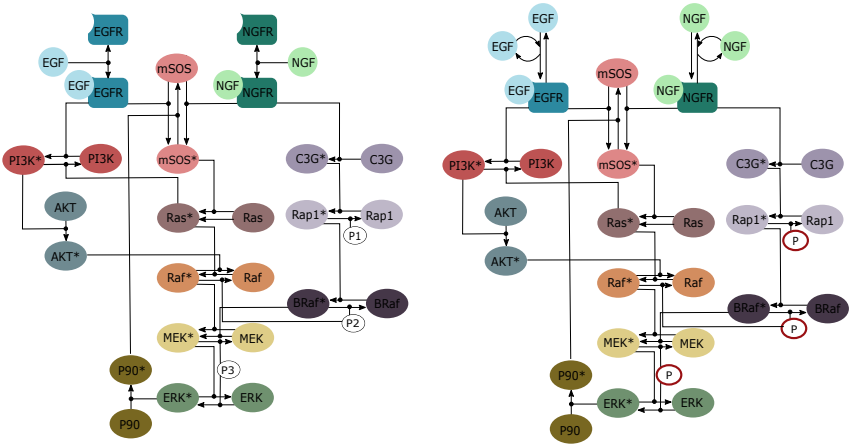
4.3.2 Aggregation of opposite states in components of a MAPK signaling cascade

Phosphorylation cascades in signal transduction can experience diverse organizational phenomenons such as sensitivity. The sensitivity is the influence of variations in the signal in the elements of the pathway; which augments with the number of layers in the circuit. The literature uses the term *ultrasensitivity* when the response is very susceptible to variations in the upper levels, such that slight variations in any level of the chain can produce large changes in the response.

Several efforts have been made in characterizing ultrasensitivity in the MAPK cascade [101, 118, 147]. In the ERK-MAPK cascade, the Ras protein acts as a critical switch in response to receptor signals that can account for biological responses such as cell growth, differentiation and proliferation. However, due to its central role, the ultrasensitivity of this molecule also dictates its carcenogenic potential [148].

The ERK – MAPK pathway starts with the extracellular activation of a RTK receptor by a signaling molecule. In response to the stimuli, RTK receptors recruit cytosolic Son of Sevenless homolog protein (SOS) to the cell surface where it stimulates GTP/GDP exchange in Ras. Active Ras recruits serine/threonine-protein kinase Raf 1 to the membrane promoting its activation and the beginning of a MAPK cascade that comprises the secuential phosphorylation of MEK 1/ 2 and ERK 1/ 2. Active ERK 1/ 2 can stimulate the kinase P 90 RSK or translocate to the nucleous and promote the transcription factors (TF) CREB, Elk 1 and c-Myc. Upon stimulation, P 90 RSK translocates to the nucleus and triggers a different response in the pathway by promoting only the transcription factors CREB and c-Fos, a TF involved in the promotion of collagenases.

Figure 20 A adapts the reaction mechanism in *BIOMOD0000000033*, a model of EGFR – NGFR signaling that concludes in ERK activation [26]. The model is formed by one central MAPK pathway, interconnected with two pathways PI 3 K and C 3 G. Upon binding of their respective signaling molecules (EGF and NGF) both receptors activate Ras by stimulation of (mSOS), which leads to the ERK phosphorylation cascade. In a dif-



(a) Original network

(b) BDE

Figure 20: (a) Adaptation of the EGFR/NGFR signaling networks in [26] were we represent the two-states of each kinase and their relations.(b) BDE reduction. The upstream input is replaced by a module describing EGF activity over receptor-bound EGF and the activity of the unregulated phosphatases $P_{1,2,3}$ is aggregated as single molecule, namely, P.

ferent path, the NGFR complex activates RAP 1-GTPase, a small GTPase from the Ras family, who activates the protein BRaf and B – Raf molecules upregulate ERK 1 / 2 by positive stimulation of MEK 1 / 2. The signal received in the EGFR receptor also activates units of PI 3 K in a bifurcation motif that starts the PI 3 K – AKT signaling pathway [154]. Active AKT downregulates ERK 1 / 2 via negative regulation of RAF 1. Independently, the MAPK pathway is regulated by P 90 in a negative feedback loop to mSOS.

The complete dynamics of the model is represented using 26 reactions and 32 species, where the active and inactive states of all the kinases involved are modeled as independent units. To facilitate visualization, components with the same color in Figure 20 belong to the same molecule and the symbol "$*$" indicates the active state, i.e., pink circles represent mSOS* and mSOS, the active and inactive forms of molecular SOS. BDE simplifies the step of EGF binding the free EGFR receptor unit,

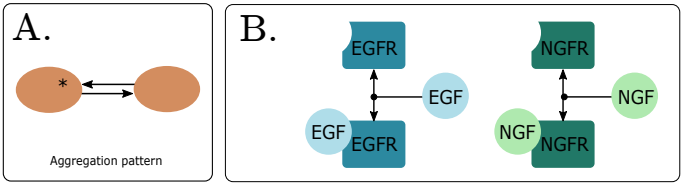


Figure 21: (A) Aggregation pattern found in the FDE-equivalence classes of [26] that equates the two forms of each molecular unit. (B) FDE reduction. The switch-like dynamics of the pathway is simplified producing a reduced model given just in terms of the binding/unbinding mechanism of EGF to EGFR.

to a mechanism where EGF interacts directly with the bound EGFR. The phosphatases $P_{(1-3)}$ are aggregated in a single equivalence class, which is reasonable, considering that these species exhibit a purely catalytic behavior and their concentration is not affected by any reaction. This yields the reduced model in Figure 20 B, characterized by 27 species and 26 reactions.

The FDE reduced model is described using 18 species and 4 reactions. Upon inspection of the FDE equivalence classes , we discover the aggregation pattern in Figure 21 A. This pattern holds for the two states of each species in Figure 20 A, such that the active and inactive forms of each molecule are equivalent and lumped to a macrovariable given by the sum of both states, i.e., $\mathrm{Raf1}^{*} = \mathrm{Raf1}$. The dynamics of the active and inactive variants are identical but with opposite sign which results in lumped variables with *zero* contribution. As consequence, the phosphorylation pathway is collapsed and we obtain the model in Figure 21 B, where the only dynamic preserved is the one corresponding to the free EGF molecule, and both EGF-bound and free units of EGFR. Interestingly, we capture the same aggregation pattern (aggregation of opposite states of a molecule) in the FDE reductions of other signaling models such as [13, 163].

4.3.3 Aggregation of diverse molecular units with symmetric behavior

As discussed in the previous case studies, equivalence-based reduction methods can capture symmetries among molecular components. In both of the examples above the equivalence is given between different variants of the same component, i.e., in 4.3.1 we relate two MAPK complexes that differ only in the binding site, whereas in 4.3.2 the relation is established between the two opposite states of a kinase. Here we use *MODEL1112260000*[3], a model for the analysis of the response of FOXO transcription factor, to show that SMB can also establish equivalences between different components that exhibit symmetric behavior.

The Forkhead Box (FOX) is a family of transcription factors with more than 100 members in the human organism classified from FOXA to FOXR. These transcription factors are involved in several biological processes such as apoptosis, cell metabolism, differentiation, and drug resistance [129]. Members of the type "O" (FOXO) are regulated by the Insulin/PI-3 K/AKT signaling pathway, but are also known for their multiple interconnections within the cellular signaling network [97]. Upon different stimuli, FOXO proteins use their DNA-binding domain in the nucleus to activate or repress target genes [206]. The transcriptional activity is regulated through post-translational modifications such as acetylation, ubiquitination or phosphorylation [214].

The model we analyze consists of 56 species and 135 reactions describing processes such as basal transcription, export, translation, and degradation of RNA and proteins. Figure 22 (A) adapts the FOXO dependent synthesis mechanism. The mRNA-bound FOXO translocates the mRNA units of IsnR (green) and SOD 2 (blue) to the cytoplasm where both target mRNAs are translated to their final protein structure. From the figure is easy to observe that the synthesis processes of IsnR and SOD2 are symmetric. Furthermore, both proteins share the same kinetic parameters in each reaction step. SE equates all the components which are at the same level, i.e., the species enclosed in the dashed boxes of Fig-

[3]This model has been moved to the curated branch under the id *BIOMD0000000705*

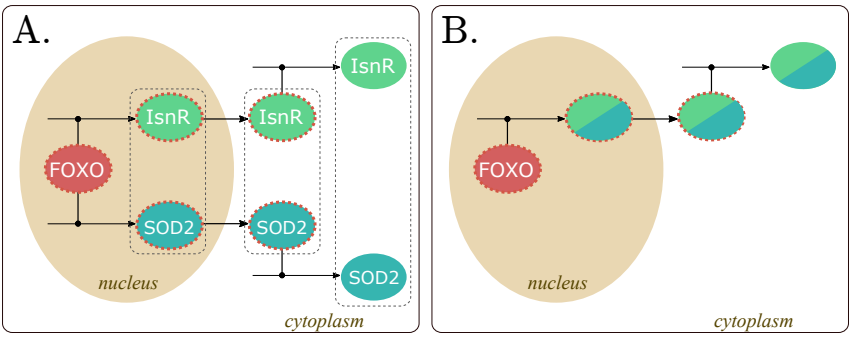


Figure 22: (A) Adaptation of the FOXO dependent protein synthesis mechanism in [188]. Dashed circles represent mRNA-bound molecules. Dashed boxes represent the SE equivalence classes. (B) SE reduced model.

ure 22 (A). This produces a reduced model in which the two synthesis mechanisms are collapsed into a single pathway where FOXO mediates the transcription and synthesis of a macrovariable (blue/green circles) given by the sum of its lumped IsnR and SOD2 members.

4.4 Concluding remarks

In the preprocessing phase (Section 4.1.3), we found 300 models not supported by ERODE. Among the reasons for incompatibility one can mention the use of exponential expressions in rate functions. This is not accepted by FDE/BDE because the underlying theory is not decidable. A workaround has been sketched in [38, 39] and builds on a systematic technique which transforms an initial value problem for an ODE system with derivatives containing rational and exponential expressions into an equivalent problem with polynomial derivatives [92], to which BE and FE can be applied. In future work we plan to implement such a transformation in order to extend the range of applicability of species equivalences.

Overall, we found exact model reductions effective in terms of both the number of cases in which a CRN could be reduced by at least one technique (40% considering both mass-action and non mass-action models) and the overall compression ratio achieved on average (32% for the number of species and 25% for the number of reactions). Unfortunately, the analysis of FDE on a rather appreciable number of models (196) was not conclusive due to timeouts, because of the relative complexity of the SMT checks that are required. This challenges the practical applicability of FDE to realistic case studies (BDE, on the other hand, timed out only twice in our tests whereas BE, FE, and SE are supported by minimization algorithms that enjoy polynomial time and space complexity), prompting alternative approaches to computing FDE, for example by parallelizing the computations, or randomized algorithms.

In the selected case studies herein presented, the exact model reductions have revealed that symmetries in certain signalling pathways carry over to equivalences at the level of the underlying quantitative semantics. Given their moderate size, the considered models would be computationally treatable even without reduction. However, the equivalences can be used as an aid in developing more complex models where such symmetries are present in some components. In addition, we remark that exact model reduction can still be useful when the complexity is due to the many repetitions that are required (e.g., for sensitivity analysis or for simulation with tight confidence intervals) or for particularly difficult analyses such as parametric inference [183].

This empirical study suggests potential benefits in the application of exact model reduction techniques in biological models. This motivates the development of the translator tool into a more mature tool to be further integrated with BioModels/SBML. The availability of ready-to-use model conversions in a simple CRN format such as ERODE's might stimulate similar assessments with other model reduction techniques (e.g., [18, 37]).

At the moment we focus on reducing models with parameterizations given as in the respective original publications. If we wish to draw more general conclusions about the relevance of the reductions and the pres-

ence of certain symmetrical patterns in signalling pathways, it becomes important to test their *robustness* with respect to the model parameters. Theoretically, this does not seem to be particularly difficult, at least for CRNs with deterministic semantics. For example, model parameters could be interpreted as further variables in the SMT formulas used for checking FDE and BDE. Such an extension is currently not implemented in ERODE and is subject to the aforementioned caveats about the scalability of SMT-based reduction techniques, hence left for future work.

Chapter 5

Computation of equilibrium points in biochemical regulatory networks

In systems biology, the analysis of the equilibrium points of ODE systems that represent biochemical reaction networks carries important physical implications that are related to various regulatory phenomena; because of this, it is an area that has received considerable attention [121, 202]. Here we discuss a method for the computation of equilibrium points, with the guarantee of yielding the unique equilibrium of the ODE under defined conditions.[1]. We provide a formal introduction of our technique and show its performance through the evaluation of two models of signaling pathways.

[1] The mathematical framework of this chapter is as in the original paper. For matters of completeness, we also included the corresponding proofs, which are the result of the work of my co-author M. Tschaikowski.

5.1 Equilibria in signaling networks

Here we shall present the main ideas using a running example. We consider a computational model which can be schematically represented as the regulatory network (RN) depicted in Figure 23 (A). The three levels of the cascade are highlighted using different colored boxes. Each level consists of a cycle involving two or more interconvertible mass-preserving forms of a species.

The dynamics of the model is described as follows. A signal (S) mediates the activation of species X_i into X_i^*. The newly activated species mediates the activation of the subsequent species downstream X_{i+1}. This sequence repeats n times, until the activation of the last species in the pathway namely X_n^*, which is responsible for triggering a cellular response R. A negative feedback loop from the last active species acts as a regulator of the pathway by inhibiting S.

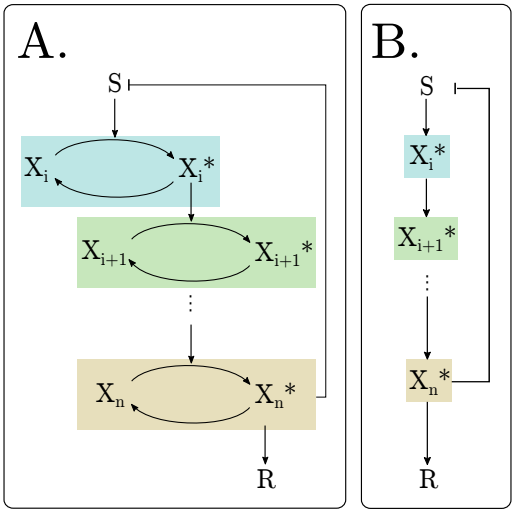


Figure 23: (A) Network representing a phosphorylation cascade in a signaling network. The number of levels is given by $i = 0, 1, \ldots, n$. The most active form of each species is indicated by the symbol $*$ next to the species identifier. (B) Signaling network where each level in (a) is represented using the most-active species

We decompose the behavior into a local dynamics that describes the evolution between the different forms of a given component (i.e., a gene, a protein, a metabolite, and so on) and the network dynamics that describes the influence (that is, activation or inhibition) that one form of each such component exerts on the other components in the system. In doing so, we are essentially following a standard biochemical representation, akin to that used, among others, by Cardelli [31] as well as Tyson and Novak [202], remarking that we are not committing to specific kinetic mechanisms.

Overall, this leads to a more abstract graphical representation of the RN (that will be mathematically formalized later) as depicted in Figure 23 (B), where we replace each box of Figure 23 (A) with a single vertex (labeled with the most active form of the component) and connect the vertices through activation or inhibition arcs. The vertex label indicates the *observable*, that is, the ODE variable related to the specific component form (for instance, the most active form) that appears in the ODEs of the other components.

Our main contribution is a graph-theoretic analysis of the RN for the computation of the equilibrium points of its associated ODE system. To see this with our example, let us make the assumption of mass-action dynamics. Then, the ODE equations associated with the vertices labeled S and X_i^* are thus:

$$\frac{d[S]}{dt} = \lambda - \beta_{11}[X_n^*](t)S(t) - \gamma[S](t) \tag{5.1}$$

$$\frac{d[X_i]}{dt} = \beta_2[X_i^*](t) - \alpha_1[X_i](t)[S](t) \tag{5.2}$$

$$\frac{d[X_i^*]}{dt} = \alpha_1[X_i](t)[S](t) - \beta_2[X_i^*](t) \tag{5.3}$$

where, following standard notation, species names in square brackets refer to concentrations. In these equations, α_i, λ, β_i, and γ are positive parameters, respectively (in particular, we assume that the signal arrives to the system at a constant rate).

