

On the kinetics and thermodynamics of templated
polymerisation:
Graphs, paths, and proofreading motifs

Benjamin Javed Qureshi
Imperial College London
Department of Bioengineering
PhD Thesis

Statement of originality

This thesis is the work of Benjamin Javed Qureshi. Anything which is not the work of Benjamin Javed Qureshi is clearly referenced.

Copyright declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law

Acknowledgements

I would like primarily like to thank my supervisor, Thomas Ouldrige, for his exceptional support both academically and pastorally and for his generosity.

I would also like to thank my past and present colleagues in the Principles of Biomolecular Systems group. In particular, I'd like to thank Jenny Poulton, Jordan Juritz, Jeremy Guntoro, and Alex Dack for many insightful conversations and problem solving sessions. Additionally, I would like to thank Rakesh Mukherjee and Križan Jurinovic for teaching me everything about the experimental side of the work in the group.

I would like to give particular mention to my friends and fellow PhD students Tomasz Procter and Alexander Yip. Our group chat was vital for the completion of this PhD.

Finally, I'd like to thank my family, particularly my mother Angela Qureshi and her partner Robert Chew, and my father Javed Qureshi; and friends outside of academia especially Louis Metrebian. Finally, I'd like to thank dogs, Harper and Bella; cat, Poppy; and lizard, Blueberry for their essential emotional support.

Abstract

Biological systems use template molecules, such as DNA or mRNA, to catalyse the production of specific, sequence dependent polymers, such as proteins. Templated polymerisation is a powerful mechanism for the production of complex molecules, which can store information. As such, templated polymerisation is both widely utilised in biology, and promises to be useful for synthetic systems.

Templated polymerisation is used to create daughter polymers with specific sequences. The newly created daughter polymer needs to be separated from the template, meaning no long-term interactions with the template may bias the distribution of daughter polymers. This observation makes templating a highly non-trivial process, with interesting thermodynamic consequences. Templated polymerisation processes can have complex reaction steps leading to monomer incorporation, such as kinetic proofreading, to improve their accuracy. These complex sub-steps often involve coupling to fuel reservoirs, leading to trade-off between free-energy costs, speed and accuracy. Further, these sub-steps can lead to correlations between monomers within the daughter copolymer.

In this thesis, I first introduce a framework for solving a broad class of detailed polymerisation models. This framework allows, for the first time, for the solving of thermodynamically consistent, complex polymerisation models with dependence on monomer incorporation rates on previous monomers in the growing polymer. After having studied detailed polymerisation models with this framework, I zoom out to study a class of enzymatic templating systems, focusing on the effects of multiple pathways leading to creation/degradation of polymers. Here, I derive fundamental bounds on the accuracy of these multi-pathway polymerisation systems as a function of the free-energies changes along the pathways.

Contents

1	Introduction to molecular self-assembly and templating <i>in vivo</i> and <i>in vitro</i>	8
1.1	The need for self-assembly	8
1.2	Synthetic self-assembly	10
1.3	Templating in biology	13
1.4	Synthetic templating	16
1.5	Thesis overview	19
2	Background theory	21
2.1	Information theory	21
2.2	Markov chains, graph theory and spanning trees	26
2.2.1	The matrix tree theorem	26
2.2.2	Markov chains	27
2.2.3	Collapsing nodes	29
2.2.4	The Markov chain tree theorem	31
2.2.5	Absorbing Markov chains	34
2.2.6	Expectations of an absorbing process are steady-state averages of a “closed process”	35
2.2.7	Steady state currents	36
2.2.8	Absorbing probabilities	38
2.2.9	Counting edge and cycle transitions	39

2.2.10	Gillespie’s algorithm	43
2.3	Chemical reaction networks	44
2.3.1	Definitions and notation	44
2.3.2	Deterministic CRNs	45
2.3.3	Stoichiometry	47
2.3.4	Stochastic CRNs	50
2.3.5	Graphical features of CRNs	52
2.4	Stochastic thermodynamics	55
2.4.1	Generalised free energy and entropy production	56
2.4.2	Time evolution and the second law	57
2.4.3	Entropy production and cycles	59
3	Copolymerisation models	62
3.1	Types of copolymerisation	62
3.2	Free copolymerisation	65
3.3	Templated polymerisation models	67
3.4	Kinetic proofreading	69
3.5	Solving polymerisation models	74
3.5.1	Thermodynamics of copolymerisation	78
3.6	Simple copolymerisation with separation of template and product	79
3.7	Outline of the rest of the thesis	81
4	A method for the analysis of generic polymerisation models	82
4.1	Introduction	82
4.2	Identifying propensities of coarse-grained models	85
4.3	Extracting properties of the fine-grained model from the solution of the coarse-grained model	89
4.3.1	Stalled growth	95
4.3.2	Limiting behaviour	96
4.3.3	Simplification for factorisable propensities	99

4.4	Example applications	101
4.4.1	Stalling behaviour in a polymerisation model with no neighbour-neighbour interactions	101
4.4.2	Reproducing the trivial internal reaction network polymerisation results	103
4.4.3	Balanced models of templated polymerisation with autonomous separation	105
4.4.4	Hopfield’s kinetic proofreading in a model of templated copying with autonomous separation	110
4.4.5	A multi-loop generalisation of Hopfield’s kinetic proofreading . . .	117
4.5	Conclusion	123
5	Bounds on the accuracy of catalytic templating networks with multiple pathways	125
5.1	Introduction	125
5.2	Models and methods	131
5.2.1	Physical model	131
5.2.2	Graphical methods for linear CRNs with a null complex	133
5.2.3	Information theoretic measures of product distributions	135
5.2.4	Simulation of specific networks	137
5.3	Free-energy differences between pathways bound accuracy for a simple network with two catalysts	137
5.4	General bounds on the properties of product distributions	141
5.4.1	Bounds on the concentrations of products	142
5.4.2	Bounds on the maximum probability and minimum entropy of the product distributions	143
5.4.3	The bounds in the presence of kinetic proofreading	146
5.5	Physical significance of the bounds	147
5.5.1	Entropy minimisation versus specificity maximisation	147

5.6	Relationship of the bounds to previously postulated costs	153
5.7	The achievability of the bounds	156
5.8	Examples	157
5.8.1	Example 1: Templated polymerisation with a destructive catalyst	157
5.8.2	Example 2: Hopfield-like kinetic proofreading with a destructive catalyst	162
5.9	Conclusion	169
6	Conclusion	177
6.1	Final remarks	181

Chapter 1

Introduction to molecular self-assembly and templating *in vivo* and *in vitro*

1.1 The need for self-assembly

Biological systems are incredibly complex, consisting of many, highly specific molecular structures and machines, which perform highly specific jobs. However, biological systems exist in the physical world and therefore are subject to the same physical laws as any non-living system. In particular, biological systems need to be able to produce these specific structures in messy, relatively high temperature environments, where thermal fluctuations are highly relevant [1]. Importantly, these biological structures need to form without the aid of a sentient, macroscopic agent. One such mechanism biological systems can employ to create these structures is self-assembly, whereby the constituent components self-assemble into the desired molecules and structures. Self assembly is common in biology, allowing, for example, for the formation of cell membranes, ribosomes [2], and bacterial flagella [3].

A prototypical and easy to grasp example of such self-assembly is that of virus capsids

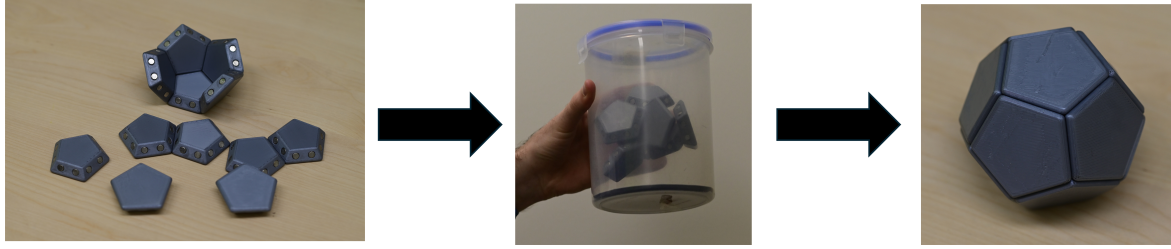


Figure 1.1: A physical, 3D printed model of virus capsid self assembly. This model was 3D printed to reproduce the model of [4]. Individual sub-units have magnets embedded in them. When the sub-units are put into a container and shaken, they will assemble into the complete dodecahedron.

(figure 1.1), whereby protein complexes come together to form the “shell” of some viruses which encloses their RNA. Figure 1.1 shows a toy model for simulating the self-assembly process at a macroscopic scale. Here, sub-units, in the form of 3D-printed pentagonal structures, are implanted with magnets such that faces of the sub-units with magnets can only stick to each other in one orientation. Then, by shaking the sub-units in a container to simulate the random motion of a thermally excited environment, the sub-units will collide with each other, their magnetic faces will stick and the final, dodecahedral structure can be formed. Similarly, certain viruses [5] consist of protein sub-units which have specific interactions with each other. In the noisy environment, in which the virus is forming, the protein sub-units undergo random motion, colliding with each other. When they collide, they may stick together due to the specific interactions and, if enough sub-units coalesce, they may form a closed shell - the virus capsid. Virus capsid self-assembly in this way has two important features which aid its function. Firstly, the process consists of a number of sub-units coming together to form a product. Thus, the product is able to be encoded as the lowest energy state of the sub-units and as such, the self-assembly may proceed as the approach to equilibrium of the inputs [6]. Secondly, virus capsids of this kind have a simple role, being to enclose the RNA, as such they are able to be highly symmetric, being a platonic solid [7], and so may consist of only a single type, or a few types, of repeating, identical sub-units.

1.2 Synthetic self-assembly

Engineers and scientists also wish to make use of self-assembling systems in order to create, more easily, complex structures at small scales. Such complex small-scale structures have generated much excitement and been widely studied for their potential applications as nanomachines [8, 9], and in medicine, in particular as drug delivery systems [10–12].

A simple set of examples utilising self-assembly mechanisms would include synthetic copolymers, such as rubbers and plastics [13, 14]. Copolymers are polymers consisting of more than one monomer type. One method of forming such copolymers is through free copolymerisation, whereby the different monomer types may be mixed in solution to form growing polymer chains [15]. Similarly to the biological examples, such systems are able to self-assemble into specific copolymers, due to specific interactions between monomers. In this case, the specific interactions take the form of certain reactive groups on the monomers. For such chemistries, forming simple polymers, such as homopolymers with only one monomer type or alternating polymers, where a monomer of type one is followed by a monomer of type two which is followed by a monomer of type one *etc.*, is relatively simple [16], but forming more specific sequences of monomers accurately requires more monomer types.

For a more modern setting, much of the field of DNA nanotechnology is built on the concept of self-assembly [17]. DNA serves as a useful basis for synthetic self-assembling systems due to the specificity of Watson-Crick base pairs [17]. In the early 1980s, Seeman and his group were able to produce synthetic, rigid, four-strand complexes known as Holliday junctions [18], as in figure 1.2(a). Along side, the observation that these junctions could be combined into square “tiles” with overhanging DNA domains [19] (figure 1.2(b,c)), Seeman was able to demonstrate that DNA could be shaped into a sub-unit with specific free binding sites. These works formed the basis of DNA nanotechnology and, in particular, tile assembly. Tile assembly relies on the use of DNA tiles, similar to those introduced by Seeman [19], which are two (or more) dimensional complexes of DNA with overhanging “sticky ends”, allowing for specific interactions between tiles [17].

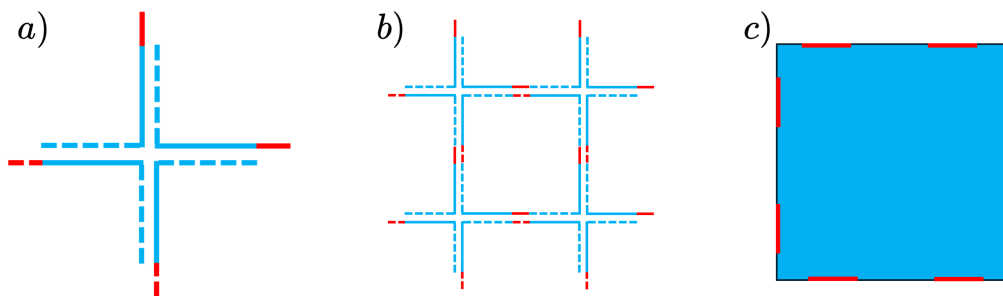


Figure 1.2: Depiction of (a) the Holliday junction with overhanging domains; (b) an assembly of four Holliday junctions arranged into a tile; (c) an abstraction of the DNA tile. For (a) and (b), DNA strands are depicted as lines. Different domains in said DNA strands are shown in different colours and as dashed or solid line segments. Dashed and solid line segments of the same colour depict complementary domains. To avoid over complication, only two colours have been used for the domains. The overhanging “sticky ends” of the DNA tiles are shown in red.

Since its inception, much interest has been paid both experimentally and theoretically to the types of structures that could be created from such tile assembly models [20–22].

Another major strategy for construction nano-structures was introduced by Rothemund who suggested that a long DNA “scaffold” strand could be used with a number of shorter “staple” strands to fold the scaffold strand into a desired shape [23], in a process he named DNA origami. Similar to tile assembly, DNA origami could be conducted as a one-pot experiment, whereby the components are added to the reaction environment and relax to their desired structure, encoded as the lowest energy state. However, where the components for tile assembly are of similar complexity, DNA origami consists of one highly complicated component, the scaffold, and many, less complicated components, the staples.

Given the autonomous nature of these synthetic DNA-based systems as well as the complexity of the target structures, self-assembly is not always simple. Problems such as misfolds and the formation of undesired structures as well as kinetic traps, whereby the system can be stuck in local energy minima for long periods of time, are relatively commonplace [24,25]. Such problems make designing effective, fast, high-yield synthetic

self-assembly systems challenging and non-trivial.

As has been hinted throughout this introduction thus far, the complexity of the desired shape and the complexity of the required input components are closely related. Synthesising complex input components can be difficult in its own right. As such, much effort has been devoted towards minimising the inputs required, both in number and complexity [26–28].

Aside from forming target structures, DNA has also shown promise as a mechanism for computation. The field of DNA computation began with Adleman [29] in the 1990s. Here, Adleman was able to encode a directed graph as a number of strands of DNA and utilise the base-pairing of the DNA to determine whether or not the encoded graph contained a Hamiltonian path. If a Hamiltonian path exists, the strands encoding the graph would assemble into a certain structure. The algorithm presented by Adleman was not autonomous, requiring many labour-intensive steps, and was designed to solve a single specific problem. However, the use of DNA to make a computation in this sense inspired the field of DNA computing. This interest in computing extended to self-assembly; in particular tile assembly models have had much theoretical attention paid to their ability to perform computations. Tracking more closely with traditional computation than the original DNA computation of Adleman [29], bits can be encoded as interactions between the tiles and thus the tiles present in the system can encode the steps of the computation [30]. The theoretical computer science community has been particularly interested in such models due to their Turing universality [31,32] (ability to simulate a Turing machine).

The introduction of DNA strand displacement reactions by Yurke *et al.* [33] further expanded the scope of what synthetic systems could be achieved. DNA strand displacement reactions are a class of reactions whereby a single-stranded DNA strand replaces one of the DNA strands in a double-stranded DNA complex. Such reactions are particularly useful in controlling the interactions of different DNA complexes [34]. Thus, such strand displacement reactions may serve as a medium for chemical implementation of neural networks [35,36].

1.3 Templating in biology

For all these self-assembled systems, the desired molecules, structures or computations are programmed into the components that make them up. Thus, simpler, more symmetric structures can be made with simpler sets of input molecules. Not all the complex macromolecules or structures in biology are able to have such symmetries. This lack of symmetries is true of the molecules key to life - DNA, RNA and proteins. Each of these molecules are copolymers of different monomers: nucleotides for DNA and RNA, and amino acids for proteins. The sequence of those monomers can be arbitrary and is vital to the functioning of these molecules.

Proteins are the main functional molecules of living organisms. Proteins mediate chemical reactions, provide structure, and allow for signalling in biological systems [2]. Interactions between the amino acids constituting the protein, as well as between the amino acids and the environment, in turn determine the shape and function of the protein. Thus, the specific sequence of amino acids is vital to the protein.

In the human body, there are estimated to be at least 20,000 different fundamental protein sequences (diversified further by chemical modifications to these proteins) necessary for our function [37]. As such, proteins are at the heart of biological complexity. However, like all biological molecules, proteins need to be assembled from their constituent monomers without a sentient architect. In order to self-assemble these molecules freely, when required, with mechanisms similar to those already described in this introduction, the cell would need a vast pool of specific sub-units. These sub-units would need to encode their relative position to other sub-units. If one does not have a vast number of different sub-units, one type of sub-unit would need to be in multiple places within the structure, with different neighbouring sub-units. Those different neighbouring sub-units would each need to have a binding site that could bind to the same sub-unit. Such a situation leaves open the opportunity for sub-units to bind in the wrong places [24,28,38].

Biological systems are not able to create proteins by pure self-assembly. Instead, they use templating to create the proteins. Templates are separate molecules which encode the

information required to create the sequence of amino acids in the protein. For proteins, these molecules are strands of RNA called mRNAs. Thus, the amino acids themselves do not need to contain this information. Using templates, biological systems are able to make all the required proteins from only a set of 20 amino acids [2] in a process called translation.

The chains of amino acids produced from translation need to fold into a protein. The shape of the folded chain is encoded into interactions between the amino acids. Incorrect binding of amino acids is known in the protein setting as misfolding. Misfolding could potentially be a large issue due to the small number of amino acid types relative to the number of proteins required. However, the specific, one-dimensional sequence of the amino acid sequence is also one of the factors assisting the correct folding of the three dimensional protein structure [28], as it restricts which amino acids in the chain may interact in a given protein.

The cell further complicates this story of producing proteins from templates by asking that the templates for proteins (mRNAs) also be created first from templated reactions from DNA in a process called transcription. DNA is the medium the cell uses to store the information for creating proteins. This full process of information in DNA being transcribed into mRNAs, whose information is then translated into proteins (figure 1.3) is so key to our understanding of life, it is often given the grandiose title of “the central dogma of molecular biology” [39]. In such biological templating reactions, the template molecules act as catalysts, and may be reused such that one template molecule may produce many downstream molecules. This observation is key as, if a new template molecule were needed for each downstream molecule, the cell would not be able to store the information for the proteins anywhere. The cell would need to self-assemble a new template, freely, for each individual protein; and thus, would still require a vast pool of specific sub-units to assemble said templates [40].

That the templates need to be reusable is both vitally important to the function of templating systems and has significant effects on their dynamics and thermodynamics. However, this detail is often overlooked in the literature. Any interactions between the

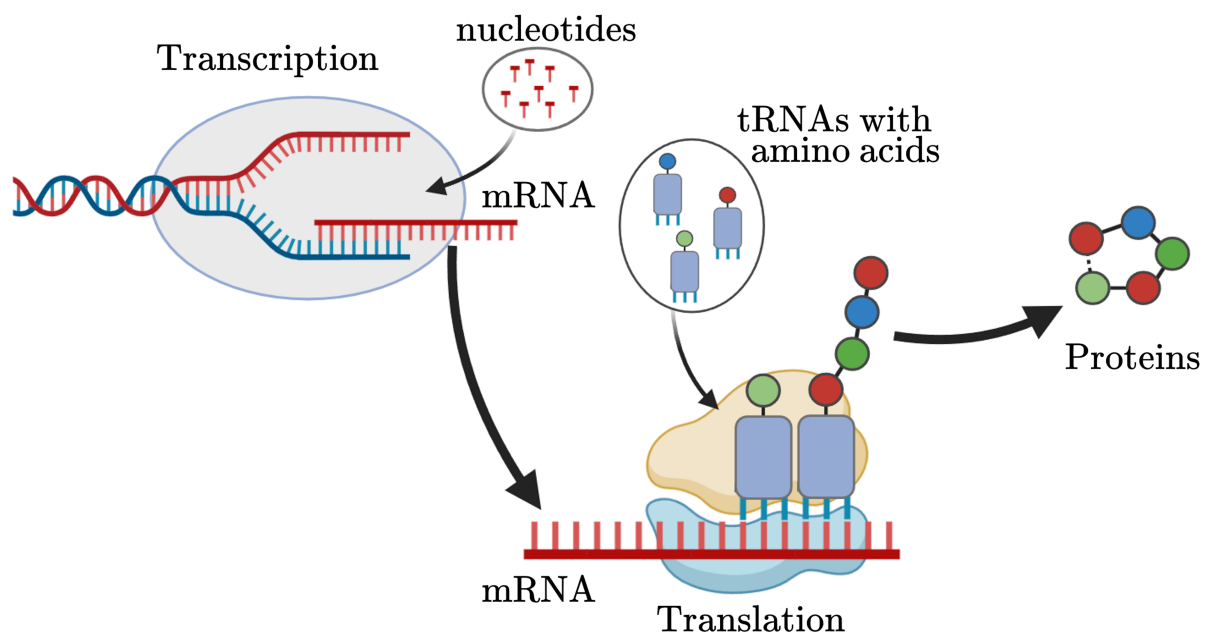


Figure 1.3: Depiction of the templated polymerisation information transfer in the central dogma of molecular biology. In the transcription, a section of DNA (a gene) acts as a template for the polymerisation of RNA nucleotides into an mRNA. This mRNA then acts as a template for the polymerisation of amino acids (with the use of tRNAs) into a protein in a process called translation. The protein can fold into its 3D structure due to amino acid-amino acid interactions (represented here by the dashed line).

template and product must be fleeting. Any bonds formed between template and product must be broken. As such templates cannot affect the stability of products.

The self-assembly mechanisms described in the sections 1.1, 1.2 allow the system to relax to an equilibrium encoding the desired, structure, molecule or computation. Such a methodology is not possible for templating systems. Often the different copolymers are of similar stability (*e.g.* strands of mRNA of equal length). More prominently, the cell needs to use the same mechanism to create different proteins, regardless of their free energy. Thus, it is not possible to encode the target copolymers as the minimum energy state. And so, templating cannot produce generic products accurately, while at thermodynamic equilibrium. In order to produce specific molecules accurately and reliably, these templated mechanisms must therefore be far from equilibrium [40]. This fact makes studying the physics of such processes more difficult and means they are far less understood. The non-equilibrium nature of templating is explored more concretely in chapters 4 and 5.

1.4 Synthetic templating

Templating is also exciting for its prospects as a mechanism for making non-symmetric, synthetic molecules or structures, such as proteins not found in nature, or any other molecules which might be too complex to assemble by other means [41]. Additionally, templating may allow for the discovery of new therapeutic molecules [42, 43]. The ability to utilise templating will likely stem from first being able to reproduce biological templating mechanisms fully synthetically.

Commercial methods for producing synthetic polynucleotides or proteins exist. However, these methods are often highly non-autonomous and therefore labour and time demanding as well as producing low yields for longer molecules. For example, common commercial methods include solid state synthesis methods [44, 45], whereby monomers (nucleotides or amino acids) are added one by one to a growing chain. These monomers are partially blocked to ensure only one monomer binds to the polymer at a time [46].

In between monomer addition steps, purification and washing steps are involved to remove excess unreacted monomers. Efforts to automate these processes often involve flows and time varying conditions [46, 47]. Other, more autonomous methods such as cell-free protein synthesis [48] extract enzymes, ribosomes *etc.* from live cells to make the proteins and as such, although practically useful, fall outside the scope of synthetic protein synthesis. Our inability to create fully synthetic, autonomous templated polymerisation reactions, similar to those occurring in the cell shows we do not understand the role of all the components, such as enzymes and ribosomes, present in said cellular reactions.

The desire to understand such enzyme-free templating reactions has led to recent experimental efforts to produce such systems [49–54]. Synthetic templating mechanisms to produce dimers have been experimentally synthesised [55, 56]. However, progress in producing longer polymers has been slow due to the problem of product inhibition [57], whereby the products block the templates slowing down reactions. Product inhibition further emphasises the importance of separation of products and templates.

Let us focus here on the work of Cabello *et al.* [58] during their time working on the experimental side of our research group. It is worth discussing in some detail the mechanisms used here by Cabello *et al.* as they form part of the inspiration for some of our theoretical models and intuitions. [58] describes a templated dimerisation reaction making use of DNA strand displacement reactions [59]. Strand displacement reactions, as depicted in figure 1.4(a), involve a DNA duplex and are the displacement of one of the strands of the duplex by another DNA strand. Cabello *et al.* introduce a new strand displacement mechanism they call handhold mediated strand displacement which they utilise to template the reaction of dimers as depicted in figure 1.4(b). Here, the monomers are themselves strands of DNA with multiple domains. Domains are short segments of a DNA strand, containing a number of nucleotides. Further, for the dimers produced, the backbone of the molecule is a section of double stranded DNA. Additionally, it is not difficult to see how one might extend such a mechanism to templating of longer molecules [60], although the detailed chemistry is more subtle. A key aspect of the polymerisation mechanism here introduced by Cabello *et al.* is that the formation of the

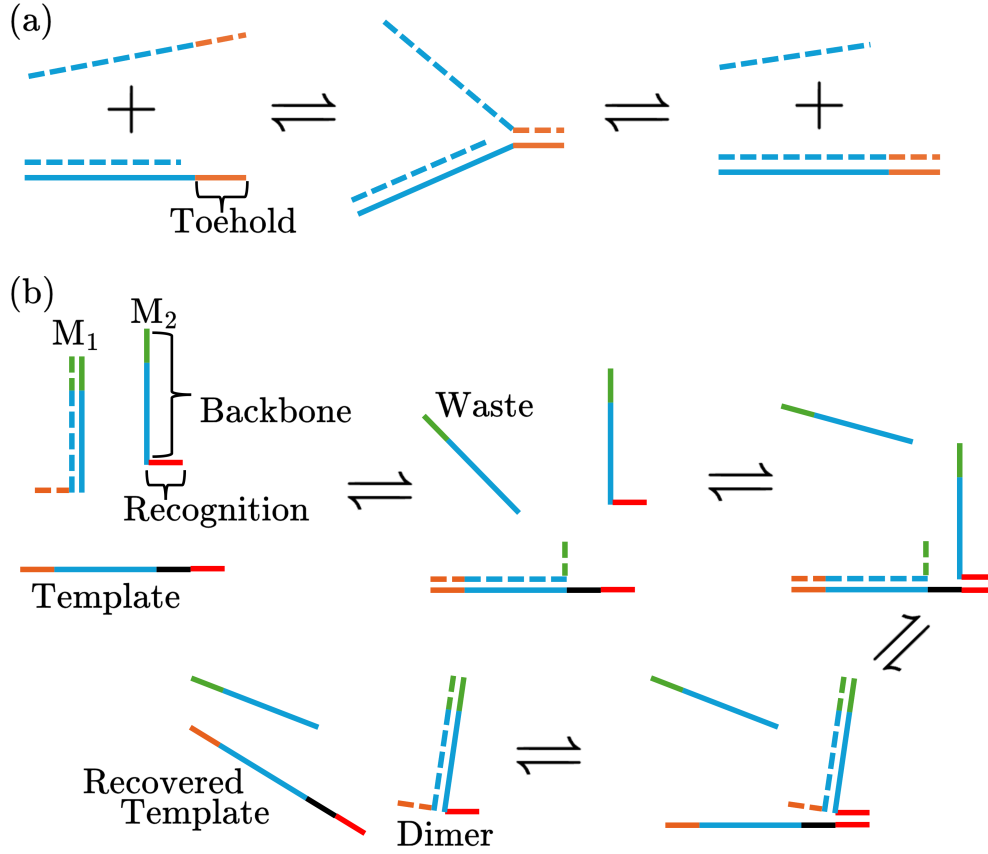


Figure 1.4: (a) A prototypical (toehold mediated) strand displacement reaction. As in figure 1.2, DNA strands are depicted as lines with different line segments representing domains. During toehold mediated strand displacement here, the orange toehold domain binds, then the blue domain displaces the strand already bound in the duplex. (b) A schematic of Cabello *et al.*'s templated dimerisation reaction [58]. The two monomers are labelled M_1 and M_2 , with orange and red recognition domains respectively. The backbone of the dimer consists of a green handhold domain and a blue branch migration domain. After the reaction, a dimer and a waste molecule have been produced and the template has been recovered.

backbone bond of the polymer is coupled to the detachment of the previous monomer from the template. Such a mechanism imposes a directionality to the copolymerisation. More importantly, this mechanism makes it easier for fully formed polymers to detach from the template.

Understanding enzyme-free templated polymer synthesis is also interesting from a pure scientific perspective. For the question of the origin of life, the earliest chemical soups would have needed self-replication mechanisms before the evolution of enzymes [61]. On the other end of the spectrum, self-replication is one of the defining features of life [62] and so, any attempt to imagine artificial living systems will likely feature such enzyme-free templated reactions.

Synthetic templating systems, as with DNA tiles, have much potential for computation. At some level, one can think of the central dogma as a two step computation, where there is an input of a gene, the gene is processed to give an mRNA and that mRNA is processed to output a protein. One can imagine then a multi-step synthetic computation system inspired by the central dogma. In such a system, each step of the computation might output a molecule that acts as a template for more, downstream reactions. Such computation systems might be particularly useful for “smart therapeutics”, where, *in vivo*, the system could compute and output an appropriate drug cocktail. From a theoretical perspective templated reactions seem to be an exciting medium for studying the thermodynamics of computation. As we shall see through this thesis, detailed models amenable to analytic or numeric mathematical calculations are possible.

