Combinatorics and Stochasticity for Chemical Reaction Networks

Thesis by

Andrés Ortiz-Muñoz

In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2022
Defended November 12, 2021

Andrés Ortiz-Muñoz
ORCID: 0000-0003-1824-3230

Dicen que no soy poeta,
tiene razón quien lo diga.
A veces escribo versos
cuando la tristeza obliga,
para bendecir a Dios
o elogiar a una hormiga.

Tan sólo digo que escribo,
jamás diré que compongo,
pues los francotiradores
dirán que los descompongo.
Mis versos son una caricia,
a veces es un rezongo.

Cuando un versillo concibo
lo confío a un papel,
luego veo que lo levanta
algún remolino cruel,
otros coleccionan polvo
en un antiguo anaquel.

Así voy por las veredas
pepenando consonantes,
a mis versos no llegó
la retórica brillante;
si algo brilla en mis palabras
es un vidrio, no un diamante.

Igual que Ponce de León
que buscaba aquella fuente
de la eterna juventud
y falleció por impotente,
busco yo la inspiración
y cada día estoy más ausente.

Seguiré escribiendo versos
no porque me crea poeta,
sólo para denunciar
profunda inquietud secreta
que en mi viaje por la vida
se introdujo en mi maleta.

Mi raído gabán lleva consigo
el polvo de todos los caminos.
En mi peregrinar abro un paréntesis:
no dejé rastro, examino,
mis huellas y mis remendados versos
se los llevaron los impíos remolinos.

Pascual Ortiz Saucedo,
"No Soy Poeta" Antología

ACKNOWLEDGEMENTS

The process of writing this thesis was an exercise in humility. I had a vision for what I wanted it to be and it was way too ambitious for the time I had available. Despite that, I insisted on pursuing that vision, going through several iterations in which I essentially rewrote my thesis in light of new ideas and understanding I had acquired. As a result, the process lasted much longer than it needed to be, and at the end I did not succeed in fleshing out what I had envisioned. At the end, I did what I needed to have done from the beginning, writing a more modest thesis that was faithful to my Ph.D. experience. It took me a while to realize that, but I am glad I finally did. I suspect that my original vision, a novel foundation of mathematics for biology, will take a lifetime to develop, if I ever succeed. But you better believe that despite the odds, I will still try.

A recurrent topic in conversations with my father is how amazing it is that I find myself in my current position. I aspire to become the first one in my extended family to get a Ph.D., let alone from an institution like Caltech. I eluded the chaos that besets so many people's dreams in my home town of Juárez, Mexico. Having as mentors and advocates professors from Caltech and Harvard and being an Omidyar Fellow at the Santa Fe Institute, all while getting to make a living out of something I love, surpass all expectations for someone from my background. When I was half my current age, I set my mind on obtaining a Ph.D. in mathematics. Today I am not a mathematician, and hardly a biologist, but I made the most of what I was given, and it was much. I am blessed to have stumbled upon many people whose generosity towards me made for a path of least resistance to the position in which I find myself now.

I am humbled by the kind of support that my advisor, Erik Winfree, has given me. He has fostered and inspired the development of my own ideas despite our having very different ways of doing mathematics. The nature of our interactions was often that of friendly competition. I would make some mathematical claim and he would claim it was false, which pressed me to prove him wrong. I learned a tremendous amount from those exchanges, not to mention how much I enjoyed them. Needless to say, I was repeatedly crushed as chemical reaction networks seem to be Erik's native language. The sparse occasions in which I prevailed are my badges of honor and are scattered throughout this thesis. Erik's perspective on molecular computing has been a tremendous source of inspiration for me as I aspire to develop my own vision. I am also deeply grateful for Erik's patience and support through the process of writing my thesis or, should I say, theses.

Meeting my mentor Walter Fontana in 2011 during a summer internship at the Novartis Institutes for Biomedical Research in Cambridge, Massachusetts was one of the most momentous events in my life. After that summer, Walter assured me that if I was ever looking for work, the doors of his

lab in Harvard would be open. And so it was. It has become somewhat of a tradition for me to spend some time in Boston each year to work on some combinatorial assembly problem. Each of those visits has been crucial to the development of my thinking. I have the privilege of having met a mentor with whom I have an unusual degree of intellectual affinity. Walter has known exactly what kind of problems and subjects to stimulate my mind with and I have benefited immensely from his guidance. Walter is not officially my Ph.D. co-advisor, but he may as well be.

The support that my mentors, Erik Winfree and Walter Fontana have showed me over the years has enabled me to accomplish what, as a naive kid, I dared to believe was possible.

During the course of my Ph.D., I have had the fortune to meet a number of kind and interesting people from which I have learned a great deal and who played an important role in my experience as a graduate student. Many of the ideas that I worked on as a graduate student were inspired by conversations with friends and collaborators Daniele Cappelletti, David Anderson, Manoj Gopalkrishnan, Tom Ouldridge, William Poole, and Abhishek Behera, as well as with friends, lab-mates, and former lab members Robert Johnson, Chris Thachuck, Dave Doty, Damien Woods, Frits Dannenberg, and Stephan Badelt. Some of my best experiences as a graduate student were the conferences and workshops on chemical reaction networks that I was invited to by David Anderson and Daniele Cappelletti, as well as the DNA Computing and Molecular Programming Conferences that I attended. It is a unique experience to be in a room where most people know the deficiency-zero theorem, among other curiosities.

I feel honored to have had the privilege of belonging to Caltech's unique intellectual environment and hanging out in its beautiful campus. I would like to thank professors Paul Sternberg, Lea Goentoro, Paul Rothemund, Justin Bois, and Rob Phillips for helping me believe that I could make it in Caltech's competitive environment. The work and generosity of the Division for Biology and Biological Engineering, as well as all of the Caltech staff, played an essential role in ensuring that my Ph.D. experience was as smooth and pleasant as possible. My time at Caltech would not have been the same without Caltech Arts. I am gratful to them for encouraging me to make art and for showcasing my work. I thank the Caltech Center for Diversity, and especially Taso Dimitriadis and Monique Richards, for their friendship and support through difficult times. I also like to thank my therapist Georgina Moncho for helping me stay afloat throughout this journey that was graduate school.

An integral component of my graduate school experience was the countless good times and conversations I had with my fellow Mexican Caltechers Manuel Razo, Porfirio Quintero, Jorge Castellanos, Alejandro Granados, David Larios, Manolo Flores and Enrique Amaya. Thank you for making it feel like we had a little bit of home in Pasadena. Manuel, Porfirio, and I started this journey together and their friendship was a constant without which I could not have prescinded. I have deep

admiration for both of them and I am proud to call them my friends. I am grateful to Manuel for introducing me to many fascinating ideas that have played an important role in my thinking and for the many great conversations we have had over the years. I have also developed a close friendship with David, Manolo, and Enrique. I will never forget the glorious days of philosophy, laughter, and creativity I have spent with them. I developed an especially close friendship with Enrique as we were living together when the world stopped in March of last year. He was also there throughout most of the process of writing my thesis. I will be forever grateful for the warmth and openness he demonstrated to me when I most needed it.

I am grateful to the Santa Fe Institute (SFI) for allowing me to join them as a postdoctoral fellow despite my not actually having a PhD yet. I would also like to thank the people at SFI for being such an amazing and welcoming community.

The University of Texas at El Paso (UTEP), that Bhutanese-style fortress decorating the Franklin mountain and overseeing the border region, was no less than a spiritual and intellectual haven for me. At the time when I was an undergraduate at UTEP my beloved natal city of Juárez was effectively a war zone. I had the privilege and honor to be welcomed at UTEP where I could escape the chaos and violence that afflicted my hometown. I benefited immensely from the generosity of professors who selflessly fostered my passion for intellectual work. As far as I know Professor Piotr Wojciechiowski initiated me into the mysteries of abstract mathematics and mathematical foundations. Professor Vladik Kreinovich was the wise master who first helped me bring order to my mathematical ideas and place them in the context of the real world. The generosity, passion and encouragement of Professors Min-Ying Leung, Jorge López, Vivian Incera, and Luis Valdez was the catalyst that enabled me to raise above the expectations of someone from my background. I will be forever grateful to that beautiful institution.

I warmly thank my family, including my sisters Veronica, Daniela, and Adriana, who have always been there for me. My grandparents, and all my aunts and uncles who have always believed in me, even when I did not. Finally I thank my parents for their unconditional love, for believing in me, and for all their hard work in the name of giving us an opportunity to get ahead. May it be so. My father who from an early age instilled in me a sense of wonder about the world and my own mind, and which still drives me to this day. My mother, who with her hard-earned pesos helped me pay for books, tuition, food, or whatever I needed when I was an undergraduate at UTEP. The love of my family is the force that drives me and that lets me rise again every time I fall.

I would especially like to thank my best friend and life partner Adam Gomez, for staying with me throughout this journey, even at the most difficult and uncertain times. I could not have done this without him.

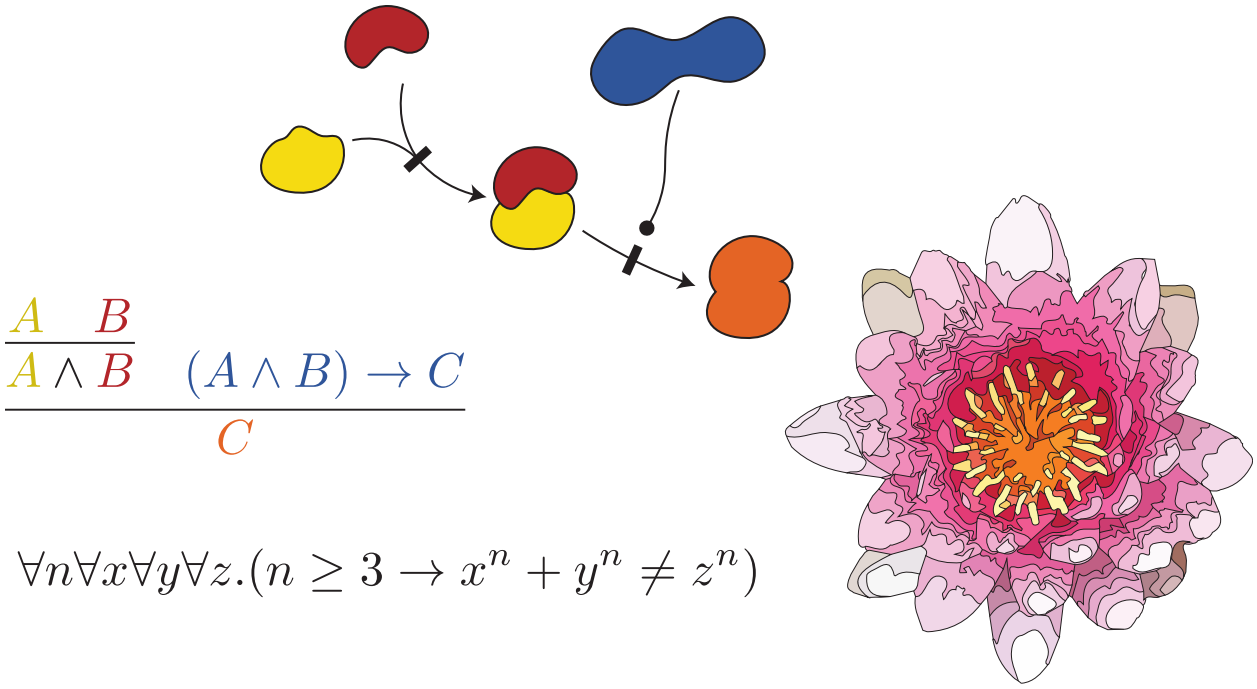


Figure 0.1: Flowers are theorems. My vision of mathematical biology consists of blurring the distinction between mathematics and biology. Mathematical theorems emerge from alphabets, syntax, and deduction. Biological structures emerge from biomolecules, chemistry, and catalysis.

I dedicate this thesis to the memory of my uncles Oscar Muñoz and Iván Muñoz, whom so suddenly left us last year as a result of the ongoing pandemic. Their music and laughter are well alive in the hearts of all of us who loved them. At an early age when I learned the division algorithm my uncle Iván taught me its *proof*: multiply the quotient by the divisor and add the remainder, the result should be the dividend. The thrill I felt when doing that became the same passion that continues to lead my mathematical career. It is the same thrill I felt when I proved the theorems contained in this thesis.

ABSTRACT

Stochastic chemical reaction networks (SCRNs) are a mathematical model which serves as a first approximation to ensembles of interacting molecules. SCRNs approximate such mixtures as always being well-mixed and consisting of a finite number of molecules, and describe their probabilistic evolution according to the law of mass-action. In this thesis, we attempt to develop a mathematical formalism based on formal power series for defining and analyzing SCRNs that was inspired by two different questions. The first question relates to the equilibrium states of systems of polymerization. Formal power series methods in this case allow us to tame the combinatorial complexity of polymer configurations as well as the infinite state space of possible mixture states. Chapter 1 presents an application of these methods to a model of polymerizing scaffolds. The second question relates to the expressive power of SCRNs as generators of stochasticity. In Chapter 2, we show that SCRNs are universal approximators of discrete distributions, even when only allowing for systems with detailed-balance. We further show that SCRNs can exactly simulate Boltzmann machines. In Chapter 3, we develop a formalism for defining the semantics of SCRNs in terms of formal power series which grew as a result of work included in the previous chapters. We use that formulation to derive expressions for the dynamics and stationary states of SCRNs. Finally, we focus on systems that satisfy complex balance and conservation of mass and derive a general expressions for their factorial moments using generating function methods.

PUBLISHED CONTENT AND CONTRIBUTIONS

Cappelletti, Daniele et al. (2020). "Stochastic chemical reaction networks for robustly approximating arbitrary probability distributions". In: *Theoretical Computer Science* 801, pp. 64–95. ISSN: 0304-3975. DOI: 10.1016/j.tcs.2019.08.013.
A.O.M. participated in the writing of the manuscript, designed the *Point Mass* CRN, carried out the proofs in sections 1-3, and helped in the conception of the constructions in the paper.

Ortiz-Muñoz, Andrés, Héctor F. Medina-Abarca, and Walter Fontana (2020). "Combinatorial protein–protein interactions on a polymerizing scaffold". In: *Proceedings of the National Academy of Sciences* 117.6, pp. 2930–2937. ISSN: 0027-8424. DOI: 10.1073/pnas.1912745117.
A.O.M. developed the mathematical formalism of the paper, and participated in writing of the manuscript and in the conception of the project.

Poole, William et al. (2017). "Chemical Boltzmann Machines". In: *DNA Computing and Molecular Programming*. Ed. by Robert Brijder and Lulu Qian. Cham: Springer International Publishing, pp. 210–231. ISBN: 978-3-319-66799-7. DOI: 10.1007/978-3-319-66799-7_14.
A.O.M. participated in the conception of the project, in writing of the manuscript, and developed the *edge species chemical Boltzmann machine* model.

TABLE OF CONTENTS

Acknowledgements . . . iv
Abstract . . . viii
Published Content and Contributions . . . ix
Table of Contents . . . x
List of Illustrations . . . xi
Overview . . . 1
Chapter I: Combinatorial Protein-Protein Interactions on a Polymerizing Scaffold . . . 8
Preface . . . 8
Abstract . . . 16
1.1 Introduction . . . 17
1.2 The polymerizing scaffold system . . . 17
1.3 The chemostatted case . . . 19
1.4 The continuum case in equilibrium . . . 20
1.5 The discrete case in equilibrium . . . 26
1.6 Main conclusions . . . 30
1.7 Supplementary information . . . 31
Chapter II: Universal Approximation of Discrete Distributions . . . 65
Abstract . . . 70
2.1 Introduction . . . 71
2.2 Preliminaries . . . 73
2.3 Dynamics . . . 76
2.4 Reacting systems . . . 80
2.5 Boltzmann machines . . . 83
2.6 Finite-support distributions . . . 85
2.7 Universality . . . 88
2.8 Discussion . . . 89
Chapter III: Formal Semantics for Stochastic Chemical Reaction Networks . . . 93
Abstract . . . 99
3.1 Introduction . . . 100
3.2 Preliminaries . . . 101
3.3 Dynamics . . . 105
3.4 Stationarity . . . 115
3.5 Factorial moments . . . 117
3.6 Assembly . . . 120
3.7 Discussion . . . 122

LIST OF ILLUSTRATIONS

Number *Page*

0.1 Flowers are theorems. My vision of mathematical biology consists of blurring the distinction between mathematics and biology. Mathematical theorems emerge from alphabets, syntax, and deduction. Biological structures emerge from biomolecules, chemistry, and catalysis. . . . vii

1.1 Model of asymmetric linear polymerization. Polymers of arbitrary length bind asymmetrically and polymers can break along any of their bonds. . . . 9

1.2 State space for linear polymerization system with 4 protomers. Each of these states corresponds to a class of microstates in which the constituting protomers are distinguishable from ona another. . . . 12

1.3 Assembled structure consisting of a scaffold joined to 4 particles via 4 bonds. . . . 13

1.4 Enzyme-substrate interaction on a polymeric scaffold . . . 18

1.5 Catalysis in a chemostatted polymerizing scaffold system . . . 20

1.6 Catalysis in a closed polymerizing scaffold system . . . 21

1.7 Multivalent scaffolds and polymerizing scaffold . . . 24

1.8 Maximer . . . 27

1.9 Multivalent and polymerizing scaffolds in the discrete case . . . 29

1.10 Scaffold polymers and multivalent agents . . . 36

1.11 The dependence of the length distribution on the protomer concentration t_S and the affinity σ . . . 38

1.12 Concentrations within length classes . . . 39

1.13 Mixtures of multivalent scaffolds . . . 40

1.14 Polymerizing scaffold and multivalent scaffolds . . . 42

1.15 Interaction horizon . . . 43

1.16 Interaction horizon scenarios . . . 46

1.17 The impact of the interaction horizon . . . 47

1.18 Length distribution in continuum and discrete polymerization . . . 51

1.19 Finite size 1D bond percolation and polymerization . . . 54

1.20 Scaling behavior of the maximer distribution . . . 55

1.21 Effects in discrete and continuum polymerizing scaffold systems . . . 57

1.22 Catalytic potential of multivalent scaffolds (discrete case) . . . 58

1.23 Catalytic potential of multivalent scaffolds (continuum case) . . . 58

1.24 Stochastic simulations 59
1.25 Variance and noise 60
1.26 Distributions of catalytic potential 61
2.1 Graphical representation of a CRN. Each black rectangle represents a reaction with the tails of the arrows being connected to the reactants of the reaction and the head being connected to the products. Next to each reaction we show the corresponding rate constant. 75
2.2 Example stochastic CRN. This system has 2 species, represented by a blue circle and a yellow triangle, and two reactions, given by the reaction diagram. On the top right is the master equation for the probability of being in state $(2, 2)$. At the bottom left corner is a diagram representing the state space of the system, with the incoming and outgoing transitions in blue and red. The plot on the right shows the probability distribution for the reachable states, with transition arrows as before. 78
2.3 Illustration of detailed balance for CRNs. Here we denote the β in Definition 2.4.2 with ε to highlight its connection to energy. In particular, the negative of its natural log corresponds to chemical potential. 83
2.4 Illustration of Lemma 2.4.1. The stationary distributions of two detailed-balanced CRNs are shown. On the top, all states are reachable and the stationary distribution is a multivariate Poisson. On the bottom the reactions give rise to reachability classes and hence the stationary distribution is a truncated multivariate Poisson. 84
2.5 On the left is shown a Boltzmann machine transition. Nodes 3, 4, 5, and 6 are initially on as well as their joining edges. Subsequently, node 1 is turned on and the additional joining edges are turned on as well. On the right is shown the corresponding reaction in the Boltzmann CRN. The circles correspond to the species X_i either on or off, and the rectangles correspond to the species $W_{i,j}$, also on or off. The grayed-out species correspond to the species that do not participate in the reaction but that are present in the mixture. 86
2.6 Illustration of a variant in the construction in Definition 2.6.1. The construction illustrated here follows the same principle as that of Definition 2.6.1, which is that each point in the support is associated with a hidden species, shown in black, and a number of reactions that inter-convert between the hidden species, while updating the counts of the visible species appropriately. The rate constants are set up so that the desired probabilities are obtained in steady state. 87
2.7 A distribution over two-dimensional state space. 88

3.1 Illustration of the combinatorics of generating functions. One way to see generating functions is as each monomial representing a structure. In particular, we can interpret x^n as a unary string of length n. In this example we see how we can exploit algebraic tricks in order to obtain combinatorial insights. 95
3.2 Illustration of probability generating functions. On the top is a general probability generating function and its combinatorial interpretation. Below is the probability generating function of a multivariate Poisson, which is given by a product of exponentials. 109
3.3 Illustration of reaction operator. The terms of the infinitesimal stochastic operator are reaction operators. Their combinatorial interpretation is shown here for an example reaction. 110
3.4 Pictorial representation of the exponential solution to the chemical master equation and its derivation. 113
3.5 Pictorial representation of the waiting operator and of the derivation of the path integral solution to the CME. 115
3.6 Example of a conservation class function. Here the conservation class function is denoted with W instead of ω. The function W was featured in Chapter 1 and it is the generating function of scaffold complexes. Each variable is color-coded to correspond to a unit in the structures. 121

OVERVIEW

Whereas mathematical physics has succeeded in making impressive predictions about the phenomena it is meant to model, biology has not yet seen mathematical models with such degree of predictive power, let alone the unification that theories of physics have achieved. The lack of predictive *and* unifying models in biology in comparison with physics reflects the bewildering difference in the complexity of the phenomena they are concerned with.

Still, the study of molecular interaction networks has helped theorists gain some perspective on and understanding of the inner workings of living cells. The mathematical machinery used in the analysis of such networks is that of *chemical reaction networks* (CRNs) (Feinberg, 1972; Horn, 1972; Horn and Jackson, 1972). A key feature of CRNs is their high level of abstraction in which molecules are devoid of internal structure and represented by real-valued variables reporting their concentration. The qualifier *chemical* is therefore slightly inappropriate since a CRN does not capture the combinatorial and generative aspects that are characteristic of actual chemistry. Rather, in a CRN, the set of molecular types is merely a list of proper names and all reactions must be specified at the outset; they are explicitly stated in the model as opposed to being implicit by virtue of chemical reactivity linked to structure. A CRN is chemical only in that its kinetics are based on the *law of mass action*. CRNs are therefore inherently *phenomenological* models (Gunawardena, 2014). In other words, they constitute falsifiable hypotheses about mechanisms and reactivities of a mixture. The predictive power of CRNs is limited by the extent to which their assumptions are correct about the systems they are meant to describe.

An important consideration in modeling biomolecular systems is their inherent *stochasticity* as a result of small molecular counts. This requires a stochastic generalization of the traditional ordinary differential equation semantics for CRNs. Such models are known as *stochastic chemical reaction networks* (SCRNs) (Anderson and Kurtz, 2015; D. Gillespie, 1976; D. T. Gillespie, 1977; Van Kampen, 1992). SCRNs are *more* fundamental than deterministic CRNs in the sense that the latter can be derived from the former as the right limit of large volumes (Kurtz, 1972). Yet, SCRNs remain agnostic to the combinatorial nature of biochemical structures. A description of reaction mixtures that can address both stochasticity and combinatorics would be more faithful to the reality of molecular biology and hence more predictive. Already a number of abstract models of reaction networks exist that incorporate both stochastic and combinatorial aspects (Benkö, Flamm, and Stadler, 2003; Blinov et al., 2004; Danos et al., 2007; Johnson and Winfree, 2020; Phillips and Cardelli, 2009). In order to understand what these systems are capable of and how they relate to one another, we will need a general theory of stochastic/combinatorial reaction networks that can accommodate the various existing models.

In this thesis, I develop a formal approach to CRNs that serves as a precursor to a general, combinatorial theory of reaction networks. I conceived these ideas in the context of applications that incorporate stochastic and combinatorial aspects (Cappelletti et al., 2020; Ortiz-Muñoz, Medina-Abarca, and Fontana, 2020; Poole et al., 2017). The approach is based on *formal power series*, which are often used for counting general classes of combinatorial objects such as trees, graphs, strings, etc. (Flajolet and Sedgewick, 2009; Wilf, 1994). I believe that the versatility of formal power series can be harnessed to incorporate general graphical models of reaction networks. We will now proceed to overview the structure of the thesis.

Chapter 1 consists of a publication that arose from explorations in the context of abstract models of polymerization aimed at elucidating the role polymerizing scaffolds might play in cellular signaling (Ortiz-Muñoz, Medina-Abarca, and Fontana, 2020). A system capable of polymerization can in principle generate an unlimited number of different molecule types, rendering a traditional CRN model infeasible. Although it would be possible to write equations describing the polymerization dynamics, the infinitely many distinct types of possible molecules would require more careful considerations of limits that standard CRN theory is not equipped to handle. In the paper, we restrict our attention to stationary states so that we can circumvent some of the difficulty inherent in computing dynamic solutions. My contribution to this paper was in the development of the mathematical formalism for computing equilibrium concentrations when the model is conceived deterministically (i.e. with continuous concentrations and hence no limit as to the maximal polymer length), as well as probabilities and expectation values when conceived stochastically (i.e. with discrete particle numbers and hence a limit on the maximal size of polymers). The bulk of complexity of the problem lies in considering mass conservation constraints. Methods based on generating functions are capable of handling the combinatorial complexity of polymers as well as their stochastic equilibria. I further develop these techniques in Chapter 3.

In addition to using CRNs as a means for understanding chemical or biological phenomena, another conception of CRNs is as models of computation (Chen, Doty, and Soloveichik, 2014a,b; Cook et al., 2009; Cummings, Doty, and Soloveichik, 2014; Soloveichik et al., 2008). In this case, the assumptions of the CRN model are taken to be *ideal* and real systems as approximations of the resulting ideal behavior. From this perspective if a computation can be embedded into the ideal mathematical behavior of a CRN, then an engineered chemical system that approximates its assumptions will also approximate that computation. Hence, in this context, CRNs are not so much meant to be predictive than prescriptive. Of particular interest is the ability of stochastic CRNs to perform probabilistic inference as this would enable engineered chemical systems to act "intelligently" in some loose sense. Since probabilistic inference depends on the ability to represent distributions, an important matter is the scope of distributions that can be observed in SCRNs.

Chapter 2 reports my perspective and contributions to two publications aimed at exploring the probabilistic expressive power of SCRNs (Cappelletti et al., 2020; Poole et al., 2017). In Cappelletti et al., 2020, we show that in fact SCRNs are capable of approximating any desired discrete distribution. This remains true even in the case when the CRNs are required to satisfy *detailed balance*. A hypothetical combinatorial reaction network model would be at least as expressive as CRNs so that it suffices to show the universality of CRNs in order to establish that of more general combinatorial models. In Poole et al., 2017, we show that SCRNs are also capable of faithfully reproducing the equilibria of Boltzmann Machines (BM). Although the model used there is that of CRNs, the existence of an underlying graphical structure in the BM invites a combinatorial model. My contribution to this paper was in the form of a CRN model that exploits this combinatorial aspect to simulate BMs exactly while preserving detailed balance.

In Chapter 3, I develop a formalism based on formal power series to define the stochastic semantics of CRNs following the proposal in Baez and Biamonte, 2018. Most of the material in this chapter was conceived as applications to the projects I was a part of as a graduate student. Of central importance to my motivation was the fact that the combinatorics and stochasticity of reaction networks can be readily handled by formal power series methods. The first part of the chapter focuses on formal expressions for representing the dynamics of SCRNs. The second half focuses on stationary solutions to complex-balanced SCRNs. I develop a generalization of the formalism presented in Chapter 1 with the purpose of computing equilibrium factorial moments of general assembly systems.

Each chapter in the thesis is preceded by a preface in which I present my personal perspective of the chapter as well as the way in which some of the ideas were conceived. This is in addition to the technical introduction that places the work in its scholarly context.

I believe that in order to move in the direction of a more unified theory of mathematical biology we may need to rehash the current foundations of mathematics to bring them closer to the objects of biology. Just as metabolic pathways, such as the citric acid cycle, are conserved across species, there are mathematical themes that are pervasive throughout mathematics. Category theory (CT) is a mathematical discipline and a foundation of mathematics aimed at the study of such universal themes and analogies between mathematical disciplines (Mac Lane, 2013; Spivak, 2014). Naturally CT has been recognized for its ability to unify concepts and theories. One example of this kind of unification is Lawvere's fixed-point theorem, which has as special cases all the classical paradoxes of self reference such as Cantor's theorem, Russell's paradox, Turing's halting problem, and Gödel's incompleteness theorem (Lawvere, 1969; Yanofsky, 2003). The theory of *species* uses CT methods to clarify the streamlining role that formal power series play in combinatorics (Bergeron, Labelle, and Leroux, 1997; Joyal, 1981). Although not mentioned in the thesis, the formal power series

semantics developed in Chapter 3 lends combinatorial semantics to SCRNs via species theory and places SCRNs in the context of CT. This was proposed by Baez and Dolan, 2001, in the context of quantum mechanics in order to formalize the concept of Feynman diagrams. For a while I have speculated about the role that Lawvere's fixed point theorem could play in the mathematical understanding of recursive phenomena in biology such as replication, the origins of life, and even novel models of chemical computation. Formulating the theory of SCRNs in categorical language would allow me to explore such questions. These are ideas that I intend to explore in the future.

BIBLIOGRAPHY

Anderson, David and Thomas Kurtz (2015). *Stochastic Analysis of Biochemical Systems*. Stochastics in Biological Systems. Springer International Publishing. ISBN: 9783319168951. DOI: 10.1007/978-3-319-16895-1.

Baez, John and Jacob Biamonte (2018). *Quantum Techniques in Stochastic Mechanics*. World Scientific. DOI: 10.1142/10623.

Baez, John and James Dolan (2001). "From Finite Sets to Feynman Diagrams". In: *Mathematics Unlimited — 2001 and Beyond*. Ed. by Björn Engquist and Wilfried Schmid. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 29–50. DOI: 10.1007/978-3-642-56478-9_3.

Benkö, Gil, Christoph Flamm, and Peter F. Stadler (2003). "A Graph-Based Toy Model of Chemistry". In: *Journal of Chemical Information and Computer Sciences* 43.4. PMID: 12870897, pp. 1085–1093. DOI: 10.1021/ci0200570.

Bergeron, François, Gilbert Labelle, and Pierre Leroux (1997). Trans. by Margaret Readdy. Encyclopedia of Mathematics and its Applications. Cambridge University Press. DOI: 10.1017/CBO9781107325913.

Blinov, Michael L. et al. (2004). "BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains". In: *Bioinformatics* 20.17, pp. 3289–3291. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bth378.

Cappelletti, Daniele et al. (2020). "Stochastic chemical reaction networks for robustly approximating arbitrary probability distributions". In: *Theoretical Computer Science* 801, pp. 64–95. DOI: 10.1016/j.tcs.2019.08.013.

Chen, Ho-Lin, David Doty, and David Soloveichik (2014a). "Deterministic function computation with chemical reaction networks". In: *Natural Computing* 13.4, pp. 517–534. DOI: 10.1007/s11047-013-9393-6.

– (2014b). "Rate-Independent Computation in Continuous Chemical Reaction Networks". In: ITCS '14. Princeton, New Jersey, USA: Association for Computing Machinery, pp. 313–326. ISBN: 9781450326988. DOI: 10.1145/2554797.2554827.

Cook, Matthew et al. (2009). "Programmability of Chemical Reaction Networks". In: *Algorithmic Bioprocesses*. Ed. by Anne Condon et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 543–584. ISBN: 978-3-540-88869-7. DOI: 10.1007/978-3-540-88869-7_27.

Cummings, Rachel, David Doty, and David Soloveichik (2014). "Probability 1 Computation with Chemical Reaction Networks". In: *DNA Computing and Molecular Programming*. Ed. by Satoshi Murata and Satoshi Kobayashi. Cham: Springer International Publishing, pp. 37–52. ISBN: 978-3-319-11295-4. DOI: 10.1007/978-3-319-11295-4_3.

Danos, Vincent et al. (2007). "Rule-Based Modelling of Cellular Signalling". In: *CONCUR 2007 – Concurrency Theory*. Ed. by Luís Caires and Vasco T. Vasconcelos. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 17–41. DOI: 10.1007/978-3-540-74407-8_3.

Feinberg, Martin (1972). "Complex balancing in general kinetic systems". In: *Archive for Rational Mechanics and Analysis* 49.3, pp. 187–194.

Flajolet, Philippe and Robert Sedgewick (2009). *Analytic Combinatorics*. Cambridge University Press. ISBN: 9781139477161.

Gillespie, Daniel (1976). "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions". In: *Journal of Computational Physics* 22.4, pp. 403–434.

Gillespie, Daniel T (1977). "Exact stochastic simulation of coupled chemical reactions". In: *The Journal of Physical Chemistry* 81.25, pp. 2340–2361.

Gunawardena, Jeremy (2014). "Models in biology:'accurate descriptions of our pathetic thinking'". In: *BMC biology* 12.1, p. 29.

Horn, Fritz (1972). "Necessary and sufficient conditions for complex balancing in chemical kinetics". In: *Archive for Rational Mechanics and Analysis* 49.3, pp. 172–186.

Horn, Fritz and Roy Jackson (1972). "General mass action kinetics". In: *Archive for Rational Mechanics and Analysis* 47.2, pp. 81–116. DOI: 10.1007/BF00251225.

Johnson, Robert F. and Erik Winfree (2020). "Verifying polymer reaction networks using bisimulation". In: *Theoretical Computer Science* 843, pp. 84–114. ISSN: 0304-3975. DOI: 10.1016/j.tcs.2020.08.007.

Joyal, André (1981). "Une théorie combinatoire des séries formelles". In: *Advances in Mathematics* 42.1, pp. 1–82. ISSN: 0001-8708. DOI: 10.1016/0001-8708(81)90052-9.

Kurtz, Thomas (1972). "The relationship between stochastic and deterministic models for chemical reactions". In: *The Journal of Chemical Physics* 57.7, pp. 2976–2978.

Lawvere, F. William (1969). "Diagonal arguments and cartesian closed categories". In: *Category Theory, Homology Theory and their Applications II*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 134–145. ISBN: 978-3-540-36101-5. DOI: 10.1007/BFb0080769.

Mac Lane, Saunders (2013). *Categories for the working mathematician*. Vol. 5. Springer Science & Business Media.

Ortiz-Muñoz, Andrés, Héctor F. Medina-Abarca, and Walter Fontana (2020). "Combinatorial protein–protein interactions on a polymerizing scaffold". In: *Proceedings of the National Academy of Sciences* 117.6, pp. 2930–2937. ISSN: 0027-8424. DOI: 10.1073/pnas.1912745117.

Phillips, Andrew and Luca Cardelli (2009). "A programming language for composable DNA circuits". In: *Journal of the Royal Society Interface* 6, S419–S436.

Poole, William et al. (2017). "Chemical Boltzmann Machines". In: *DNA Computing and Molecular Programming*. Ed. by Robert Brijder and Lulu Qian. Cham: Springer International Publishing, pp. 210–231. DOI: 10.1007/978-3-319-66799-7_14.

Soloveichik, David et al. (2008). "Computation with finite stochastic chemical reaction networks". In: *Natural Computing* 7.4, pp. 615–633. DOI: 10.1007/s11047-008-9067-y.

Spivak, David I. (2014). *Category Theory for the Sciences*. MIT Press. ISBN: 9780262320535.

Van Kampen, Nicolaas Godfried (1992). *Stochastic processes in physics and chemistry*. Vol. 1. Elsevier.

Wilf, Herbert S. (1994). In: *generatingfunctionology*. Ed. by Herbert S. Wilf. 2nd ed. San Diego: Academic Press. ISBN: 978-0-08-057151-5. DOI: 10.1016/B978-0-08-057151-5.50003-4.

Yanofsky, Noson S. (2003). "A Universal Approach to Self-Referential Paradoxes, Incompleteness and Fixed Points". In: *Bulletin of Symbolic Logic* 9.3, pp. 362–386. DOI: 10.2178/bsl/1058448677.

Chapter 1

COMBINATORIAL PROTEIN-PROTEIN INTERACTIONS ON A POLYMERIZING SCAFFOLD

Andrés Ortiz-Muñoz[a,1], Héctor F. Medina-Abarca[b,1], and Walter Fontana[b,2]

[a] California Institute of Technology, Pasadena, CA 91125

[b] Systems Biology, Harvard Medical School, Boston, MA 02115

[1] both authors contributed equally

[2] to whom correspondence should be addressed: walter_fontana@hms.harvard.edu

PREFACE

In the summer of 2013, I was a summer intern at Harvard Systems Biology under the mentorship of Professor Walter Fontana. My project that summer revolved around roughly the following question: suppose you have a system of particles that can polymerize linearly and asymmetrically. If the total number of particles in the system is conserved, what is the equilibrium concentration of a polymer of length n? Below I will describe the steps I followed in answering the question and the insights I obtained in the process.

We will consider the system defined by the following set of reactions

$$A_m + A_n \underset{k_-}{\overset{k_+}{\rightleftharpoons}} A_{m+n},$$

where A_i denotes a polymer of length i. This model assumes that any two polymers may bind to form a longer polymer, and that a polymer may break into any two polymers with the same total length (See Figure 1.1). Furthermore, the model assumes that the binding and dissociation rate constants are independent of the lengths of the polymers involved. Other mechanisms of polymerization are possible, but this is the one I felt was sufficiently complex while remaining mathematically tractable. The system of ordinary differential equations (ODEs) that results from this mechanism and assuming the law of mass action is

$$\frac{da_n}{dt} = \sum_{i=1}^{n-1} k_+ a_i a_{n-i} + 2\sum_{i=1}^{\infty} k_- a_{i+n} - 2\sum_{i=1}^{\infty} k_+ a_i a_n - \sum_{i=1}^{n-1} k_- a_n, \tag{1.1}$$

where a_i denotes the concentration of A_i at time t. This system bears similarities to Smoluchowski's coagulation equation, with the exception that said equation does not incorporate polymer fission (Smoluchowski, 1916). My summer project amounted to solving the system in Equation 1.1 in equilibrium, where the derivatives vanish.

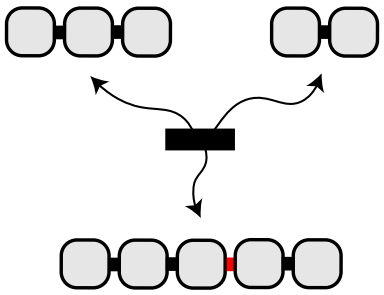


Figure 1.1: Model of asymmetric linear polymerization. Polymers of arbitrary length bind asymmetrically and polymers can break along any of their bonds.

Equation 1.1 simplifies significantly with the observation that, in equilibrium, it admits solutions satisfying for all natural numbers $m, n \geq 1$

$$k_- a_{m+n} = k_+ a_m a_n,$$

which is a much simpler system to solve. In this case, a_n denotes the equilibrium concentration of A_n. That the system can be cast in this simple form is a result of of the fact that it satisfies *detailed balance* (Horn and Jackson, 1972; Onsager, 1931; Wegscheider, 1902). The resulting system of equations can be solved via a recursion in which one obtains the concentration of polymers A_n in terms of the concentration of monomers A_1 as follows

$$a_n = \kappa^{n-1} a_1^n,$$

where $\kappa = k_+/k_-$ is an *association constant*. For convenience, let us use the unit-less variables $\alpha_n = \kappa a_n$ so that the above equation becomes

$$\alpha_n = \alpha_1^n. \tag{1.2}$$

Since all concentrations are expressed in terms of the (unit-less) monomer concentration α_1, it only remains to determine its value in order to solve for the concentrations of all polymer lengths. We assume that the total number of protomers is conserved and known. Since each polymer of length i contains i protomers, we must have that

$$\alpha = \sum_{i=1}^{\infty} i\alpha_i, \tag{1.3}$$

where α denotes the (unit-less) total protomer concentration in the mixture, i.e. the concentration of monomers when no polymer has formed yet. Notice that using Equation 1.2, we can express the summand in Equation 1.3 above in terms of the following derivative

$$i\alpha_i = i\alpha_1^i = \alpha_1 \frac{d\alpha_1^i}{d\alpha_1}.$$

Applying the same derivative and multiplication by α_1 to a geometric series of powers of α_1, we can find an expression for the known total protomer concentration α_1 in terms of the unknown variable α_1

$$\alpha = \alpha_1 \frac{d}{d\alpha_1} \sum_{i=1}^{\infty} \alpha_1^i = \alpha_1 \frac{d}{d\alpha_1} \left(\frac{\alpha_1}{1-\alpha_1} \right) = \frac{\alpha_1}{(1-\alpha_1)^2}. \tag{1.4}$$

This equation is quadratic in α_1 so it can be solved explicitly yielding

$$\alpha_1 = \alpha \left(\frac{1 - \sqrt{1+4\alpha}}{2\alpha} \right)^2.$$

Finally, the concentration of polymers of length n in terms of the unit-less total concentration α is given by[1]

$$\alpha_n = \alpha^n \left(\frac{1 - \sqrt{1 + 4\alpha}}{2\alpha} \right)^{2n}.$$

Pleased with the outcome of my summer project, I returned to my undergraduate institution, the University of Texas at El Paso, for one last semester. I graduated in December of that year, which meant that I had a few months before starting graduate school, whcih I spent working in the Fontana lab. Motivated by my summer project, I decided that in those months I would pursue a stochastic generalization of the same polymerization system whose deterministic formulation I had previously cracked.

The stochastic analog to Equation 1.1 is the *chemical master equation* (CME) (Van Kampen, 1992). The CME is an equation that defines the evolution of a probability distribution over discrete counts of molecules, or *states*, according to the reaction mechanism of a system. The state space of a system of linear polymerization is fairly complex. Given a finite amount of protomers the state space is the set of different ways of distributing those protomers into polymers. The size of that set corresponds to what in number theory is known as the *integer partitions* of n (Andrews, 1984). For example, if a total of 4 protomers exist in a mixture, the state space corresponds to the 5 different ways of expressing the number 4 as a sum of positive integers: $1 + 1 + 1 + 1$, $1 + 1 + 2$, $1 + 3$, $2 + 2$, and 4. These 5 sums can also be seen as states where each term is the length of a polymer. For example, the sum $1 + 1 + 2$ corresponds to a state with two monomers and a dimer.

As opposed to writing down a general expression for the CME, which would have been extraordinarily complex, I opted to solving a small case explicitly and sought generalizations of that simple solution. As we have already seen, a system with 4 protomers has 5 possible configurations, which in equilibrium results in a system of 5 linear equations. The state space of the system with its transitions is summarized in Figure 1.2. The solution to the resulting system of equations is a 5-dimensional vector of probabilities for each of the states of the state space. It can be obtained through standard linear algebra methods. Rather than writing down here the full 5-dimensional vector solution, we will focus our attention on the following common denominator to all 5 entries of that vector

$$Z_4 = 1 + 12\kappa + 36\kappa^2 + 24\kappa^3. \tag{1.5}$$

[1] An interesting property of this system is that when we have $\alpha = 1$, the equilibrium concentrations of polymers can be written in terms of even powers of the golden ratio

$$\alpha_n = \left(\frac{1 - \sqrt{5}}{2} \right)^{2n}.$$

I have never made much of this curious result, but it was a lovely way to conclude my summer project.

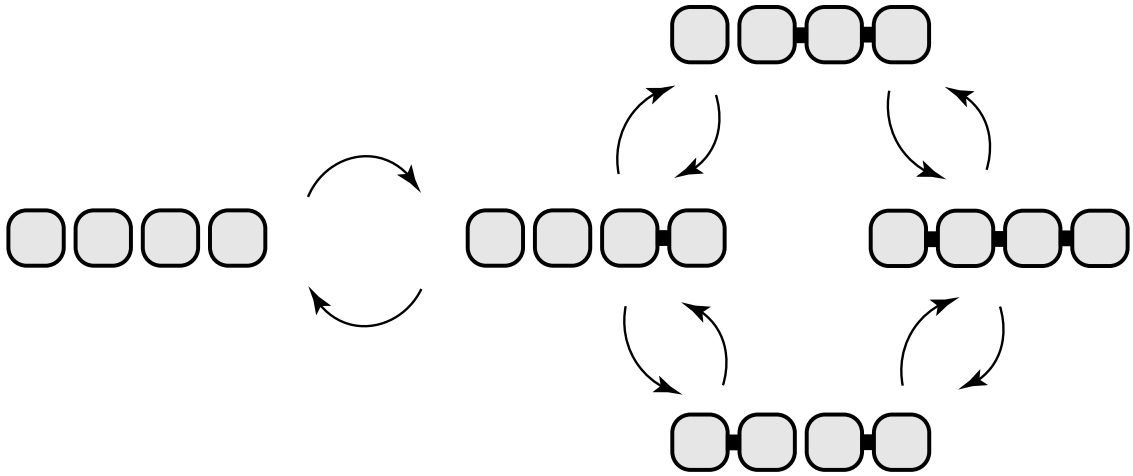


Figure 1.2: State space for linear polymerization system with 4 protomers. Each of these states corresponds to a class of microstates in which the constituting protomers are distinguishable from ona another.