To cope with the length of some expressions, in the remainder of this chapter we use $\dot{A}(t)$ instead dA/dt to indicate the derivative of A.

5.1.1 *Open* and *closed*-loop networks

For simplicity, let us assume that our example CRN has three levels and the least-active species in each level are X_0, X_1 and X_2. The key idea of our method builds upon [4, 177] and consists in transforming a *closed-loop* RN into an *open-loop* one by suppressing one or several edges of the original network. Vertices whose incoming edges are suppressed are treated as *input* vertices. Figure 24 visualizes this process in the case of our running example.

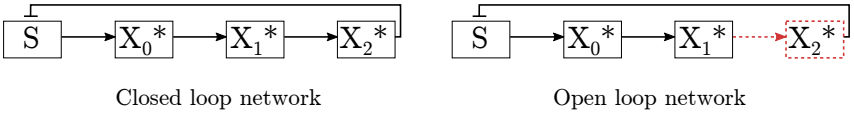


Figure 24: The feedback from X_1^* to X_2^* is suppressed and vertex X_2^* is interpreted as an input vertex.

Under certain assumptions, the dynamical system underlying the open loop network has a unique equilibrium and can be efficiently computed *whenever a constant input is provided*. For instance, by removing the feedback loop from X_2^* to S, we treat $[X_2^*]$ as a constant input in the ODE (5.1) of S. With this, we can compute the equilibrium of S by setting its derivative to zero as

$$S = \frac{\lambda}{\beta_{11}[X_2^*] + \gamma} =: \phi_S([X_2^*]) \tag{5.4}$$

where the ϕ function maps the input $[X_2^*]$ to the unique equilibrium point of $[S]$. We can continue this reasoning along the pathway, by interpreting $[S]$ as a constant input in the ODEs (5.2)-(5.3) of X_1 and X_1^*. This leads to the equilibrium

$$[X_1^*] = \frac{\alpha_1 c_{X_1^*}}{\alpha_1[S] + \beta_2} =: \phi_{X_1^*}([S]), \tag{5.5}$$

where we have set

$$c_{X_1^*} := [X_1](0) + [X_1^*](0)$$

thanks to the mass conservation property $[\dot{X_1}](t) + [\dot{X_1^*}](t) = 0$. By continuing in a similar fashion, one can derive the equilibrium formula $\phi_{X_2^*}([X_1^*])$.

While the above discussion implies that the equilibrium point of the open-loop network given in Figure 24 can be computed efficiently, its relation to the equilibria of the original closed-loop network in Figure 24 is not obvious. We can address this problem by *closing* the open loop network by reactivating the previously suppressed edge. Formally, this corresponds to the feedback condition $[X_2^*] = \phi_{X_2^*}([X_1^*])$, where $\phi_{[X_2^*]}([X_1^*])$ denotes the equilibrium value of X_2^* in the case of a given constant input X_1^*. Overall, we obtain the fixed-point equation

$$\begin{aligned}[X_2^*] &= \phi_{X_2^*}\Big(\phi_{X_1^*}\Big(\phi_{X_0^*}\left(\phi_S([X_2^*])\right)\Big)\Big) \\ &=: f_{X_2^*}([X_2^*]) \end{aligned} \tag{5.6}$$

Therefore, the equilibria of the original ODEs system stands in an one-to-one correspondence with the solution of the *single* nonlinear equation (5.6). More formally, if $[X_2^*]$ is the value of the X_2^*-coordinate of an equilibrium of the ODE system, then $[X_2^*]$ solves (5.6). Conversely, every solution of (5.6) induces the equilibria $[S]$, $[X_0^*]$, $[X_1^*]$ and $[X_2^*]$ for the ODEs of S, X_0^*, X_1^* and X_2^*, respectively, through the functions ϕ_S, $\phi_{X_0^*}$, and $\phi_{X_1^*}$. Together with the equilibrium conditions imposed by the ODEs and the conservation of mass, this determines the equilibria of the remaining five ODEs.

The arguments from above have been already observed in [176, 177]. More specifically, in [176, 177] the above discussion is generalized and the equilibria of a nonlinear ODE system is expressed as fixed points of a nonlinear vector function of smaller size. Unfortunately, the computation of the solution set of a system of nonlinear equations is computationally prohibitive and does not scale, see concluding remarks of [176].

We address this problem by solving the system of nonlinear equations via a fixed-point iteration algorithm. More specifically, we identify graph-theoretic conditions under which an RN induces a fixed-point equation $x = f(x)$ such that f is *anti-monotonic*, i.e., $x \leq x'$ implies $f(x') \leq f(x)$, where $\leq$ is to be interpreted component wise if x and $f(x)$

are vectors. For instance, let us denote by $c_{X_2^*}$ the maximal concentration attainable by X_2 in any of its forms. Then, for any $0 \leq x \leq x' \leq c_{X_2^*}$, where it holds that $f_{X_2^*}(x') \leq f_{X_2^*}(x)$. In the case of the running example, this should is not a surprise because it essentially states that an increase in inhibition leads to a decrease in activation.

While anti-monotonicity does not imply in general that a fixed-point iteration $x, f(x), f(f(x)), \ldots$ converges to a fixed-point of f, it ensures that the composition $g := f \circ f$ is *monotonic*, i.e., $0 \leq x \leq x' \leq c$ implies $g(x) \leq g(x')$, with c being the vector of all maximal attainable species concentrations. This and Kleenes's fixed-point theorem ensure that the sequences $0, g(0), g(g(0)), \ldots$ and $c, g(c), g(g(c)), \ldots$ converge to the least and the greatest fixed-point of g, respectively.

Noting that any fixed-point of f has to be necessarily a fixed-point of $g = f \circ f$, this can be used to prove the following two statements:

1. Let $x^\perp$ and $x^\top$ be the least and greatest fixed-point of g, respectively. Then, any equilibrium of the ODE system underlying the RN is contained in $[x^\perp; x^\top]$.

2. If $x^* = x^\perp = x^\top$, then x^* is the unique equilibrium of the ODE system underlying the RN.

The first statement allows one to efficiently estimate the set of equilibria of the ODE system, while the second statement allows one to decide whether the ODE system has a unique equilibrium and, in the case it does, allows for an efficient computation of it. If applied to the running example, the above discussion ensures that we obtain the unique solution of the nonlinear equation (5.6) when the sequences

$$0, \qquad g_{X_2^*}(0), \qquad g_{X_2^*}\big(g_{X_2^*}(0)\big), \qquad \ldots$$

and

$$c_{X_2^*}, \qquad g_{X_2^*}(c_{X_2^*}), \qquad g_{X_2^*}\big(g_{X_2^*}(c_{X_2^*})\big), \qquad \ldots$$

converge to the same value, with $g_{X_2^*} := f_{X_2^*} \circ f_{X_2^*}$.

5.2 Formal definition of the RN semantics

We can now formalize the ideas and results presented in the previous section, to this end, we start by fixing the notation for a RN. Lets denote by $\mathbb{R}^I$ vectors with index set I. For $x \in \mathbb{R}^I$ and $I_0 \subseteq I$, the restriction of x to the index set I_0 is denoted by $x_{|I_0}$.

Definition 10 (Regulatory Network)**.** *A Regulatory Network is a directed graph (V, E) where V is the set of vertices and E is the set of labeled edges, i.e., $E \subseteq V \times \{+, -\} \times V$.*

For each RN:

1. *The set of all outgoing neighbors of vertex $i \in V$ is denoted by $\mathtt{out}(i)$, that is $\mathtt{out}(i) = \{j \in V \mid (i, \cdot, j) \in E\}$. Similarly, $\mathtt{in}(i)$ denotes all incoming neighbors of i.*

2. *A vertex is called* activator *(resp.,* inhibitor*) if all its outgoing edges are activating (resp., inhibiting). We let V^+ and V^- be the sets of activator and inhibitor vertices, respectively.*

3. *The set of inhibitors with incoming edges defines the set of core observables $\mathcal{I}$, that is, $\mathcal{I} = \{i \in V^- \mid \mathtt{in}(i) \neq \emptyset\}$.*

The symbols in the label set of the edges denote activation and inhibition in the obvious way. The set of core observables $\mathcal{I}$ corresponds essentially to those vertices whose incoming edges are all suppressed and which act as exogenous inputs in the open network. Throughout the remainder of this thesis we exclude the case of vertices with no outgoing edges. For a more compact notation, the set of vertices is assumed to be $V = \{1, 2, \ldots, n\}$ for some $n \geq 1$.

We now define the semantics of a RN in terms of a system of coupled ODEs. In particular, with a given vertex i we associate a set of m_i ODEs over variables x_i^1, …, $x_i^{m_i}$. Essentially, each of the m_i variables x_i^1, …, $x_i^{m_i}$ related to a vertex i may represent the different forms that a component can exhibit, e.g., X_1 and X_1^* in Figure 23 (A). One such variable (i.e., the first component x_i^1 without loss of generality) to represent the *observable* — the only variable that may appear in the set of ODE related to the other vertices; this formalizes the idea of the labels used in the pictorial representation of the RN of, for example, Figure 23 (B).

Definition 11 (RN Semantics)**.** *Given a RN, its underlying ODE is given by associating each vertex $i \in V$ with a system of $m_i \geq 1$ ODEs over variables $x_i^1, \ldots, x_i^{m_i}$. The ODE system associated with vertex i is given by*

$$\dot{\vec{x}}_i = F_i(x_1^1, \ldots, x_{i-1}^1, \vec{x}_i, x_{i+1}^1, \ldots, x_n^1)$$

with $\vec{x}_j := (x_j^1, \ldots, x_j^{m_j})$ for $1 \leq j \leq n$. The ODE system underlying the RN is given by

$$\dot{\vec{x}} = F(\vec{x}), \qquad \text{where } \vec{x} = (\vec{x}_1, \ldots, \vec{x}_n) \text{ and } F = (F_1, \ldots, F_n).$$

The initial condition of the RN is denoted by $\vec{x}(0)$ and assumed to be non-negative. Moreover, we assume that the solution of $\dot{\vec{x}} = F(\vec{x})$ remains non-negative if initialized with a non-negative initial condition.

In the remainder, we avoid the use of the superscript for the observable of vertex i, i.e., we write x_i to indicate x_i^1.

5.2.1 *Well-Posed* RN

We next introduce well-posed RNs. Intuitively, this identifies a local property of the network whereby the ODE system associated with each vertex enjoys a unique equilibrium when the observables of the other vertices that act as inhibitors or activators are treated as exogenous constant inputs.

Definition 12 (Well-Posed RN)**.** *An RN (V, E) is called* well-posed *when the following conditions are satisfied.*

1) *V^+ and V^- form a partition of V, i.e., there are no vertices that are both activators and inhibitors.*

2) *For every $i \in V$, the maximal value of x_i attainable across all non-negative initial conditions $\hat{\vec{x}}(0)$ satisfying $\|\hat{\vec{x}}(0)\|_1 = \|\vec{x}(0)\|_1$ exists and is denoted by c_i.*[2]

3) *For every vertex $i \in V$, define*

   - *$\mathcal{U}_i^+ := \prod_{j \in V^+ \cap in(i)} [0; c_j]$ the set of admissible activation inputs; and*

[2]Following standard notation, $\|\cdot\|_1$ denotes the L^1 norm.

- $\mathcal{U}_i^- := \prod_{j \in V^- \cap \mathtt{in}(i)} [0; c_j]$ *the set of admissible inhibition inputs.*

Then, for every $i \in V$, $u^+ \in \mathcal{U}_i^+$ *and* $u^- \in \mathcal{U}_i^-$, *it must hold that the equilibrium equation*

$$0 = F_i(x_1, \ldots, x_{i-1}, \vec{x}_i, x_{i+1}, \ldots, x_n)$$

admits a non-negative solution that satisfies $x_j = u_j^+$ *for all* $j \in V^+ \cap \mathtt{in}(i)$ *and* $x_k = u_k^-$ *for all* $k \in V^- \cap \mathtt{in}(i)$.

Moreover, the equilibrium equation must characterize the value $\vec{x}_i$. *That is, given two non-negative solutions*

$$0 = F_i(x_1, \ldots, x_{i-1}, \vec{x}_i, x_{i+1}, \ldots, x_n)$$
$$0 = F_i(\hat{x}_1, \ldots, \hat{x}_{i-1}, \hat{\vec{x}}_i, \hat{x}_{i+1}, \ldots, \hat{x}_n)$$

satisfying $x_j = \hat{x}_j = u_j^+$ *for all* $j \in V^+ \cap \mathtt{in}(i)$ *and* $x_k = \hat{x}_k = u_k^-$ *for all* $k \in V^- \cap \mathtt{in}(i)$, *then* $\vec{x}_i = \hat{\vec{x}}_i$.

The uniqueness property ensures the well-definedness of $\phi_i(u^+, u^-) := x_i$.

Essentially, the equilibrium of vertex i is uniquely determined in the case one is given constant activation and inhibition inputs u^+ and u^-, respectively. Since inputs are described by the values of the observables $x = (x_1, \ldots, x_n)$, one can use x instead of the redundant (at least as far the steady-state regime is concerned) vector $\vec{x}$.

For instance, in the case of the running example from Figure 23, equations (5.4) and (5.5) show that the vertices associated with S and X_1^* in Figure 23 (B) do satisfy the requirement above.

5.2.2 Anti-monotonic RN

Armed with the notion of well-posedness, we are ready to introduce anti-monotonic RNs, the core concept of this approach. The graph-theoretic conditions describing an anti-monotonic network ensure that the corresponding nonlinear fixed-point equation characterizing the equilibria, $y = f(y)$, can be constructed and is anti-monotonic, that is, it satisfies $f(y') \leq f(y)$ for all $y \leq y'$.

Definition 13 (Anti-monotonic RN)**.** *A well-posed network* (V, E) *is called* anti-monotonic *if the following properties are satisfied.*

$i)$ *There are no inhibitors that are inhibited.*

$ii)$ *All vertices with incoming neighbors can be reached from the set of core observables* $\mathcal{I}$*.*

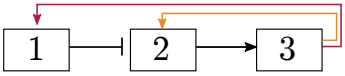
$iii)$ *The graph which arises from* (V, E) *by removing all inhibitors is acyclic, that is,* $(V^+, E \setminus (V^- \times \{-,+\} \times V \cup V \times \{-,+\} \times V^-))$ *is acyclic.*

$iv)$ *With* $\phi_i(u^+, u^-)$ *being as in Definition 12,* ϕ_i *is a continuous function that is monotonic and anti-monotonic in* u^+ *and* u^-*, respectively. That is:*

- ϕ_i *is continuous as a function of* $(u^+, u^-) \in \mathcal{U}_i^+ \times \mathcal{U}_i^-$*;*
- $\phi_i(u^-, u^+) \leq \phi_i(u^-, \hat{u}^+)$ *if* $u^- \in \mathcal{U}_i^-$ *and* $u^+, \hat{u}^+ \in \mathcal{U}_i^+$ *with* $u^+ \leq \hat{u}^+$*;*
- $\phi_i(u^-, u^+) \geq \phi_i(\hat{u}^-, u^+)$ *if* $u^-, \hat{u}^- \in \mathcal{U}_i^-$ *and* $u^+ \in \mathcal{U}_i^+$ *with* $u^- \leq \hat{u}^-$*.*

5.2.3 Algebraic manipulations over the RN

We first remark that $ii)$ requires that all $i \notin \mathcal{I}$ depend on the core observables $\mathcal{I}$. Intuitively, this is needed to ensure that the equilibrium of the opened network can be computed from the values assigned to $\mathcal{I}$. Condition $iii)$, instead, is needed to exclude dependency deadlocks. For instance, the network depicted in Figure 25 cannot be handled because the equilibrium of 2 depends on the equilibrium of 3 and vice versa. Overall, $ii)$ and $iii)$ ensure that f in the fixed-point equation $y = f(y)$ is well-defined, while condition $i)$ and $iv)$ imply that f is anti-monotonic.