1.5 Thesis overview

Thus, as we have seen in sections 1.3 and 1.4, progress has been limited, both from a biological and synthetic perspective, to understand templating. This lack of understanding suggests there is a need to expand the theory of templating further. As such, the aims of this thesis are to develop, in more detail, fundamental models and accompanying theoretical techniques to illuminate the complicated processes of templated polymerisation

systems. I shall look at models that are applicable to understanding features of both biological and synthetic systems. In order to introduce properly the existing literature on models of templating, more background mathematical and physical theory is required. Thus, the layout of this thesis is as follows. In chapter 2, I introduce the necessary mathematical background. In chapter 3, I review the existing literature on models of templated polymerisation. At the end of this chapter I can describe more precisely the niche this thesis aims to fill. In chapter 4, I introduce a class of complex models of polymerisation and develop the methodology to solve and analyse these models. With the models in this chapter and the newly developed methodology, I am able to answer the question of how complex chemical reaction motifs in polymerisation systems behave in terms of the accuracy of producing polymers, their energetic costs, and the time taken, and how such quantities compare to simpler models. In chapter 5, I explore a related class of models, which explicitly consider the effects of multiple pathways in systems where polymers are created and degraded. Through these models I ask how multiple pathways can allow for biased distributions of polymers to be maintained. In this chapter, I derive fundamental bounds describing how narrow distributions of products may be in such systems. Finally, in chapter 6, I provide a conclusion and outlook.

Chapter 2

Background theory

In the following chapter, I shall introduce the necessary mathematical background for motivating, modelling, analysing, and interpreting the main results of this thesis.

2.1 Information theory

I shall begin with that of information theory as this serves as the mathematical foundation for quantitatively stating and measuring the goals of achieving accuracy for these biochemical systems. The playground for information theory is that of probability. Throughout this subsection, I shall use capital letters X, Y to refer to random variables; calligraphic letters $\mathcal{X}, \mathcal{Y}$ for their state spaces; and lowercase letters x, y for specific realisations of the random variables. I shall assume the state spaces of these random variables are discrete. Additionally, I shall use blackboard bold $\mathbb{P}$ to refer to probability measures, and functions $p_X(x)$, $q_X(x)$ *etc.* to refer to probability distributions, *i.e.* $p_X(x) = \mathbb{P}[X = x]$. Subscripts such as $p_X, p_{X/Y}$ may be used to specify certain distributions. However, where there is no confusion, such subscripts will be dropped. I shall use $\mathbb{E}$ to refer to the expectation of a given random variable. Similarly to the probability measures, I may make use of subscripts to specify distributions where there may be confusion. The main quantity of interest for information theory is that of entropy. Given a probability distribution,

such as p_X for random variable X , the entropy of this distribution is given as [63]

$$H[X] = - \sum_{x \in \mathcal{X}} p_X(x) \log p_X(x) = \mathbb{E}_X[-\log p_X(X)], \quad (2.1)$$

with the convention that $p_X(x) \log p_X(x) = 0$ when $p_X(x) = 0$. Note that, although entropy is a functional of the probability distribution p_X , common notation is to write $H[X]$. In this thesis, I may write either $H[X]$ or $H[p_X]$. The base of the logarithm defines the units of the entropy. In pure mathematical or computer science literature, $\log_2$ is the most common, giving units of bits. In physics literature, however, $\ln$ is most common, giving units of nats. Changing the base of the logarithm only changes the numerical value of the entropy by a constant multiplier.

Note that, for a discrete random variable, the entropy is bounded $H[X] \in [0, \log |\mathcal{X}|]$, where $|\mathcal{X}|$ is the number of elements of the state space. The lower bound is achieved only by a distribution with one element of the state space having probability 1 and all others having probability 0, and the upper bound being achieved by the uniform distribution over the state space. As such, the entropy gives a measure of the spread of the distribution. Another way one may see how the entropy describes the uncertainty of a distribution is as follows. Let x be a sample of a discrete random variable X , the entropy of the distribution of X , when measured in bits, gives a lower bound on the expected minimum number of yes or no questions required for an optimum strategy to ascertain the value of x [63]. For example, let $\mathcal{X} = \{1, 2, 3, 4\}$ with two distributions, an unbiased one: $p(x) = 1/4$ and a biased distribution q , taking values $\{1/2, 1/4, 1/8, 1/8\}$ on $\{1, 2, 3, 4\}$. For the unbiased distribution, the entropy is 2 bits, and indeed one can always work out the value of a sample with two questions. For example one might ask, is the value odd or even as the first question, and then, for the second question, can ask whether or not the value is 1 if the previous answer were odd, or 2 if the previous answer were even. For the biased distribution, the entropy may be calculated to be 1.75 bits. An optimal questioning strategy would be asking first “is the value 1?”, then “is the value 2?” and finally “is the value 3?”. This indeed gives us an average of 1.75 questions needed to get

the answer.

Entropy can equally be defined for joint probability distributions, for example, $p_{X,Y}(x, y)$, as [63]:

$$H[X, Y] = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log p_{X,Y}(x, y) = \mathbb{E}_{X,Y}[-\log p_{X,Y}(X, Y)]. \quad (2.2)$$

If X and Y are independent, such that $p_{X,Y}(x, y) = p_X(x)p_Y(y)$, then $H[X, Y] = H[X] + H[Y]$, but in general $H[X, Y] \leq H[X] + H[Y]$. For conditional probability distributions, entropy is defined as [63]:

$$H[Y|X] = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log p_{Y|X}(x, y) = \mathbb{E}_{X,Y}[-\log p_{Y|X}(X, Y)]. \quad (2.3)$$

Note that for the conditional entropy, the expectation is taken over the joint probability distribution. The conditional entropy may be written in terms of other entropies as $H[Y|X] = H[X, Y] - H[X]$. Thus, the conditional entropy may be interpreted as how much X reduces the uncertainty of the joint distribution.

Mutual information is the final quantity of interest of information theory I shall discuss here. Mutual information is a measure of two random variables, X, Y , which may be written in terms of entropies as:

$$I[X; Y] = H[X] + H[Y] - H[X, Y]. \quad (2.4)$$

Recalling that $H[X] + H[Y]$ is the uncertainty of two independent random variables X, Y , the mutual information describes how much less uncertain the two random variables are when considered together as opposed to each being considered in isolation. It is often helpful to think of one of the random variables as an input and the other as the output of some process. In this case, an informative way of writing the mutual information is as $I[X; Y] = H[Y] - H[Y|X]$, such that the mutual information may be interpreted as how much the uncertainty of output Y is reduced by input X .

One of the major successes of Shannon's information theory is the channel coding theorem [64]. The aim of the channel coding theorem is to study the limitations of communication. Suppose there is some input X , which one wishes to convey to a receiver. p_X then describes the distribution of possible messages one much wish to send. However, there is a noisy channel through which the random variable must be sent. Such a channel leads to noise in the output variable Y , that the receiver receives. Formally, define a communication channel to be the $p_{Y|X}$. The channel is shown pictorially in figure 2.1. For an information channel, the channel capacity is then defined as:

$$C = \max_{p_X(x)} I(X; Y), \quad (2.5)$$

where the maximum is taken over all input distributions. Shannon's channel coding theorem states that the channel capacity sets a limit on the rate at which information can be sent via the channel. I shall not go into the specifics of the channel coding theorem for this thesis, but rather use it as further justification for the use of mutual information and channel capacity as measures of the accuracy of polymerisation.

It is worth calculating the channel capacity of channels with a certain symmetry. Suppose there is an input random variable X and output Y . For a fixed input $X = x$, think of the conditional probability distribution $p_{Y|X}(x, y)$ as a list of numbers describing the probabilities of output Y . Let this list of numbers for any specific y be some permutation of the same set of numbers. Similarly for any output $Y = y$, thinking of $p_{Y|X}(x, y)$ as a list of numbers, this list is simply a permutation of some set of numbers. More precisely, for a specific $x_1 \in \mathcal{X}$, write the conditional probabilities as a vector $\mathbf{p} = (p_{Y|X}(x_1, y_1), \dots, p_{Y|X}(x_1, y_{|\mathcal{Y}|}))$. Then for any $x_i \in \mathcal{X}$, there exists a permutation of $\mathcal{Y}$, $\Theta_{x_i} : \mathcal{Y} \mapsto \mathcal{Y}$, such that $(p_{Y|X}(x_i, \Theta_{x_i}(y_1)), \dots, p_{Y|X}(x_i, \Theta_{x_i}(y_{|\mathcal{Y}|}))) = \mathbf{p}$. And similarly for a specific $y_1 \in \mathcal{Y}$, setting $\mathbf{q} = (p_{Y|X}(x_1, y_1), \dots, p_{Y|X}(x_{|\mathcal{X}|}, y_1))$, there exists a set of permutations of $\mathcal{X}$, Θ_{y_i} , such that $(p_{Y|X}(\Theta_{y_i}(x_1), y_i), \dots, p_{Y|X}(\Theta_{y_i}(x_{|\mathcal{X}|}), y_i)) = \mathbf{q}$. A channel with these properties is called a symmetric channel [63]. In this case, the channel

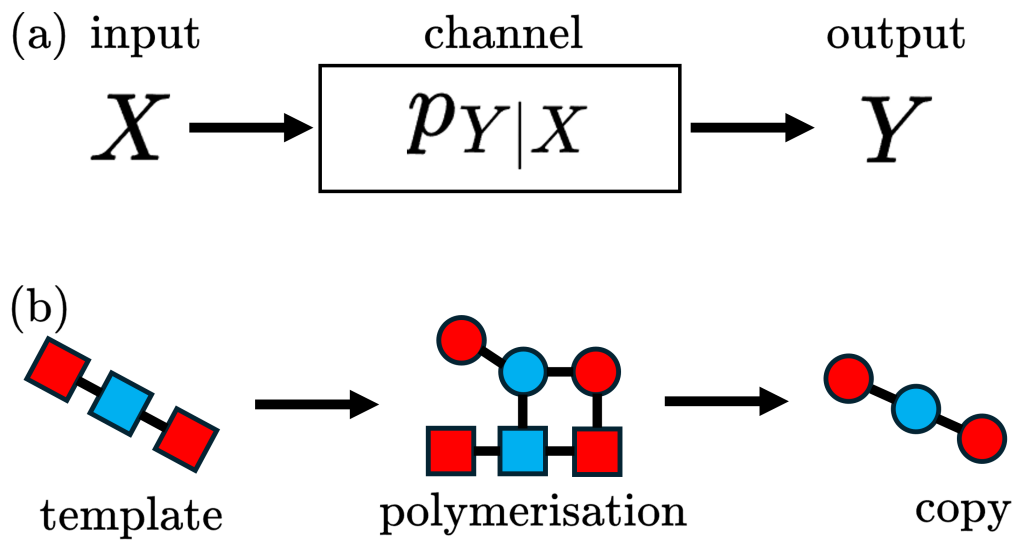


Figure 2.1: Depiction of an information channel (a) and its mapping to templated polymerisation (b). The channel is any process which takes in input(s) X and produces some output(s) Y with some noise. For templated polymerisation the inputs are template sequences and the outputs are the polymer sequences produced. I shall call these polymers the copies.

capacity may be calculated to be [64]

$$C = \log |\mathcal{Y}| + \sum_{y \in \mathcal{Y}} p_{Y|X}(x, y) \log p_{Y|X}(x, y), \quad (2.6)$$

where x is any fixed input. The maximising input distribution (comparing to eqn. 2.5) in this case, is the uniform distribution.

2.2 Markov chains, graph theory and spanning trees

2.2.1 The matrix tree theorem

A vital theorem that will underpin the methods of this thesis is that of the Matrix tree theorem. Originating with Tutte [65], this theorem provides a method for evaluating the determinant of a matrix as a sum over certain sub-graphs, called spanning trees, of a certain graph. Thus, to present this theorem, I shall introduce some graph theoretic quantities and notation.

Let $\mathcal{G} = (\mathcal{X}, \mathcal{E}, w)$ represent a weighted directed simple graph where $\mathcal{X}$ is the set of nodes and $\mathcal{E}$ is the set of edges. Let us assume there is a mapping $\mathcal{X} \mapsto \{1, \dots, |\mathcal{X}|\}$ such that the nodes can be assumed to take integer values. This graph is equipped with a weight function $w : \mathcal{E} \rightarrow \mathbb{R}^+$, where $\mathbb{R}^+$ is the set of positive real numbers. One may write the weight of an edge as $w(e)$ or $w(i, j)$ for the edge $e = (i, j) = i \rightarrow j$. The adjective simple here describes that between any two nodes there is no more than one edge present and the graph contains no self-loops (edges of the form $i \rightarrow i$).

A spanning tree of a weighted directed graph $\mathcal{G}$ is a subgraph $T = (\mathcal{X}_T, \mathcal{E}_T, w)$ which is connected, contains all the nodes of $\mathcal{G}$ (*i.e.* $\mathcal{X} = \mathcal{X}_T$), and contains no (undirected) cycles. A spanning tree is further said to be rooted at a node x , if all other nodes apart from x have an out-degree of one. Equivalently, for any node in the spanning tree, there exists a directed path towards the root. For a spanning tree T of a weighted graph, assign

a weight to the spanning tree as:

$$w(T) = \prod_{e \in \mathcal{E}_T} w(e). \quad (2.7)$$

Note that for a given root node, assuming the overarching graph $\mathcal{G}$ is finite, the set of spanning trees rooted at said node is finite.

Given a weighted graph, define the so called Laplacian matrix L as an $|\mathcal{X}| \times |\mathcal{X}|$ matrix with off-diagonal elements $L_{ij} = w(i, j)$, where edge $i \rightarrow j$ exists, and 0 otherwise; and diagonal elements $L_{ii} = -\sum_{j \neq i} L_{ji}$. Since the columns of the matrix sum to zero, the Laplacian matrix has at least one zero eigenvalue. It can be further shown (*e.g.* [66]) that if the graph is strongly connected, such that between any two pairs of nodes, i, j , there is a path of edges $i \rightarrow k, k \rightarrow l, \dots m \rightarrow j$ in the graph, then the nullity of the Laplacian is 1.

The matrix tree theorem then states that the determinant of the matrix formed by removing row and column i from the Laplacian matrix of $\mathcal{G}$, which I shall denote as $\det(L/i)$, can be found as a sum over spanning trees of $\mathcal{G}$ rooted at node i , specifically

$$\det(L/i) = \pm \sum_{T \in \mathcal{T}(i)} w(T), \quad (2.8)$$

where $\mathcal{T}(i)$ for $i \in \mathcal{X}$ represents the set of spanning trees of $\mathcal{G}$ rooted at i . The sign can be determined, but it is not necessary to go into the details for this thesis.

2.2.2 Markov chains

Much of the framework for models I shall introduce in this thesis will be built on the foundation of continuous time Markov chains (CTMCs). Let us write a continuous time stochastic process as X_t , where t is a time variable and indexes the random process, such that, for each $t \geq 0$, $X_t \in \mathcal{X}$. $\mathcal{X}$ is the state space of the CTMC. Assume this state space is finite. For such a continuous time stochastic process, the Markov property may

be written as [67]:

$$\mathbb{P}[X_{t_{n+1}} = x | X_{t_1} = x_1, X_{t_2} = x_2, \dots, X_{t_n} = x_n] = \mathbb{P}[X_{t_{n+1}} = x | X_{t_n} = x_n] \quad (2.9)$$

where $t_1 \leq t_2 \leq \dots \leq t_{n+1}$. In words, a process obeys the Markov property if it only depends on its past through its most recent event. A CTMC is a stochastic process on a discrete state space and continuous time which obeys the Markov property. A CTMC may be defined by a transition function W , such that in some small time window dt , $\mathbb{P}[X_{t+dt} = x_2 | X_t = x_1] = W_{x_2 x_1}(t)dt$, for $x_1 \neq x_2$ [67]. The transition function describes the probability of jumping from state x_1 to state x_2 in some small time window. Let us write the transition function for the CTMC as the matrix $W(t)$, such that $W_{ji}(t)$ is the rate of the transition from state i to state j . Then I denote the Markov process as the tuple, $(\mathcal{X}, W)$.

Let $p_i(t) = \mathbb{P}[X_t = i]$. One may write the so called master equation [68] describing how the probability distribution $p_i(t)$ changes in time:

$$\frac{dp_i(t)}{dt} = \sum_{j=1}^{|\mathcal{X}|} W_{ij}(t)p_j(t) - W_{ji}(t)p_i(t). \quad (2.10)$$

Define a matrix with entries

$$K_{ij}(t) = \begin{cases} W_{ij}(t) & i \neq j, \\ -\sum_{k=1}^{|\mathcal{X}|} W_{ki}(t) & i = j. \end{cases} \quad (2.11)$$

Then one may rewrite the master equation as a matrix equation $dp_i(t)/dt = \sum_{j=1}^{|\mathcal{X}|} K_{ij}(t)p_j(t)$. The matrix $K(t)$ is known as the rate matrix. Let us now assume that the rate matrix is constant $K(t) = K$. Such a process is described as time-homogeneous or simply homogeneous. Written in matrix form, one can easily see the formal solution to the master equation may be written:

$$\mathbf{p}(t) = \exp(tK)\mathbf{p}(0), \quad (2.12)$$

where $\exp(\cdot)$ is the matrix exponential and $\mathbf{p}(t)$ is a vector with i^{th} entry $p_i(t)$. However, calculating the matrix exponential for most processes is a difficult and cumbersome task, and the resulting closed form will be very complicated. It is most often easier to work with the steady state solution when possible. Let us write the steady state solution as $\boldsymbol{\pi} = \lim_{t \rightarrow \infty} \mathbf{p}(t)$. Then, one can see that $0 = K\boldsymbol{\pi}$, and so $\boldsymbol{\pi}$ belongs to the right nullspace of K . Note that, by construction, the columns of K sum to zero and therefore the nullspace of K is at least one dimensional. If the nullspace of K is exactly one dimensional, the CTMC has a unique steady state, independent of its initial conditions [67].

2.2.3 Collapsing nodes

A technique I shall be making use of throughout this thesis is manipulating the states of Markov processes to create new, related processes. I shall refer to this technique as collapsing sets of nodes together. Suppose there is a CTMC $(\mathcal{X} \cup \mathcal{Y}, W)$ such that the state space is split into $\mathcal{X} = \{1, \dots, |\mathcal{X}|\}$ and $\mathcal{Y} = \{|\mathcal{X}| + 1, \dots, |\mathcal{X}| + |\mathcal{Y}|\}$. One can collapse the set of states $\mathcal{Y}$ into a single state y to form a new process $(\mathcal{X} \cup \{y\}, W')$. The state y could either be one of the states of $\mathcal{Y}$ or some new state, whichever seems more appropriate for the application. The transition function for this new process is such that

$$\begin{aligned} W'_{yi} &= \sum_{z \in \mathcal{Y}} W_{zi}, \\ W'_{iy} &= \sum_{z \in \mathcal{Y}} W_{iz}, \\ W'_{ij} &= W_{ij}, \end{aligned} \tag{2.13}$$

for $i, j \in \mathcal{X}$. The rates into (or out of) the collapsed state in the new process are the sum over rates which originally went into (or out of) the states which were collapsed. For example, figure 2.2(c) depicts the process in figure 2.2(a) with states $\{1, 2, 3, A, B\}$ collapsed into one node. This example process will be explored in more detail in section 2.2.5. Note also that the definition eqn. 2.13 removes any possible self-loops that you might have expected from edges of the form $m \rightarrow n$ for $m, n \in \mathcal{Y}$ in the original process.

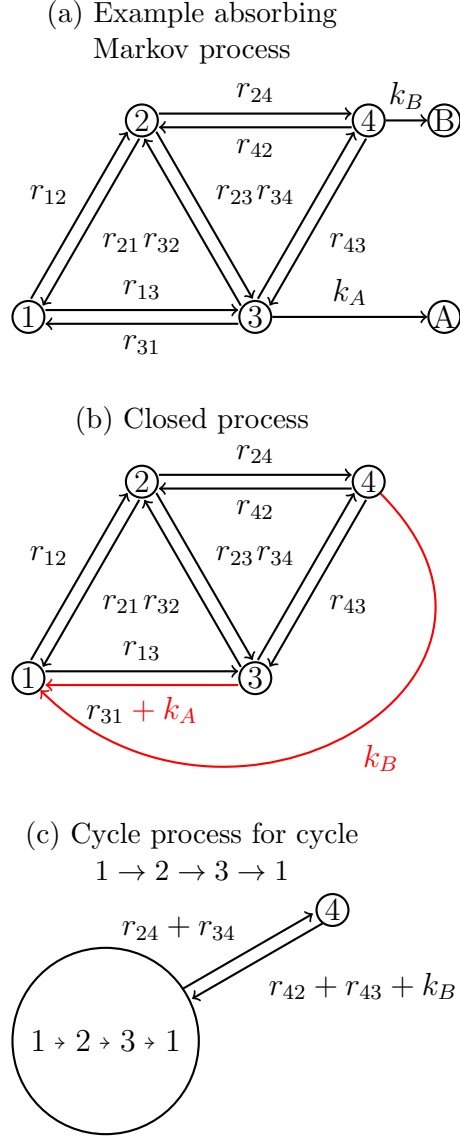


Figure 2.2: Graphical representations of an absorbing Markov process to illustrate the methodology outlined in section 2.2.9. (a) Example absorbing Markov process $(\mathcal{X} \cup \mathcal{A}, K)$, with two absorbing states, $\mathcal{A} = \{A, B\}$, and four transient states $\mathcal{X} = \{1, 2, 3, 4\}$. (b) The closed process starting at state 1, $(\mathcal{X}, W^1)$. (c) The cycle process $(\{c\} \cup \mathcal{X}/\{1, 2, 3\} = \{c, 4\}, W^{1,C})$ for the cycle $C = 1 \rightarrow 2 \rightarrow 3 \rightarrow 1$ or $C' = 1 \rightarrow 3 \rightarrow 2 \rightarrow 1$.

2.2.4 The Markov chain tree theorem

A useful visualisation of a CTMC can be made by noticing that the rate matrix looks like the Laplacian matrix of a weighted directed graph. With such an interpretation of the rate matrix, the CTMC may be drawn as a graph with nodes representing states, and weighted edges representing allowed transitions, with their weights being the transition rates. Such a representation as a graph is more than just useful as a visualisation aide thanks to an application of the matrix tree theorem known as the Markov chain tree theorem [69, 70] (MCTT). More precisely, let $\mathcal{G} = (\mathcal{X}, \mathcal{E}, w)$ be a weighted directed graph. $\mathcal{G}$ can be said to represent a CTMC $(\mathcal{X}, W)$, if an edge $i \rightarrow j \in \mathcal{E}$ if and only if $W_{ji} > 0$ for $i, j \in \mathcal{X}$; and the weight function obeys $w(i \rightarrow j) = W_{ji}$ where $W_{ji} > 0$. Let $\mathcal{G}$ be a weighted directed graph representing a CTMC whose rate matrix has nullity one. The Markov chain tree theorem gives a formula for the steady state solution to the master equation as

$$\pi_i = \frac{\sum_{T \in \mathcal{T}(i)} w(T)}{\sum_{j \in \mathcal{X}} \sum_{T \in \mathcal{T}(j)} w(T)}. \quad (2.14)$$

A proof for this result can be seen in [70]. The denominator here is simply a normalisation term to ensure the resultant vector is a probability distribution. More importantly, the steady state probability to be in a given state is proportional to the sum over spanning trees rooted at that state with proportionality constant being some normalisation factor. Sums over spanning trees will be a reoccurring concept in this thesis. As such it is important to emphasise, when I write “sum over the spanning trees”, I am referring to a quantity like the numerator of eqn. 2.14 - a sum over a set of spanning trees where the summand is the weight of the given tree. The normalisation factor can be cumbersome to calculate since it requires finding a sum over spanning trees for each node in the graph. However, in many applications, knowledge of the full steady state distribution is not necessary and only ratios of steady state probabilities are needed. The ratio of the steady state probability to be in state i to the steady state probability to be in state j

can be written as:

$$\frac{\pi_i}{\pi_j} = \frac{\sum_{T \in \mathcal{T}(i)} w(T)}{\sum_{T \in \mathcal{T}(j)} w(T)}, \quad (2.15)$$

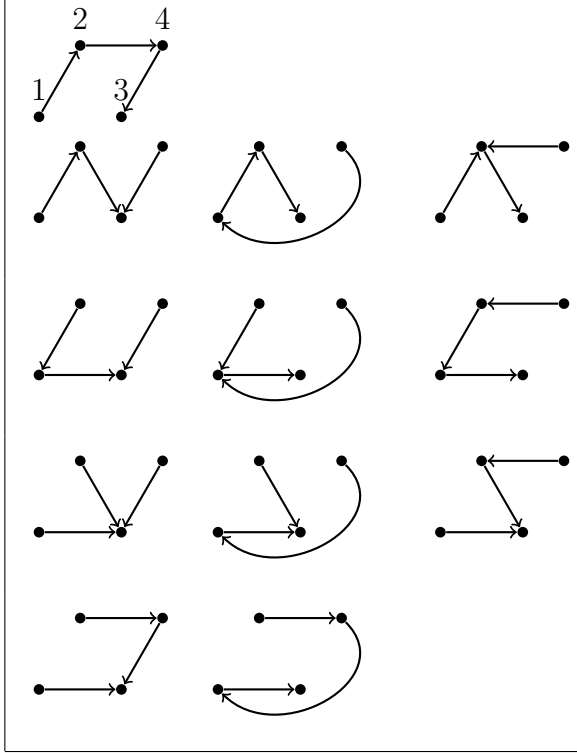
with the normalisation terms cancelling.

I note here that sums of spanning trees can be factorised in terms of self-avoiding walks (SAWs), a result which is both useful for generating sets of spanning trees and, later, will help in making statements about sums over spanning trees. For a directed graph, an SAW is a sequence of nodes $(x_1, x_2, \dots, x_n)$ such that between each consecutive pair of nodes, there exists an edge in the graph $x_i \rightarrow x_{i+1}$, and no node appears more than once in the sequence. I shall refer to the SAW as going from node x_1 to node x_n .

Suppose one has a process $(\mathcal{X}, W)$, being represented by graph $(\mathcal{X}, \mathcal{E}, w)$ for which one wishes to find the sum of spanning trees rooted at $x_1 \in \mathcal{X}$. This sum may be factorised in terms of SAWs between two vertices in the graph. Select some other arbitrary vertex $x_2 \in \mathcal{X}/\{x_1\}$ and let $\mathcal{S}(x_2 \rightarrow x_1)$ be the set of SAWs from x_2 to x_1 . Each spanning tree of the graph rooted at x_1 contains exactly one SAW from $\mathcal{S}(x_2 \rightarrow x_1)$. However, multiple different spanning trees can contain the same SAW. Thus, the spanning trees rooted at x_1 can be grouped by which SAW from $\mathcal{S}(x_2 \rightarrow x_1)$ they contain. For example, figure 2.3 depicts some spanning trees for the closed example process (figure 2.2(b)) described in section 2.2.5. In figure 2.3(a), the spanning trees are arranged in terms of SAWs from node 1 to node 3, with the first row for SAW: $1 \rightarrow 2 \rightarrow 4 \rightarrow 3$; the second row for $1 \rightarrow 2 \rightarrow 3$; and the last three rows for $1 \rightarrow 3$. Similarly for figure 2.3(b), the trees are arranged in terms of SAWs from node 1 to node 4 with row one for $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$; row two for $1 \rightarrow 2 \rightarrow 4$; row three for $1 \rightarrow 3 \rightarrow 4$; and row four for $1 \rightarrow 3 \rightarrow 2 \rightarrow 4$.

For each SAW $S \in \mathcal{S}(x_2 \rightarrow x_1)$, let us also form a new process $(\mathcal{X}/S \cup \{s\}, W^S)$, by collapsing the states in S to a new state s . Further, let w^S be the weight function on the appropriate graph representing the new process, with $\mathcal{T}_S(s)$ being the set of spanning trees rooted at the s node. Then, the sum over spanning trees of the original process can

(a) Spanning trees rooted at node 3.



(b) Spanning trees rooted at node 4.

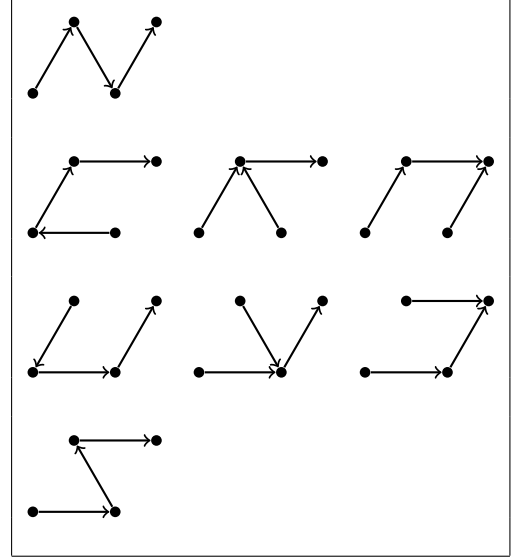


Figure 2.3: Spanning trees of the closed processes rooted at (a) node 3 and (b) node 4 derived from figure 2.2 (b), with nodes labelled in the first spanning tree and all other trees following the same positioning. The spanning trees have been arranged in terms of the self-avoiding walk between nodes 1 and 3 for the trees rooted at node 3 and arranged in terms of the self-avoiding walks between nodes 1 and 4 for the trees rooted at node 4.

be written, factoring out the sum over SAWs, as:

$$\sum_{T \in \mathcal{T}(x_1)} w(T) = \sum_{S \in \mathcal{S}(x_2 \rightarrow x_1)} \underbrace{\left[\sum_{e \in S} w(e) \right]}_{\text{SAW term}} \underbrace{\left[\sum_{T \in \mathcal{T}_S(s)} w^S(T) \right]}_{\text{Spanning tree term}}. \quad (2.16)$$

The spanning tree term looks rather intimidating to define and write down. However, it is quite intuitive and in practice one need not go through the formal process of writing down a collapsed process to find it. A crucially useful part of this expansion is that if one now wishes to find the sum over spanning trees rooted at x_2 , the sum may be factored into SAW from x_1 to x_2 . In this factorisation, the spanning tree term from eqn. 2.16 will be the same as for the spanning trees rooted at x_1 , and the SAW term will be very similar, just flipping the direction of all the edges. For now I shall leave this factorisation as a way to organise spanning trees.