This expression is known as the *partition function* of the system. The partition function is a normalization factor consisting of the sum of weights associated to each state. Each term in Equation 1.5 corresponds to a class of states sharing the same total number of bonds, indicated by the power of κ. The coefficient of the power of κ is the *degeneracy* of the class, given by the number of ways of achieving the corresponding number of bonds and taking into consideration the distinguishability of the protomers. For example, the term $36\kappa^2$ is the contribution to the partition function from states that have a total of 2 bonds, which are represented by the sums 2 + 12 and 1 + 3. Those two states can be collectively realized in 36 different ways using 4 distinguishable protomers. We also say that the class consists of 36 *microstates*.

Since the weight of a single state is easy to calculate —it is given by κ raised to the power of the total number of bonds in the state— knowing the partition function is enough to compute all probabilities. The partition function we calculated above corresponds to a single case, that of having a total of 4 protomers, but its form suggests a hypothesis about the solution to the general case: the partition function of a system with n protomers is given by the sum of Boltzmann terms over all energy states weighted by their degeneracies. This kind of combinatorial reasoning became the basis with which I solved for the equilibria of a number of different assembly systems. Later I would learn that the form of this partition function arises from a general product-of-Poisson pattern of equilibrium distributions for systems with detailed balance (Anderson, Craciun, and Kurtz, 2010; Whittle, 1986).

In addition to working on the stochastic formulation of linear polymerization, I spent my time after undergraduate and before graduate school solving a variety of simple molecular assembly systems

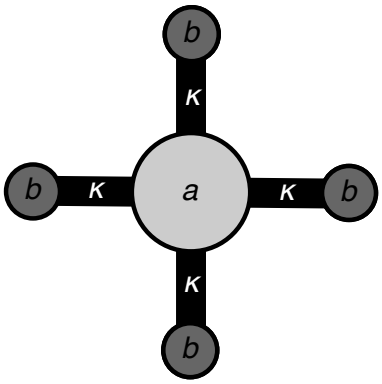


Figure 1.3: Assembled structure consisting of a scaffold joined to 4 particles via 4 bonds.

such as dimerization, scaffolds, branching structures, rings, etc, always assuming that the systems satisfied detailed balance and that the total numbers of protomers were conserved. I observed that, in their deterministic formulations, the equilibrium concentration of an assembled structure was given by the exponential of its energy of formation and the product of the concentrations of monomers of the components it is made of. This was uncanny since it revealed a correspondence between the graphical structure of an assembly and its equilibrium concentration. For example, the assembly in Figure 1.3 would have a concentration of

$$c = \kappa^4 a b^4,$$

where c is the concentration of the assembly, κ is an association constant, a is the concentration of the middle component, and b is the concentration of the outer component. Notice that each of the components of the assembly contributes one factor to the concentration. *At this point, I began to understand that I could reason graphically about the algebraic structures I was working with* and I began to wonder how far I could take that kind of reasoning. For stochastic systems, I observed that the partition function could always be written as a sum over bond counts of the corresponding energy term multiplied by the degeneracy of those bond counts. The complexity of these problems lay mainly in counting those degeneracies for the different assembly systems.

Already as a graduate student, I spent on the order of a week each year in Boston working on progressively more general models of linear polymerization in both their deterministic and stochastic formulations. As already noted, the equilibrium concentration of a polymer can be expressed as the product of concentrations of its constituent monomers and the affinities of its bonds. The problem in those cases was therefore mainly to compute the concentrations of monomers in terms of the known concentrations of total monomers. Of central importance to these problems was the expression for the the total concentration of the system in equilibrium, i.e. the concentration of the mixture regardless of molecular species. Its centrality was owed to the fact that the different

total protomer concentrations appeared as its derivatives, as it was the case in Equation 1.4. To see this notice that, for example, for the assembly in Figure 1.3, which has concentration given by $c = \kappa^4 ab^4$, its contribution to the total concentration of b is given by $4c = b\frac{dc}{db} = 4\kappa^4 ab^4$. As a result, the total concentration of a protomer x in an assembly system can be obtained by applying the operator $x\frac{d}{dx}$ to the sum of the concentrations of all polymers —the mixture concentration. Given the above-mentioned correspondence between the concentration of a polymer and its graphical structure it turns out that the mixture concentration plays the role of a generating function of polymers. The complexity for the deterministic formulations in equilibrium lay mainly in finding simple expressions for this generating function and computing its derivatives.

For their stochastic formulations, I employed the technique delineated above of computing partition functions as sums of Boltzmann terms weighted by degeneracies. The complexity in this case was mainly in computing those degeneracies. Initially I employed a number of different forms of combinatorial reasoning to compute the degeneracies of the various polymerization systems I was considering. Given the increasing complexity of those systems, the bulk of my time in Boston was initially spent in deriving combinatorial schemes for computing degeneracies. Some time in 2016 my mentor Walter Fontana suggested I read the book *Analytic Combinatorics* (Flajolet and Sedgewick, 2009). The book presents generating function methods for counting combinatorial classes of objects as well as the insights that complex analysis brings in approximating their coefficients. I focused mostly on the *combinatorics*, formal methods for writing and manipulating generating functions, and not so much on the *analytics*, methods from complex analysis for extracting numbers from those functions. Still, using those methods I was able to derive in 5 minutes the partition functions of assembly systems that had previously taken me days of hard work to derive. Needless to say being able to compute partition functions so swiftly felt like magic compared to the brute-force methods I had been using before. One of the main insights I obtained was that by taking the exponential of the mixture concentration function from the deterministic formulations, which is also the polymer generating function, I could obtain a generating function for the partition functions themselves. When interpreted combinatorially, the exponential function is the generating function of finite sets, and hence the exponential of the generating function for polymers gives the generating function of multisets of polymers, which correspond to states of the polymerization system.

Not much later, I learned that what I had been doing informally —working with algebraic expressions as if they were combinatorial objects— was in fact elegantly made rigorous in the theory of *combinatorial species* (Bergeron, Labelle, and Leroux, 1997). My desire to better understand the correspondence between algebra and combinatorics led me down a rabbit hole of progressively more fundamental perspectives on combinatorics and formal power series culminating in a formu-

lation in terms of *homotopy type theory* (HoTT) (Univalent Foundations Program, 2013; Yorgey, 2014). Such perspectives lie at the vanguard of contemporary mathematics and it is my hope to gain a better understanding of them so I can contribute to facilitating their assimilation into the standard methods of mathematical biology. None of that deeper, more fundamental perspective perspective appears in the formalism of the article that follows. Rather, the mathematics used there is at the level of what can be found in Flajolet and Sedgewick, 2009. That is the case also for the further development performed in Chapter 3. Formulation of these techniques in the context of HoTT is, however, underway and will be part of my research agenda for years to come.

My contribution to this article was mainly in the form of the mathematical methods, theorems, proofs, and analysis of equations. The writing was mainly done by my mentor Walter Fontana, with the exception of sections 1.7.7 and 1.7.8 of the supplementary material, which were done by me. Although I participated in all aspects of the conception of the project, the biological perspective is mainly due to my collaborators. All plots and simulations were done also by Walter Fontana.

ABSTRACT

Scaffold proteins organize cellular processes by bringing signaling molecules into interaction, sometimes by forming large signalosomes. Several of these scaffolds are known to polymerize. Their assemblies should therefore not be understood as stoichiometric aggregates, but as combinatorial ensembles. We analyze the combinatorial interaction of ligands loaded on polymeric scaffolds, in both a continuum and discrete setting, and compare it with multivalent scaffolds with fixed number of binding sites. The quantity of interest is the abundance of ligand interaction possibilities—the catalytic potential Q—in a configurational mixture. Upon increasing scaffold abundance, scaffolding systems are known to first increase opportunities for ligand interaction and then to shut them down as ligands become isolated on distinct scaffolds. The polymerizing system stands out in that the dependency of Q on protomer concentration switches from being dominated by a first order to a second order term within a range determined by the polymerization affinity. This behavior boosts Q beyond that of any multivalent scaffold system. In addition, the subsequent drop-off is considerably mitigated in that Q decreases with half the power in protomer concentration than for any multivalent scaffold. We explain this behavior in terms of how the concentration profile of the polymer length distribution adjusts to changes in protomer concentration and affinity. The discrete case turns out to be similar, but the behavior can be exaggerated at small protomer numbers because of a maximal polymer size, analogous to finite-size effects in bond percolation on a lattice.

1.1 Introduction

Protein-protein interactions underlying cellular signaling systems are mediated by a variety of structural elements, such as docking regions, modular recognition domains, and scaffold or adapter proteins (Bhattacharyya et al., 2006; Good, Zalatan, and Lim, 2011). These devices facilitate both the evolution and control of connectivity within and among pathways. Since the scaffolding function of a protein can be conditional upon activation and also serve to recruit other scaffolds, the opportunities for plasticity in network architecture and behavior are abundant.

Scaffolds are involved in the formation of signalosomes —transient aggregations of proteins that process and propagate signals. A case in point is the machinery that tags β-catenin for degradation in the canonical Wnt pathway. β-catenin is modified by CK1α and GSK3β without binding any of these kinases directly, but interacting with them through the Axin scaffold (Ikeda et al., 1998; Liu et al., 2002). In addition, the DIX domain in Axin allows for oriented Axin polymers (Fiedler et al., 2011), while APC (another scaffold) can bind multiple copies of Axin (Behrens et al., 1998), yielding Axin-APC aggregates to which kinases and their substrates bind.

By virtue of their polymeric nature, scaffold assemblies like these have no defined stoichiometry and may only exist as statistical ensembles rather than a single stoichiometrically well-defined complex (Deeds et al., 2012; Suderman and Deeds, 2013). As a heterogeneous mixture of aggregates with combinatorial state, the β-catenin destruction system thus appears to be an extreme example of what has been called a "pleiomorphic ensemble" (Mayer, Blinov, and Loew, 2009).

Scaffold-mediated interactions are characteristically subject to the prozone or "hook" effect. At low scaffold concentrations, adding more scaffold facilitates interactions between ligands. Beyond a certain threshold, however, increasing the scaffold concentration further prevents interactions by isolating ligands on different scaffold molecules (Bray and Lay, 1997; Ferrell, 2000; Levchenko, Bruck, and Sternberg, 2000). For a scaffold S that binds with affinity α an enzyme A and a substrate B, present at concentrations t_A and t_B, the threshold is at $1/\alpha + (t_A + t_B)/2$.

In this contribution, we define and analyze a simple model of enzyme-substrate interaction mediated by a polymerizing scaffold. The model does not take into account spatial constraints of polymer chains and therefore sits at a level of abstraction that only encapsulates combinatorial aspects of a pleiomorphic ensemble and briefly peeks down the trail of critical phenomena often associated with phase-separation (Bergeron-Sandoval, Safaee, and Michnick, 2016; Li et al., 2012).

1.2 The polymerizing scaffold system

Let S (the scaffold) be an agent with four distinct binding sites {a,b,x,y}. At site y, agent S can reversibly bind site x of another S with affinity σ, forming (oriented) chains. For the time being, we exclude the formation of rings. Sites a and b can reversibly bind an agent of type A (the enzyme)

and of type B (the substrate) with affinities α and β, respectively. All binding interactions are independent. When the system is closed, the total concentrations of A, B, and S are given by t_A, t_B, and t_S. This setup allows for a variety of configurations as shown on the left of the arrow in Fig. 1.4. We posit that each enzyme A can act on each substrate B bound to the same complex. We refer to the number pq of potential interactions enabled by a configuration with sum formula $A_pS_nB_q$ as that configuration's "catalytic potential" Q. By extension we will speak of the catalytic potential Q of a mixture of configurations as the sum of their catalytic potentials weighted by their concentrations.

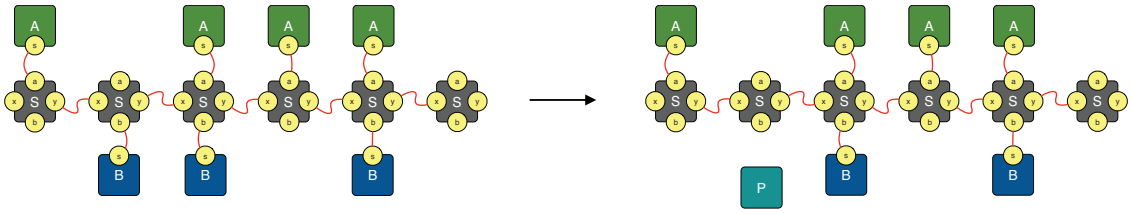


Figure 1.4: Enzyme-substrate interaction on a polymeric scaffold. In the polymerizing model, scaffold protomers S binding each other with affinity σ yield a distribution of polymers of varying length to which enzymes A and substrates B bind with affinities α and β, respectively. For each configuration, the rate of conversion to product is a function of the configuration's catalytic potential Q, which is the number of possible interactions between bound A and B agents. Here, each of the four As can interact with each of the three Bs for a total of $Q = 12$ possible interactions.

If we assume that the assembly system equilibrates rapidly, the rate of product formation is given by Qk_{cat} with k_{cat} the catalytic rate constant and Q the equilibrium abundance of potential interactions between A- and B-agents. Rapid equilibration is a less realistic assumption than a quasi-steady state but should nonetheless convey the essential behavior of the system. In the following we first provide a continuum description of equilibrium Q in terms of concentrations (which do not imply a maximum polymer length) and then a discrete statistical mechanics treatment for the average equilibrium Q (where t_S is a natural number and implies a maximum length).

In the present context, molecular species Y_i that assemble from T distinct building blocks ("atoms") X_j through reversible binding interactions have a graphical (as opposed to geometric) structure that admits two descriptors: ω_i, the number of symmetries of Y_i (here $\omega_i = 1$ because the polymers are oriented), and $\mu_{i,j}$, the number of atoms X_j in Y_i. The equilibrium concentration y_i of any Y_i is given by $y_i = \varepsilon_i \prod_{j=1}^{T} (x_j)^{\mu_{i,j}}$, where $\varepsilon_i = 1/\omega_i \prod_{r\in P} K_r$ is the exponential of the free energy content of Y_i, with $K_r \in \{\alpha, \beta, \sigma\}$ the equilibrium constant of the rth reaction along some assembly path P. The x_j are the equilibrium concentrations of free atoms of type j (here $T = 3$). Hence, $\varepsilon_i = \alpha^p \beta^q \sigma^r$ for a Y_i that contains p bonds between A and S, q bonds between B and S, and r bonds between S protomers.

Consider first the polymerization subsystem. From what we just laid out, the equilibrium concentration of a polymer of length l is $\sigma^{l-1}s^l$, where s is the equilibrium concentration of monomers of S. Summing over all polymer concentrations yields the total abundance of entities in the system, $W(s) = \sum_{l=1}^{\infty} \sigma^{l-1}s^l = s/(1-\sigma s)$. $W(s)$ gives us a conservation relation, $t_S = s\,dW(s)/ds$, from which we obtain s as:

$$s = \frac{1}{4\sigma}\left(\sqrt{4 + 1/(\sigma t_S)} - \sqrt{1/(\sigma t_S)}\right)^2. \tag{1.6}$$

Using (1.6) in $\sigma^{l-1}s^l$ yields the dependence of the polymer size distribution on parameters t_S and σ. $W(s)$ has a critical point at $s_{\mathrm{cr}} = 1/\sigma$, at which the concentrations of all length classes become identical. It is clear from (1.6) that s can never attain that critical value for finite σ and t_S.

1.3 The chemostatted case

In a chemostatted system, s can be clamped at any desired value, including the critical point $1/\sigma$ at which ever more protomers are drawn from the reservoir into the system to feed polymerization. We next include ligands A and B at clamped concentrations a and b. Let $A_pS_nB_q$ be the sum formula of a scaffold polymer of length n with p A-agents and q B-agents. There are $\binom{n}{p}\binom{n}{q}$ such configurations, each with the same catalytic potential $Q = pq$. Summing up the equilibrium abundances of all configurations yields

$$W(s,a,b) = a + b + \frac{s(1+\alpha a)(1+\beta b)}{1-\sigma s(1+\alpha a)(1+\beta b)}. \tag{1.7}$$

(1.7) corresponds to the $W(s)$ of ligand-free polymerization by a coarse-graining that only sees scaffolds regardless of their ligand-binding state, i.e. by dropping terms not containing s and substituting $s(1+\alpha a)(1+\beta b) \to s$. (1.7) indicates that, at constant chemical potential for A, B and S, the presence of ligands lowers the critical point of polymerization to $s_{\mathrm{cr}} = 1/(\sigma(1+\alpha a)(1+\beta b))$ because, in addition to polymerization, free S is also removed through binding with A and B.

Q_{poly}, the Q of the system, is obtained by summing up the Q of each configuration weighted by its equilibrium concentration (SI section 1). Using W, we compute Q_{poly} as

$$Q_{\mathrm{poly}} = ab\frac{\partial^2}{\partial a \partial b}W = \alpha a \beta b\, s\, \frac{1+\sigma\, s(1+\alpha a)(1+\beta b)}{(1-\sigma s(1+\alpha a)(1+\beta b))^3}. \tag{1.8}$$

Note that Q_{poly} inherits the critical point of W. The behavior of the chemostatted continuum model is summarized in Fig. 1.5.

Q_{poly} (red) diverges as the polymerization system approaches the critical point. The inset of Fig. 1.5A shows the scaffold length distribution at the black dot on the Q_{poly}-profile. The red dotted curve reports the length distribution in the presence of ligands, $[\{A_*S_kB_*\}] = \sigma^{-1}(\sigma s(1+\alpha a)(1+\beta b))^k$,

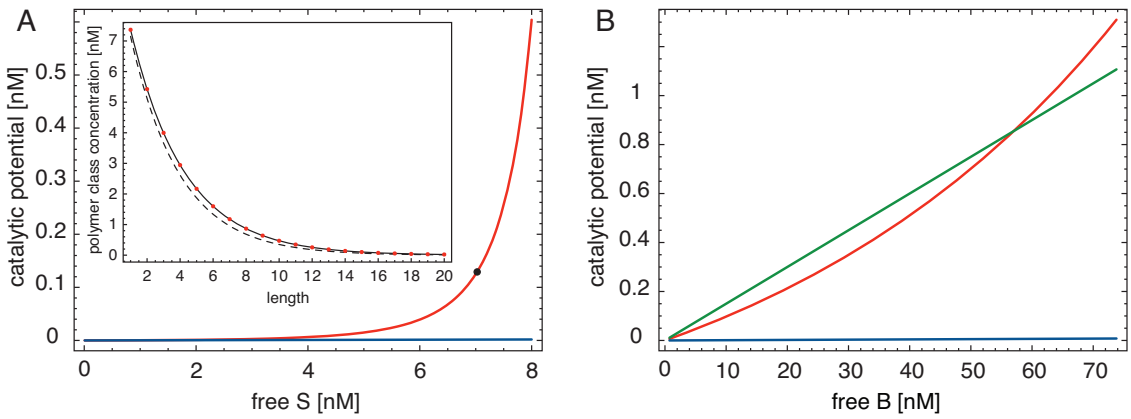


Figure 1.5: Catalysis in a chemostatted polymerizing scaffold system. A: The red graph shows the catalytic potential Q as a function of chemostatted s according to (1.8) for $\alpha = \beta = 10^6$ M^{-1}, $\sigma = 10^8$ M^{-1}, and $a = b = 15 \cdot 10^{-9}$ M (about $2\,10^4$ molecules in 10^{-12} L). The blue curve is the special case of $\sigma = 0$, which is the monovalent scaffold system, $Q = \alpha a \beta b\, s$. The inset shows the scaffold length distribution at $s = 7.15$ nM, corresponding to Q at the black filled circle. The critical point in this example is $s_{cr} \sim 9.7$ nM. Panel B: The catalytic potential at $s = 7.15$ nM as a function of clamped b (the substrate); other parameters as in A. Red: polymerizing scaffold system; blue: monovalent scaffold; green: chemostatted Michaelis-Menten in which A binds directly to B with affinity α.

whereas the black dotted curve reports the length distribution in the absence of ligands, $s_k \equiv [S_k] = \sigma^{k-1} s^k$. The presence of A and B shifts the distribution to longer chains. The blue curve in Fig. 1.5A shows the catalytic potential of the monovalent scaffold, $\sigma = 0$. It increases linearly with s, but at an insignificant slope compared with the polymerizing case, which responds by raising the size (surface) distribution, thus drawing in more S from the reservoir to maintain a given s; this, in turn, draws more A and B into the system. In Fig. 1.5B, s is fixed and b, the substrate concentration, is increased. The green straight line is the Michaelis-Menten case, which consists in the direct formation of an AB complex and whose $Q = \alpha\, a\, b$ is linear in b. The red line is the polymerizing scaffold system whose $s_{\rm cr}$ can be attained by just increasing b, (1.8). All else being equal, there is a b at which more substrate can be processed than through direct interaction with an enzyme. The slope of the monovalent scaffold (blue) is not noticeable on this scale.

1.4 The continuum case in equilibrium

We turn to the system with fixed resources t_S, t_A and t_B, expressed as real-valued concentrations. (1.8) for $Q_{\rm poly}$ is now evaluated at the equilibrium concentrations s, a and b of the free atoms. These are obtained by solving the system of conservation equations, $t_S = s\,\partial W/\partial s$, $t_A = a\,\partial W/\partial a$, $t_B = b\,\partial W/\partial b$ (solutions in SI, section 1). The orange curve in Fig. 1.6A depicts the saturation curve of the catalytic potential $Q_{\rm direct}$ of the Michaelis-Menten mechanism for a fixed concentration t_A of

enzyme as a function of substrate t_B. The green curves are saturation profiles of the polymerizing scaffold system at varying protomer abundances t_S under the same condition. As in the chemostatted case, beyond some value of t_S, the catalytic potential of the polymerizing system exceeds that from direct interaction.

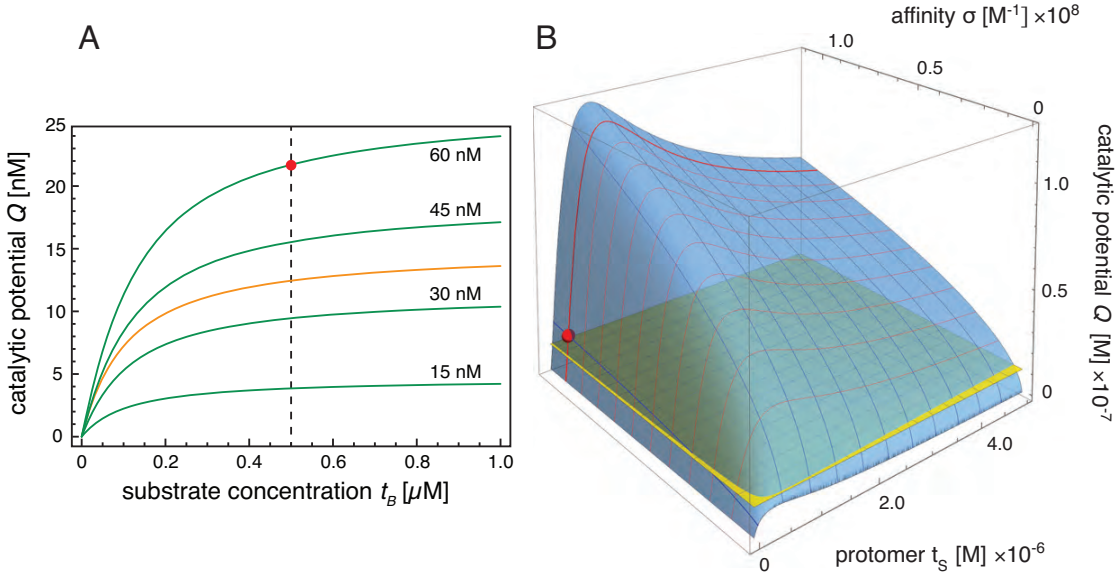


Figure 1.6: Catalysis in a closed polymerizing scaffold system. A: The orange curve shows the saturation of catalytic potential Q of the direct enzyme (A)-substrate (B) interaction, a classic Michaelis-Menten mechanism, as a function of t_B for $\beta = 10^7$ M^{-1} and $t_A = 15 \cdot 10^{-9}$ M. The green curves depict the saturation curves for Q of the poly-scaffold with affinities $\alpha = \beta = 10^7$ M^{-1} and $\sigma = 10^8$ M^{-1} at various protomer abundances t_S. B: The catalytic potential surface for the poly-scaffold as a function of t_S and σ; other parameters as in panel A. The red ball corresponds to the conditions marked by the red dot in panel A ($t_B = 5 \cdot 10^{-7}$ M). The flat yellow surface is the Q for the direct enzyme-substrate interaction (i.e. the intersection of the vertical dotted line in panel A with the orange curve). See text for discussion.

Q_{poly} can be modulated not only by the protomer concentration t_S, but also the protomer affinity σ (Fig. 1.6B). Increasing t_S improves Q_{poly} dramatically at all affinities up to a maximum after which enzyme and substrate become progressively separated due to the prozone effect. At all protomer concentrations, in particular around the maximizing one, Q_{poly} always increases with increasing affinity σ. Fig. 1.6B suggests that for the modulation through σ to be most effective the protomer concentration should be close to the maximizing t_S.

1.4.1 Comparison with multivalent scaffold systems

With regard to Q, a polymer chain of length n is equivalent to a multivalent scaffold agent $S_{(n)}$ with n binding sites for A and B each. It is therefore illuminating to compare the polymerizing system with multivalent scaffolds and their mixtures.

It is straightforward to calculate the equilibrium concentration of configurations $A_pS_{(n)}B_q$ for an n-valent scaffold by adopting a site-oriented view that exploits the independence of binding interactions. The calculation (SI section 2) yields as a general result that the catalytic potential for an arbitrary scaffolding system, assuming independent binding of A and B, consists of two factors:

$$Q = \underbrace{p(t_{\mathrm{sit}}, t_A, \alpha)p(t_{\mathrm{sit}}, t_B, \beta)}_{I}\underbrace{Q_{\max}(\vec{t}_S)}_{II}. \tag{1.9}$$

The dimensionless function $p(t_{\mathrm{sit}}, t_X, \gamma)$ denotes the equilibrium fraction of X-binding *sites*, with total concentration t_{sit}, that are occupied by ligands of type X, with total concentration t_X:

$$p(t_{\mathrm{sit}}, t_X, \gamma) = \frac{\gamma t_X - \gamma t_{\mathrm{sit}} - 1 + \sqrt{4\gamma t_X + (\gamma t_X - \gamma t_{\mathrm{sit}} - 1)^2}}{\gamma t_X - \gamma t_{\mathrm{sit}} + 1 + \sqrt{4\gamma t_X + (\gamma t_X - \gamma t_{\mathrm{sit}} - 1)^2}}.$$

This expression is the well-known dimerization equilibrium, computed at the level of sites rather than scaffolds and taken relative to t_{sit} (SI section 2).

Factor I depends on the total concentration of ligand binding sites (for each type) but not on how these sites are partitioned across the agents providing them. For example, a multivalent scaffold $S_{(n)}$, present at concentration $t_{S_{(n)}}$, provides $t_{\mathrm{sit}} = nt_{S_{(n)}}$ binding sites and the probability that a site of any particular agent is occupied is the same as the probability that a site in a pool of $nt_{S_{(n)}}$ sites is occupied. For a heterogeneous mixture of multivalent scaffold agents, we have $t_{\mathrm{sit}} = \sum_{i=1}^{n} i\, t_{S_{(i)}}$; for a polymerizing system in which each protomer S exposes one binding site, we have $t_{\mathrm{sit}} = t_S$.

Factor II is the maximal Q attainable in a scaffolding system. This factor depends on how sites are partitioned across scaffold agents with concentrations $\vec{t}_S = (t_{S_{(1)}}, \ldots, t_{S_{(n)}})$, but does not depend on ligand binding equilibria. For example, a system of multivalent agents at concentrations $\vec{t}_S$ has $Q_{\max} = \sum_{i=1}^{n} i^2 t_{S_{(i)}}$. The polymerizing scaffold system is analogous, but $n = \infty$ and the $t_{S_{(i)}}$ are determined endogenously by aggregation: $t_{S_{(i)}} = s_i = \sigma^{i-1}s^i$. This yields simple expressions for the catalytic potential of a polymerizing scaffold, Q_{poly}, and multivalent scaffold, Q_{multi}:

$$Q_{\mathrm{poly}} = p(t_S, t_A, \alpha)p(t_S, t_B, \beta)\frac{s(1+\sigma s)}{(1-\sigma s)^3} \tag{1.10}$$

$$Q_{\mathrm{multi}} = p(n\, t_{S_{(n)}}, t_A, \alpha)p(n\, t_{S_{(n)}}, t_B, \beta)n^2 t_{S_{(n)}}$$

with s in (1.10) given by (1.6). (1.10) is equivalent to (1.8). While (1.8) requires solving a system of mass conservation equations to obtain a, b, and s, Q_{poly} as given by (1.10) does not refer to a

and b, but only to s as determined by the ligand-free polymerization subsystem. The Q that shapes the Michaelis-Menten rate law under the assumption of rapid equilibration of enzyme-substrate binding has the same structure as (1.9): $Q_{\text{direct}} = p(t_A, t_B, \alpha)t_A$, where t_A and t_B are the total enzyme and substrate concentration, respectively. The presence of a second concurrent binding equilibrium in (1.9) characterizes the prozone effect.

Adding sites, all else being equal, necessarily decreases the fraction p of sites bound. Specifically, factor I tends to zero like $1/t_{\text{sit}}^2$ for large t_{sit}. In contrast, Q_{max} increases monotonically, since adding sites necessarily increases the maximal number of interaction opportunities between A and B. For a multivalent scaffold Q_{max} diverges linearly with t_{sit}. For the polymerizing system Q_{max} diverges like $t_{\text{sit}}^{3/2}$ (SI section 5).

Fig. 1.7A provides a wide-range comparison of Q_{poly} (red) with Q_{multi} for various valencies (blue) at the same site concentration $t_{\text{sit}} = t_S$.

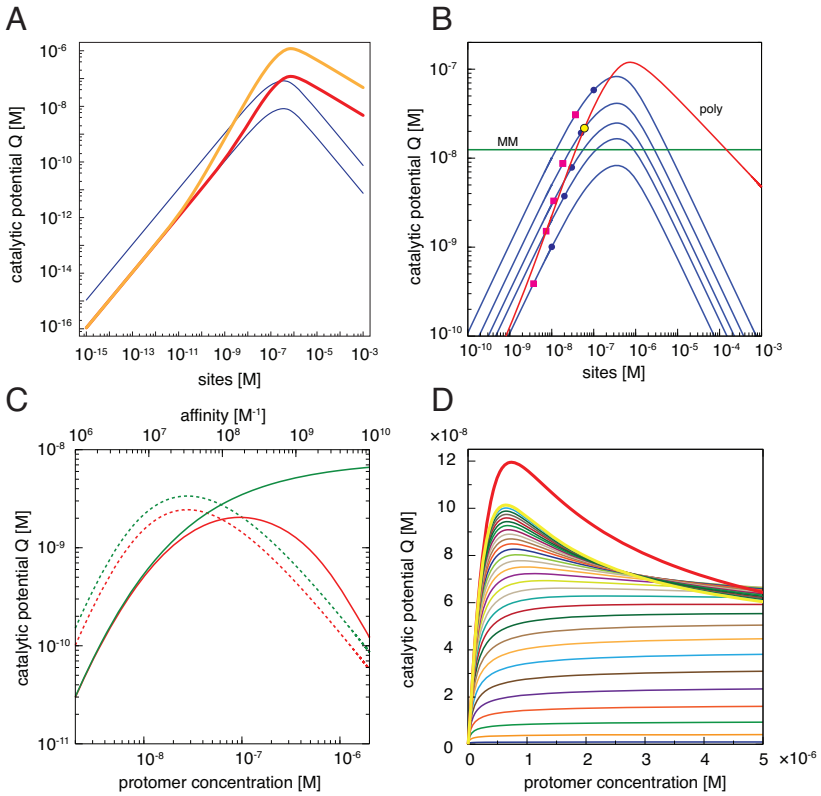


Figure 1.7: Multivalent scaffolds and polymerizing scaffold. A: Large-scale view of the catalytic potential Q as a function of site concentration t_S. The blue curves depict Q_{multi} for n-valent scaffolds (lower: $n = 1$, higher: $n = 10$). The location of the peak of Q_{multi} is independent of the valency n when expressed as a function of $t_{\text{sit}} = t_S$ (SI section 5, Eq. 38). The red and orange curves depict Q_{poly} for two affinities (red: $\sigma = 10^8$ M^{-1}, orange: $\sigma = 10^{10}$ M^{-1}). Other parameters: $\alpha = \beta = 10^7$ molecules^{-1}, $t_A = 1.5 \cdot 10^{-8}$ M, $t_B = 5 \cdot 10^{-7}$ M. On a log-log scale, the up-slope of Q_{poly} is 1 initially—the same as for multivalent scaffolds—and increases to 2 prior to reaching the prozone peak. The down-slope is $-1/2$, whereas it is -1 for multivalent scaffolds (SI section 5). B: Close-up of the peak region in panel A for the red curve; multivalent scaffolds were added for $n = 2, 3, 5$. The slight asymmetry in the Q profiles of multivalent scaffolds stems from the differences in ligand concentrations of our running example; see also SI, section 11. The yellow dot on the Q_{poly} curve corresponds to the red dot in Fig. 1.6. A pink square on a blue curve of valency n marks Q_{multi} when the scaffold concentration $t_{S(n)}$ is the same as the concentration of polymers of size n (s_n) at the t_S at which the length class n dominates the polymerizing system (SI section 3 Fig. S2B). A blue dot indicates the Q_{multi} when the scaffold concentration $t_{S(n)} = 1/\sigma$, which is the asymptotic (and maximal) value of s_n, for all n, in the limit of infinite t_S. These markers serve to show that within the most populated length classes the prozone peak is never reached. MM labels the Michaelis-Menten case of Fig. 1.6 for comparison. See text for details. C: The solid lines in the graph exemplify the absence of a prozone within an isolated length class n, here $n = 3$, and the presence of a prozone for the same class in the context of all other classes. Green solid: Q_{multi} for $n = 3$ using $t_{S_{(3)}} = s_3$ and $t_{\text{sit}} = 3\, t_{S_{(3)}}$. Red solid: Q_{multi} for $n = 3$ using $t_{S_{(3)}} = s_3$ but $t_{\text{sit}} = t_S$. The dotted lines illustrate the situation for the length class $n = 3$ as a function of affinity σ (upper abscissa, same ordinate). In this dimension, the bending of the curves is *not* due to a prozone effect, since the number of sites does not increase; see text. D: Cumulative sums from $i = 1$ to $n = 30$ of Q_{multi} with $t_{S_{(i)}} = s_i$ and $t_{\text{sit}} = \sum_{i=1}^{n} i\, t_{S_{(i)}}$.

On a log-log scale, scaffolds of arbitrary valency n exhibit a Q_{multi} whose slope as a function of t_{sit} is 1, with offset proportional to n, until close to the peak. For the polymerizing scaffold, the first order term of the series expansion of Q_{poly} is independent of the affinity σ (SI section 5), whereas the second order term is linear in σ. Hence, for small t_{sit}, the polymerizing system behaves like a monovalent scaffold and any multivalent scaffold offers a better catalytic potential. However, as t_S increases, the equilibrium shifts markedly towards polymerization, resulting in a slope of 2, which is steeper than that of any multivalent scaffold. The steepening of Q_{poly} is a consequence of longer chains siphoning off ligands from shorter ones (SI, section 4). All n-valent scaffolds reach their maximal Q_{multi} at the same abundance of sites $t_{\text{sit}} = n\, t_{S_{(n)}} = t_S$ and before Q_{poly}. The superlinear growth in Q_{max} of the polymerizing system softens the decline of Q_{poly} to an order $t_S^{-1/2}$ for large t_S. In contrast, the decline of Q_{multi} is of order t_{sit}^{-1}. In sum, the polymerizing scaffold system catches up with any multivalent scaffold, reaches peak-Q later, and declines much slower.

The mitigation of the prozone effect begs for a mechanistic explanation, since a prozone could occur not only within each length class but also between classes. To assess the within-class prozone, we think of a length class k as if it were an *isolated* k-valent scaffold population at concentration $t_{S_{(k)}} = s_k = \sigma^{k-1} s^k$ with $Q_{\text{multi}} = p(k\, s_k, t_A, \alpha) p(k\, s_k, t_B, \beta) k^2 s_k$. For all k, s_k approaches monotonically the limiting value $1/\sigma$ as $t_S \to \infty$ (SI section 2, Fig. S1A). Assuming equal affinity α for both ligands A and B, peak-Q_{multi} for a k-valent scaffold occurs at $t_{S_{(k)}}^{\text{peak}} = k^{-1}(\alpha^{-1} + (t_A + t_B)/2)$. Thus, when established through a polymerization system, $t_{S_{(k)}}$ can never exceed the concentration required for peak-Q_{multi} for any k up to $k = \sigma/\alpha$ (Fig. 1.7B, blue dots). For the α used in the red curve of Fig. 1.7B this lower bound is $k = 10$ and the actual value, given employed values of t_A and t_B, is about $k = 35$. At the yellow marker and at peak-Q_{poly} in Fig. 1.7B 98% and 68%, respectively, of all sites are organized in length classes below 10. Thus, the most populated lengths avoid the within-class prozone entirely (for example $k = 3$ as depicted in Fig. 1.7C, green solid line). Yet, the actual behavior of the kth length class occurs in the context of all other classes, i.e. at site concentration t_S, not just $k\, s_k$. In this frame, the class indeed exhibits a prozone (Fig. 1.7C, red solid line). The overall prozone of the polymerizing scaffold system is therefore mainly due to the spreading, and ensuing isolation, of ligands *between* length classes. This "entropic" prozone becomes noticeable only when including all length classes up to relatively high k because the majority of sites are concentrated at low k where they are even jointly insufficient to cause a prozone, Fig. 1.7D.

At constant t_S and in the limit $\sigma \to \infty$, s_k tends toward zero for all k (SI, Fig. S3C). In the σ-dimension, unlike in the t_S-dimension, the class s_k itself has a peak. As σ increases, the k of the class that peaks at a given σ increases. Consequently, the Q_{multi} of each length-class in isolation will show a "fake" prozone with increasing σ, due entirely to the polymerization wave passing through class k as it moves towards higher k while flattening (Fig. 1.7C, dotted lines). Since there

is no site inflation, the overall Q_{poly} increases monotonically.

Effects of ligand imbalance and unequal ligand binding affinities are discussed in the SI, section 11.

1.4.2 Interaction horizon

The assumption that every A can interact with every B attached to the same scaffold construct is unrealistic. It can, however, be tightened heuristically without leaving the current level of abstraction. We introduce an "interaction horizon," $q_{max}(l, h)$, defined as the radius h in terms of scaffold bonds within which a bound A can interact with a bound B on a polymer of size l. In this picture, an A can interact with at most $2h + 1$ substrate agents B: h to its "left," h to its "right" and the one bound to the same protomer. The interaction horizon only modulates the Q_{max} of a polymer of length l, replacing the interaction factor l^2 with (SI section 6):

$$q_{max}(l, h) = \begin{cases} l(2h + 1) - h(h + 1), & \text{for } 0 \leq h \leq l - 1 \\ l^2, & \text{for } h \geq l \end{cases} .$$

The horizon h could be a function of l. One case, in which h covers a constant *fraction* of a polymer, is treated in section 6 of the SI. In a more restrictive scenario, we assume a fixed horizon independent of length, which could reflect a constant local flexibility of a polymer chain. With the assumption of a constant h, (1.10) becomes (SI section 6)

$$Q_{\text{poly}} = p(t_S, t_A, \alpha)p(t_S, t_B, \beta)\frac{s\left(1 + \sigma s - 2(\sigma s)^{h+1}\right)}{(1 - \sigma s)^3}. \tag{1.11}$$

In (1.11), the numerator of the Q_{max} term of (1.10) is corrected by $-2s(\sigma s)^{h+1}$. Since $\sigma s < 1$ for all finite t_S and σ, even moderate values of h yield only a small correction to the base case of a limitless horizon.

1.5 The discrete case in equilibrium

Replacing concentrations with particle numbers $t_S, t_A, t_B \in \mathbb{N}$ in a specified reaction volume yields the discrete case. In this setting, we must convert deterministic equilibrium constants, such as σ to corresponding "stochastic" equilibrium constants σ_s through $\sigma_s = \sigma/(AV)$, where A is Avogadro's constant and V the reaction volume to which the system is confined. For simplicity, we overload notation and use σ for σ_s.

The basic quantity we need to calculate is the average catalytic potential $\langle Q_{\text{poly}} \rangle = \sum_{l,i,j} i\, j\, \langle n_{lij} \rangle$, where $\langle n_{lij} \rangle$ is the average number of occurrences of a polymer of length l with i and j ligands of type A and B, respectively. Conceptually, $\langle n_{lij} \rangle$ counts the occurrences of an assembly configuration $A_i S_l B_j$ in every possible state of the system weighted by that state's Boltzmann probability. In the SI (section 7), we show that $\langle n_{lij} \rangle$ is given by the number of ways of building one copy of $A_i S_l B_j$

from given resources (t_S, t_A, t_B) times the ratio of two partition functions—one based on a set of resources reduced by the amounts needed to build configuration $A_i S_l B_j$, the other based on the original resources. The posited independence of all binding processes in our model implies that the partition function is the product of the partition functions of polymerization and dimerization, which are straightforward to calculate (SI section 8). While exact, the expressions we derive for $\langle Q_{\text{poly}} \rangle$ (SI, section 8, Eq. 66) and $\langle Q_{\text{multi}} \rangle$ (SI, section 8, Eq. 69) are sums of combinatorial terms and therefore not particularly revealing. For numerical evaluation of these expressions, we change the size of the system by a factor ξ (typically $\xi = 0.01$), i.e. we multiply volume and particle numbers with ξ and affinities with $1/\xi$. Such re-sizing preserves the average behavior. Our numerical examples therefore typically deal with 10-1000 particles and stochastic affinities on the order of 10^{-2} to 10 molecules^{-1}.

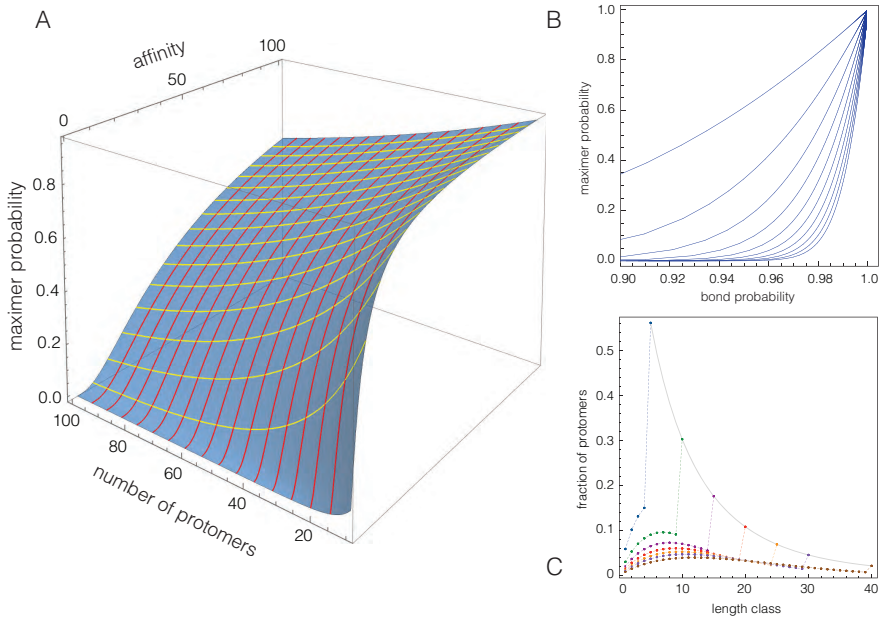


Figure 1.8: Maximer. A: The surface depicts the probability of observing the maximer as a function of t_S and σ. B: Here the maximer probability is graphed as function of the probability p that a bond exists between two protomers. p is a function of t_S and σ and can be calculated exactly. Each curve corresponds to a particular t_S with varying σ. t_S ranges from 10 (topmost curve) to 100 (bottom curve) in increments of 10, while σ ranges from 1 to 1000. C: Mass distributions in the polymerizing scaffold model. Any curve depicts the fraction of protomers in all length classes n, computed as $n\,\sigma^{n-1} t_S!/(t_S - n)!\, Z^{(\text{poly})}_{t_S-n}/Z^{(\text{poly})}_{t_S}$ with $Z^{(\text{poly})}_{t_S}$ the partition function for polymerization with t_S protomers (SI, section 8). Each curve corresponds to a given number of protomers: $t_S = 5$ (blue), 10 (green), 15 (plum), 20 (red), 25 (orange), 30 (purple), 40 (brown); affinity $\sigma = 3$ in all cases. When t_S is small, the longest possible polymer—the "maximer"—is realized with appreciable frequency and dominates the mass distribution. As t_S increases, at fixed σ, the maximal length class increases too but its dominance fades.