Anti-monotonic RN

Figure 25: An RN that is not anti-monotonic because is violates condition $iii)$.

The graph-theoretic conditions $i) - iii)$ boil down to verifying that a subgraph of (V, E) has no loops, a well-known problem that can be

solved in polynomial time. Instead, condition $iv)$ depends on the actual ODE system underlying the RN. For instance, recall that in the case of the example network from Figure 23 (B), we have $\phi_{X_0^*}([S]) = \alpha_1 c_{X_0^*}/(\alpha_1[S]+\beta_2)$. Hence, (5.5) fulfills condition $iv)$. More in general, the following can be proven.

Lemma 1. *The example network from Figure 23 (B) satisfies $i) - iv)$.*

Lemma 1. Since ϕ_S and $\phi_{X_0^*}$ have been covered in Section 5.1, we are left to derive the expressions $\phi_{X_1^*}$ and $\phi_{X_2^*}$. Noting that $\phi_{X_2^*}$ is an instance of $\phi_{X_1^*}$ with modified activation and inhibition parameters, it suffices to focus on $\phi_{X_1^*}$ only. Standard algebraic manipulations reveal that

$$\phi_{X_1^*}([X_0^*]) = c_{X_1^*} \frac{\alpha_3\alpha_4[X_0^*]^2}{\alpha_3\alpha_4[X_0^*]^2 + \alpha_3\beta_4[X_0^*] + \beta_5\beta_6}$$

For completeness, we remark here that

$$\phi_{X_1}([X_0^*]) = c_{X_1^*} \frac{\beta_5\beta_6}{\alpha_3\alpha_4[X_0^*]^2 + \alpha_3\beta_4[X_0^*] + \beta_5\beta_6}$$
$$\phi_{X_1^*}([X_0^*]) = c_{X_1^*} \frac{\alpha_3\beta_4[X_0^*]}{\alpha_3\alpha_4[X_0^*]^2 + \alpha_3\beta_4[X_0^*] + \beta_5\beta_6}$$

Since α_i and β_i are positive, it suffices to show that the function $h(y) := a_1y^2/(a_1y^2 + a_2y + a_3)$ is monotonic in y when $a_1, a_2, a_3 > 0$. A differentiation of h with respect to y yields

$$(\partial_y h)(y) = \frac{2a_1y}{a_1y^2 + a_2y + a_3} - \frac{a_1y^2(a_2 + 2a_1y)}{(a_1y^2 + a_2y + a_3)^2}$$

Algebraic manipulations reveal that $\partial_y h$ has the roots $y = 0$ and $y = -\frac{2a_3}{a_2} < 0$. Since h is non-negative and $h(0) = 0$, we infer the claim. □

In our framework, we suppress the incoming edges of the core observables $\mathcal{I}$ and show that the equilibrium of the so-obtained open-loop network can be computed from the values assigned to $\mathcal{I}$. To do so, we will work with three index sets. First, there is the full vector $\vec{x}$. The subvector x of $\vec{x}$ which lives in $\mathbb{R}^V$ and tracks the observables. Finally, $x_{|\mathcal{I}}$ provides only the values of those vertices who become dangling in the open the network. Our theorems work on this coarsest set $\mathcal{I}$ and show

that the knowledge of those values characterizes the equilibria $\vec{x}$ of the ODE system $\dot{\vec{x}} = F(\vec{x})$ in full.

In particular, the next result ensures that Algorithm 1 computes, for a given vector $y \in \prod_{i\in\mathcal{I}}[0; c_i]$, a value $f(y) \in \prod_{i\in V}[0; c_i]$ such that the fixed-points of $y \mapsto f_{|\mathcal{I}}(y)$ stand in an one-to-one correspondence with the equilibria of the ODE system $\dot{\vec{x}} = F(\vec{x})$ of the RN. Thus, if $y = f_{|\mathcal{I}}(y)$, then the unique equilibrium $\vec{x}$ associated to y can be computed from y. For instance, in the case of the network from Figure 23 (B), $\mathcal{I} = \{X_2^*\}$ and $y \mapsto f_{|\mathcal{I}}(y)$ rewrites to $[X_2^*] \mapsto f_{X_2^*}([X_2^*])$, thus resembling (5.6). The function value $f([X_2^*])$, instead, has the four components

$$\begin{aligned}
f_S([X_2^*]) &= \phi_S([X_2^*]) \\
f_{X_0^*}([X_2^*]) &= \phi_{X_0^*}(f_S([X_2^*])) \\
f_{X_1^*}([X_2^*]) &= \phi_{X_1^*}(f_{X_0^*}([X_2^*])) \\
f_{X_2^*}([X_2^*]) &= \phi_{X_2^*}(f_{X_1^*}([X_2^*])),
\end{aligned}$$

where ϕ_S and $\phi_{X_0^*}$ are given in 5.4 and 5.5, respectively (while $\phi_{X_1^*p}$ and $\phi_{X_2^*}$ are derived in the proof of Lemma 1). If $[X_2^*]$ satisfies $[X_2^*] = f_{X_2^*}([X_2^*])$, we set

$$\begin{aligned}
[S] &:= f_S([X_2^*]) & [X_0^*] &:= f_{X_0^*}([X_2^*]) \\
[X_1^*] &:= f_{X_1^*p}([X_2^*]) & [X_2^*] &:= f_{X_2^*}([X_2^*])
\end{aligned}$$

With this, the remaining components of the equilibrium, $[X_0]$, $[X_1]$, and $[X_2]$, are determined by the corresponding equilibrium equations. In the case of $[X_0]$ and $[X_0^*]$ for instance, it holds that (see proof of Lemma 1)

$$\begin{aligned}
[X_1] &= c_{X_1^*p}\frac{\beta_5\beta_6}{\alpha_3\alpha_4[X_0^*]^2 + \alpha_3\beta_4[X_0^*] + \beta_5\beta_6} \\
[X_1^*] &= c_{X_1^*p}\frac{\alpha_3\alpha_4[X_0^*]^2}{\alpha_3\alpha_4[X_0^*]^2 + \alpha_3\beta_4[X_0^*] + \beta_5\beta_6},
\end{aligned}$$

where $c_{X_1^*} = [X_1](0) + [X_1^*](0)$. Noting that the above values $[X_1]$ and $[X_1^*]$ depend only on $[X_0^*]$, we observe that the equilibrium of vertex X_1^* is fully determined by $[X_0^*]$. Hence, vertex $[X_0^*]$ satisfies condition 3) of Definition 12 which requires that the equilibrium equation of vertex

$i \in V$,

$$0 = F_i(x_1, \ldots, x_{i-1}, \vec{x}_i, x_{i+1}, \ldots, x_n),$$

characterizes the value of $\vec{x}_i$ (with respect to the given activation and inhibition inputs).

Theorem 2. *Assume that (V, E) is an anti-monotonic RN. Then, Algorithm 1 computes, for any vector $y \in \prod_{i\in\mathcal{I}}[0; c_i]$, a value $f(y) \in \prod_{i\in V}[0; c_i]$. The function f enjoys the following properties.*

- *It is anti-monotonic, i.e., $0 \leq y \leq y' \leq c_{|\mathcal{I}}$ implies $0 \leq f(y') \leq f(y) \leq c$.*
- *There is an one-to-one correspondence between the equilibria of F and the fixed-points of $f_{|\mathcal{I}}$. More formally*
  - *If $\vec{x}$ is such that $0 = F(\vec{x})$, then $f(\vec{x}_{|\mathcal{I}}) = \vec{x}_{|\mathcal{I}}$.*
  - *If $y \in \prod_{i\in\mathcal{I}}[0; c_i]$ satisfies $f_{|\mathcal{I}}(y) = y$, then there exists a unique $\vec{x}$ such that that $0 = F(\vec{x})$ and $y = \vec{x}_{|\mathcal{I}}$. Moreover, $f(y) = \vec{x}_{|V}$.*

The above result ensures that it suffices to find all fixed-points of $f_{|\mathcal{I}}$ in order to determine the equilibria of $\dot{\vec{x}} = F(\vec{x})$. In particular, given a fixed-point y of $f_{|\mathcal{I}}$, the vector $f(y)$ provides the observables V with values, i.e., $x_i = f_i(y)$. Hence, $\vec{x}_i$ can be obtained by solving the equilibrium equations, see discussion preceding Theorem 2.

5.2.4 Computation of the anti-monotonic function underlying an anti-monotonic RN

As for Algorithm 1, any vector $y \in \prod_{i\in\mathcal{I}}[0; c_i]$ is processed in three stages. In the first stage, the for loop from line 2, the algorithm sets the f value of each vertex in $\mathcal{I}$ (i.e., $f_i(y) := y_i$ for all $i \in \mathcal{I}$) and computes the f values of all vertices that have no incoming neighbors (note that the equilibrium function of any vertex $i \in V$ with $\texttt{in}(i) = \emptyset$ has no inputs, hence ϕ_i is merely a constant).

In the second stage, the algorithm computes the f values of all remaining vertices, that is vertices that are activators with at least one incoming neighbor. The underlying computation is carried out using the while loop from line 10. This is because the f values in question have

Algorithm 1 Computation of the anti-monotonic function f underlying an anti-monotonic RN.

Require: Anti-monotonic network (V, E) and $y \in \prod_{i \in \mathcal{I}} [0; c_i]$
set $\texttt{done}$ to $V^- \cup \{i \in V^+ \mid \texttt{in}(i) = \emptyset\}$
for all $i \in \texttt{done}$ **do**
  if $i \in \mathcal{I}$ **then**
    set $f_i(x)$ to y_i
  else
    set $f_i(x)$ to ϕ_i
  end if
end for
set $\texttt{left}$ to $\bigcup_{i \in \texttt{done}} \texttt{out}(i)$
while $\texttt{left} \neq \emptyset$ **do**
  set $\texttt{tmp}$ to $\texttt{left}$
  for all $i \in \texttt{tmp}$ **do**
    if $\texttt{in}(i) \subseteq \texttt{done}$ **then**
      set $f_i(x)$ to $\phi_i(u^+, u^-)$, where u^+ and u^- are computed from $\{f_j(x) \mid j \in \texttt{in}(i)\}$
      set $\texttt{done}$ to $\texttt{done} \cup \{i\}$
      set $\texttt{left}$ to $\texttt{left} \cup (\texttt{out}(i) \setminus \texttt{done})$
    end if
  end for
end while
for all $i \in \mathcal{I}$ **do**
  set $f_i(x)$ to $\phi_i(u^+)$, where u^+ is computed from $\{f_j(x) \mid j \in \texttt{in}(i)\}$
end for
return $f(x) \in \prod_{i \in V} [0; c_i]$

to be computed in a specific order that is dictated by the graph. For instance, in the case of the network from Figure 23 (B), $f_S([X_2^*])$ has to be computed before $f_{X_0^*}([X_2^*])$ can be computed. Because of this, $f_{X_0^*}([X_2^*])$ is computed during the second iteration, while $f_S([X_2^*])$ is obtained in the first iteration.

The third and final stage of the algorithm is given by the for loop in line 20. There, the algorithm computes the f values of all $i \in \mathcal{I}$ using the previously computed f values. This intuitively corresponds to a closing of the opened network, see Section 5.1. In particular, if the f values computed during the for loop in line 20 coincide with $y \in \prod_{i\in\mathcal{I}}[0; c_i]$, then $y = f_{|\mathcal{I}}(y)$.

If applied to the network from Figure 23 (B), Algorithm 1 repeats the computational steps from Section 5.1 by deriving, in each iteration of the while loop from line 10, $f_S([X_2^*]) := \phi_S([X_2^*])$, $f_{X_0^*}([X_2^*]) := \phi_{X_0^*}(f_S([X_2^*]))$ and $f_{X_1^*}(X_2^*) := \phi_{X_1^*}(f_{X_0^*}([X_2^*]))$, respectively. Instead, the for loop from line 20 sets $f_{X_2^*}([X_2^*]) := \phi_{X_2^*}(f_{X_1^*}([X_2^*]))$.

Armed with Theorem 2, we are in a position to state our main result.

Theorem 3. *Assume that (V, E) is an anti-monotonic RN and let $\mathcal{I}$ and f be as in Theorem 2. Define $g(y) := f_{|\mathcal{I}}(f_{|\mathcal{I}}(y))$ for any $y \in \prod_{i\in\mathcal{I}}[0; c_i]$. Then, the following holds true.*

1) *Function $f_{|\mathcal{I}} : \prod_{i\in\mathcal{I}}[0; c_i] \to \prod_{i\in\mathcal{I}}[0; c_i]$ has at least one fixed-point.*
2) *Function g is monotonic, i.e., $0 \leq y \leq y' \leq c_{|\mathcal{I}}$ implies $0 \leq g(y) \leq g(y') \leq c$.*
3) *Sequence $0_{|\mathcal{I}}, g(0_{|\mathcal{I}}), g(g(0_{|\mathcal{I}})), \ldots$ converges to $x^\perp$, the least fixed-point of g.*
4) *Sequence $c_{|\mathcal{I}}, g(c_{|\mathcal{I}}), g(g(c_{|\mathcal{I}})), \ldots$ tends to $x^\top$, the greatest fixed-point of g.*
5) *If $\vec{x}$ is such that $0 = F(\vec{x})$, then $\vec{x}_{|\mathcal{I}} \in [x^\perp; x^\top]$.*
6) *If $x^* = x^\perp = x^\top$, then $0 = F(\vec{x})$ has a unique solution $\vec{x}$ and $\vec{x}_{|\mathcal{I}} = x^*$.*

Theorem 3 formalizes the claims made in Section 5.1 and provides an analysis framework for the equilibria of an ODE underlying an anti-monotonic RN. Having all these elements, we can formalize the proof for Theorem 2.

Proof. Let us first assume that every iteration of the while loop from line 10 has at least one $i \in \texttt{tmp}$ for which the if-clause in line 13 becomes true. Under this assumption, every iteration of the while loop adds at least one element to `done`, hence Algorithm 1 terminates. We next show that f is anti-monotonic. To this end, let $V_k \subseteq V$ be the set of vertices that is added to `done` during the k-th iteration of the while loop, where $1 \leq k \leq \nu$. Additionally, let V_0 coincide with the value of `done` at line 2. Let $f_i^k(y)$ denote the value assigned to vertex i, $f_i(y)$, at the beginning of the k-th iteration of the while loop in the case where the algorithm has been invoked with $y \in \prod_{i \in \mathcal{I}}[0; c_i]$. Exploiting the fact that $y \leq y'$, we next show that

a) If $i \in V_0 \cup \ldots \cup V_{k-1}$ is an inhibitor, then $f_i^k(y) \leq f_i^k(y')$

b) If $i \in V_0 \cup \ldots \cup V_{k-1}$ is an activator, then $f_i^k(y') \leq f_i^k(y)$

Please note that a) does not contradict the final claim because the values assigned to inhibitors with incoming neighbors will be overwritten in line 21. To see a)-b) for $k = 1$, we first observe that, for any $i \in V_0$ with no incoming neighbors, line 6 sets $f_i(y)$ to the constant value ϕ_i. Hence, $f_i(y') = \phi_i = f_i(y)$ for all such i. Instead, line 4 ensures that $f_i(y) = y_i \leq y'_i = f_i(y')$ for all $i \in \mathcal{I} \subseteq V_0$. This shows a)-b) for $k = 1$. Let us now assume that a)-b) hold true at the beginning of the k-th iteration. The definition of $V_0, \ldots, V_{k-1}$ implies that $\texttt{done} = V_0 \cup \ldots \cup V_{k-1}$ at the beginning of the k-th iteration. This, $iv)$ and the validity of a)-b) at the beginning of the k-th iteration then imply that every vertex added to `done` in line 15 will satisfy a)-b).