2.2.5 Absorbing Markov chains

Let us now review and adapt some diagrammatic techniques introduced by Hill to analyse absorbing Markov chains [69, 71] and treated more mathematically rigorously by Qian, Qian and Qian in [72, 73]. An absorbing state of a CTMC is a state whose out degree is zero. Hence, any trajectory which reaches this state will get stuck and be absorbed into said state. An absorbing Markov chain is a Markov chain for which any trajectory through its state space with arbitrary initial conditions will reach an absorbing state in finite time almost surely [74]. The state space of an absorbing Markov chain can be decomposed into absorbing states, $\mathcal{A}$, and transient states, $\mathcal{X}$, such that the state space is $V = \mathcal{A} \cup \mathcal{X}$. Let us assume this state space is mapped onto the integers such that $\mathcal{X} = \{1, \dots, |\mathcal{X}|\}$ and $\mathcal{A} = \{|\mathcal{X}| + 1, \dots, |\mathcal{X}| + |\mathcal{A}|\}$.

Throughout this section, I shall refer to the absorbing Markov chain given in figure 2.2(a), which possesses two absorbing states and non-trivial cycles, for illustrative purposes.

2.2.6 Expectations of an absorbing process are steady-state averages of a “closed process”

I shall derive expressions for four main quantities: the probabilities of reaching certain absorbing states, the expected time taken to absorption, the expected net number of times traversing a given edge before absorption and the expected number of times that a trajectory goes round a cycle before absorption. These quantities in general depend on the starting distribution. I shall assume that the initial conditions for the Markov chain are such that the initial probability distribution has a single state with probability one and all others probability zero. Let us denote this starting state as $\sigma \in \mathcal{X}$. It is worth noting that all the methods I show here may be adapted to initial distributions which have support in more than one state. In such a case, for each state for which the initial distribution has support, one may calculate the relevant quantities for a deterministic starting state. Then one may take the expected value of a given relevant quantity using the initial probability distribution.

The quantities one wishes to calculate can be found in terms of the “closed” process [71]. The closed process is a modified version of an absorbing Markov chain in which transitions to the absorbing states are redirected to the starting state. Figure 2.2(b) shows the closed process starting at state 1 of the example absorbing chain of figure 2.2(a). The closed process for a Markov process $(\mathcal{X} \cup \mathcal{A}, W)$ starting at state $\sigma \in \mathcal{X}$ is a new Markov process $(\mathcal{X}, W^\sigma)$ with a rate function given by:

$$W_{\sigma x}^\sigma = W_{\sigma x} + \sum_{A \in \mathcal{A}} W_{Ax} \quad (2.17)$$

for $x \in \mathcal{X}/\{\sigma\}$ and agreeing with W for indices $\mathcal{X}/\{\sigma\} \times \mathcal{X}$. Equivalently, this process may be thought of as collapsing the nodes $\mathcal{A} \cup \{\sigma\}$ to the node σ .

The closed process has a unique stationary distribution for the following reasons. From the definition of an absorbing Markov chain, there exists a path from any state to an absorbing state, taking finite time. Thus, in the closed process, there is a path from any state to the starting state, taking finite time. The set of states including the starting

state and all those that may be reached from the starting state is therefore positive recurrent and further, this set is the only recurrent set of states and will be reached from any other state. Since there is only one recurrent set of states, there is a unique stationary distribution [74].

Expected quantities of an absorbing Markov chain, such as the probability that a given absorbing state is reached, can be found in terms of steady state quantities of the closed process. Whenever a trajectory of the original process reaches an absorbing state, in the closed process that same trajectory would have been reset back to the starting state. Hence, running the closed process for long times is equivalent to generating many independent trajectories to absorption for the original chain. Thus, averaging quantities in the steady state of the closed process is equivalent to taking expectations over independent trials of quantities in the absorbing chain [71]. It is worth noting that the dependence of expected quantities on the starting state is encoded in the definition of the closed process.

2.2.7 Steady state currents

Steady state currents can be defined from the steady state distribution of the closed process that correspond to expected currents of the absorbing chain. Let a superscript σ denote quantities in the closed process starting at state σ . Thus, let π^σ be the steady state probability distribution and W^σ the rate function. The current along a given edge, $e = i \rightarrow j$, is given by the probability to be in state i , which I write as π_i^σ , multiplied by the rate along said edge. Therefore, the steady state current along all edges that originally led to absorbing states can be written as:

$$J_{\text{Tot}}(\sigma) = \sum_{A \in \mathcal{A}} \sum_{i \in \mathcal{X}} \pi_i^\sigma W_{Ai}, \quad (2.18)$$

where, as before, $\mathcal{X}$ is the set of transient states and $\mathcal{A}$, the set of absorbing states. $J_{\text{Tot}}(\sigma)$ is the expected total current to absorbing states from state σ , and, therefore, its reciprocal is the expected time to absorption:

$$\mathbb{E}[\text{Time}] = \frac{1}{J_{\text{Tot}}(\sigma)}. \quad (2.19)$$

For the example process shown in figure 2.2(a), I present the spanning trees of the corresponding closed process rooted at nodes 3 and 4 in figure 2.3. Given the spanning trees, the total current to absorbing states can be written down directly as:

$$\begin{aligned} J_{\text{Tot}}(1) = & \frac{1}{\mathcal{N}} \left[k_A \left[r_{12} r_{24} r_{43} + r_{12} r_{23} (r_{42} + r_{43} + k_B) \right. \right. \\ & + r_{13} ((r_{43} + k_B)(r_{21} + r_{23} + r_{24}) + r_{42}(r_{21} + r_{23})) \Big] \\ & + k_B \left[r_{12} r_{23} r_{34} + r_{12} r_{24} (r_{31} + k_A + r_{32} + r_{34}) \right. \\ & \left. \left. + r_{13} r_{34} (r_{21} + r_{23} + r_{24}) + r_{13} r_{32} r_{24} \right] \right], \end{aligned} \quad (2.20)$$

where $\mathcal{N}$ is the normalisation term:

$$\begin{aligned} \mathcal{N} = & \left[r_{34} r_{42} r_{21} + r_{32} r_{24} k_B + r_{32} r_{21} (r_{43} + r_{42} + k_B) + r_{34} k_B (r_{24} + r_{23} + r_{21}) \right. \\ & + (r_{31} + k_A) (r_{21} r_{43} + r_{21} k_B + r_{42} r_{21} + r_{23} r_{43} + r_{23} k_B + r_{42} r_{23} + r_{24} r_{43} + r_{24} k_B) \Big] \\ & + \left[r_{13} r_{34} r_{42} + r_{13} r_{32} (r_{43} + r_{42} + k_B) + r_{12} (r_{34} r_{42} + r_{43} r_{32} + r_{32} k_B + r_{32} r_{42} + r_{43} (r_{31} + k_A) \right. \\ & + k_B (r_{31} + k_A) + r_{42} (r_{31} + k_A) + r_{34} k_B) \Big] \\ & + \left[r_{12} r_{24} r_{43} + r_{12} r_{23} (r_{42} + r_{43} + k_B) + r_{13} ((r_{43} + k_B)(r_{21} + r_{23} + r_{24}) + r_{42}(r_{21} + r_{23})) \right] \\ & + \left[r_{13} r_{34} (r_{21} + r_{23} + r_{24}) + r_{13} r_{32} r_{24} + r_{12} r_{23} r_{34} + r_{12} r_{24} (r_{31} + k_A + r_{32} + r_{34}) \right]. \end{aligned} \quad (2.21)$$

I have factorised each of the terms in terms of SAWs. The first square bracket corresponds to the trees rooted at node 1, organised by SAWs from node 3; the second to trees rooted at 2 organised by SAWs from 1; the third to trees rooted at 3 organised by SAWs from 1 and the fourth to trees rooted at 4 organised by SAWs from 1.

In eqn. 2.20, the terms multiplied by k_A are the partial current to absorbing state A , *i.e.* the current along transition $3 \rightarrow A$, coming from the trees rooted at node 3, and equivalently for k_B with state B , *i.e.* the current along transition $4 \rightarrow B$, coming from trees rooted at node 4.

2.2.8 Absorbing probabilities

Given a Markov chain with multiple absorbing states, one can ask for the probability of absorption in each absorbing state in the long time limit. The probability that a trajectory eventually ends in a specific absorbing state can be calculated from the closed process, by dividing the expected current along transitions that originally led to the absorbing state in question by $J_{\text{Tot}}(\sigma)$ (eqn. 2.18). Therefore, the absorption probabilities can be written

$$\mathbb{P}[\sigma \rightarrow A] = \frac{\sum_{x \in \mathcal{X}} \pi_x^\sigma W_{Ax}}{\sum_{B \in \mathcal{A}} \sum_{x \in \mathcal{X}} \pi_x^\sigma W_{Bx}}, \quad (2.22)$$

using the notation $\mathbb{P}[\sigma \rightarrow A]$ to denote probability of being absorbed to A given that the trajectory started in state σ . It is worth noting here that given that this quantity is a ratio of currents, there is a factor of π^σ in both the denominator and the numerator of the expression. In practice, the normalisation factor from the MCTT (eqn. 2.14) cancels out, which simplifies the quantities in the calculation.

For the example process shown in figure 2.2, the partial currents to absorbing states A and B can be used to write down the absorbing probabilities:

$$\begin{aligned} \mathbb{P}[1 \rightarrow A] &= \frac{1}{\mathcal{N}J_{\text{Tot}}(1)} k_A [r_{12}r_{24}r_{43} + r_{12}r_{23}(r_{42} + r_{43} + k_B) \\ &\quad + r_{13}((r_{43} + k_B)(r_{21} + r_{23} + r_{24}) + r_{42}(r_{21} + r_{23}))], \\ \mathbb{P}[1 \rightarrow B] &= \frac{1}{\mathcal{N}J_{\text{Tot}}(1)} k_B [r_{13}r_{34}(r_{21} + r_{23} + r_{24}) + r_{13}r_{32}r_{24} \\ &\quad + r_{12}r_{23}r_{34} + r_{12}r_{24}(r_{31} + k_A + r_{32} + r_{34})]. \end{aligned} \quad (2.23)$$

The normalisation factor, $\mathcal{N}$, propagated through from eqn. 2.14, conveniently cancels out with the $1/\mathcal{N}$ implicit in J_{tot} .

2.2.9 Counting edge and cycle transitions

In this subsection, I shall calculate the expected net number of times traversing a given edge of an absorbing Markov process before absorption. Additionally, I shall calculate the expected number of times a non-recurrent cycle of an absorbing Markov process is traversed before absorption. Both of these will be of use later in defining a notion of chemical work.

To calculate the net number of times crossing a given edge, find the expected (net) current along the transition, $i \rightleftharpoons j$, between states $i, j \in \mathcal{X}$ of an absorbing process, $(\mathcal{X} \cup \mathcal{A}, W)$, where the direction of this current is from state i to j . The expected (net) current through this edge, which I denote $J_{i \rightleftharpoons j}(\sigma)$, given starting in state $\sigma \in \mathcal{X}$, can be calculated from the closed process, $(\mathcal{X}, W^\sigma)$, as in eqn. 2.17, as the difference between the steady state probability to be in state i multiplied by the rate from $i \rightarrow j$ and the steady state probability to be in state j multiplied by the rate from $j \rightarrow i$,

$$J_{i \rightleftharpoons j}(\sigma) = \pi_i^\sigma W_{ji} - \pi_j^\sigma W_{ij}. \quad (2.24)$$

The net number of times traversing this edge (number of observed transitions $i \rightarrow j$ - number of observed transitions $j \rightarrow i$) before absorption is then just the ratio between this current and total current to absorbing states:

$$N_{i \rightleftharpoons j}(\sigma) = \frac{J_{i \rightleftharpoons j}(\sigma)}{J_{\text{Tot}}(\sigma)}. \quad (2.25)$$

One often also wishes to find the expected number of times traversing a non-recurrent cycle. Define a non-recurrent cycle for a Markov chain to be a cycle of states $(x_1 \rightarrow x_2 \rightarrow \cdots \rightarrow x_1)$, where each state, aside from the originating state, does not appear more than once in the cycle. For example, the cycle $A \rightarrow B \rightarrow C \rightarrow D \rightarrow A$ is non-recurrent, but $A \rightarrow B \rightarrow C \rightarrow D \rightarrow B \rightarrow A$ is recurrent. Note that the originating state is arbitrary, and so $A \rightarrow B \rightarrow C \rightarrow D \rightarrow A$ is equivalent to $B \rightarrow C \rightarrow D \rightarrow A \rightarrow B$. Note, I define a cycle as going in a single direction.

For a stationary process, the expected frequency with which a cycle is completed in a given direction can be calculated from the one-way cycle current [71, 75], which is the probability current going around the cycle. Note this one-way current is not a net current.

For a chosen non-recurrent cycle, the one-way cycle current can be calculated diagrammatically from three terms. First, a cycle term given by the product of rates around the cycle in the chosen direction. Second, a spanning tree term that can be found by collapsing the nodes in the cycle into a cycle node (in figure 2.2(c), the cycle $1 \rightarrow 2 \rightarrow 3 \rightarrow 1$ has been collapsed in this way) and finding the sum of spanning trees of this new graph rooted at the collapsed cycle node. Finally, there is a normalisation factor, which is the same normalisation factor as for the current, $\mathcal{N}$. Explicitly, consider an absorbing Markov chain, $(\mathcal{X} \cup \mathcal{A}, W)$ and its closed process starting at $\sigma \in \mathcal{X}$, $(\mathcal{X}, W^\sigma)$, with w and w^σ representing the corresponding weight functions for graphs representing them. Let $C = (\mathcal{C}, \mathcal{E}_C)$ be a subgraph representing a cycle in the closed process, with $\mathcal{C}$ the set of nodes in the cycle and $\mathcal{E}_C$ the set of edges. To calculate the spanning tree term for the one-way cycle current, construct a new Markov chain, the cycle process, $(\{c\} \cup (\mathcal{X}/\mathcal{C}), W^{\sigma, C})$, by collapsing the states of the cycle C to a single node c . Thus $\{c\} \cup (\mathcal{X}/\mathcal{C})$ is the set of transient states of the original process with the states in the cycle replaced by the single node, c , and $W^{\sigma, C}$ is given by:

$$\begin{aligned} W_{cx}^{\sigma, C} &= \sum_{i \in C} W_{ix}^\sigma \\ W_{xc}^{\sigma, C} &= \sum_{i \in C} W_{xi}^\sigma \end{aligned} \tag{2.26}$$

for $x \in \mathcal{X}/\mathcal{C}$ and agreeing with W^σ elsewhere. The cycle process for the cycle $C = 1231$ (or $C' = 1321$) of the example system in figure 2.2(a) is shown in figure 2.2(c). Let $\mathcal{T}_\sigma(x), \mathcal{T}_C(x)$ be the sets of spanning trees rooted at x of the closed process and cycle

process, respectively. Then, the cycle current is given by [75]:

$$J_{\text{Cyc}}(\sigma, C) = \frac{\overbrace{\left(\prod_{e \in C} w(e) \right)}^{\text{Cycle}} \overbrace{\sum_{T \in \mathcal{T}_C} w^{\sigma, C}(T)}^{\text{Spanning Trees}}}{\underbrace{\sum_{x \in \mathcal{X}} \sum_{T \in \mathcal{T}_\sigma(x)} w^\sigma(T)}_{\text{Normalisation}}}. \quad (2.27)$$

Note for the cycle term, the edges are taken from the original process rather than the closed process. Note also that for the cycle in the opposite orientation $C' = (\mathcal{C}, \mathcal{E}'_C)$, the cycle process, formed by collapsing the states of C' is the same as for the original cycle C . Thus, the spanning tree and normalisation terms are the same as for C , only the cycle term is different.

Given the cycle current for the closed process, the expected number of circulations of the cycle before absorption, $N_{\text{Cyc}}(\sigma, C)$, is the ratio of the cycle current to the total current to absorbing states:

$$N_{\text{Cyc}}(\sigma, C) = \frac{J_{\text{Cyc}}(\sigma, C)}{J_{\text{Tot}}(\sigma)}. \quad (2.28)$$

For the example process, the expected number of circulations of $C = 1231$ is

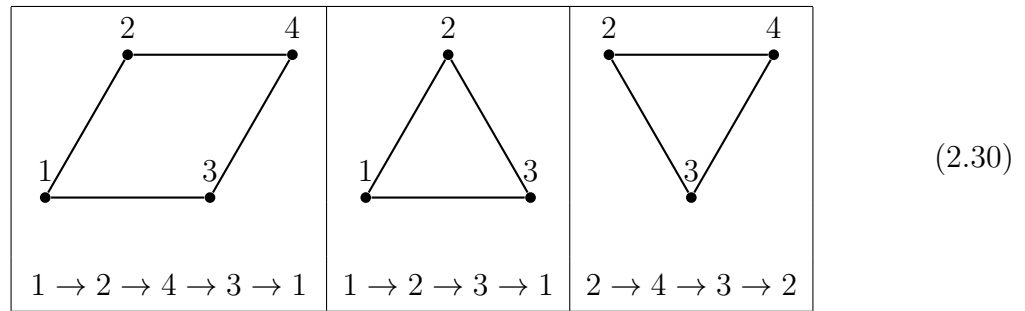
$$N_{\text{Cyc}}(1, C) = \frac{(r_{12}r_{23}r_{31})(r_{42} + r_{43} + k_B)}{\mathcal{N}J_{\text{Tot}}(1)}, \quad (2.29)$$

with the same implicit cancellation of normalisation as before, since $J_{\text{Tot}} \propto \frac{1}{\mathcal{N}}$.

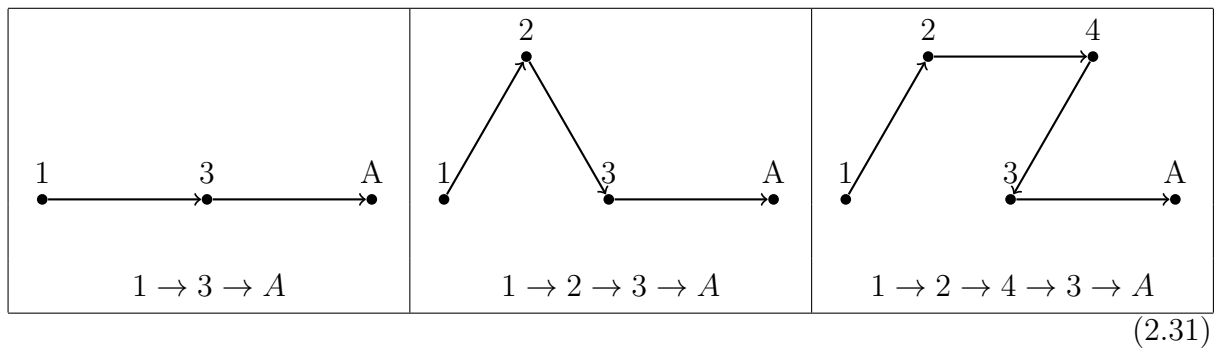
For an absorbing process starting at a given state, the cycles may be divided into internal and external cycles. External cycles are those which appear in the closed process and involve edges which were absorbing edges in the original process. Internal cycles are those which were present in the original process.

For the example process, figure 2.2(a), the cycles may be written down as follows. Let's divide these cycles into internal cycles, external cycles to absorbing state A and

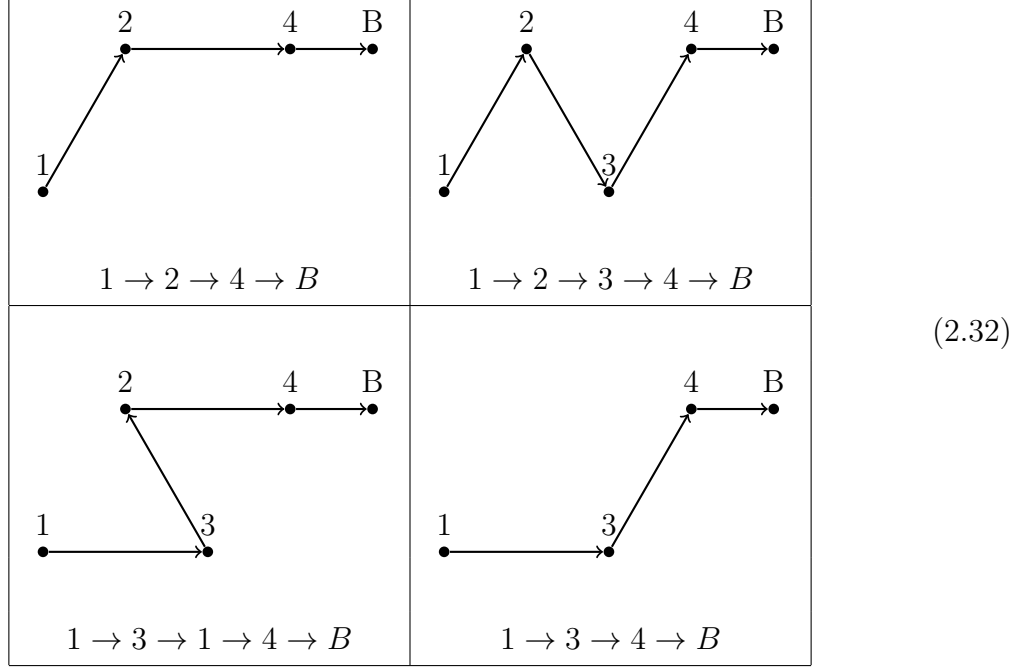
external cycles to absorbing state B . Firstly, the internal cycles are:



where the cycle is written out below in the clockwise direction. Similarly, the external cycles to state A are:



Finally, the external cycles to absorbing state B are:



The external cycles correspond to the pathways from the starting state to an absorbing state. Therefore, the expected number of times traversing an external cycle before absorption will be at most one and corresponds to the probability of following a given path to absorption. Further, the sum of eqn. 2.28 over all external cycles will be one.

2.2.10 Gillespie's algorithm

I end this brief stint through CTMCs with a practical tool for simulating such Markov processes exactly. In 1976, Gillespie [76] introduced an algorithm to sample trajectories of CTMCs. Consider a process $(\mathcal{X}, W)$. Let us continue to work with homogeneous processes. Let the process have initial probability distribution $\mathbf{p}(0)$. For each state i in the system, define a total out rate

$$k_{\text{out}}(i) = \sum_{j \in \mathcal{X}} W_{ji} = -K_{ii}, \quad (2.33)$$

where K is the rate matrix. Also define a transition probability distribution $\mathbb{P}[j|i]$ for the probability of the next state j given the trajectory is currently in state i . This probability is given by

$$\mathbb{P}[j|i] = \frac{W_{ji}}{k_{\text{out}}(i)}. \quad (2.34)$$

Note that $\mathbb{P}[i|i] = 0$. The algorithm is then as follows.

1. Draw an initial state according to the probability distribution $\mathbf{p}(0)$. Define this to be the current state. Set the time t to zero.
2. Draw a random variable $\tau \sim \text{Exp}(1/k_{\text{out}}(i))$, an exponential random variable with mean $1/k_{\text{out}}(i)$. Update the time $t = t + \tau$.
3. Draw a next state according to the transition probability distribution $\mathbb{P}[\cdot|\text{current state}]$. Set this next state to be the current state.
4. Repeat from step 2 onwards until you wish to stop.

Performing such an algorithm while recording the list of states visited in order $x_0, \dots, x_n$ as well as the times $t_0 = 0, t_1 = \tau, \dots$, yields a trajectory sampled according to the CTMC, which is in state x_0 for time t_0 to t_1 , state x_1 from t_1 to t_2 *etc.*

2.3 Chemical reaction networks

2.3.1 Definitions and notation

Let us move on to introduce a different, yet closely related class of processes. Chemical reaction networks (CRNs) are a mathematical framework for studying systems of chemical reactions. A chemical reaction network is the triple $(\mathcal{X}, \mathcal{C}, \mathcal{R})$ consisting of a set of chemical species $\mathcal{X}$, a set of chemical complexes $\mathcal{C}$ and a set of reactions $\mathcal{R}$. Chemical complexes are collections of chemical species. There is a special complex I shall call the null complex (which may or not be present in a given CRN) and denote by $\emptyset$ representing the complex with no chemical species. It is also worth noting that each chemical species

can appear more than once in a given complex. A chemical reaction is something which turns one chemical complex into another. I might represent a chemical reaction as $(\mathbf{c}^+, \mathbf{c}^-)$ or $\mathbf{c}^+ \rightarrow \mathbf{c}^-$. In such a representation, the complex $\mathbf{c}^+$ is known as the reactants and the complex $\mathbf{c}^-$ the products of the reaction. I shall use C for the number of chemical species $|\mathcal{X}|$, and R for the number of reactions $|\mathcal{R}|$. Further, I shall assume that the chemical species are mapped onto the integers $\{1, \dots, C\}$. If the chemical species are represented as integers, complexes may be represented as vectors $\mathbf{c} \in \mathbb{N}^C$. Such a vector represents the complex by having in the i^{th} position, the number of chemical species i present in said complex. These vectors are known as the stoichiometric vectors.

2.3.2 Deterministic CRNs

To determine the dynamics represented by a CRN, I need to give a few more assumptions. The first assumption I shall make in this section is that each chemical species is present in large enough numbers that fluctuations in the number of a given species are negligible. Thus, the CRN can be treated deterministically, with each chemical species being represented by a (continuous) concentration, which I shall denote by z_i for chemical species i . One is then interested in how the concentration of each species varies in time due to the CRN and as such wishes to write down a system of ordinary differential equations (ODEs) describing the rate of change of concentrations.

The second assumption I shall make (which shall apply throughout this work) is that of mass action kinetics. By assuming that the reactions take place in a roughly homogeneous (well-mixed) environment, one may assume that the rate of a given reaction is proportional to the amount of each of its reactants present. For deterministic CRNs this takes the form that the rate of reaction $\mathbf{c}^+ \rightarrow \mathbf{c}^-$

$$\text{Rate} \propto \mathbf{z}^{\mathbf{c}^+} = \prod_{i=1}^C z_i^{(c^+)_i}, \quad (2.35)$$

where $(c^+)_i$ is the i^{th} component of $\mathbf{c}^+$. Further, I introduce notation, common to CRN literature, that vector to the power of another vector is taken to mean a product of

component-wise exponents. The constant of proportionality for a given reaction is known as its rate constant and is often denoted by k with a subscript identifying the reaction. Reaction $\mathbf{c}^+ \rightarrow \mathbf{c}^-$, with rate constant k , feeds into the change in concentration of chemical species i by contributing the terms:

$$\frac{dz_i}{dt} = \cdots + [(c^-)_i - (c^+)_i] k \mathbf{z}^{\mathbf{c}^+} + \cdots \quad (2.36)$$

such that this term is positive if there is more of species i in the products of the reaction than the reactants. Summing over all such terms from all the reactions gives the full rate of change of concentration z_i . Such a set of equations is known as the reaction rate equations (RREs). As one can observe from this equation, the units of a reaction rate should be concentration per time. Thus, the units of rate constants are concentration to some power per time. To avoid messes with keeping track of these various concentrations, I shall assume there is some overall reference concentration [77], which can be used as the standard unit of concentration, such that concentrations are dimensionless.

Some chemical species one may wish to keep at constant concentration. These species are said to be chemostatted. A common assumption as to how a chemical species might be chemostatted is that there exists some CRN outside of what is being explicitly modelled [78]. This external CRN has a steady state such that the chemostatted species is at its chemostatted concentration. Further, the external CRN acts on a much faster timescale compared to the CRN of interest. Under such a timescale separation, any change in the concentration of the chemostatted species due to the CRN of interest is quickly reverted by the external CRN. Practically, if a chemical species is being treated as chemostatted, it can be removed from $\mathcal{X}$ and its contribution to the ODEs for concentrations of other species (eqn. 2.36) may be absorbed into the corresponding rate constants.