The key aspect of the discrete case is the existence of a largest polymer consisting of all t_S protomers. We refer to it as the "maximer"; no maximer exists in the continuum case because of the infinite fungibility of concentrations (Fig. S9). Since there is only one maximer for a given t_S, its expectation is the probability of observing it: $\langle s_{\max} \rangle = t_S!\, \sigma^{t_S-1} / Z_{t_S}^{(\text{poly})}$, where $Z_{t_S}^{(\text{poly})}$ is the partition function of polymerization (SI, sections 8 and 9). This probability is graphed as a function of t_S and σ in Fig. 1.8A. At any fixed t_S, the probability of observing the maximer will tend to 1 in the limit $\sigma \to \infty$. This puts a ceiling to $Q_{\max}$ that is absent from the continuum description. In the t_S-dimension, the maximer probability decreases as t_S increases at constant σ.

Polymerization as considered here has a natural analogy to bond percolation on a 1-dimensional lattice (SI, section 9). The probability of percolation (in which the entire lattice becomes one connected component) is parametrized by the probability p of a bond between adjacent lattice sites. In the case of polymerization we can compute the probability p that any two protomers are linked by a bond as a function of t_S and σ. For continuum but not for discrete polymerization, the analogy to percolation on an infinite 1D lattice is actually an exact correspondence (SI, section 9). For the present purpose, the percolation perspective is useful in that it combines the two main model parameters t_S and σ in the single quantity p (Fig. 1.8B). As in finite-size percolation, the salient observation is that for small t_S the maximer has a significant probability of already occurring at modest affinities; for example, given 10 protomers and discrete binding affinity 1, p is already 0.78 and the maximer probability a respectable 0.06. For larger t_S, the maximer loses significance unless the affinity is scaled up correspondingly (SI section 10). This is also reflected in the mass distribution, Fig. 1.8C.

Fig. 1.9A compares the discrete polymerizing scaffold system with discrete multivalent scaffolds, much like Fig. 1.7A for the continuum case. The behavior of the discrete case is essentially similar to that of the continuum case—with a few nuances that are prominent at low particle numbers and high affinities, such as the topmost orange curve. Its $\langle Q_{\text{poly}} \rangle$-profile does not hug the monovalent profile (bottom green chevron curve) to then increase its slope into the prozone peak as in the continuum case (Fig. 1.7A). A behavior like in the continuum case is observed for the lower orange and red curves, for which σ is much weaker. In the continuum case, the affinity does not affect slope—the slope always shifts from 1 to 2 within some region of protomer abundance; rather, the affinity determines where that shift occurs (Fig. 1.7A). The higher the affinity, the earlier the shift. The topmost orange curve could be seen as realizing an extreme version of the continuum

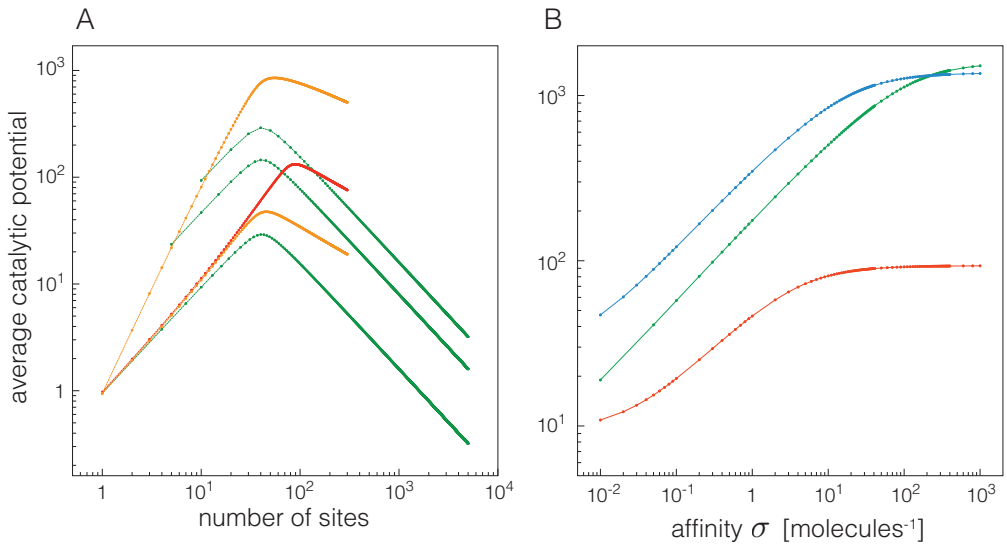


Figure 1.9: Multivalent and polymerizing scaffolds in the discrete case. A: Comparison of polymerizing scaffold (orange and red) with multivalent systems of various valencies (green). Orange: $t_A = t_B = 40$, $\alpha = \beta = 0.9$, $\sigma = 10$ (upper) $\sigma = 0.01$ (lower). All affinities in units of molecules^{-1}. Red: $t_A = t_B = 80$, $\alpha = \beta = 0.9$, $\sigma = 0.01$. Green: $t_A = t_B = 40$, $\alpha = \beta = 0.9$, valency $n = 10$ (top), $n = 5$ (middle), $n = 1$ (bottom). B: $\langle Q_{\text{poly}} \rangle$ as a function of affinity σ. $t_A = t_B = 40$, $\alpha = \beta = 0.9$, $t_S = 300$ (green), $t_S = 10$ (red), $t_S = 50$ (blue).

behavior in which an exceptionally high affinity causes a shift to slope 2 at unphysically low protomer concentrations. That such a scenario can be easily realized in the discrete case is due to the significant probability with which the maximer occurs at low particle numbers, similar to finite-size percolation. It bears emphasis that, as the number t_S of protomers increases, the maximer probability decreases (Fig. 1.8C), since the length of the maximer is t_S. Yet, once the maximer has receded in dominance, the increased number of length classes below it have gained occupancy and control the catalytic potential much like in the continuum case. Likewise, affinity does not appear to affect the slope of the downward leg as t_S increases.

The discrete multivalent scaffold system behaves much like its continuum counterpart.

In the affinity dimension, Fig. 1.9B, the discrete system shows a behavior similar to the continuum case with the qualification that $\langle Q_{\text{poly}} \rangle$ must level off to a constant, rather than increasing indefinitely. This is because, at constant t_S, an ever increasing affinity will eventually drive the system into its maximer ceiling. Because of the volume-dependence of stochastic equilibrium constants, such an increase in affinity at constant protomer number can be achieved by any physical reduction of the effective reaction volume, for example by confinement to a vesicle or localization to a membrane raft.

We determined standard deviations using stochastic simulations of the cases presented in Fig. 1.9A

(SI, section 12). For a given $\langle Q \rangle$, the standard deviation is larger after the prozone peak than before. Upon adding ligand binding sites, the ratio of standard deviation to mean (noise) increases much slower for the polymerizing system than for multivalent scaffolds.

1.6 Main conclusions

Our theoretical analysis of a polymerizing scaffold system shows that, at constant chemical potential, the system can be driven into criticality not only by increasing protomer concentration or affinity, but by just increasing ligand concentrations.

In equilibrium, the system stands out in how the prozone effect plays out. Compared with multivalent scaffolds, the polymerizing system boosts catalytic potential on the upward leg beyond a certain protomer concentration; delays the prozone peak; and dramatically mitigates the collapse on the downward leg. We explain this behavior by how the polymer length distribution adjusts to changes in protomer concentration and affinity. The discrete case behaves likewise, but, at small protomer numbers, the existence of a maximal polymer manifests itself in behavior only attainable at extreme parameter values in the continuum case.

A polymerizing scaffold could be viewed as a programmable surface whose extent can be regulated by varying parameters such as protomer concentration, polymerization affinity and, in a discrete setting, reaction volume. The system effectively concentrates interacting ligands, much like a vesicle would, but through a simpler mechanism. Given the pervasive potential for scaffold polymerization through DIX domains and the like, we suspect that many systems of this kind will be discovered.

Our model is a stylized vignette amenable to analytic treatment and exploitable for insight. Adding a bond distance constraint to the interaction among ligands did not alter the fundamental picture. Taking into account conformational aspects of polymeric chains would be a useful step, as would generalizations in which scaffolding units of distinct types form multiply interconnected aggregates facilitating diverse ligand interactions. We would expect variations in the concentration of scaffold units to have wide ranging effects on the equilibrium mixture of assemblies and the overall catalytic potential.

Acknowledgements. We gratefully acknowledge discussions with Tom Kolokotrones, Eric Deeds, and Daniel Merkle.

1.7 Supplementary information

1.7.1 W and Q in the polymerizing scaffold model

In this section, we step through the treatment of the polymerizing scaffold model with more granularity.

A polymerizing scaffold protomer S has 1 binding site for each ligand A and B. Let $\{A_pS_nB_q\}$ be the set of complexes (configurations) consisting of a scaffold polymer with n protomers, p agents of type A and q agents of type B; let $[\{A_pS_nB_q\}]$ denote their aggregate equilibrium concentration. The equilibrium concentration of any particular representative $A_pS_nB_q$ of that class is given by

$$[A_pS_nB_q] = \sigma^{n-1}\alpha^p\beta^q s^n a^p b^q = \sigma^{n-1}s^n(\alpha a)^p(\beta b)^q, \tag{1.12}$$

where a, b, s are the equilibrium concentrations of *free* A, B, and S, respectively; α denotes the equilibrium constant of A binding to S and, similarly, β and σ are the equilibrium constants for B binding to S and for S binding to S, respectively. All binding interactions are posited to be mechanistically independent of one another.

In an equilibrium treatment, a system of reactions only serves to define a set of reachable complexes and could be replaced with any other mechanism, no matter how unrealistic, as long as it produces the same set of reachable configurations. Hence we could posit that a polymer of length n is generated by a reversible "reaction" in which all constituent protomers come together at once. The equilibrium constant of such an imaginary reaction must be the exponential of the energy content of a polymer of length n, which in our case is simply $(n-1)$ times the energy content of a single bond, i.e. $\ln\sigma$. Thus, the equilibrium constant of the fictitious one-step assembly reaction is σ^{n-1} and (1.12) follows.

To aggregate the equilibrium concentrations of all molecular configurations in the class $\{A_pS_nB_q\}$ we note that the set $\{A_pS_nB_q\}$ includes $\binom{n}{p}\binom{n}{q}$ configurations with the same energy content $\sigma^{n-1}\alpha^p\beta^q$. Summing over all p and q, yields the contribution of the polymer length class n, $\{A_*S_nB_*\}$

$$[\{A_*S_nB_*\}] = \sigma^{n-1}s^n\left[\sum_{p=1}^{n}\binom{n}{p}\alpha^p\,a^p\right]\left[\sum_{q=1}^{n}\binom{n}{q}\beta^q\,b^q\right] = \sigma^{n-1}s^n(1+\alpha\,a)^n(1+\beta\,b)^n$$

$$= \frac{1}{\sigma}\left(\sigma\,s\,(1+\alpha\,a)(1+\beta\,b)\right)^n. \tag{1.13}$$

Summing over all equilibrium concentrations defines a function W:

$$W = a + b + \frac{1}{\sigma}\sum_{n=1}^{\infty}\left(\sigma\,s\,(1+\alpha a)\,(1+\beta b)\right)^n = a + b + s(1+\alpha a)(1+\beta b)\sum_{n=0}^{\infty}\left(\sigma s(1+\alpha a)(1+\beta b)\right)^n. \tag{1.14}$$

When viewing a, b, and s as formal variables, W acts as a generating function of energy-weighted configurational counts. By differentiating W with respect to s, each s-containing term gets multiplied with the exponent of s, which is the S-content of the respective configuration. Multiplying by s then restores the exponent and recovers the equilibrium concentration of the respective configuration. Summing over all configurations so treated yields the total amount of S protomers in the system and thus a conservation relation. This holds for all formal variables representing the "atoms," or building blocks, of the system:

$$t_A = a\frac{\partial W(a,b,s)}{\partial a}, \quad t_B = b\frac{\partial W(a,b,s)}{\partial b} \quad t_S = s\frac{\partial W(a,b,s)}{\partial s}. \tag{1.15}$$

By solving the equations (1.15), we obtain the equilibrium concentrations of free A, B, and S needed to compute the equilibrium concentration of any configuration:

$$a = \frac{\alpha t_A - \alpha t_S - 1 + \sqrt{(\alpha t_A + \alpha t_S + 1)^2 - 4\alpha t_A \alpha t_S}}{2\alpha} \tag{1.16}$$

$$b = \frac{\beta t_B - \beta t_S - 1 + \sqrt{(\beta t_B + \beta t_S + 1)^2 - 4\beta t_B \beta t_S}}{2\beta} \tag{1.17}$$

$$s = \frac{2}{\sigma^2 t_S} \frac{2\sigma t_S + 1 - \sqrt{4\sigma t_S + 1}}{\left(\alpha t_A - \alpha t_S + 1 + \sqrt{(\alpha t_A + \alpha t_S + 1)^2 - 4\alpha t_A \alpha t_S}\right)\left(\beta t_B - \beta t_S + 1 + \sqrt{(\beta t_B + \beta t_S + 1)^2 - 4\beta t_B \beta t_S}\right)}. \tag{1.18}$$

Carrying out the geometric sum in (1.14) yields equation (2) in the main text:

$$W(a,b,s) = a + b + \frac{s(1+\alpha a)(1+\beta b)}{1 - \sigma s(1+\alpha a)(1+\beta b)}. \tag{1.19}$$

The same manipulation of W used to obtain (1.15) can be carried out twice, once for a and once for b, to yield the catalytic potential of the system:

$$Q = a\, b\, \frac{\partial^2}{\partial a \partial b} W(a,b,s), \tag{1.20}$$

given as equation (3) in the main text.

By setting $a = b = 0$, we recover the standalone polymerization system with

$$W(s) = \frac{s}{1 - \sigma s} \tag{1.21}$$

and s obtained from solving $t_S = dW(s)/ds$:

$$s = \frac{1}{4\sigma}\left(\sqrt{4 + \frac{1}{\sigma t_S}} - \sqrt{\frac{1}{\sigma t_S}}\right)^2, \tag{1.22}$$

as in equation (1) of the main text. We discuss the main properties of the standalone polymerization system in section 1.7.3 of this Appendix. In an equilibrium setting, the critical point of the model with ligands A and B should be the same as that of the polymerization system without ligands, namely $t_S \to \infty$ or $\sigma \to \infty$. This is not obvious from W (whose critical point Q inherits) as given in (1.19) with solutions (1.16)-(1.18). However, it is made explicit in an alternative, more insightful derivation of the equilibrium catalytic potential Q given in section 1.7.2 of this Appendix.

1.7.2 Derivation of the general expression for the catalytic potential

In this section, we derive expression (4) of the main text.

We consider a *multivalent* scaffold agent S with n_A binding sites for A and n_B binding sites for B. Our goal is to calculate the catalytic potential Q_{multi} of a system consisting of A-agents at concentration t_A, B-agents at concentration t_B, and S-agents at concentration t_S.

The function $W(a, b, s)$, introduced in the main text for the polymerizing scaffold system, sums up the equilibrium concentrations of all possible entities in the system. The same concept applies to a multivalent scaffold:

$$W_{\text{multi}}(a, b, s) = a + b + s(1 + \alpha a)^{n_A}(1 + \beta b)^{n_B} \tag{1.23}$$

with a, b, and s the equilibrium concentrations of the free A, B, and S, respectively. The catalytic potential Q_{multi} of the multivalent scaffold system is

$$Q_{\text{multi}} = a\, b\, \frac{\partial^2}{\partial a \partial b} W_{\text{multi}}(a, b, s) = s\, \alpha\, \beta\, a\, b\, n_A\, n_B\, (1 + \alpha a)^{n_A - 1}(1 + \beta b)^{n_B - 1}. \tag{1.24}$$

The equilibrium concentrations a, b, and s are determined by the system of conservation equations

$$a\frac{\partial}{\partial a} W = t_A, \quad b\frac{\partial}{\partial b} W = t_B, \quad s\frac{\partial}{\partial s} W = t_S. \tag{1.25}$$

However, we can bypass solving these equations by calculating the concentrations directly, which serendipitously gives us an intelligible expression for the catalytic potential Q in general.

We first calculate the equilibrium concentration of the fully occupied scaffold configuration, $[A_{n_A} S B_{n_B}]$ by reasoning at the level of binding *sites*. The concentration of *sites available* for binding to S are denoted by a, which is also the concentration of free A-agents. Since each A-binding site on S is independent, the equilibrium fraction of S-agents that are fully occupied with A-agents is simply

$$\frac{[\{A_{n_A} S\}]}{t_S} = \left(\frac{\alpha a}{1 + \alpha a}\right)^{n_A}. \tag{1.26}$$

The expression in parentheses is the single-site binding equilibrium. Likewise, let $[s]$ be the concentration of free A-binding sites on S-agents and $[as]$ the concentration of bonds between A- and S-agents. In equilibrium, we have that

$$\alpha a\,[s] = [as], \quad n_A t_S = [s] + [as], \quad t_A = a + [as]. \tag{1.27}$$

Hence, $a = [as]/(\alpha[s])$ or $a = (t_A - a)/(\alpha[s]) = (t_A - a)/(\alpha(n_A t_S - t_A + a))$, which yields a quadratic in a whose solution is

$$a = \frac{1}{2\alpha}\left(\alpha t_A - n_A \alpha t_S - 1 + \sqrt{(\alpha t_A - n_A \alpha t_S - 1)^2 + 4\alpha t_A}\right). \tag{1.28}$$

We plug (1.28) into (1.26) to obtain

$$\frac{[\{A_{n_A} S\}]}{t_S} = \left(\frac{\alpha t_A - n_A \alpha t_S - 1 + \sqrt{(\alpha t_A - n_A \alpha t_S - 1)^2 + 4\alpha t_A}}{\alpha t_A - n_A \alpha t_S + 1 + \sqrt{(\alpha t_A - n_A \alpha t_S - 1)^2 + 4\alpha t_A}}\right)^{n_A}. \tag{1.29}$$

The same reasoning holds for the (independent) binding of B to S:

$$\frac{[\{SB_{n_B}\}]}{t_S} = \left(\frac{\beta t_B - n_B \beta t_S - 1 + \sqrt{(\beta t_B - n_B \beta t_S - 1)^2 + 4\beta t_B}}{\beta t_B - n_B \beta t_S + 1 + \sqrt{(\beta t_B - n_B \beta t_S - 1)^2 + 4\beta t_B}}\right)^{n_B}. \tag{1.30}$$

At this point, it is useful to abbreviate

$$\begin{aligned} a_\pm \equiv a_\pm(t_A, t_S, \alpha, n_A) &= \alpha t_A - n_A \alpha t_S \pm 1 + \sqrt{(\alpha t_A - n_A \alpha t_S - 1)^2 + 4\alpha t_A} \\ b_\pm \equiv b_\pm(t_B, t_S, \beta, n_B) &= \beta t_B - n_B \beta t_S \pm 1 + \sqrt{(\beta t_B - n_B \beta t_S - 1)^2 + 4\beta t_B} \end{aligned}. \tag{1.31}$$

Note that these abbreviations are dimensionless functions of the parameters t_A, t_S, α, and $n_{A/B}$. Because A and B bind independently, we can combine (1.29) and (1.30) to obtain:

$$[A_{n_A} S B_{n_B}] = t_S \frac{a_-^{n_A}}{a_+^{n_A}} \frac{b_-^{n_B}}{b_+^{n_B}} = (\alpha a)^{n_A} (\beta b)^{n_B} s, \tag{1.32}$$

where the last equation is the equilibrium concentration in terms of free A, free B, and free S, as mentioned in the Introduction of the main text (and section 1.7.1 of this Appendix). The expression a for free A is given by (1.28), or $a = a_-/(2\alpha)$. The expression b for free B is analogous, $b = b_-/(2\beta)$. Equation (1.32) now yields s:

$$s = t_S \frac{1}{(\alpha a)^{n_A} (\beta b)^{n_B}} \frac{a_-^{n_A}}{a_+^{n_A}} \frac{b_-^{n_B}}{b_+^{n_B}} = t_S \frac{2^{n_A} 2^{n_B}}{a_+^{n_A} b_+^{n_B}}. \tag{1.33}$$

To summarize, using abbreviations (1.31):

$$a = \frac{a_-}{2\alpha}, \quad b = \frac{b_-}{2\beta}, \quad s = t_S \left(\frac{2}{a_+}\right)^{n_A} \left(\frac{2}{b_+}\right)^{n_B}. \tag{1.34}$$

Keep in mind that $a_{+/-}$ and $b_{+/-}$ are not constants, but functions of the system parameters. We now insert (1.34) into (1.24) to obtain

$$\begin{aligned} Q_{\text{multi}} &= n_A\, n_B\, s \left(\frac{\alpha a}{1+\alpha a}\right)\left(\frac{\beta b}{1+\beta b}\right)(1+\alpha a)^{n_A}(1+\beta b)^{n_B} \\ &= n_A\, n_B\, t_S \left(\frac{2}{a_+}\right)^{n_A}\left(\frac{2}{b_+}\right)^{n_B}\left(\frac{\alpha a}{1+\alpha a}\right)\left(\frac{\beta b}{1+\beta b}\right)(1+\alpha a)^{n_A}(1+\beta b)^{n_B} \\ &= n_A\, n_B\, t_S \left(\frac{\alpha a}{1+\alpha a}\right)\left(\frac{\beta b}{1+\beta b}\right)\left(\frac{2+2\alpha a}{a_+}\right)^{n_A}\left(\frac{2+2\beta b}{b_+}\right)^{n_B} \\ &= n_A\, n_B\, t_S \left(\frac{\alpha a}{1+\alpha a}\right)\left(\frac{\beta b}{1+\beta b}\right) \\ &= n_A\, n_B\, t_S \frac{a_-}{a_+}\frac{b_-}{b_+}. \end{aligned} \tag{1.35}$$

The cancellations are due to $2\alpha a = a_-$ (from (1.34)) and $a_+ = a_- + 2$ (from (1.31)).

Return to equation (1.29) and set $n_A = 1$. This gives the fraction of A-binding sites (of monovalent scaffold agents) that are occupied, that is, the probability that an A is bound:

$$p(t_S, t_A, \alpha) = \frac{a_-(t_A, t_S, \alpha, 1)}{a_+(t_A, t_S, \alpha, 1)} = \frac{\alpha t_A - \alpha t_S - 1 + \sqrt{(\alpha t_A - \alpha t_S - 1)^2 + 4\alpha t_A}}{\alpha t_A - \alpha t_S + 1 + \sqrt{(\alpha t_A - \alpha t_S - 1)^2 + 4\alpha t_A}}. \tag{1.36}$$

In the site-oriented view, it does not matter whether an A-binding site belongs to a monovalent scaffold agent or to an n-valent scaffold agent. At the same agent concentration t_S, the n-valent agent simply provides n times more sites. Thus, the probability that an A is bound if the scaffolds are n-valent is

$$p(nt_S, t_A, \alpha) = \frac{a_-(t_A, t_S, \alpha, n)}{a_+(t_A, t_S, \alpha, n)} = \frac{a_-(t_A, nt_S, \alpha, 1)}{a_+(t_A, nt_S, \alpha, 1)}, \tag{1.37}$$

since the number of binding sites only scales t_S in (1.31). With these observations, we can rephrase (1.35) as the product of two terms:

$$Q_{\text{multi}} = \underbrace{p(n_A t_S, t_A, \alpha)p(n_B t_S, t_B, \beta)}_{I}\underbrace{n_A\, n_B\, t_S}_{II}. \tag{1.38}$$

Term (I) is the probability that a site of *some* S is occupied by A and a site of *some* S is occupied by B. Term (II) counts the maximal number of possible interactions between A and B agents in the system.

Let $S_{(i)}$ denote an agent of valency i for both ligands and let $t_{S_{(i)}}$ denote its concentration. In a mixture of multivalent scaffold types of distinct valencies $i = 1, \ldots, n$ present at concentrations

$t_{S_{(i)}}$, the catalytic potentials of each type add up to that of the mixture, Q_{mix}:

$$Q_{\text{mix}} = p\left(\textstyle\sum_{i=1}^{n} i\, t_{S_{(i)}}, t_A, \alpha\right) p\left(\textstyle\sum_{i=1}^{n} i\, t_{S_{(i)}}, t_B, \beta\right) \displaystyle\sum_{i=1}^{n} i^2 t_{S_{(i)}}. \tag{1.39}$$

Generally, we can write Q_{mix} as

$$Q_{\text{mix}} = p(t_{\text{sit}}, t_A, \alpha) p(t_{\text{sit}}, t_B, \beta)\, Q_{\text{max}}(\vec{t}_S). \tag{1.40}$$

In (1.40), t_{sit} is the total concentration of binding sites, regardless of how they are partitioned across scaffold agents, $\vec{t}_S = (t_{S(i)}, \ldots, t_{S(n)})$ is a partition of sites across scaffold molecules of different valencies, and Q_{max} is the maximal attainable number of enzyme-substrate interactions in the system, which depends on the concentration of scaffolds and their valency.

If the mixture results from a polymerization process between monovalent scaffolds $S \equiv S_{(1)}$, we identify a polymer of length l with an l-valent scaffold agent (Figure 1.10).

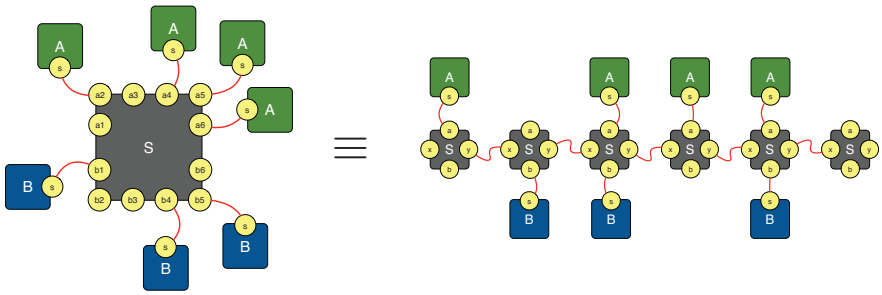


Figure 1.10: A multivalent scaffold agent can be thought as representing a particular scaffold polymer configuration.

The concentrations $t_{S_{(l)}}$ are endogenously determined by polymerization at equilibrium:

$$t_{S_{(l)}} = \sigma^{l-1} s^l,$$

where the expression for s is given by the expression for the equilibrium concentration of free monomer in the polymerization system *absent ligands*, expression (1.22) in section 1.7.1 (equation (1) in the main text). Using these $t_{S_{(l)}}$ in the sum (1.39), which in the continuum case runs to $n = \infty$, yields the expression (5) for Q_{poly} in the main text:

$$Q_{\text{poly}} = p(t_S, t_A, \alpha) p(t_S, t_B, \beta) \sum_{n=1}^{\infty} n^2 \sigma^{n-1} s^n = p(t_S, t_A, \alpha) p(t_S, t_B, \beta) \frac{s(1+\sigma s)}{(1-\sigma s)^3}, \tag{1.41}$$

with $p(\cdots)$ given by (1.36).

1.7.3 Overview of the polymerization system

In this section, we summarize some combinatorial properties of the polymerization subsystem. Understanding the concentration profile of the polymer length distribution is useful for rationalizing the overall behavior with respect to catalytic potential, because we can view the polymerizing scaffold system as a mixture of multivalent scaffolds whose concentration is set by polymerization. Since this is the simplest conceivable polymerization system, it would surprise us if anything being said here isn't already known in some form or another. Some of the features described can be found in Flory (Flory, 1936).

Let S_n be a polymer of length n and let s_n denote the equilibrium concentration of polymers in length class n. To conform with our previous notation, we shall refer to the equilibrium concentration of the monomer as $s \equiv s_1$ and to the monomer species as $S \equiv S_1$. As stated repeatedly,

$$s_n = \sigma^{n-1} s^n \quad \text{with} \quad s = \frac{1}{4\sigma}\left(\sqrt{4 + \frac{1}{\sigma t_S}} - \sqrt{\frac{1}{\sigma t_S}}\right)^2. \tag{1.42}$$

Figure 1.11 shows the dependency of s_n on the total protomer concentration t_S (panels A and B) and the affinity σ (panels C and D). Obviously, s_n is a geometric progression, thus linear in a lin-log plot for all parameter values (insets of panel A and C).

In the t_S dimension, s_n approaches $1/\sigma$ from below for each n and there is no value of t_S that maximizes s_n. In the σ dimension, s_n approaches 0 like $1/\sigma$ (in the lin-log plot, inset of panel C, the straight lines become less tilted and sink toward 0); see also expansions (1.47) and (1.48) below. However, for any given length class n, there is a σ that maximizes the concentration of that class:

$$\sigma = \frac{n^2 - 1}{4t_S}. \tag{1.43}$$

At that σ, the respective s_n is the most frequent, i.e. the most dominant, length class. It does not mean that s_n is at its most frequent, for s_n rises to $1/\sigma$ as $t_S \to \infty$. In the continuum description, the most frequent polymer class is always the monomer, for any t_S or σ. This is much more pronounced in the t_S dimension than the σ dimension.

Panels B and D of Figure 1.11 show the "mass" distribution, ns_n, i.e. the concentration of protomers in each length class. For all values of t_S and σ, the mass exhibits a maximum at some class length.

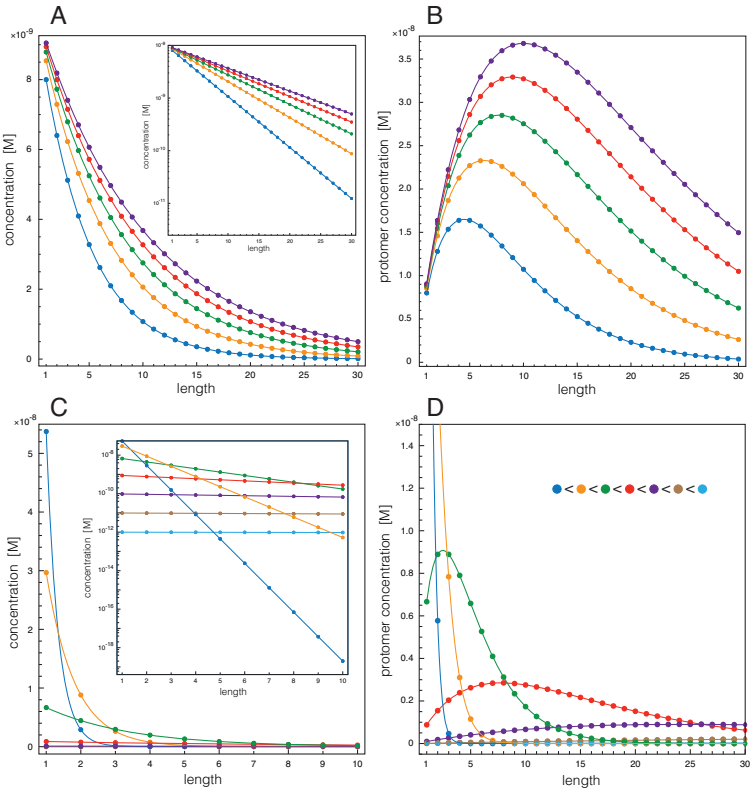


Figure 1.11: The dependence of the length distribution on the protomer concentration t_S and the affinity σ. A: The curves depict the length distribution s_i of the linear polymerization subsystem with varying t_S at $\sigma = 10^8$ M^{-1}. Blue: $t_S = 2 \cdot 10^{-7}$ M, orange: $t_S = 4 \cdot 10^{-7}$ M, green: $t_S = 6 \cdot 10^{-7}$ M, red: $t_S = 8 \cdot 10^{-7}$ M, purple: $t_S = 1 \cdot 10^{-6}$ M. The inset plots the same curves in lin-log. B: The curves depict the concentrations of protomers in each length class, that is, the "mass" distribution $i\, s_i$ under the same conditions as in panel A. C: The curves depict the length distribution s_i with varying polymerization affinity σ at $t_S = 6 \cdot 10^{-8}$ M. Blue: $\sigma = 10^6$ M^{-1}, orange: $\sigma = 10^7$ M^{-1}, green: $\sigma = 10^8$ M^{-1}, red: $\sigma = 10^9$ M^{-1}, purple: $\sigma = 10^{10}$ M^{-1}, brown: $\sigma = 10^{11}$ M^{-1}, light blue: $\sigma = 10^{12}$ M^{-1}. D: As in panel B, but with varying affinity σ (as in panel C) at $t_S = 6 \cdot 10^{-8}$. For all panels $\alpha = \beta = 10^7$ M^{-1}, $t_A = 15 \cdot 10^{-9}$ M and $t_B = 5 \cdot 10^{-7}$ M.

This maximum wanders towards ever larger n with increasing t_S and σ, while its value steadily increases with t_S, whereas it decreases with increasing σ. The length class n whose mass is maximized at a given t_S and σ is

$$n_{\max} = \left\lceil \log\left(\frac{4 t_S \sigma}{\left(\sqrt{1 + 4 t_S \sigma} - 1\right)^2} \right) \right\rceil^{-1}, \tag{1.44}$$

and, for given σ and n, the t_S at which the class n becomes the most massive of all classes is given

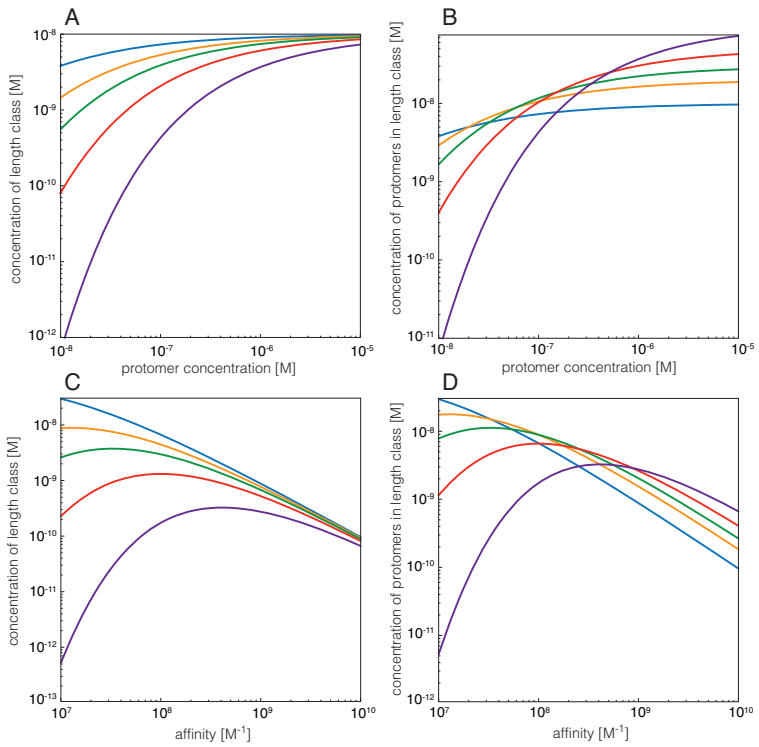


Figure 1.12: Concentrations within length classes. These panels are complementary to those in Figure 1.11. Each curve tracks the concentration of a particular length class n as protomer concentration t_S and affinity σ are varied, effectively following the changes along a vertical cut across the curves in Figure 1.11. Blue: $n = 1$, orange: $n = 2$, green: $n = 3$, red: $n = 5$, purple: $n = 10$. All other parameters as in Figure 1.11. A: Concentration s_n of length class n with varying t_S. B: Concentration ns_n of the mass in length class n with varying t_S. Panel C: Concentration s_n of length class n with varying σ. Panel D: Concentration ns_n of the mass in length class n with varying σ.

by

$$t_S = \frac{\exp(1/n)}{\sigma(1 - 2\exp(1/n) + \exp(2/n))}. \tag{1.45}$$

The pink squares on the blue multivalent scaffold curves in Figure 4B of the main text correspond to the catalytic potential Q that obtains at this concentration of sites. The same expression obtains for σ by swapping t_S and σ. At the t_S at which the mass in class n peaks, the concentration of the class is

$$s_{n_{\max}} = \frac{1}{e\sigma}, \tag{1.46}$$

independent of $n_{\max}$. Equation (1.44) assumes a continuous n; thus, to account for the discrete nature of polymer length, the actual $n_{\max}$ should be the nearest integer to the $n_{\max}$ given in (1.44). Accordingly, the actual value of $s_{n_{\max}}$ in expression (1.46) will wobble slightly.

Switching perspective from the length distribution to the behavior within a length class yields Figure 1.12. The expansion of s_n shows how each length class approaches its limit as $t_S \to \infty$ or $\sigma \to \infty$ (multiply by n for the mass distribution):

$$\text{As } t_S \to \infty,\ s_n \to \frac{1}{\sigma} \text{ with } \frac{1}{\sigma} - \frac{n}{\sigma^{3/2}} \frac{1}{t_S^{1/2}} + O\left(\frac{1}{t_S}\right) \tag{1.47}$$

$$\text{As } \sigma \to \infty,\ s_n \to 0 \text{ with } \frac{1}{\sigma} - \frac{n}{t_S^{1/2}} \frac{1}{\sigma^{3/2}} + O\left(\frac{1}{\sigma^2}\right). \tag{1.48}$$

1.7.4 Mixtures of multivalent scaffolds

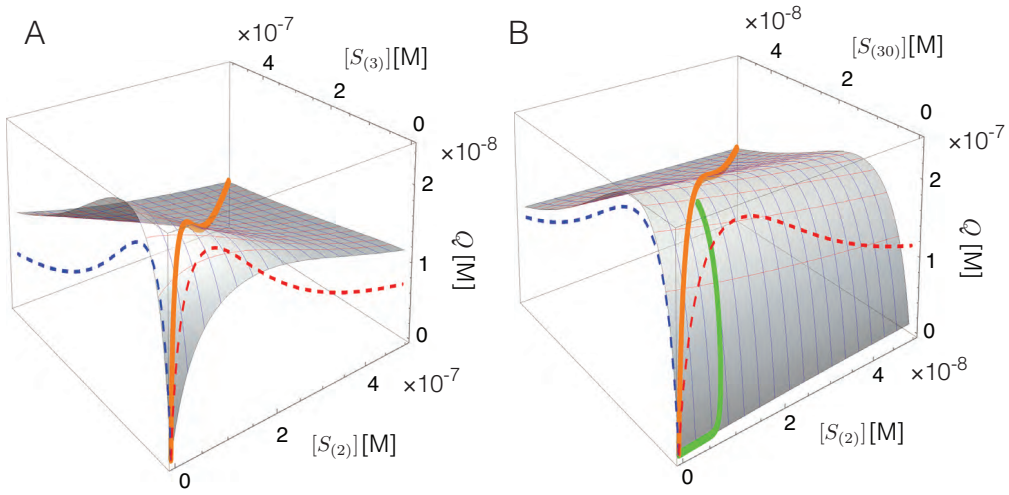


Figure 1.13: Mixtures of multivalent scaffolds. A: The graphics renders the Q_{multi}-surface of a mixture of a bivalent and trivalent scaffold. The orange line is the Q-profile when both agents are added in equal amounts to the mix. The dotted lines are projections of the orange line for comparison with the homogeneous scaffold systems. B: Same as in panel A but for a mixture of $S_{(2)}$ and $S_{(30)}$; only the portion of the surface at low scaffold concentrations is shown. The green curve shows the Q-trajectory for the binary mixture that would obtain when $[S_{(2)}]$ and $[S_{(30)}]$ are set by the polymerizing scaffold system with increasing t_S. The green curve is the *whole* trajectory, because both $[S_{(2)}]$ and $[S_{(30)}]$ converge to $1/\sigma = 10^{-8}$ M (Figure 1.12). Other parameters: $\alpha = \beta = 10^7$ M^{-1}, $t_A = 15 \cdot 10^{-9}$ M, $t_B = 5 \cdot 10^{-7}$ M.

Figure 1.13A shows the Q_{mix}-surface (1.39) of a bivalent and trivalent scaffold mixture. The main observation is the asymmetry in the effect on Q upon adding $S_{(3)}$ to a fixed amount of $S_{(2)}$ compared to the other way around—blue versus red mesh lines in Figure 1.13. Upon adding $S_{(3)}$, the ligands A and B re-equilibrate over the available binding sites. Over a range of $[S_{(2)}]$, this equilibration is more likely to result in A and B agents ending up on the same $S_{(3)}$ scaffold than on the same

$S_{(2)}$ scaffold. This is most pronounced at small $[S_{(2)}]$ and disappears gradually as the addition of binding sites drives the system past the prozone peak due to the p^2 term in (1.39). The orange curve shows the Q-profile of a mixture in which $S_{(3)}$ and $S_{(2)}$ are increased in equal amounts. The dotted curves are the projections of the mixture curve on each component axis for the purpose of comparison with the Q-curves of each component in isolation. This behavior is more dramatic in binary mixtures of multivalent scaffolds with large valency differences (Figure 1.13B).

In a polymerizing scaffold system, the concentrations $s_i \equiv [S_{(i)}]$ and $s_j \equiv [S_{(j)}]$ do not increase in equal amounts when t_S is increased, but are related by a factor $(\sigma s)^{i-j}$. Since $\sigma s < 1$ for $t_S < \infty$, there is a lag between the rise of $S_{(i)}$ and $S_{(j)}$, where $S_{(i)}$ increases before $S_{(j)}$ for $i < j$; this lag is more dramatic the bigger the difference $|i - j|$ (Figure 1.13B, green curve). In the polymerizing system, as t_S increases, the ratio of $S_{(i)}$ and $S_{(j)}$ will tend to 1, but by then the between-class prozone is taking its toll. In sum, the "stealing" of ligands by higher length classes from lower ones is the reason for the turn towards a steeper slope of Q_{poly} at t_S values at which polymerization becomes effective (Figure 4A in the main text). Incidentally, the shift of ligands from lower towards higher valency classes also tends to flatten the intrinsic slope of the downward leg of lower valency classes after the prozone peak, contributing further to prozone mitigation in the overall system.

1.7.5 Comparison between polymerizing and multivalent scaffold systems

In the main text, Figure 4A and 4B, we compare multivalent scaffolds with the polymerizing scaffold system. Figure 1.14 places that comparison in the context of the full Q_{poly} surface to show the effectiveness of regulating the affinity σ.

While even for $n_A = n_B = n$ and $\alpha = \beta$, Q_{multi} is a cumbersome expression, determining the concentration of scaffold agents t_S for which $dQ_{\text{multi}}/dt_S = 0$ yields a simple solution

$$t_S = \frac{1}{n}\left(\frac{1}{\alpha} + \frac{t_A + t_B}{2}\right). \tag{1.49}$$

Equation (1.49) shows that when plotting Q_{multi} against the concentration of sites $t_{\text{sit}} = nt_S$, as in Figure 1.14 and Figure 4A of the main text, the prozone peaks line up for all valencies n.