Overall, we obtain the validity of a)-b) at the beginning of the $(k+1)$-th iteration. Since the while condition is false at the beginning of the last iteration and $ii)$ holds true, we infer that $V_\nu = \emptyset$ and $V_0 \cup \ldots \cup V_\nu = V$ at the beginning of the last iteration. This shows that all vertices satisfy the inequality conditions of a)-b) after the while loop is completed. Thanks to the fact that inhibitors can only be activated (see i)), this allows us to conclude that the values assigned to all inhibitors i in line 21 satisfy $f_i(y') \leq f_i(y)$.

The above discussion shows that $f_i(y) = y_i$ for all $i \in \mathcal{I}$ if and only if, for every $i \in \mathcal{I}$, the value $f_i(y)$ computed in line 21 coincides with the value $f_i^\nu(y)$. Since $f_i^\nu(y)$ are computed via the functions $(\phi_i)_i$, this and 3) of Definition 12 ensure the one-to-one correspondence.

We are left with showing that every iteration of the while loop features one $i \in \texttt{left}$ such that $\texttt{in}(i) \subseteq \texttt{done}$ holds true in line 13. To see this, let us assume towards a contradiction that the algorithm gets stuck, i.e., during an iteration of the while loop, we end up in the situation where for all $i \in \texttt{left}$ it holds true that $\texttt{in}(i) \not\subseteq \texttt{done}$. With this, fix any $i' \in \texttt{left}$ and let $j_1, \ldots, j_k$ be a path in $V \setminus \texttt{done}$ that connects $\texttt{done}$ to i', that is

- $j_k = i'$
- $\texttt{in}(j_1) \cap \texttt{done} \neq \emptyset$
- $j_l \in V \setminus \texttt{done}$ for all $1 \leq l \leq k$
- $(j_l, j_{l+1}) \in E$ for all $1 \leq l \leq k-1$

The existence of such a path is guaranteed by $ii)$ and $\mathcal{I} \subseteq V_0 \subseteq \texttt{done}$ (also, notice that in the trivial case $\texttt{in}(i') \cap \texttt{done} \neq \emptyset$ we can simply set $k = 1$). Since $\texttt{in}(j_1) \cap \texttt{done} \neq \emptyset$, it must hold that $j_1 \in \texttt{left}$. This, in turn, implies that $\texttt{in}(j_1) \not\subseteq \texttt{done}$, as otherwise the algorithm would not be stuck. Hence, we can pick some $i'' \in \texttt{in}(j_1) \cap (V \setminus \texttt{done})$. Overall, we were able to construct a path from some $i'' \in (V \setminus \texttt{done})$ to $i' \in (V \setminus \texttt{done})$ that remains in $V \setminus \texttt{done}$ and whose length is positive because $i' \neq i''$. By applying the very same argument to i'', we can construct a path from some $i''' \in (V \setminus \texttt{done})$ to $i'' \in (V \setminus \texttt{done})$ that remains in $V \setminus \texttt{done}$ and that has positive length. By repeating the argument at most $|V \setminus \texttt{done}|$ times, we thus can construct a cycle in $(V \setminus \texttt{done}) \subseteq (V \setminus \mathcal{I})$ which contradicts $iii)$. □

We conclude the section by noting that one may be tempted to check $x^{\perp} = x^{\top}$ by solving the ODE system for different initial conditions (that respect the maximal attainable concentration vector c). However, while such ad-hoc approach can be used to discover the presence of different equilibria, it cannot be used to prove their absence because the number of possible initial conditions is infinite.

5.3 Evaluation of the algorithm for computing the equilibria

Together with the equilibria expressions from Section 5.1 and the proof of Lemma 1, we will provide the complete configuration of Algorithm 1 through two case studies.

5.3.1 MAPK signaling cascade

As discussed in previous chapters, MAPK is a pathway that can be started by several receptors upon external stimuli. The signal triggers the consecutive activation of several downstream protein kinases, where the last kinase can promote cellular responses such as growth, development, differentiation, proliferation, inflammation, and aptosis.

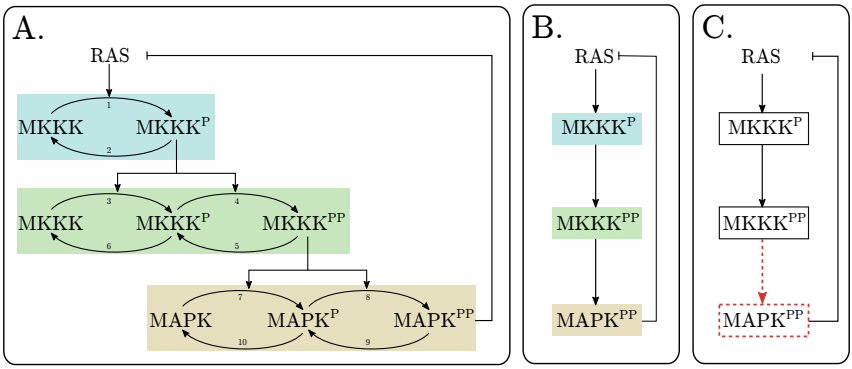


Figure 26: (A). MAPK signaling pathway adapted from [119], where each level of the pathway is represented using a different color. (B) Linear pathway replacing each family of states (e.g., $\{[\mathrm{MKK}], [\mathrm{MKK_P}], [\mathrm{MKK_{PP}}]\}$) by the most active species (e.g., $[\mathrm{MKK_{PP}}]$). (C) Closed loop RN obtained over (B)

Each level in Figure 26 (A) consists of a phosphorylation cycle involving two or more interconvertible mass-preserving forms of a kinase where the suffixes '-P' and '-PP' denote the *partial* and *total* phosphorylated forms of such kinase, respectively. The most active form in each

level regulates its downstream kinases acting as a catalyst in a phosphorylation reaction (e.g., the reaction with label 3). The dephosphorylation process instead occurs spontaneously (e.g., label 6). The dynamical system behind the MAPK schematics in Figure 26 (A) is given by the following nine differential equations:

$$
\begin{aligned}
[\dot{\mathrm{RAS}}](t) &= \lambda - \beta_0[\mathrm{MAPK_{PP}}](t)[\mathrm{RAS}](t) - \gamma[\mathrm{RAS}](t) \\
[\dot{\mathrm{MKKK}}](t) &= \beta_2[\mathrm{MKKK_P}](t) - \alpha_1[\mathrm{MKKK}](t)[\mathrm{RAS}](t) \\
[\dot{\mathrm{MKKK_P}}](t) &= \alpha_1[\mathrm{MKKK}](t)[\mathrm{RAS}](t) - \beta_2[\mathrm{MKKK_P}](t) \\
[\dot{\mathrm{MKK}}](t) &= \beta_6[\mathrm{MKK_P}](t) - \alpha_3[\mathrm{MKKK_P}](t)[\mathrm{MKK}](t) \\
[\dot{\mathrm{MKK_P}}](t) &= \alpha_3[\mathrm{MKKK_P}](t)[\mathrm{MKK}](t) + \beta_5[\mathrm{MKK_P}](t) \\
&\quad - \beta_6[\mathrm{MKK_P}](t) - \alpha_4[\mathrm{MKKK_P}](t)[\mathrm{MKK_P}](t) \\
[\dot{\mathrm{MKK_{PP}}}](t) &= \alpha_4[\mathrm{MKKK_P}](t)[\mathrm{MKK_P}](t) - \beta_5[\mathrm{MKKK_P}](t) \\
[\dot{\mathrm{MAPK}}](t) &= \beta_{10}[\mathrm{MAPK_P}](t) - \alpha_7[\mathrm{MKK_{PP}}](t)[\mathrm{MAPK}](t) \\
[\dot{\mathrm{MAPK_P}}](t) &= \alpha_7[\mathrm{MKK_{PP}}](t)[\mathrm{MAPK}](t) + \beta_9[\mathrm{MAPK_{PP}}](t) \\
&\quad - \alpha_8[\mathrm{MKK_{PP}}](t)[\mathrm{MAPK_P}](t) - \beta_{10}[\mathrm{MAPK_P}](t) \\
[\dot{\mathrm{MAPK_{PP}}}](t) &= \alpha_8[\mathrm{MKK_{PP}}](t)[\mathrm{MAPK_P}](t) - \beta_9[\mathrm{MAPK_{PP}}](t),
\end{aligned}
$$

where the initial conditions are given by

$$
\begin{array}{lll}
[\mathrm{RAS}](0) = 0 & [\mathrm{MKKK}](0) = 90 & [\mathrm{MKKK_P}](0) = 10 \\
[\mathrm{MKK}](0) = 280 & [\mathrm{MKK_P}](0) = 10 & [\mathrm{MKK_{PP}}](0) = 10 \\
[\mathrm{MAPK}](0) = 280 & [\mathrm{MAPK_P}](0) = 10 & [\mathrm{MAPK_{PP}}](0) = 10
\end{array}
$$

Together with the ODEs, the initial conditions led to the following maximal attainable species concentrations:

$$
\begin{aligned}
c_{\mathrm{RAS}} &= \lambda/\gamma \\
c_{\mathrm{MKKK_P}} &= [\mathrm{MKKK}](0) + [\mathrm{MKKK_P}](0) = 100 \\
c_{\mathrm{MKK_{PP}}} &= [\mathrm{MKK}](0) + [\mathrm{MKK_P}](0) + [\mathrm{MKK_{PP}}](0) = 300 \\
c_{\mathrm{MAPK_{PP}}} &= [\mathrm{MAPK}](0) + [\mathrm{MAPK_P}](0) + [\mathrm{MAPK_{PP}}](0) = 300
\end{aligned}
$$

As in Section 5.1, by removing the feedback loop from $\mathrm{MAPK_{PP}}$ to RAS, we treat $[\mathrm{MAPK_{PP}}]$ as a constant input in the ODE of RAS. With this, we can compute the equilibrium of RAS by setting its derivative to zero as

$$[\mathrm{RAS}] = \frac{\lambda}{\beta_{11}[\mathrm{MAPK_{PP}}] + \gamma} =: \phi_{\mathrm{RAS}}([\mathrm{MAPK_{PP}}])$$

Reinterpreting $[\mathrm{RAS}]$ we obtain the equilibrium for $[\mathrm{MKKK_P}]$ given by

$$[\mathrm{MKKK_P}] = \frac{\alpha_1 c_{\mathrm{MKKK_P}}}{\alpha_1[\mathrm{RAS}] + \beta_2} =: \phi_{\mathrm{MKKK_P}}([\mathrm{RAS}])$$

which is consistent with Equation 5.5. By continuing in a similar fashion, one can derive the equilibrium formula $\phi_{\mathrm{MKK_{PP}}}([\mathrm{MKKK_P}])$. We *close* the open loop network in Figure 26 (B) by reactivating the previously suppressed edge (Figure 26 (C)), this corresponds to the feedback condition $[\mathrm{MAPK_{PP}}] = \phi_{\mathrm{MAPK_{PP}}}([\mathrm{MKK_{PP}}])$, where $\phi_{\mathrm{MAPK_{PP}}}([\mathrm{MKK_{PP}}])$ denotes the equilibrium value of $\mathrm{MAPK_{PP}}$ in the case of a given constant input $[\mathrm{MKK_{PP}}]$.

$$\begin{aligned}[\mathrm{MAPK_{PP}}] &= \phi_{\mathrm{MAPK_{PP}}}\Big(\phi_{\mathrm{MKK_{PP}}}\Big(\phi_{\mathrm{MKKK_P}}\left(\phi_{\mathrm{RAS}}([\mathrm{MAPK_{PP}}])\right)\Big)\Big) \\ &=: f_{\mathrm{MAPK_{PP}}}([\mathrm{MAPK_{PP}}]) \end{aligned} \tag{5.7}$$

Therefore, the equilibria of the original ODEs system consisting of *nine* equations stands in a one-to-one correspondence with the solution of the *single* nonlinear equation (5.7).

If applied to the MAPK network from Figure 26 (B), Algorithm 1 repeats the computational steps from Section 5.1 by deriving, in each iteration of the while loop from line 10, $f_{\mathrm{RAS}}([\mathrm{MAPK_{PP}}]) := \phi_{\mathrm{RAS}}([\mathrm{MAPK_{PP}}])$, $f_{\mathrm{MKKK_P}}([\mathrm{MAPK_{PP}}]) := \phi_{\mathrm{MKKK_P}}(f_{\mathrm{RAS}}([\mathrm{MAPK_{PP}}]))$ and $f_{\mathrm{MKK_{PP}}}(\mathrm{MAPK_{PP}})$ $:= \phi_{\mathrm{MKK_{PP}}}(f_{\mathrm{MKKK_P}}([\mathrm{MAPK_{PP}}]))$, respectively. Instead, the for loop from line 20 sets $f_{\mathrm{MAPK_{PP}}}([\mathrm{MAPK_{PP}}]) := \phi_{\mathrm{MAPK_{PP}}}(f_{\mathrm{MKK_{PP}}}([\mathrm{MAPK_{PP}}]))$. Table 6 shows the result of the algorithm in each iteration.

Table 6: Application of Algorithm 1 to the MAPK model. The values are stated with respect to the beginning of the while loop iteration one, two, three, four and line 20, respectively. All function terms are meant to be evaluated with respect to x_4, e.g., $f_1 \equiv f_1(x_4)$, $\phi_1 \equiv \phi_1(x_4)$ and $\phi_2(f_1) \equiv \phi_2(f_1(x_4))$. This is because the feedback from 3 to 4 is suppressed during the while loop computation and 4 is treated as an input vertex, compare Figure 24. However, after line 20 the feedback from vertex 3 to 4 is taken into account, thus allowing one to obtain $f_4(x_4)$ from the values computed by the while loop.

Value	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Line 20
f_1	—	ϕ_1	ϕ_1	ϕ_1	ϕ_1
f_2	—	—	$\phi_2(f_1)$	$\phi_2(f_1)$	$\phi_2(f_1)$
f_3	—	—	—	$\phi_3(f_2)$	$\phi_3(f_2)$
f_4	x_4	x_4	x_4	x_4	$\phi_4(f_3)$
`done`	$\{4\}$	$\{1,4\}$	$\{1,2,4\}$	$\{1,2,3,4\}$	$\{1,2,3,4\}$
`left`	$\{1\}$	$\{2\}$	$\{3\}$	$\emptyset$	$\emptyset$

5.3.2 EGFR signaling pathway

We next discuss a model for the EGFR signaling pathway from [26], whose associated RN and a compact set of labels is given in Figure 27 A and B respectively. From now on, we will use a set of labels to identify the diverse variables from the model. We use a mass-action semantics that follows [31], where each vertex is associated with a variable triplet.