2.3.3 Stoichiometry

A given reaction $\mathbf{c}^+ \rightarrow \mathbf{c}^-$, causes a change in the numbers of chemical species. This change in stoichiometry due to the reaction may be represented as a vector $\mathbf{c}^- - \mathbf{c}^+$. Putting the stoichiometry vectors all together, much of the CRN can be captured by a single matrix. This stoichiometric matrix S is an $C \times R$ matrix with each column being the stoichiometric change due to a reaction [79]. This stoichiometric matrix allows the RREs for the CRN to be written down very compactly as:

$$\frac{d\mathbf{z}}{dt} = S\mathbf{J}(\mathbf{z}), \quad (2.37)$$

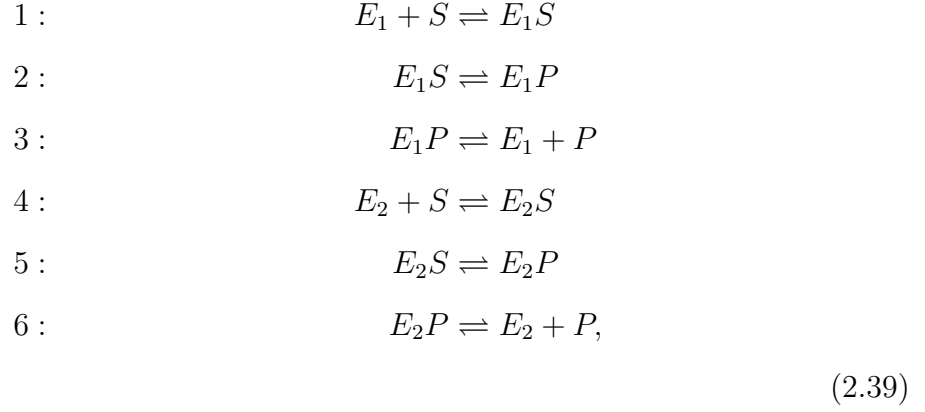
where $\mathbf{J}(\mathbf{z})$ is a vector of reaction currents. Assuming mass action kinetics, the reaction currents take the form $J_\rho(\mathbf{z}) = k_\rho \mathbf{z}^{\mathbf{c}^+}$ for reaction ρ representing $\mathbf{c}^+ \rightarrow \mathbf{c}^-$ with rate constant k_ρ .

A common assumption is that of microreversibility, which in the context of CRNs means that if a given reaction $\mathbf{c}^+ \rightarrow \mathbf{c}^-$ can occur, its reverse reaction $\mathbf{c}^- \rightarrow \mathbf{c}^+$ can also occur. Under such an assumption, both the forwards and backwards reactions can be packaged into a single reaction for the stoichiometric matrix and the pair of reactions can be encoded with a modified reaction current

$$J_\rho(\mathbf{z}) = k_\rho^+ \mathbf{z}^{\mathbf{c}^+} - k_\rho^- \mathbf{z}^{\mathbf{c}^-}, \quad (2.38)$$

where k_ρ^+ is the rate constant for the forwards reaction and k_ρ^- for the reverse reaction. Such a form is not affected by which reaction is defined to be the forwards reaction as swapping forwards and backwards reactions changes the sign of both the reaction current and the corresponding column of the stoichiometric matrix. I shall assume microreversibility throughout this thesis unless otherwise stated.

Throughout the rest of this subsection I shall refer back to the following example CRN, representing two enzymes E_1 , E_2 which can both catalyse the conversion of a substrate S into a product P .



with reaction number written on the left. A list of reactions written like this, alongside the rate constants of each reaction, contains all the information required for the CRN (chemical species and chemical complexes). Thus such lists of reactions are often stated to be the CRN. Each of these reactions is reversible with the forward reaction going from left to right. As such I shall use reaction currents of the form eqn. 2.38. This CRN has eight chemical species, which I shall list in the order $\{S, P, E_1, E_1S, E_1P, E_2, E_2S, E_2P\}$. The stoichiometric matrix for this CRN is

$$S = \begin{pmatrix} -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}.
\tag{2.40}$$

The stoichiometric matrix allows us to make some statements about the CRN without considering the reaction rates. Two such types of statement come from the nullspaces

of the stoichiometric matrix. Firstly, suppose $\mathbf{l}$ is a left null vector of S (*i.e.* $\mathbf{l} \cdot S = 0$). Then $\mathbf{l}$ corresponds to a conserved quantity or conservation law of the CRN [79]. More precisely, $L = \mathbf{l} \cdot \mathbf{z}$ is conserved as

$$\frac{dL}{dt} = \mathbf{l} \cdot S\mathbf{J} = 0. \quad (2.41)$$

If one finds such a conserved quantity whose corresponding vector contains all positive entries, that vector corresponds to conserved moieties. Moieties are parts of chemical complexes. In the example CRN, there are three left null vectors, which can be written

$$\mathbf{l}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{l}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{l}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}. \quad (2.42)$$

$\mathbf{l}_1$ corresponds to the conserved moiety of substrate/product, whereas $\mathbf{l}_2$ and $\mathbf{l}_3$ correspond to the conservation of enzymes (E_1 and E_2 respectively). Often when one writes out a CRN to try to model a system, they expect the CRN to have certain conserved moieties, for example corresponding to various catalysts in the system.

If there exists a left null vector of the stoichiometric matrix with all strictly positive entries, then the CRN has an overall conservation of mass. For the example CRN, $\mathbf{l}_1 + \mathbf{l}_2 + \mathbf{l}_3$ is such a vector. If the CRN does not have an overall conservation of mass, it might not be possible to write down linearly independent left null vectors with all non-negative entries.

Since the quantities corresponding to the left null vectors are conserved, they do not change along a trajectory. The values of all conserved quantities specifies a stoichiometry

compatibility class [80].

Next, look at the right nullspace of the stoichiometric matrix. Vectors in the right nullspace correspond to cycles of reactions which lead to no change in stoichiometry. For the example (eqn. 2.39), the right nullspace is one dimensional and is spanned by

$$\mathbf{v} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \end{pmatrix}. \quad (2.43)$$

Such a vector corresponds to the cycle of doing reactions 1,2,3 in the forwards direction and 4,5,6 in the backwards direction once each. Tracking the numbers of each chemical species as they change according to these reactions, one would indeed find all species remain unchanged. These cycles are interesting when studying the steady state. It is clear from eqn. 2.37, that at a steady state, the current vector is in the right nullspace of the stoichiometric matrix. If the steady state of a microreversible CRN is $\mathbf{J} = 0$, the steady state is said to be detailed balanced [81] as the current for each individual reaction is balanced between forwards and backwards reactions. If the right nullspace is 0-dimensional, then only detailed balanced steady states are possible.

2.3.4 Stochastic CRNs

I shall have a brief discussion here of stochastic modelling of CRNs. Although in this thesis I shall be mainly concerned with deterministic CRNs, it is worth including the following discussion for completeness and to note when stochastic CRNs behave like deterministic ones on average.

Stochastically, one may use a CRN to define a CTMC. This CTMC has a state space defined by the number of each chemical species. Thus, one could represent the states

as a C -dimensional vector with non-negative integer entries. The state space of such a CTMC is, in general, infinite. However, the specifics of the CRN such as a conservation of mass, might restrict a given trajectory to a finite subspace.

Transitions occur between states according to the reactions. Assuming the trajectory in state $\mathbf{c}$, a reaction $\mathbf{c}^+ \rightarrow \mathbf{c}^-$, changes the state to $\mathbf{c} - \mathbf{c}^+ + \mathbf{c}^-$. Assuming mass action kinetics, and that the reaction has rate constant k , the rate of such a reaction is

$$k \prod_{i=1}^C \frac{c_i!}{(c_i - (c^+)_i)!}, \quad (2.44)$$

if $c_i \geq (c^+)_i$ for all species i and zero otherwise. The rate constants here differ from the deterministic case by a factor of some volume term. If one wishes to make a clear connection to the deterministic limit, it is necessary to consider this volume more carefully.

Let $P(\mathbf{c}, t)$ describe the probability of being in state $\mathbf{c}$ at time t and let each of the reactions be indexed by ρ such that reactions look like $\mathbf{c}_\rho^+ \rightarrow \mathbf{c}_\rho^-$ and have rate constants k_ρ . Bringing this all together, the master equation [67] (for a general, not necessarily reversible CRN) may be written as

$$\begin{aligned} \frac{dP(\mathbf{c}, t)}{dt} = \sum_{\rho=1}^R \left(- P(\mathbf{c}, t) k_\rho \prod_{i=1}^C \frac{c_i!}{(c_i - (c_\rho^+)_i)!} \right. \\ \left. + P(\mathbf{c} - (\mathbf{c}_\rho^- - \mathbf{c}_\rho^+), t) k_\rho \prod_{i=1}^C \frac{(c_i - ((c_\rho^-)_i - (c_\rho^+)_i))!}{(c_i - (c_\rho^-)_i)!} \right), \end{aligned} \quad (2.45)$$

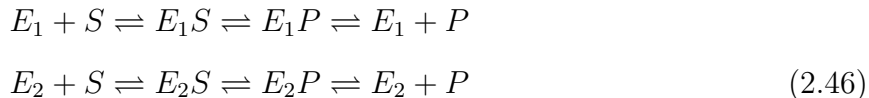
where, implicitly, any terms with negative factorials are neglected. Such an equation is often referred to as the chemical master equation. In general, the chemical master equation is difficult to solve, even at steady state. The CRN might also impose restrictions due to the discreteness of the numbers of chemical species. For example, a species might only change by an even number in any reaction. Thus, its parity might be restricted to match its initial conditions. These are in addition to any conservation laws, which might restrict the accessible states to a given stoichiometric compatibility class for a given initial

condition.

Nonetheless, an important question is whether one can say that the mean number of chemical species at steady state is proportional to a solution to the RREs (eqn. 2.37) in the appropriate stoichiometric compatibility class. In general, such a proportionality need not hold. However, one class of CRNs for which this statement is true are the deficiency zero CRNs [82], which I shall describe in the next subsection.

2.3.5 Graphical features of CRNs

There are many ways to visualise CRNs, one natural way is with the complexes graph [81]. Here, the complexes $\mathcal{C}$ become the nodes of the graph and the reactions become the edges. The complexes graph is the natural object one would come to if one wished to write down the reactions of the CRN compactly. For the example CRN (eqn. 2.39), the complexes graph is just



where for clarity, I use $\rightleftharpoons$ to represent edges and write out the complexes in terms of their constituent chemical species. The complexes graph, however, does not make obvious some of the connections in the CRN. The complexes graphs will often be disconnected, even though the complexes in different connected components of the graph can share chemical species. As such cycles of reactions may be hidden in the complexes graph.

One formulation of the deficiency of chemical reaction networks makes explicit the references to hidden cycles. One can look at the number of cycles of reactions as defined by the right nullspace of the stoichiometric matrix and compare this to the number of cycles in the complexes graph. Letting I be the incidence matrix of the complexes graph, then this formulation of the deficiency can be written [83] as

$$\delta = \dim(\ker(S)) - \dim(\ker(I)) \tag{2.47}$$

where S is the stoichiometric matrix. Other formulations of the deficiency exist, however, this one is visually tractable, makes clear the connections to cycles and is easy to apply to certain classes of CRNs. The example CRN has one cycle according to the stoichiometric matrix and zero in the complexes graph, thus $\delta = 1$.

Deficiency is a useful quantity to consider due to the existence of a number of theorems which have been proven for CRNs which are deficiency zero or one. The most famous of these deficiency theorems is the aptly named deficiency zero theorem [84], which states that (for the reversible CRNs being considered) that there is a unique stable equilibrium in each stoichiometric compatibility class. Thus, such networks cannot show any long time behaviour other than converging to an equilibrium. Another such deficiency result is what was alluded to in the previous subsection: that stochastic realisations of deficiency zero CRNs agree with their deterministic counterparts at steady state. More specifically, Anderson, Craciun, and Kurtz [82] prove that the stationary distribution is Poissonian in each stoichiometric compatibility class with mean proportional to the appropriate deterministic solution. For the networks I shall consider in this thesis, these deficiency theorems are used implicitly such that the distinction between stochastic and deterministic CRNs can be ignored and it can be assumed that any CRN will tend to a steady state.

Finally, I shall discuss linear CRNs. Linear CRNs are such that each complex contains at most one chemical species and if a complex does contain a chemical species, then it has stoichiometry one. Linear CRNs are particularly nice to work with. Linear CRNs without a null complex have the benefit that their stoichiometric matrix is the incidence matrix for their complexes graph. Thus, any cycles in their stoichiometry are also in their complexes graph and so they are deficiency zero.

In fact, linear CRNs are stronger than just deficiency zero, with, for example, solutions to their chemical master equation being Poissonian for all time, if their initial conditions are Poisson [85]. The discussion can now be split into two categories: linear CRNs with and without a null complex.

Firstly, linear CRNs without a null complex are guaranteed to have a conservation

of mass. Each connected component of their complex graph contains an independent set of chemical species. Thus in each connected component, a chemical species can only undergo a number of morphological changes and never increase the number of chemicals. Further, due to this conservation of mass, linear CRNs without a null complex behave like CTMCs, with states being the chemical species and transitions the reactions. At the deterministic level, their RREs for concentrations look like master equations for the probabilities in a CTMC. Thus solutions, with concentrations normalised by the total concentration, can be thought of as probabilities obeying a CTMC, allowing similar use of the matrix tree theorem (eqn. 2.8). On a stochastic level, with an initial condition containing n molecules, each of these molecules undergoes an independent realisation of the CTMC.

Linear CRNs with a null complex are slightly more complicated, but not by much. The null complex breaks the conservation of mass and so their RREs are no longer a CTMC master equation. However, their reaction rate equations can still be written as

$$\frac{d\mathbf{z}}{dt} = A\mathbf{z} + \mathbf{b}, \quad (2.48)$$

where A is a rate-matrix-like matrix of rate constants with off diagonal elements equal to rate constants for reactions between chemical species and diagonal elements equal to minus the sum of rate constants from a species. $\mathbf{b}$ is a vector of rate constants for reactions from the null complex to chemical species. Such an equation yields nice properties. Formally, the solution looks like

$$\mathbf{z}(t) = \exp(At) (\mathbf{z}(0) + A^{-1}\mathbf{b}) - A^{-1}\mathbf{b} \quad (2.49)$$

with the first term giving the transient trajectory and the last term giving the steady state. However, I shall go into more detail on the properties of such CRNs in chapter 5.

2.4 Stochastic thermodynamics

Finally in this chapter, I shall discuss stochastic thermodynamics. Arising from the need for a thermodynamic theory for describing non-equilibrium systems in which fluctuating quantities are important, stochastic thermodynamics proves particularly useful for studying biochemical systems [1].

In this section, I shall focus on the stochastic thermodynamics of (bio)chemical systems. As such, I shall consider the following setup where there is some system of interest coupled to a large environment. The environment and system together are isolated. The environment contains various components such as chemical buffers to supply chemical species or fuels; as well as being a heat bath to either supply or absorb energy as heat. This environment is constant in the sense that the concentrations of any chemical reservoirs and its temperature are not affected by the system of interest.

The system of interest has a microscopic state space fully defined by its microstates specifying the positions and momenta of all particles of the system. Such a microscopic description is, in general, not helpful. It is often more useful to consider a coarse-graining of the state space into a smaller number of macrostates, where each macrostate is a non-overlapping set of microstates. The microstates within a macrostate will generally have some defining feature which make them useful. For example, in a system consisting of two particles, one might consider a macrostate corresponding to a dimer. Such a macrostate would consist of any microstates where the two particles were within some distance of each other.

I shall assume some level of timescale separation between the microstates within a macrostate and the macrostates themselves, such that transitions between microstates within a given macrostate are much more likely than transitions between microstates in different macrostates. Such an assumption means that trajectories, upon entering a given macrostate, have enough time to equilibrate among the microstates within that macrostate. In such an equilibration, information about where the trajectory entered the macrostate is lost and which macrostate the trajectory visits next is independent of the

previous macrostates. In other words, the trajectory at the level of macrostates becomes Markovian [77, 86]. The details of the microstates can be mostly ignored, and the system can be treated as a CTMC $(\mathcal{X}, W)$ with a state space defined by the macrostates. As I shall only consider coupling to a single heat bath at a fixed temperature in this thesis, I use units such that temperature and Boltzmann's constant are both unity.

As will be seen in this section, in order to for the system to be amenable to thermodynamic analysis, the CTMC describing the system must be microreversible. Such a system is referred to as being thermodynamically consistent.

2.4.1 Generalised free energy and entropy production

There is a comprehensive framework of results for studying equilibrium thermodynamic systems. One wishes to extend this framework to study these far-from-equilibrium systems. As such one wishes to find generalised thermodynamic quantities as functions of the well defined thermodynamic quantities of each macrostate. Assume that there is a distribution over the macrostates which, at some time may, be written $p(i)$ where the index i runs over the macrostates. Certain quantities, such as the internal energy translate as expected to macrostates. The generalised internal energy of the system may be written as an average over the internal energies of the macrostates as $\mathcal{U}_{\text{sys}}[p(i)] = \sum_i p(i) U_{\text{sys}}(i) = \langle U_{\text{sys}} \rangle$. Such a decomposition is not guaranteed for all quantities. Importantly, considering the entropy of the system, one finds two contributions:

$$\mathcal{S}_{\text{sys}}[p(i)] = \sum_i p(i) S_{\text{sys}}(i) - \sum_i p(i) \ln p(i) = \langle S_{\text{sys}} \rangle + H[p(i)]. \quad (2.50)$$

The first term is the average entropy of the macrostates as one would expect. However, there is an additional informational term arising here which is the entropy of the distribution of macrostates. This informational term also feeds into the free energy of the system. As such, denoting this generalised free energy as $\mathcal{G}_{\text{sys}}[p(i)]$, it may be written [87]

$$\mathcal{G}_{\text{sys}}[p(i)] = \langle G_{\text{sys}} \rangle - H[p(i)], \quad (2.51)$$

where $\langle G_{\text{sys}} \rangle$ is the average free energy of the macrostates.

These generalised quantities described so far describe the ensemble of macrostates. Given that each of these quantities can be written as an expectation over macrostates, one may use the generalised quantities to define macrostate level (stochastic) quantities. The descriptor stochastic is used for these quantities as they are random variables whose averages give the ensemble quantities. At the level of these stochastic quantities, one is often most interested in changes along some trajectory starting in state i_0 and ending in state i_τ . One such quantity of interest is the total change in entropy of both the system and environment:

$$\Delta_{i_0 i_\tau} \sigma = \Delta_{i_0 i_\tau} S_{\text{env}} + \Delta_{i_0 i_\tau} S_{\text{sys}} - \ln \frac{p(i_\tau)}{p(i_0)}, \quad (2.52)$$

where $\Delta_{ij} Q = Q(j) - Q(i)$ is the change in quantity Q from state i to state j . The quantity σ is called the stochastic entropy production. Under the assumptions made here, that the system and environment are isolated from the rest of the universe, any increase in the internal energy of the system must be due to a depositing of heat from the environment. As such, at the level of macrostates, $\Delta_{i_0 i_\tau} U_{\text{sys}} = -\Delta_{i_0 i_\tau} S_{\text{env}}$. The stochastic entropy production is the negative of the stochastic free-energy change of the system, $\Delta_{i_0 i_\tau} \sigma = -\Delta_{i_0 i_\tau} G_{\text{sys}}$.

2.4.2 Time evolution and the second law

Equilibrium statistical mechanics tells us that the equilibrium distribution of macrostates is the Boltzmann distribution, $p_{\text{eq}}(i) \propto e^{-G(i)}$ where $G(i)$ is the free energy of state i . One of the defining features of equilibrium is that there are no net currents. For every pair of states, $W_{ij} p_{\text{eq}}(j) = W_{ji} p_{\text{eq}}(i)$, where the matrix element W_{ij} gives the transition rate for $j \rightarrow i$. Substituting in the Boltzmann distribution and rearranging, it can be seen that

$$\frac{W_{ij}}{W_{ji}} = \exp[-(G(i) - G(j))], \quad (2.53)$$

for a system which reaches equilibrium, the ratio of forwards to backwards transition is set by the free-energy difference between the states. This observation can be taken and

applied as an assumption for non-equilibrium systems, with a slight extension:

$$\frac{W_{ij}}{W_{ji}} = \exp [-(G_{\text{sys}}(i) - G_{\text{sys}}(j)) + \delta_{ij}], \quad (2.54)$$

where δ_{ij} is an additional driving term, biasing the transition due to fuel reservoirs in the environment. That the ratio of transition rates takes this form is known as local detailed balance. If the driving is zero for all reactions, the system is (globally) detailed balanced and will eventually relax to equilibrium. Local detailed balance is invaluable in interpreting the thermodynamics of chemical systems. If the drivings are not all zero, the system may still reach a steady state, known as a non-equilibrium steady state. Non-equilibrium steady states may have non-zero net currents, compared to detailed balanced steady states where no net currents may flow.

Assuming the trajectory starts in some known initial distribution, one can consider the stochastic entropy production (eqn. 2.52) to be a stochastic process up to time τ , whose average is the entropy production. Similarly one can consider the free-energy change up to time τ . One can look at the time evolution of the free-energy change. Making use of the master equation (eqn. 2.10) and local detailed balance (eqn. 2.54),

$$\frac{d\mathcal{G}_{\text{sys}}}{dt} = - \sum_{ij} W_{ij} p_j \ln \frac{W_{ij} p_j}{W_{ji} p_i} - \sum_{ij} W_{ij} p_j \delta_{ij}. \quad (2.55)$$

the negative of the first term

$$\dot{\sigma} = \sum_{ij} W_{ij} p_j \ln \frac{W_{ij} p_j}{W_{ji} p_i} \quad (2.56)$$

can be identified as the entropy production rate (EPR). Further, the form of the equation implies, mathematically, $\dot{\sigma} \geq 0$, which is the second law of thermodynamics. Further one can see that the EPR is zero if and only if $W_{ij} p_j = W_{ji} p_i$, which is equilibrium. As such the EPR gives an indicator of a non-equilibrium system.

If the driving δ_{ij} is zero, then one can see how the entropy production rate gives the

rate of decrease of generalised free-energy. However, when the driving is non-zero, this second term, $\sum_{ij} W_{ij} p_j \delta_{ij}$ is, in general, non-zero and gives the decrease in free energy of the fuel baths. This driving, therefore, can pay for an increase in the free energy of the system.

2.4.3 Entropy production and cycles

The EPR can be further split into two terms, an entropy change and a free-energy term,

$$\begin{aligned} \dot{\sigma} &= \underbrace{\sum_{ij} W_{ij} p_j \ln \frac{W_{ij}}{W_{ji}}}_{\text{free energy}} + \underbrace{\sum_{ij} W_{ij} p_j \ln \frac{p_j}{p_i}}_{\text{entropy}} \\ &= \frac{1}{2} \sum_{ij} J_{ij} \ln \frac{W_{ij}}{W_{ji}} + \frac{1}{2} \sum_{ij} J_{ij} \ln \frac{p_j}{p_i}, \end{aligned} \quad (2.57)$$

utilising the currents ($J_{ij} = W_{ij} p_j - W_{ji} p_i$). The free-energy term here can be seen as a sum over edges of the graphical representation of the CTMC. This sum over edges is also intimately linked to the notion of cycles [88], at least at steady state. The equivalence of this free-energy term calculated by summing over edges versus summing over cycles can be shown for a general process. Let $(\mathcal{X}, W)$ be the process, let w be the weight function for the graph induced, and let π_i be the steady state probability to be in state i for this process. For an edge $j \rightarrow i$, the net current at steady state through this edge is:

$$J_{ij}^{\text{ss}} = W_{ij} \pi_j - W_{ji} \pi_i. \quad (2.58)$$

π_i can be written in terms of spanning trees by MCTT (eqn. 2.14), and the sum over spanning trees may be expanded by SAWs from j to i . For π_j , the sum over spanning trees may be expanded by SAWs from i to j such the spanning tree terms of both expansions are the same and only the direction of edges in the SAW term are flipped. The net current

may then be written:

$$J_{ij} = \frac{1}{\mathcal{N}} \sum_{S \in \mathcal{S}(i \rightarrow j)} \left[W_{ij} \prod_{e \in S} w(e) - W_{ji} \prod_{e' \in S} w(e') \right] \left[\sum_{T \in \mathcal{T}_S(s)} w^S(e) \right] \quad (2.59)$$

where $\mathcal{S}(i \rightarrow j)$ is the set of SAWs from node i to node j ; remembering that $w(e)$ gives the weight (transition rate) of edge e ; $\mathcal{N}$ is the normalisation as in eqn. 2.14, and e' is the edge in the opposite direction, *i.e.* if $e = j \rightarrow i$, $e' = i \rightarrow j$; and the last bracketed term is the spanning tree part for SAW, S , as in eqn. 2.16. One of the SAWs from i to j will simply be the single transition $i \rightarrow j$. However, this term will cancel out from the sum leaving just the non-trivial SAWs. Taking a non-trivial SAW from i to j and multiplying by the rate W_{ij} gives a cycle containing the edge $j \rightarrow i$. Therefore, the current may be written as a sum over cycle currents, as in section 2.2.9, of cycles which contain the edge $j \rightarrow i$ minus those which contain $i \rightarrow j$.

$$J_{ij} = \sum_{\substack{\text{cycles } C \\ \text{containing } j \rightarrow i}} J(C) - J(C'), \quad (2.60)$$

where $J(C)$ is a cycle current defined analogously to eqn. 2.27 and C' is the cycle in the opposite direction.

For the EPR, each of the edges contains a contribution to the free-energy term $\ln \left(\frac{W_{ij}}{W_{ji}} \right)$. Since, the J_{ij} may be split up as a sum over cycles, the parts of this corresponding to given cycles may be collected and the sum over edges may be converted into a sum over cycles. Doing so, one finds the contribution to the free-energy term from cycle, C , to be $A(C)$, where

$$A(C) = \ln \prod_{e \in C} \frac{w(e)}{w(e')} \quad (2.61)$$

are known as the cycle affinities. Thus, the free-energy term has been rewritten in terms of cycles as

$$\dot{\sigma} = \sum_{\text{cycles } C} A(C)(J(C) - J(C')). \quad (2.62)$$

Whether or not it is easier to think about this free-energy term in terms of cycles or edges depends on the system.

Chapter 3

Copolymerisation models

In chapter 4, I shall introduce a general class of models of polymerisation. As such, it is necessary to review existing studies of polymerisation, to show how the models presented fit into the literature.

3.1 Types of copolymerisation

As mentioned in chapter 1, copolymers are polymers consisting of multiple types of monomer. A certain class of copolymerisation models I shall review here typically consist of a growing polymer where monomers are added or removed from the end to extend or shrink the polymer. Such models can be divided into three categories based on their interactions with a template. These categories are free copolymerisation, templated self-assembly and templated copolymerisation with simultaneous separation. I illustrate the difference between such models with figure 3.1(a, b, c). Here, I imagine an arbitrarily long polymer to the left, of which only showing the last few monomers are shown. Free copolymerisation involves no template and the polymer grows by new monomers binding directly to the growing polymer.

Templated self-assembly by comparison does make use of a template. For such models, new monomers bind to the template first and are subsequently polymerised into the growing copolymer. For templated self-assembly models, the separation of template and

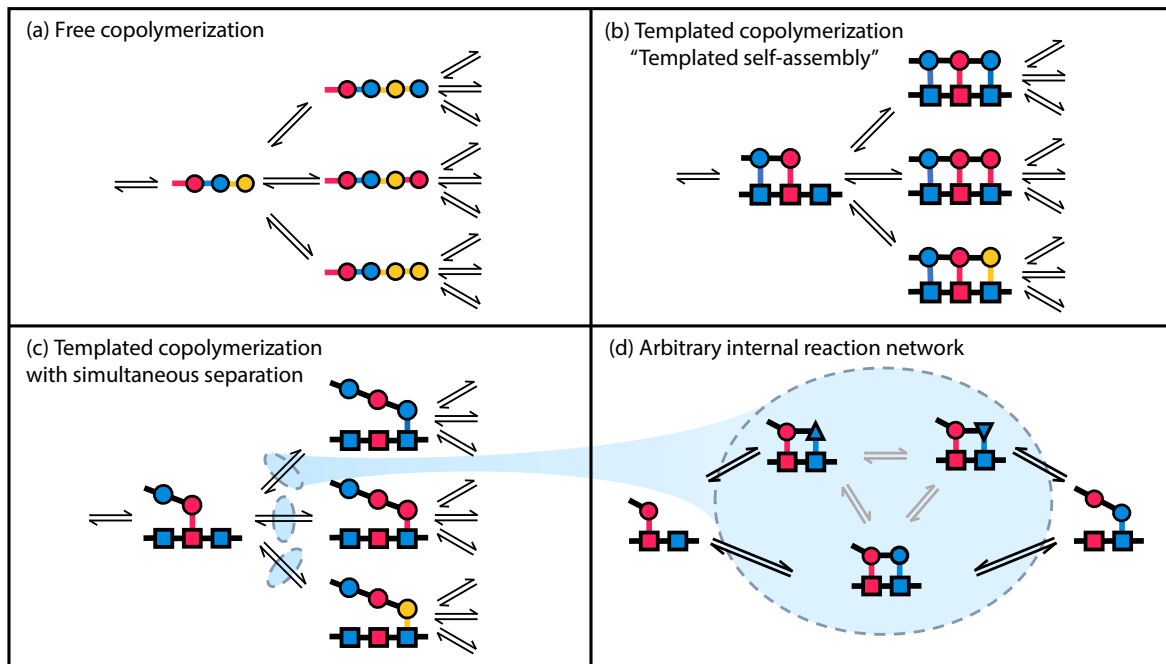


Figure 3.1: Figure reproduced from Qureshi *et al.* [89]. Comparison of the different types of copolymerisation mechanism with three types of monomer (blue, red and yellow). (a) shows free copolymerisation, (b) templated self assembly, and (c) templated copolymerisation with simultaneous separation. In (b) and (c) the template is shown with squares and the growing polymer with circles. (d) An example of a more detailed reaction scheme used to select the next monomer. In each of these sub-figures, different colours represent different monomer types, with bonds coloured accordingly when their strength might depend on the monomer type. In (d), different activation states of the monomer undergoing incorporation are represented by different shapes. The dashed bubble indicates how the arbitrary set of reactions in (d) may replace the simple reaction surrounded by a dashed bubble in (c) for a more complex model.

product is not directly coupled to the growth of the polymer. Separation may occur by some distinct mechanism after the copolymerisation has completed, but this separation process is not modelled. For example, in DNA replication, each of the newly-created daughter strands remains attached in a double-stranded duplex at the end of the process. Separation of the two strands of the double stranded DNA occurs is coupled to the next round of replication and is mediated by helicase enzymes [2].