Expanding Q_{multi} (assuming $n_A = n_B = n$) in t_S near zero, yields

$$Q_{\text{multi}} = \frac{\alpha t_A \beta t_B}{1 + \alpha t_A + \beta t_B + \alpha\beta t_A t_B} n^2 t_S + O(t_S^2). \tag{1.50}$$

Hence in a log-log plot, the up-leg of $Q_{\text{multi}}(n)$ has, to leading order, slope 1 and offset n when plotted against sites $t_{\text{sit}} = nt_S$ as in Figure 4A of the main text. Similarly, expanding Q_{multi} in t_S near infinity, yields

$$Q_{\text{multi}} = t_A t_B \frac{1}{t_S} + O(1/t_S^2), \tag{1.51}$$

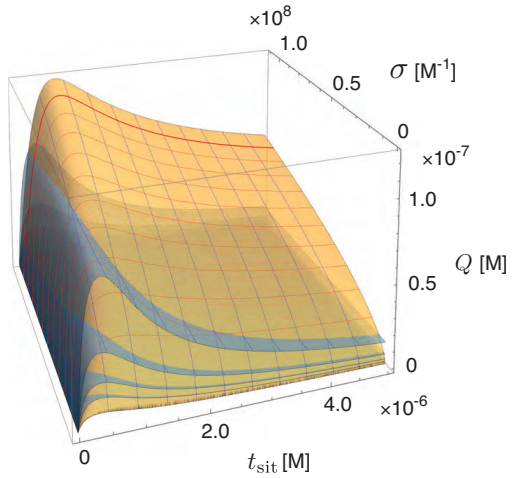


Figure 1.14: Polymerizing scaffold and multivalent scaffolds. The surface shows Q_{poly} as function of t_S and σ, giving more context to Figure 4B in the main text. The emphasized mesh line (red) at $\sigma = 10^8$ corresponds to the Q-function of the polymerizing scaffold system shown in Figure 4B of the main text. $\alpha = \beta = 10^7$ M^{-1}, $t_A = 15 \cdot 10^{-9}$ M, $t_B = 5 \cdot 10^{-7}$ M.

and hence, to leading order, a slope of -1 in a log-log plot in the down-leg after the prozone peak and an offset of n when plotted against t_{sit} as in Figure 4A of the main text.

The expansion of Q_{poly} in t_S (= t_{sit}) around zero yields

$$Q_{\text{poly}} = \frac{\alpha t_A \beta t_B}{1 + \alpha t_A + \beta t_B + \alpha\beta t_A t_B} t_S + [f(\alpha, \beta, t_A, t_B) + g(\alpha, \beta, t_A, t_B)\sigma]\, t_S^2 + O(t_S^3) \tag{1.52}$$

with $f()$ and $g()$ functions of the indicated parameters. The leading-order term is the same as the Q_{multi} of the monovalent scaffold, *and is independent of* σ, which enters the second-order term. Accordingly, for small t_S, Q_{poly} hugs the Q of the monovalent scaffold as if there was no polymerization; as t_S increases, σ (i.e. polymerization) becomes effective and Q_{poly} doubles its slope upward. This is clearly seen in Figure 4A of the main text. Some microscopic consequences from building up a length distribution as t_S increases are discussed in section 1.7.4.

Expanding Q_{poly} in t_S at infinity yields

$$Q_{\text{poly}} = 2 t_A t_B \sqrt{\sigma} \sqrt{\frac{1}{t_S}} + O(1/t_S^{3/2}), \tag{1.53}$$

where the $p(t_S, t_A, \alpha)p(t_S, t_B, \beta)$ component scales with $t_A t_B / t_S^2$ and the Q_{max} component with $2t_S^{3/2}\sqrt{\sigma}$ to leading order. As a result, the slope of the down-leg of Q_{poly} after the prozone peak in a log-log plot is $-1/2$.

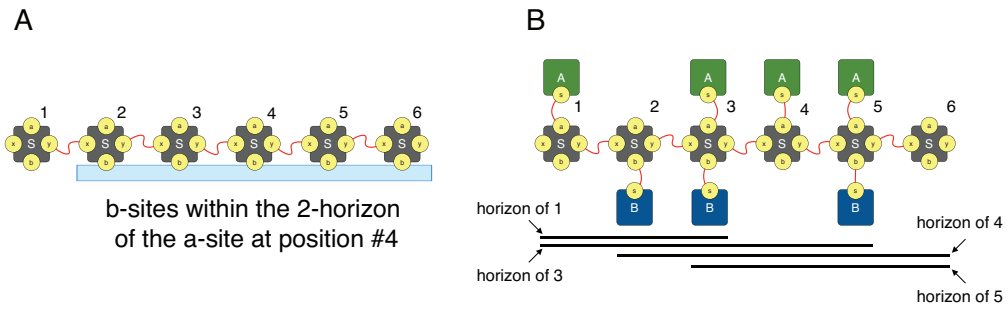


Figure 1.15: Interaction horizon. The schematic illustrates the case in which the horizon h is less than the polymer length n. In this case, each A-binding position can interact with at most h B-binding positions on its "left" or "right" side. When $h \geq n$, every A-position can interact with every B-position.

1.7.6 Interaction horizon

Structural constraints might prevent every catalyst A on a polymeric scaffold from interacting with all substrates B bound to the same polymer. To obtain a rough sense of how such constraints could impact the catalytic potential Q, we define an "interaction horizon," h, Figure 1.15. The horizon h is the farthest distance in terms of scaffold bonds that a bound A can "reach." This means that a given bound enzyme A can interact with at most $2h + 1$ substrate agents B: h to its "left," h to its "right" and the one bound to the same protomer, Figure 1.15A. For example, in Figure 1.15B, the 2-horizon of the A at position 1 includes the Bs at positions 2 and 3, but not at position 5. Likewise, the B at position 2 is outside the 2-horizon of the A at position 5, whereas all Bs are within reach of the A at position 3. Clearly, the interaction horizon only modulates the Q_{max} in equation (1.40) of a polymer of length n; more precisely, it modulates the interaction factor—the n^2 in the first equation of (1.41). We now write this factor as $q_{max}(n, h)$; it replaces the n^2 in (1.41).

To reason about the catalytic combinations, we first consider the case $0 \leq h \leq \lfloor n/2 \rfloor$:

$$q_{max}(n,h) = \underbrace{(n-2h)(2h+1)}_{\text{I}} + \underbrace{2h(h+1)}_{\text{II}} + \underbrace{2\sum_{k=1}^{h-1}(h-k)}_{\text{III}} = \quad n(2h+1) - h(h+1). \tag{1.54}$$

Term I refers to the $n - 2h$ positions in the middle region of the chain that can interact with the full complement of $2h + 1$ sites within its horizon. Term II refers to the h positions at each end of the chain and accounts for all $h + 1$ sites reachable towards the interior of the chain. Term III accounts for the remaining $h - k$ locations towards the end of the chain that can be reached from a position considered in term II; these locations depend on that position's distance k from the end of

the chain. For $\lfloor n/2 \rfloor < h \leq n-1$, we obtain

$$q_{max}(n,h) = \underbrace{(2h-n)n}_{\text{I'}} + \underbrace{2(n-h)(h+1)}_{\text{II'}} + \underbrace{2\sum_{k=1}^{n-h}(k-1)}_{\text{III'}} = n(2h+1) - h(h+1). \quad (1.55)$$

In analogy to (1.54), Term I' refers to the $2h-n$ positions that can access the whole chain; term II' accounts for the $h+1$ locations spanned by the inward-facing side of the remaining $n-h$ positions at each end of the chain. Finally, term III' accounts for the locations covered by the outward facing side of these $n-h$ positions.

If the horizon h is larger than the polymer length n, then every A-position can interact with every B-position on the polymeric scaffold and $q_{max}(n,h) = n^2$. Merging this with (1.54) and (1.55) yields

$$q_{max}(n,h) = \begin{cases} n(2h+1) - h(h+1), & \text{for } 0 \leq h \leq n-1 \\ n^2, & \text{for } h \geq n \end{cases} \quad (1.56)$$

which appears in the main text. The corner cases are covered correctly: $q_{max}(n,0) = n$ and $q_{max}(n,n-1) = n^2$. (Note that $h = n$ yields the same result as $h = n-1$, which is useful below.)

We use (1.56) to calculate two scenarios. In scenario 1, h is a simple linear function of the length n: $h = \xi n$ with $0 \leq \xi \leq 1$. In other words, every A can monitor the same fraction ξ of B-binding sites on a polymer of any size. This seems rather unrealistic (and makes h a continuous variable, although that appears to work just fine). However, scenario 1 may serve as a comparison with the subsequent, more realistic scenario 2.

When $h = \xi n$, h is always less or equal than n and the first case of (1.56) applies. Using $q_{max}(n,h)$ with $h = \xi n$ instead of n^2 in the first equation of (1.41) yields

$$\begin{aligned} Q_{max}(\xi) &= \sum_{n=1}^{\infty} \left[n(2h+1) - h(h+1)\right] \sigma^{n-1} s^n = \sum_{n=1}^{\infty} \left[n(2\xi n+1) - \xi n(\xi n+1)\right] \sigma^{n-1} s^n \\ &= \frac{1}{\sigma}\left[\xi(2-\xi)\sum_{n=1}^{\infty} n^2 \sigma^n s^n + (1-\xi)\sum_{n=1}^{\infty} n\sigma^n s^n\right] = \xi(2-\xi)\frac{s(1+\sigma s)}{(1-\sigma s)^3} + (1-\xi)\frac{s}{(1-\sigma s)^2}, \end{aligned} \quad (1.57)$$

which leads to

$$Q = p(t_S, t_A, \alpha) p(t_S, t_B, \beta) \left(\xi(2-\xi)\frac{s(1+\sigma s)}{(1-\sigma s)^3} + (1-\xi)\frac{s}{(1-\sigma s)^2}\right). \quad (1.58)$$

For $\xi = 1$, the expression (1.58) becomes (1.41), as a horizon that equals the length of any polymer does not affect Q_{max}. For $\xi = 0$, we get

$$Q = p(t_S, t_A, \alpha)p(t_S, t_B, \beta)\frac{s}{(1-\sigma s)^2} = p(t_S, t_A, \alpha)p(t_S, t_B, \beta)t_S, \tag{1.59}$$

because of $t_S = s\,dW/ds$ for the polymer-only system. Thus, for $\xi = 0$, we recover the Q of the simple monovalent scaffold, since in this case the organization of protomers into polymers does not affect catalytic potential. Scenario 1 is shown in Figure 1.16, panels A and B.

In scenario 2, h = const for all lengths n, which means a "hard" horizon independent of polymer size. This scenario is more realistic. $Q_{max}(h)$ becomes

$$\begin{aligned}
Q_{max}(h) &= \sum_{n=1}^{\infty} q_{max}(n,h)\sigma^{n-1}s^n = \sum_{n=1}^{h} n^2\sigma^{n-1}s^n + \sum_{n=h+1}^{\infty}[n(2h+1) - h(h+1)]\sigma^{n-1}s^n \\
&= \frac{1}{\sigma}\left\{\sum_{n=1}^{h} n^2(\sigma s)^n + (2h+1)\sum_{n=h+1}^{\infty} n(\sigma s)^n - h(h+1)\sum_{n=h+1}^{\infty}(\sigma s)^n\right\} \\
&= \frac{1}{\sigma}\left\{\frac{\sigma s(1+\sigma s) - (\sigma s)^{h+1}[(h+1)^2 - (2h^2+2h-1)\sigma s + h^2(\sigma s)^2]}{(1-\sigma s)^3}\right. \\
&\quad \left. + (2h+1)\frac{(\sigma s)^{h+1}(h+1-h\sigma s)}{(1-\sigma s)^2} - h(h+1)\frac{(\sigma s)^{h+1}}{1-\sigma s}\right\} \\
&= \frac{s\left(1+\sigma s - 2(\sigma s)^{h+1}\right)}{(1-\sigma s)^3},
\end{aligned} \tag{1.60}$$

yielding

$$Q = p(t_S, t_A, \alpha)p(t_S, t_B, \beta)\frac{s\left(1+\sigma s - 2(\sigma s)^{h+1}\right)}{(1-\sigma s)^3}, \tag{1.61}$$

which is equation (6) of the main text. Expression (1.61) becomes (1.59) for $h = 0$, as we would expect. As h increases, (1.61) quickly converges to the infinite horizon case (1.41), since $\sigma s < 1$ raised to the power of h becomes negligible. Scenario 2 is shown in Figure 1.16, panels B and D. As suggested in Figure 1.17, even restrictive structural constraints (small h) make only a relatively modest dent in the catalytic potential of the polymerizing scaffold when compared to that of the plain Michaelis-Menten scenario.

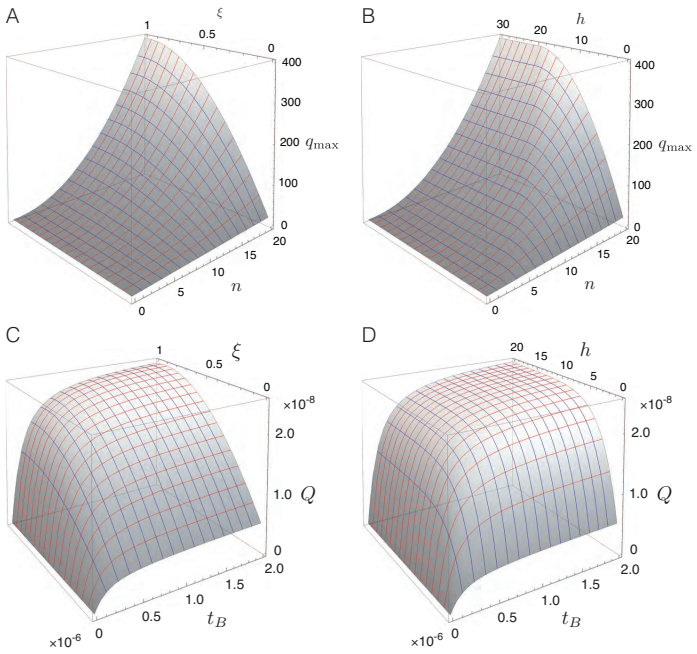


Figure 1.16: Interaction horizon scenarios. A: $q_{max}(n, h)$, equation (1.56), for scenario 1 when $h = \xi n$ ($0 \leq \xi \leq 1$). B: $q_{max}(n, h)$, equation (1.56), for scenario 2 when h is a constant independent of n. The difference to panel A is that the surface of scenario 2, once h exceeds n, is a quadratic extension of the surface of scenario 1 in panel A at $\xi = 1$. C: The Q-surface (1.58) for scenario 1 as a function of substrate concentration t_B. D: The Q-surface (1.61) for scenario 2 as a function of substrate concentration t_B. In Figure 1.17, this surface is compared against the Michaelis-Menten case. The parameter values in C and D are: $\alpha = \beta = 10^7$ M and $\sigma = 10^8$ M, $t_A = 15 \cdot 10^{-9}$ M, and $t_S = 60 \cdot 10^{-9}$ M.

1.7.7 The discrete case

While we strive for a reasonably self-contained exposition, some details are only asserted for brevity and are developed in a forthcoming manuscript providing a more general treatment of equilibrium assembly.

In the following, we use the same symbols for the binding affinities α, β, and σ as in the continuum case, but they must now be understood as "stochastic affinities." Specifically, if γ' is a binding affinity in the continuum case, the stochastic affinity γ (in units of molecules^{-1}) is related as $\gamma = \gamma'/(AV)$, where V is the effective volume hosting the system and A is Avogadro's constant. Thus a polymerization affinity of 3 molecules^{-1} in the discrete case corresponds to about $1.8 \cdot 10^{12}$ M^{-1} in a cell volume of 10^{-12} L in the continuum setting.

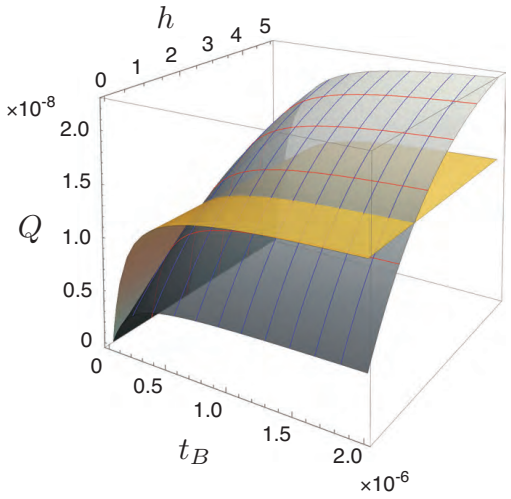


Figure 1.17: The impact of the interaction horizon. The Q-surface (1.61) with hard horizon h, gray, and the plain dimerization (Michaelis-Menten) surface, orange, for the parameter settings corresponding to Figure 3A in the main text ($\alpha = \beta = 10^7$ M^{-1} and $\sigma = 10^8$ M^{-1}, $t_A = 15 \cdot 10^{-9}$ M, $t_S = 60 \cdot 10^{-9}$ M). At $t_S = 60$ nM (the curve with the red dot in Figure 3A of the main text) a horizon $h = 2$ is already sufficient to achieve a higher catalytic potential than the direct binding of enzyme to substrate. This suggests that structural constraints forcing a small interaction horizon might not undermine the efficacy of a polymerizing scaffold.

Average catalytic potential

Our objective is to calculate the average catalytic potential $\langle Q \rangle$ of a scaffold mixture, defined as

$$\langle Q \rangle = \sum_{i=0}^{\min(t_A,n)} \sum_{j=0}^{\min(t_B,n)} i\, j\, \langle S_{ij} \rangle, \tag{1.62}$$

where S_{ij} is any scaffold (polymer or multivalent) with n A-binding sites, of which i are occupied, and n B-binding sites, of which j are occupied. More precisely, S_{ij} is the set of all configurations, or molecular species, with i and j agents of type A and B bound, respectively. $\langle S_{ij} \rangle$ is the average or expected total number of such configurations in an equilibrium system with resource vector $\vec{t} = (t_A, t_B, t_S)' \in \mathbb{N}_0^3$. The $'$ means a transpose. (t_S is typically the number of scaffolds of a given valency n or the number of protomers in a polymerizing system. When considering mixtures of scaffolds of different valencies i, t_S is generalized accordingly.)

This raises the need to compute $\langle S_{ij} \rangle$, which requires a little detour. We start by defining a few well-known quantities.

Assume a system of molecular interactions with a set of atomic building blocks, or atoms for short, $\{X_1, \ldots, X_T\}$ (in the main text typically $T = 3$, namely A, B, and S) that give rise to a set

of configurations $\{Y_1, \ldots, Y_C\}$. Since we are interested in equilibrium, the precise nature of the interactions is irrelevant as long as the resulting systems have the same set of reachable molecular species. The assembly scenarios considered in the main text only require binding and unbinding interactions.

Boltzmann factor of a molecular species

Each molecular species Y_i has a Boltzmann factor given by

$$\varepsilon_i = \prod_r \gamma_r, \tag{1.63}$$

where $\gamma_r = \exp(-\frac{\Delta G_r^0}{kT})$ is the binding constant of the r-th reaction and the product runs over a series of reactions r that constitute an assembly path from atomic components (A, B, and S). Note that, in the discrete case, ε_i is not divided by the number of symmetries ω_i as in the continuum case (main text leading up to Eq. [1]). The effect of symmetries is accounted for in the state degeneracy, Eq. (1.65) below, which considers all instances of Y_i in a given state. As a consequence, $-kT \log \varepsilon_i$ is not the free energy of formation, but just the internal energy due to bond formation.

Boltzmann factor of a state

By extension, the Boltzmann factor of a system *state* $\vec{n} = (n_1, n_2, \ldots, n_C)'$, where n_i is the number of particles of species Y_i, is given by

$$\varepsilon(\vec{n}) = \prod_{i=1}^{C} (\varepsilon_i)^{n_i}. \tag{1.64}$$

More precisely, (1.64) is the Boltzmann factor associated with a particular *realization* of the state $\vec{n}$, as all atoms are labelled (distinguishable).

Degeneracy of a state

A state $\vec{n}$ is the specification of a multiset of species in which atom labels are ignored. The degeneracy $d(\vec{t}, \vec{n})$ of a state $\vec{n}$ with resource vector $\vec{t} = (t_1, \ldots, t_T)$ is the number of distinct ways of realizing it by taking into account atom labels. Let $\mu_{i,j}$ denote the number of atoms of type X_j contained in one instance of Y_i. For a given resource vector $\vec{t}$ the set $\Sigma(\vec{t})$ of states $\vec{n}$ that are compatible with it satisfy $t_j = \sum_{i=1}^{C} \mu_{i,j} n_i$ for every atom type X_j. Hence, the degeneracy of a state $\vec{n} \in \Sigma(\vec{t})$ is given by

$$d(\vec{t}, \vec{n}) = \frac{\prod_{i=1}^{T} t_i!}{\prod_{i=1}^{C} n_i! \prod_{i=1}^{C} (\omega_i)^{n_i}}. \tag{1.65}$$

The numerator counts all permutations of the atoms that constitute the system, the first product in the denominator corrects for all orderings among the n_i copies of species Y_i and the second product corrects for all symmetries associated with Y_i.

The partition function for a given resource vector

As usual,

$$Z(\vec{t}) = \sum_{\vec{n}\in\Sigma(\vec{t})} d(\vec{t},\vec{n})\varepsilon(\vec{n}), \tag{1.66}$$

where the sum runs over all admissible states given resource vector $\vec{t}$. The equilibrium probability of a state $\vec{n}$ is given by

$$p(\vec{t},\vec{n}) = \frac{d(\vec{t},\vec{n})\varepsilon(\vec{n})}{Z(\vec{t})}. \tag{1.67}$$

The average number of instances of a specific configuration in equilibrium

For a given resource vector $\vec{t}$ a species Y_i occurs in various numbers n_i across the states $\vec{n}$ in the admissible set $\Sigma(\vec{t})$. The average abundance of Y_i, $\langle n_i\rangle$ then is

$$\langle n_i\rangle = \sum_{\vec{n}\in\Sigma(\vec{t})} n_i p(\vec{t},\vec{n}) = \frac{1}{Z(\vec{t})}\sum_{\vec{n}\in\Sigma(\vec{t})} n_i d(\vec{t},\vec{n})\varepsilon(\vec{n}). \tag{1.68}$$

The workhorse for the discrete treatment of the scaffolding systems discussed in the main text is the following Theorem.

Theorem:
The average equilibrium abundance $\langle n_i\rangle$ of species Y_i in an assembly system with resource vector $\vec{t}$ is given by

$$\langle n_i\rangle = \varrho(\vec{t},Y_i)\varepsilon_i\frac{Z(\vec{t}-\vec{\mu}_i)}{Z(\vec{t})}, \tag{1.69}$$

where $\vec{\mu}_i = (\mu_{i,1},\ldots,\mu_{i,T})'$ is the atomic content vector of species Y_i; $\varrho(\vec{t},Y_i)$ is the number of distinct realizations of a single instance of Y_i given resources $\vec{t}$; and $Z(\vec{t}-\vec{\mu}_i)$ is the partition function of a system in which the atomic resources have been decreased by the amount needed to build one instance of Y_i.

It is immediate from (1.65) that

$$\varrho(\vec{t},Y_i) = d(\vec{t},\vec{Y}_i) = \frac{\prod_{j=1}^{T} t_j!}{\prod_{j=1}^{T}(t_j-\mu_{i,j})!\,\omega_i}, \tag{1.70}$$

where $\vec{Y}_i$ denotes a unit vector in the Y_i direction. We provide a proof of the theorem using generating functions elsewhere. However, to see why the claim holds, we reason as follows. The subset of $\Sigma(\vec{t})$ in which we restrict ourselves to states $\vec{n}$ that contain at least one copy of Y_i stands in a 1-1 correspondence to the unrestricted state space $\Sigma(\vec{t}-\vec{\mu}_i)$, because any realization of Y_i in $\Sigma(\vec{t})$ occurs in all possible contexts and these contexts are precisely the states of $\Sigma(\vec{t}-\vec{\mu}_i)$. The question then is how the degeneracy and the energy content of a state $\vec{n} \in \Sigma(\vec{t}-\vec{\mu}_i)$ change by adding $\vec{\mu}_i$ atoms to realize one instance of Y_i. The degeneracy of state $\vec{n} \in \Sigma(\vec{t}-\vec{\mu}_i)$ is amplified (multiplied) by $\varrho(\vec{t}, Y_i)$ realizations of Y_i, but one instance of Y_i is added to those the state already had and so we also need to divide by $n_i + 1$ to compensate for indistinguishable permutations within the instances of Y_i, see (1.65). Thus, $d(\vec{t}, \vec{n} + \vec{Y}_i) = (\varrho(\vec{t}, Y_i)/(n_i + 1))d(\vec{t} - \vec{\mu}_i, \vec{n})$ and the Theorem follows as summarized symbolically:

$$\frac{1}{Z(\vec{t})} \sum_{\substack{\vec{n}\in\Sigma(\vec{t}) \\ n_i \geq 1}} n_i d(\vec{t}, \vec{n})\varepsilon(\vec{n}) = \frac{1}{Z(\vec{t})} \sum_{\vec{n}\in\Sigma(\vec{t}-\vec{\mu}_i)} (n_i + 1)\frac{\varrho(\vec{t}, Y_i)}{n_i + 1} d(\vec{t} - \vec{\mu}_i, \vec{n})\varepsilon_i\varepsilon(\vec{n}) = \varrho(\vec{t}, Y_i)\varepsilon_i \frac{Z(\vec{t} - \vec{\mu}_i)}{Z(\vec{t})}. \tag{1.71}$$

It remains to compute the partition function of the assembly systems discussed in the main text, which is not too difficult and provided in the subsequent section 1.7.8.

1.7.8 Partition functions and average catalytic potential

Polymerizing scaffold without ligands

Let a state contain i bonds (not necessarily in the same polymer). Any such state has a Boltzmann factor σ^i, where σ is the binding affinity between two scaffold protomers. We count the number of ways to realize i bonds as follows. Line up the t_S (labelled) protomers and observe that there are $t_S - 1$ slots between protomers where a bond could be inserted. Thus there are $\binom{t_S-1}{i}$ ways of inserting i bonds and the insertion of i bonds always creates $t_S - i$ molecules. For each choice of i slots there are $t_S!$ permutations of the protomers. Since the order in which a choice of bond locations creates the $t_S - i$ molecules is irrelevant, we must reduce the label permutations by $(t_S - i)!$ object permutations to obtain the degeneracy d_i of a state with i bonds. The partition function is therefore

$$Z_{t_S}^{\text{poly}} = \sum_{i=0}^{t_S-1} \sigma^i \binom{t_S - 1}{i} \frac{t_S!}{(t_S - i)!}. \tag{1.72}$$

The number of possible realizations of a single polymer s_n of length n is $t_S!/(t_S - n)!$, which yields with (1.69) for the average number of polymers of length n, $\langle s_n \rangle$:

$$\langle s_n \rangle = \frac{t_S!}{(t_S - n)!}\sigma^{n-1}\frac{Z_{t_S-n}^{\text{poly}}}{Z_{t_S}^{\text{poly}}}. \tag{1.73}$$

Figure 1.18 compares the length distributions of equivalent continuum and discrete polymerization systems.

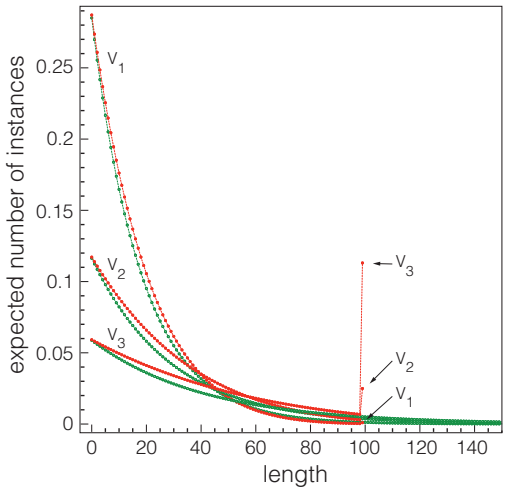


Figure 1.18: Length distribution in continuum and discrete polymerization. A continuum and discrete polymerization system are set up with equivalent parameters assuming a base volume $V = 10^{-15}$ L (the order of magnitude of a bacterial cell). Their length distributions are compared for three volumes: $V_1 = 0.05V$, $V_2 = 0.02V$, $V_3 = 0.01V$. A change in volume means a change in affinity for the discrete system and a change in protomer concentration for the continuum system, i.e. $t_S = 100$ protomers or $t_S = 100/(AV_i)$ M; discrete affinity $\sigma_s = 10^8/(AV_i)$ molecules^{-1} or continuum affinity $\sigma_d = 10^8$ M^{-1}. The green curves are associated with the continuum system (equation 1.42 and the red ones with the discrete case (equation 1.73. Associated volumes are as indicated in the graph. Since the curves cross, the maximer is also marked with the corresponding volume. The continuum distribution is cut off at 150.

Average catalytic potential of the polymerizing scaffold with ligands

Because of binding independence, the partition function of this system is the product of three partition functions: $Z^{\text{poly}}_{t_S} Z^{\text{dimer}}_{t_S,t_A} Z^{\text{dimer}}_{t_S,t_B}$, with $Z^{\text{dimer}}_{t_S,t_X}$ the partition function of a system in which S-agents and X-agents can dimerize with affinity γ. $Z^{\text{dimer}}_{t_S,t_X}$ is simple to obtain: choose i agents of type A, i agents of type S, and pair them:

$$Z^{\text{dimer}}_{t_S,t_X} = \sum_{i=0}^{\min(t_S,t_X)} \gamma^i \binom{t_S}{i}\binom{t_X}{i} i!. \tag{1.74}$$

Putting this together yields the partition function for resource vector $\vec{t} = (t_A, t_B, t_S)$

$$Z(\vec{t}) = \left[\sum_{k=0}^{t_S-1} \sigma^k \binom{t_S-1}{k} \frac{t_S!}{(t_S-k)!}\right] \left[\sum_{i=0}^{\min(t_S,t_A)} \alpha^i \binom{t_A}{i}\binom{t_S}{i} i!\right] \left[\sum_{j=0}^{\min(t_S,t_B)} \beta^j \binom{t_B}{j}\binom{t_S}{j} j!\right]$$
$$= Z^{\text{poly}}_{t_S} Z^{\text{dimer}}_{t_S,t_A} Z^{\text{dimer}}_{t_S,t_B}. \tag{1.75}$$

The total number of realizations, $\varrho(\vec{t}, \{A_i S_l B_j\})$ of polymers of length l with i A-agents and j B-agents attached, and thus each with Boltzmann factor $\sigma^{l-1}\alpha^i\beta^j$, is given by

$$\varrho(\vec{t}, \{A_i S_l B_j\}) = \frac{t_S!}{(t_S-l)!}\binom{l}{i}\binom{t_A}{i} i! \binom{l}{j}\binom{t_B}{j} j! = \binom{l}{i}\binom{l}{j}\frac{t_S!}{(t_S-l)!}\frac{t_A!}{(t_A-i)!}\frac{t_B!}{(t_B-i)!}$$
$$= \binom{l}{i}\binom{l}{j}\frac{\vec{t}!}{(\vec{t}-\vec{v})!} \tag{1.76}$$

where $\vec{v} = (i, j, l)$ is the composition vector of the configuration and we define for brevity the factorial of a vector as the product of the factorials of its components. Putting all this together yields the average catalytic potential $\langle Q \rangle$

$$\langle Q_{\text{poly}} \rangle = \sum_{l=1}^{t_S} \sum_{i=0}^{\min\{l,t_A\}} \sum_{j=0}^{\min\{l,t_B\}} \underbrace{i\,j}_{\substack{\#\text{ of}\\ \text{interactions}}} \underbrace{\underbrace{\binom{l}{i}\binom{l}{j}\frac{\vec{t}!}{(\vec{t}-\vec{v})!}}_{\substack{\text{total realizations of}\\ \text{configurations with } \vec{v}}} \sigma^{l-1}\alpha^i\beta^j \frac{Z(\vec{t}-\vec{v})}{Z(\vec{t})}}_{\text{average total counts}}. \tag{1.77}$$

Average catalytic potential of the multivalent scaffold with ligands

The case of a multivalent scaffold with m binding sites for A and n binding sites for B follows the lines of section 1.7.8. For each type of binding sites, one can formulate a partition function in full analogy to $Z^{\text{dimer}}_{t_S,t_X}$, but with $m\,t_S$ (or $n\,t_S$) sites available to bind i agents of type A (or j agents of type B) to yield a state with Boltzmann factor $\alpha^i\beta^j$. Thus, the partition function for a multivalent scaffold system is

$$Z^{\text{multi}}_{t_A,t_B,t_S} = \sum_{i=0}^{\min(m\,t_S,t_A)} \sum_{j=0}^{\min(n\,t_S,t_B)} \alpha^i\,\beta^j \binom{t_A}{i}\binom{m\,t_S}{i} i! \binom{t_B}{j}\binom{n\,t_S}{j} j! \tag{1.78}$$

The average number of scaffolds loaded with i ligands of type A and j ligands of type B *in a particular configuration* then becomes

$$\langle n_{ij} \rangle = \frac{t_A!}{(t_A-i)!}\frac{t_B!}{(t_B-j)} t_S\,\alpha^i\,\beta^j \frac{Z^{\text{multi}}_{t_A-i,t_B-j,t_S-1}}{Z^{\text{multi}}_{t_A,t_B,t_S}}. \tag{1.79}$$

Finally, for the average catalytic potential we have

$$\langle Q_{\text{multi}} \rangle = \sum_{i=0}^{\min(t_A,m)} \sum_{j=0}^{\min(t_B,n)} i\, j \binom{m}{i}\binom{n}{j}\langle n_{ij}\rangle. \tag{1.80}$$

1.7.9 Remarks on numerical evaluation

While expressions (1.77) and (1.80) are explicit, their use with large particle numbers—t_S, t_A and t_B—is limited by numerical instabilities (even after efficiency rearrangements). In a separate paper, we connect assembly systems with the theory of analytic combinatorics (Flajolet and Sedgewick, 2009), which provides direct approximations based on viewing generating functions as analytic functions over the complex numbers. In our hands, these approximations are not accurate enough over the entire parameter range for the present context. Our figures were therefore generated using the exact expressions (1.77) and (1.80), using arbitrary-precision calculations (to 100 significant digits) in Mathematica (Wolfram Research Inc., 2019), and employing relatively modest particle numbers to keep computation times reasonable.

1.7.10 The maximer probability and 1D percolation

The probability of observing the longest possible polymer, given protomer resources, is obtained from (1.73) by setting $n = t_S$:

$$\langle s_{\max} \rangle = \frac{t_S!\, \sigma^{t_S-1}}{Z^{\text{poly}}_{t_S}}. \tag{1.81}$$

This probability is graphed as a function of t_S and σ in Figure 5A of the main text.

There is an analogy between 1D bond percolation and polymerization at our level of abstraction. The analogy is an exact correspondence in the case of continuum polymerization and bond percolation on an infinite 1D lattice.

A basic quantity in 1D percolation is the mean number of chains (clusters) of size n normalized per lattice site, which is given by $p^{n-1}(1-p)^2$, where p is the probability of a bond between adjacent lattice sites and functions as a parameter. The same expression obtains in terms of the concentration of polymers of length n normalized per protomer (Flory, 1936; Reynolds, Stanley, and Klein, 1977):

$$\frac{s_n}{t_S} = p^{n-1}(1-p)^2. \tag{1.82}$$

In the context of polymers, the bond probability is not the primary parameter, but a function of the basic parameters t_S and σ. Following Flory (Flory, 1936), we can express p as

$$p = \frac{t_S - W}{t_S} = 1 - \frac{1}{t_S}\frac{s}{1-\sigma s}, \tag{1.83}$$

with W the concentration of all polymers as defined in (1.14) for $a = b = 0$ and given more compactly by (1.21). The first equality defines p in terms of the difference between the maximal possible concentration of objects in the system (t_S) and the actual concentration of objects; this difference is the concentration of bonds. Using (1.42) for s yields

$$p = 1 - \frac{2}{1 + \sqrt{1 + 4\sigma t_S}}. \tag{1.84}$$

Together, expressions (1.82) and (1.84) are equivalent to (1.42) and connect simple polymerization to percolation. As well-known, in the infinite/continuum case, percolation can only occur at $p = 1$, which is to say in the limit of $t_S \to \infty$ or $\sigma \to \infty$.

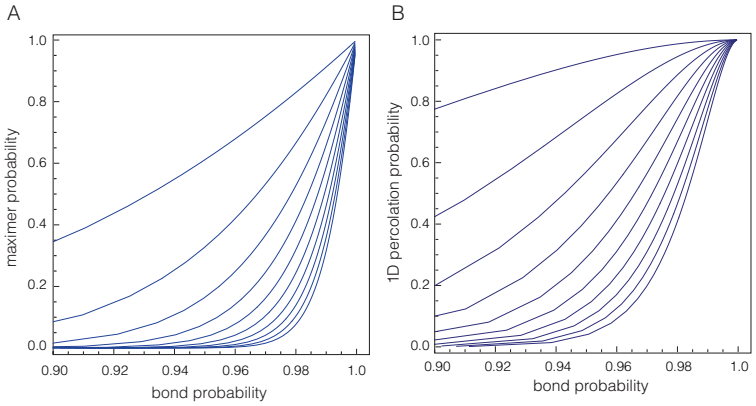


Figure 1.19: Finite size 1D bond percolation and polymerization. A: This panel is panel B of Figure 5 in the main text. It depicts the probability of the maximer (1.81) as a function of p_{bond} as given by (1.85). Each curve represents a particular t_S-value for which σ sweeps from 1 to 1000 molecules $^{-1}$. t_S ranges from 10 (topmost curve) to 100 (lowest curve) in increments of 10. B: The plot depicts the 1D bond percolation probability (1.86) as a function of the same bond probabilities used in panel A. The comparison serves to illustrate the difference between 1D bond percolation and polymerization while also emphasizing the analogy. On the other hand, bond percolation on an infinite 1D lattice is equivalent to polymerization described in terms of continuous concentrations.

The analogy persists, but the exact correspondence breaks down in the finite, i.e. discrete, case. The percolation probability in the polymerization case is $\langle s_{\text{max}} \rangle$ as given by (1.81). The bond probability, p_{bond}, is the expected fraction of bonds and can be computed following the arguments that led to (1.72). We obtain

$$p_{\text{bond}} = \frac{1}{t_S - 1} \frac{\sum_{i=1}^{t_S - 1} i\sigma^i \binom{t_S - 1}{i} \frac{t_S!}{(t_S - i)!}}{Z_{t_S}^{\text{poly}}}. \tag{1.85}$$

In 1D bond percolation, the percolation probability is

$$p_{\text{perc}} = 1 - (1-p)^2 \sum_{i=0}^{t_S-2} i p^{i-1} = p^{t_S-2}(t_S - p(t_S-2) - 1), \tag{1.86}$$

with t_S the size of the lattice and p the bond probability.

In Figure 5B of the main text, we sweep across a range for t_S and σ. For each (t_S, σ) pair, we calculate the corresponding p_{bond} via (1.85) as the abscissa and $\langle s_{\max}\rangle$ via (1.81) as the ordinate. This graph is reproduced as Figure 1.19B for comparison with finite-size bond percolation, Figure 1.19A. Clearly in (1.86) p is just a parameter, but in Figure 1.19A, we compute it via (1.85) using the same sweep over t_S and σ as for Figure 1.19B to make comparison meaningful. The view from percolation is useful because it packages the dependency on t_S and σ into the single quantity p (or p_{bond}).

1.7.11 Scaling behavior

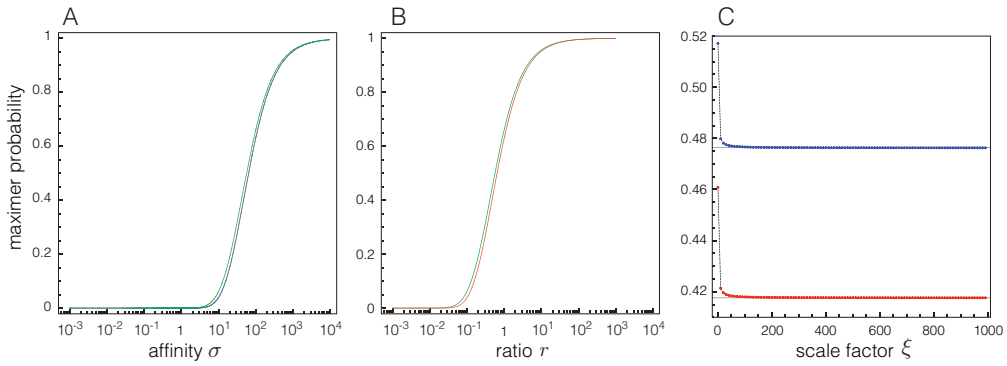


Figure 1.20: Scaling behavior of the maximer distribution. The panels illustrate the approximate scaling behavior of $\langle s_{\max}\rangle$ from different perspectives implied by (1.88). In all three panels, the ordinate is the maximer probability as given by (1.81). A: The graph exemplifies the relation (1.88) by plotting three curves, blue: $\langle s_{\max}\rangle[10, 0.1\sigma]$, red: $\langle s_{\max}\rangle[100, \sigma]$, and green: $\langle s_{\max}\rangle[1000, 10\sigma]$ as a function of the affinity σ. The blue and green graphs are related to the (arbitrary) red baseline graph by scale factors $\xi = 0.1$ and $\xi = 10$, respectively. The red and blue graphs sit on top of each other, while green has a slight (and slightly σ-dependent) shift to the left. B: This panel illustrates the scaling version (1.89), comparing red: $\langle s_{\max}\rangle[1000, r\,1000]$ with green: $\langle s_{\max}\rangle[10, r\,10]$, sweeping along r. C: The graph in this panel shows an integer sweep of the scale factor ξ, as per (1.88), for two pairs, $[t_S, \sigma] = [10, 5]$ (red), $[t_S, \sigma] = [10, 6]$ (blue). The scaling relation is well fulfilled except for very small particle numbers.

We refine the notation for the maximer probability (1.81) to emphasize the dependence on the parameters t_S and σ,

$$\langle s_{\max}\rangle[t_S, \sigma] \equiv \langle s_{\max}\rangle, \tag{1.87}$$

in order to note an approximate scaling relation that we observe numerically:

$$\langle s_{\max}\rangle[t_S, \sigma] \approx \langle s_{\max}\rangle[\xi t_S, \xi\sigma], \tag{1.88}$$

with $\xi > 0$ a dimensionless scale factor. Two systems are approximately equivalent if their protomer numbers and affinities are related by the same scale factor: $t_S^{(1)} = \xi t_S^{(2)}$ and $\sigma^{(1)} = \xi\sigma^{(2)}$. This implies that $t_S^{(1)}/t_S^{(2)} = \sigma^{(1)}/\sigma^{(2)}$ or $r = \sigma^{(1)}/t_S^{(1)} = \sigma^{(2)}/t_S^{(2)}$. The latter says that two systems behave approximately the same if the ratio r of their respective affinity to protomer number is the same, which yields another way of expressing the scaling observation as

$$\langle s_{\max}\rangle[t_S^{(1)}, r\, t_S^{(1)}] \approx \langle s_{\max}\rangle[t_S^{(2)}, r\, t_S^{(2)}]. \tag{1.89}$$

These relations are depicted in Figure 1.20.

1.7.12 Unequal ligand concentrations and ligand binding affinities

Polymerizing scaffold system

As in Figure 6 of the main text, Figure 1.21A evidences the σ-dependence of the initial slope in the discrete system and illustrates the effect of ligand imbalance: Once the scarcer ligand, here A, is mostly bound up and the number of scaffold protomers increases further, A-ligands must spread across an increasingly wider range of length classes, thereby reducing the likelihood of multiple occupancy on the same polymer. As a result, although the binding opportunities for the more abundant ligand, here B, increase (up to the overall prozone peak), B-particles bound to a particular polymer are less likely to encounter any As bound to it. The result is a slope reduction compared to a situation in which both ligands are present in equal numbers. A substantive difference between ligand binding constants causes not only a slope reduction prior to the prozone but has, in particular, the effect of delaying the prozone peak considerably beyond what one would expect based on particle numbers alone. It is worth noting that in the Wnt signaling cascade, ligand affinities——enzyme-scaffold, i.e. GSK3β–Axin, and substrate-scaffold, i.e. β-catenin–Axin—are regulated by the signaling process (Luo et al., 2007; Willert, Shibamoto, and Nusse, 1999).