Here we already express it in the more compact notation which exploits the preservation of mass among the three forms. Thus we use two ODE variables x_i^* and x_i, which represent the active and the passive form of each component, respectively, and denote by c_i the total concentration for vertex i. Let us denote by $\Omega_i^+ \subseteq V$ and $\Omega_i^- \subseteq V$ the set of activators and inhibitors of vertex i, respectively. The ODEs are given by:

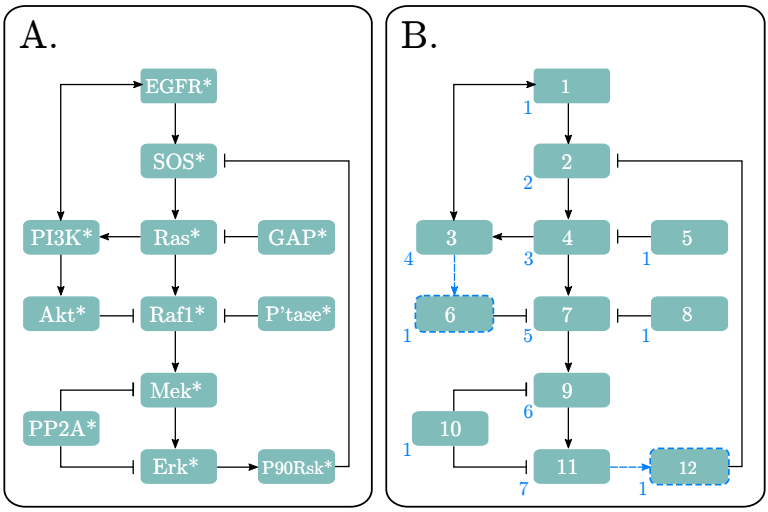


Figure 27: (A) Regulatory network adapted from the EGFR pathway from citeBrown2004. (B) The corresponding open network which arises by suppressing the activations from 3 to 6 and 11 to 12, respectively. The open network has the core observables 6 and 12. The blue numbers give the while loop iteration at whose beginning the corresponding f value becomes available for the first time.

$$\dot{x}_i^* = -\Big(\sum_{I_k \in \Omega_i^-} \beta_{k,i} x_{I_k} + \sum_{A_l \in \Omega_i^+} \alpha_{k,i} (c_{A_l} - x_{A_l}) \Big) x_i^* + \Big(\sum_{A_l \in \Omega_i^+} \alpha_{k,i} x_{A_l} + \sum_{I_k \in \Omega_i^-} \beta_{k,i} (c_{I_k} - x_{I_k}) \Big) (c_i - x_i^* - x_i)$$

$$\dot{x}_i = -\Big(\sum_{A_l \in \Omega_i^+} \alpha_{k,i} x_{A_l} + \sum_{I_k \in \Omega_i^-} \beta_{k,i} (c_{I_k} - x_{I_k}) \Big) x_i + \Big(\sum_{I_k \in \Omega_i^-} \beta_{k,i} x_{I_k} + \sum_{A_l \in \Omega_i^+} \alpha_{k,i} (c_{A_l} - x_{A_l}) \Big) (c_i - x_i^* - x_i) \tag{5.8}$$

where the parameters α and β are positive constants that give the strengths of inhibition and activation, respectively. The set of core ob-

servables is $\mathcal{I} = \{6, 12\}$ and conditions $i) - iii)$ can be easily seen to hold true. Instead, using

$$a_i(A, I) = \sum_{A_l} \alpha_{l,i} x_{A_l} - \sum_{I_k} \beta_{k,i} x_{I_k} + \sum_{I_k} \beta_{k,i} c_{I_k}$$
$$b_i(A, I) = \sum_{I_k} \beta_{k,i} x_{I_k} - \sum_{A_l} \alpha_{l,i} x_{A_l} + \sum_{A_l} \alpha_{l,i} c_{A_l},$$

it can be shown that $\phi_i = c_i a_i^2/(a_i^2 + a_i b_i + b_i^2)$ is the unique equilibrium point of x_i^*. Moreover, it can be proven that the roots of $(\partial_{A_l} \phi_i)(A_l)$ are either less than or equal zero or strictly greater than c_{A_l}. By considering $A_l \mapsto (\partial_{A_l}^2 \phi_i)(A_l)$ at $A_l = 0$, we infer that ϕ_i is monotonic in A_l on $[0; c_{A_l}]$. A similar argumentation ensures that ϕ_i is anti-monotonic in every I_k on $[0; c_{I_k}]$, thus yielding the following.

Lemma 2. *The semantics (5.8) satisfies condition iv).*

The dynamical system is induced by Figure 27 and the semantics discussed in the proof of Lemma 2. For instance, since 2 is inhibited by 12 and activated by 1, the ODEs of 2 are given by

$$\dot{x}_2^* = -(\beta_{12} x_{12}^* + \alpha_1 (c_1 - x_1^*)) x_2^* + (\alpha_1 x_1^* + \beta (c_{12} - x_{12}^*))(c_2 - x_2^* - x_2)$$
$$\dot{x}_2 = -(\alpha_1 x_1^* + \beta_{12} (c_{12} - x_{12}^*)) x_2 + (\beta_{12} x_{12} + \alpha_1 (c_1 - x_1^*))(c_2 - x_2^* - x_2)$$

If a vertex has more than one activator or inhibitor, the above formula has to be adjusted, see proof of Lemma 2. In the case a vertex has no activators, all its α coefficients are set to zero. Likewise, all β coefficients are set to zero if a vertex has no inhibitors. Since this implies that the ODEs of $i \in \{1, 5, 8, 10\}$ are zero, we set $\phi_i := c_i := x_i^*(0)$, with $x_i^*(0)$ being the initial condition from [26] (the computation of ϕ_1 is actually more complicated than that because ϕ_1 is the steady-state of an ODE system that is independent from all other vertices of the regulatory network; ϕ_1 was computed using the parameters and initial conditions from [26]). The initial conditions from [26] and the dynamics outlined above led to the

following maximal attainable concentrations:

$$
\begin{aligned}
c_{\text{EGFR}^*} &= 79955 & c_{\text{SOS}^*} &= 120000 & c_{\text{PI3K}^*} &= 120000 \\
c_{\text{Ras}^*} &= 120000 & c_{\text{GAP}^*} &= 120000 & c_{\text{Akt}^*} &= 120000 \\
c_{\text{Raf1}^*} &= 120000 & c_{\text{Ptase}^*} &= 120000 & c_{\text{MEK}^*} &= 600000 \\
c_{\text{PP2A}^*} &= 12000 & c_{\text{ERK}^*} &= 600000 & c_{\text{P90Rsk}^*} &= 120000
\end{aligned}
$$

Together with the formula $\phi_i = c_i a^2/(a^2 + ab + b^2)$ from the proof of Lemma 2, this provides a complete configuration of Algorithm 1. Table 7 additionally provides the corresponding f values.

Table 7: Application of Algorithm 1 to the model of the epidermal growth factor from Figure 27. The variable values are stated with respect to the beginning of the while loop iteration one, two, three, four and five, respectively. All function terms are meant to be evaluated with respect to x_6 and x_{12}, e.g., $f_2 \equiv f_2(x_6, x_{12})$ and $\phi_2(f_1, f_{12}) \equiv \phi_2(f_1(x_6, x_{12}), f_{12}(x_6, x_{12}))$.

Value	*Iteration 1*	*Iteration 2*	*Iteration 3*	*Iteration 4*	*Iteration 5*	*Iteration 6*
f_1	ϕ_1	ϕ_1	ϕ_1	ϕ_1	ϕ_1	ϕ_1
f_2	—	$\phi_2(f_1, f_{12})$	$\phi_2(f_1, f_{12})$	$\phi_2(f_1, f_{12})$	$\phi_2(f_1, f_{12})$	$\phi_2(f_1, f_{12})$
f_3	—	—	—	$\phi_3(f_1, f_4)$	$\phi_3(f_1, f_4)$	$\phi_3(f_1, f_4)$
f_4	—	—	$\phi_4(f_2, f_5)$	$\phi_4(f_2, f_5)$	$\phi_4(f_2, f_5)$	$\phi_4(f_2, f_5)$
f_5	ϕ_5	ϕ_5	ϕ_5	ϕ_5	ϕ_5	ϕ_5
f_6	x_6	x_6	x_6	x_6	x_6	x_6
f_7	—	—	—	$\phi_7(f_4, f_6, f_8)$	$\phi_7(f_4, f_6, f_8)$	$\phi_7(f_4, f_6, f_8)$
f_8	ϕ_8	ϕ_8	ϕ_8	ϕ_8	ϕ_8	ϕ_8
f_9	—	—	—	—	$\phi_9(f_7, f_{10})$	$\phi_9(f_7, f_{10})$
f_{10}	ϕ_{10}	ϕ_{10}	ϕ_{10}	ϕ_{10}	ϕ_{10}	ϕ_{10}
f_{11}	—	—	—	—	—	$\phi_{11}(f_9, f_{10})$
f_{12}	x_{12}	x_{12}	x_{12}	x_{12}	x_{12}	x_{12}

The order in which the algorithm computes the entries of the vector function f is provided in Figure 27 B (light-blue numbers alongside the nodes). For instance, f_6 and f_{12} are initialized during the for loop in line 2 and are thus available at the beginning of the first while loop iteration. This is because 6 and 12 are treated as inputs during the while loop. The for loop from line 20, instead, takes the suppressed feedbacks to 6 and 12 into account and assigns $f_6 := \phi_6(f_3)$ and $f_{12} := \phi_{12}(f_{11})$. On the other hand, f_2 is computed during the first while loop iteration.

5.4 Concluding remarks

Our method for the computation of equilibrium points in ODE systems encoding RNs is based on the idea of cutting feedback loops in the network, similarly to other approaches in the literature, most notably [4] and [176]. Unlike [4] we do not focus on systems with positive feedback. In fact, while positive feedback is related to the presence of multiple equilibria, we were able to prove that our models of MAPK and EGFR signaling pathways with negative feedback admit a unique equilibrium point.

We evaluated our approach on the MAPK and EGFR model from Figure 26 and Figure 27, respectively. To this end, we ran 100 experiments for each model in which the parameters were uniformly sampled from the compact interval $[1; 1000]$, covering thus a variety of possible scenarios; for both models we used the initial conditions stated in [119] and [26], respectively. In each experiment, the ODE system could be shown to enjoy a unique equilibrium because the least and the greatest fixed-point of g from Theorem 3 were identical.

A possible line of future research is to combine our computational approach with [4, 71]. Specifically, the goal would be to provide sufficient conditions ensuring that the set of equilibria contains, for instance, only attractors. To this end, we plan to relate the anti-monotonic regulatory networks to the graphs that are induced by systems of differential equations in [3–5, 71]. Moreover, we intend to investigate whether the requirement that each vertex is either an inhibitor or activator can be dropped by introducing artificial vertices that preserve equilibria. Finally, aiming at a more efficient computation of equilibria, we also plan to investigate if the current approach can be combined with model reduction techniques such as [33, 36, 38, 39, 57, 77].

Chapter 6

Conclusion

This thesis provides a detailed analysis of techniques for the reduction of quantitative systems of biochemical networks together with the interpretation of the biological and functional characteristics of reductions obtained over a variate collection of case studies.

We provide a framework for the automatic analysis of large-scale quantitative repositories of biological models, with special support for models based in SBML specifications [169]. We consider a family of techniques based on equivalence relations over the species in the network [36, 38, 40], leading to a coarse graining which provides the exact aggregate time-course evolution for each equivalence class. All the reductions were performed in ERODE [35], a state-of-the-art tool for the reduction of dynamical systems in which the selected equivalences are implemented. In addition, we provide a tool for translation of models encoded in System Biology Markup Language (SBML) to CRN-like specifications, which is the input format of ERODE.

For networks with stochastic dynamics, we provide equivalences at the level of the Markov chain that can be applied to biological networks, and complex networks depicting epidemic processes. Here we show that SE can be seen as complementary, exact aggregation method for epidemic processes on complex networks where the maximal SE induces a partition on the nodes of the graph which is a refinement of the orbit

partition, i.e., the partition of nodes where each block is a distinct orbit. Interestingly, we also show that SE can be applied to models that do not satisfy the conditions required in [185].

These exact model reduction techniques produce comprehensible reduced models where the collapse of several species into one may carry a physical interpretation that increases our understanding of the system under study. In the case studies discussed in this dissertation, the exact model reductions revealed symmetries in signaling pathways that carry over to equivalences at the level of the underlying quantitative semantics. Given their moderate size, the considered models would be computationally treatable even without reduction. However, the equivalences can be used as an aid in developing more complex models where such symmetries are present in some components.

In a different stream, we approach model reduction by focusing on the steady-state behavior of the system. To this end, we provide a method for the computation of equilibrium points, with the guarantee of yielding the unique equilibrium of the ODE under defined conditions [170]. We apply this algorithm for the simplification of two signaling pathways [26, 147] characterized by different dynamic mechanisms. The EGFR circuit in [26] models the interaction among the two states (active/inactive) of each molecule involved in the ERK-MAPK and PI3K pathways. Whereas the mechanism in [147] studies in more detail the phosphorylation and dephosphorylation mechanisms in MAPK by building a cascade that includes all the states of MAPK in each level on the pathway. In both cases we were able to collapse the behavior of the intermediate components which allows to characterize the dynamics in a single fixed point equation. Importantly, our algorithm does not require the availability of the Jacobian of the ODE vector field, thus, we are not limited by the system's dimension.

Interestingly, the analysis performed in Chapters 3-5 reveals reduction *motifs* embedded in the structure of the network, i.e., structures that where compressed in each case found.

Overall we could identify 3 functional/structural motifs:

- *Motif 1* is a two-component motif which represents the interaction between two species X and X*, where X* is the active variant of X. This mechanism is a recurrent element in signaling networks where several layers of switch-like proteins regulate the cell's response to external stimuli. Our aggregation techniques collapse the two states of the protein in a macrovariable $\tilde{X}$, represented by the sum of the active and inactive states, i.e. $\tilde{X_{(t)}} = [X]_{(t)} + [X*]_{(t)}$.

- *Motif 2* is a three-component motif depicting the interaction between three species ${X_0}^0, {X_0}^1, {X_1}^0$, so that X can transform into ${X_0}^1$ or ${X_1}^0$. Species X changes to a state where one of its sites is occupied, e.g., by a phosphatase to model X's phosphorylation or by another component to form a complex. In this scenario, our techniques can merge the symmetric species ${}^1_0$ and ${}^0_1$ into a macrovariable $\hat{X}$ such that $\hat{X}_{(t)} = [X^1_0]_{(t)} + [X^0_1]_{(t)}$.

- *Motif 3* is a single-component motif, where we establish relationships among the different states of each binding sites. The existence of multiple binding sites is a characteristic of in several members of signal transduction pathways such as scaffold proteins and multisite receptors. In either way, the multisite property allows to recruit diverse molecules to form molecular complexes with specific functions. The motif that we identified is a multisite receptor with two protein-binding docks (β, γ) and one ligand-binding site (α), i.e., a RTK receptor. Often times the dynamics involved in RTK signaling follows a hierarchical configuration where the propensity of binding adapter species is influenced by the state of the receptor's binding sites. In such circumstances, we preserve the hierarchy in the order of the reactions but we abstract from the different states of the receptor in each level, thus producing a reduced description where each stage is represented by a macrovariable that relates the involved receptor's states.

These findings provide the intuition that our techniques abstract beyond the behavior of the system to capture structural/functional segments of the network that can be simplified with no (or little) effect in the dynamics of the model. This becomes useful when we are dealing with large networks, where to check a priori for the existence of such motifs is less costly than the computation of the complete network.

We plan to study further the network structures of diverse biological systems to characterize a wider selection of reducible motifs. It would be interesting to investigate if these motifs also appear in other networks beyond the scope of systems biology. A natural follow up is of course to build an algorithm to check the presence of reducible motifs in the network which would allow to check if a network is reducible without actually computing the reduction.

Another future line of research is to explore the performance of other lumping schemes for the reduction of ODE systems, for instance improper lumping techniques. The fact that the network transformation is done by means of an improper lumping implies that the original variables can be mapped to more than one macrovariable in the reduced model. As with the exact techniques discussed in this dissertation, we will evaluate the performance of such improper lumping scheme by observing the effectiveness and the intelligibility of the obtained reductions.

Finally, while considerable work has been done on abstraction techniques that preserve certain desired original dynamics in an exact way [57, 85, 200], it is known that such abstractions are sensitive to the specific values of the model parameters. However such values are usually difficult to measure, and are therefore characterized by a considerable degree of uncertainty. Hence, another line of future research is to study the effect of parametric tolerances in the abstraction of biological models.

Bibliography

[1] Luis U Aguilera et al. "A new efficient approach to fit stochastic models on the basis of high-throughput experimental data using a model of IRF7 gene expression as case study". In: *BMC systems biology* 11.1 (2017), p. 26 (cit. on p. 57).

[2] David Angeli and Eduardo Sontag. "Interconnections of monotone systems with steady-state characteristics". In: *Optimal control, stabilization and nonsmooth analysis*. Springer, 2004, pp. 135–154 (cit. on p. xxii).

[3] David Angeli and Eduardo D. Sontag. "Multi-stability in monotone input/output systems". In: *Systems & Control Letters* 51.3-4 (2004), pp. 185–202. DOI: 10.1016/j.sysconle.2003.08.003 (cit. on p. 106).

[4] David Angeli et al. "Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems". In: *Proceedings of the National Academy of Sciences* 101.7 (2004), pp. 1822–1827. DOI: 10.1073/pnas.0308265100 (cit. on pp. 85, 106).

[5] David Angeli et al. "Graph-theoretic characterizations of monotonicity of chemical networks in reaction coordinates". In: *Journal of Mathematical Biology* 61.4 (2010). DOI: 10.1007/s00285-009-0309-0 (cit. on p. 106).