Contrary to templated self-assembly, there is templated copolymerisation with simultaneous separation. These copolymerisation models do have separation of template and product coupled to the copolymerisation process such that at any time only a finite, small section of polymer is bound to the template. In figure 3.1(c), I depict the monomers bound to the template at one time as being only the last monomer of the growing copolymer. In general, however, models may consider the last few monomers being bound to the template.

The main differences between these categories of model occur in the free-energies of the states as the polymer grows and in correlations generated between monomers within the copolymer products. I shall describe these differences in more detail throughout this thesis.

Orthogonally, polymerisation models can also be described by the complexity of the sub-steps leading to the incorporation of monomers. A simple (CTMC) model of polymerisation will have a state space consisting of polymers of length l where transitions occur between polymers of length l and length $l \pm 1$. Such a model describes a growing polymer with monomers added or removed in a single step as in figure 3.1(a, b, c). However, a most general model of polymerisation should include complex sub-steps. Figure 3.1(d) depicts how a polymerisation model might have multiple sub-steps leading to the incorporation of monomers. However, models in the literature often neglect multiple sub-steps due to the complexity of the models making solving such models difficult.

One can also describe polymerisation models based on the dependence of rates on monomers. In the simplest models, the rates of monomer addition or removal from a growing polymer depend only on the monomer being added or removed (and potentially

the corresponding monomer in the template). More generally, the rates of monomer addition or removal might depend on monomers further back from the tip of the growing polymer. Again, such dependencies are often neglected in models in the literature due to the difficulty they bring to solving said models.

Finally, models can be split into categories based on whether monomers can be removed from a growing copolymer. If the monomers cannot be removed, the model has irreversible incorporation of monomers. As was seen in section 2.4, reversible reactions are required for thermodynamically consistent models. However, some of the models make use of irreversible reactions to simplify the mathematics.

All these categories of polymerisation models shall be described in more detail in this chapter and expanded on in chapter 4.

3.2 Free copolymerisation

The first copolymerisation models focused on free copolymerisation. Mayo and Lewis [15] wrote one of the earliest studies of copolymerisation. Here, they considered a chemical reaction network model for a growing copolymer of two monomer types, M_1 and M_2 . They considered chemical species representing polymers whose last monomer is M_1 or M_2 , and considered reaction rates such that the rate of adding a new monomer depends only on that new monomer and the last monomer of the growing polymer. Notably also, they considered irreversible reactions, such that once a monomer binds to the growing polymer, the monomer may not detach. Nonetheless, Mayo and Lewis were able to identify a relationship between the depletion rate of each monomer type and the reactivity ratios, defined for a polymer ending in M_1 as the ratio of the rate of adding monomer M_1 to the rate of adding M_2 and similarly for polymer ending in M_2 as the ratio of the rate of adding monomer M_2 to the rate of adding M_1 . Such a relationship is useful from an experimental perspective for measuring the reactivity ratios.

The reactivity ratios in turn are useful quantities as they relate to the lengths of runs of the same monomer within the polymer. As such one can extract some useful information

about the structure of the polymers such as the expected fraction of monomer one or two, or the expected lengths of regions of pure monomer one or two. These quantities, however, highlight a key aspect of free copolymerisation, that it is difficult to use this mechanism to produce information bearing polymers. If one's goal is to produce a copolymer with some general properties, which are not sensitive to variations in the specific sequence of monomers, then they may use mechanisms described by the models of Mayo and Lewis. However, if one wishes to produce a copolymer with a very specific sequence of monomers, such as those which transmit information, such a mechanism is not suitable.

An additional point to note about Mayo and Lewis' model is that they consider irreversible polymerisation of monomers. Hence, the models are not thermodynamically consistent. Further, such an assumption means that the statistics of monomers at the tip of the growing copolymer are the same as those far behind the growing tip. Such a distinction is important for the consideration of models with reversible binding, determining which mechanisms are most effective at reducing the probability of errors in the sequence of monomers.

Since Mayo and Lewis' early studies, thermodynamically consistent free copolymerisation models have been analysed [90,91]. Such models allow for detachment of monomers after polymerisation with rates defined by the binding free energies of the monomers. Similar to Mayo and Lewis, properties of the sequences of monomers produced could be analysed. However, as mentioned in chapter 1, if one wishes to encode a specific sequence, they need a large number of monomer types. However, even with a large number of monomer types, free copolymerisation is unsuitable for producing specific sequences of monomers. The sequence of monomers produced by such a mechanism is Markovian. So if ever a monomer which is not the one required by the desired sequence is polymerised into the copolymer, monomers later on in sequence of the copolymer produced will most likely not be correct.

3.3 Templated polymerisation models

Templated copolymerisation has been studied in much detail in the past two decades. The existence of a template in the form of another polymer adds a number of conceptual differences to copolymerisation models. The first of these conceptual differences is that monomers can be defined to be correct relative to the template. This relative correctness is different to free copolymerisation, where it makes most sense for monomers to be considered correct only relative to the other monomers in their vicinity within the daughter copolymer. A common assumption made in the literature related to this concept is a mapping of monomers to being right or wrong relative to the template [92]. Such a mapping is viable if one assumes a symmetry in the monomer interactions that the monomer interactions with the template only come in two types corresponding to right and wrong. As such these interactions cannot discriminate between different types of monomer which have the same interactions with the template. For example, one would not be able to make such an assumption for a full model of transcription where an A binding to an A in the template might have a different binding energy to a G binding to an A in the template even though both interactions are wrong. This assumption is made frequently in the literature as it greatly simplifies models and calculations, since one does not need to consider the sequence of monomers in the template. After such a mapping of the polymer sequence, a key quantity of interest is the error probability - the probability that a wrong monomer is present in the new copolymer produced. Being able to ignore the sequence of the template is a great advantage not only for studying models given a specific template sequence, but also since one does not need to consider the distribution of templates. However, frameworks do exist in which the sequence of monomers in the template is explicitly considered for example [93].

One of the earliest models introduced was that of Gaspard and Andrieux [94], whereby monomers would be added to or removed from a growing polymer. The models they considered encompassed different types of copolymerisation, both free and templated as in figure 3.1(a, b, c). Gaspard and Andrieux studied these models in various regimes [95,96].

The models considered allow for the dependence of monomer binding rates on previous monomers in the growing copolymer. Further, monomer binding was reversible, allowing for a thermodynamically consistent description. However, the binding of monomers occurred in a single, simple step. A full solution to the model (in the long polymer limit) was presented in [97]. This solution is important to this thesis going forwards, as such a more in depth review is given in section 3.5.

Sartori and Pigolotti provided a similar model for copolymerisation [98,99]. However, a key difference in Sartori and Pigolotti’s models is the consideration of arbitrary set of internal reactions leading to polymerisation, similar to figure 3.1(d). In these works, the authors considered polymers of just right or wrong monomers. They were able to solve for the error probability for various networks of reactions. The authors further presented thermodynamic analysis of their models, in one study considering the difference between kinetic discrimination, whereby right monomers bind/unbind more quickly than wrong monomers, and energetic discrimination, whereby right monomers bind more strongly to the template [98]; and in another study, considering the errors in various regimes of dissipation [99]. In both cases, the authors were able to utilise their models and analysis to study the trade-offs between dissipation, accuracy and speed key to these biological processes.

A key assumption to the solution presented by Sartori and Pigolotti was the independence of rates of reactions on all previous monomers. Such an assumption ensures that each monomer binds independently of previous monomers and therefore the probability of a given sequence of monomers factorises neatly into a product of probabilities for each site on the template. This assumption necessarily breaks down for templated polymerisation with simultaneous separation, as a physical implementation naturally leads to dependence of rates on previous monomers. However, even for templated self assembly one could imagine models whereby a wrong monomer at the tip affects the binding of subsequent monomers.

Throughout the literature, other such templated models have been developed. These models typically focus on templated self-assembly [92, 99–112], although, some models

looking at the effect of polymer separation have been analysed [40, 113, 114]. Many of these studies allowed for complex mechanisms leading to polymerisation [92, 99–104, 109–112] and considered the effects of reversible binding of monomers [99, 101–108, 113, 114]. However, only [113, 114] considered the effects of dependence of binding on previous monomers. Further, to the best of my knowledge, no previous studies have analysed reversible systems with complex sub-steps leading to polymerisation and dependence on previous monomers.

The types of assumptions or simplifications that are made will depend on the aims of the paper in question. More biologically focussed papers tend to have the goal of reproducing a specific model. For example, Sahoo and Klumpp [100, 115] study specific, complex mechanisms to probe certain biologically realistic enzymatic models.

More mathematically focussed papers tend to be interested in methodology for example the methods of Gaspard and Andrieux [97, 106–108, 116], which I shall discuss in more detail in section 3.5. Chiuchiù *et al.* [117] by contrast introduce a more novel methodology utilising continuous reaction-coordinate based mathematics. In contrast, the more physically focussed ones are often interested in the limits of polymerisation processes, for example in relation to the thermodynamic uncertainty relation [101, 104]; or with respect to trade-offs between speed, energy and accuracy [103, 110, 112]. However, there is often overlap in the aims of such studies, for example in introducing new, mathematically-interesting methodologies, for the purposes of probing physical questions on trade-offs and how models can produce qualitatively different behaviours with relation to accuracy and energy dissipation in different regions of parameter space [98, 99, 102]. Where this thesis aims to reside is discussed in more detail in section 3.7.

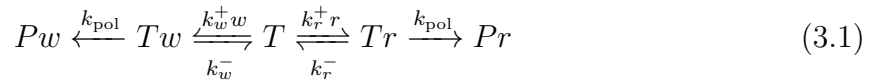
3.4 Kinetic proofreading

One might ask why complex sub-steps leading to polymerisation would be necessary and therefore an important aspect of models to consider. For systems for which the target structure is encoded as the lowest free energy state, one way to increase the accuracy

of self-assembly is by decreasing the free energy of said target state. For templated copolymerisation, such a strategy is not necessarily possible, the target products are not often the lowest free energy state. Additionally, the basic chemistry of the reactions may not allow for the free energies to become arbitrarily low. Further, the requirement that the template needs to separate from the product [40], means that increasing the binding energy of products to templates is not necessarily the best way to increase accuracy. Thus, one must look to more complex methods to increase accuracy of templated copolymerisation.

Hopfield [118] and Ninio [119] independently and contemporaneously proposed such a complex mechanism while studying translation. They noticed that the binding energy difference between correct and incorrect codons (which encode how amino acids bind to the mRNA template), was not enough to explain the accuracy observed in biological templating via models of a simple mechanism for polymerisation.

Consider a simple model for the selection of monomers as depicted in figure 3.2(a). Here for conceptualisation and illustrative purposes, I allow for irreversible reactions. Consider two monomer types, a right monomer and wrong monomer, each chemostatted (section 2.3.2) at concentrations r and w respectively. As such, the model may be represented as a CTMC:



where T represents the state of template and growing polymer up to this point, Tr and Tw represent a right or wrong monomer binding to said template-polymer complex, and Pr and Pw representing the template-polymer complex with the right or wrong monomer polymerised into the chain. The rate of polymerisation k_{pol} is assumed to be the same for both the right and wrong monomers. Local detailed balance implies $k_i^+/k_i^- = e^{-\Delta G_i}$, where ΔG_i is the free energy of binding of monomer i . I assume there is no additional driving. It is usual to assume the binding rates are the same for both right and wrong monomers $k_r^+ = k_w^+$ [118]. Hence the rate constants may be written $k_i^- = k e^{-\Delta G_i}$ for some

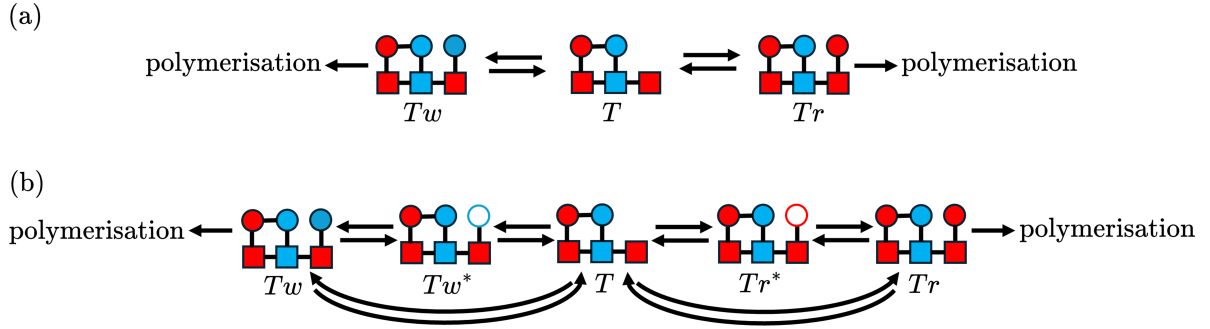


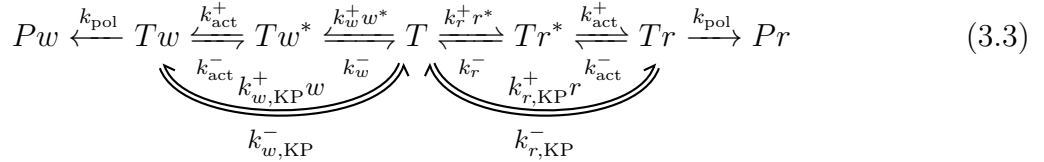
Figure 3.2: Comparison of (a) a simple polymerisation step to (b) a kinetic proofreading one. (a) provides a visual representation of the CRN in eqn. 3.1, and (b) a representation of the CRN in eqn. 3.4. In both cases, a small section of template and polymer is depicted. The template is depicted with squares and the polymer copy with circles with bonds represented by lines. The different monomer types are represented by different colours. Arrows represent allowed reactions. The monomer which binds could be either red or blue. In (a) monomers can bind and unbind until eventually a bound monomer is polymerised into the chain. For kinetic proofreading (b) inactivated monomers are represented by the empty circles for which the border colour represents the monomer type. From the T state, an inactive monomer can bind. When bound, this inactive monomer can either unbind or become activated. When there is an activated monomer bound to the template, this activated monomer can either unbind from the template or be polymerised into the chain.

monomer-independent rate constant k . Further, assume the right and wrong monomers are held at the same concentration. The question is then, given starting in state T , what is the probability of being absorbed into state Pr or Pw . Utilising the methods from chapter 2, the ratio of being absorbed into Pw to being absorbed into Pr can be found to be

$$\frac{\mathbb{P}[Pw]}{\mathbb{P}[Pr]} = \frac{k_{\text{pol}} + ke^{-\Delta G_r}}{k_{\text{pol}} + ke^{-\Delta G_w}}, \quad (3.2)$$

also known as the error ratio. This ratio is less than one so long as $\Delta G_r > \Delta G_w$, *i.e.* the right monomer binds more strongly to the template. Further, the ratio is maximised, for fixed binding energies, to $\exp(-(\Delta G_r - \Delta G_w))$ in the limit $k_{\text{pol}} \rightarrow 0$. The binding energy difference sets a limit on the maximum error ratio.

Compare this simple model to a more complex model whereby the monomers must first be activated before polymerisation as in figure 3.2(b). Such a model performs kinetic proofreading. Let us assume the non-activated monomers, are chemostatted at concentrations r^* and w^* and the activated monomers are chemostatted at concentrations r and w . We may similarly represent this more complex model as a CTMC:



Here, there are additional states, Tr^* and Tw^* for the template bound to non-activated monomers. The activation of the monomers occurs via the reactions with rate $k_{\text{act}}^\pm$. Assume these reactions are the same for both monomer types and are driven such that $k_{\text{act}}^+/k_{\text{act}}^- = e^{\Delta G_{\text{act}}}$. This driving could come from, for example, coupling to the conversion of ATP to ADP, which is not explicitly modelled. The reactions labelled KP are for the binding and unbinding of activated monomers. Assume the activation of the monomer does not affect the binding energy between monomer and template such that $\frac{k_{i,\text{KP}}^+}{k_{i,\text{KP}}^-} = e^{-\Delta G_i}$. Similarly to before, assume the binding rates are the same, $k_{r,\text{KP}}^+ = k_{w,\text{KP}}^+ = k_{\text{KP}}^+$.

Analogously to the simple model, the error ratio may be calculated to be

$$\frac{(k_{\text{act}}^+ k w^* + k_{KP} w (k e^{-\Delta G_w} + k_{\text{act}}^+))}{(k_{\text{act}}^+ k r^* + k_{KP} r (k e^{-\Delta G_r} + k_{\text{act}}^+))} \frac{(k_{\text{act}}^- k e^{-\Delta G_r} + (k_{\text{pol}} + k_{KP} e^{-\Delta G_r})(k e^{-\Delta G_r} + k_{\text{act}}^+))}{(k_{\text{act}}^- k e^{-\Delta G_w} + (k_{\text{pol}} + k_{KP} e^{-\Delta G_w})(k e^{-\Delta G_w} + k_{\text{act}}^+))}. \quad (3.4)$$

Let us assume that right monomers are held at the same concentration as wrong ones, such that $r = w$ and $r^* = w^*$. Then, first take the limits that $r, w \rightarrow 0$ assuming the cell is able to break down the activated monomers that are not bound to the template, quickly; and the limit $k_{\text{act}}^- \rightarrow 0$ ($\Delta G_{\text{act}} \rightarrow \infty$) perhaps by the cells quickly removing ADP after being formed from the breakdown of ATP. After this, consider the limit that $k_{\text{pol}}, k_{\text{act}}^+ \rightarrow 0$, *i.e.* binding or unbinding of monomers is much faster than activation or polymerisation. In these limits, the error ratio reduces to $\exp(-2(\Delta G_r - \Delta G_w))$ which is the square of error ratio from the simple case. Thus, by forcing the monomers to go through two discrimination steps, where they can unbind from the template either before activation (via $k e^{-\Delta G_i}$) or after (via $k_{KP} e^{-\Delta G_i}$), the effective discrimination between binding energies has been doubled. This doubling of discrimination is only achieved in this limit of irreversible and slow reactions. Away from this limit, enhanced accuracy can still be achieved, but will be reduced.

The kinetic proofreading motif has introduced a cycle of reactions compared to the simple model. This cycle has a non zero affinity, both due to the driving ΔG_{act} and the irreversible reactions caused by chemical species held at zero concentration. As such, the system has non-zero heat flow and the kinetic proofreading cycle makes a trade-off, increasing the accuracy of polymerisation at the cost of fuel consumption. This idea was revolutionary in introducing a mechanism whereby increased accuracy is achieved in exchange for more free energy. As such, the idea inspired many others to study the trade-offs between free-energy cost and accuracy; and to produce and analyse more complicated mechanisms which utilise such free-energy consuming cycles to increase accuracy. Since its first description, there have been many models developed to study the effects of kinetic proofreading in polymerisation [92, 99, 101–103, 110–112]. Indeed, kinetic proofreading serves as an excellent first example system to study in any framework attempting to

analyse general complex polymerisation models.

3.5 Solving polymerisation models

The models discussed thus far only consider the rates of monomer inclusion to depend on the monomer being added. In particular, for free copolymerisation (figure 3.1(a)) and templated self-assembly with separation (figure 3.1(c)) dependence of rates on previous monomers is a key feature. Thus here I shall review and reproduce some work by Gaspard and Andrieux [97], which will serve as the base I shall extend in chapter 4.

For simplicity, I shall review the method for the case of two monomer types, right r and wrong w . As mentioned in section 3.3, the mapping of templated polymerisation sequences to right and wrong monomers is common. Consider a CTMC with state space given by $\{x_1x_2\cdots x_l \mid x_i \in \{r, w\}, l \in \mathbb{N}\}$, that is the sequences of r and w of l elements where l is a natural number. l can be zero for the initial, empty sequence $\emptyset$. These sequences represent polymers of length l with monomers x_i . I shall use the notation $\&$ to refer to an arbitrary sequence with which we are not concerned. For example, $\&x_{l-1}x_l$ is an polymer with the last two monomers being $x_{l-1}x_l$, but arbitrary monomers before this tip. Transitions occur between polymers of length l and $l + 1$ as:

$$\&x_l \xrightleftharpoons[\omega_{-x_{l+1}x_l}]{\omega_{+x_{l+1}x_l}} \&x_lx_{l+1}, \quad (3.5)$$

only depending on the two monomers at the end of the polymer. The methodology can easily (albeit messily) extended to dependence on the last k monomers at the tip of polymer [116], although such an extension is not discussed in detail in this thesis. The initial reaction to and from $\emptyset$ in general are not important for the longtime behaviour to be considered in this thesis. With state space and transitions defined, the CTMC has

been defined. The master equation for this CTMC can be written down as:

$$\begin{aligned} \frac{d}{dt} \mathbb{P}_t[x_1 \cdots x_l] &= \omega_{+x_l x_{l-1}} \mathbb{P}_t[x_1 \cdots x_{l-1}] + \sum_{y=r,w} (\omega_{-y x_l} \mathbb{P}_t[x_1 \cdots x_l y]) \\ &- \left(\omega_{-x_l x_{l-1}} + \sum_{y=r,w} \omega_{+y x_l} \right) \mathbb{P}_t[x_1 \cdots x_l], \end{aligned} \quad (3.6)$$

where $\mathbb{P}_t[x_1 \cdots x_l]$ is used to refer to the probability of being in state $x_1 \cdots x_l$ at time t . Such a system of equations is difficult to work with in general.

The length of the polymer undergoes a biased random walk such that the expected length of the polymer grows linearly with some average velocity v . Thus, a simplifying assumption can be made that the sequence of monomers reaches some stationary distribution, $\mu(x_1 \cdots x_l)$ such that the probability may be split as

$$\mathbb{P}_t[x_1 \cdots x_l] \approx \mu(x_1 \cdots x_l) p_t(l), \quad (3.7)$$

where $p_t(l)$ is probability of being any polymer of length l ,

$$p_t(l) = \sum_{x_1, \dots, x_l} \mathbb{P}_t[x_1 \cdots x_l]. \quad (3.8)$$

Importantly, all of the time dependence is in the length probability. The factorisation of the probability defines a probability distribution over polymers of a given length, $\mu(x_1 \cdots x_l)$. Summing this distribution over the first $l - n$ elements of the polymer gives a hierarchy of probabilities, which in an abuse of notation, we denote all as μ , such that $\mu(x_{l-(n-1)}, \cdots x_l)$ gives the probability to be in a state with the n monomers at the end of the polymer specified. The key assumption that allows for the solving of these models is that, for large enough l , this hierarchy is independent of l . Thus, this distribution gives a new way to view the copolymerisation process, focusing on the monomers at the tip of the growing polymer. I shall refer to these probabilities as the tip distribution.

Substituting the factorisation of eqn. 3.7 into the master equation eqn. 3.6 and sum-

ming over sequences, gives a master equation for the length of a copolymer with rates in terms of the original rates of the process $\omega_{\pm yx}$ and the tip distribution. The resultant set of ODEs are linear and thus can be solved as a linear superposition of functions of the form $\exp(iql + s_q t)$ for $q \in [-\pi, \pi]$ and s_q a function to be determined. A quick calculation, assuming an initial condition with length $l = 0$ and assuming the expected length grows linearly $\mathbb{E}_t[l] = vt$, shows $\left. \frac{\partial s_q}{\partial q} \right|_{q \rightarrow 0} = -iv$. Thus, the growth velocity may be read off in terms of $\omega_{\pm yx}$ and μ as

$$v = \sum_{x,y} \omega_{+yx} \mu(x) - \omega_{-yx} \mu(x, y). \quad (3.9)$$

This relation I shall come back to.

I now need to solve for the tip distribution. I shall look for Markovian solutions for the tip distribution, such that

$$\mu(x_{l-n}, \dots, x_l) = \mu(x_{l-n}|x_{l-n+1}) \cdots \mu(x_{l-1}|x_l) \mu(x_l), \quad (3.10)$$

where $\mu(x|y)$ are conditional probabilities that a given monomer is x given that the one after it in the polymer is y . By considering the $q \rightarrow 0$ contribution of $\exp(iql + s_q t)$ to the master equation (eqn. 3.6); summing over the first $l - 1$ monomers in the sequence; separately only summing over the first $l - 2$ monomers; and utilising the Markov property, we arrive at a set of equations only in terms of $\mu(x)$ and $\mu(x|y)$ and the rates. These equations fully specify the tip probabilities and can in theory be solved although this is messy and cumbersome (hence they are not stated here). Instead, I shall do some more work to set up a workable procedure for solving the polymerisation processes.

Define partial velocities v_x as the rate of growth of the polymer given there is a monomer x at the tip, *i.e.* $v = \sum_x v_x \mu(x)$. These partial velocities can be read off from the full velocity (eqn. 3.9), by looking at the coefficients of $\mu(x)$. Substituting these partial velocities into the relations for the tip probabilities gives a set of equations for the

conditional tip probabilities

$$\mu(x|y) = \frac{\omega_{+yx}\mu(x)}{(\omega_{-yx} + v_y)\mu(y)}. \quad (3.11)$$

This relation can in turn be re-substituted into the partial velocities to give a closed set of equations for the partial velocities only in terms of the rates:

$$v_x = \sum_y \frac{\omega_{+yx}v_y}{(\omega_{-yx} + v_y)}. \quad (3.12)$$

Finally, rearranging eqn. 3.11 and summing over x gives

$$\mu(y) = \sum_x \frac{\omega_{+yx}}{(\omega_{-yx} + v_y)}\mu(x). \quad (3.13)$$

For rates and partial velocities both known, this equation is linear in $\mu(x)$.

The tip probabilities, however, are not the quantity of most interest. The quantity of interest for us is the probability for a given monomer to be in the main body of the polymer, far behind the tip, as this will dominate the final state of the full polymer. As stated, the tip probabilities form a Markov chain, moving backwards from the tip of the polymer. Thus, the steady state of this Markov chain (defined with transition probabilities $\mu(x|y)$ for a monomer x behind a monomer y) gives us these probabilities to find given monomers in the body of the polymer. I refer to these as the bulk probabilities and denote them by ε . Using eqn. 3.11 and the eigenvalue equation for the steady state of a Markov chain, it can be seen that $\varepsilon(x)/\mu(x) \propto v_x$. Finally, by normalisation, it can be seen that the bulk probabilities may be written

$$\varepsilon(x) = \frac{\mu(x)v_x}{\sum_y \mu(y)v_y} = \frac{\mu(x)v_x}{v}. \quad (3.14)$$

Thus, I have reviewed a procedure for solving for the bulk probabilities in terms of the rates. Firstly, solve for the partial velocities using eqn. 3.12. Then, one may solve

for the tip probabilities μ using eqn. 3.13. Finally, one may combine them into the bulk probabilities eqn. 3.14. I shall explore more of the properties of these solutions later in chapter 4 as well as certain limits and my extension.

3.5.1 Thermodynamics of copolymerisation

Having found the bulk probabilities $\varepsilon(x)$, one can ask how the accuracy depends on the free energies of the polymerisation process. A key quantity for this thermodynamic analysis is the entropy rate, defined as

$$H = \lim_{n \rightarrow \infty} \frac{1}{n} H[\mu(x_1 \cdots x_n)] \quad (3.15)$$

where $H[\mu(x_1 \cdots x_n)]$ is the Shannon entropy of the polymer distribution. Since the produced polymer is a Markov chain, this entropy rate can be calculated as [63]

$$\begin{aligned} H = & - \varepsilon(\varepsilon_w \ln \varepsilon_w + (1 - \varepsilon_w) \ln(1 - \varepsilon_w)) \\ & - (1 - \varepsilon)(\varepsilon_r \ln \varepsilon_r + (1 - \varepsilon_r) \ln(1 - \varepsilon_r)), \end{aligned} \quad (3.16)$$

where ε_i is the probability of an error (wrong monomer) after a monomer i . The entropy rate gives the average entropy each monomer contributes to the uncertainty of the polymer. Since this entropy rate is bounded by $\ln 2$ and the goal of these polymerisation system is to produce a polymer with low error, $\ln 2 - H$ gives a measure of the accuracy of the polymer.

The increase in entropy production per net monomer added to the polymer takes the form $\Delta G_{\text{pol}} + H + \text{H.O.T}$, where H.O.T are the higher order terms in the length L of the polymer. These higher order terms include a contribution to the uncertainty in the length. Since the length of the polymer obeys a central limit theorem with variance which scales linearly in average length, there is a contribution which is proportional to $\ln(L)/L$. There will also be energetic contributions due to the binding energy of the tip monomer to the template although these contributions scale as $1/L$.

Additionally, one can define the free-energy change per forward polymerisation step ΔG_{pol} . For simple polymerisation models, this polymerisation free energy will consist of the binding energy of the backbone of the polymer as well as some contributions from the concentrations of chemical species such as monomers, and potentially any permanent interactions with the template.

In the limit of long polymers which are considered in this thesis, the change in the entropy production per net monomer added is $\Delta G_{\text{pol}} + H$. This entropy change needs to be positive, by the second law, for a growing polymer, such that $\Delta G_{\text{pol}} + H \geq 0$. Adding $\ln 2$ to both sides of the inequality and rearranging, it can be seen that $\ln 2 - H \leq \ln 2 + \Delta G_{\text{pol}}$. The left hand side of this inequality is a measure of the accuracy of the polymerisation system. One wishes to make this term as large as possible for an accurate system. The right hand side of the inequality is a measure of how much free energy is being put into the system to drive the creation of polymers. $-\ln 2$ is also the minimum free energy per monomer required to assemble a polymer at equilibrium. Thus this inequality may be used to define a efficiency of polymerisation [113]

$$\eta = \frac{\ln 2 - H}{\ln 2 + \Delta G_{\text{pol}}} \in [0, 1], \quad (3.17)$$

which gives a measure of how much of the free energy put into driving the system corresponds to an increase in accuracy.