In the continuum case, unlike the discrete case, the initial slope is independent of the polymerization constant σ until a level of protomer abundance is reached sufficient for making polymerization

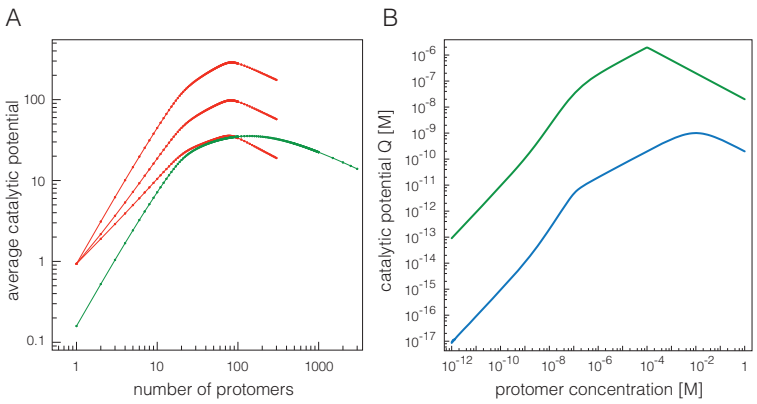


Figure 1.21: Effects in discrete and continuum polymerizing scaffold systems. A: The panel illustrates the effects of the polymerization constant σ, of ligand imbalance, and of unequal ligand affinities on discrete polymerization. Red, ligand imbalance: $t_A = 20$, $t_B = 80$, $\alpha = \beta = 0.9$ molecules^{-1}, $\sigma = 0.01$ (lower), $\sigma = 0.1$ (middle), $\sigma = 1$ (upper). Green, unequal ligand affinities: $t_A = t_B = 20$, $\alpha = 0.01$, $\beta = 1$ molecules^{-1}, $\sigma = 1$ molecules $^{-1}$. t_S on the abscissa. B: This panel illustrates the effects of ligand imbalance and of unequal ligand binding constants on continuum polymerization. Blue, unequal binding constants: $\alpha = 10^2$ M^{-1}, $\beta = 10^9$ M^{-1}, $t_A = t_B = 10^{-7}$ M, $\sigma = 10^8$ M^{-1}. Green, ligand imbalance: $t_A = 10^{-8}$ M, $t_B = 10^{-4}$ M, $\alpha = \beta = 10^7$ M^{-1}, $\sigma = 10^8$ M^{-1}.

effective, as discussed in section 1.7.5 (equation 1.52). The inflection point at which the slope changes from 1 to 2 (in a log-log plot) will shift accordingly. After that slope change, the responses to ligand imbalance and to differences between ligand binding constants are analogous to the discrete case, as seen in Figure 1.21B.

Neither ligand imbalance or differences in binding constants appear to affect the downward slope at large t_S in the continuum or the discrete case.

Multivalent scaffold system

The responses to ligand and affinity imbalances in a multivalent scaffold system follow similar lines as in the polymerizing case. When both ligand types are present with the same number of particles, the ligand with higher affinity experiences the prozone later, since the amount of scaffold-bound ligand is higher compared to the other type. This is seen in Figure 1.22B with the steepening of the downward slope associated with the stronger binding ligand. The situation with ligand imbalance is analogous. The ligand with higher abundance keeps binding while the scarcer ligand is undergoing its prozone; thus the subdued effect on catalytic potential, which, in the example of Figure 1.22C is mainly holding a constant level until the prozone for the more abundant ligand

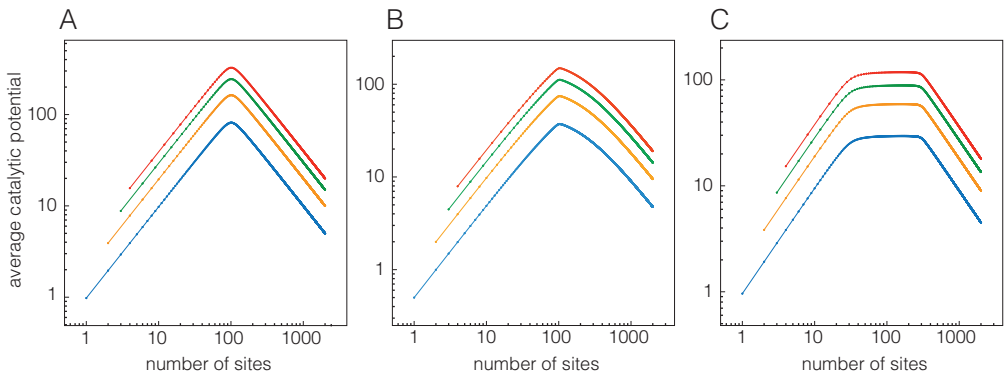


Figure 1.22: Catalytic potential of multivalent scaffolds (discrete case). A: $\langle Q_{\text{multi}} \rangle$, equation (1.80), when particle numbers and binding affinities are the same for both ligand types: A and B are 100 particles each, binding affinities are 0.9 molecules^{-1}. Valencies: 1 (blue), 2 (orange), 3 (green), 4 (red). The abscissa shows the total number of sites, but $\langle Q_{\text{multi}} \rangle$ is calculated for site increments that reflect the valency of each scaffold. B: Like panel A, but unequal ligand binding affinities: $\alpha = 0.01$ and $\beta = 9$ molecules^{-1}. C: Like panel A, but unequal numbers of ligand particles: $A = 30$ and $B = 300$, binding affinities for both are 0.9 molecules^{-1}. Colors indicate valencies as in panel A.

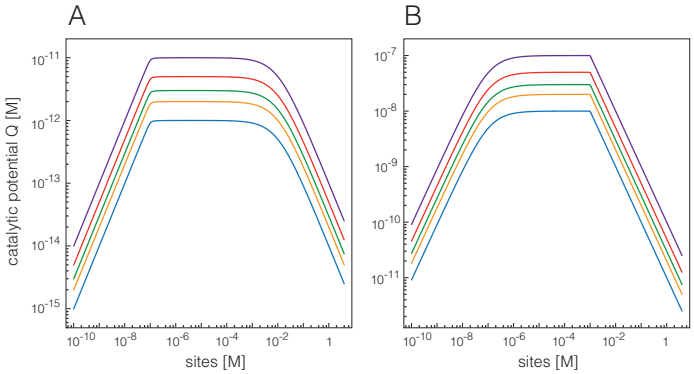


Figure 1.23: Catalytic potential of multivalent scaffolds (continuum case). A: The panel provides an example for the effect of unequal ligand binding affinity. $t_A = t_B = 10^{-7}$ M, $\alpha = 10^2$ M^{-1}, $\beta = 10^9$ M^{-1}, valencies: 1, 2, 3, 4. B: The panel illustrates the effect of ligand concentration imbalance. $t_A = 10^{-8}$ M, $t_B = 10^{-3}$ M, $\alpha = \beta = 10^7$ M^{-1}, valencies: 1, 2, 3, 4.

sets in. Although affinity and number imbalance mimic each other, the affinity imbalance exhibits a much less pronounced plateau around the prozone peak and consequently the drop-off is less sharp than in the case of number imbalance. Extremely high affinity differences would be required to generate a plateau similar to number imbalance. This is seen in the continuum case, shown in Figure 1.23A, where affinities differ by 7 orders of magnitude. The concentration imbalance in the continuum case yields a similar picture as in the discrete case (Figure 1.23B).

1.7.13 Stochastic simulations

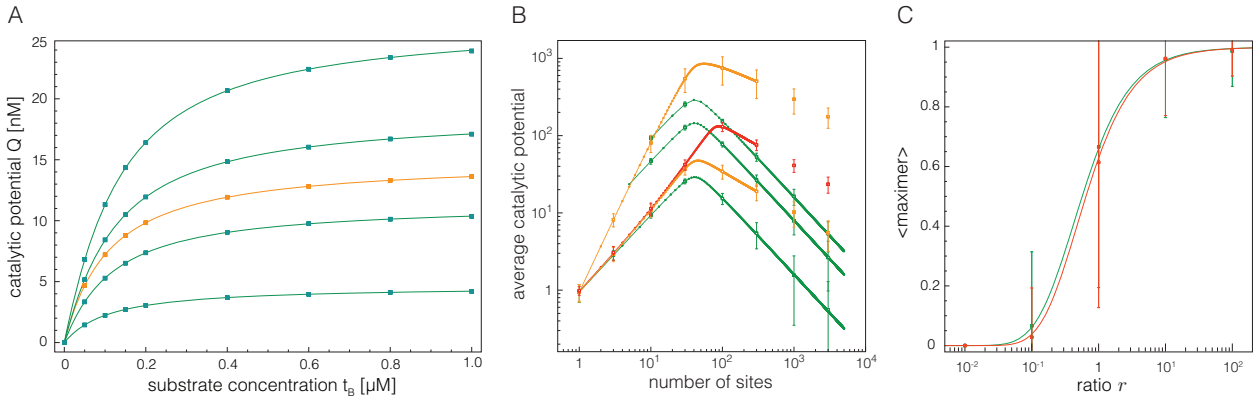


Figure 1.24: Stochastic simulations. For all stochastic simulations, we used a volume on the order of a human erithrocyte, $V = 10^{-12}$ L. All summary statistics were computed with 500 samples, each an independent and equilibrated state. A: The solid curves in this panel are identical to those in Figure 3A of the main text. Stochastic simulations were performed by converting deterministic affinities into stochastic affinities as described in the main text (section "The discrete case in equilibrium") and by converting concentrations into particle numbers at the given volume V. Averages of catalytic potnetial are indicated by filled squares. Green: polymerizing system at various protomer numbers, descending from top: 36120 molecules (60 nM), 27090 molecules (45 nM), 18060 molecules (30 nM), 9030 molecules (15 nM). Orange: reference Michaelian system with 60200 (100nM) enzymes. Because of the large numbers of particles, the standard deviation is smaller than the squares at the chosen scale. This panel is meant as a sanity check that simulations at large particle numbers indeed reproduce the continuum picture as we derived it analytically. B: The curves in this panel are identical to those in Figure 6A of the main text and refer to discrete scaffolding systems. Stochastic simulations were performed using the same parameters listed in that Figure. The squares mark the average catalytic potential, which coincides with the theoretical calculations; the error bars mark one standard deviation. In the polymerizing scaffold case, the simulation allowed us to extend the range of the rather time-consuming calculations using the analytical expression 1.77. Note the log-log scale of the axes distorting the error bars; for a linear-log scale see Figure 1.25. Green: multivalent scaffolds of valencies $n = 10$ (upper), $n = 5$ (middle), and $n = 1$ (lower). Orange: polymerizing scaffold system with polymerization affinities $\sigma = 10$ (upper) and $\sigma = 0.01$ (lower). Red: polymerizing scaffold system at the same affinity as the lower orange curve, but with twice the number of ligand particles. C: The curves are identical to those in Figure 1.20B. As in that Figure, r is the ratio of affinity to the number of protomers. Squares mark the average number of maximers and error bars mark one standard deviation. Green: system with 10 protomers. Red: system with 1000 protomers.

Our analysis of the discrete case focuses on average behavior. Analytic techniques for higher moments are beyond the scope of this contribution and will be presented elsewhere. In lieu

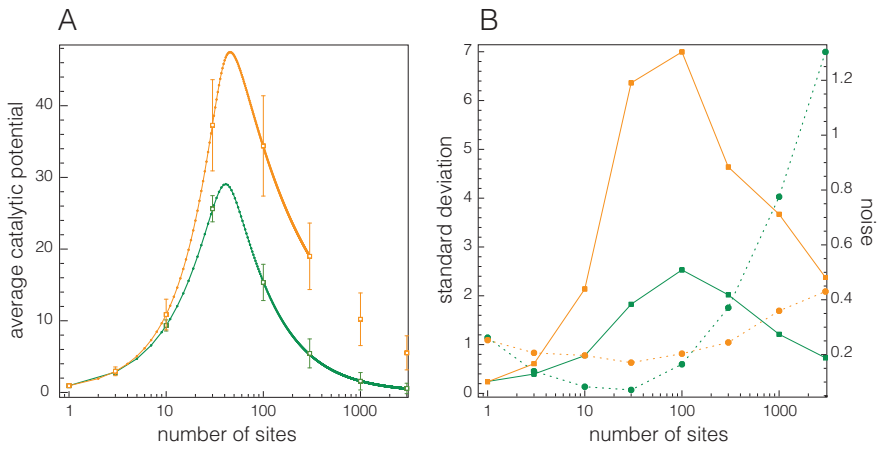


Figure 1.25: Variance and noise. A: This panel reproduces a subset of data from Figure 1.24B on a linear-log scale to enable a more direct visual interpretation of fluctuations. The green curve in this panel corresponds to the lowest green curve in Figure 1.24B. It belongs to a system of multivalent scaffolds with valency 1. The orange curve belongs to the polymerizing scaffold system and corresponds to the lowest orange curve in Figure 1.24B. Because the valency of individual scaffolds in both systems is 1, the number of sites on the abscissa corresponds to the number of scaffold agents, polymerizing or not. The main observation is that for the same average catalytic potential $\langle Q \rangle$ the standard deviation is larger after the prozone peak than prior to it. B: This panel recasts the information in panel A by directly displaying the standard deviation (solid curves). The dashed curves (right ordinate) depict the noise, i.e. the ratio of standard deviation to the mean. The main observation here is that the polymerizing system (orange) is significantly less noisy than the monovalent scaffold system (green).

of an analytic treatment, we performed several stochastic simulations using the Kappa platform (Boutillier, Feret, et al., 2018; Boutillier, Maasha, et al., 2018) and GNU Parallel (Tange, 2018). Figure 1.24 displays the essential observations in the context of Figures 3A and 6A of the main text and 1.20B of this Supplement.

Fluctuations in the binding of ligands translate into Q-fluctuations on the basis of how sites are partitioned into agents. There are three regimes, which we describe in the case of a monovalent scaffold system for simplicity (lowest green curve in Figure 1.24; green curve in Figure 1.25; and Figure 1.26): (i) At low scaffold numbers, prior to the prozone peak, most scaffolds are fully occupied by both ligands. Fluctuations cause transitions between system states with similar Q and variance is therefore low (see red distributions in Figure 1.26). (ii) Just past the prozone peak, many

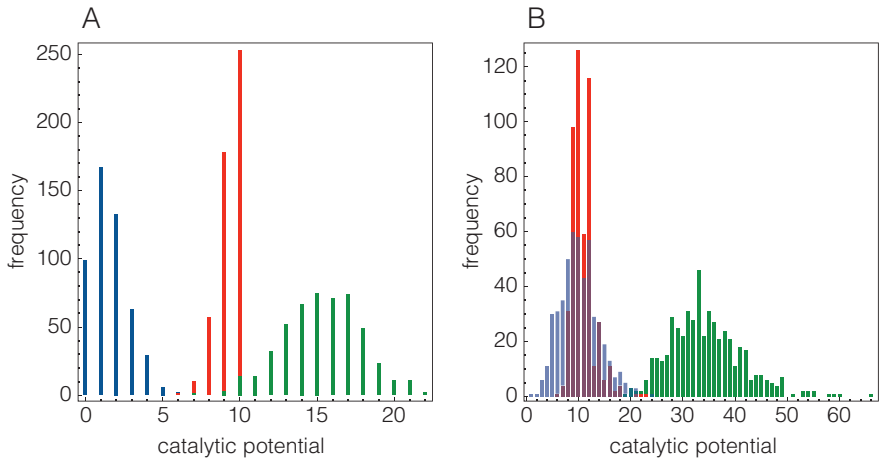


Figure 1.26: Distributions of catalytic potential. Panels A (monovalent scaffold system) and B (polymerizing scaffold system) depict the distribution of catalytic potential for a state sampled prior to the prozone peak (10 scaffold particles, red), just past the peak (100 particles, green) and well past the peak (1000 particles, blue). Other parameters as in Figure 6A of the main text.

scaffolds are still occupied by both ligands, but there is an increasing number of singly bound and some empty scaffolds. Unbinding from a fully occupied scaffold is statistically offset by re-binding to the pool of singly-bound scaffolds, which yields a net effect similar to situation (i). However, in addition, singly-bound scaffolds may also lose their ligand. This event is neutral in Q, but free ligands may re-bind an already singly-bound scaffold, thereby increasing Q. Likewise, dissociation from a fully occupied scaffold an re-association with an empty one will decrease Q. As a result of this expanded Q-range, the variance has increased compared to a situation with similar average Q prior to the prozone peak (see green distributions in Figure 1.26). (iii) Well past the prozone peak, a number of scaffolds are bound by one ligand and many have no ligands at all. Ligand binding fluctuations will mainly shift ligands from singly-bound scaffolds to empty scaffolds with no effect on Q. As a result, Q-variance is now decreasing again (see blue distributions in Figure 1.26).

BIBLIOGRAPHY

Anderson, David, Gheorghe Craciun, and Thomas Kurtz (2010). "Product-form stationary distributions for deficiency zero chemical reaction networks". In: *Bulletin of Mathematical Biology* 72.8, pp. 1947–1970. DOI: 10.1007/s11538-010-9517-4.

Andrews, George E. (1984). *The Theory of Partitions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press. DOI: 10.1017/CBO9780511608650.

Behrens, Jürgen et al. (1998). "Functional Interaction of an Axin Homolog, Conductin, with β-Catenin, APC, and GSK3β". In: *Science* 280.5363, pp. 596–599. DOI: 10.1126/science.280.5363.596.

Bergeron, François, Gilbert Labelle, and Pierre Leroux (1997). Trans. by Margaret Readdy. Encyclopedia of Mathematics and its Applications. Cambridge University Press. DOI: 10.1017/CBO9781107325913.

Bergeron-Sandoval, Louis-Philippe, Nozhat Safaee, and Stephen W. Michnick (2016). "Mechanisms and Consequences of Macromolecular Phase Separation". In: *Cell* 165.5, pp. 1067–1079. DOI: 10.1016/j.cell.2016.05.026.

Bhattacharyya, Roby P. et al. (2006). "Domains, Motifs, and Scaffolds: The Role of Modular Interactions in the Evolution and Wiring of Cell Signaling Circuits". In: *Annual Review of Biochemistry* 75.1, pp. 655–680. DOI: 10.1146/annurev.biochem.75.103004.142710.

Boutillier, Pierre, Jérôme Feret, et al. (2018). *The Kappa Language and Kappa Tools: A User Manual and Guide [v4]*. https://kappalanguage.org.

Boutillier, Pierre, Mutaamba Maasha, et al. (2018). "The Kappa platform for rule-based modeling". In: *Bioinformatics* 34.13, pp. i583–i592. DOI: 10.1093/bioinformatics/bty272.

Bray, Dennis and Steven Lay (1997). "Computer-based analysis of the binding steps in protein complex formation". In: *Proceedings of the National Academy of Sciences* 94.25, pp. 13493–13498. DOI: 10.1073/pnas.94.25.13493.

Deeds, Eric et al. (2012). "Combinatorial complexity and compositional drift in protein interaction networks." In: *PLOS ONE* 7.3, e32032. DOI: 10.1371/journal.pone.0032032.

Ferrell, James E. (2000). "What Do Scaffold Proteins Really Do?" In: *Science's STKE* 2000.52, pe1–pe1. DOI: 10.1126/stke.522000pe1.

Fiedler, Marc et al. (2011). "Dishevelled interacts with the DIX domain polymerization interface of Axin to interfere with its function in down-regulating β-catenin". In: *Proceedings of the National Academy of Sciences* 108.5, pp. 1937–1942. DOI: 10.1073/pnas.1017063108.

Flajolet, Philippe and Robert Sedgewick (2009). *Analytic Combinatorics*. Cambridge University Press. ISBN: 9781139477161.

Flory, Paul J. (1936). "Molecular size distribution in linear condensation polymers I". In: *Journal of the American Chemical Society* 58.10, pp. 1877–1885. DOI: 10.1021/ja01301a016.

Good, Matthew C., Jesse G. Zalatan, and Wendell A. Lim (2011). "Scaffold Proteins: Hubs for Controlling the Flow of Cellular Information". In: *Science* 332.6030, pp. 680–686. DOI: 10.1126/science.1198701.

Horn, Fritz and Roy Jackson (1972). "General mass action kinetics". In: *Archive for Rational Mechanics and Analysis* 47.2, pp. 81–116. DOI: 10.1007/BF00251225.

Ikeda, Satoshi et al. (1998). "Axin, a negative regulator of the Wnt signaling pathway, forms a complex with GSK-3β and β-catenin and promotes GSK-3β-dependent phosphorylation of β-catenin". In: *The EMBO Journal* 17.5, pp. 1371–1384. DOI: 10.1093/emboj/17.5.1371.

Levchenko, Andre, Jehoshua Bruck, and Paul W. Sternberg (2000). "Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties". In: *Proceedings of the National Academy of Sciences* 97.11, pp. 5818–5823. DOI: 10.1073/pnas.97.11.5818.

Li, Pilong et al. (2012). "Phase transitions in the assembly of multivalent signalling proteins". In: *Nature* 483.7389, pp. 336–340. DOI: 10.1038/nature10879.

Liu, Chunming et al. (2002). "Control of β-Catenin Phosphorylation/Degradation by a Dual-Kinase Mechanism". In: *Cell* 108.6, pp. 837–847. DOI: 10.1016/S0092-8674(02)00685-2.

Luo, Wen et al. (2007). "Protein phosphatase 1 regulates assembly and function of the β-catenin degradation complex". In: *The EMBO Journal* 26.6, pp. 1511–1521. DOI: 10.1038/sj.emboj.7601607.

Mayer, Bruce J., Michael L. Blinov, and Leslie M. Loew (2009). "Molecular machines or pleiomorphic ensembles: signaling complexes revisited". In: *Journal of Biology* 8.9, p. 81. DOI: 10.1186/jbiol185.

Onsager, Lars (1931). "Reciprocal Relations in Irreversible Processes. I." In: *Physical Review* 37 (4), pp. 405–426. DOI: 10.1103/PhysRev.37.405.

Reynolds, Peter J., H. Eugene Stanley, and William Klein (1977). "Ghost fields, pair connectedness, and scaling: exact results in one-dimensional percolation". In: *Journal of Physics A: Mathematical and General* 10.11, pp. L203–L209. DOI: 10.1088/0305-4470/10/11/007.

Smoluchowski, Marian V. (1916). "Drei Vortrage uber Diffusion, Brownsche Bewegung und Koagulation von Kolloidteilchen". In: *Zeitschrift fur Physik* 17, pp. 557–585.

Suderman, Ryan and Eric Deeds (2013). "Machines vs. Ensembles: Effective MAPK Signaling through Heterogeneous Sets of Protein Complexes". In: *PLOS Computational Biology* 9.10, pp. 1–11. DOI: 10.1371/journal.pcbi.1003278.

Tange, Ole (2018). *GNU Parallel 2018*. Ole Tange. DOI: 10.5281/zenodo.1146014.

Univalent Foundations Program, The (2013). *Homotopy Type Theory: Univalent Foundations of Mathematics*. Institute for Advanced Study: https://homotopytypetheory.org/book.

Van Kampen, Nicolaas Godfried (1992). *Stochastic processes in physics and chemistry*. Vol. 1. Elsevier.

Wegscheider, Rudolf (1902). "Über simultane Gleichgewichte und die Beziehungen zwischen Thermodynamik und Reaktionskinetik homogener Systeme:" in: *Zeitschrift für Physikalische Chemie* 39U.1, pp. 257–303. DOI: doi:10.1515/zpch-1902-3919.

Whittle, Peter (1986). *Systems in Stochastic Equilibrium*. Probability and Statistics. Chichester: Wiley.

Willert, Karl, Sayumi Shibamoto, and Roel Nusse (1999). "Wnt-induced dephosphorylation of Axin releases β-catenin from the Axin complex". In: *Genes & Development* 13.14, pp. 1768–1773. DOI: 10.1101/gad.13.14.1768.

Wolfram Research Inc. (2019). *Mathematica, Version 12.0*. Champaign, IL.

Yorgey, Brent (2014). "Combinatorial species and labelled structures". PhD thesis. University of Pennsylvania.

Chapter 2

UNIVERSAL APPROXIMATION OF DISCRETE DISTRIBUTIONS

PREFACE

Up to when I started my Ph.D. all I had ever studied in the way of stochastic chemical reaction networks (SCRNs) was their stationary distributions when *detailed balance* is satisfied. In such cases, the stationary distributions can be taken to describe a system in equilibrium and the mathematical expressions describing equilibria are as simple as they can get. In particular, they can be expressed as restrictions of multidimensional Poisson distributions, which are well understood. When formulated as generating functions, the equilibrium distributions can be written as exponentials of linear functions. In my contributions to the article in Chapter 1, I harnessed this simplicity in order to analyze the equilibria of systems of polymerization, which due to the complexity of their sets of species and their state spaces would be otherwise intractable.

When I joined the Winfree lab Erik gave me a project that challenged my understanding of SCRNs as it required that I look beyond detailed balance. The project was based on the following question: given an appropriate measure of descriptive complexity for SCRNs and probability distributions over an integer lattice, is it possible to show that any given distribution can be produced as the stationary distribution of some SCRN with complexity matching that of the distribution? Does it remain true when restricting to certain classes of SCRNs such as detailed-balanced, bimolecular, unit rate constants, etc.? A related question, which is the one I ended up focusing on and which gave rise to the content of this chapter is the following: given a probability distribution over a positive integer lattice, henceforth a *discrete distribution*, is it possible to find a SCRN with stationary distribution approximating the given distribution? I liked this question because it was mathematically precise and because I saw it as an opportunity to venture beyond detailed balance into the more general stationary distributions of SCRNs. I had many failed attempts at attacking this problem, but each of those failures was nevertheless of immense value for me because they broadened my perspective on SCRNs and helped me develop an original vision and style that I hope to further develop in the coming years. Needless to say, each of my failed ventures was successful in bringing me a great deal of joy.

The first venture that I embarked upon made me dream of a kind of algebra of SCRNs whose algebraic operations would translate neatly into corresponding operations for their stationary distributions. This way, I imagined, one would be able to translate a constructive approach for discrete distributions into some constructive image in the context of SCRNs and their algebraic operations. I still believe that such a theory could be developed rigorously and elegantly and I am fascinated by such possibility. My starting point was the following observation: for a given discrete distribution q the set of SCRNs with q as a stationary distribution form a vector space, let us refer to it as the q-space. This can be seen by observing that if A and B are the transition matrices of two SCRNs,

if $Aq = 0$ and $Bq = 0$, meaning q is a stationary distribution of both SCRNs, then we have that $(A + B)q = 0$. I imagined that finding a basis for the q-space would amount to finding a set of optimal SCRNs that generate q. Further reduction could be achieved if the q-space was an *algebra*[1], as would be the case if one could define a tensor operation for SCRNs. In that case a *subbasis* —a generating set for the basis with respect to multiplication— would play the role that the prime numbers play for the integers but in this case for the q-space. A related idea to computing the q-space of a distribution q is that of finding a space of SCRNs capable of producing distributions q parametrized by some vector of parameters x, let us call them $q(x)$-spaces. One example of this is the case where the distributions $q(x)$ are multidimensional Poisson distributions with parameters x. The $q(x)$-space of these parametrized distributions is known to be the space of detailed-balanced SCRNs with energies provided by the vector of logarithms $-\log x_i$ (Cappelletti and Wiuf, 2016). We give a simple proof of this result in Chapter 3 using generating functions.

Some time before joining the Winfree lab, I came across an early version of the book *Quantum Techniques for Stochastic Mechanics* by Baez and Biamonte, 2018. The book formulates SCRNs in terms of their probability generating functions in order to establish a number of correspondences between SCRN theory and quantum field theory. I do not understand quantum field theory so the comparison was not all that useful for me; however, I have since been inspired by the elegance that generating functions lend to SCRN theory. For example, the *chemical master equation* (CME) becomes a partial differential equation on the probability generating function. Since the generating functions of multidimensional Poisson distributions are exponentials of linear functions, their partial derivatives are particularly simple so the CME is easy to evaluate. One can use this in order to prove the above-mentioned result that SCRNs with stationary multidimensional Poisson distributions are generated by detailed-balanced SCRNs, or more generally, by complex-balanced SCRNs, as we do in Chapter 3. Given the often tractable interaction between derivatives and exponential functions one of the many threads, I wish to follow in the future is that of characterizing the structure of SCRNs that generate distributions $q(x)$ with generating functions given by exponentials of, say, quadratic functions, or other classes of functions beyond linear ones. In another one of my ventures, I sought to understand what stationary distributions are like in general. My question was, if the stationary distributions of complex-balanced SCRNs are multidimensional Poisson distributions, what is the form of stationary distributions for SCRNs in general? I thought that if I knew the general form of stationary distributions I could harness it to produce distributions at will. I began with simple SCRNs, for example with a finite and small reachable state space. In one occasion, I set out to solve one of those systems by hand, which amounts to solving a linear system of equations. Inspired by the graphical methods I had discovered when I worked with polymerization systems, I wrote down the linear system and left the entries of the transition matrix as formal variables. What

[1] In the sense of a vector space with a multiplication operation that is bilinear.

I obtained was that each of the terms in the solution corresponded to a spanning tree in the state space transition graph. The result is well known, but my having arrived at it independently was a reassuring sign that my ideas were at least catching up to history.

As it may transpire from the previous paragraphs, I have a penchant for closed-form expressions. I am not so much interested in numerical approximations to solutions as much as I am interested in the structure of solutions themselves. It is this proclivity that led me to the graphical methods, I have described in this preface and that of Chapter 1. Closed-form expression are, however, a tall order so my insistence on them was one of my main obstacles when it came to tackling the problem I described before regarding the ability of SCRNs to generate discrete distributions, henceforth the *generating distributions* problem. It turns out that generating distributions is better framed in the language of approximations. I was always better at, or at least I enjoyed more, algebra and discrete math than analysis[2]. For this problem it arrived the time when I realized that I had no option but to confront those dreaded "epsilons" and "deltas."

We can formulate the problem of generating distributions as follows. First, we say that a discrete distribution can be approximated by SCRNs if for each positive real number ε there is a SCRN, and an initial condition, that converges to a stationary distribution that is, at most, a distance ε away from q. The question is then whether all discrete distributions can be approximated by SCRNs in this way. This formulation requires that one define a metric for discrete distributions. Any choice of metric furbishes the set of discrete distributions with the structure of a topological space. The set of SCRNs with initial conditions can also be given topological structure. Each reaction can be associated with the space comprised of the positive real numbers and their usual topology as this is the set of values that its rate constant can take. A set of reactions can similarly be associated with the product topology of the space of rate constant values for each of its reactions. The set of SCRNs in this way becomes a topological space consisting of many disconnected components, one for each finite set of reactions. If we restrict our attention to some set of well-behaved SCRNs, such as those with well-defined stationary distributions, the function that maps a SCRN and its initial condition to their limit distribution is likely to be continuous. This is since, intuitively, a small change in rate constants will give rise to a small change in stationary distribution. The generating distributions problem can be therefore phrased topologically without any mention to the approximation parameter ε as follows: is the image of the space of SCRNs and initial conditions *dense*[3] in the space of discrete distributions? It is this question that we explore in the present chapter.

[2]Results such as Stone duality, Isbell duality, and Gelfand duality, which are forms of duality between algebra and geometry would suggest that such a preference is simply that of a choice of language (nLab authors, 2021).

[3]A subset A of a topological space X is said to be dense if every open set containing a point in X also contains a point in A. This means that any point of X can be approximated arbitrarily well with points of A.

In addition the chapter contains my contribution to a publication (Poole et al., 2017) that explores a related but more particular question: can SCRNs generate the scope of distributions that can be generated with Boltzmann machines (BMs)? BMs are early thermodynamically-inspired models of neural networks with the ability to perform inference (Hinton and Sejnowski, 1983). Mimicking the behavior of BMs renders SCRNs as kinds of inference machines and thus potentially seen as acting intelligently in some way. For this problem, we sought to devise SCRN constructions inspired by the formalism of BMs. My contribution to the article was in the form of a SCRN construction that generated exactly the distributions of BMs, while satisfying detailed balance. I arrived at the construction by following graphical reasoning, this time translating the structure of a BM graph into molecules and their corresponding energies.

As I have already alluded to, my multiple approaches to the generating distributions problem were not fruitful for the problem itself. The construction provided in the chapter for approximating discrete distribution was devised by my advisor Erik Winfree. My job was simply to perform proofs wherever they were needed. One contribution that is original to me is that of a SCRN that can approximate *point mass distributions*, i.e. discrete distributions where all the probability is concentrated in a single point. That construction is particularly simple and analytically tractable and it necessarily requires that a system violates detailed balance, as opposed to the above-mentioned construction. Furthermore, the construction is robust in the sense that it is independent of initial conditions. That was one of my first results regarding the generating distributions project. Those point-mass SCRN can be "combined" to generate any discrete distribution point by point. Although Erik and I had come up with the construction that would achieve that, I was never able to generate the required proof. For that we had to recruit the help of SCRN experts David Anderson and his then-postdoc Danielle Cappelletti. The proof itself was performed by Daniele and it comprises the bulk of the 15-page appendix to Cappelletti, Ortiz-Muñoz, et al., 2020. Although the proof established the desired proposition, that our combination scheme for point-mass SCRNs works, its style is much different from the approaches I had attempted. I wish to one day go back to this problem and try to find an approach that is more aesthetically pleasing to me and hopefully more insightful. Some of the techniques in Chapter 3 for computing the dynamics of SCRNs were developed with that desire in mind.

ABSTRACT

We show that the set of discrete distributions that arise as the marginals of limit distributions of stochastic chemical reaction networks (CRNs) is dense in the space of all distributions. We do so by providing a construction of a class of detailed-balanced CRNs that can produce arbitrary discrete distributions with finite support. Given that the set of distributions with finite support is already dense in the space of distributions, we conclude that stochastic CRNs are universal approximators of discrete distributions. In addition we provide a CRN construction that faithfully captures the equilibrium distributions of Boltzmann machines.

2.1 Introduction

Cells are the fundamental units of multicellular organisms, meaning that they are the building blocks out of which all living things are assembled. As such, the world of cells and their interaction reflects on the world of macroscopic organisms, like ourselves. For example, understanding of cellular processes has been fundamental for a greater understanding of human disease. If something goes wrong at the level of cells it is likely to manifest at the organism level. Similarly cellular processes reflect on the correct functioning of organism and therefore help in the understanding of the processes that constitute a healthy organism. A cell is a complex object on its own, and in an analogous way to the cellular world is reflected by the whole organism, the world of molecules and their interactions is reflected by the workings of a single cell and its interactions with other cells. Due to their fundamental role in biology, an understanding of cells as complex molecular machines is an important part of an understanding of biological phenomena.

Mathematizing real phenomena, meaning the development of mathematical theories capable of modeling and predicting measurable data about a real phenomenon, liberates us from the confines of reality and puts us in the relatively freer world of abstract mathematics. Mathematics, too, is a complex machinery. Like molecules and cells, mathematical propositions interact with one another through the rules of mathematics in order to create complex mathematical theories. The difference between biological and mathematical machines is that mathematical experiments can be carried out on a piece of paper, in our heads, or in a computer, which are generally vastly more accessible than actual biological systems. Metaphorically speaking, a mathematical theory of biology should then allow us to perform approximate biological experiments in a piece of paper. Of course, the extent to which these mathematical experiments say anything about the biological world depends on how closely a mathematical model reflects the biological phenomena it is intended to model.

We use *chemical reaction networks* (CRNs) as models for thinking about the behavior of a hypothetical cell conceived as a complex chemical system. One way to think about CRNs is as mathematical models of well-mixed chemical mixtures. A cell is hardly a well-mixed chemical reactor; it is highly compartmentalized and the medium is dynamical and complex. Yet, the well-mixing assumption is a good first approximation to the behavior of a cell. The complexity of well-mixed chemical systems stems from the idiosyncratic ways in which molecules interact with one another, its "interaction network". We use CRNs as general models for such networks of interactions of molecules in a well-mixed medium.

We understand well-mixing as the assumption that under some spatial and temporal scales, each fixed number of molecules in the system has essentially the same probability of all coming into physical proximity with one another. When a number of molecules in the system can react, they will do so if they come together in just the right way, yielding a small change in the chemical

composition of the system. Hence, due to the irrelevance of spatial information for well-mixed systems we can describe them in terms of the concentrations of each of the different types of molecules in the system. Furthermore, even if molecular interactions were to be deterministic, since we lose information by forgetting everything about the system but the concentrations, its behavior will appear *stochastic* in general. In the limit of large concentrations random fluctuations vanish in comparison and the behavior of the system becomes deterministic. Although some small molecules in a cell do exist in high concentrations, key large molecules like proteins and nucleic acids exist in low concentrations. Therefore, when we define the dynamics of CRNs we opt for a model with a discrete state space and stochastic dynamics, which we refer to simply as a *stochastic CRN*. To the extent that cells are approximated by such chemical systems, the study of stochastic CRNs is a worthwhile pursuit for mathematical biology.

The mathematics of CRNs is based on empirical observations of real chemical reactors rather than on first principles. As such the value of CRNs is not so much in prediction but as falsifiable models of biochemical reaction networks that, along with experiments, can help us uncover the vast complexity of chemical processes tht constitute a cell. Beyond biology, a different perspective on CRNs is that one can begin with a system that we can prove has a certain behavior, and then engineer a real chemical system that recapitulates that behavior. In this case, we can see CRNs as constituting a formal language or a model of computation, and the closer we can get a chemical system to match the assumptions of a model, the better they will carry out the computations that we can prove a CRN simulates.

One perspective on biological systems is that they resist dissipation by virtue of their encoding beliefs about their environment and minimizing an *informational free energy* functional of those beliefs, approximating a process of Bayesian inference (Conant and Ashby, 1970; Friston, 2012). Each of the internal states of the system can be seen as a hypothesis about the state of its environment, and a probability distribution over such states can be seen as prior beliefs about the state of the environment. Given this, it is reasonable to hypothesize that a system that can effectively encode a large number of probability distributions would be potentially capable of handling a wide range of different environments. The question then arises as to what kinds of distributions can be produced by chemical systems.

In this document, we will provide a general CRN construction that is capable of approximating any discrete probability distribution. The idea is that if we can engineer a physical system that reproduces the behavior of our construction, then such system would in principle be capable of representing any prior belief about its environment. We will only focus on the mechanism for generating distributions and not on mechanisms for updating prior beliefs or for decision-making. The construction we provide here was included in a publication, which also explores more general

frameworks for encoding distributions (Cappelletti, Ortiz-Muñoz, et al., 2020). In particular that paper places emphasis on universal approximation under a *robustness* condition, which dictates that the encoded distribution be independent of initial conditions.

The question of the range of distributions attainable as stationary distributions of SCRNs has been explored by other authors. In Cardelli, Kwiatkowska, and Laurenti, 2018, the authors provide a construction that is capable of exactly generating any discrete distribution with finite support. Their construction, however, does not satisfy detailed balance and it is in fact not even ergodic. Similarly, Fett, Bruck, and Riedel, 2007, provide a construction that, although it can also approximate arbitrary distributions, it is highly dependent on initial conditions and it is also not ergodic. In Plesa et al., 2018, they develop methods for controlling the noise of SCRNs while preserving their deterministic behavior. It is not clear, however, if their methods are capable of encoding arbitrary distributions as this is not their focus.

In addition to the general construction for approximating arbitrary distributions, we include a construction that is designed to exactly simulate Boltzmann machines (BMs). This construction appears, among other ones, in Poole et al., 2017. Although the previous results already imply that the equilibria of BMs can be produced with SCRNs, this construction is crafted to resemble a BMs in important ways, such as probabilistic inference.

Our constructions consist of abstract chemical systems, however, the ability of a chemical system to actually be able to produce arbitrary probability distributions will depend on that system's ability to produce rates and interactions that match our constructions. Still, our results imply that the more programmable a biochemical system is, either by humans or by evolution, the larger the class of steady-state distributions that are accessible to them. Our results for our chemical models of BMs constitute one of potentially many ways in which a cell may employ its stationary distribution as a generative, inferential mechanism. Although our constructions are arbitrary they demonstrate that programming stochastic behavior using interaction network topology is possible and suggest that the networks of real biological systems could have been selected by evolution partly for their ability to produce complex probability distributions.

2.2 Preliminaries

Our basic objects of study are chemical reaction networks (CRNs). A CRN is a specification of a set of species and a set of reactions, each of which has a corresponding stoichiometry, which itself consists of pairs of multisets of species (Gunawardena, 2003; Horn and Jackson, 1972). Furthermore, each reaction has a corresponding rate constant, which is a positive real number. We proceed to define CRNs formally below.

For convenience, in the following we will denote the set of natural numbers with $\mathbb{N}$, the set of real

numbers with $\mathbb{R}$, and the set of positive real numbers with $\mathbb{R}^+$. For sets A, B, we use the notations $A \to B$ and B^A for the set of functions from A to B interchangeably.

Definition 2.2.1. A *chemical reaction network* is a quadruple (S, R, n, k), where S is a set of *species*, R is a set of *reactions*, $n : 2 \times R \times S \to \mathbb{N}$ is a *stoichiometric function*, and $k : R \to \mathbb{R}^+$ is a *rate constant function*. We denote the collection of all CRNs with CRN.

Let $\Lambda = (S, R, n, k)$ be a CRN. We refer to multisets $x : S \to \mathbb{N}$ as *states* or *counts*. For each species $s \in S$, we denote the number of molecules of type s in x as x_s. We write states as formal sums of species as follows

$$x = \sum_{s \in S} x_s s.$$

The stoichiometric function n says, for each reaction, the number of molecules of each species that participate in that reaction. Hence, $n_{0,r,s}$ is the number of molecules of type s that participate in the reactants of r, while $n_{1,r,s}$ is the number of molecules of type s that are produced by reaction r. We can summarize the stoichiometry and rate constant of a reaction as follows

$$n_{0,r} \xrightarrow{k_r} n_{1,r} \equiv \sum_{s \in S} n_{0,r,s} s \xrightarrow{k_r} \sum_{s \in S} n_{1,r,s} s.$$

Let us consider a concrete example, which will hopefully make the above abstract definitions clearer. We consider a CRN $\Lambda = (S, R, n, k)$, where $S = \{A, B\}$, $R = \{0, 1\}$. We summarize the stoichiometric function as a pair of matrices $n_0, n_1 : R \times S \to \mathbb{N}$, where the rows correspond to reactions and columns to species, as follows

$$n_0 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \qquad n_1 = \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}.$$

We can also summarize the rate constant function as a matrix, in this case of dimension $R \times 1$ as $k = [1, 2]$. Using the reaction notation we defined above, we can summarize the stoichiometric and rate constant functions as follows

$$2A \underset{2}{\overset{1}{\rightleftharpoons}} B. \tag{2.1}$$

Here, the double arrows denote a pair of reactions that are the reverse of each other.

We are interested in how the chemical system specified by a CRN evolves over time after it is initialized in some initial condition. We will assume that the description of an instantaneous state is given by the discrete counts of molecules present in a mixture, which consequently gives rise to stochasticity in the time evolution of the system. We will consider here probability distributions over all multisets of species. A probability distribution can represent uncertainty with regards to the state of a system, or a statistical ensemble of systems. In either case, the probability distributions we

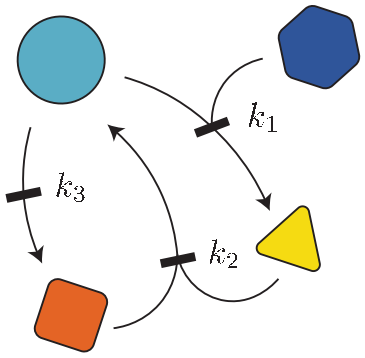


Figure 2.1: Graphical representation of a CRN. Each black rectangle represents a reaction with the tails of the arrows being connected to the reactants of the reaction and the head being connected to the products. Next to each reaction we show the corresponding rate constant.

consider are functions that assign to each state, i.e. multiset of species, a non-negative real number corresponding to the probability that one will find the system in such state, and where the sum of the probabilities of all states must add up to 1. We will ultimately be interested in approximating multiset distributions, not necessarily originating from a CRN, using the probability distributions obtained from CRNs. In order to be able to speak of approximation, we must have a means of comparing two distributions and saying how far apart they are from each other. We will do this by taking two distributions, then, for each state, consider the absolute value of the difference between their probabilities, and finally the distance between the distributions is given by the supremum of the set of values obtained this way. Given that the element-wise difference between two probability distributions is not necessarily a probability distribution, meaning it may have negative values or it may not add up to 1, we will consider a general notion of norm defined over all real-valued functions over multisets.