[6] Uri M. Ascher and Linda R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. SIAM, 1988 (cit. on p. 3).

[7] Michael Ashburner et al. "Gene ontology: tool for the unification of biology". In: *Nature genetics* 25.1 (2000), p. 25 (cit. on pp. 60, 62).

[8] Amos Bairoch. "The ENZYME database in 2000". In: *Nucleic acids research* 28.1 (2000), pp. 304–305 (cit. on p. 61).

[9] Sourav Bandyopadhyay et al. “A human MAP kinase interactome”. In: *Nature methods* 7.10 (2010), p. 801 (cit. on p. 3).

[10] Dipak Barua and Byron Goldstein. “A mechanistic model of early FcεRI signaling: lipid rafts and the question of protection from dephosphorylation”. In: *PLoS One* 7.12 (2012), e51669 (cit. on pp. xxi, 38).

[11] Dipak Barua and William S Hlavacek. “Modeling the effect of APC truncation on destruction complex function in colorectal cancer cells”. In: *PLoS computational biology* 9.9 (2013), e1003217 (cit. on p. 38).

[12] Dipak Barua et al. “A bipolar clamp mechanism for activation of Jak-family protein tyrosine kinases”. In: *PLOS Computational Biology* 5.4 (2009), e1000364 (cit. on p. 38).

[13] Fortunato Bianconi et al. “Computational model of EGFR and IGF1R pathways in lung cancer: a systems biology approach for translational oncology”. In: *Biotechnology advances* 30.1 (2012), pp. 142–153 (cit. on p. 77).

[14] M. L. Blinov et al. “BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains”. In: *Bioinformatics* 20.17 (2004), pp. 3289–3291 (cit. on pp. 21, 55).

[15] J. Joseph Blum et al. “A mathematical model quantifying GnRH-induced LH secretion from gonadotropes”. In: *Am J Physiol Endocrinol Metab* 278.2 (2000), E263–272. ISSN: 0193-1849. URL: `http://ajpendo.physiology.org/content/278/2/E263.long` (cit. on pp. xxi, 57).

[16] Marián Boguñá et al. “Models of Social Networks based on Social Distance Attachment”. In: *Phys. Rev. E* 70.5 (2004), p. 056122 (cit. on p. 51).

[17] Paolo Boldi et al. “UbiCrawler: A Scalable Fully Distributed Web Crawler”. In: *Software: Practice & Experience* 34.8 (2004), pp. 711–726 (cit. on p. 51).

[18] Michele Boreale. “Algebra, Coalgebra, and Minimization in Polynomial Differential Equations”. In: *20th International Conference on Foundations of Software Science and Computation Structures (FOSSACS)*. 2017, pp. 71–87 (cit. on p. 80).

[19] Nikolay M Borisov et al. “Domain-oriented reduction of rule-based network models”. In: *IET systems biology* 2.5 (2008), pp. 342–351 (cit. on pp. 38–40).

[20] Nikolay M Borisov et al. “Signaling through receptors and scaffolds: independent interactions reduce combinatorial complexity”. In: *Biophysical journal* 89.2 (2005), pp. 951–966 (cit. on p. 5).

[21] Nikolay M Borisov et al. “Trading the micro-world of combinatorial complexity for the macro-world of protein interaction domains”. In: *Biosystems* 83.2-3 (2006), pp. 152–166 (cit. on p. 20).

[22] Carline M J Brands and Martinus a J S Van Boekel. “Kinetic modelling of reactions in heated disaccharide-casein systems”. In: *Journal of Agricultural and Food Chemistry* 50 (2002), pp. 6725–6739 (cit. on p. 57).

[23] Carline MJ Brands and Martinus AJS van Boekel. “Reactions of monosaccharides during heating of sugar- casein systems: Building of a reaction network model”. In: *Journal of Agricultural and Food Chemistry* 49.10 (2001), pp. 4667–4675 (cit. on p. 56).

[24] George Edward Briggs and John Burdon Sanderson Haldane. “A note on the kinetics of enzyme action”. In: *Biochemical journal* 19.2 (1925), p. 338 (cit. on pp. 4, 16, 17).

[25] Lihi Brocke et al. “Functional implications of the subunit composition of neuronal CaM kinase II”. In: *Journal of Biological Chemistry* 274.32 (1999), pp. 22713–22722 (cit. on p. 46).

[26] K S Brown et al. “The statistical mechanics of complex signaling networks: nerve growth factor signaling”. In: *Physical Biology* 1.3 (2004), p. 184 (cit. on pp. 75–77, 102, 104, 106, 108).

[27] P. Buchholz. “Exact and Ordinary Lumpability in Finite Markov Chains”. In: *Journal of Applied Probability* 31.1 (1994), pp. 59–75. ISSN: 00219002 (cit. on pp. 8, 22, 31, 32).

[28] Saulius Butenas et al. “The significance of circulating factor IXa in blood”. In: *Journal of Biological Chemistry* 279.22 (2004), pp. 22875–22882. ISSN: 00219258. DOI: 10.1074/jbc.M400531200 (cit. on p. 57).

[29] Daniele Cappelletti and Carsten Wiuf. “Elimination of intermediate species in multiscale stochastic reaction networks”. In: *Ann. Appl. Probab.* 26.5 (Oct. 2016), pp. 2915–2958. DOI: 10.1214/15-AAP1166. URL: https://doi.org/10.1214/15-AAP1166 (cit. on p. 58).

[30] L. Cardelli. “On process rate semantics”. In: *Theor. Comput. Sci.* 391.3 (3 2008), pp. 190–215. ISSN: 0304-3975 (cit. on p. 42).

[31] Luca Cardelli. “Morphisms of reaction networks that couple structure to function”. In: *BMC Systems Biology* 8.1 (2014) (cit. on pp. 10, 84, 102).

[32] Luca Cardelli Cardelli et al. “Aggregation of species with equivalent stochastic dynamics in chemical reaction networks by partition refinement”. Submitted to Nature Scientific Reports. 2019 (cit. on p. 35).

[33] Luca Cardelli et al. “Comparing Chemical Reaction Networks: A Categorical and Algorithmic Perspective”. In: *LICS*. 2016, pp. 485–494. DOI: 10.1145/2933575.2935318 (cit. on p. 106).

[34] Luca Cardelli et al. “Efficient Syntax-Driven Lumping of Differential Equations”. In: *TACAS*. 2016, pp. 93–111. DOI: 10.1007/978-3-662-49674-9_6 (cit. on p. 20).

[35] Luca Cardelli et al. “ERODE: a tool for the evaluation and reduction of ordinary differential equations”. In: *International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*. Springer. 2017, pp. 310–328 (cit. on pp. 24, 107).

[36] Luca Cardelli et al. “Forward and Backward Bisimulations for Chemical Reaction Networks”. In: *26th International Conference on Concurrency Theory, CONCUR*. 2015, pp. 226–239. DOI: 10.4230/LIPIcs.CONCUR.2015.226 (cit. on pp. 5, 6, 20, 22, 34, 106, 107).

[37] Luca Cardelli et al. “Guaranteed Error Bounds on Approximate Model Abstractions through Reachability Analysis”. In: *15th International Conference on Quantitative Evaluation of Systems (QEST)*. 2018 (cit. on pp. 6, 9, 80).

[38] Luca Cardelli et al. “Maximal aggregation of polynomial dynamical systems”. In: *Proceedings of the National Academy of Sciences* 114.38 (2017), pp. 10029–10034 (cit. on pp. 6, 7, 9, 10, 20, 22–24, 32, 38, 79, 106, 107).

[39] Luca Cardelli et al. “Symbolic Computation of Differential Equivalences”. In: *POPL*. 2016, pp. 137–150. DOI: 10.1145/2837614.2837649 (cit. on pp. 6, 9, 10, 22, 28, 69, 79, 106).

[40] Luca Cardelli et al. “Syntactic Markovian bisimulation for chemical reaction networks”. In: *Models, Algorithms, Logics and Tools*. Springer, 2017, pp. 466–483 (cit. on pp. 8–10, 20, 22, 29, 31, 33, 38, 58, 107).

[41] Ron Caspi et al. “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases”. In: *Nucleic acids research* 42.D1 (2013), pp. D459–D471 (cit. on pp. 60, 61).

[42] Claudio Castellano et al. “Statistical physics of social dynamics”. In: *Reviews of modern physics* 81.2 (2009), p. 591 (cit. on p. 50).

[43] E. Cator and P. Van Mieghem. “Second-order mean-field susceptible-infected-susceptible epidemic threshold”. In: *Phys. Rev. E* 85 (5 May 2012), p. 056111. DOI: 10.1103/PhysRevE.85.056111. URL: https://link.aps.org/doi/10.1103/PhysRevE.85.056111 (cit. on p. 53).

[44] Ayse Koca Caydasi et al. “A dynamical model of the spindle position checkpoint”. In: *Molecular systems biology* 8.1 (2012), p. 582 (cit. on pp. xxi, 48, 49, 57).

[45] Manash S Chatterjee et al. “Systems biology of coagulation initiation: kinetics of thrombin generation in resting and activated human blood”. In: *PLOS Computational Biology* 6.9 (2010), e1000950 (cit. on p. xxi).

[46] Vijayalakshmi Chelliah et al. “BioModels database: a repository of mathematical models of biological processes”. In: *Encyclopedia of Systems Biology* (2013), pp. 134–138 (cit. on p. 62).

[47] MY Chou and TC Ho. “Continuum theory for lumping nonlinear reactions”. In: *AIChE journal* 34.9 (1988), pp. 1519–1527 (cit. on p. 19).

[48] J. Colvin et al. “RuleMonkey: software for stochastic simulation of rule-based models”. In: *BMC Bioinformatics* 11 (2010), p. 404. DOI: 10.1186/1471-2105-11-404 (cit. on p. 38).

[49] J. Colvin et al. “Simulation of large-scale rule-based models”. In: *Bioinformatics* 25.7 (2009), pp. 910–917 (cit. on p. 38).

[50] UniProt Consortium. “UniProt: a hub for protein information”. In: *Nucleic acids research* 43.D1 (2014), pp. D204–D212 (cit. on p. 62).

[51] Holger Conzelmann et al. “A domain-oriented approach to the reduction of combinatorial complexity in signal transduction networks”. In: *BMC bioinformatics* 7.1 (2006), p. 34 (cit. on pp. 4, 20).

[52] Holger Conzelmann et al. “Exact model reduction of combinatorial reaction networks”. In: *BMC systems biology* 2.1 (2008), p. 78 (cit. on pp. 4, 7).

[53] Dawn Cotter et al. “Lmpd: lipid maps proteome database”. In: *Nucleic acids research* 34.suppl_1 (2006), pp. D507–D510 (cit. on p. 61).

[54] Mélanie Courtot et al. “Controlled vocabularies and semantics in systems biology”. In: *Molecular systems biology* 7.1 (2011), p. 543 (cit. on p. 62).

[55] Gheorghe Craciun et al. “Understanding bistability in complex enzyme-driven reaction networks”. In: *Proceedings of the National Academy of Sciences* 103.23 (2006), pp. 8697–8702. ISSN: 0027-8424. DOI: 10.1073/pnas.0602767103. eprint: https://www.pnas.org/content/103/23/8697.full.pdf. URL: https://www.pnas.org/content/103/23/8697 (cit. on p. xxi).

[56] Sune Danø et al. “Reduction of a biochemical model with preservation of its basic dynamic properties”. In: *The FEBS journal* 273.21 (2006), pp. 4862–4877 (cit. on p. 20).

[57] Vincent Danos et al. “Abstracting the Differential Semantics of Rule-Based Models: Exact and Automated Model Reduction”. In: *LICS*. 2010, pp. 362–381. DOI: 10.1109/LICS.2010.44 (cit. on pp. 106, 110).

[58] Vincent Danos et al. “Rule-based modelling of cellular signalling”. In: *International conference on concurrency theory*. Springer. 2007, pp. 17–41 (cit. on p. 21).

[59] Leonardo De Moura and Nikolaj Bjørner. “Z3: An Efficient SMT Solver”. In: *TACAS*. 2008, pp. 337–340 (cit. on p. 29).

[60] Daniela Degenring et al. “Sensitivity analysis for the reduction of complex metabolism models”. In: *Journal of Process Control* 14.7 (2004), pp. 729–745 (cit. on pp. 6, 17).

[61] Kirill Degtyarenko et al. “ChEBI: a database and ontology for chemical entities of biological interest”. In: *Nucleic acids research* 36.suppl_1 (2007), pp. D344–D350 (cit. on p. 62).

[62] Domitilla Del Vecchio and Eduardo D Sontag. “Engineering principles in bio-molecular systems: from retroactivity to modularity”. In: *European Journal of Control* 15.3-4 (2009), pp. 389–397 (cit. on p. 3).

[63] Aristides Dokoumetzidis and Leon Aarons. “Proper lumping in systems biology models”. In: *IET systems biology* 3.1 (2009), pp. 40–51 (cit. on p. 20).

[64] Christian Doll et al. “Agonist-selective patterns of μ-opioid receptor phosphorylation revealed by phosphosite-specific antibodies”. In: *British journal of pharmacology* 164.2 (2011), pp. 298–307 (cit. on p. 39).

[65] Mirela Domijan and Markus Kirkilionis. “Graph theory and qualitative analysis of reaction networks”. In: *NHM* 3.2 (2008), pp. 295–322 (cit. on p. 1).

[66] Maddalena Donzelli et al. “Hierarchical order of phosphorylation events commits Cdc25A to Beta-TrCP-dependent degradation”. In: *Cell Cycle* 3.4 (2004), pp. 467–469 (cit. on p. 39).

[67] Andreas Dräger et al. “JSBML: a flexible Java library for working with SBML”. In: *Bioinformatics* 27.15 (June 2011), pp. 2167–2168. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr361. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/27/15/2167/13846361/btr361.pdf (cit. on p. 62).

[68] Omer Dushek et al. “Ultrasensitivity in multisite phosphorylation of membrane-anchored proteins”. In: *Biophysical journal* 100.5 (2011), pp. 1189–1197 (cit. on p. 38).

[69] R. W. Eash et al. “Equilibrium Traffic Assignment on an Aggregated Highway Network for Sketch Planning”. In: *Transportation Research Record* 994 (1983), pp. 30–37 (cit. on p. 51).

[70] Michael B. Elowitz et al. “Stochastic Gene Expression in a Single Cell”. In: *Science* 297.5584 (2002), pp. 1183–1186 (cit. on p. 7).

[71] G.A. Enciso et al. “Nonmonotone systems decomposable into monotone systems with negative feedback”. In: *Journal of Differential Equations* 224.1 (2006), pp. 205–227. DOI: https://doi.org/10.1016/j.jde.2005.05.007 (cit. on p. 106).

[72] Peter Erdi and Janos Toth. *Mathematical models of chemical reactions: theory and applications of deterministic and stochastic models*. Manchester University Press, 1989 (cit. on p. 1).

[73] Antonio Fabregat et al. “The reactome pathway knowledgebase”. In: *Nucleic acids research* 46.D1 (2017), pp. D649–D655 (cit. on p. 60).

[74] J. R. Faeder et al. “Investigation of Early Events in FcεRI-Mediated Signaling Using a Detailed Mathematical Model”. In: *The Journal of Immunology* 170.7 (2003), pp. 3769–3781 (cit. on p. 38).

[75] Scott Federhen. “The NCBI taxonomy database”. In: *Nucleic acids research* 40.D1 (2011), pp. D136–D143 (cit. on p. 62).

[76] M. Feinberg. *Lectures on Chemical Reaction Networks*. Tech. rep. University of Wisconsin, 1979. URL: https://crnt.osu.edu/LecturesOnReactionNetworks (cit. on p. 42).

[77] J. Feret et al. “Internal coarse-graining of molecular systems”. In: *Proc. of the National Academy of Sciences* 106.16 (2009), pp. 6453–6458. DOI: 10.1073/pnas.0809908106 (cit. on p. 106).