3.6 Simple copolymerisation with separation of template and product

An important study into polymerisation which I shall be building on in this thesis came from [113]. Here, Poulton *et al.* considered a thermodynamically consistent, simple class of models of templated polymerisation explicitly focusing on the effects of simultaneous separation. They were particularly interested in correlations generated within the copolymer sequence caused by coupling separation of polymer and template to the growth of

the polymer.

Poulton *et al.* wished to find a set of reaction rates which would allow them to emulate the statistics of a physical model of copolymerisation with simultaneous separation. As such, they sought out single set reaction rates with ratios of the forwards to backwards considered describing the free energy changes in the addition or removal of monomers. They noticed by these free energy changes that the rates naturally had to depend on previous monomers. Thus, they considered the CTMC described by Gaspard and Andrieux [97] with rates given by

$$\begin{aligned}\omega_{+rr} &= e^{\Delta G_K}, & \omega_{-rr} &= e^{\Delta G_K} e^{-\Delta G_{\text{pol}}} \\ \omega_{+wr} &= 1, & \omega_{-wr} &= e^{-\Delta G_{\text{pol}}} e^{\Delta G_{\text{TT}}} \\ \omega_{+rw} &= e^{\Delta G_K}, & \omega_{-rw} &= e^{-\Delta G_{\text{pol}}} e^{-\Delta G_{\text{TT}}} \\ \omega_{+ww} &= 1, & \omega_{-ww} &= e^{-\Delta G_{\text{pol}}},\end{aligned}$$

introducing three free energies ΔG_{pol} for the polymerisation of a monomer into the chain; ΔG_K which acts to speed up reactions for the binding/unbinding of a new right monomer; and ΔG_{TT} as the free energy binding difference between right and wrong monomers. I shall revisit these reaction rates in chapter 4. There I shall describe how to derive such rates from a microscopic model describing the reactions, after I have introduced the methodology.

With these rates defined, Poulton *et al.* were able to use Gaspard and Andrieux's method [97] to solve for the probabilities of finding monomer w in the bulk of the polymer. Such a probability is the error probability ε . Additionally, since the sequence of monomers in the polymer is itself a (stationary) Markov chain with states being r and w , one just needs to calculate transition probabilities ε_w and ε_r , defined as the probability of observing a wrong monomer after a wrong monomer (ε_w) or after a right monomer (ε_r), to specify the statistics of the sequence of monomers. In particular, one can calculate the entropy rate of the polymer. With the entropy rate and the polymerisation free energy, ΔG_{pol} , they were able to study the efficiency of such polymerisation models.

The analysis presented by Poulton *et al.* [113], gave a good description of a simple polymerisation model. However, they were not able to analyse more complex polymerisation processes such as those which include proofreading. Thus, in chapter 4, I shall extend the models and methodology of Gaspard and Andrieux, and Poulton *et al.* to be able to account for more complex polymerisation processes.

3.7 Outline of the rest of the thesis

As has been seen throughout this section, there is a wide literature analysing different polymerisation models. Each of these previous studies into copolymerisation needed to make various restrictive assumptions which prevented them from being able to analyse generic polymerisation models. For example, these studies might assume simple steps leading to the polymerisation of monomers; they might assume irreversible reactions, making their models not amenable to thermodynamic analysis; or they might assume the rates of addition of new monomers is independent of monomers already polymerised into the copolymer. The main aims of this thesis are to be able to relax these assumptions.

In chapter 4, I shall introduce a framework which allows for the analysis of generic polymerisation models with arbitrarily complex sub-steps leading to monomer incorporation; fully reversible reactions; and dependence on previous monomers in the polymer. Further, within this framework, one is able to extract not only statistical quantities such as the probabilities of monomers in the polymer bulk, but also thermodynamic and kinetic quantities such as free energies and polymerisation speed. Finally, I show that this framework may be used to analyse some rather complex example systems.

In chapter 5, I address the questions of how multiple polymerisation/depolymerisation pathways (of arbitrary complexity) affect the information propagation of said systems. In particular, I am able to derive fundamental bounds on how the accuracy of these multi-pathway polymerisation systems depends on free-energy drivings along each of the pathways.

Chapter 4

A method for the analysis of generic polymerisation models

The following chapter has been adapted from the publication “A universal method for analyzing copolymer growth” [89] and expanded to fit this thesis. The content included in this this chapter is the work of myself, Benjamin Qureshi.

4.1 Introduction

As has been seen so far in chapter 3, there is a need for a framework in which we may solve thermodynamically consistent models, which allows both for the dependence of binding rates on previous monomers and for arbitrary networks of reactions leading to polymerisation (figure 3.1(d)). In this chapter, I introduce such a framework. The framework presented here makes use of a novel coarse-graining which preserves steady state properties of the processes. However, the fine-grained details such as entropy production and time taken, while initially hidden by the coarse-graining, are not lost and can be recovered afterwards. It is important to emphasise that to the best of my knowledge this is the first such framework which allows for analysis of polymerisation models with all these features. Further, I believe this framework is mathematically pleasing, allowing for

analytical results in many cases. Finally, we shall see in section 5.8.2 that this framework can be used to analyse large, complex polymerisation mechanisms.

In section 3.5, I reviewed a method presented by Gaspard and Andrieux [97] to analyse polymerisation processes in which each monomer is added in a single step (*i.e.* if the internal reaction network shown in figure 3.1(d) were trivial), assuming long polymers. I shall present a method for mapping more complex models for the individual polymerisation step onto coarse-grained descriptions that can be analysed using this framework, and then subsequently show how to back out the behaviour of the full model from the results. As will be seen, the techniques reviewed in chapter 2 for solving absorbing Markov chains, will prove essential both for this mapping and backing out the full behaviour.

Consider a growing copolymer with M monomer types, which are assumed to be chemostatted in the environment. At a coarse-grained level, define a state space of finite length sequences $\{x_1x_2\cdots x_l \mid x_i \in \{1, 2, \cdots M\}, l \in \mathbb{N}\}$, where l is the length of the sequence. I shall refer to the coarse-grained states in this state space as *completed states*. On this coarse-grained level, a sequence of length l may increase in length by one unit by polymerising one of M units at the growing tip ($x_1x_2\cdots x_l \rightarrow x_1x_2\cdots x_lx_{l+1}$), or it may decrease in length by one unit ($x_1x_2\cdots x_l \rightarrow x_1x_2\cdots x_{l-1}$). Such a coarse-grained model is depicted in figure 3.1(a, b, c) for free polymerisation, templated self-assembly and templated polymerisation with simultaneous separation.

In general, copolymerisation processes may be best described by models in which the underlying copolymerisation reaction networks are complex, featuring multiple sub-steps in arbitrarily complex (connected) networks connecting the completed states, as suggested in figure 3.1(d). Hence, overall, one could consider a copolymerisation process as having a tree-like structure with networks of reactions connecting completed states, as in figure 4.1. Such a class of models is wide-reaching, with many examples from the literature included in this class [15, 91, 92, 99–104, 109–113, 115, 120–127]. Note, the generality of the models that are being analysed here. There are only two main restrictions that must be applied to the arbitrary reaction networks, namely that they must be connected and that the network connecting a polymer in state $\&x$ and one in state $\&xy$ (where $\&$ means an

arbitrary polymer before the specified tip monomers) only depends on monomers x and y . This second restriction corresponds to the assumption that the rate of adding a monomer y to a polymer ending in x can only depend on these two monomers in line with [15,90,91,97,105–108,113,124,125,128]. However, as shall be mentioned in chapter 6, this second assumption could be dropped at the expense of simplicity. Note then that the reaction network connecting states $\&x$ and $\&xy$, and the one connecting states $\&x$ and $\&xz$ for $y \neq z$ need not even share the same graphical structure (*i.e.* nodes and edges). However, for additional symmetry, the examples presented in this thesis do share the same graphical structure between different monomer types, with only the reaction rates changing.

I shall define a Markov process at the level of the coarse-grained completed states that, by construction, preserves probabilities of transitions between the completed states of the fine-grained process, and therefore preserves the statistics of the sequences produced. The coarse-grained Markov process does not preserve the distribution of transition times between completed states implied by the fine-grained model, which will, in general, be non-Markovian. Moreover, it does not provide fine-grained information on trajectories between the coarse-grained completed states. However, temporal details and information about the fine-grained dynamics can be added back in at a later stage, once statistics have been analysed at the coarse-grained level.

I shall demonstrate that this method is indeed able to reproduce prior results for simple polymerisation models as well as show how the results simplify for polymerisation models with certain assumptions on the nature of the dependence on previous monomers. Additionally, I shall exhibit this method for the analysis of a Hopfield-style proofreading network and an extension to Hopfield proofreading networks with N activation stages. Through this analysis, I am able to derive results for these proofreading networks as a function of N . With these results, statements can be made about said proofreading networks such as optimal N for minimising errors.

4.2 Identifying propensities of coarse-grained models

The transition rates of the coarse-grained model (hereafter labelled propensities to avoid confusion with the underlying rates of the fine-grained process) can be found by considering first passage problems between completed states. From a given completed state, there are $M + 1$ completed states that may be reached, corresponding to the M possible additions of a monomer and the removal of the monomer currently at the tip of the copolymer. For a first passage problem, each of these reachable completed states can be converted into an absorbing state by removing the transitions out of said state, as in figure 4.1(a), in the same vein as Cady and Qian [129]. I shall refer to this absorbing Markov process as the *step-wise process* and define step to mean the addition/removal of a monomer.

Since, I shall work with the assumption that the transition rates depend on the two monomers at the growing tip of the copolymer, there will therefore be M^2 flavours of this process corresponding to the combinations of the two terminal monomers of the copolymer, comprising the central state $\&xy$ (recalling $\&$ represents an arbitrary sequence). One then wishes to find the absorbing probabilities, $\mathbb{P}[\&xy \rightarrow \&xyz]$, $z \in \{1, \dots, M\}$, and $\mathbb{P}[\&xy \rightarrow \&x]$ given an initial condition of the central state $\&xy$. As outlined in section 2.2.5 (in particular eqn. 2.22), these probabilities can be found by constructing the closed process and finding sums over spanning trees rooted at different states.

A useful result I can state and prove here regards a different way to view the numerator of eqn. 2.22. Ignoring a normalisation inside the π^σ , this numerator is a sum over all edges that originally lead to an absorbing state A in the unclosed process. In each of the terms of this sum is the sum over spanning trees rooted at the tail of this edge. The alternate way to view this numerator, again ignoring the normalisation in π^σ , is by unravelling those edges that lead to absorbing state A and then finding the spanning trees rooted at the node A . More precisely, let $\mathcal{G} = (\mathcal{X} \cup \mathcal{A}, \mathcal{E})$ be an absorbing process, with $\mathcal{A}$ the set of absorbing states. Let $\mathcal{G}^\sigma$ be the closed process for starting state $\sigma \in \mathcal{X}$. Suppose the full set edges leading to a particular absorbing state in $\mathcal{G}$ is $\{x_1 \rightarrow A, \dots, x_n \rightarrow A\}$

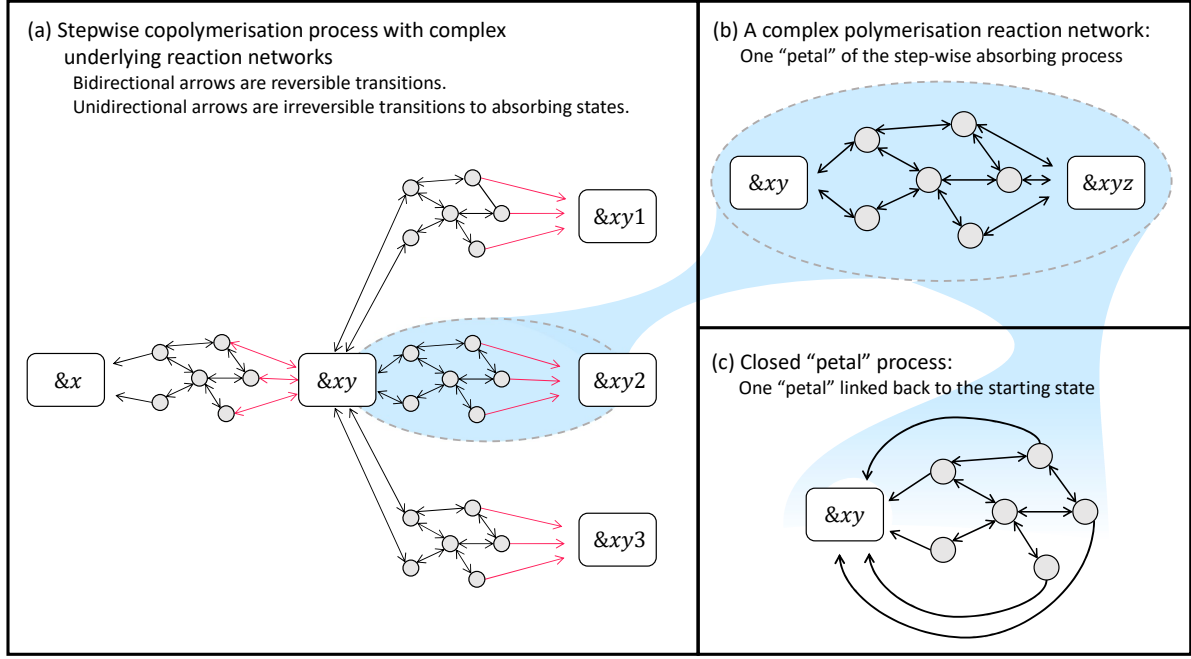


Figure 4.1: (a) The step-wise process for an arbitrary model with 3 monomer types. This step-wise process is for a copolymer $&xy$. The edges coloured red are the *completion* edges. The flower like structure of the step-wise process can be seen with 4 petals each connected at the starting state, $&xy$. (b) One of the petals of the step-wise process, which we use to define $\Lambda^\pm(z, y)$. $\Lambda^+(z, y)$ is defined as the sum of spanning trees rooted at the rightmost state, $&xyz$ and $\Lambda^-(z, y)$ the sum of trees rooted at the leftmost state, $&xy$. Note that in (b) arrows which are irreversible in (a) are depicted as reversible. This reversibility does not cause any issues as the would be reversed reactions do not appear in any of the sums over spanning trees. However, including the reversed reactions allows for the symmetry between forwards and backwards petals to be seen. (c) One of the petals (connecting $&xy$ to $&xyz$, as in (b)) which has been linked back to the starting state. This graph is used to define $Q(z, y)$ as the sum of spanning trees rooted at the leftmost state, $&xy$.

for $x_1, \dots, x_n \in \mathcal{X}$, $A \in \mathcal{A}$. Further, let $\mathcal{N} = \sum_{j \in \mathcal{X}} \sum_{T \in \mathcal{T}^\sigma(j)} w(T)$ where $\mathcal{T}^\sigma(j)$ is the set of spanning trees rooted at state $j \in \mathcal{X}$ in the closed process. The numerator of eqn. 2.22 can then be written:

$$\frac{1}{\mathcal{N}} \sum_{i=1}^n W_{Ax_i} \sum_{T \in \mathcal{T}^\sigma(x_i)} w(T). \quad (4.1)$$

The $\frac{1}{\mathcal{N}}$ cancels with equivalent normalisations in the denominator and thus can be ignored here. Thus, we wish to prove:

$$\sum_{i=1}^n W_{Ax_i} \sum_{T \in \mathcal{T}^\sigma(x_i)} w(T) = \sum_{T \in \mathcal{T}^{\sigma,A}(A)} w(T), \quad (4.2)$$

where $\mathcal{T}^{\sigma,A}(A)$ is the set of spanning trees rooted at state A for the closed process where the edges leading to state A have been unfurled, (*i.e.* the original process with only $\{\sigma\} \cup \mathcal{A}/\{A\}$ collapsed to a single state).

To prove eqn. 4.2, I shall provide a bijection between trees in $\bigcup_{i=1}^n \mathcal{T}^\sigma(x_i)$ and those in $\mathcal{T}^{\sigma,A}(A)$. Consider a tree $T_i = (\mathcal{X}, \mathcal{E}^\sigma \cup \mathcal{E}_A) \in \mathcal{T}^\sigma(x_i)$, where $\mathcal{E}_A = \{x_{i_1} \rightarrow \sigma, \dots, x_{i_m} \rightarrow \sigma\}$ for $\{x_{i_1}, \dots, x_{i_m}\} \subseteq \{x_1, \dots, x_n\}/\{x_i\}$. $(\mathcal{X}, \mathcal{E}^\sigma)$ is then a spanning forest¹ rooted at $\{x_i, x_{i_1}, \dots, x_{i_m}\}$. Let $\mathcal{E}'_A = \{x_i \rightarrow A, x_{i_1} \rightarrow A, \dots, x_{i_m} \rightarrow A\}$, then $T_A = (\mathcal{X}, \mathcal{E}^\sigma \cup \mathcal{E}'_A)$ is a spanning tree in $\mathcal{T}^{\sigma,A}(A)$. Thus, swapping the set of edges $\mathcal{E}_A$ for $\mathcal{E}'_A$ maps a tree in $\bigcup_{i=1}^n \mathcal{T}^\sigma(x_i)$ to one in $\mathcal{T}^{\sigma,A}(A)$. To go back from a tree in $\mathcal{T}^{\sigma,A}(A)$ to one in $\bigcup_{i=1}^n \mathcal{T}^\sigma(x_i)$, just replace $\mathcal{E}'_A$ by $\mathcal{E}_A$. For a tree in $\mathcal{T}^{\sigma,A}(A)$, one needs to identify that one of the edges in $\mathcal{E}'_A$, is the special edge equivalent to the $x_i \rightarrow A$. To identify this edge, note there exists a unique path from σ to A since it is a spanning tree. The final edge of this path will be the special edge $x_i \rightarrow A \in \mathcal{E}'_A$ which has no equivalent in $\mathcal{E}_A$ and such that x_i is the root of the spanning tree to which the tree is mapped (*i.e.* identifying which of the $\mathcal{T}^\sigma(x_i) \subseteq \bigcup_{j=1}^n \mathcal{T}^\sigma(x_j)$ to which the mapped tree belongs). Further, $\frac{w(T_A)}{w(T_i)} = w(x_i \rightarrow A)$, and so the equivalence of sum over spanning trees (eqn. 4.2) holds and the result is proven.

¹A spanning forest of a directed graph $\mathcal{G}$ is a subgraph which has the same set of nodes as $\mathcal{G}$ and has no cycles. A spanning forest of a directed graph is said to be rooted at a set of nodes if all nodes outside of this set have out-degree one.

The step-wise process has $M + 1$ petal-like graphs each connected to the central state, but disconnected from each other. Due to this structure, any sums over spanning trees of the full process will factorise into a product of sums over spanning trees of the petals (see [103] for a similar factorisation). Thus, the absorbing probabilities take the following form:

$$\begin{aligned}\mathbb{P}[\&xy \rightarrow \&xyz] &= \frac{1}{\mathcal{N}} \Lambda^+(z, y) \left[\prod_{z' \neq z} Q(z', y) \right] Q(y, x), \\ \mathbb{P}[\&xy \rightarrow \&x] &= \frac{1}{\mathcal{N}} \Lambda^-(y, x) \prod_z Q(z, y),\end{aligned}\tag{4.3}$$

Here $z \in \{1, \dots, M\}$, $\mathcal{N}$ is the normalisation factor from π^σ in eqn. 2.14. $\Lambda^+(z, y)$ is the sum over spanning trees of the petal connecting states $\&xy$ and $\&xyz$, rooted at the forward completed state, $\&xyz$ and similarly, $\Lambda^-(y, x)$ is the sum over spanning trees of the petal connecting states $\&x$ and $\&xy$, rooted at the backwards completed state, $\&x$ (figure 4.1(b)). Finally, $Q(y, x)$ is the sum over spanning trees of the petal connecting states $\&x$ and $\&xy$, where edges to state $\&xy$ have been linked back to the state $\&x$ and the spanning trees are rooted at that state $\&x$, as in figure 4.1(c). Since Q is a sum over spanning trees rooted at the node to which edges have been redirected, the sum takes the same form for both the forwards and backwards petals, only depending on which two completed states it is connecting. Thus, $\Lambda^+(z, y) \left[\prod_{z' \neq z} Q(z', y) \right] Q(y, x)$ is the sum over spanning trees of the closed stepwise process, with edges to $\&xyz$ unfurled, rooted at node $\&xyz$; and $\Lambda^-(y, x) \prod_z Q(z, y)$ is the sum over spanning trees of the closed stepwise process, with edges to $\&x$ unfurled, rooted at node $\&x$.

We wish to find a set of propensities $\omega_{\pm yx}$ for the transitions $\&x \xrightarrow{\omega_{+yx}} \&xy$ and $\&xy \xrightarrow{\omega_{-yx}} \&x$ which provide the same transition probabilities. Thus, we require:

$$\begin{aligned}\frac{\mathbb{P}[\&xy \rightarrow \&xyz]}{\mathbb{P}[\&xy \rightarrow \&x] + \sum_{z'} \mathbb{P}[\&xy \rightarrow \&xyz']} &= \frac{\omega_{+zy}}{\omega_{-yx} + \sum_{z'} \omega_{+z'y}} \\ \frac{\mathbb{P}[\&xy \rightarrow \&x]}{\mathbb{P}[\&xy \rightarrow \&x] + \sum_{z'} \mathbb{P}[\&xy \rightarrow \&xyz']} &= \frac{\omega_{-yx}}{\omega_{-yx} + \sum_{z'} \omega_{+z'y}}\end{aligned}\tag{4.4}$$

for $z \in \{1, \dots, M\}$. By dividing the numerator and denominator of the left hand side of these equations by $\frac{1}{N} [\prod_{z'} Q(z', y)] Q(y, x)$, it can be seen that choosing propensities such that

$$\omega_{\pm yx} = \frac{\Lambda^{\pm}(y, x)}{Q(y, x)} \quad (4.5)$$

not only preserves the ratios of probabilities of transitions to completed states, but also ensures that $\omega_{\pm yx}$ only depends on monomers x and y . I note here that this coarse-graining process is different from lumping [74, 130], in which the state space is reduced while attempting to retain trajectory dynamics. In the approach of this chapter, the coarse-grained process does not reproduce the dynamics of the fine-grained process, only the statistics of the completed states that are visited, meaning that the sequence of states visited in the trajectories for the coarse-grained process and the sequence of completed states visited in the fine-grained process (ignoring the other fine-grained states) have the same distribution. However, dynamic quantities may be extracted exactly from the step-wise process. That the trajectories have the same distribution follows from that the transition probabilities for the coarse-grained process are chosen to be the same as the probabilities of reaching the next completed state in the fine grained process.

Eqn. 4.5 is a key result of this thesis. It describes a coarse-graining procedure that replaces a subgraph connecting two states by a single edge. The coarse-graining here preserves the steady state properties of the graph outside of this subgraph. Further, this coarse-graining can be applied outside of just the framework presented in this chapter.

4.3 Extracting properties of the fine-grained model from the solution of the coarse-grained model

With these propensities $\omega_{\pm yx}$, the method from Gaspard and Andrieux (section 3.5) may be used to solve for the statistics of the polymers. In particular one needs to solve for the probability (or frequency) of a given monomer x in the bulk of the copolymer $\varepsilon(x)$

as well as the tip probabilities μ . The easiest quantities to extract are the frequencies of monomers in the bulk of the copolymer. These quantities are identical in the coarse-grained and fine-grained models, since the coarse-graining preserves the statistical distribution of sequences produced. Therefore $\varepsilon(x)$ as defined in eqn. 3.14 and $\varepsilon(y|x)$, obtained from the coarse-grained process, apply directly to the fine-grained process.

The tip probabilities μ (eqn. 3.13) give the fraction of time spent in each tip state in the coarse-grained model. However, the coarse-grained model will not reproduce the time series of the fine-grained model, only the sequences of completed states visited. The coarse-grained process is still a CTMC. However, the time dependence of this coarse-grained CTMC is essentially meaningless. The process that is important is the embedded discrete time process (see *e.g.* [131]), which can be found by the standard procedure of quotienting out the expected lifetime of the coarse-grained tip state defined by $\&xy$, which I denote $\tau(x, y)$, giving the frequency with which the tip states are given in the coarse-grained model as

$$\xi(x, y) = \frac{1}{\sum_{x', y'=1}^M \frac{\mu(x', y')}{\tau(x', y')}} \frac{\mu(x, y)}{\tau(x, y)}, \quad (4.6)$$

$$\tau(x, y) = \frac{1}{\omega_{-yx} + \sum_{z=1}^M \omega_{+zy}}. \quad (4.7)$$

This frequency defines a new tip distribution ξ . $\xi(x, y)$ is the frequency that a given pair of monomers x, y is observed at the tip of the growing copolymer in the sequence of transitions. This distribution $\xi(x, y)$ applies to both the coarse-grained model and the sequence of completed states visited in the full, fine-grained model. It can therefore be used to find averages of key dynamic properties in the fine-grained process.

For example, the probability, P , that a growing copolymer increases in length at each step of the step-wise process is an important quantity that can be calculated. P is calculated by averaging the probability of adding a monomer over the possible states

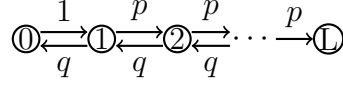


Figure 4.2: Graphical representation of the random walk process considered.

&xy:

$$P = \sum_{x,y=1}^M \xi(x,y) \frac{\sum_{z=1}^M \omega_{+zy}}{\omega_{-yx} + \sum_{z=1}^M \omega_{+zy}}. \quad (4.8)$$

Upon averaging out the sequence information the growth of a polymer can be treated as a random walk with probability P of stepping forwards and $1 - P$ of stepping back. The expected number of monomer inclusion/removal steps per net forward step can then be found.

Set up a random walk as follows. Let the state space be the nodes $\{0, 1, \dots, L\}$, where L is the length of the walk (polymer). Let the transition $0 \rightarrow 1$ have probability 1, $i \rightarrow i + 1$ for $i = 1, \dots, L - 1$ have probability p , $i \rightarrow i - 1$ for $i = 1, \dots, L - 1$ have probability $q = 1 - p$ and let state L be an absorbing state as in figure 4.2.

One then wishes to find the expected number of steps to absorption, given starting in state 0, for which the spanning tree methods with eqn. 2.19 can be utilised. Since the total rate out of any state sums to one, the expected number of steps equals the expected time to absorption. Thus, form the closed process starting at 0. Let $f(n)$ be the sum over spanning trees rooted at node n for this closed process. $f(n)$ is given by:

$$f(n) = \begin{cases} \sum_{i=0}^{L-1} p^i q^{L-1-i} & \text{for } n = 0 \\ p^{n-1} \sum_{i=0}^{L-1-n} p^i q^{L-1-n-i} & \text{for } n = 1, \dots, L - 1. \end{cases} \quad (4.9)$$

From this expression, the expected number of steps before absorption is:

$$\mathbb{E}[\text{steps}] = \frac{\sum_{n=0}^{L-1} f(n)}{pf(L-1)}. \quad (4.10)$$

By utilising the formulae for finite geometric series, the expected number of steps can be found to be:

$$\mathbb{E}[\text{steps}] = \frac{1}{2p-1} \left(L - 1 - \frac{q}{p^L} \left(\frac{p^L - q^L}{p - q} \right) + \frac{q^L}{p^L} + \frac{p^L - q^L}{p^{L-1}} \right). \quad (4.11)$$

Most of this expression is sub-linear in L , and as such:

$$\lim_{L \rightarrow \infty} \frac{\mathbb{E}[\text{steps}]}{L} = \frac{1}{2p-1}, \quad (4.12)$$

which is the expected total number of steps per net forward step.

A number of quantities scale with the total number of steps rather than the net number of steps, making the number of steps per net forward step a necessary quantity. For example, in order to find the expected time taken per net forward step, one can find the expected time to absorption for the step-wise process, figure 4.1(a), $T(x, y)$, for a copolymer in state $\&xy$. This may be calculated as $1/J_{\text{Tot}}(\&xy)$ for the step-wise process using eqn. 2.18. The expected time per net forward step is then

$$\tau_{\text{step}} = \frac{1}{2P-1} \sum_{x,y=1}^M \xi(x, y) T(x, y) = \frac{1}{2P-1} \sum_{x,y=1}^M \frac{\xi(x, y)}{J_{\text{Tot}}(\&xy)}. \quad (4.13)$$

$1/\tau_{\text{step}}$ is the physical average growth rate of the copolymer in the fine-grained model.

The chemical work done by the system in producing the copolymer is another key quantity. In a purely chemical system, with no time-varying externally applied protocols, the entropy increase of the universe is given by the decrease in the generalised free energy of the chemical system, including any coupled reservoirs of fuel molecules [77], section 2.4. Since the total free energy must decrease, any increase in one contribution must be paid

for by a decrease of at least the same magnitude in another contribution.

As in eqn. 2.51, the generalised free energy can be split into a term corresponding to the chemical free energy of the system, averaged over the uncertain state of the system, and a term related to the entropy arising due to the uncertainty of the state occupied [132].