Definition 2.2.2. A *multiset distribution* is a pair (A, p), where A is a set, and $p : \mathbb{N}^A \to \mathbb{R}$ is a *probability function* satisfying $p_x \geq 0$ and $\sum p_x = 1$. We denote the collection of all distributions with Distr, and the set of all distributions with underlying set A as Distr_A. For each pair (A, f) of *multiset functions*, where A is a set and $f : \mathbb{N}^A \to \mathbb{R}$, its *norm* is given by

$$\|f\| = \max_{x \in \mathbb{N}^A} |f_x|.$$

For each set A, we will regard the set $\mathbb{N}^A \to \mathbb{R}$ of multiset functions as a vector space, meaning that we can add multiset functions and we can multiply them by scalars. Recall that multiset functions comprise not only probability distributions but all real-valued functions over the set $\mathbb{N}^A$ of multisets of A. Each multiset $x \in \mathbb{N}^A$ gives rise to a basis vector $[x] : \mathbb{N}^A \to \mathbb{R}$ defined by $[x]_y = \delta_{x,y}$. We

can write a function $f : \mathbb{N}^A \to \mathbb{R}$ in terms of basis vectors as follows

$$f = \sum_{x \in \mathbb{N}^A} f_x[x].$$

Notice that a scalar multiple of a basis vector has the general form $c[x]$, where $c \in \mathbb{R}$ is a scalar and $x \in \mathbb{N}^A$ is a multiset, which could give rise to confusion as it could be interpreted as a function c evaluated at x. In this chapter, however, a state x appearing inside square brackets universally denotes the basis vector corresponding to x.

2.3 Dynamics

Given some CRN, we are interested in how a probability distribution over its set of states evolves according to the reactions and rate constants of the CRN. The probability distribution must evolve according to a *master equation*. Below we will consider as an example the CRN in Equation 2.1 and derive its master equation in order to give an idea of what the general case should look like. We then proceed to give definitions of operators from which we can extract master equations of general CRNs. The style of presentation and the choice of notation here are suitable for use in a framework in which multiset distributions are presented as generating functions. We will develop that framework in Chapter 3.

In what follows, the word *operator* will simply mean a linear transformation $\mathscr{T} : \mathbb{R}^{\mathbb{N}^S} \to \mathbb{R}^{\mathbb{N}^S}$ from the space $\mathbb{R}^{\mathbb{N}^S}$ of multiset functions to itself. If a family $\{\mathscr{T}_i\}_{i \in I}$ of operators indexed by a set I commutes, we will use the product notation $\prod_{i \in I} \mathscr{T}_i$ to denote their composition. Finally, we denote the repeated application of an operator with $\mathscr{T}^n$, where n is the number of times that the operator is applied. Notice that the two notations are related by $\mathscr{T}^n = \prod_{i=1}^{n} \mathscr{T}$.

Let us consider the CRN in Equation 2.1. A general state of such system is a multiset $aA + bB$, for natural numbers a and b. The master equation, with p_x being the time-dependent probability of state x, in this case would be the following

$$\frac{dp_{aA+bB}}{dt} = (a+2)(a+1)p_{(a+2)A+(b-1)B} + 2(b+1)p_{(a-2)A+(b+1)B} - a(a-1)p_{aA+bB} - 2bp_{aA+bB}. \quad (2.2)$$

The first term corresponds to probability flowing into the state $aA+bB$ from state $(a+2)A+(b-1)B$ when applying the reaction $2A \to B$, which removes two A and creates one B, leading to the state $aA + bB$. Although multisets only account for counts but not identity of molecules, we assume that a multiset is a coarse-grained description of a state in which each molecule is identifiable. In that case, the number of ways of applying a reaction will take into consideration the distinguishability of molecules present in a state. Hence the first term has a factor of $(a+2)(a+1)$, which corresponds to the number of ways of applying reaction $2A \to B$ when the counts are given by $(a+2)A+(b+1)B$. The second term corresponds to probability flowing in from state $(a-2)A+(b+1)B$ when applying

$B \to 2A$. In this case, we multiply the probability of $(a-2)A + (b+1)B$ by $2(b+1)$, taking into consideration that the reaction has rate constant 2 and that there are $b+1$ ways of applying it. Notice that the first two terms are positive, indicating that they contribute to the increase of probability at state $aA + bB$. The last two terms, which are negative, correspond to probability flowing out of the state due to the action of both reactions. The third term is multiplied by $a(a-1)$ and the last term by $2b$, again, taking into consideration the rate constants and the number of ways that the reactions can be applied in the current state.

Notice that the master equation above gives the rate of change of probability at a given state as a linear combination of probabilities of other states. This means that it should be possible to find some linear transformation $\mathscr{A}$ such that, when it is applied to the time-dependent probability distribution vector p, it gives a vector of corresponding rates of change. In other words, we would like to find a linear transformation $\mathscr{A} : \mathbb{R}^{\mathbb{N}^S} \to \mathbb{R}^{\mathbb{N}^S}$ such that

$$\frac{dp}{dt} = \mathscr{A}p.$$

If we integrate the above equation and make use of the fundamental theorem of calculus, we can obtain an equivalent formulation of the master equation in integral form

$$p = p_0 + \int_0^t \mathscr{A}p dt,$$

where p_0 is the distribution at $t = 0$. The integral form reveals the recursive nature of the master equation, as it expresses the dynamics of a CRN in terms of itself. Moreover, it makes explicit the dependence on the initial distribution p_0. As it turns out, it is possible to define a time-dependent operator $\mathscr{E} = e^{\mathscr{A}t}$ such that when applied to an initial distribution p_0 it returns the time evolution of the distribution. In other words, $\mathscr{E}$ satisfies $p = \mathscr{E}p_0 = e^{\mathscr{A}t}p_0$. We will deal with this expression more in detail in Chapter 3.

Below we define the dynamics of a stochastic CRN in terms of creation and annihilation operators and then use them to extract the usual notion of propensities and of the chemical master equation. The idea behind an annihilation operator is that it removes one instance of some species in a state, while the creation operator inserts an instance of a species. Multiple application of either operator results in the removal or insertion of multiple instances of a species.

Definition 2.3.1. Let $\Lambda = (S, R, n, k)$ be a CRN. We define the *annihilation operator* ∂_s and the *creation operator* $\partial_s^\dagger$ for each $s \in S$ by

$$\partial_s[x] = x_s[x - s], \qquad \partial_s^\dagger[x] = [x + s].$$

Recall that $[x]$ denotes a basis vector for state x so that a linear operator can be defined by its action on basis vectors, which extends uniquely to all vectors by linearity. For each multiset $x \in \mathbb{N}^S$, we

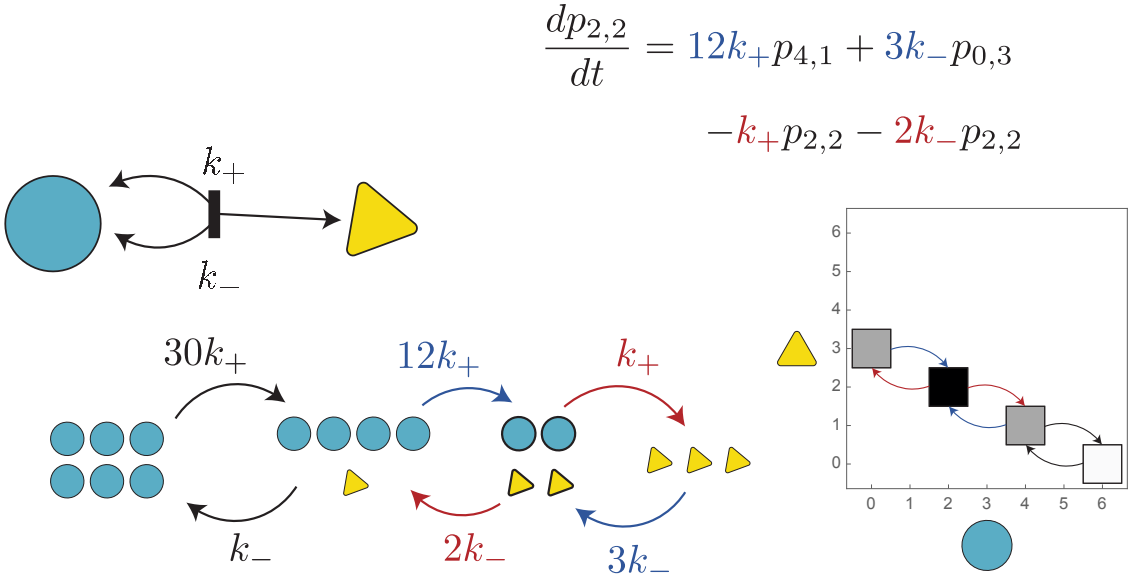


Figure 2.2: Example stochastic CRN. This system has 2 species, represented by a blue circle and a yellow triangle, and two reactions, given by the reaction diagram. On the top right is the master equation for the probability of being in state $(2, 2)$. At the bottom left corner is a diagram representing the state space of the system, with the incoming and outgoing transitions in blue and red. The plot on the right shows the probability distribution for the reachable states, with transition arrows as before.

define multiset annihilation and creation operators as follows

$$\partial^x = \prod_{s \in S} \partial_s^{x_s}, \qquad \partial^{\dagger x} = \prod_{s \in S} \partial_s^{\dagger x_s}.$$

We define the *infinitesimal stochastic operator* $\mathscr{A}_\Lambda : \mathbb{R}^{\mathbb{N}^S} \to \mathbb{R}^{\mathbb{N}^S}$ in terms of annihilation and creation operators as follows

$$\mathscr{A}_\Lambda = \sum_{r \in R} k_r (\partial^{\dagger n_{1,r}} - \partial^{\dagger n_{0,r}}) \partial^{n_{0,r}}.$$

Finally, we define the *evolution operator* $\mathscr{E}_\Lambda : \mathsf{Distr}_S \times (0, \infty) \to \mathsf{Distr}_S$ as follows

$$\mathscr{E}_\Lambda = 1 + \int_0^t \mathscr{A}_\Lambda \mathscr{E}_\Lambda dt,$$

where 1 denotes an identity operator. For each distribution $p \in \mathsf{Distr}_S$, we refer to $\mathscr{E}_\Lambda p$ as the *stochastic dynamics* of Λ with *initial distribution* p.

Having defined the infinitesimal stochastic operator $\mathscr{A}_\Lambda$, we will now confirm that it does indeed generalize the example we gave in Equation 2.2. As we mentioned previously, the annihilation operator ∂_s applied to the basis vector $[x]$ gives the number of ways of annihilating one instance of

s in state x. There are x_s different choices of s to annihilate, and after annihilating we are left with one s less, meaning we end up in state $x - s$. If it happens that $x_s = 0$, the result will simply be the zero vector, since there are no s available to annihilate. The creation operator $\partial_s^\dagger$ applied to x creates one instance of s within x. There is only one way of creating one s and the resulting state is $x + s$. Repeated application of ∂_s results in the annihilation of multiple instances of s, which can happen in a number of different ways. For example, consider the triple application

$$\partial_s^3[x] = x_s \partial_s^2[x - s] = x_s(x_s - 1)\partial_s[x - 2s] = x_s(x_s - 1)(x_s - 2)[x - 3s].$$

The coefficient of $[x - 3s]$ is a combinatorial factor indicating the number of ways of annihilating one s, then another one, and then one more. Equivalently, the combinatorial factor gives the number of ways of permuting 3 elements out of x_s elements. In general, we will have

$$\partial_s^m[x] = \frac{x_s!}{(x_s - n)!}[x - ms],$$

where m is the number of times the operator is applied. For multiset powers of ∂, we have

$$\partial^n[x] = \frac{x!}{(x - n)!}[x - n],$$

where n is a vector of natural numbers and where the multiset factorials are given by

$$x! = \prod_{s \in S} x_s!.$$

We define the *propensity* $\varrho_{r,x}$ of reaction r at state x as follows

$$\varrho_{r,x} = k_r \frac{x!}{(x - n_{0,r})!}.$$

The propensity gives the probability per unit time that a reaction will take place in a given state. We have the following identity involving annihilation operators and propensity

$$k_r \partial^{n_{0,r}}[x] = \varrho_{r,x}[x - n_{0,r}].$$

The above identity allows us to write the action of the infinitesimal stochastic operator as follows

$$\mathscr{A}_\Lambda[x] = \sum_{r \in R} \varrho_{r,x}([x - n_{0,r} + n_{1,r}] - [x]). \tag{2.3}$$

Let $p_0 \in \mathsf{Distr}_S$ be some initial condition and let $p = \mathscr{E}_\Lambda p_0$ be the corresponding stochastic dynamics. By taking a time derivative of the evolution operator, we obtain the following expression

$$\frac{dp}{dt} = \mathscr{A}_\Lambda p.$$

Using our expression in Equation 2.3 for the action of $\mathscr{A}_\Lambda$ in terms of propensities and rearranging, we obtain

$$\mathscr{A}_\Lambda p = \sum_{x\in\mathbb{N}^S} p_x \mathscr{A}_\Lambda[x] = \sum_{x\in\mathbb{N}^S} p_x \sum_{r\in R} \varrho_{r,x}([x - n_{0,r} + n_{1,r}] - [x])$$

$$= \sum_{x\in\mathbb{N}^S} [x] \sum_{r\in R} \varrho_{r,x+n_{0,r}-n_{1,r}} p_{x+n_{0,r}-n_{1,r}} - \varrho_{r,x} p_x.$$

Finally, we can use the above expression to write the rate of change of probability at a particular state

$$\frac{dp_x}{dt} = \sum_{r\in R} \varrho_{r,x+n_{0,r}-n_{1,r}} p_{x+n_{0,r}-n_{1,r}} - \varrho_{r,x} p_x.$$

The above expression is also known as the *chemical master equation*. We conclude that $\mathscr{A}_\Lambda$ behaves as we expect it to do.

For the CRN in Equation 2.1, the infinitesimal stochastic operator takes the form

$$\mathscr{A}_\Lambda = (\partial_B^\dagger - \partial_A^{\dagger 2})\partial_A^2 + 2(\partial_A^{\dagger 2} - \partial_B^\dagger)\partial_B.$$

The action of the operator is given by

$$\mathscr{A}_\Lambda[aA+bB] = a(a-1)[(a-2)A+(b+1)B] - a(a-1)[aA+bB] + 2b[(a+2)A+(b-1)B] - 2b[aA+bB].$$

The propensities at a generic state $aA + bB$ are given by

$$\varrho_{A\to B,aA+bB} = a(a-1), \qquad \varrho_{B\to A,aA+bB} = 2b.$$

Finally, the chemical master equation is given by

$$\frac{dp_{aA+bB}}{dt} = (\mathscr{A}_\Lambda p)_{aA+bB}$$

$$= (a+2)(a+1)p_{(a+2)A+(b-1)B} + 2(b+1)p_{(a-2)A+(b+1)B} - a(a-1)p_{aA+bB} - 2bp_{aA+bB},$$

as desired.

2.4 Reacting systems

We would like to model a scenario in which a hypothetical experimenter is able to prepare a chemical mixture with specified counts of molecules. As the experimenter allows the system to evolve, they are able to measure the counts of a set of species that are visible to them. Such a set may consist of all the species, or of a proper subset of species. By performing multiple trials with the same initial state, the experimenter is able to determine the stationary probability distribution that the system converges to. The definition below is intended to model such a scenario.

In order to model the long-term behavior of the system when given an initial state, we will consider the limit of the evolution operator when it is applied to the given initial state.

Definition 2.4.1. For each distribution (A, p) and subset $V \subseteq A$, we define the *marginal* $\mathscr{M}_V p \in \mathsf{Distr}_V$ of p onto V as follows

$$(\mathscr{M}_V p)_x = \sum_{y \in x + \mathbb{N}^{A-V}} p_y,$$

where $x + \mathbb{N}^{A-V} = \{x + y \mid y \in \mathbb{N}^{A-V}\}$. A *reacting system* is a triple (Λ, x, V), where $\Lambda = (S, R, n, k)$ is a CRN, $x : S \to \mathbb{N}$ is an *initial condition*, and $V \subseteq S$ is a set of *visible* species. We denote the collection of all reaction systems with RxnSys. For each reacting system (Λ, x, V), its *visible limit distribution* $\pi_{\Lambda,x,V} \in \mathsf{Distr}_V$ is given by

$$\upsilon_{\Lambda,x,V} = \mathscr{M}_V \lim_{t \to \infty} \mathscr{E}[x].$$

Let us consider the CRN in Equation 2.1. Suppose that we are given the initial condition $8A$. With this initial condition, we get that the set of reachable states and their connectivity are described by the following

$$8A \underset{2}{\overset{56}{\rightleftharpoons}} 6A + B \underset{4}{\overset{30}{\rightleftharpoons}} 4A + 2B \underset{6}{\overset{12}{\rightleftharpoons}} 2A + 3B \underset{8}{\overset{2}{\rightleftharpoons}} 4B,$$

where the numbers in the arrows correspond to propensities. Given that this system is ergodic, the limiting distribution coincides with the stationary distribution whose support is the set of reachable states. Notice that the set of reachable states is also the set of states $aA + bB$ with the conserved property that $a + 2B = 8$. We can therefore express the limit distribution as follows

$$\left(\lim_{t \to \infty} \mathscr{E}[8A]\right)_{aA+bB} = \begin{cases} 0, & a + 2b \neq 8 \\ \pi_{aA+bB}, & a + 2b = 8 \end{cases}$$

where the stationary distribution π satisfies

$$\mathscr{A}\pi = 0.$$

The above equation for the stationary distribution π is equivalent to the linear system

$$\mathscr{A}_\Lambda[8A] = -56\pi_{8A} + 2\pi_{6A+B} = 0$$

$$\mathscr{A}_\Lambda[6A + B] = 56\pi_{8A} - 32\pi_{6A+B} + 4\pi_{4A+2B} = 0$$

$$\mathscr{A}_\Lambda[4A + 2B] = 30\pi_{6A+B} - 16\pi_{4A+2B} + 6\pi_{2A+3B} = 0$$

$$\mathscr{A}_\Lambda[2A + 3B] = 12\pi_{4A+2B} - 8\pi_{2A+3B} + 8\pi_{4B} = 0$$

$$\mathscr{A}_\Lambda[4B] = 2\pi_{2A+3B} - 8\pi_{4B} = 0.$$

We can solve this system explicitly after which we obtain the solution

$$\pi = \frac{1}{764}[8A] + \frac{28}{764}[6A + B] + \frac{210}{764}[4A + 2B] + \frac{420}{764}[2A + 3B] + \frac{105}{764}[4B]. \tag{2.4}$$

Let $V = \{B\}$ be a set of visible species consisting only of the species B. The visible limit distribution of the resulting reacting system with initial condition $x = 8A$ is given by

$$\upsilon_{\Lambda,x,V} = \frac{1}{764}[0] + \frac{28}{764}[B] + \frac{210}{764}[2B] + \frac{420}{764}[3B] + \frac{105}{764}[4B].$$

What we have done is simply to project the distribution onto states involving only B by forgetting the counts of A. In this case, for each of the counts of B appearing in the reachable states, the counts of A are uniquely determined. It is possible, however, that in a given system with initial condition, there are multiple states with the same counts of a set of visible species. In such case, the projection will not be as simple as in our example, as we will need to add the probabilities of all states that share the same counts of visible species.

Before we present our construction for approximating arbitrary distributions, we will briefly review detailed-balanced CRNs.

Definition 2.4.2. A CRN (S, R, n, k) satisfies *detailed balance* if it is *reversible*, meaning that for each reaction $r \in R$, there uniquely exists $r_- \in R$ such that $r_{-\_} = r$, $n_{0,r} = n_{1,r_-}$ and $n_{1,r} = n_{0,r_-}$, and if there exists a function $\beta : S \longrightarrow \mathbb{R}^+$ satisfying

$$k_r \beta^{n_{0,r}} = k_{r_-} \beta^{n_{1,r}},$$

where $\beta^n = \prod_{s \in S} \beta_s^{n_s}$.

CRNs that satisfy detailed balance are well-behaved and, in particular, they have nice stationary distributions, namely, they are product-of-Poisson form[4] (Whittle, 1986). Furthermore, given an initial condition, they are ergodic in the set of reachable states, meaning that the limit distribution coincides with the unique stationary distribution whose support is the set of reachable states. We capture this result in the following lemma.

Lemma 2.4.1. *Let $\Lambda = (S, R, n, k)$ be a CRN satisfying detailed balance with some $\beta : S \longrightarrow \mathbb{R}^+$, and $x \in \mathbb{N}^S$ be some initial condition. Then, the limit distribution with initial condition x is given by*

$$\pi_\Lambda = \lim_{t \to \infty} \mathscr{E}_\Lambda[x] = M \sum_{x \twoheadrightarrow y} \frac{\beta^y}{y!}[y],$$

[4] More generally, complex-balanced CRNs also have product-of-Poisson form of stationary distributions (Anderson, Craciun, and Kurtz, 2010).

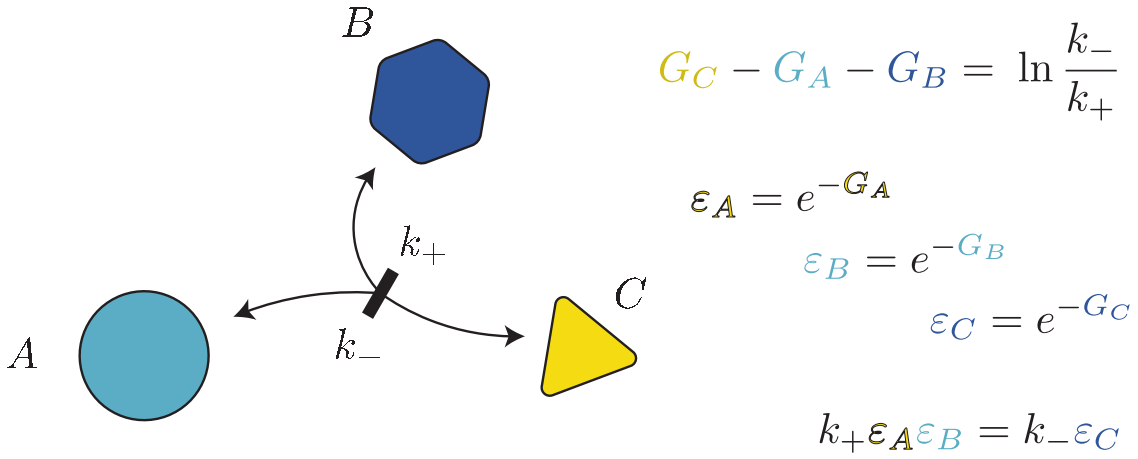


Figure 2.3: Illustration of detailed balance for CRNs. Here we denote the β in Definition 2.4.2 with ε to highlight its connection to energy. In particular, the negative of its natural log corresponds to chemical potential.

where $x \twoheadrightarrow y$ denotes that x can reach y through reaction transitions, and where the quantity M is a normalization constant

$$M = \left(\sum_{x \twoheadrightarrow y} \frac{\beta^y}{y!} \right)^{-1}.$$

Let us consider again the CRN in Equation 2.1. Notice that that CRN satisfies detailed balance since it is reversible, and for β defined by $\beta_A = \sqrt{2}$ and $\beta_B = 1$, we have

$$k_{2A \to B} \beta_A^2 = 2 = k_{B \to 2A} \beta_B.$$

According to the lemma, for initial condition $8A$, we must have

$$\pi = M \left(\frac{2^4}{8!0!}[8A] + \frac{2^3}{6!1!}[6A + B] + \frac{2^2}{4!2!}[4A + 2B] + \frac{2}{2!3!}[2A + 3B] + \frac{1}{0!4!}[4B] \right).$$

If we multiply and divide by $2^4/8!$, we can rewrite this solution as follows

$$\pi = \frac{2^4}{8!} M \left([8A] + 28[6A + B] + 210[4A + 2B] + 420[2A + 3B] + 105[4B] \right),$$

which clearly corresponds to the solution we had obtained before in Equation 2.4.

2.5 Boltzmann machines

In this section, we will show that CRNs can be used to reproduce the equilibrium distributions of Boltzmann machines.

Definition 2.5.1. A *Boltzmann machine* is a triple $B = (n, w, \theta)$, where $n \in \mathbb{N}$ is a natural number indicating the number of *nodes*, $w \in \mathbb{R}^{n \times n}$ is a *weight matrix* satisfying $w_{i,j} = w_{j,i}$ and $w_{i,i} = 0$, and

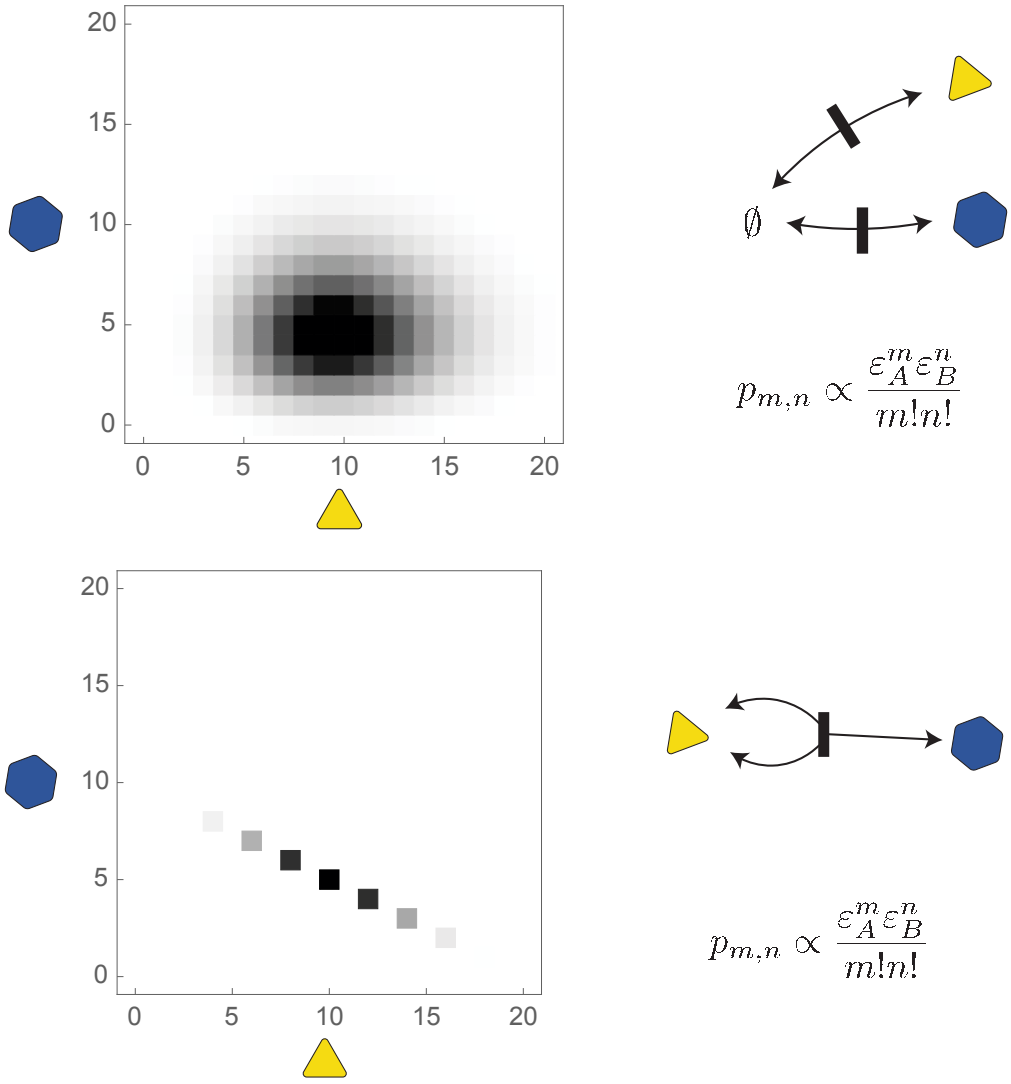


Figure 2.4: Illustration of Lemma 2.4.1. The stationary distributions of two detailed-balanced CRNs are shown. On the top, all states are reachable and the stationary distribution is a multivariate Poisson. On the bottom the reactions give rise to reachability classes and hence the stationary distribution is a truncated multivariate Poisson.

$\theta : \mathbb{R}^X$ is a *bias vector*. A *state* of a Boltzmann machine is a vector $x \in 2^n$ with binary entries. The *energy function* $\varepsilon^B : 2^n \to \mathbb{R}$ of a Boltzmann machine define is defined as follows

$$\varepsilon_x^B = -\sum_{i<j} x_i x_j w_{i,j} - \sum_{i=1}^{n} x_i \theta_i.$$

The *equilibrium distribution* $\pi^B : 2^n \to \mathbb{R}$ of a Boltzmann machine is defined as follows

$$\pi_x^B = \frac{e^{-\varepsilon_x^B}}{Z_B},$$

where $Z_B = \sum_{x \in 2^n} e^{-\varepsilon_x^B}$ is the *partition function*.

Definition 2.5.2. For each Boltzmann machine $B = (n, w, \theta)$, we define the reaction system $\alpha_{\mathsf{BCRN}}(B) = (\Lambda, x, X)$, where Λ is a CRN with species given by the sets $X = \{X_i^0\}_{i=1}^n \cup \{X_i^1\}_{i=1}^n$and $W = \{W_{i,j}^0\}_{i<j} \cup \{W_{i,j}^1\}_{i<j}$, reactions given by

$$X_i^0 + \sum_{j \neq i} X_j^{\sigma_j} + \sum_{\substack{j \neq i \\ \sigma_j = 1}} W_{i,j}^0 \underset{1}{\overset{k_{i,\sigma}}{\rightleftharpoons}} X_i^1 + \sum_{j \neq i} X_j^{\sigma_j} + \sum_{\substack{j \neq i \\ \sigma_j = 1}} W_{i,j}^{\sigma_j}, \text{ for } 1 \leq i \leq n \text{ and } \sigma \in 2^{n-1},$$

initial condition given by $x = \sum_{i=1}^n X_i^0 + \sum_{i<j} W_{i,j}^0$, and visible species given by the set X. We assume that $W_{i,j}^s = W_{j,i}^s$ for $s \in 0, 1$ (See Figure 2.5 for an illustration of the reactions.). The rate constants $k_{i,\sigma}$ are given by

$$\ln k_{i,\sigma} = \theta_i + \sum_{i \neq j} \sigma_j w_{i,j}.$$

Notice that the above CRN satisfies detailed balance with

$$\beta_{X_i^0} = 1, \quad \beta_{X_i^1} = e^{\theta_i}, \quad \beta_{W_{i,j}^0} = 1, \quad \beta_{W_{i,j}^1} = e^{w_{i,j}}.$$

Indeed, we have that

$$k_{i,\sigma} \beta_{X_i^0} \prod_{j \neq i} \beta_{X_j^{\sigma_j}} \beta_{W_{i,j}^0} = \exp\left(\theta_i + \sum_{j \neq i} \sigma_j (\theta_j + w_{i,j})\right) = \beta_{X_i^1} \prod_{j \neq i} \beta_{X_j^{\sigma_j}} \beta_{W_{i,j}^{\sigma_j}}.$$

Theorem 2.5.1. *For each Boltzmann machine $B = (x, n, \theta)$ we have that the visible limit distribution $\upsilon_{\alpha(x,n,\theta)} = \pi^B$.*

2.6 Finite-support distributions

In this section we will show that, in principle, for any given distribution, an experimenter can always find a reacting system whose visible limit distribution is the desired distribution. We will begin by showing that each distribution with finite support is the visible limit distribution of some reacting

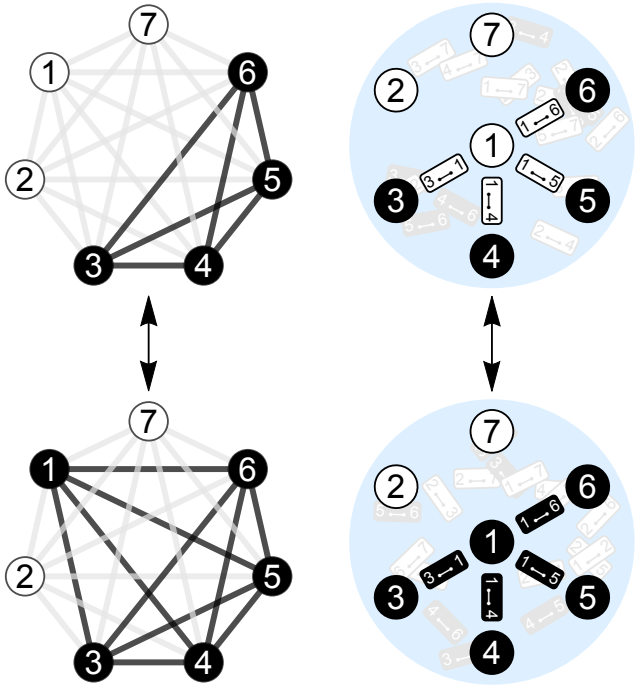


Figure 2.5: On the left is shown a Boltzmann machine transition. Nodes 3, 4, 5, and 6 are initially on as well as their joining edges. Subsequently, node 1 is turned on and the additional joining edges are turned on as well. On the right is shown the corresponding reaction in the Boltzmann CRN. The circles correspond to the species X_i either on or off, and the rectangles correspond to the species $W_{i,j}$, also on or off. The grayed-out species correspond to the species that do not participate in the reaction but that are present in the mixture.

system. In more detail, each distribution is mapped to a reacting system whose underlying CRN satisfies detailed balance (defined below), and whose visible species coincide with the underlying set of the target distribution. Then, we will show that distributions with finite support get arbitrarily close to any distribution, so that in fact every distribution can be approximated by the construction we provide below.

We will now present our construction for approximating distributions.

Definition 2.6.1. For each distribution $(V, p) \in \mathsf{Distr}$ with finite support and with $x \in \mathsf{supp}p$ a state in the support, we define the reacting system $\alpha(V, p, x) = (\Lambda, x + C_x, V) \in \mathsf{RxnSys}$, where $\Lambda = (S, R, n, k)$ is given by $S = V \cup \{C_y\}_{y \in \mathsf{supp}p}$; the reactions, stoichiometry, and rate constants given by

$$C_y + y \underset{z!p_z}{\overset{y!p_y}{\rightleftharpoons}} C_z + z,$$

for $y, z \in \mathsf{supp}p$ and $y \neq z$; and with initial condition $x + C_x$.

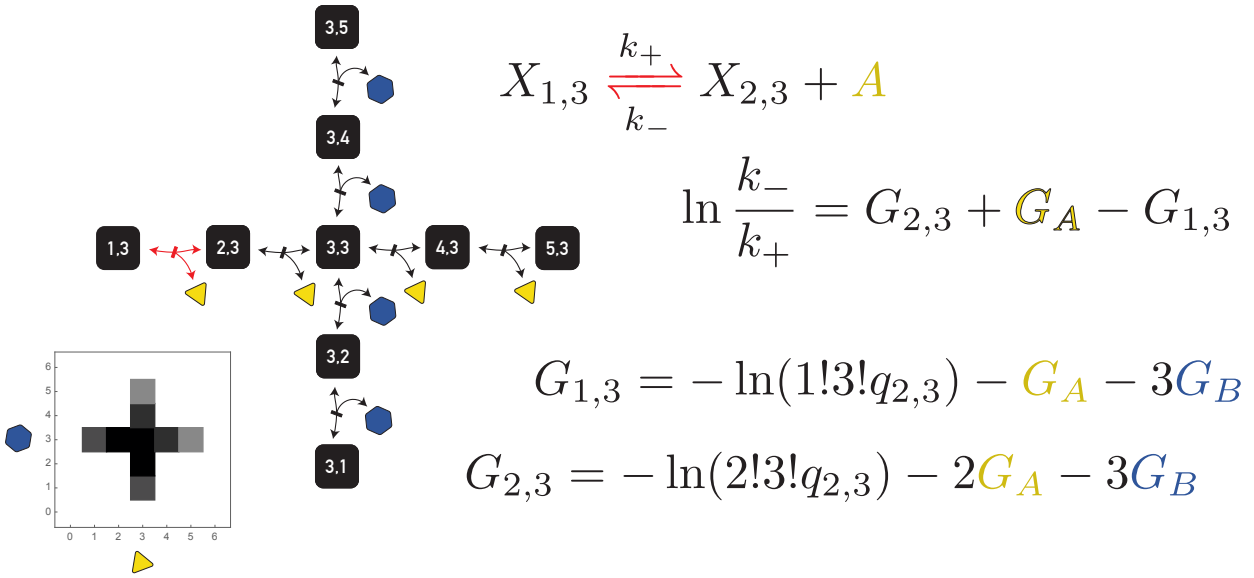


Figure 2.6: Illustration of a variant in the construction in Definition 2.6.1. The construction illustrated here follows the same principle as that of Definition 2.6.1, which is that each point in the support is associated with a hidden species, shown in black, and a number of reactions that interconvert between the hidden species, while updating the counts of the visible species appropriately. The rate constants are set up so that the desired probabilities are obtained in steady state.

The idea behind the above construction is the following. Each of the states y in the support of p gets a species C_y. By initializing the system at state $C_x + x$, the reactions guarantee that there are only ever one C_y species present. Moreover, if the species C_y is present, the counts of V species will be precisely y. The rate constants are chosen so that the CRN satisfies detailed balance with

$$\beta_s = 1, \qquad \beta_{C_y} = y! p_y,$$

for $s \in V$ and $y \in \mathsf{supp}p$. Indeed, we have that

$$k_{C_y+y \to C_z+z} \beta^{C_y+y} = (y! p_y)(z! p_z) = k_{C_z+z \to C_x+x} \beta^{C_z+z}.$$

Now we show that the visible limit distribution of the reacting system defined in the construction is the desired distribution.

Lemma 2.6.1. *For each distribution* $(V, p) \in \mathsf{Distr}$ *with finite support and with* $x \in \mathsf{supp}p$ *an element of the support, we have that* $\upsilon_{\alpha(V,p,x)} = p$.

Proof. Let (V, p) be a distribution with finite support with $x \in \mathsf{supp}p$, and $\alpha(V, p, x) = (\Lambda, C_x + x, V)$. We know that Λ satisfies detailed balance and that the set of reachable states is $\{C_y + y\}_{y \in \mathsf{supp}p}$. Therefore, we have that the limit distribution of Λ with initial condition $[C_x + x]$ satisfies

$$\lim_{t \to \infty} \mathscr{E}_\Lambda[C_x + x] = \sum_{y \in \mathsf{supp}p} \frac{\beta^{C_y+y}}{(C_y + y)!}[C_y + y] = \sum_{y \in \mathsf{supp}p} p_y [C_y + y].$$

Finally, by taking the marginal onto V, we obtain

$$\upsilon_{\alpha(V,p,x)} = \mathscr{M}_V \lim_{t\to\infty} \mathscr{E}_\Lambda[C_x + x] = \sum_{y\in \mathsf{supp}p} p_y[y] = p.$$

□

Let us consider the distribution in Figure 2.7, which we can think of as a distribution $p \in \mathsf{Distr}_{\{A,B\}}$ over multisets of two species A and B. This distribution has finite support, which means that by the Lemma above, we can generate it with some reacting system. The reacting system will have set of visible species $V = \{A, B\}$. Furthermore, for each non-zero pixel x in the "mushroom," we will have a corresponding species C_x. We will not write the reactions since in this case they are of the order of 16^2, but there will be a pair of reversible reactions

$$C_x + x_A A + x_B B \underset{x_A!x_B!p_x}{\overset{y_A!y_B!p_y}{\rightleftharpoons}} C_y + y_A A + y_B B$$

for each non-zero pixels x, y in the image. The initial condition can be $C_x + x$ for any non-zero pixel x.

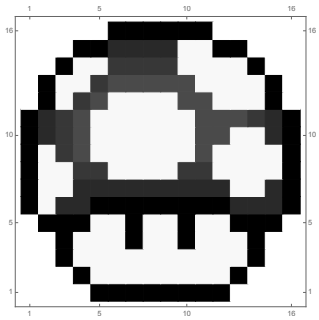


Figure 2.7: A distribution over two-dimensional state space.

2.7 Universality

Having showed that our construction is able to produce all distributions with finite support, it only remains to show that an arbitrary distribution can be approximated with finite-support distributions. To be more precise, we want to establish that for any distribution and positive real number there exists a distribution with finite support such that the norm of the difference of the distributions is less than the given positive real number.

Theorem 2.7.1. *For each distribution* $(V, p) \in \mathsf{Distr}$ *and* $\varepsilon > 0$*, there exists a reacting system* $(\Lambda, x, V) \in \mathsf{RxnSys}$ *such that*

$$\|\upsilon_{\Lambda,x,V} - p\| < \varepsilon.$$

Proof. Let $(V, p) \in \mathsf{Distr}$ and $\varepsilon > 0$. Consider a bijection[5] $x : \mathbb{N} \to \mathbb{N}^V$ with the property that $p_{x_i} \leq p_{x_j}$ whenever $j \leq i$. Let $n \in \mathbb{N}$ be a natural number with the property that $\sum_{i=0}^{n-1} p_{x_i} > 1 - \varepsilon$. Let us define the distribution (V, q) given by

$$q_y = \begin{cases} p_{x_i}, & y = x_i, \quad 0 \leq i \leq n-1 \\ \sum_{i=0}^{n} p_{x_i}, & y = x_n \\ 0, & y = x_i, \quad i \geq n+1 \end{cases}.$$

We know from Lemma 2.6.1 that $\upsilon_{\alpha(V,q,x)} = q$ for any $x = x_i$, where $i \leq n$. Furthermore, we have that

$$\|p - q\| = |p_n - q_n| = \sum_{i=n+1}^{\infty} p_{x_i} < \varepsilon.$$

□

2.8 Discussion

We have established that any arbitrary distribution can be approximated with the visible limit distribution of some reacting system. We say that the set of CRNs is *universally approximating*. In fact, we have proved a stronger result, namely, that the set of detailed-balanced CRNs is universally approximating. In Cappelletti, Ortiz-Muñoz, et al., 2020, we show in addition that universal approximation is also possible if we require that the stationary distribution be independent of initial conditions, a property we refer to as *robustness*. When the target distribution is a point mass distribution, the detailed-balanced construction in this chapter yields a trivial CRN with no reactions and with initial condition simply being centered at the desired point. The robust constructions, however, are more meaningful since we must be able to approximate a point mass distribution starting at any state.

Some results exist that establish limits to the degree to which fluctuations in the quantity of a molecule can be reduced (Lestas, Vinnicombe, and Paulsson, 2010) and therefore appear to be in conflict with our robustness results mentioned above, which allow us to approximate point mass distributions, and thus reduce noise, arbitrarily. The claims of Lestas, Vinnicombe, and Paulsson, 2010, however, can apply only to a restricted class of reaction networks as, for example, the system given by[6]

$$(n+1)X \to nX, \qquad \emptyset \to X, \tag{2.5}$$

is capable of reducing noise arbitrarily by appropriately modulating the rate constants. Other unpublished constructions that arose when working on our article can also achieve arbitrary noise

[5] Such a bijection is guaranteed to exist since V is finite and thus $\mathbb{N}^V$ has the same cardinality as $\mathbb{N}$.

[6] This simple system was one of my original contributions to Cappelletti, Ortiz-Muñoz, et al., 2020.

reduction without the need for high molecularity reactions as they incorporate only reactions that are at most bimolecular. Three assumptions in Lestas, Vinnicombe, and Paulsson, 2010, are that the regulated molecule X decays linearly, that X regulates the creation of a molecule Y, and that some arbitrary system depending on Y can create X. These conditions are encoded by the following schema

$$X \to \emptyset, \qquad u(Y) \to u(Y) + X, \qquad f(X) \to f(X) + Y,$$

where $f(Y)$ represents some general catalytic action of Y for producing X and $u(Y)$ is some "demon" that can use the whole history of Y to control the production of X. The explicit construction in Equation 2.5 as well as the unpublished ones mentioned above each violate at least one of those assumptions.

The results in this chapter as well of those in Cappelletti, Ortiz-Muñoz, et al., 2020, that do not appear here are relevant in a biological context. Cells and biological systems in general are believed to resist decay with the aid of an internal model of their surrounding (Friston, 2012). Such a model appears as a probability distribution representing beliefs about the state of the environment, which by following a process similar to Bayesian inference a system can make predictions about its surroundings. It is therefore of interest to understand which kinds of beliefs, in the form of probability distributions, can be held by a chemical system such as a cell. Our results indicate that, in principle, all internal models can be produced by stochastic chemical systems. Our work focuses on the generation of distributions but has the drawback that those distributions are fixed. More work would be necessary to incorporate ways of updating the target distributions according to measurements of the environment. Preliminary forms of this appear in Poole et al., 2017, where by "clamping" the values of certain species the system can perform probabilistic inference.

It is important to note that our approach for representing probability distributions with chemical systems is one of many potential methods that can be implemented. In particular, we propose to represent distributions by means of stationary distributions of stochastic chemical systems; however, it is also possible to, for example, represent them with a deterministic CRN in which the concentrations of species are proportional to concentrations. In Baez and Pollard, 2016, a framework is proposed that places deterministic CRNs in the context of the replicator equation used in evolutionary game theory and relates the concept of free energy to that of fitness. Other forms of representing distributions with CRNs, deterministic or stochastic, are possible, and more work in that direction could bring insight into the way chemistry relates to biology.