[78] Jerome Feret et al. “Lumpability abstractions of rule-based systems”. In: *Theoretical Computer Science* 431 (2012), pp. 137–164 (cit. on pp. 20, 21).

[79] Simon J Fraser. “The steady state and equilibrium approximations: A geometrical picture”. In: *The Journal of chemical physics* 88.8 (1988), pp. 4732–4738 (cit. on p. 16).

[80] Gregor F. Fussmann et al. “Crossing the Hopf Bifurcation in a Live Predator-Prey System”. In: *Science* 290.5495 (2000), pp. 1358–1360. ISSN: 0036-8075. DOI: 10.1126/science.290.5495.1358. eprint: http://science.sciencemag.org/content/290/5495/1358.full.pdf (cit. on p. xxi).

[81] Arnab Ganguly et al. “Markov chain aggregation and its applications to combinatorial reaction networks”. In: *Journal of mathematical biology* 69.3 (2014), pp. 767–797 (cit. on p. 22).

[82] Pascale Gaudet et al. “Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium”. In: *Briefings in bioinformatics* 12.5 (2011), pp. 449–462 (cit. on p. 60).

[83] Carel van Gend and Jacky L Snoep. “Systems biology model databases and resources”. In: *Essays in biochemistry* 45 (2008), pp. 223–236 (cit. on p. 59).

[84] Kristina M. Giantsos-Adams et al. “Heparan sulfate regrowth profiles under laminar shear flow following enzymatic degradation”. In: *Cellular and Molecular Bioengineering* 6.2 (2013), pp. 160–174. ISSN: 18655025. DOI: 10.1007/s12195-013-0273-z (cit. on p. 57).

[85] Daniel T Gillespie. “Exact stochastic simulation of coupled chemical reactions”. In: *The journal of physical chemistry* 81.25 (1977), pp. 2340–2361 (cit. on pp. 3, 110).

[86] Anne-Claude Gingras et al. “Hierarchical phosphorylation of the translation inhibitor 4E-BP1”. In: *Genes & development* 15.21 (2001), pp. 2852–2864 (cit. on p. 39).

[87] Carlos A. Gómez-Uribe et al. “Enhanced identification and exploitation of time scales for model reduction in stochastic chemical kinetics”. In: *The Journal of Chemical Physics* 129.24 (2008), p. 244112. DOI: 10.1063/1.3050350. eprint: https://doi.org/10.1063/1.3050350. URL: https://doi.org/10.1063/1.3050350 (cit. on p. 58).

[88] Marko Gosak et al. “Network science of biological systems at different scales: a review”. In: *Physics of life reviews* 24 (2018), pp. 118–135 (cit. on p. 1).

[89] John Goutsias and Garrett Jenkinson. “Markovian dynamics on complex reaction networks”. In: *Physics reports* 529.2 (2013), pp. 199–264 (cit. on p. 8).

[90] Ulrike Gruneberg et al. “Nud1p links astral microtubule organization and the control of exit from mitosis”. In: *The EMBO Journal* 19.23 (Dec. 2000), pp. 6475–6488 (cit. on p. 48).

[91] Philip A Gruppuso et al. “Insulin receptor tyrosine kinase domain auto-dephosphorylation”. In: *Biochemical and biophysical research communications* 189.3 (1992), pp. 1457–1463 (cit. on p. 44).

[92] C. Gu. “QLMOR: A Projection-Based Nonlinear Model Order Reduction Approach Using Quadratic-Linear Representation of Nonlinear Systems”. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 30.9 (2011), pp. 1307–1320. DOI: 10.1109/TCAD.2011.2142184 (cit. on p. 79).

[93] Jeremy Gunawardena. "Multisite protein phosphorylation makes a good threshold but can be a poor switch". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.41 (2005), pp. 14617–14622. DOI: 10.1073/pnas.0507322102. eprint: http://www.pnas.org/content/102/41/14617.full.pdf (cit. on p. 43).

[94] An Chi Guo et al. "ECMDB: the E. coli Metabolome Database". In: *Nucleic acids research* 41.D1 (2012), pp. D625–D630 (cit. on p. 60).

[95] Purnananda Guptasarma. "Does replication-induced transcription regulate synthesis of the myriad low copy number proteins of Escherichia coli?" In: *BioEssays* 17.11 (1995), pp. 987–997 (cit. on p. 7).

[96] Yuji Hayashi et al. "Regulation of neuronal nitric-oxide synthase by calmodulin kinases". In: *Journal of Biological Chemistry* 274.29 (1999), pp. 20597–20602 (cit. on p. 57).

[97] Stephen M Hedrick et al. "FOXO transcription factors throughout T cell biology". In: *Nature Reviews Immunology* 12.9 (2012), p. 649 (cit. on p. 78).

[98] Reinhart Heinrich et al. "Mathematical models of protein kinase signal transduction". In: *Molecular Cell* 9.5 (2002), pp. 957–970. ISSN: 10972765. DOI: 10.1016/S1097-2765(02)00528-2 (cit. on p. xxi).

[99] Thomas A Henzinger et al. "Sliding window abstraction for infinite Markov chains". In: *International Conference on Computer Aided Verification*. Springer. 2009, pp. 337–352 (cit. on pp. 8, 58).

[100] Bo Kyeng Hou et al. "BioSilico: an integrated metabolic database system". In: *Bioinformatics* 20.17 (2004), pp. 3270–3272 (cit. on p. 61).

[101] Chi-Ying Huang and James E Ferrell. "Ultrasensitivity in the mitogen-activated protein kinase cascade". In: *Proceedings of the National Academy of Sciences* 93.19 (1996), pp. 10078–10083 (cit. on p. 75).

[102] H Huang et al. "A systematic lumping approach for the reduction of comprehensive kinetic models". In: *Proceedings of the Combustion Institute* 30.1 (2005), pp. 1309–1316 (cit. on p. 20).

[103] Michael Hucka et al. "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models". In: *Bioinformatics* 19.4 (2003), pp. 524–531 (cit. on pp. xxii, 9, 61, 64).

[104] Michael Hucka et al. “The Systems Biology Markup Language (SBML): language specification for level 3 version 2 core”. In: *Journal of integrative bioinformatics* 15.1 (2018) (cit. on pp. 64, 65).

[105] Michael Hucka et al. “Systems biology markup language (SBML) level 2: structures and facilities for model definitions”. In: *Nature Precedings* (2007), pp. 1–1 (cit. on pp. 64, 65).

[106] Michael Hucka et al. “The Systems Biology Markup Language (SBML): language specification for level 3 version 1 core”. In: *Journal of integrative bioinformatics* 12.2 (2015), pp. 382–549 (cit. on pp. 64, 65).

[107] Andy Hudmon and Howard Schulman. “Structure–function of the multifunctional Ca2+/calmodulin-dependent protein kinase II”. In: *Biochemical Journal* 364.3 (2002), pp. 593–611 (cit. on p. 46).

[108] Giulio Iacobelli et al. “Differential bisimulation for a Markovian process algebra”. In: *International Symposium on Mathematical Foundations of Computer Science*. Springer. 2015, pp. 293–306 (cit. on p. 6).

[109] Lisa Jeske et al. “BRENDA in 2019: a European ELIXIR core data resource”. In: *Nucleic acids research* 47.D1 (2018), pp. D542–D549 (cit. on p. 61).

[110] Qian Jiang et al. “Quorum Sensing: A Prospective Therapeutic Target for Bacterial Diseases”. In: *BioMed research international* 2019 (2019) (cit. on p. xxi).

[111] David Jullesson et al. “Dominant negative inhibition data should be analyzed using mathematical modeling–re-interpreting data from insulin signaling”. In: *The FEBS journal* 282.4 (2015), pp. 788–802 (cit. on p. 57).

[112] Nick Juty et al. “Identifiers. org and MIRIAM Registry: community resources to provide persistent identification”. In: *Nucleic acids research* 40.D1 (2011), pp. D580–D586 (cit. on p. 62).

[113] Minoru Kanehisa and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30 (cit. on pp. 60, 61).

[114] Hye-Won Kang and Thomas G. Kurtz. “Separation of time-scales and model reduction for stochastic reaction networks”. In: *The Annals of Applied Probability* 23.2 (2013), pp. 529–583. ISSN: 10505164. URL: http://www.jstor.org/stable/23474859 (cit. on p. 58).

[115] J.G. Kemeny and J.L. Snell. *Finite Markov Chains*. Berlin: Springer New York, Heidelberg, 1976 (cit. on pp. 8, 22, 32).

[116] W. O. Kermack and A. G. McKendrick. “A Contribution to the Mathematical Theory of Epidemics”. In: *Proceedings of the Royal Society of London. Series A.* 115.772 (1927), pp. 700–721 (cit. on p. 50).

[117] Ingrid M Keseler et al. “The EcoCyc database: reflecting new knowledge about Escherichia coli K-12”. In: *Nucleic acids research* 45.D1 (2016), pp. D543–D550 (cit. on p. 60).

[118] Boris N Kholodenko. “Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades”. In: *European journal of biochemistry* 267.6 (2000), pp. 1583–1588 (cit. on p. 75).

[119] Boris N. Kholodenko. “Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades”. In: *European Journal of Biochemistry* 267.6 (2000), pp. 1583–1588. DOI: 10.1046/j.1432-1327.2000.01197.x (cit. on pp. xxi, 99, 106).

[120] Zachary A King et al. “BiGG Models: A platform for integrating, standardizing and sharing genome-scale models”. In: *Nucleic acids research* 44.D1 (2015), pp. D515–D522 (cit. on p. 61).

[121] Hiroaki Kitano. “Biological robustness”. In: *Nature Reviews Genetics* 5 (Nov. 2004), 826 EP - (cit. on p. 82).

[122] Hiroaki Kitano. “Systems biology: a brief overview”. In: *science* 295.5560 (2002), pp. 1662–1664 (cit. on pp. xxi, 1).

[123] Katrin Kolczyk et al. “The Process-Interaction-Model: a common representation of rule-based and logical models allows studying signal transduction on different levels of detail”. In: *BMC bioinformatics* 13.1 (2012), p. 251 (cit. on p. 38).

[124] Markus Koschorreck et al. “Reduced modeling of signal transduction–a modular approach”. In: *BMC bioinformatics* 8.1 (2007), p. 336 (cit. on p. 20).

[125] James Kuo J. C. W.; Wei. "Lumping Analysis in Monomolecular Reaction Systems. Analysis of Approximately Lumpable System". In: *Industrial and Engineering Chemistry Fundamentals* 8 (1 Feb. 1969). DOI: 10.1021/i160029a020. URL: http://gen.lib.rus.ec/scimag/index.php?s=10.1021/i160029a020 (cit. on pp. 18, 19).

[126] T. G. Kurtz. "The Relationship between Stochastic and Deterministic Models for Chemical Reactions." In: *Journal of Chemical Physics* 57.7 (1972) (cit. on pp. 2, 42).

[127] Elena Kutumova et al. "Model composition through model reduction: a combined model of CD95 and NF-κB signaling pathways". In: *BMC systems biology* 7.1 (2013), p. 13 (cit. on p. xxi).

[128] Cho Kwang-Hyun et al. "Mathematical modeling of the influence of RKIP on the ERK signaling pathway". In: *International Conference on Computational Methods in Systems Biology*. Springer. 2003, pp. 127–141 (cit. on p. xxi).

[129] Eric W. F. Lam et al. "Forkhead box proteins: tuning forks for transcriptional harmony". In: *Nature Reviews Cancer* 13 (June 2013), 482 EP - (cit. on p. 78).

[130] K. G. Larsen and A. Skou. "Bisimulation through probabilistic testing". In: *Information and Computation* 94.1 (1991), pp. 1–28 (cit. on p. 6).

[131] Nicolas Le Novère et al. "Minimum information requested in the annotation of biochemical models (MIRIAM)". In: *Nature biotechnology* 23.12 (2005), p. 1509 (cit. on p. 62).

[132] Nicolas Le Novère et al. "JSBML 1.0: providing a smorgasbord of options to encode systems biology models". In: *Bioinformatics* 31.20 (June 2015), pp. 3383–3386. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv341. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/31/20/3383/17087774/btv341.pdf (cit. on p. 62).

[133] Chang Jun Lee et al. "A revisit to the one form kinetic model of prothrombinase". In: *Biophysical chemistry* 149.1-2 (2010), pp. 28–33 (cit. on p. 56).

[134] Jure Leskovec et al. "Graph Evolution: Densification and Shrinking Diameters". In: *ACM Trans. Knowledge Discovery from Data* 1.1 (2007), pp. 1–40 (cit. on p. 51).

[135] Daniel J. Lew and Daniel J. Burke. "The Spindle Assembly and Spindle Position Checkpoints". In: *Annual Review of Genetics* 37.1 (2003). PMID: 14616062, pp. 251–282 (cit. on p. 48).

[136] Chen Li et al. "BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models." In: *BMC Systems Biology* 4 (June 2010), p. 92 (cit. on pp. 55, 61).

[137] Genyuan Li and Herschel Rabitz. "A general analysis of approximate lumping in chemical kinetics". In: *Chemical engineering science* 45.4 (1990), pp. 977–1002 (cit. on pp. xxii, 9).

[138] Genyuan Li and Herschel Rabitz. "A general analysis of exact lumping in chemical kinetics". In: *Chemical Engineering Science* 44.6 (1989), pp. 1413–1430. ISSN: 0009-2509. DOI: http://dx.doi.org/10.1016/0009-2509(89)85014-6 (cit. on p. 5).

[139] Genyuan Li et al. "A general analysis of exact nonlinear lumping in chemical kinetics". In: *Chemical Engineering Science* 49.3 (1994), pp. 343–361 (cit. on pp. 18, 19).

[140] Jun Li et al. "A stochastic model of Escherichia coli AI-2 quorum signal circuit reveals alternative synthesis pathways". In: *Molecular Systems Biology* 2 (2006). ISSN: 17444292. DOI: 10.1038/msb4100107 (cit. on p. 57).

[141] Wolfram Liebermeister et al. "Biochemical network models simplified by balanced truncation". In: *The FEBS journal* 272.16 (2005), pp. 4034–4043 (cit. on p. 3).

[142] John Lisman et al. "Mechanisms of CaMKII action in long-term potentiation". In: *Nature Reviews Neuroscience* 13 (2012), pp. 169–182 (cit. on p. 47).

[143] John Lisman et al. "The molecular basis of CaMKII function in synaptic and behavioural memory". In: *Nature Reviews Neuroscience* 3 (2002), pp. 175–190 (cit. on p. 46).

[144] Verónica Lloréns-Rico et al. "Bacterial antisense RNAs are mainly the product of transcriptional noise". In: *Science Advances* 2.3 (2016), pp. 1–9. ISSN: 23752548. DOI: 10.1126/sciadv.1501363. arXiv: arXiv:1011.1669v3 (cit. on p. 57).

[145] Ulrich Maas and Stephen B Pope. "Implementation of simplified chemical kinetics based on intrinsic low-dimensional manifolds". In: *Symposium (International) on Combustion*. Vol. 24. 1. Elsevier. 1992, pp. 103–112 (cit. on p. 16).

[146] Daniel P Maki and Maynard Thompson. *Mathematical models and applications: with emphasis on the social life, and management sciences*. Prentice-Hall Englewood Cliffs, 1973 (cit. on p. 50).

[147] Nick I Markevich et al. "Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades". In: *The Journal of cell biology* 164.3 (2004), pp. 353–359 (cit. on pp. xxi, 2, 3, 74, 75, 108).

[148] NI Markevich et al. "Signal processing at the Ras circuit: what shapes Ras activation patterns?" In: *Systems biology* 1.1 (2004), pp. 104–113 (cit. on p. 75).

[149] Angélica S Mata and Silvio C Ferreira. "Pair quenched mean-field theory for the susceptible-infected-susceptible model on complex networks". In: *EPL (Europhysics Letters)* 103.4 (2013), p. 48003 (cit. on p. 53).

[150] Julian McAuley and Jure Leskovec. "Learning to Discover Social Circles in Ego Networks". In: *Advances in Neural Information Processing Systems*. 2012, pp. 548–556 (cit. on p. 51).