For the systems studied here, in which there are infinitely long copolymers that have reached steady state growth, the only relevant contribution to the entropy term is the increase in Shannon entropy of the copolymer sequence produced as the polymer gets longer. This entropy term can be calculated as in section 3.5, using the statistics of the completed states.

Since, as in section 3.5, the purpose of a copolymerisation system is often to produce a low entropy (or “accurate”) sequence, it is reasonable to think of the chemical free-energy decrease per net forward step as the chemical work done to reduce the information entropy of eqn. 3.16 below that of a uniform, random polymer. Extending the definition provided by Poulton *et al.* [113], (eqn. 3.17) the efficiency of copolymerisation may be defined to be:

$$\eta = \frac{\ln M - H}{\ln M + \mathcal{W}_{\text{chem}}} \in [0, 1], \quad (4.14)$$

where $\ln M$ is the entropy per monomer (or entropy rate) of a uniform, random copolymer with M monomer types, and $\mathcal{W}_{\text{chem}}$ is the average decrease in chemical free energy per net forward step. Similarly to the simple copolymerisation efficiency (eqn. 3.17), this efficiency is the ratio between the entropy drop due to the accuracy of the copolymer compared to a random one ($\ln M - H$) and the chemical work used to drive the system ($\mathcal{W}_{\text{chem}}$) above that required to make a random copolymer in equilibrium ($-\ln M$) [133]. Another way to view the numerator of the efficiency is through eqn. 2.6. If one makes an assumption that there exists a template for each possible sequence of monomers and that those templates act symmetrically on the set of possible sequences, then the numerator represents the mutual information between output polymer and template per length of polymer.

The expected work done during a transition adding or removing a monomer can be found from the step-wise process (figure 4.1), given starting in completed state $\&xy$. The expected work done before absorption can be calculated as follows, making use of the free-energy term of eqn. 2.57,

$$w_{\text{chem}}(x, y) = -\Delta G_{\text{chem}}(x, y) = \sum_{b>a} \ln \left(\frac{W(a, b)}{W(b, a)} \right) N_{a\rightleftharpoons b}(\&xy), \quad (4.15)$$

where $W(a, b)$ is the rate of the reaction $a \rightarrow b$ and $N_{a\rightleftharpoons b}(\&xy)$ is the expected net number of times traversing edge $a \rightleftharpoons b$ before absorption giving starting in the central state of the step-wise process, $\&xy$, as in eqn. 2.25. This sum will also require contributions from edges which lead to absorbing states. For such edges, the rate for the reverse transition in the logarithm of eqn. 4.15 is the rate from the full process.

Equivalently, however, as outlined in section 2.2.9, this chemical work may be found by considering the non-recurrent cycles of the process [88]. For a given internal cycle, C , recall the affinity [77],

$$A(C) = \ln \frac{\prod_{e \in C} W(e)}{\prod_{e \in C'} W(e)}, \quad (4.16)$$

where the sum is over the edges, e , composing the cycle and C' is the cycle with edges in the reversed direction. For external cycles, the affinity may be defined in the same way, inferring the rate for the reversed edge of the transition to absorbing states from the full process. The expected work done before absorption (in term of the cycles), given starting in the state $\&xy$ is

$$w_{\text{chem}}(x, y) = -\Delta G_{\text{chem}}(x, y) = \sum_C A(C) \frac{J_{\text{Cyc}}(\&xy, C) - J_{\text{Cyc}}(\&xy, C')}{J_{\text{Tot}}(\&xy)}, \quad (4.17)$$

where J_{Cyc} is the cycle current (eqn. 2.27) and the sum runs over all the cycles of the (closed) step-wise process. Averaging $w_{\text{chem}}(x, y)$ with ξ and multiplying by the expected number of steps per net forward step gives the expected chemical work done per net

forward step,

$$\mathcal{W}_{\text{chem}} = \frac{1}{2P-1} \sum_{x,y=1}^M \xi(x,y) w_{\text{chem}}(x,y). \quad (4.18)$$

Further, the forms of eqns. 4.13 and 4.18 may be applied to an arbitrary quantity for which one can find the expected value in the step-wise process starting in state $\&xy$ (*e.g.* $A(x,y)$). One can then average this arbitrary quantity using the distribution, ξ , to obtain the expected value of the quantity per step. Then, if appropriate, multiplying by $1/(2P-1)$, gives the expected value of the quantity per net forward step. In practice, as shall be seen in section 4.4.4, since the quantities one wishes to calculate may be written in terms of sums over spanning trees, the quantities for the step-wise process may be written as a sum over the terms per petal, with the quantity for a given petal factorising into some quantity, which depends only on that petal, multiplied by Q 's for the other petals.

4.3.1 Stalled growth

Explicit simulation of copolymer growth is particularly challenging in regimes where $P \approx 0.5$, since many backward and forwards steps are taken per net forwards step. At $P = 0.5$, then the process will not reliably produce copolymers; for $P < 0.5$ polymers will tend to shrink. In general, for $P = 0.5$, the model can be described as having stalled. The approach presented in this thesis is particularly beneficial in this case; indeed, it is possible to check whether a model is at the stall point by considering an $M \times M$ dimensional matrix of the ratios of forward to backwards propensities [97], $Z_{yx} = \left(\frac{\omega_{+yx}}{\omega_{-yx}} \right) = \frac{\Lambda^+(y,x)}{\Lambda^-(y,x)}$. The model is at the stall point if and only if [97]:

$$\det(Z - \mathbb{1}_M) = 0, \quad (4.19)$$

where $\mathbb{1}_M$ is the $M \times M$ identity matrix, and shrinking if this determinant is negative. Heuristically, since Z gives the ratios of adding a monomer to removing one, this condition essentially says that models will stall if the total rate of adding a monomer is equal to

the total rate of removing one.

In a typical model, there exists at least one parameter that controls the driving. Often this parameter is related to the backbone strength of the polymer produced: *e.g.* the free energy drop associated with the formation of a generic backbone bond, ΔG_{pol} . This parameter will be present in the rates of each external cycle so that, by tuning it, the model can be moved all the way from stalling to irreversible driving, whereby monomers cannot be removed once polymerised. If such a parameter exists, the stall condition, eqn. 4.19, may be rephrased in terms of this parameter. For example, for the case of the parameter being ΔG_{pol} , one may find some threshold value Γ such that the model will stall for $\Delta G_{\text{pol}} = \Gamma$.

4.3.2 Limiting behaviour

I shall note two limits for which analytic expressions for the frequency of monomer types in the copolymer bulk for all models may be given.

First, consider the case that the system is at the stall point (eqn. 4.19). In general, entropy production can still occur within cycles in the step-wise process; therefore, these frequencies cannot be determined from equilibrium arguments and are non-trivial. Nonetheless, at the stall point, the monomer frequencies in the bulk may be expressed relatively simply as the following. The frequency of monomer x , $\varepsilon_{\text{stall}}(x)$, is proportional (up to normalisation) to the cofactor of the diagonal element (corresponding to monomer x) of the matrix $(\mathbb{1}_M - Z)$. More explicitly, at stall,

$$\varepsilon_{\text{stall}}(x) \propto A_{xx}, \quad (4.20)$$

where A_{ij} is the cofactor of element i, j of the matrix $\mathbb{1} - Z$.

For example, for $M = 2$,

$$\begin{aligned} \varepsilon_{\text{stall}}(1) &\propto 1 - \frac{\omega_{+22}}{\omega_{-22}}, \\ \varepsilon_{\text{stall}}(2) &\propto 1 - \frac{\omega_{+11}}{\omega_{-11}}, \end{aligned} \quad (4.21)$$

and for $M = 3$, we have

$$\begin{aligned}
\varepsilon_{\text{stall}}(1) &\propto \left(1 - \frac{\omega_{+22}}{\omega_{-22}}\right) \left(1 - \frac{\omega_{+33}}{\omega_{-33}}\right) - \frac{\omega_{+23}}{\omega_{-23}} \frac{\omega_{+32}}{\omega_{-32}}, \\
\varepsilon_{\text{stall}}(2) &\propto \left(1 - \frac{\omega_{+11}}{\omega_{-11}}\right) \left(1 - \frac{\omega_{+33}}{\omega_{-33}}\right) - \frac{\omega_{+13}}{\omega_{-13}} \frac{\omega_{+31}}{\omega_{-31}}, \\
\varepsilon_{\text{stall}}(3) &\propto \left(1 - \frac{\omega_{+11}}{\omega_{-11}}\right) \left(1 - \frac{\omega_{+22}}{\omega_{-22}}\right) - \frac{\omega_{+12}}{\omega_{-12}} \frac{\omega_{+21}}{\omega_{-21}}.
\end{aligned} \tag{4.22}$$

To show this relation (eqn. 4.20), I shall rely on the relationship between cofactors and vectors of the nullspace of a matrix. Let B be an arbitrary matrix with a one dimensional nullspace, and let A be its matrix of cofactors. Recall that

$$BA^T = \det(B)\mathbb{1} = 0. \tag{4.23}$$

Thus, any column of A^T is in the nullspace of B . In anticipation, let $\boldsymbol{\mu}$ be a vector in the nullspace of B and $\mathbf{v}$ be a vector in the nullspace of B^T . Since B has a one dimensional nullspace, then

$$\frac{\mu_x}{\mu_y} = \frac{A_{ix}}{A_{iy}}, \tag{4.24}$$

for some arbitrary i . Similarly,

$$\frac{v_x}{v_y} = \frac{A_{xj}}{A_{yj}}, \tag{4.25}$$

for arbitrary j .

Looking at eqns. 3.12, and 3.13, noting that near the stall point, $v_z \ll \omega_{\pm y, x}$, one observes that, the tip probabilities, $\mu(x)$, form a vector in the nullspace of $\mathbb{1}_M - Z$ and the tip velocities, v_x , form a vector in the nullspace of $\mathbb{1}_M - Z^T$. Hence,

$$\frac{\mu(y)v_y}{\mu(x)v_x} = \frac{A_{jy}A_{yi}}{A_{jx}A_{xi}}, \tag{4.26}$$

for arbitrary i, j . Thus, choosing $j = y$ and $i = x$ leads to cancellation such that

$$\frac{\mu(x)v_x}{\mu(y)v_y} = \frac{A_{xx}}{A_{yy}}. \quad (4.27)$$

Since

$$\varepsilon(x) = \frac{\mu(x)v_x}{\sum_y \mu(y)v_y} = \frac{A_{xx}}{\sum_y A_{yy}}, \quad (4.28)$$

the required result, eqn. 4.20, is obtained.

On the other end of the spectrum, monomer bulk frequencies in the irreversible limit, where $\omega_{-yx} = 0$ for all x, y , can also be solved for analytically. Intuitively, one could consider the Markov process on the state space $\{1, \dots, M\}$ representing copolymers with a given monomer at its tip, and transitions between those states with rates, $W_{\text{irrev}}(x \rightarrow y) = \omega_{+yx}$. The steady state of this process will give the time dependent frequencies of having a given monomer at the tip of the copolymer. Therefore, dividing by the time spent in each state will give the bulk frequencies. A nice way to write out these frequencies in the style of the methods described thus far is as a sum over the spanning trees on the complete graph on M vertices with rate functions $W_{\text{irrev}}(x, y) = \omega_{+yx}$. Explicitly, we may write these frequencies (up to normalisation) as:

$$\varepsilon_{\text{irrev}}(x) \propto \left(\sum_{T \in \mathcal{T}(x)} \prod_{e \in T} W_{\text{irrev}}(e) \right) \sum_{y=1}^M \omega_{+yx}, \quad (4.29)$$

where $\mathcal{T}(x)$ is the set of spanning trees of the complete graph on M vertices.

For example, with $M = 2$,

$$\begin{aligned} \varepsilon_{\text{irrev}}(1) &= \frac{\omega_{+12}(\omega_{+11} + \omega_{+21})}{\omega_{+12}(\omega_{+11} + \omega_{+21}) + \omega_{+21}(\omega_{+12} + \omega_{+22})}, \\ \varepsilon_{\text{irrev}}(2) &= \frac{\omega_{+21}(\omega_{+12} + \omega_{+22})}{\omega_{+12}(\omega_{+11} + \omega_{+21}) + \omega_{+21}(\omega_{+12} + \omega_{+22})}. \end{aligned} \quad (4.30)$$

I shall derive this more formally now. We wish to find an expression for the frequency with which monomer x appears in the bulk of the copolymer in the irreversible limit. This limit is such that the backwards propensities are zero, $\omega_{-yx} = 0$. With this assumption, from eqn. 3.12, the partial velocities become

$$v_x = \sum_y \omega_{+yx}. \quad (4.31)$$

With this form for the velocities, eqn. 3.13 may be manipulated to

$$\begin{aligned} \mu(x) &= \sum_y \frac{\omega_{+yx}}{\sum_z \omega_{+zx}} \mu(y), \\ \sum_{z \neq x} \omega_{+zx} \mu(x) + \omega_{+xx} \mu(x) &= \sum_{y \neq x} \omega_{+xy} \mu(y) + \omega_{+xx} \mu(x). \end{aligned} \quad (4.32)$$

This last line is the equation for steady state of a Markov process with probability $\mu(x)$ to be in state x and rate ω_{+yx} of transition from state x to state y . Thus, set $\mu(x)$ to be the steady state probability distribution of the Markov process on M states with transition rates from state x to y given by ω_{+yx} , and $v_x = \sum_y \omega_{+yx}$. Then, calculating

$$\varepsilon(x) \propto \mu(x) v_x, \quad (4.33)$$

gives the required result. Finding the distribution, $\mu(x)$, in terms of spanning trees of the complete graph on M elements gives eqn. 4.29.

4.3.3 Simplification for factorisable propensities

The presented method applies to arbitrary complex copolymerisation models obeying the structure of figure 4.1. However, if one makes some further common assumptions, much of the analysis simplifies. For example, consider the case in which the ratios of propensities

may be factored:

$$\frac{\omega_{+yx}}{\omega_{-yx}} = \frac{\Lambda^+(y, x)}{\Lambda^-(y, x)} = Y(y)X(x), \quad (4.34)$$

where Y is a function of monomer y only and X is a function of monomer x only. Intuitively, such a condition holds in the cases where there is no direct, monomer-type-dependent interactions between monomers in the growing polymer, such as when monomers only interact with a template [92, 99–108, 113, 114, 123, 134]. Under such an assumption, multiple calculations simplify.

Two such simplifications are to do with stall. Firstly, the stall condition for such models may be rewritten:

$$\sum_x X(x)Y(x) = 1. \quad (4.35)$$

Secondly, the bulk frequency at stall is just:

$$\varepsilon_{\text{stall}}(x) = X(x)Y(x). \quad (4.36)$$

These two results can be shown as follows. Consider the ratio of propensities factorising as in eqn. 4.34. Thinking of the functions X and Y as column vectors, since they each have a discrete domain, the matrix Z may be written,

$$Z = \mathbf{X}\mathbf{Y}^T. \quad (4.37)$$

By a well known result [135],

$$\det(\mathbb{1}_M - \mathbf{X}\mathbf{Y}^T) = 1 - \mathbf{Y}^T\mathbf{X} = 1 - \sum_x X(x)Y(x). \quad (4.38)$$

At stall, this determinant is zero. Thus, rearranging gives the stall condition eqn.4.35.

The model is at stall if this equation is satisfied. As shown, the frequency of monomer

x in the bulk of the copolymer is given by the cofactor of the diagonal elements of $\mathbb{1}_M - \mathbf{X}\mathbf{Y}^T$. The cofactor, A_{xx} , may be written:

$$A_{xx} = \det(\mathbb{1}_{M-1} - \mathbf{X}_{[x]}\mathbf{Y}_{[x]}^T) = 1 - \sum_{y \neq x} X(y)Y(y) = X(x)Y(x), \quad (4.39)$$

using the stall condition, and where $\mathbf{X}_{[x]}$ is the vector $\mathbf{X}$, missing element $X(x)$, *i.e.* $\mathbf{X}_{[x]} = (X(1), \dots, X(x-1), X(x+1), \dots, X(M))^T$. Eqn. 4.39 is the required result. Additionally, because of the stall condition, $\varepsilon(x) = X(x)Y(x)$ is already a normalised probability distribution.

4.4 Example applications

I shall now consider some exemplar classes of models to: provide examples of how to utilise the methods introduced in section 4.2; validate their accuracy; and to show the types of quantities and information that may be extracted.

A useful initial classification of models is into those which I shall call balanced. I shall refer to a model as being balanced if its petals (see figure 4.1(b)) are detailed balanced. Such models are useful baseline checks as their cycles all have zero affinity, meaning no chemical work is done internally and hence the only contributions to chemical work are from external cycles. Further, these models exhibit a proper equilibrium at the stall point, and as such allow for equilibrium arguments to validate the method at this point. It is worth noting that although related to the notion of detailed balance, the full model with its infinite state space is not detailed balanced.

4.4.1 Stalling behaviour in a polymerisation model with no neighbour-neighbour interactions

I shall start with the simplest case, where the propensities in the coarse-grained model only depend on the monomer type being added/removed: $\omega_{\pm yx} = \omega_{\pm y}$, such as in a simple

model for templated self assembly, figure 3.1(b). Assume there exists a backbone free energy, ΔG_{pol} controlling the driving as in section 4.3.1. Any spanning tree in $\Lambda^\pm$ must involve at least one incidence of ΔG_{pol} , since it appears in every external cycle. Therefore, the ratio of propensities can be split as follows:

$$\frac{\omega_{+y}}{\omega_{-y}} = e^{\Delta G_y} e^{\Delta G_{\text{pol}}}, \quad (4.40)$$

where ΔG_y encompasses the rest of the details about the models. We note in general, ΔG_y may be a function of ΔG_{pol} , however in many cases, it is not. These cases include when there is only one completion reaction (highlighted in red in figure 4.1(a)) that contains the dependence on ΔG_{pol} or if the model is balanced. $-\Delta G_y$ may then be interpreted as an effective binding free energy of monomer y . If one thinks of ΔG_{pol} as the free energy drive of the model away from stall, one is looking for a threshold value $\Delta G_{\text{pol}} = \Gamma$, above which the model will not stall. Using eqn. 4.35, it can be seen that

$$\Gamma = -\ln \left(\sum_y e^{\Delta G_y} \right) = -\ln \mathcal{Z}, \quad (4.41)$$

where $\mathcal{Z}$ is the partition function for a system with one state for each monomer type, each state labelled by y and with free energy $-\Delta G_y$. Furthermore, using eqn. 4.36, the bulk frequencies at the stall point may be written:

$$\varepsilon_{\text{stall}}(y) = \frac{e^{\Delta G_y}}{\sum_x e^{\Delta G_x}} = \frac{1}{\mathcal{Z}} e^{\Delta G_y}, \quad (4.42)$$

which is the probability of selecting a state y , with free energy, $-\Delta G_y$ as predicted by equilibrium statistical mechanics. In these results, $-\Delta G_y$ looks like the equilibrium contribution to free energy, and the results follow fairly directly in equilibrium. However, these results hold even if the process involves fuel-consuming cycles: entropy may still be being produced at stall. In such cases, the effect of breaking equilibrium will be to change the effective free energies of selecting a given monomer type.

4.4.2 Reproducing the trivial internal reaction network polymerisation results

Let us now reproduce the coarse-grained results of Poulton *et al.* from a microscopic model. Consider a model with two monomer types, right r and wrong w (as in section 3.3) and consider the internal reaction network for the addition of monomer y to a polymer in state $\&x$ as in figure 4.3.

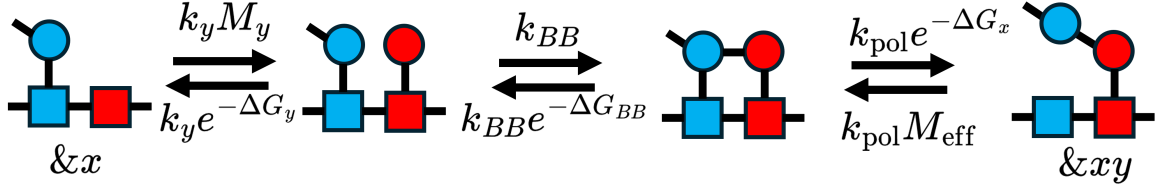


Figure 4.3: Reaction rates of the simple model to reproduce the trivial internal reaction network of section 4.4.2. In this model a monomer y binds to a free site on the template, with rate $k_y M_y$ (or unbinds with rate $k_y e^{-\Delta G_y}$). This monomer forms a backbone with the previous monomer in the polymer x with rate k_{BB} (or breaks this backbone bond with rate $k_{BB} e^{-\Delta G_{BB}}$). Finally, the previous monomer in the polymer detaches from the template with rate $k_{pol} e^{-\Delta G_x}$ (or reattaches to the template with rate $k_{pol} M_{eff}$)

Assume that monomer y is chemostatted at M_y in solution. k_y is the binding rate constant for monomer y to a free slot on the template. Further, assume this binding rate constant is such that $k_r = k e^{\Delta G_K}$ and $k_w = k$. ΔG_y is the binding free energy of monomer y . Additionally, assume that the binding/unbinding rates are such that the rate of breaking a bond are determined by the binding free energy. The leftmost reaction is thus the binding/unbinding of monomer y from solution. k_{BB} is the rate constant for the formation/breaking of the polymer backbone and ΔG_{BB} the binding energy of the backbone, both of which are assumed to be independent of monomer type. Thus the middle reaction is the formation of the polymer backbone. Finally, M_{eff} is an effective concentration. This last reaction involves the unbinding/binding of the penultimate monomer in the polymer. The forwards reaction is the unbinding of penultimate monomer x . This penultimate monomer is held in close proximity to the template binding site. Thus, the rebinding of the penultimate monomer is the same as the binding of a monomer with an

enhanced concentration (M_{eff}).

With these reactions, the method (section 4.2) can be used by calculating the sums over spanning trees:

$$\begin{aligned}\Lambda^+(y, x) &= k_y k_{BB} k_{\text{pol}} M_y e^{-\Delta G_x}, \\ \Lambda^-(y, x) &= k_y k_{BB} k_{\text{pol}} M_{\text{eff}} e^{-\Delta G_{BB}} e^{-\Delta G_y}, \\ Q(y, x) &= k_y k_{BB} e^{-\Delta G_y} e^{-\Delta G_{BB}} + k_{\text{pol}} e^{-\Delta G_x} (k_{BB} + k_y e^{-\Delta G_{-y}}).\end{aligned}\quad (4.43)$$

If one assumes that the initial binding/unbinding of monomer y is much slower than the other reactions, $k_y \ll k_{BB}, k_{\text{pol}}$, then the propensities (eqn. 4.5) reduce to:

$$\begin{aligned}\omega_{+yx} &\approx k_y M_y \\ \omega_{-yx} &\approx k_y M_{\text{eff}} e^{-\Delta G_{BB}} e^{\Delta G_x - \Delta G_y}.\end{aligned}\quad (4.44)$$

These propensities reproduce the results of Poulton *et al.* (eqn. 3.18), where the concentration of monomers in solution are set to unity ($M_y = 1$), and the polymerisation free energy is identified to be $\Delta G_{\text{pol}} = \Delta G_{BB} - \ln M_{\text{eff}}$.

One notices that this model is factorisable with

$$\frac{\omega_{+yx}}{\omega_{-yx}} = e^{\Delta G_{\text{pol}}} e^{\Delta G_y} e^{-\Delta G_x}, \quad (4.45)$$

such that the stall condition is $2e^{-\Delta G_{\text{pol}}} = 1$ so $\Delta G_{\text{pol}} = -\ln 2$ in agreement with [113]. Further, the frequency of monomer x in the bulk is just 0.5 at stall as one would expect.

One trade-off that can be easily discussed with this framework is a limit observed by Poulton *et al.* whereby, if the difference between right and wrong monomers diverges (either by $\Delta G_K \rightarrow \infty$ or $\Delta G_{TT} = \Delta G_r - \Delta G_w \rightarrow \infty$), then the efficiency at $\Delta G_{\text{pol}} = 0$ tends to unity.

The trade-off made here is only between speed and energy, as the accuracy will be approximately perfect. When one calculates the forwards probability (eqn. 4.8) in this limit, one can see $P = 0.5$ in this limit for all $\Delta G_{\text{pol}} \in [-\ln 2, 0]$.

For all ΔG_{pol} where $P = 0.5$, the number of steps taken per net addition of monomer diverges. However, the time to absorption $T(x, y)$ remains finite. So, the time per net step (eqn. 4.13) diverges. For $\Delta G_{\text{pol}} > 0$, the random walks described in the previous paragraph become biased to step forwards ($P > 0.5$). As such the number of steps taken per net forwards step becomes finite, allowing for eqn. 4.13 also to be finite. The polymer will consist of only r monomers in the bulk ($\varepsilon(r) = 1$). It is in this sense that a trade-off is being made only between the energy cost (ΔG_{pol}) and speed.

For all $\Delta G_{\text{pol}} \in [-\ln 2, 0)$ in this limit, the time taken per net step is divergent. Further, accuracy $\ln 2 - H$ decreases to zero as $\Delta G_{\text{pol}} \rightarrow -\ln 2$. The trade off being made in this regime is between accuracy and free energy only.

4.4.3 Balanced models of templated polymerisation with autonomous separation

Next I shall discuss a class of models where the ratio of propensities may be written:

$$\frac{\omega_{+yx}}{\omega_{-yx}} = e^{\Delta G_y} e^{-\Delta G_x} e^{\Delta G_{\text{pol}}}. \quad (4.46)$$

As before, ΔG_{pol} , coming from the polymerisation reactions represents the driving of this process. Such a class of models includes, most notably, balanced models of templated polymerisation with autonomous separation, for example the simple model from Poulton *et al.* just discussed [113]. In these cases, the breaking of the previous copy-template bond every time a new bond is formed enforces the structure in eqn. 4.46. I shall assume, as in ref. [113], that ΔG_y is independent of ΔG_{pol} .

Using eqn. 4.35 and eqn. 4.36, the stall point can be found to be $\Delta G_{\text{pol}} = \Gamma = -\ln M$ and bulk frequencies at stall, $\varepsilon_{\text{stall}}(y) = \frac{1}{M}$, where M the number of monomer types. Physically, these results can be understood by considering balanced models of templated polymerisation with autonomous separation. For such models, by definition there is no entropy production in internal cycles and therefore, the stall point must be thermodynamic equilibrium. In such models, the only driving comes from the polymerisation,

ΔG_{pol} , and the entropic effect having M monomers to choose. These two effect balance at equilibrium [133].

Next consider the limit that the completion reactions, highlighted in red in figure 4.1(a), are much slower than the other reactions. Explicitly, let k be some rate constant at the same order of magnitude of the rates of the process that are not the rates for the completion transitions indicated in red in figure 4.1(a). Write the completion rates as $k_{\text{com}} R_{\text{com}}^+(y, x)$, where $k_{\text{com}} \ll k$ is a rate constant controlling the overall speed of the completion reactions and $R_{\text{com}}^+(y, x)$ provides any sequence dependence. Similarly, the reverse transitions along the completion edges have the rate $k_{\text{com}} R_{\text{com}}^-(y, x)$. Further, let there be n_{com} such completion reactions in a given petal of the step-wise process (assume this number is the same for all pairs of monomers, x, y).

Assume for simplicity that all completion reactions, $R_{\text{com}}^\pm(y, x)$, take the same form in a given petal. Then, the sum over spanning trees $Q(y, x)$ can be written as

$$Q(y, x) = \frac{1}{n_{\text{com}}} \frac{\Lambda^-(y, x)}{k_{\text{com}} R_{\text{com}}^-(y, x)} + \mathcal{O}\left(\frac{k_{\text{com}}}{k}\right), \quad (4.47)$$

since $\Lambda^-(y, x)$ has first order terms in k_{com}/k . This fact can be seen from noting that the leading order terms in $Q(y, x)$ are the trees with no completion reactions and the leading order terms in $\Lambda^-(y, x)$ are those same leading order trees of $Q(y, x)$, except with one completion reaction added in. There are n_{com} such completion reactions and each adds the same leading order term to $\Lambda^-(y, x)$. With $Q(y, x)$ taking this form, and remembering eqn. 4.46, the propensities take the following form:

$$\begin{aligned} \omega_{+yx} &= n_{\text{com}} k_{\text{com}} R_{\text{com}}^-(y, x) e^{\Delta G_y - \Delta G_x + \Delta G_{\text{pol}}} + \mathcal{O}\left(\frac{k_{\text{com}}}{k}\right)^2, \\ \omega_{-yx} &= n_{\text{com}} k_{\text{com}} R_{\text{com}}^-(y, x) + \mathcal{O}\left(\frac{k_{\text{com}}}{k}\right)^2. \end{aligned} \quad (4.48)$$

The $n_{\text{com}} k_{\text{com}}$ term cancels in ratios of $\omega_{\pm yx}$ variables, and therefore does not affect the sequence statistics. Thus, in the slow completion limit, such models are only affected by the binding free energy differences ($\Delta G_y - \Delta G_x$), the driving (ΔG_{pol}), and the nature of

the final completion step (R_{com}^-). Therefore, the fine details do not affect the statistics of the polymers.