BIBLIOGRAPHY

Anderson, David, Gheorghe Craciun, and Thomas Kurtz (2010). "Product-form stationary distributions for deficiency zero chemical reaction networks". In: *Bulletin of Mathematical Biology* 72.8, pp. 1947–1970. DOI: 10.1007/s11538-010-9517-4.

Baez, John and Jacob Biamonte (2018). *Quantum Techniques in Stochastic Mechanics*. World Scientific. DOI: 10.1142/10623.

Baez, John and Blake S. Pollard (2016). "Relative Entropy in Biological Systems". In: *Entropy* 18.2. DOI: 10.3390/e18020046.

Cappelletti, Daniele, Andrés Ortiz-Muñoz, et al. (2020). "Stochastic chemical reaction networks for robustly approximating arbitrary probability distributions". In: *Theoretical Computer Science* 801, pp. 64–95. DOI: 10.1016/j.tcs.2019.08.013.

Cappelletti, Daniele and Carsten Wiuf (2016). "Product-form poisson-like distributions and complex balanced reaction systems". In: *SIAM Journal on Applied Mathematics* 76.1, pp. 411–432. DOI: 10.1137/15M1029916.

Cardelli, Luca, Marta Kwiatkowska, and Luca Laurenti (2018). "Programming discrete distributions with chemical reaction networks". In: *Natural Computing* 17.1, pp. 131–145. DOI: 10.1007/s11047-017-9667-5.

Conant, Roger C. and W. Ross Ashby (1970). "Every good regulator of a system must be a model of that system". In: *International Journal of Systems Science* 1.2, pp. 89–97. DOI: 10.1080/00207727008920220.

Fett, Brian, Jehoshua Bruck, and Marc D. Riedel (2007). "Synthesizing Stochasticity in Biochemical Systems". In: *2007 44th ACM/IEEE Design Automation Conference*, pp. 640–645.

Friston, Karl (2012). "A Free Energy Principle for Biological Systems". In: *Entropy* 14.11, pp. 2100–2121. ISSN: 1099-4300. DOI: 10.3390/e14112100.

Gunawardena, Jeremy (2003). *Chemical reaction network theory for in-silico biologists*. http://vcp.med.harvard.edu/papers/crnt.pdf.

Hinton, Geoffrey E. and Terrence J. Sejnowski (1983). "Optimal perceptual inference". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. Vol. 448.

Horn, Fritz and Roy Jackson (1972). "General mass action kinetics". In: *Archive for Rational Mechanics and Analysis* 47.2, pp. 81–116. DOI: 10.1007/BF00251225.

Lestas, Ioannis, Glenn Vinnicombe, and Johan Paulsson (2010). "Fundamental limits on the suppression of molecular fluctuations". In: *Nature* 467.7312, pp. 174–178. DOI: 10.1038/nature09333.

nLab authors (2021). *duality between algebra and geometry*. http://ncatlab.org/nlab/show/duality+between+algebra+and+geometry. Revision 7.

Plesa, Tomislav et al. (2018). “Noise control for molecular computing”. In: *Journal of The Royal Society Interface* 15.144, p. 20180199. DOI: 10.1098/rsif.2018.0199.

Poole, William et al. (2017). “Chemical Boltzmann Machines”. In: *DNA Computing and Molecular Programming*. Ed. by Robert Brijder and Lulu Qian. Cham: Springer International Publishing, pp. 210–231. DOI: 10.1007/978-3-319-66799-7_14.

Whittle, Peter (1986). *Systems in Stochastic Equilibrium*. Probability and Statistics. Chichester: Wiley.

Chapter 3

FORMAL SEMANTICS FOR STOCHASTIC CHEMICAL REACTION NETWORKS

PREFACE

This chapter is a shadow of the thesis I had originally conceived. I decided that I would leave my original vision for my postdoctoral work, where I will be able to develop it adequately and thoughtfully. What I present here is the starting point of a generalization that will result in a combinatorial/topological framework for formal power series and stochastic chemical reaction networks in particular. I believe such a framework could provide a more adequate mathematical foundation for biology that is at the same time inherently computational.

The idea behind a lot of this material is that of never evaluating variables —treating variables as *formal* rather than *numerical*. Doing so results in a picture in which functions acquire a combinatorial interpretation as ways of organizing formal variables into discrete structures. For example, if one restrains from using commutativity, the expression $(z+1)^3$ can be expanded as

$$(z+1)^3 = z\cdot z\cdot z + z\cdot z\cdot 1 + z\cdot 1\cdot z + 1\cdot z\cdot z + z\cdot 1\cdot 1 + 1\cdot z\cdot 1 + 1\cdot 1\cdot z + 1\cdot 1\cdot 1.$$

Written this way, this expression reveals that, for example, the terms containing two z's correspond to ways of choosing two positions out of three, which we know from the binomial formula that their number is given by a binomial coefficient. Working formally often reveals combinatorial patterns behind algebraic expressions and can be helpful in reasoning with them more clearly (see Figure 3.1).

Using generating functions in my work has led me on a research trajectory that I am likely to follow for many years ahead. It began when I worked on polymers, the subject of Chapter 1, and I realized that manipulating certain functions mathematically paralleled structural manipulations of polymer-related structures. When I worked on the universality of SCRNs, the subject of Chapter 2, I discovered, among other things, that general stationary distributions can be expressed in terms of spanning trees, which are combinatorial objects[1]. This observation makes me wonder about whether there is a general way of interpreting the objects of SCRN theory combinatorially and whether doing so can be fruitful for mathematical biology, which is rife with combinatorial structures.

In an attempt to gain a better understanding of generating functions and their usefulness for combinatorics, I came across the theory *combinatorial species* (Bergeron, Labelle, and Leroux, 1997; Joyal, 1981). Combinatorial species make use of category theory in order to give a clear picture of the combinatorial interpretation of formal power series. Despite its elegance and ability

[1]This result is well-known and can be seen as a version of Kirchoff's matrix tree theorem (Chaiken and Kleitman, 1978). In the context of continuous time Markov-chains, of which SCRNs are an example, it is known as the Markov chain tree theorem (Leighton and Rivest, 1983).

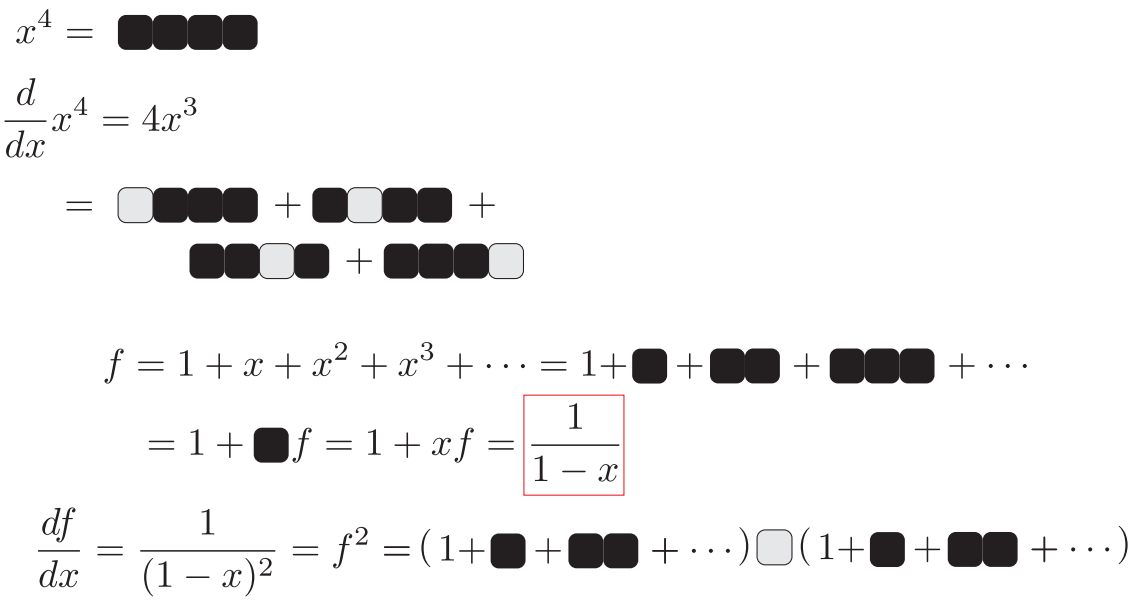


Figure 3.1: Illustration of the combinatorics of generating functions. One way to see generating functions is as each monomial representing a structure. In particular, we can interpret x^n as a unary string of length n. In this example we see how we can exploit algebraic tricks in order to obtain combinatorial insights.

to explain many aspects of combinatorial generating functions, combinatorial species theory falls short of being able to encompass all formal power series. In particular, the only power series that can arise from combinatorial species are of the form

$$f(z) = \sum_{n=0}^{\infty} \frac{a_n}{n!} z^n, \tag{3.1}$$

where a_n is the cardinality of a finite set of structures, and thus a natural number. This is insufficient for my work as I make use of probability generating functions, which can have coefficients with values anywhere in the unit interval.

In order to move to more general frameworks in which all power series can be interpreted combinatorially, one needs generalizations of combinatorial species. One of such generalizations was proposed as a way of formalizing the combinatorics of Feynman diagrams in quantum mechanics (Baez and Dolan, 2001). The framework rests on the concept of *groupoid*, which can be seen as generalizations of sets where each element can potentially have non-trivial symmetries (Brown, 1987; Weinstein, 1996). Better still, groupoids are generalizations of equivalence relations in which transitivity, reflexivity and symmetry are upgraded to compositionality, existence of identities, and existence of inverses, respectively. By introducting *groupoid cardinality*, which as opposed to the cardinality of sets can acquire values spanning all positive real numbers, the authors show that formal power series can be seen as projections of structures defined in terms of groupoids. The

result is that in Equation 3.1 the a_n are now the cardinalities of some groupoid of structures, rather than a set of structures, and thus can be any positive real number. I need to at least upgrade sets to groupoids in order to develop a structural theory of chemical reaction networks.

This line of inquiry has led me naturally to *homotopy type theory* (HoTT) (Univalent Foundations Program, 2013). HoTT is a foundation of mathematics based on *homotopy types*, which can be seen as infinite-dimensional generalizations of groupoids. Just as groupoids are generalizations of sets, 2-groupoids are similarly a generalization of groupoids, and this trend can of generalization be carried out all the way to ∞-groupoids. ∞-groupoids can be seen as consisting of points, paths between points, 2-dimensional paths between paths, and so on. Hence their connection to topology. Formulating formal power series and SCRNs in the framework of HoTT would give them not only combinatorial but also topological interpretation. I believe that this perspective could be useful for mathematical biology. For example, whereas we usually speak of a set of conformations of a molecule it would be more adequate to speak of a *space* of conformations. Similarly, it would be more useful to speak of a space of species and a space of reactions in a chemical reaction network rather than merely sets. Furthermore, the computational aspect of HoTT could be harnessed to understand computation in chemical and biological systems.

The purpose of this chapter is to serve as a starting point for a combinatorial/homotopical interpretation of the objects of SCRN theory. The starting point is, as proposed in Baez and Biamonte, 2018, to reformulate SCRNs in terms of their probability generating functions. Following that to incorporate combinatorial generating functions in order to describe structured collections of species. Once formulated in this way, the idea is to use the generalizations of combinatorial species in the language of HoTT (Yorgey, 2014) in order to place SCRN theory in such a context.

It is well known that linear systems described by an eqeuation of the form

$$\frac{dp}{dt} = Ap,$$

where p is some vector and A a linear transformation, admit solutions of the form $p = e^{At}p_0$, where p_0 is an initial condition and e^{At} is a matrix exponential (Moya-Cessa and Soto-Eguibar, 2011). In the case of SCRNs, and continuous-time Markov chains (CTMCs) in general, A is what is known as the *transition rate matrix*(McQuarrie, 1967). Expanding the matrix exponential, we obtain

$$e^{At} = 1 + At + A^2\frac{t^2}{2!} + \cdots = \sum_{n=0}^{\infty} A^n \frac{t^n}{n!}. \tag{3.2}$$

In Baez and Biamonte, 2018, the authors describe the terms of the above sum conceptually as describing "sums over histories." In particular, the term with A^n would correspond to "histories" where four reactions took place. I found this description illuminating but at the same time conflicting with other notions of history that I had learned in different places.

In my first year at Caltech, I took a course called *order of magnitude biology* with professors Rob Phillips and Justin Bois. One of our homework sets included a problem that asked us to calculate the probability that a kinesin protein would take one step on a microtubule. Such an event consists of kinesin first detaching from the microtubule and then reattaching. This calculation is equivalent to computing the probability that a generic CTMC will undergo two given transitions in a given amount of time, which is proportional to

$$p_{i\to j\to k}(t) = a_{i\to j}a_{j\to k}\int_0^t e^{-a_i t}e^{-a_j(t-\tau)}d\tau = a_i a_j \frac{e^{-a_i t} - e^{a_j t}}{a_j - a_i}, \tag{3.3}$$

where i, j, and k are states, $a_{i\to j}$ and $a_{j\to k}$ are transition rates, and a_i and a_j are the sums of the transition rates for transitions starting at i and j, respectively. This exercise motivated me to generalize the formula for an arbitrary number of transitions as I understood that such a calculation would ultimately allow me to write down a generic solution to the *chemical master equation* (CME) for SCRNs in terms of sums of those integrals ranging over all possible trajectories of the system. Yet, a solution written in that form looked very different from the exponential solution referred to above, and the "histories" given by the terms in Equation 3.2 were certainly not the ones I could obtain from generalizations of the integral in Equation 3.3. The material in Section 3.3 was my attempt at reconciling these two views. Although generating functions were not essential, formal methods were what allowed me to arrive at my desired result.

I spent a great deal of time during my PhD thinking about detailed-balanced SCRNs and their mathematical properties. Their equilibria are mathematically elegant and very pleasant to work with. One thing I was always curious about was whether there were more general systems that displayed the same kind of equilibria. I knew that indeed complex-balanced SCRNs had the same form of stationary distributions (Anderson, Craciun, and Kurtz, 2010), but I wondered whether there were a yet more general class of systems that had the same form of stationary distributions. What is more, I wondered whether it was possible to characterize the class of *all* SCRNs that had stationary distributions with product-of-Poisson form. By considering The generating functions of Poisson distributions are exponential functions. Furthermore, the CME becomes a partial differential equation when translated into the language of generating functions. Since it is easy to take derivatives of exponential functions, I set out to characterize the class of SCRNs that could have exponentials as stationary probability generating functions. What I obtained was basically the definition of complex balance, which revealed to me that complex-balanced SCRNs were in fact the most general class of systems that admitted product-of-Poisson-form stationary distributions. The same result was established in Cappelletti and Wiuf, 2016, using standard SCRN semantics in terms of CTMCs, but the proof using generating functions is much more concise. This is the subject of Section 3.4.

When we were working on the article Cappelletti, Ortiz-Muñoz, et al., 2020, our collaborator David Anderson came up with a construction aimed at controlling the noise in the counts of some molecular species Y through a collection of catalysts X_i. In his analysis of the construction, he made use of *Dynkin's formula* (Øksendal, 2010), which in the context of SCRNs describes the time evolution of the expectation value of a quantity. In particular, for the construction in question, we wanted to know what the variance of Y was. David's proof inspired me to search for general expressions for other higher-order moments of SCRNs. It was then that I came across the notion of *factorial moments*. I discovered that factorial moments were much more elegantly described for SCRN than regular moments. I found out also that it was possible to completely characterize the dynamics of SCRNs in terms of their factorial moments, as it is the case for regular moments. I derived an expression analogous to the CME for factorial moments, but it was very complicated. In an attempt at simplifying it I encountered that the probability generating function $f(z)$ and the factorial moment generating function $m(z)$ are related by $m(z) = f(z+1)$. The finding was startlingly simple, especially since the CME for factorial moments was pretty complicated. A special case of this relationship for the case of there being only one species appears in Behr, Duchamp, and Penson, 2017. Others have explored methods for solving factorial moment hierarchies, which is not something we do here (Krishnamurthy and Smith, 2017; Smadbeck and Kaznessis, 2012, 2013; Smith and Krishnamurthy, 2017, 2021; Sotiropoulos and Kaznessis, 2011). For some time I have dreamt of working out the combinatorial interpretation of factorial moments. The fact that z^n is mapped to $(z+1)^n$ tells me that it must have something to do with subsets and binomial coefficients. I intend to work on this problem in the future. I report my findings about factorial moments in Section 3.5.

The last section focuses on methods for computing the stationary distributions of SCRNs for which a rough notion of structure and molecular content exists. I refer to these as *assembly systems* because they are intended to model systems composed of atomic units that assemble into more complex structures. I exploit the results on complex balance from Section 3.4 in order to extract partition functions from the exponential stationary generating functions. Finally, I apply these methods in conjunction from those of Section 3.5 in order to derive general expressions for the factorial moments of assembly systems. I see this section as a starting point for a generalization of SCRNs in which species are not merely names of molecules but discrete structures composed of units, such as molecules composed of atoms or protein complexes made up of proteins. A formalism parallel to mine with use of generating functions for assembly systems appears in Whittle, 1986, but they do not focus on partition functions or factorial moments.

ABSTRACT

We define the semantics of stochastic chemical reaction networks (SCRNs) in terms of formal power series. The chemical master (CME) equation becomes a formal partial differential equation on the probability generating function of a SCRN. We define a class of regular solutions to the CME as ones where the probability of infinite paths vanishes at all finite times. Equivalently, regular solutions to the CME are those that can be expressed as the exponential of an infinitesimal stochastic operator. We focus on stationary solutions to the CME, which are time-independent. We show that complex-balanced SCRNs are precisely those that admit an exponential power series as a stationary solution. We define factorial moments and their generating function and derive a simple relationship between the factorial moment generating function and the probability generating function. Finally, we define assembly systems, which are complex-balanced SCRNs where each species has a composition, and for which all reactions preserve composition.

3.1 Introduction

As we saw in the previous chapter, the sole structure of a network of interactions in a chemical system can produce highly complex behavior, that is not even considering the additional expressive power that we would gain by considering spatial heterogeneity. In a CRN, we need to specify the complete sets of molecules and interactions as part of their definition. For a real chemical system such as a biological cell it can be hard, or impossible, to know a priori the complete set of molecules and reactions due to the combinatorial complexity of ever larger molecular structures and potential interactions between them. Instead, the process is generative, meaning that from a set of constituent atoms and rules that govern their local interactions we can inductively generate larger and more complex molecules. As such we can see the programmable part of a chemical system not as the complete interaction network itself but as the set of atoms and their local interactions. The species of a CRN would then become derived, structured entities themselves. Due to the discrete nature of molecules these structures are of a combinatorial nature.

In the case of biology the set of atoms from which other molecules are generated are not atoms in the sense of physics; rather, they are the basic constituents of a given level of description, such as nucleotides, amino acids, or proteins. A mathematical model of such systems would be in the form of a CRN *generated* by a set of atomic constituents and local interactions. Such models would give rise to a higher level of precision in terms of elucidating biomolecular phenomena by producing combinatorial interaction models that can be compared to experiment.

Formal power series in the form of generating functions are used in combinatorics and probability theory to summarize collections of numbers such as probability distributions or counts of combinatorial structures (Flajolet and Sedgewick, 2009; Wilf, 1994). Once distributions and counts are represented in the power series form one can usually use the familiar operations of calculus and algebra in order to convert a power series into a closed-form that can be more easily manipulated than mere collections of numbers. In this chapter we will develop a formalism for stochastic chemical reaction networks based on formal power series. The ideas are inspired by Baez and Biamonte, 2018, where they introduce such formalism under the name of *stochastic mechanics*, and they use it to draw parallels with quantum mechanics.

We use the stochastic mechanics formalism in order to define the corresponding *chemical master equation* (CME) and define expressions for the stochastic dynamics. The CME has a standard universal solution in terms of a matrix exponential (Moya-Cessa and Soto-Eguibar, 2011). We use formal methods to show that the standard operator exponential solution can be expressed in a form where probabilities of specific paths can be more readily extracted.

We show that *complex-balanced* SCRNs can be identified with those that admit an exponential function as the stationary distribution. We do this by first showing that complex-balanced systems

admit a stationary distribution whose generating function is an exponential. This was also done in Baez and Biamonte, 2018. This result is well known for SCRNs, though the proofs do not always employ generating functions (Anderson, Craciun, and Kurtz, 2010; Whittle, 1986). We show the converse, namely that if a SCRN admits an exponential stationary solution, then it must be complex balanced. This characterizes complex-balanced SCRNs as those that admit exponential functions as stationary solutions. This result was proved in Cappelletti and Wiuf, 2016, using standard SCRN methods, but the proof we provide here is much shorter and concise.

We will also define *factorial moments* and their generating function. We show that the probability and factorial moment generating functions are related to one another by a simple transformation: adding a unit to each variable. A special case of this result appears in Behr, Duchamp, and Penson, 2017, for the case of systems having a single species. This result is hinted at in Krishnamurthy and Smith, 2017; Smith and Krishnamurthy, 2017, but it does not appear explicitly.

We define assembly systems as SCRNs that conserve mass and are complex balanced. We develop a method for computing the partition function of the different conservation classes, which can be used to compute probabilities. Similar ideas appear in Whittle, 1986, however, they do not provide rigorous proofs and their methods are restricted to detailed balance. We derive a formula for computing the factorial moments of assembly systems.

3.2 Preliminaries

We are interested in well-mixed chemical solutions with a small number of molecules. At any given moment one of such solutions is described by the positions and velocities of all the molecules in the mixture but such a description is too specific for our purposes. In particular, since the solution is well-mixed, we consider not a low level description of the mixture but a high level one in which we only specify the number of molecules present. This is because the well-mixing assumption means that the states of the solution do not spend more time in any one region of phase space than another. As a result, the only informative variable regarding the chemical solution is the number of molecules it contains. Mathematically this means that the state of a system is described by a function that assigns to each kind of molecule, or species, a whole number, its counts. We formalize this below with the concept of a multiset.

For sets A and B, we will use the notations $A \to B$ and B^A for the set of functions from A to B interchangeably. We will denote the set of natural numbers, i.e. non-negative integers, with $\mathbb{N}$, the set of real numbers with $\mathbb{R}$, and the set of positive real numbers with $\mathbb{R}^+$.

Let A be a set. A *multiset* over A is a function $x : A \to \mathbb{N}$ with the property that the sum $\sum_{a \in A} x_a$ is finite. Although the set $\mathbb{N}^A$ may include functions that are not multisets because their total is not finite, whenever we write the symbol $\mathbb{N}^A$, we will hereafter mean the set of multisets over A. We

write multisets $x \in \mathbb{N}^A$ as *formal sums* of elements of A as follows

$$x = \sum_{a \in A} x_a a.$$

The *sum* of two multisets $x, y \in \mathbb{N}^A$ is given by

$$x + y = \sum_{a \in A} (x_a + y_a) a.$$

Example 3.2.1. Let $A = \{X, Y\}$. Examples of multisets over A are

$$0, \quad X, \quad Y, \quad X + Y, \quad 10X + 3Y.$$

An example of a multiset sum is

$$(3X + 2Y) + (5X + 10Y) = 8X + 12Y.$$

We will be interested in how our chemical mixtures evolve over time. Given that our description of a mixture is in terms of multisets of species, we choose to model the evolution as a random process. At any given state of the mixture, given by a multiset of species, any of a number of reactions may occur, each with a different probability. As a result, the time evolution of the mixture is non-deterministic. We choose to model this situation by means of probability distributions, which assign a probability to each of the multisets that are reachable from some initial multiset or ensemble of multisets. Mathematically a probability distribution over multisets is a function that assigns a real number, its probability, to each multiset. In general, we will need more than probability distributions as there may be calculations that involve real-valued functions over multisets that either do not add up to 1 or that have negative values. We therefore consider not just probability distributions but general functions that assign real numbers to multisets.

Formal Power Series

Instead of working with simple functions over multisets, we will be using the notion of a formal power series. In a formal power series, we assign a formal variable to each element of a given set. Then, to each multiset over that set, we associate a monomial consisting of the product of powers of formal variables, where the powers of each variable are given by the values of the multiset. Note that since multisets use natural numbers, the powers of the formal variables in a formal power series are always nonnegative. The value in doing this is that we can take advantage of the algebra of formal power series in order to perform various manipulations of functions over multisets. We will now define formal power series and their algebra.

Definition 3.2.1. Let A be some set. The ring $\mathbb{R}[[z]]$ of *formal power series* over the *formal variables* $z = (z_a)_{a\in A}$, or simply a *power series* over A, consists of functions $f : \mathbb{N}^A \to \mathbb{R}$. For a power series $f \in \mathbb{R}[[z]]$, we use the notation

$$f = \sum_{x\in\mathbb{N}^A} f_x z^x,$$

where $z^x = \prod_{a\in A} z_a^{x_a}$. For power series $f, g \in \mathbb{R}[[z]]$ over A, we define *addition* and *multiplication* as follows

$$f + g = \sum_{x\in\mathbb{N}^A} (f_x + g_x) z^x,$$

$$fg = \sum_{x\in\mathbb{N}^A} \sum_{0\le y\le x} f_y g_{x-y} z^x,$$

where the order of multisets is component-wise, i.e. $y \le x$ iff $y_a \le x_a$ for each $a \in A$.

Although formal power series are very similar to regular power series, which are functions of one or multiple real variables, a difference is that since the variables of a formal power series are not to be thought of as real numbers, we do not need to worry about issues of convergence. For example, the formal power series $\sum_{i=0}^{\infty} i! z^i$ is a well-defined object, whereas the corresponding power series diverges everywhere except at $z = 0$. A regular power series is the representation of a *partial* function $f : \mathbb{R}^n \to \mathbb{R}$ of n variables, which assigns to vectors $z \in \mathbb{R}^n$ of real numbers a real number $f(z)$, whenever $f(z)$ is defined. The reason a power series representation is in general a partial function is that a power series may not converge for some values of z and hence it is not defined for all input values. The concept of a radius of convergence is precisely that of establishing the range with which a power series converges around a point. A formal power series, on the other hand, does not encounter the convergence issue since it is not a function of a real variable. Instead a formal power series $f : \mathbb{N}^A \to \mathbb{R}$, where $|A| = n$, is a *total* function that assigns to each multiset $x \in \mathbb{N}^A$ a real number f_x. A more suggestive notation for f being a formal power series is $f \in \mathbb{R}^{\mathbb{N}^A}$, because it highlights the vector-like nature of f. In this vector picture, $\mathbb{R}^{\mathbb{N}^A}$ is a vector space of infinite dimension $|\mathbb{N}^A|$, f_x is the x-th entry of f, and z^x is the basis vector in the x direction. Hence, unless we invoke an infinite sum of formal power series, a formal power series obtained by applying the operations of formal power series a finite number of times will always be well-defined, as long as we start with well-defined formal power series. This is the case for the operations of addition and multiplication we defined above, as well as for some of the ones we will define below in this chapter.

To be more concrete, let us consider an example in which an operation fails to produce a legitimate formal power series. As mentioned above, if we begin with a set of well-defined formal power

series and use them to construct new formal power series by applying the addition and multiplication operations a finite number of times, what we obtain will always be a well-defined formal power series. It is possible to prove this by induction but we will not do that here. For a proof of this statement see, for example, Flajolet and Sedgewick, 2009. One operation between formal power series that arises naturally in many contexts is that of *composition*. The composition of formal power series should generalize to composition of analytic functions (without requiring the formal power series to be analytic). Let us consider two formal power series on a single variable, $f, g : \mathbb{R}^{\mathbb{N}}$. It would be tempting to define composition as follows

$$f \circ g = \sum_{n \in \mathbb{N}} f_n g^n,$$

where g^n denotes the n-fold multiplication of g times itself, which is well-defined by the discussion above. Notice that the constant term of $f \circ g$ is a number given by the following infinite sum

$$(f \circ g)_0 = \sum_{n \in \mathbb{N}} f_n g_0^n,$$

which we cannot guarantee will converge in general. Hence, in order to yield a well-defined operation, composition of formal power series requires that the right term of the composition, in this case g, has a null constant term, namely $g_0 = 0$. If this is the case then it is possible to show that the summations arising in $f \circ g$ are all of a finite number of terms (see for example Flajolet and Sedgewick, 2009 or Wilf, 1994).

In spite of the difference between regular and formal power series, the notation for formal power series invites us to use our intuitions about regular power series to think about formal ones. For example, the notation z^x for unit vectors resembles a monomial, i.e. a regular power series with one term. The multiplication of formal power series is designed to match the behavior of multiplication of regular power series. For example $z^x z^y = z^{x+y}$. One must be careful, however, not to be too liberal in the application of regular power series concepts to formal power series. In particular, the notions of continuity, radius of convergence, and derivative, are meaningless for formal power series. One example of using old intuitions for formal power series is the *geometric series*. Let us define a geometric series in a single variable as the formal power series

$$\frac{1}{1-z} \equiv 1 + z + z^2 + \cdots = \sum_{i=0}^{\infty} z^i.$$

Notice that the expression on the left is purely formal. We did not give a definition of the reciprocal of a formal power series. Instead, we use that notation to emphasize the fact that if we multiply the geometric series by $1 - z$ we obtain the following by applying the definitions of addition and multiplication of formal power series

$$(1-z)\frac{1}{1-z} \equiv (1-z)(1 + z + z^2 + \cdots) = (1 + z + z^2 + \cdots) - (z + z^2 + z^3 + \cdots) = 1.$$

We must therefore be careful to remember that in the expression $1/(1-z)$ we are not actually dividing by $1-z$; instead, it is reminding us that $1-z$ is the multiplicative inverse of the geometric series. Furthermore, since z is not a number, to say that $1/(1-z)$ is not well-defined for $z=1$ is meaningless.

Henceforth, by power series we will mean a formal power series unless we state otherwise.

Example 3.2.2. Let $A = \{1, 2\}$. Examples of formal power series over A are

$$0, \quad 1, \quad z_1, \quad z_2, \quad 1 + 2z_1 + 2z_2 + 4z_1z_2, \quad \sum_{x \in \mathbb{N}^A} \frac{z_1^{x_1}}{x_1!} \frac{z_2^{x_2}}{x_2!}, \quad \sum_{x \in \mathbb{N}^A} z_1^{x_1} z_2^{x_2}.$$

Examples of additions and multiplications of power series over A are the following

$$(1 + 2z_1) + (1 + z_2) = 2 + 2z_1 + 2z_2,$$

$$(1 + 2z_1)(1 + 2z_2) = 1 + 2z_1 + 2z_2 + 4z_1z_2,$$

$$\left(\sum_{x \in \mathbb{N}^A} \frac{z_1^{x_1}}{x_1!} \frac{z_2^{x_2}}{x_2!} \right) + \left(\sum_{x \in \mathbb{N}^A} z_1^{x_1} z_2^{x_2} \right) = \sum_{x \in \mathbb{N}^A} \left(1 + \frac{1}{x_1! x_2!} \right) z_1^{x_1} z_2^{x_2},$$

$$\left(\sum_{x \in \mathbb{N}^A} \frac{z_1^{x_1}}{x_1!} \frac{z_2^{x_2}}{x_2!} \right) \left(\sum_{x \in \mathbb{N}^A} z_1^{x_1} z_2^{x_2} \right) = \sum_{x \in \mathbb{N}^A} \sum_{0 \leq y \leq x} \frac{1}{y_1! y_2!} z_1^{x_1} z_2^{x_2}.$$

3.3 Dynamics

So far we have talked about states of well-mixed chemical solutions, which we have chosen to model mathematically as multisets, and probability distributions over states, which we have chosen to model as formal power series. We are ultimately interested in how these states or distributions evolve over time according to reactions, so we need to nail down a mathematical model for reactions. For our purposes, a reaction will be a rule that specifies a number of reactants, which are a collection of species, which can be converted into a number of products, also a collection of species. In addition, each reaction will have a probability rate of occurring at a given state of the mixture, and that probability rate will be proportional to the number of ways in which the reactants can be chosen from the mixture. Formally then a reaction is a pair of multisets of species, its reactants and products, along with a positive real number that gives the proportionality of the probability rate of the reaction to the number of ways of picking out the reactants from a mixture. In a given chemically-reacting mixture, there may be a number of reactions possible, and we refer to the collection of reactions as a network of reactions. We will now provide the formal definition of a chemical reaction network.

Definition 3.3.1. A *chemical reaction network* is a quadruple (S, R, n, k), where S is a set of *species*, R is a set of *reactions*, $n : 2 \times R \times S \to \mathbb{N}$ is a *stoichiometric function*, and $k : R \to \mathbb{R}^+$ is a *rate constant function*.

In order to know the reactants of a given reaction $r \in R$, we evaluate the stoichiometric function at the pair $(0, r)$, which gives the reactant multiset $n_{0,r} : S \to \mathbb{N}$. Similarly, the evaluation $n_{1,r} : S \to \mathbb{N}$ gives the product multiset of r. We will use the following notation for the stoichiometry and rate constant of reactions

$$\sum_{s \in S} n_{0,r,s} s \xrightarrow{k_r} \sum_{s \in S} n_{1,r,s} s.$$

As we discussed, the rate constant is a constant of proportionality that when multiplied times the number of ways of selecting the reactants of a reaction out of a mixture it gives the probability per unit time that that reaction will take place within such mixture. The number of ways the reactants of a reaction can be selected out of a mixture is a combinatorial factor that we now explore. Suppose that the state of a mixture is described by a multiset $x \in \mathbb{N}^S$, and that we want to know how many ways we can pick out the reactants of some reaction $r \in R$. Although the multiset description only provides the amounts of each species in a mixture and hence does not take into consideration the fact that the molecules contained in the mixture are all distinct, we must take the latter into consideration in order to obtain the correct combinatorial factor we are after. Let us focus on one species $s \in S$ for the moment. There are a total of x_s instances of s in the mixture, and the reaction requires $n_{0,r,s}$ of them. There are $\binom{x_s}{n_{0,r,s}} = \frac{x_s!}{n_{0,r,s}!(x_s - n_{0,r,s})!}$ number of ways of choosing $n_{0,r,s}$ molecules out of the x_s that are available. However, since the level of abstraction that we have chosen for reactions is such that they only specify how many molecules are required but not *how* they are required, we choose to assume that all molecules in the reactant play a different role in the reaction. As a result, we must consider not combinations but permutations, and, therefore, the number of ways of picking out the species in the reactants is $\frac{x_s!}{(x_s - n_{0,r,s})!}$. If in a fine-grained description of a reaction it turns out that all the molecules participate identically, or if there are any symmetries, the rate constants at our level of abstraction must absorb that information. Our analysis was for a single species, but the combinatorial factor giving the total number of ways of picking out the reactants from the mixture is the product of the respective permutations for each species. With our combinatorial factor in hand, we can now compute the probability rate, also known as the *propensity*, at which the reaction will take place within the given mixture. Hence, the propensity of reaction r for a multiset x, which we denote as $\varrho_{r,x}$ is given by

$$\varrho_{r,x} = k_r \frac{x!}{(x - n_{0,r})!},$$

where $x! \equiv \prod_{s \in S} x_s!$ for any multiset $x \in \mathbb{N}^S$.

Example 3.3.1. Let (S, R, n, k) be a CRN with $S = \{X, Y\}$, $R = \{1, 2\}$, and with stoichioemtries and rate constants given by

$$2X + Y \xrightarrow{1.2} X, \quad 0 \xrightarrow{0.3} X + Y.$$

The propensities at multiset $10X + 6Y$ are given by

$$\varrho_{1,10X+6Y} = 1.2(10 \cdot 9)(6) = 648, \quad \varrho_{2,10X+6Y} = 0.3.$$

There is a nice way of extracting the propensities of a reaction by exploiting the formal power series formalism. First, let us notice that if we take multiple derivatives of a monomial z^n, where for the moment z is a single formal variable and not a tuple of formal variables, we obtain permutations as coefficients. For example, for a third derivative we obtain

$$\frac{d^3}{dz^3}(z^n) = \frac{d^2}{dz^2}(nz^{n-1}) = \frac{d}{dz}(n(n-1)z^{n-2}) = n(n-1)(n-2)z^{n-3} = \frac{n!}{(n-3)!}z^{n-3}.$$

If we let $z = (z_a)_{a \in A}$ again be an S-tuple of formal variables, we see that the multiple application of derivatives obtains the combinatorial factor of the form appearing in the formula for propensities. For example, if $S = \{a, b, c\}$, the multiple derivative below obtains

$$\frac{\partial^6}{\partial z_a^3 \partial z_b^2 \partial z_c}(z^x) = x_a(x_a-1)(x_a-2)x_b(x_b-1)x_c z_a^{x_a-3} z_b^{x_b-2} z_c^{x_b-1} = \frac{x!}{(x-3a-2b-c)!}z^{x-3a-2b-c},$$

where $x \in \mathbb{N}^S$ is a multiset. Let us now give a more general and formal treatment of the preceding ideas.

For the rest of the chapter, an *operator* $\mathscr{O} : \mathbb{R}[[z]] \to \mathbb{R}[[z]]$ will be a linear function from power series to power series. We would like to define a differential operator that recapitulates the behavior of the familiar differential operator in calculus. In particular, a differential operator must satisfy the *power rule*

$$\frac{\partial}{\partial z_a}(z^x) = x_a z^{x-a},$$

as we have already made use of in the above discussion. Since an operator is linear and every power series is a linear combination of monomials, the power rule in fact uniquely determines the action of the differential operator. We will now provide the formal definition of the differential operators. In order to keep notation more succinct, we opt to use the symbol ∂_a in place of $\frac{\partial}{\partial z_a}$.

For each $a \in A$, we define a *differential operator* $\partial_a : \mathbb{R}[[z]] \to \mathbb{R}[[z]]$ as follows

$$\partial_a f = \sum_{x \in \mathbb{N}^A} (x_a + 1) f_{x+a} z^x.$$

For each multiset $x \in \mathbb{N}^A$ we define the *multi-differential operator* $\nabla^x : \mathbb{R}[[z]] \to \mathbb{R}[[z]]$ as follows

$$\nabla^y f = \sum_{x \in \mathbb{N}^A} \frac{(x+y)!}{x!} f_{x+y} z^x.$$

The reason for the notation for the multi-differential operator is that it can be seen as multiset powers of a *gradient operator* $\nabla = (\partial_a)_{a \in A}$. More explicitly, since the differential operators commute with one another, we can write the expression

$$\nabla^x = \prod_{a \in A} \partial_a^{x_a},$$

where the product indicates repeated application of an operator. We leave it to the reader to verify that our definition of the multi-differential operator does in fact coincide with the above product of differential operators.

With the multi-differential operator in hand, we observe that propensities appear as coefficients of the following operator application

$$k_r \nabla^{n_{0,r}} f = \sum_{x \in \mathbb{N}^S} k_r \frac{(x+n_{0,r})!}{x!} f_{x+n_{0,r}} z^x = \sum_{x \in \mathbb{N}^S} \varrho_{r,x+n_{0,r}} z^x.$$

Applied to a single monomial we have

$$k_r \nabla^{n_{0,r}} z^x = \varrho_{r,x} z^{x-n_{0,r}}.$$

As we discussed previously, the propensity of a reaction at a state gives the probability per unit time that the reaction will take place at the given state. When a reaction takes place, it changes the state, which results in a corresponding decrease of probability of being at that state. Accordingly, if a given state is the result of applying a reaction at some other state, the probability of the given state will increase when such a reaction takes place. By taking into consideration all the reactions that may lead out of a state, as well as all the reactions that may lead into that state, we can use the propensities to compute the rate of change of probability at a state. The following equation for the rate of change of probability is known as the *chemical master equation*

$$\partial_t f_x = \sum_{r \in R} \varrho_{r,x+n_{0,r}} f_{x+n_{0,r}} - \varrho_{r,x} f_x,$$

where f_x denotes the time-dependent probability of being in state x. The chemical master equation characterizes the dynamics of a probability distribution described by a chemical reaction network.

The chemical master equation above involves time-dependent probabilities, as well as a time derivative, which we have not yet defined. In the spirit of formal semantics, rather than defining

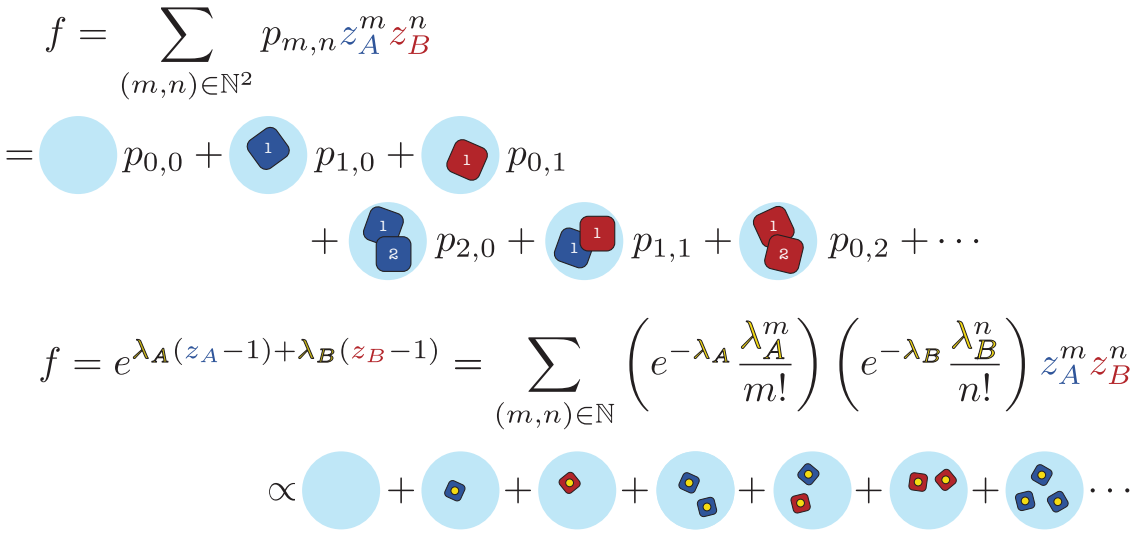


Figure 3.2: Illustration of probability generating functions. On the top is a general probability generating function and its combinatorial interpretation. Below is the probability generating function of a multivariate Poisson, which is given by a product of exponentials.

time-dependence as dependence on a real-valued time variable, we will treat time itself as a formal variable. What this means is that we will regard time-dependent power series as power series with an additional formal variable corresponding to time. As we will see later, making time a formal variable will come in handy, especially when taking integrals with respect to time. As opposed to the z variables, which we think of as strictly formal, the formal time variable will eventually be evaluated in order to obtain probabilities at given times. In later sections, we will introduce formal devices for evaluating the formal time variable at desired positive real numbers. We will now provide the formal definition of time-dependent formal power series.

A *dynamic power series* over A is a power series over the variables $z = (z_a)_{a \in A}$ and the *formal time variable* t. We denote the ring of dynamic power series by $\mathbb{R}[[z, t]]$. For a dynamic power series $f \in \mathbb{R}[[z, t]]$, we use the notation

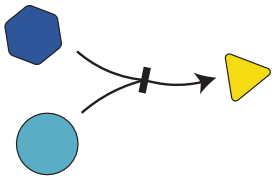
$$f = \sum_{\substack{i \in \mathbb{N} \\ x \in \mathbb{N}^A}} f_{i,x} t^i z^x.$$

We will also make use of the following notation

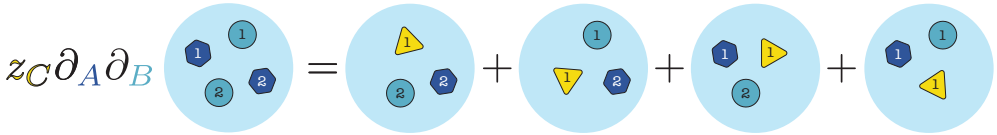
$$f = \sum_{i=0}^{\infty} f_i t^i,$$

where each $f_i \in \mathbb{R}[[z]]$ is a power series over A. Finally, we will also use the notation

$$f = \sum_{x \in \mathbb{N}^A} f_x z^x,$$



$$z_C \partial_A \partial_B (z_A^2 z_B^2) = 4 z_A z_B z_C$$



$$\mathcal{A}_r \, f = k_r (z_C - z_A z_B) \partial_A \partial_B f$$

Figure 3.3: Illustration of reaction operator. The terms of the infinitesimal stochastic operator are reaction operators. Their combinatorial interpretation is shown here for an example reaction.

where each $f_x \in \mathbb{R}[[t]]$ is a formal power series on the single variable t. These last two notations will come into conflict as the symbol f_0 may mean $i = 0$ or $x = 0$. Whenever ambiguity may arise, we will write $f_{0,\cdot}$ or $f_{\cdot,0}$ in order to distinguish each case; otherwise, we will deduce the meaning from context. The differential operator for t will be analogous to the differential operator for the z variables

$$\partial_t f = \sum_{i=0}^{\infty} (i+1) f_{i+1} t^i .$$

Having defined dynamic power series and explored the link between differential operators and propensities, we are ready to define the stochastic dynamics of a CRN in terms of formal power series. We will do so by introducing an operator whose action on a dynamic power series is to give the time derivative of the dynamics and recovers the chemical master equation.