[151] Melanie Meister et al. "Mitogen-activated protein (MAP) kinase scaffolding proteins: a recount". In: *International journal of molecular sciences* 14.3 (2013), pp. 4854–4884 (cit. on p. 43).

[152] Hongyu Miao et al. "Modeling the dynamics and migratory pathways of virus-specific antibody-secreting cell populations in primary influenza infection". In: *PLOS One* 9.8 (2014), e104781 (cit. on p. 57).

[153] Andrew K Miller et al. "An overview of the CellML API and its implementation". In: *BMC bioinformatics* 11.1 (2010), p. 178 (cit. on p. 61).

[154] Karin Moelling et al. "Regulation of Raf-Akt cross-talk". In: *Journal of Biological Chemistry* 277.34 (2002), pp. 31099–31106 (cit. on p. 76).

[155] Deborah K Morrison. "KSR: a MAPK scaffold of the Ras pathway?" In: *Journal of cell science* 114.9 (2001), pp. 1609–1612 (cit. on p. 43).

[156] Brian Munsky and Mustafa Khammash. "The finite state projection algorithm for the solution of the chemical master equation". In: *The Journal of Chemical Physics* 124.4 (2006), p. 044104 (cit. on pp. 8, 58).

[157] Ambarish Nag et al. “Aggregation of membrane proteins by cytosolic cross-linkers: theory and simulation of the LAT-Grb2-SOS1 system”. In: *Biophysical journal* 96.7 (2009), pp. 2604–2623 (cit. on p. 38).

[158] RI Nicholson et al. “EGFR and cancer prognosis”. In: *European journal of cancer* 37 (2001), pp. 9–15 (cit. on p. 1).

[159] Darryl Nishimura. “BioCarta”. In: *Biotech Software & Internet Report: The Computer Software Journal for Scient* 2.3 (2001), pp. 117–120 (cit. on p. 61).

[160] Raffaella Ocone and Gianni Astarita. “Kinetics and thermodynamics in multicomponent mixtures”. In: *Advances in Chemical Engineering*. Vol. 24. Elsevier, 1998, pp. 1–78 (cit. on p. 19).

[161] Miles S Okino and Michael L Mavrovouniotis. “Simplification of mathematical models of chemical reaction systems”. In: *Chemical reviews* 98.2 (1998), pp. 391–408 (cit. on pp. 17, 20).

[162] Brett G Olivier and Jacky L Snoep. “Web-based kinetic modelling using JWS Online”. In: *Bioinformatics* 20.13 (2004), pp. 2143–2144 (cit. on pp. 55, 61).

[163] Richard J Orton et al. “Computational modelling of cancerous mutations in the EGFR/ERK signalling pathway”. In: *BMC Systems Biology* 3.1 (2009), p. 100 (cit. on p. 77).

[164] R. Paige and R. Tarjan. “Three Partition Refinement Algorithms”. In: *SIAM Journal on Computing* 16.6 (1987), pp. 973–989 (cit. on p. 23).

[165] Francesco Pappalardo et al. “Computational modeling of PI3K/AKT and MAPK signaling pathways in melanoma cancer”. In: *PLoS One* 11.3 (2016), e0152104 (cit. on p. xxi).

[166] Romualdo Pastor-Satorras et al. “Epidemic processes in complex networks”. In: *Reviews of modern physics* 87.3 (2015), p. 925 (cit. on pp. 50, 51, 53).

[167] Shirley Pepke et al. “A dynamic model of interactions of Ca2+, calmodulin, and catalytic subunits of Ca2+/calmodulin-dependent protein kinase II”. In: *PLoS computational biology* 6.2 (2010), e1000675 (cit. on pp. 46, 47, 57).

[168] Gislene Pereira and Elmar Schiebel. “Kin4 Kinase Delays Mitotic Exit in Response to Spindle Alignment Defects”. In: *Molecular Cell* 19.2 (2005), pp. 209–221 (cit. on p. 48).

[169] Isabel Cristina Pérez-Verona et al. “A Large-Scale Assessment of Exact Model Reduction in the BioModels Repository”. In: *International Conference on Computational Methods in Systems Biology*. Springer. 2019, pp. 248–265 (cit. on pp. 9, 24, 29, 31, 107).

[170] Isabel Cristina Pérez-Verona et al. “Fixed-Point Computation of Equilibria in Biochemical Regulatory Networks”. In: *International Workshop on Hybrid Systems Biology*. Springer. 2019, pp. 45–62 (cit. on pp. 9–11, 108).

[171] Tatjana Petrov. “Markov chain aggregation and its application to rule-based modelling”. In: *Modeling Biomolecular Site Dynamics*. Springer, 2019, pp. 297–313 (cit. on pp. 21, 22).

[172] Dexter Pratt et al. “NDEx, the network data exchange”. In: *Cell systems* 1.4 (2015), pp. 302–305 (cit. on p. 61).

[173] Thomas P Prescott and Antonis Papachristodoulou. “Layered decomposition for the model order reduction of timescale separated biochemical reaction networks”. In: *Journal of theoretical biology* 356 (2014), pp. 113–122 (cit. on p. 4).

[174] Thomas Prescott and Antonis Papachristodoulou. “Signal propagation across layered biochemical networks”. In: *American Control Conference (ACC), 2014*. IEEE. 2014, pp. 3399–3404 (cit. on p. 3).

[175] Carole J. Proctor and Douglas A. Gray. “Explaining oscillations and variability in the p53-Mdm2 system”. In: *BMC Systems Biology* 2.2006 (2008), pp. 1–20. ISSN: 17520509. DOI: 10.1186/1752-0509-2-75 (cit. on pp. xxi, 57).

[176] N. Radde. “Fixed point characterization of biological networks with complex graph topology”. In: *Bioinformatics* 26.22 (2010), pp. 2874–2880. DOI: 10.1093/bioinformatics/btq517 (cit. on pp. 86, 106).

[177] Nicole Radde. “Analyzing fixed points of intracellular regulation networks with interrelated feedback topology”. In: *BMC Systems Biology* 6.1 (June 2012), p. 57. DOI: 10.1186/1752-0509-6-57 (cit. on pp. 85, 86).

[178] Ovidiu Radulescu et al. “Reduction of dynamical biochemical reactions networks in computational biology”. In: *Frontiers in genetics* 3 (2012), p. 131 (cit. on p. 17).

[179] Thomas B Rasmussen and Michael Givskov. "Quorum-sensing inhibitors as anti-pathogenic drugs". In: *International Journal of Medical Microbiology* 296.2-3 (2006), pp. 149–161 (cit. on p. xxi).

[180] Carlos Salazar and Thomas Höfer. "Multisite protein phosphorylation – from molecular mechanisms to kinetic models". In: *FEBS Journal* 276.12 (2009), pp. 3177–3198. ISSN: 1742-4658. DOI: 10.1111/j.1742-4658.2009.07027.x. URL: http://dx.doi.org/10.1111/j.1742-4658.2009.07027.x (cit. on pp. 3, 43, 44).

[181] Satoru Sasagawa et al. "Prediction and validation of the distinct dynamics of transient and sustained ERK activation". In: *Nature cell biology* 7.4 (2005), p. 365 (cit. on p. 3).

[182] M Schauer and R Heinrich. "Quasi-steady-state approximation in the mathematical modeling of biochemical reaction networks". In: *Mathematical biosciences* 65.2 (1983), pp. 155–170 (cit. on p. 4).

[183] David Schnoerr et al. "Approximation and inference methods for stochastic biochemical kinetics—a tutorial review". In: *Journal of Physics A: Mathematical and Theoretical* 50.9 (2017), p. 093001 (cit. on pp. 58, 80).

[184] Julia M. Shifman et al. "Ca2+/calmodulin-dependent protein kinase II (CaMKII) is activated by calmodulin with two bound calciums". In: *Proceedings of the National Academy of Sciences* 103.38 (2006), pp. 13968–13973 (cit. on p. 47).

[185] Péter L Simon et al. "Exact epidemic models on graphs using graph-automorphism driven lumping". In: *Journal of mathematical biology* 62.4 (2011), pp. 479–508 (cit. on pp. 50, 51, 53, 108).

[186] N. A. Sinitsyn et al. "Adiabatic coarse-graining and simulations of stochastic biochemical networks". In: *Proceedings of the National Academy of Sciences* 106.26 (2009), pp. 10546–10551. ISSN: 0027-8424. DOI: 10.1073/pnas.0809340106. eprint: http://www.pnas.org/content/106/26/10546.full.pdf. URL: http://www.pnas.org/content/106/26/10546 (cit. on p. 58).

[187] Rex T Skodje and Michael J Davis. "Geometrical simplification of complex kinetic systems". In: *The Journal of Physical Chemistry A* 105.45 (2001), pp. 10356–10365 (cit. on p. 16).

[188] Graham R Smith and Daryl P Shanley. "Modelling the response of FOXO transcription factors to multiple post-translational modifications made by ageing-related signalling pathways". In: *PLOS One* 5.6 (2010), e11092 (cit. on pp. 57, 79).

[189] Michael W Sneddon et al. "Efficient modeling, simulation and coarse-graining of biological complexity with NFsim". In: *Nature methods* 8.2 (2011), p. 177 (cit. on pp. 3, 38, 39, 44, 45, 60).

[190] Thomas J Snowden et al. "Methods of model reduction for large-scale biological systems: a survey of current methods and trends". In: *Bulletin of mathematical biology* 79.7 (2017), pp. 1449–1486 (cit. on pp. 16, 17, 20).

[191] Guenter Stoesser et al. "The EMBL nucleotide sequence database". In: *Nucleic acids research* 30.1 (2002), pp. 21–26 (cit. on p. 62).

[192] Manish Sud et al. "Lmsd: Lipid maps structure database". In: *Nucleic acids research* 35.suppl_1 (2006), pp. D527–D532 (cit. on p. 61).

[193] R. Suderman and E. J. Deeds. "Machines vs. Ensembles: Effective MAPK Signaling through Heterogeneous Sets of Protein Complexes". In: *PLOS Comput Biol* 9.10 (Oct. 2013), e1003278. DOI: 10.1371/journal.pcbi.1003278 (cit. on p. 38).

[194] Mikael Sunnåker et al. "Zooming of states and parameters using a lumping approach including back-translation". In: *BMC systems biology* 4.1 (2010), p. 28 (cit. on pp. 20, 21).

[195] Peter S. Swain et al. "Intrinsic and extrinsic contributions to stochasticity in gene expression". In: *Proceedings of the National Academy of Sciences* 99.20 (2002), pp. 12795–12800 (cit. on p. 7).

[196] Stephanie R Taylor et al. "Oscillator model reduction preserving the phase response: application to the circadian clock". In: *Biophysical journal* 95.4 (2008), pp. 1658–1673 (cit. on p. 3).

[197] Shannon E Telesco et al. "A multiscale modeling approach to investigate molecular mechanisms of pseudokinase activation and drug resistance in the HER3/ErbB3 receptortyrosine kinase signaling network". In: *Molecular Biosystems* 7.6 (2011), pp. 2066–2080 (cit. on p. xxi).

[198] Matthew Thomson and Jeremy Gunawardena. "Unlimited multistability in multisite phosphorylation systems". In: *Nature* 460.7252 (July 2009), pp. 274–277. URL: http://dx.doi.org/10.1038/nature08102 (cit. on p. 43).

[199] Mirco Tribastone and Andrea Vandin. "Speeding up stochastic and Deterministic simulation by Aggregation: an Advanced Tutorial". In: *2018 Winter Simulation Conference, WSC 2018, Gothenburg, Sweden, December 9-12, 2018*. 2018, pp. 336–350. DOI: 10.1109/WSC.2018.8632364 (cit. on pp. 2, 24).

[200] M. Tschaikowski and M. Tribastone. "Exact fluid lumpability for Markovian process algebra". In: *23rd International Conference on Concurrency Theory (CONCUR)*. LNCS. 2012, pp. 380–394 (cit. on pp. 20, 110).

[201] Tamas Turanyi and Alison S. Tomlin. *Analysis of Kinetic Reaction Mechanisms*. Springer, 2014 (cit. on pp. xxii, 5).

[202] John J. Tyson and Béla Novák. "Functional Motifs in Biochemical Reaction Networks". In: *Annual Review of Physical Chemistry* 61.1 (2010). PMID: 20055671, pp. 219–240. DOI: 10.1146/annurev.physchem.012809.103457. eprint: https://doi.org/10.1146/annurev.physchem.012809.103457 (cit. on pp. 10, 82, 84).

[203] Ravishankar R Vallabhajosyula and Herbert M Sauro. "Complexity reduction of biochemical networks". In: *Proceedings of the 38th conference on Winter simulation*. Winter Simulation Conference. 2006, pp. 1690–1697 (cit. on p. 3).

[204] Ravishankar Rao Vallabhajosyula et al. "Conservation analysis of large biochemical networks". In: *Bioinformatics* 22.3 (2005), pp. 346–353 (cit. on pp. 3, 4).

[205] A. Valmari and G. Franceschinis. "Simple $O(m \log n)$ Time Markov Chain Lumping". In: *International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*. 2010, pp. 38–52. DOI: 10.1007/978-3-642-12002-2_4 (cit. on pp. 37, 58).

[206] Lars P Van Der Heide et al. "The ins and outs of FoxO shuttling: mechanisms of FoxO translocation and transcriptional regulation". In: *Biochemical Journal* 380.2 (2004), pp. 297–309 (cit. on p. 78).

[207] N. G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. 3rd. Elsevier, 2007 (cit. on pp. 8, 16).

[208] Piet Van Mieghem. "The N-intertwined SIS epidemic network model". In: *Computing* 93.2 (Dec. 2011), pp. 147–169. ISSN: 1436-5057. DOI: 10.1007/s00607-011-0155-y. URL: https://doi.org/10.1007/s00607-011-0155-y (cit. on p. 53).

[209] D Vaudry et al. "Signaling pathways for PC12 cell differentiation: making the right connections". In: *Science* 296.5573 (2002), pp. 1648–1649 (cit. on p. xxi).

[210] Eberhard O. Voit. "Biochemical Systems Theory: A Review". In: *ISRN Biomathematics* 2013 (2013), p. 53. URL: http://dx.doi.org/10.1155/2013/897658%20%]%20897658 (cit. on pp. 2, 13).

[211] Eberhard O. Voit et al. "150 Years of the Mass Action Law". In: *PLOS Computational Biology* 11.1 (Jan. 2015), pp. 1–7. DOI: 10.1371/journal.pcbi.1004012. URL: https://doi.org/10.1371/journal.pcbi.1004012 (cit. on p. 13).

[212] Pu Wang et al. "Understanding the Spreading Patterns of Mobile Phone Viruses". In: *Science* 324.5930 (2009), pp. 1071–1076 (cit. on p. 50).

[213] Wei Wang et al. "Unification of theoretical approaches for epidemic spreading on complex networks". In: *Reports on Progress in Physics* 80.3 (2017), p. 036603 (cit. on pp. 51, 53).

[214] Ashley E Webb and Anne Brunet. "FOXO transcription factors: key regulators of cellular quality control". In: *Trends in biochemical sciences* 39.4 (2014), pp. 159–169 (cit. on p. 78).

[215] Ulrike Wittig et al. "SABIO-RK—database for biochemical reaction kinetics". In: *Nucleic acids research* 40.D1 (2011), pp. D790–D796 (cit. on p. 61).

[216] James WT Yates. "Structural identifiability of physiologically based pharmacokinetic models". In: *Journal of pharmacokinetics and pharmacodynamics* 33.4 (2006), pp. 421–439 (cit. on p. 16).

[217] R. Zafarani and H. Liu. *Social Computing Data Repository at ASU*. 2009. URL: http://socialcomputing.asu.edu (cit. on p. 51).

[218] Beichuan Zhang et al. "Collecting the Internet AS-level Topology". In: *SIGCOMM Computer Communication Review* 35.1 (2005), pp. 53–61 (cit. on p. 51).

[219] Zhike Zi. “Sensitivity analysis approaches applied to systems biology models”. In: *IET systems biology* 5.6 (2011), pp. 336–346 (cit. on pp. 9, 17).