The *infinitesimal stochastic operator* $\mathscr{A} : \mathbb{R}[[z]] \to \mathbb{R}[[z]]$ is given by

$$\mathscr{A} = \sum_{r \in R} k_r (z^{n_{1,r}} - z^{n_{0,r}}) \nabla^{n_{0,r}} .$$

A *stochastic dynamics* $f \in \mathbb{R}[[z,t]]$ of a CRN is a dynamic power series satisfying

$$\partial_t f = \mathscr{A} f .$$

We will now show that the stochastic dynamics we defined above indeed corresponds to the chemical

master equation. Using the relationship between differential operators and propensities, we obtain

$$\mathscr{A} f = \sum_{r \in R} k_r (z^{n_{1,r}} - z^{n_{0,r}}) \nabla^{n_{0,r}} f = \sum_{r \in R} \sum_{x \in \mathbb{N}^S} (z^{n_{1,r}} - z^{n_{0,r}}) \varrho_{r,x+n_{0,r}} f_{x+n_{0,r}} z^x$$

$$= \sum_{x \in \mathbb{N}^S} \sum_{r \in R} \varrho_{r,x+n_{0,r}} f_{x+n_{0,r}} (z^{x+n_{1,r}} - z^{x+n_{0,r}}) = \sum_{x \in \mathbb{N}^S} \sum_{r \in R} (\varrho_{r,x+n_{0,r}-n_{1,r}} f_{x+n_{0,r}-n_{1,r}} - \varrho_{r,x} f_x) z^x$$

$$= \sum_{x \in \mathbb{N}^S} \partial_t f_x z^x = \partial_t f.$$

Example 3.3.2. Let $S = \{A, B, C\}$ and let (S, R, n, k) be a CRN with the following reactions

$$A + B \underset{5}{\overset{7}{\rightleftharpoons}} 2C,$$

where the double arrow indicates that the reaction is reversible. The infinitesimal stochastic operator for this CRN is

$$\mathscr{A} = 7(z_C^2 - z_A z_B)\partial_A \partial_B + 5(z_A z_B - z_C^2)\partial_C^2.$$

The chemical master equation for this CRN has the form

$$\partial_t f_x = 7(x_A + 1)(x_B + 1) f_{x+A+B-2C} + 5(x_C + 2)(x_C + 1) f_{x+2C-A-B} + (7 x_A x_B - 5 x_C (x_C - 1)) f_x.$$

We will now begin to look at methods for finding solutions for stochastic dynamics of CRNs. Our first strategy will consist of integrating the CME. We begin by defining time integration for the time variable. As we did for derivatives of formal power series, we want integration to be a generalization of the usual notion of integration. As we know from calculus, the definite integral of a monomial satisfies

$$\int_0^t t^n dt = \frac{t^{n+1}}{n+1}.$$

Since all dynamic power series are linear combinations of powers of t (with coefficients in $\mathbb{R}[[z]]$), the above integral operation extends uniquely to all dynamic power series, which leads to the following definition.

The *time integral operator* $\int_0^t dt : \mathbb{R}[[z,t]] \to \mathbb{R}[[z,t]]$ on a dynamic power series $f \in \mathbb{R}[[z,t]]$ is given by

$$\int_0^t f dt = \sum_{i=1}^{\infty} \frac{f_{i-1}}{i} t^i.$$

With regards to time integration and differentiation, we should expect them to satisfy some generalized version of the fundamental theorem of calculus. If we focus on a single monomial, we see that the following is true

$$\int_0^t \partial_t t^n dt = \int_0^t n t^{n-1} dt = n \frac{t^n}{n} = t^n,$$

so long as $n \geq 1$. In the case that $n = 0$, we obtain

$$\int_0^t \partial_t t^0 dt = \int_0^t 0 dt = 0.$$

Taking these identities together, we have the *formal fundamental theorem of calculus*

$$\int_0^t \partial_t f \, dt = \sum_{i=0}^{\infty} f_i \int_0^t \partial_t t^i = \sum_{i=1}^{\infty} f_i t^i = f - f_0,$$

where $f_0 \in \mathbb{R}[[z]]$. We now use the fundamental theorem to integrate the CME. We obtain

$$f - f_0 = \int_0^t \partial_t f = \int_0^t \mathscr{A} f \, dt,$$

which leads to the following fixed point equation

$$f = f_0 + \int_0^t \mathscr{A} f \, dt. \tag{3.4}$$

We refer to this as the *integral form* of the CME. This equation gives f as the result of doing something to f itself. We can therefore apply the same transformation twice and obtain

$$f = f_0 + \int_0^t \mathscr{A} f \, dt = f_0 + \int_0^t \mathscr{A} f_0 dt + \int_0^t \mathscr{A} \int_0^t \mathscr{A} f \, dt dt = f_0 + t\mathscr{A} f_0 + \int_0^t \mathscr{A} \int_0^t \mathscr{A} f \, dt dt.$$

Notice that repeated integration of the constant series 1 yields the recursion

$$\left(\int_0^t \mathscr{A} \, dt\right)^i 1 = \left(\int_0^t \mathscr{A} \, dt\right)^{i-1} \mathscr{A} t = \left(\int_0^t \mathscr{A} \, dt\right)^{i-2} \frac{t^2}{2} \mathscr{A}^2 = \left(\int_0^t dt\right)^{i-3} \frac{t^3}{6} \mathscr{A}^3 = \cdots = \frac{t^i}{i!} \mathscr{A}^i.$$

In order to obtain a more general formula for f, we continue the recursion in Equation 3.4, applying the transformation of the integral CME an arbitrary number of times, and obtain

$$f = \left(\int_0^t \mathscr{A} \, dt\right)^{n+1} f + \sum_{i=0}^{n} \left(\int_0^t \mathscr{A} \, dt\right)^i f_0 = \left(\int_0^t \mathscr{A} \, dt\right)^{n+1} f + \sum_{i=0}^{n} \frac{t^i}{i!} \mathscr{A}^i f_0.$$

Intuitively, we would expect that, if we let n go to infinity, we would obtain

$$f = \lim_{n \to \infty} \left(\int_0^t \mathscr{A} \, dt\right)^n f + \sum_{i=0}^{\infty} \frac{t^i}{i!} \mathscr{A}^i f_0,$$

however, we do not know if the limit in the first term is well-defined. We will focus on stochastic dynamics that can be expressed as the sum in the second term. Notice that the sum in the second term has the form of the series of an exponential, which leads to the following definition.

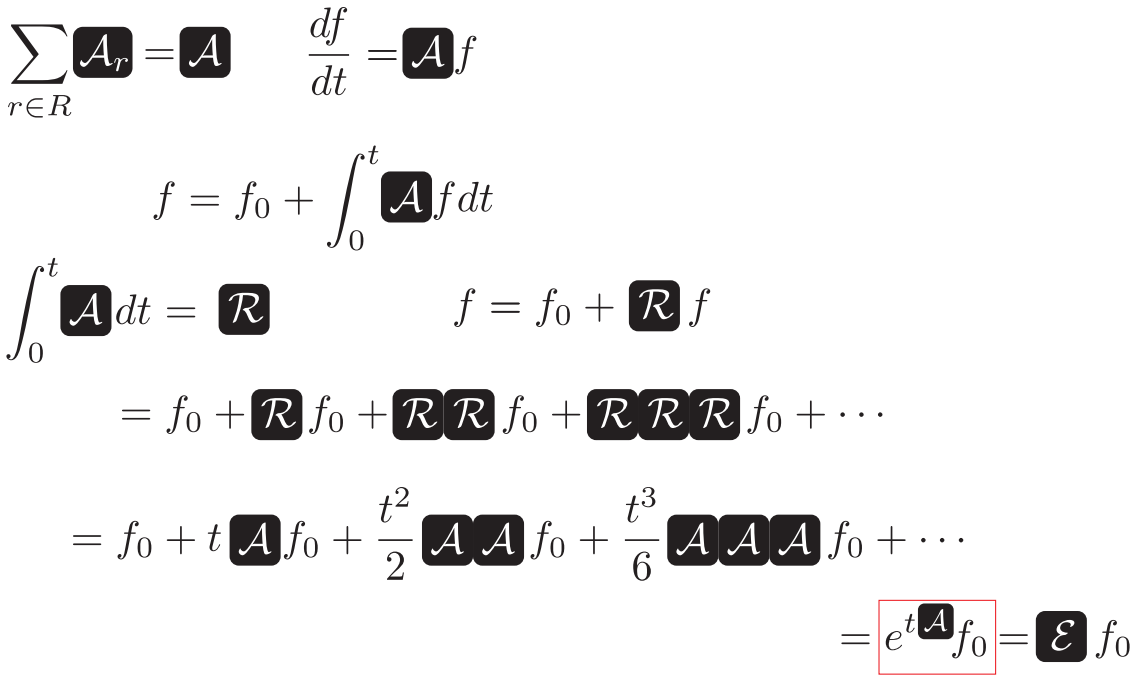


Figure 3.4: Pictorial representation of the exponential solution to the chemical master equation and its derivation.

We define the *operator exponential* as

$$e^{t\mathscr{A}} = \sum_{i=0}^{\infty} \frac{t^i}{i!} \mathscr{A}^i.$$

We say that a stochastic dynamics $f \in \mathbb{R}[[z,t]]$ is *regular* if it has the form

$$f = e^{t\mathscr{A}} f_0.$$

We will now look at an alternative way of obtaining the same exponential solution. Recall the formula for the geometric series

$$\frac{1}{1-t} = \sum_{i=0}^{\infty} t^i.$$

This suggests a definition for inverses of operators of the form $1 - \mathscr{O}$ as

$$\frac{1}{1-\mathscr{O}} = \sum_{i=0}^{\infty} \mathscr{O}^i,$$

provided the sum is well-defined. We can express the CME as follows

$$f_0 = f - \int \mathscr{A} f = \left(1 - \int \mathscr{A}\right) f,$$

where we have used $\int = \int_0^t dt$ for brevity. Using the above operator inverse formula, we obtain

$$f = \frac{1}{1 - \int \mathscr{A}} f_0 = \sum_{i=0}^{\infty} \left(\int \mathscr{A} \right)^i f_0 = \sum_{i=0}^{\infty} \frac{t^i}{i!} \mathscr{A}^i f_0 = e^{t\mathscr{A}} f_0.$$

Let us define positive and negative parts of the infinitesimal stochastic operator as follows

$$\mathscr{A} = \mathscr{A}_+ - \mathscr{A}_- = \sum_{r \in R} k_r z^{n_{1,r}} \nabla^{n_{0,r}} - \sum_{r \in R} k_r z^{n_{0,r}} \nabla^{n_{0,r}}.$$

We can write the CME as follows

$$f_0 + \int \mathscr{A}_+ f = f + \int \mathscr{A}_- f = \left(1 + \int \mathscr{A}_-\right) f.$$

Using the operator inverse formula, we obtain

$$f = \frac{1}{1 + \int \mathscr{A}_-} f_0 + \frac{1}{1 + \int \mathscr{A}_-} \int \mathscr{A}_+ f.$$

We define the *waiting operator* as

$$\mathscr{W} = \frac{1}{1 + \int \mathscr{A}_-} = \sum_{i=0}^{\infty} \left(- \int \mathscr{A}_- \right)^i.$$

The waiting operator corresponds to all the instances of no reactions happening in a given amount of time. Notice that when applied to a time-independent function, the waiting operator gives

$$\mathscr{W} f_0 = \sum_{i=0}^{\infty} \left(- \int \mathscr{A}_- \right)^i f_0 = \sum_{i=0}^{\infty} \frac{(-t)^i}{i!} \mathscr{A}_-^i f_0 = e^{-t\mathscr{A}_-} f_0,$$

so the waiting operator simplifies to the exponential operator $e^{-t\mathscr{A}_-}$. In terms of the waiting operator, we can express the CME as

$$f = e^{-t\mathscr{A}_-} f_0 + \mathscr{W} \int \mathscr{A}_+ f.$$

Example 3.3.3. Let $S = \{A, B\}$ and let (S, R, n, k) be a CRN defined by

$$A \underset{k_2}{\overset{k_1}{\rightleftharpoons}} B.$$

We will compute the probability that a system starts with $5A$, converts one A into a B, and then converts a B into an A, leading back to $5A$. We have the following transitions and propensities

$$5A \underset{k_2}{\overset{5k_1}{\rightleftharpoons}} 4A + B \xrightarrow{4k_1} 3A + 2B.$$

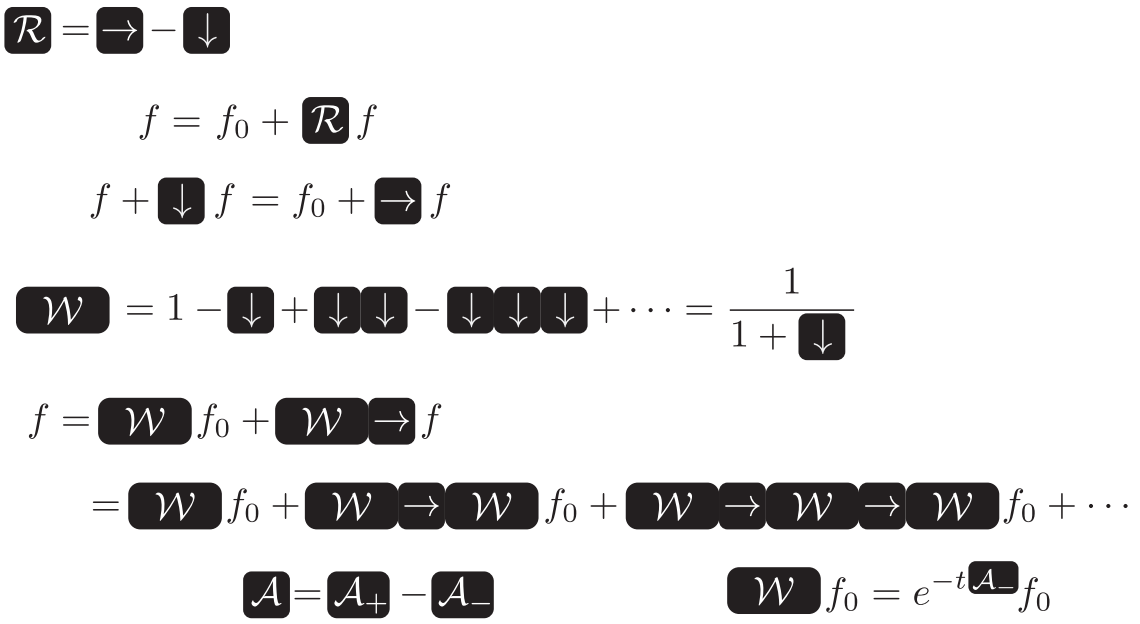


Figure 3.5: Pictorial representation of the waiting operator and of the derivation of the path integral solution to the CME.

We have the following operator sequence

$$\int k_2 z_A \partial_B \mathscr{W} \int k_1 z_B \partial_A \mathscr{W} z_A^5 = k_1 k_2 \int z_A \partial_B \mathscr{W} \int z_B \partial_A e^{-t5k_1} z_A^5$$

$$= 5k_1 k_2 \int z_A \partial_B \mathscr{W} \sum_{i=1}^{\infty} \frac{t^i}{i!} (-5k_1)^{i-1} z_A^4 z_B$$

$$= 5k_1 k_2 \int z_A \partial_B \sum_{j=0}^{\infty} \sum_{i=1}^{\infty} \left(- \int \mathscr{A}_- \right)^j \frac{t^i}{i!} (-5k_1)^{i-1} z_A^4 z_B$$

$$= 5k_1 k_2 \int z_A \partial_B \sum_{j=0}^{\infty} \sum_{i=1}^{\infty} \frac{t^{i+j}}{(i+j)!} (-4k_1 - k_2)^j (-5k_1)^{i-1} z_A^4 z_B$$

$$= 5k_1 k_2 \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \frac{t^{i+j}}{(i+j)!} (-4k_1 - k_2)^{j-1} (-5k_1)^{i-1} z_A^5$$

$$= 5k_1 k_2 \sum_{i=2}^{\infty} \sum_{j=0}^{i-2} \frac{t^i}{i!} (-4k_1 - k_2)^{i-j-2} (-5k_1)^j z_A^5.$$

3.4 Stationarity

We will now, and for the rest of this chapter, focus on stationary solutions to the CME. A stationary dynamics is one which is equal to its initial condition. In other words, the stochastic dynamics is

time invariant. Stationary solutions contain an important class of dynamics, those that are obtained as the long-term dynamics of systems initialized at a pure mixture. Such limit distributions are stationary. On the other hand, not all stationary distributions are limit distributions. In particular, any linear combination of stationary distributions is itself a stationary distribution, up to normalization, so a stationary distribution composed of two limit distributions may not in general be the limit distribution of any pure state. If we consider the limits of dynamics with arbitrary initial conditions, then limit dynamics coincide with stationary dynamics. We have already argued that limit dynamics are stationary. Conversely, if the initial condition of a stochastic dynamics is stationary then it is also the limit dynamics. For this reason, we will focus on stationary dynamics, which are easier to define.

We say that a stochastic dynamics f is *stationary* if it satisfies

$$f = f_0 + \int \mathscr{A} f = f_0.$$

A regular stochastic dynamics f is stationary if and only if

$$f = f_0, \quad \text{and} \quad \mathscr{A} f = 0.$$

As we previously discussed, a positive linear combination of stationary dynamics is a stationary dynamics. We can think of stationary dynamics obtained this way as being composite. Accordingly, we think of stationary dynamics which cannot be obtained as positive linear combinations of other stationary dynamics as being atomic, or indivisible.

Let (S, R, n, K). We say that a stationary dynamics $f \in \mathbb{R}[[z]]$ is *irreducible* if for any positive series $g, h \in \mathbb{R}[[z]]$ with $f = g + h$, if g is a stationary dynamics, then $h = 0$.

3.4.1 Complex balance

We will now focus on stationary series that can be expressed as exponential functions. The generating functions of product of Poisson distributions are exponential functions, and are also stationary solutions for the dynamics of complex-balanced CRNs. We will review the definition of complex balance and show that it is equivalent to exponential stationary series.

We say that a CRN is *complex balanced* if there exists a function $c : S \to \mathbb{R}^+$ such that for all multisets $x \in \mathbb{N}^S$, we have

$$\sum_{\substack{r \in R: \\ n_{0,r} = x}} k_r c^{n_{0,r}} = \sum_{\substack{r \in R: \\ n_{1,r} = x}} k_r c^{n_{0,r}}.$$

For a function $c : S \to \mathbb{R}^+$, we define the *exponential power series* e^{cz} as follows

$$e^{cz} = \sum_{x \in \mathbb{N}^S} \frac{c^x}{x!} z^x.$$

Let us assume that a CRN has a stationary dynamics of the form $f = e^{cz}$. Observe that applying a differential operator to an exponential series, we obtain

$$\nabla^x e^{cz} = c^x e^{cz}.$$

If we apply the infinitesimal stochastic operator to an exponential power series, we obtain

$$\mathscr{A} e^{cz} = \sum_{r \in R} k_r (z^{n_{1,r}} - z^{n_{0,r}}) \nabla^{n_{0,r}} e^{cz} = \sum_{r \in R} k_r (z^{n_{1,r}} - z^{n_{0,r}}) c^{n_{0,r}} e^{cz}$$

$$= e^{cz} \sum_{r \in R} k_r (z^{n_{1,r}} - z^{n_{0,r}}) c^{n_{0,r}} = e^{cz} \sum_{x \in \mathbb{N}^S} \left(\sum_{\substack{r \in R: \\ n_{1,r} = x}} k_r c^{n_{0,r}} - \sum_{\substack{r \in R: \\ n_{0,r} = x}} k_r c^{n_{0,r}} \right) z^x.$$

If we wish to have $\mathscr{A} e^{cz} = 0$, as required for a stationary dynamics, we must either have $e^{cz} = 0$ or

$$\sum_{\substack{r \in R: \\ n_{1,r} = x}} k_r c^{n_{0,r}} - \sum_{\substack{r \in R: \\ n_{0,r} = x}} k_r c^{n_{0,r}} = 0,$$

for each $x \in \mathbb{N}^S$, which is the definition of complex balance. Since $e^{cz} \neq 0$, we must have that a CRN is complex balanced if, and only if, there exists an exponential power series e^{cz} that is stationary

$$\mathscr{A} e^{cz} = 0.$$

3.5 Factorial moments

We will now be interested in a characterization of stochastic dynamics in terms of factorial moments. The factorial moments are the expected values of the numbers of ways of permuting a multiset onto mixtures of an ensemble. In order to formally define expected values, we introduce the inner product of power series.

The *inner product* between power series $f, g \in \mathbb{R}[[z]]$ is given by

$$\langle f, g \rangle = \sum_{x \in \mathbb{N}^A} x! f_x g_x$$

where $x! = \prod_{a \in A} x_a!$. When it is well defined, the inner product between two power series is a real number. With this definition of inner product, we can express the sum of the coefficients of a power series as the inner product with e^z

$$\langle e^z, f \rangle = \sum_{x \in \mathbb{N}^S} f_x.$$

Also, we can extract coefficients of power series using the inner products with

$$\left\langle \frac{z^x}{x!}, f \right\rangle = f_x.$$

Suppose we are interested in the expected value of a function $\alpha : \mathbb{N}^S \to \mathbb{R}$. We consider the following *exponential series* of α

$$\bar{\alpha} = \sum_{x \in \mathbb{N}^S} \frac{z^x}{x!} \alpha_x.$$

Then, the *expected value* of α with respect to a normalized power series f, if it exists, is given by

$$E_f[\alpha] = \langle \bar{\alpha}, f \rangle = \sum_{x \in \mathbb{N}^S} \alpha_x f_x.$$

Given a distribution f, the *factorial moment* of order $x \in \mathbb{N}^S$ is the expected value of α with $\alpha_y = \frac{y!}{(y-x)!}$:

$$m_x = \sum_{y \in \mathbb{N}^S} \frac{y!}{(y-x)!} f_y.$$

Let $x \in \mathbb{N}^S$ and $\alpha_y = \frac{y!}{(y-x)!}$ for $y \in \mathbb{N}^S$ so that $m_x = E[\alpha]$. Hence, we have that the exponential series of α satisfies

$$\bar{\alpha} = \sum_{y \in \mathbb{N}^S} \frac{z^y}{(y-x)!} = z^x e^z.$$

We have therefore that

$$m_x = \langle z^x e^z, f \rangle.$$

Let us define the *factorial moment operator* as follows

$$\mathscr{M} f = \sum_{x \in \mathbb{N}^S} \frac{z^x}{x!} \langle z^x e^z, f \rangle.$$

Notice that the variables z inside the inner product disappears after taking the inner product. We could therefore substitute for any equivalent tuple of variables and have the same effect. Let us assume that the inner products are taken over the tuple $z' = (z'_s)_{s \in S}$. We can now take the sum inside the inner product and obtain

$$\mathscr{M} f = \sum_{x \in \mathbb{N}^S} \frac{z^x}{x!} \langle z'^x e^{z'}, f \rangle = \langle e^{z'z} e^{z'}, f \rangle = \langle e^{z'(z+1)}, f \rangle = \sum_{x \in \mathbb{N}^S} (z+1)^x f_x.$$

If we define the evaluation of f at $z + 1 = (z_s + 1)_{s \in S}$ by

$$f(z+1) = \sum_{x \in \mathbb{N}^S} (z+1)^x f_x,$$

then we have that

$$\mathscr{M} f(z) = f(z+1).$$

We would now like to obtain the dynamics of the moment series. We will do so by taking the time derivative of the moment operator applied to a stochastic dynamics, and then use the CME to derive an expression for the dynamics of the moment series.

If we take the time derivative of the moment operator applied to a stochastic dynamics, we obtain

$$\partial_t \mathscr{M} f = \mathscr{M} \partial_t f = \mathscr{M} \mathscr{A} f.$$

Let us define the *factorial moment series* as the application of the moment operator to the stochastic dynamics

$$\mu = \mathscr{M} f.$$

We therefore have that

$$\partial_t \mu = \mathscr{M} \mathscr{A} f = \mathscr{M} \mathscr{A} \mathscr{M}^{-1} \mu.$$

The operator $\mathscr{M} \mathscr{A} \mathscr{M}^{-1}$ is analogous to the infinitesimal stochastic operator, where in this case it generates the dynamics of the moment series rather than the probability series. Let us examine the operator $\mathscr{M} \mathscr{A} \mathscr{M}^{-1}$ by looking at its action on pure mixtures

$$\mathscr{M} \mathscr{A} \mathscr{M}^{-1} z^x = \mathscr{M} \mathscr{A} (z-1)^x = \mathscr{M} \sum_{r \in R} k_r (z^{n_{1,r}} - z^{n_{0,r}}) \nabla^{n_{0,r}} (z-1)^x$$

$$= \sum_{r \in R} k_r ((z+1)^{n_{1,r}} - (z+1)^{n_{0,r}}) \nabla^{n_{0,r}} z^x.$$

We can therefore conclude that

$$\partial_t \mu = \mathscr{M} \mathscr{A} \mathscr{M}^{-1} \mu = \sum_{r \in R} k_r ((z+1)^{n_{1,r}} - (z+1)^{n_{0,r}}) \nabla^{n_{0,r}} \mu.$$

In order to derive an equivalent to the chemical master equation but for factorial moments, we use the inner product operation as before

$$\partial_t \mu_x = \left\langle \frac{z^x}{x!}, \partial_t \mu \right\rangle = \left\langle \frac{z^x}{x!}, \sum_{r \in R} k_r ((z+1)^{n_{1,r}} - (z+1)^{n_{0,r}}) \nabla^{n_{0,r}} \mu \right\rangle$$

$$= \left\langle \sum_{r \in R} ((\nabla + 1)^{n_{1,r}} - (\nabla + 1)^{n_{0,r}}) \frac{z^x}{x!}, \nabla^{n_{0,r}} \mu \right\rangle$$

$$= \left\langle \sum_{r \in R} \sum_{0 \le y \le n_{1,r}} k_r \binom{n_{1,r}}{y} \frac{z^{x-y}}{(x-y)!} - \sum_{0 \le y \le n_{0,r}} k_r \binom{n_{0,r}}{y} \frac{z^{x-y}}{(x-y)!}, \nabla^{n_{0,r}} \mu \right\rangle$$

$$= \sum_{r \in R} \left(\sum_{0 \le y \le n_{1,r}} k_r \binom{n_{1,r}}{y} \frac{(x-y+n_{0,r})!}{(x-y)!} \mu_{x-y+n_{0,r}} - \sum_{0 \le y \le n_{0,r}} k_r \binom{n_{0,r}}{y} \frac{(x-y+n_{0,r})!}{(x-y)!} \mu_{x-y+n_{0,r}} \right).$$

Finally, for the factorial moments, we obtain

$$\partial_t m_x = \sum_{r \in R} \left(\sum_{0 \le y \le n_{1,r}} k_r \binom{n_{1,r}}{y} \frac{x!}{(x-y)!} m_{x-y+n_{0,r}} - \sum_{0 \le y \le n_{0,r}} k_r \binom{n_{0,r}}{y} \frac{x!}{(x-y)!} m_{x-y+n_{0,r}} \right).$$

3.6 Assembly

We will now focus on CRNs where there is a well defined notion of composition and energy of species. The composition we see as consisting of a number of atoms of different types and reactions as rearranging atoms in a mixture, but never creating them or destroying them. We assume that the energy gives rise to a complex balance system, which then enables us to compute the factorial moments.

An *assembly system* is a complex balanced CRN (S, R, n, k, c) with a set of atoms A and an *atomic content* function $\alpha : S \longrightarrow \mathbb{N}^A$ with the following conservation property

$$\sum_{s \in S} n_{0,r,s} \alpha_s = \sum_{s \in S} n_{1,r,s} \alpha_s,$$

for each reaction $r \in R$. For $x \in \mathbb{N}^S$, we define $\alpha x \in \mathbb{N}^A$ as follows

$$\alpha x = \sum_{s \in S} x_s \alpha_s.$$

With this notation, the conservation property looks like $\alpha n_{0,r} = \alpha n_{1,r}$.

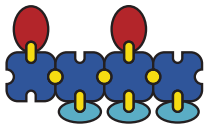
For a multiset $\tau \in \mathbb{N}^A$ of total atoms, we consider a *conservation class* Ω_τ consisting of all multisets $x \in \mathbb{N}^S$ whose total of atoms is equal to τ

$$\Omega_\tau = \{x \in \mathbb{N}^S : \alpha x = \tau\}.$$

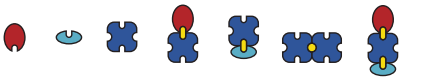
Notice that the conservation of atoms in assembly systems implies that reactions will never lead out of a given conservation class. A different question is whether it is possible to use reactions to reach each mixture in a conservation class.

For multisets $x, y \in \mathbb{N}^S$, if there exists a finite sequence of reactions that lead from x to y, we say that y is *reachable* from x and we write $x \twoheadrightarrow y$. We say that an assembly system is *thorough* if for each $\tau \in \mathbb{N}^A$, we have that for all $x, y \in \Omega_\tau$, $x \twoheadrightarrow y$. Then an assembly system is thorough if from each mixture, we can reach each state in its conservation class.

We will now focus on stationary dynamics restricted to a given conservation class. By definition, assembly systems satisfy complex balance, which means that they have a stationary dynamics of exponential form. This stationary dynamics, however, will be a mixture of stationary dynamics of all conservation classes. Since we want the stationary dynamics of specific conservation classes, we need something better. For that purpose we introduce a new set of formal variables, one for each atom type, which keep track of the atomic contents of species, and then harness them in order to extract specific stationary dynamics from the general exponential solution.

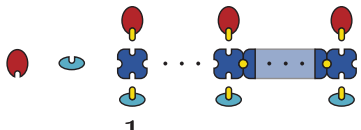


$$= \alpha^2\beta^3\sigma^3 a^2 b^3 s^4$$



$$W = \cdots$$

$$= a + b + s + \alpha as + \beta bs + \sigma s^2 + \alpha\beta abs + \cdots$$



$$= a + b + \sum_{n=1}^{\infty} \sigma^{n-1} s^n (1+\alpha a)^n (1+\beta b)^n$$

$$= a + b + \frac{s(1+\alpha a)(1+\beta b)}{1 - \sigma s(1+\alpha a)(1+\beta b)}$$

Figure 3.6: Example of a conservation class function. Here the conservation class function is denoted with W instead of ω. The function W was featured in Chapter 1 and it is the generating function of scaffold complexes. Each variable is color-coded to correspond to a unit in the structures.

We define the *conservation class function* $\omega \in \mathbb{R}[[z, w]]$, where $w = (w_a)_{a\in A}$ is an A-tuple of formal variables, as follows

$$\omega = e^{czw^\alpha},$$

where $w^\alpha = (\prod_{a\in A} w_a^{\alpha_{s,a}})_{s\in S}$. If we expand the conservation class function, we obtain

$$\omega = e^{czw^\alpha} = \sum_{x\in\mathbb{N}^S} \frac{z^x}{x!} c^x w^{\alpha x} = \sum_{\tau\in\mathbb{N}^A} \sum_{x\in\Omega_\tau} \frac{z^x}{x!} c^x w^\tau.$$

For each $\tau \in \mathbb{N}^A$, we define the following function

$$\omega_\tau = \sum_{x\in\Omega_\tau} \frac{z^x}{x!} c^x.$$

We can therefore write the conservation class function as follows

$$\omega = \sum_{\tau\in\mathbb{N}^A} \omega_\tau w^\tau.$$

Notice that for each $x \in \mathbb{N}^S$, we can use conservation of atoms and complex balance to obtain

$$\sum_{\substack{r\in R:\\ n_{1,r}=x}} k_r c^{n_{0,r}} w^{\alpha n_{0,r}} - \sum_{\substack{r\in R:\\ n_{0,r}=x}} k_r c^{n_{0,r}} w^{\alpha n_{0,r}} = w^{\alpha x} \left(\sum_{\substack{r\in R:\\ n_{1,r}=x}} k_r c^{n_{0,r}} - \sum_{\substack{r\in R:\\ n_{0,r}=x}} k_r c^{n_{0,r}} \right) = 0,$$

which implies that $\mathscr{A}\omega = 0$. We therefore have that

$$\mathscr{A}\omega = \sum_{\tau\in\mathbb{N}^A} \mathscr{A}\omega_\tau w^\tau = 0,$$

and therefore $\mathscr{A}\omega_\tau = 0$ for all $\tau \in \mathbb{N}^A$. We conclude that each ω_τ is a stationary dynamics. Furthermore, if the assembly system is thorough, then ω_τ is irreducible.

Applying the factorial moment operator to the conservation class function, we obtain

$$\mathscr{M}\omega = e^{c(z+1)w^\alpha} = e^{czw^\alpha}e^{cw^\alpha} = \sum_{x\in\mathbb{N}^S}\sum_{\tau\in\mathbb{N}^A} \frac{z^x}{x!}c^x w^{\alpha x}\langle e^z, \omega_\tau\rangle w^\tau$$

$$= \sum_{\tau\in\mathbb{N}^A}\sum_{x\in\mathbb{N}^S} \frac{z^x}{x!}c^x\langle e^z, \omega_{\tau-\alpha x}\rangle w^\tau.$$

Therefore, the moment function of a conservation class Ω_τ has the form

$$\mathscr{M}\omega_\tau = \sum_{x\in\mathbb{N}^S} \frac{z^x}{x!}c^x\langle e^z, \omega_{\tau-\alpha x}\rangle.$$

The factorial moment of order x for conservation class Ω_τ is therefore

$$m_{\tau,x} = c^x\frac{\langle e^z, \omega_{\tau-\alpha x}\rangle}{\langle e^z, \omega_\tau\rangle}.$$

3.7 Discussion

In *The barrier of objects: From dynamical systems to bounded organizations* (Fontana and Buss, 1996), the authors propose that the traditional dynamical systems approach to the study of complex, and, in particular, biological systems is inadequate because it is unable to address the fact that the objects studied cannot be reduced to numerical values given that their structural nature is dynamic and central to their function. In their words: "In Nature, interaction involves objects directly and never by a numerical value describing them. Stepping outside of conventional dynamical systems requires taking this observation seriously." This thesis and especially this chapter are the beginning of an attempt at "taking their observation seriously."

The first step in my approach to taking their observation seriously is by shifting focus from numerical evaluation of functions into structural analysis of expressions by treating variables as formal rather than numerical. The result is that of a foundation based on the concept of generating functions, or formal power series. This is the approach taken in this chapter. This is nevertheless not quite the "direct interaction of objects" that Fontana and Buss observe. A formal power series approach is a stepping stone into a theory that takes objects and their structure seriously and works with them directly rather than with their numerical projections.

As an example, let us consider finite sets and arithmetic. If we focus on the part of arithmetic concerned with the study of positive integers, with the operations of addition and multiplication, we can see that arithmetic is like a shadow of the theory of finite sets and the functions between them. Sets also have operations of addition and multiplication, but they usually go by the names of *disjoint union* and *Cartesian product*. Take, for example, sets $A = \{a, b, c\}$ and $B = \{a, \alpha\}$, their disjoint union and Cartesian product are given by the following sets

$$A \sqcup B = \{(1, a), (1, b), (1, c), (2, a), (2, \alpha)\}, \quad A \times B = \{(a, a), (a, \alpha), (b, a), (b, \alpha), (c, a), (c, \alpha)\}.$$

Notice that, despite A and B having one element in common, labeling guarantees that they are mapped to different elements in the disjoint union. At the end, we have that the cardinality of a disjoint union is the sum of cardinalities of the summands: $|A \sqcup B| = |A| + |B|$; and the cardinality of a Cartesian product is the product of the cardinalities of the multiplicands: $|A \times B| = |A||B|$. We can thus see that working with finite sets instead of natural numbers is a way of "working with objects directly rather than their numerical values." The picture of finite sets is richer than that of arithmetic as, for example, there is only one number 5 but many sets with cardinality 5. Also, two natural numbers are either equal to one another or not, i.e. equality is an equivalence relation. In contrast, two finite sets with the same cardinality are isomorphic to one another in potentially many ways, $n!$ ways for a set with cardinality n to be precise. The latter observation gives rise to the concept of a *bijective proof* (Loehr, 2011), which is a way of establishing the equality between two expressions denoting natural numbers by finding a bijection between two sets with cardinalities given by the expressions in question.

The problem of using sets as representatives of objects in chemistry, biology, or complex systems in general is that they lack any structure beyond the numerosity of their elements. This can be seen by the fact that two sets are isomorphic precisely when they have the same number of elements. In order to have a mathematical foundation that is more adequate for developing mathematical biology, we need the foundational objects to be closer to the objects of biology. One possible such foundations is homotopy type theory (HoTT). The purpose of this chapter is to prime SCRN theory for formulation in HoTT. It achieves that by formulating SCRN theory in terms of formal power series. Then, by employing existing proposals (Yorgey, 2014) for using HoTT to give formal power series meaning in terms of homotopy types, we can lift SCRNs to the level of HoTT.

Within HoTT, it is also possible to add, multiply, and exponentiate objects as it is possible for sets; however, division is also possible, which it is not the case for sets. Furthermore, the cardinalities of homotopy types span all positive real numbers rather than just the positive integers such as is the case for finite sets. I believe that using HoTT as a foundation for SCRN theory will bring us closer to taking seriously the observation made by Fontana and Buss.

One advantage of formulating SCRNs in HoTT is that the language is an *extension* rather than a *replacement* of the language of sets. What this means is that one can continue to use the usual language in which SCRNs are formulated, for example, to speak of sets of species and reactions, but lurking in the background lie a myriad of features ready for use on demand. Those features include the ability to speak of *spaces* of species and reactions, which would be appropriate when those are more structures, such is the case for virtually all biological systems.

In addition to the topological features of HoTT, another advantage is that HoTT is, as its name states, a *type theory*, which means that it can be given computational content. Formulating SCRN theory in the language of HoTT would make it possible to study questions of molecular and biological computing by exploiting its computational content.

BIBLIOGRAPHY

Anderson, David, Gheorghe Craciun, and Thomas Kurtz (2010). "Product-form stationary distributions for deficiency zero chemical reaction networks". In: *Bulletin of Mathematical Biology* 72.8, pp. 1947–1970. DOI: 10.1007/s11538-010-9517-4.

Baez, John and Jacob Biamonte (2018). *Quantum Techniques in Stochastic Mechanics*. World Scientific. DOI: 10.1142/10623.

Baez, John and James Dolan (2001). "From Finite Sets to Feynman Diagrams". In: *Mathematics Unlimited — 2001 and Beyond*. Ed. by Björn Engquist and Wilfried Schmid. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 29–50. DOI: 10.1007/978-3-642-56478-9_3.

Behr, Nicolas, Gerard HE Duchamp, and Karol A Penson (2017). "Combinatorics of chemical reaction systems". In: *arXiv preprint arXiv:1712.06575*.

Bergeron, François, Gilbert Labelle, and Pierre Leroux (1997). Trans. by Margaret Readdy. Encyclopedia of Mathematics and its Applications. Cambridge University Press. DOI: 10.1017/CBO9781107325913.

Brown, Ronald (1987). "From Groups to Groupoids: a Brief Survey". In: *Bulletin of the London Mathematical Society* 19.2, pp. 113–134. DOI: 10.1112/blms/19.2.113.

Cappelletti, Daniele, Andrés Ortiz-Muñoz, et al. (2020). "Stochastic chemical reaction networks for robustly approximating arbitrary probability distributions". In: *Theoretical Computer Science* 801, pp. 64–95. DOI: 10.1016/j.tcs.2019.08.013.

Cappelletti, Daniele and Carsten Wiuf (2016). "Product-form poisson-like distributions and complex balanced reaction systems". In: *SIAM Journal on Applied Mathematics* 76.1, pp. 411–432. DOI: 10.1137/15M1029916.

Chaiken, S and D.J Kleitman (1978). "Matrix Tree Theorems". In: *Journal of Combinatorial Theory, Series A* 24.3, pp. 377–381. ISSN: 0097-3165. DOI: 10.1016/0097-3165(78)90067-5.

Flajolet, Philippe and Robert Sedgewick (2009). *Analytic Combinatorics*. Cambridge University Press. ISBN: 9781139477161.

Fontana, Walter and Leo W. Buss (1996). "The barrier of objects: From dynamical systems to bounded organizations". In: *Boundaries and Barriers*. Ed. by John Casti and Anders Karlqvist. Reading MA: Addison-Wesley, 56–116.

Joyal, André (1981). "Une théorie combinatoire des séries formelles". In: *Advances in Mathematics* 42.1, pp. 1–82. ISSN: 0001-8708. DOI: 10.1016/0001-8708(81)90052-9.

Krishnamurthy, Supriya and Eric Smith (2017). "Solving moment hierarchies for chemical reaction networks". In: *Journal of Physics A: Mathematical and Theoretical* 50.42, p. 425002. DOI: 10.1088/1751-8121/aa89d0. URL: https://doi.org/10.1088/1751-8121/aa89d0.

Leighton, F.T. and R.L. Rivest (1983). *The Markov Chain Tree Theorem*. MIT/LCS/TM. Laboratory for Computer Science, Mass. Inst. of Technology.

Loehr, Nicholas (2011). *Bijective combinatorics*. CRC Press.

McQuarrie, Donald A (1967). "Stochastic approach to chemical kinetics". In: *Journal of applied probability* 4.3, pp. 413–478.

Moya-Cessa, Héctor and Francisco Soto-Eguibar (2011). *Differential equations: an operational approach*. Rinton Press, Incorporated. ISBN: 1-58949-060-4.

Øksendal, Bernt (2010). *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer Berlin Heidelberg.

Smadbeck, Patrick and Yiannis Kaznessis (2012). "Efficient moment matrix generation for arbitrary chemical networks". In: *Chemical Engineering Science* 84, pp. 612–618. ISSN: 0009-2509. DOI: 10.1016/j.ces.2012.08.031.

– (2013). "A closure scheme for chemical master equations". In: *Proceedings of the National Academy of Sciences* 110.35, pp. 14261–14265. ISSN: 0027-8424. DOI: 10.1073/pnas.1306481110.

Smith, Eric and Supriya Krishnamurthy (2017). "Flows, scaling, and the control of moment hierarchies for stochastic chemical reaction networks". In: *Phys. Rev. E* 96 (6), p. 062102. DOI: 10.1103/PhysRevE.96.062102.

– (2021). "Eikonal solutions for moment hierarchies of chemical reaction networks in the limits of large particle number". In: *Journal of Physics A: Mathematical and Theoretical* 54.18, p. 185002. DOI: 10.1088/1751-8121/abe6ba.

Sotiropoulos, Vassilios and Yiannis Kaznessis (2011). "Analytical derivation of moment equations in stochastic chemical kinetics". In: *Chemical Engineering Science* 66.3, pp. 268–277. ISSN: 0009-2509. DOI: 10.1016/j.ces.2010.10.024.

Univalent Foundations Program, The (2013). *Homotopy Type Theory: Univalent Foundations of Mathematics*. Institute for Advanced Study: https://homotopytypetheory.org/book.

Weinstein, Alan (1996). "Groupoids: unifying internal and external symmetry". In: *Notices of the AMS* 43.7, pp. 744–752.

Whittle, Peter (1986). *Systems in Stochastic Equilibrium*. Probability and Statistics. Chichester: Wiley.

Wilf, Herbert S. (1994). In: *generatingfunctionology*. Ed. by Herbert S. Wilf. 2nd ed. San Diego: Academic Press. ISBN: 978-0-08-057151-5. DOI: 10.1016/B978-0-08-057151-5.50003-4.

Yorgey, Brent (2014). "Combinatorial species and labelled structures". PhD thesis. University of Pennsylvania.