Assuming that all completion edges are associated with the same free energy change $-\Delta G_{\text{pol}}$, so that $R_{\text{com}}^-(y, x) = e^{-\Delta G_{\text{pol}}}$, one may solve for the statistics explicitly. Let us consider the propensities propensities given by eqn. 4.48, cancelling $n_{\text{com}}k_{\text{com}}$, and with $M = 2$ (writing monomers as 1, 2). With these propensities, eqn. 3.12 becomes

$$v_1 = \frac{v_1}{e^{-\Delta G_{\text{pol}}} + v_1} + \frac{e^{-DG}v_2}{e^{-\Delta G_{\text{pol}}} + v_2} \quad (4.49)$$

$$v_2 = \frac{e^{DG}v_2}{e^{-\Delta G_{\text{pol}}} + v_1} + \frac{v_2}{e^{-\Delta G_{\text{pol}}} + v_2}, \quad (4.50)$$

where $DG = \Delta G_1 - \Delta G_2$. These equations may be solved by using the following form the velocities:

$$v_x = e^{\Delta G_y - \Delta G_x} v_y. \quad (4.51)$$

Doing so, reduces eqn. 3.12 to a quadratic equation,

$$0 = v_1^2 + (1 + e^{DG})(e^{-\Delta G_{\text{pol}}} - 1)v_1 + e^{-\Delta G_{\text{pol}}}(e^{-\Delta G_{\text{pol}}} - 2) \quad (4.52)$$

with one positive root,

$$v_1 = \frac{1}{2} \left((1 - e^{-\Delta G_{\text{pol}}})(e^{-DG} + 1) + \sqrt{(e^{-\Delta G_{\text{pol}}} - 1)^2(e^{-DG} - 1)^2 + 4e^{-DG}} \right), \quad (4.53)$$

when the system is not stalling. v_2 can be found from in terms of v_1 as $v_2 = e^{-DG}v_1$. Further, a quick check confirms $v_1 = 0$ if $\Delta G_{\text{pol}} = -\ln 2$. Further, with v_y known, eqn. 3.13 is a simple linear equation, μ can be found as the eigenvector of the matrix

$$\begin{pmatrix} \frac{1}{e^{-\Delta G_{\text{pol}}} + v_1} & \frac{e^{-DG}}{e^{-\Delta G_{\text{pol}}} + v_2} \\ \frac{e^{DG}}{e^{-\Delta G_{\text{pol}}} + v_1} & \frac{1}{e^{-\Delta G_{\text{pol}}} + v_2} \end{pmatrix}, \quad (4.54)$$

with eigenvalue 1 and normalised to sum to 1. Combining the solutions for μ and v , using eqn. 3.14, gives:

$$\begin{aligned} \varepsilon(1) = & \left(1 - \frac{1}{2}(e^{-\Delta G_{\text{pol}}} - 1)(e^{-DG} - 1) \right. \\ & \left. + \frac{1}{2}\sqrt{(e^{-\Delta G_{\text{pol}}} - 1)^2(e^{-DG} - 1)^2 + 4e^{-DG}} \right)^{-1}, \end{aligned} \quad (4.55)$$

where $DG = \Delta G_1 - \Delta G_2$, as an analytic solution for the error. This expression is plotted in figure 4.4(a) for $DG = 4$. From this expression, the stall driving, $\Delta G_{\text{pol}} = -\ln 2$, can be substituted into this expression to confirm explicitly that the bulk frequency indeed becomes $\varepsilon(1) = \frac{1}{2}$. Further, taking the irreversible limit, $\Delta G_{\text{pol}} \rightarrow \infty$, the bulk frequency becomes:

$$\varepsilon(1) = \frac{e^{\Delta G_1}}{e^{\Delta G_1} + e^{\Delta G_2}}, \quad (4.56)$$

the equilibrium statistical mechanics probability of choosing state 1 with free energy $-\Delta G_1$, given state 2 has free energy $-\Delta G_2$. Since the completion reactions are slow and irreversible, in this limit, the process selecting the monomers is allowed to equilibrate. Therefore, copolymerisation is simply sampling from the equilibrium distribution of this process, and hence tends to the result predicted by equilibrium statistical mechanics.

Eqn. 4.48 shows that, in the slow completion limit, the fine details of the reaction network leading to selection of a specific monomer become unimportant and the models collapse onto a single accuracy curve determined by DG , ΔG_{pol} and R_{com}^- . It is worth emphasising briefly here that the graphical description I have introduced has allowed this fact to be proven easily.

Conversely, if all parameters except k_{com} are fixed, the bulk frequencies seem to tend monotonically to their limits as $k_{\text{com}}/k \rightarrow 0$, either from above or below. This fact can be used to compare bulk frequencies for certain types of model. For example, one may compare on-rate discrimination [98], where incorrect monomers bind more slowly, to off-rate discrimination [98], where incorrect monomers unbind more quickly. An example model

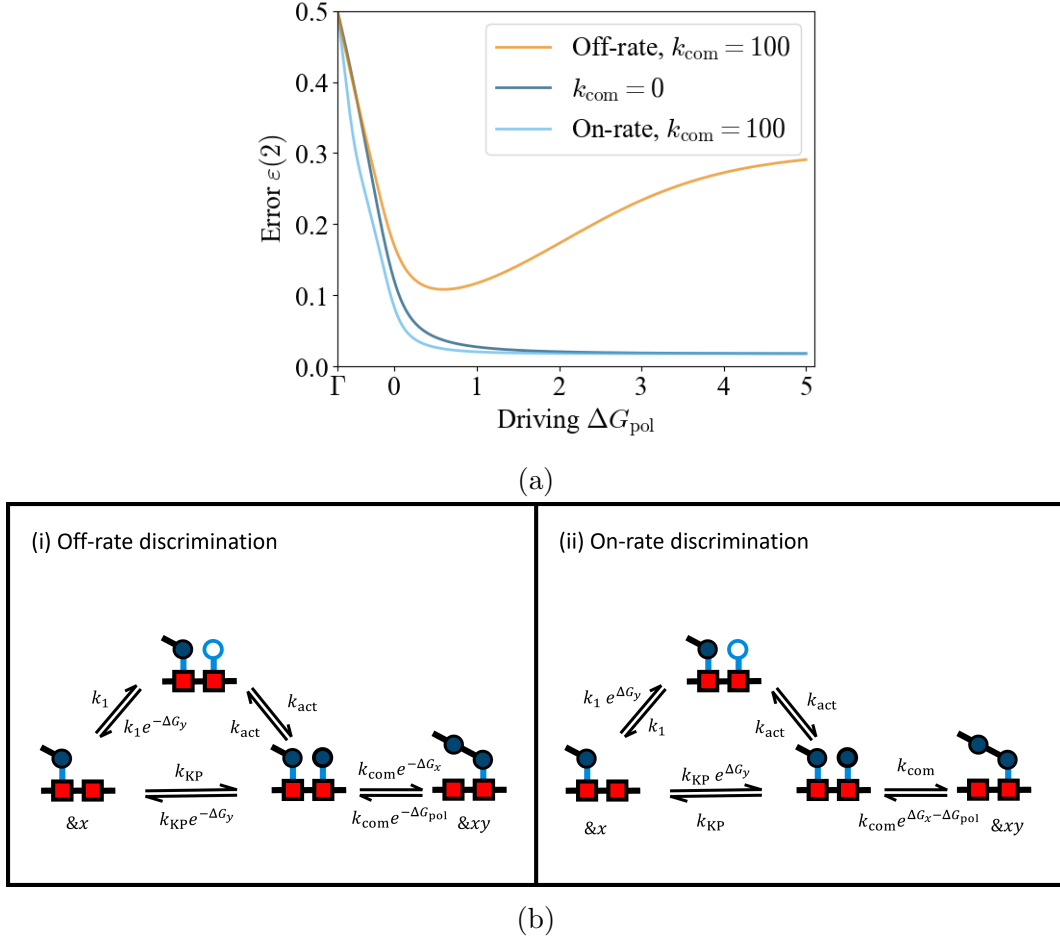


Figure 4.4: (a) Plots of the frequency of the less stably-bound (incorrect) monomer with smallest binding free energy, labelled 2 for on- and off-rate discrimination balanced models with $k_{\text{com}} = 100$ and $k_{\text{com}} \rightarrow 0$. The binding free-energy difference for these models is $DG = \Delta G_1 - \Delta G_2 = 4$. The models are topologically the Hopfield model as in figure 4.5a, with $\Delta G_{\text{act}} = 0$ and $M_{\text{in}} = M_{\text{act}} = 1$, so that they are balanced. However, for the on-rate discrimination, the free-energy terms are in the binding reactions instead of the unbinding ones. The specific models are given in (b). (b) Reaction rates of the (i) off-rate and (ii) on-rate discrimination models used to produce the results of (a). These reactions represent a single petal of the step-wise process (figure 4.1) between completed states $\&x$ and $\&xy$. For the results in (a) for the off-rate and on-rate curves, the following parameters were taken, $\Delta G_1 = 2$, $\Delta G_2 = -2$, $k_1 = k_{\text{KP}} = k_{\text{act}} = 1$, $k_{\text{com}} = 100$.

comparing on-rate and off-rate discrimination is plotted in figure 4.4. Consider the bulk frequency of an incorrect monomer. On-rate discrimination benefits from fast polymerisation and therefore tends to its slow polymerisation limit from below, whereas off-rate discrimination benefits from allowing the process selecting monomers to equilibrate and hence tends to its slow copolymerisation limit from above. This fact sets up a hierarchy for a given set of parameters, and moderate or strong driving. For the bulk frequency of incorrect monomers, off-rate discrimination $>$ slow copolymerisation $>$ on-rate discrimination. On-rate discrimination is, in general, better, although experimentally harder to achieve. This observation is consistent with the results of Sartori and Pigolotti [98] and Poulton *et al.* [113] for kinetic (on-rate) and energetic (off-rate) discrimination.

4.4.4 Hopfield’s kinetic proofreading in a model of templated copying with autonomous separation

For the final example, I shall consider an explicit model of copolymerisation, with Hopfield’s kinetic proofreading mechanism (as in section 3.4) incorporated into a templated copolymerisation system with simultaneously separating product in a thermodynamically valid way. From this setup, I shall provide a fully worked example of an explicit model, as well as demonstrating the power of the method for analysing sequences of models with recursive structures through looking at a generalised version of Hopfield’s proofreading incorporated into a model of templated polymerisation with simultaneous separation.

Explicitly, I first consider the one-loop model of kinetic proofreading shown in figure 4.5(a). There are two monomer types, the right ones r and wrong ones w . Note that the model has already been transformed so that the sequence of the copy is defined relative to that of the template [92]. These monomer types exist in inactive and active states with concentrations M_{in} and M_{act} , respectively, relative to some reference concentration, with each monomer type having the same concentration. As previously, I shall assume the environment is such that these species are chemostatted.

The monomers may bind to the template either in an active or inactive state with

binding free energies $-\Delta G_x$ for monomer type x . Inactive monomers may be activated on the template with a free-energy change of ΔG_{act} . Finally, active monomers may be polymerised into the copolymer chain, with free-energy change $-\Delta G_{\text{pol}}$, and the penultimate monomer of the copolymer unbinds from the template. Each of these reactions is assigned a forwards and reverse reaction rate consistent with the thermodynamic model; the full model is illustrated in figure 4.5(a). Conceptually, the proofreading motif functions by providing two opportunities to reject the unwanted monomer w : first, when the un-activated monomer binds, and second, after it has been activated. To be effective, a non-zero affinity is required to drive the system around the cycle of states in the correct order: unbound template site $\rightarrow$ unactivated monomer bound $\rightarrow$ activated monomer bound [77, 118]. I emphasise that this model differs from Hopfield’s original description in two important ways: firstly, I consider a full, microscopically reversible polymerisation process, rather than a single incorporation step with irreversible polymerisation; and secondly, the proofreading motif has been embedded into a non-trivial polymerisation process involving autonomous detachment from the template.

Given the model as described in figure 4.5(a), one can first identify the propensities $\omega_{\pm xy}$ connecting completed states. Due to the petal-like structure, one can follow eqn. 4.3 and simply consider spanning trees of the petal sub-processes illustrated in figure 4.5; $\Lambda^-(y, x)$ rooted at $\&x$, $\Lambda^+(y, x)$ rooted at $\&xy$, and $Q(y, x)$, for a petal connecting $\&x$ and $\&xy$. Explicitly writing out the sums of spanning trees, one obtains:

$$\Lambda_1^+(y, x) = [k_1 k_{\text{act}} M_{\text{in}} + k_{KP} M_{\text{act}} (k_1 e^{-\Delta G_y} + k_{\text{act}})] k_{\text{pol}} e^{-\Delta G_x}, \quad (4.57)$$

$$\Lambda_1^-(y, x) = [k_1 k_{\text{act}} e^{\Delta G_{\text{act}} - \Delta G_y} + k_{KP} e^{-\Delta G_y} (k_1 e^{-\Delta G_y} + k_{\text{act}})] k_{\text{pol}} e^{-\Delta G_{\text{pol}}}, \quad (4.58)$$

$$Q_1(y, x) = \left[k_1 k_{\text{act}} e^{\Delta G_{\text{act}} - \Delta G_y} + (k_{KP} e^{-\Delta G_y} + k_{\text{pol}} e^{-\Delta G_x}) (k_1 e^{-\Delta G_y} + k_{\text{act}}) \right]. \quad (4.59)$$

Here, we add a subscript 1 to denote these expressions as applying to the simple, “1-loop”, Hopfield model, which we shall extend to allow more loops later. We note that the ratio, $\Lambda^+(y, x)/\Lambda^-(y, x)$ factorises as eqn. 4.34, and so we can easily write down the

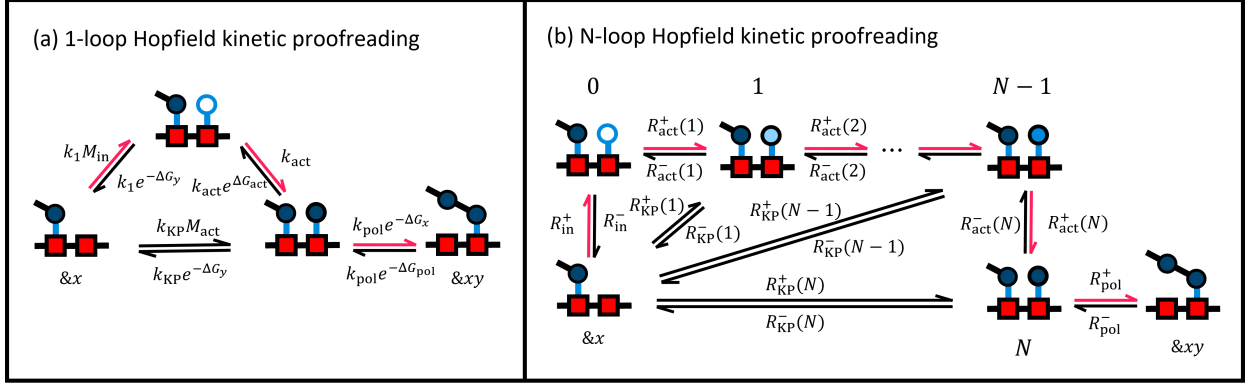


Figure 4.5: Reaction rates of the (a) 1-loop and (b) N -loop Hopfield kinetic proofreading models implemented in a templated polymerisation model with autonomous separation system. Each of these subfigures represents a single petal of the step-wise process as in figure 4.1, going from completed state $\&x \rightarrow \&xy$. In both cases, the template is represented by red squares. In (a) the inactive monomer is represented by a white circle and the activated monomers by a dark blue circle. In (b), different levels of activation are represented by increasingly dark shades of blue circles. Further, in (b) the numbers by the states represent the activation level of the monomer. In each case, the desired pathway is highlighted with red arrows.

stall condition as $\Delta G_{\text{pol}} = \Gamma$ with

$$\Gamma = -\ln \left(\frac{k_1 k_{\text{KP}} M_{\text{act}} e^{-\Delta G_r} + k_{\text{act}} (k_{\text{KP}} M_{\text{act}} + k_1 M_{\text{in}})}{k_1 k_{\text{KP}} e^{-\Delta G_r} + k_{\text{act}} (k_1 e^{\Delta G_{\text{act}}} + k_{\text{KP}})} + \frac{k_1 k_{\text{KP}} M_{\text{act}} e^{-\Delta G_w} + k_{\text{act}} (k_{\text{KP}} M_{\text{act}} + k_1 M_{\text{in}})}{k_1 k_{\text{KP}} e^{-\Delta G_w} + k_{\text{act}} (k_1 e^{\Delta G_{\text{act}}} + k_{\text{KP}})} \right). \quad (4.60)$$

Note that setting $M_{\text{in}} = M_{\text{act}} = 1$, $\Delta G_{\text{act}} = 0$, in eqn. 4.60, collapses Γ to $-\ln 2$ as these conditions reduce the system to a balanced one with a stall point at equilibrium, as in section 4.4.3.

The frequency of right and wrong monomers, $\epsilon(x = r, w)$ may be calculated from eqn. 3.14. The copying error, as represented by $\epsilon(w)$, is plotted in figure 4.6(a). This plot demonstrates that method agrees well with the results found from a Gillespie simulation [76] of the same model on a long template. The proofreading model has also been compared to a “0-loop” version of the model, in which the inactivated monomers and the inactivated monomer bound state are omitted and so the system reduces to the

simple model (figure 4.3). As can be seen, the proofreading motif generally improves accuracy when driven above its stall point $\Delta G_{\text{pol}} = \Gamma$. Indeed, one may write down expressions for the bulk frequency in the irreversible limit ($\Delta G_{\text{pol}} \rightarrow \infty$) using eqn. 4.29. In this irreversible limit, Hopfield's classic argument can be recovered by taking some further limits consistent with his analysis. Namely, letting $M_{\text{act}}, k_{\text{act}}, k_{\text{pol}} \rightarrow 0$, one finds $\varepsilon(w)/\varepsilon(r) = e^{2(\Delta G_w - \Delta G_r)}$. In this limit, the ratio of incorrect monomers to correct ones involves the square of the binding free energy difference, reflecting the fact that two steps of discrimination have occurred.

Expressions for the expected chemical work done per net step of the process may also be written down. This quantity will involve the total current to absorbing states of the step-wise process for starting with a copolymer $\&xy$, which may be written as:

$$\begin{aligned} J_{\text{Tot}}(y, x) &= \frac{1}{\mathcal{N}(y, x)} (\Lambda_1^+(r, y) Q(w, y) Q(y, x) \\ &+ \Lambda_1^+(w, y) Q(r, y) Q(y, x) + \Lambda_1^-(y, x) Q(r, y) Q(w, y)), \end{aligned} \quad (4.61)$$

where $\mathcal{N}$ is a normalisation factor that will cancel out of calculations. In order to track each of the terms here, I shall break down the contributions to the chemical work done into three parts, one for each of the petals present in the step-wise process. These three petals correspond to adding a monomer type r , adding a monomer type w or removing a monomer type y . Let us label each of these contributions to the chemical work with a subscript, $\mathcal{G}_r(y, x)$ for the transition $\&xy \rightarrow \&xyr$, $\mathcal{G}_w(y, x)$ for the transition $\&xy \rightarrow \&xyw$, and $\mathcal{G}_q(y, x)$ for the transition $\&xy \rightarrow \&x$. From the r petal:

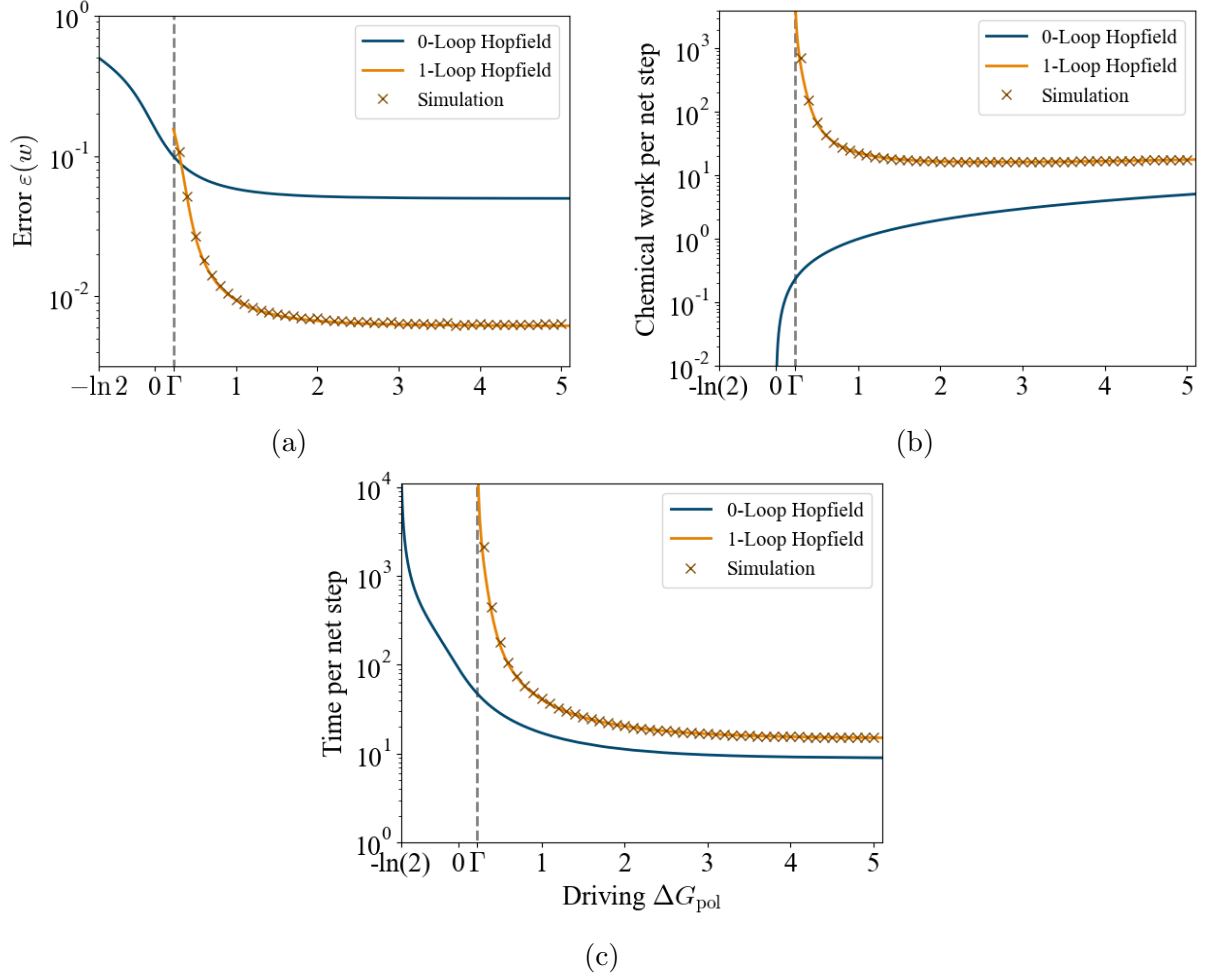


Figure 4.6: The analytical method applied to a 1-loop proofreading model (figure 4.5(a)), compared to Gillespie simulation of the same model and a simpler 0-loop model. For these data, the following parameters were used: $\Delta G_r = 2$, $\Delta G_w = -2$, $\Delta G_{\text{act}} = -1$, $M_{\text{in}} = 1$, $M_{\text{act}} = 0.01$, $k_1 = k_{\text{act}} = k_{\text{KP}} = 1$. The stall point Γ of the 1-loop proofreading model is marked on each of the plots. The Gillespie simulations used a template of length 2000 and were run till completion with the first monomer being chosen as either r or w with probability 0.5. The statistics were averaged over 2000 copolymers per data point. The chemical work was calculated from the simulation as $(\text{Inactive monomers}) * (\Delta G_{\text{act}} - \ln(M_{\text{act}}/M_{\text{in}})) + L * (\Delta G_{\text{pol}} + \ln M_{\text{act}})$ where “Inactive monomers” is the number of inactive monomers taken out of the environment and L is the length of the template.

$$\begin{aligned}
\mathcal{G}_r(y, x) = & \left[\left(-\Delta G_{\text{act}} + \ln \frac{M_{\text{in}}}{M_{\text{act}}} \right) k_1 k_{\text{act}} k_{KP} e^{-\Delta G_r} (M_{\text{in}} + M_{\text{act}} e^{\Delta G_{\text{act}}}) \right. \\
& + (\Delta G_{\text{pol}} + \Delta G_r - \Delta G_y + \ln M_{\text{in}} - \Delta G_{\text{act}}) (k_1 k_{\text{act}} k_{\text{pol}} M_{\text{in}} e^{-\Delta G_y}) \\
& \left. + (\Delta G_{\text{pol}} + \Delta G_r - \Delta G_y + \ln M_{\text{act}}) k_{KP} k_{\text{pol}} M_{\text{act}} e^{-\Delta G_y} (k_1 e^{-\Delta G_r} + k_{\text{act}}) \right] \\
& \times \frac{Q(w, y) Q(y, x)}{\mathcal{N}(y, x) J_{\text{Tot}}(y, x)}. \tag{4.62}
\end{aligned}$$

The first line of eqn. 4.62 corresponds to the chemical work associated with the internal cycle: an inactive monomer binds to the template, it is activated, and the activated monomer unbinds. The second line corresponds to an external cycle: an inactive monomer binds to the template, it is activated and polymerised into the chain with the previous monomer, y , detaching from the template. The third line corresponds to the alternative external cycle: an active monomer binds to the template and is polymerised with monomer, y , unbinding from the template. One may similarly write down $\mathcal{G}_w(y, x)$ as eqn. 4.62, except swapping r and w . Finally, the contribution to the chemical work from the petal for removing monomer y may be written:

$$\begin{aligned}
\mathcal{G}_q(y, x) = & \left[-(\Delta G_{\text{pol}} + \Delta G_y - \Delta G_x + \ln M_{\text{in}} - \Delta G_a) (k_1 k_{\text{act}} k_{\text{pol}} e^{\Delta G_{\text{act}} - \Delta G_y - \Delta G_{\text{pol}}}) \right. \\
& \left. - (\Delta G_{\text{pol}} + \Delta G_y - \Delta G_x + \ln M_{\text{act}}) k_{KP} k_{\text{pol}} e^{-\Delta G_{\text{pol}} - \Delta G_y} (k_1 e^{-\Delta G_r} + k_{\text{act}}) \right] \\
& \times \frac{Q(r, y) Q(w, y)}{\mathcal{N}(y, x) J_{\text{Tot}}(y, x)}. \tag{4.63}
\end{aligned}$$

Here, only external cycles are possible. The first line corresponds to monomer x rebinding to the template, monomer y being depolymerised, this monomer being deactivated and an inactive monomer y unbinding from the template; and the second line to x rebinding, y being depolymerised and active monomer y unbinding from the template.

The distribution, $\xi(y, x)$ may be calculated from eqn. 4.7 and P from eqn. 4.8, letting the chemical work done per net step of the the 1-loop model be written:

$$\Delta\mathcal{G} = \frac{1}{2P-1} \sum_{x,y \in \{r,w\}} \xi(y, x) (\mathcal{G}_r(y, x) + \mathcal{G}_w(y, x) + \mathcal{G}_q(y, x)). \quad (4.64)$$

This chemical work done is plotted for a certain set of parameters in figure 4.6 (b) and is also compared both to the results of direct simulation and the simpler “0-loop” model which has chemical work, ΔG_{pol} . The free-energy cost of the proofreading mechanism diverges as $\Delta G_{\text{pol}} \rightarrow \Gamma$ since there will be a finite chemical work done per monomer addition/removal step due to the proofreading internal cycle, and the number of addition/removal steps per net step diverges. Further, for large ΔG_{pol} , the work tends to be dominated by ΔG_{pol} , albeit very slowly, as shown by the orange line gradually approaching ΔG_{pol} (the blue line) in figure 4.6(b).

Additionally, one can find an expression for the time taken per net step forwards, eqn. 4.13. For this quantity, the explicit expression for the normalisation $\mathcal{N}$ is needed. Similarly to the chemical work, this term can be split into contributions from the petal adding an r , $\mathcal{N}_r(y, x)$; from the petal adding a w , $\mathcal{N}_w(y, x)$; from the petal removing monomer y , $\mathcal{N}_q(y, x)$ and a contribution from the central node. These normalisation terms come from the sums of spanning trees directed to the individual nodes in the closed step-wise process. Summing the spanning trees, we find

$$\begin{aligned} \mathcal{N}_r(y, x) &= \left[k_1 M_{\text{in}} (k_{\text{act}} e^{\Delta G_{\text{act}}} + k_{KP} e^{-\Delta G_r} + k_{\text{pol}} e^{-\Delta G_y}) \right. \\ &\quad + k_{KP} k_{\text{act}} M_{\text{act}} e^{\Delta G_{\text{act}}} + k_1 k_{\text{act}} M_{\text{in}} + k_{KP} k_{\text{act}} M_{\text{act}} + k_1 k_{KP} e^{-\Delta G_r} \Big] \\ &\times Q(y, x) Q(w, y), \end{aligned} \quad (4.65)$$

with a similar result for $\mathcal{N}_w(y, x)$ except swapping r and w . Finally, for the monomer removal petal:

$$\mathcal{N}_q(y, x) = k_{\text{pol}} e^{-\Delta G_{\text{pol}}} (k_1 e^{-\Delta G_y} + k_{\text{act}} + k_{\text{act}} e^{\Delta G_{\text{act}}}) Q(r, y) Q(w, y). \quad (4.66)$$

The total normalisation is then:

$$\mathcal{N}(y, x) = \mathcal{N}_r(y, x) + \mathcal{N}_w(y, x) + \mathcal{N}_q(y, x) + Q(y, x)Q(r, y)Q(w, y), \quad (4.67)$$

with the last term being the contribution from the starting, central node. This normalisation can be used in eqn. 4.61 to give the current to absorbing states, which can be used in eqn. 4.13 to find the expected time per net step. This time is plotted in figure 4.6(c), alongside a simulation of the same model and the simplified, 0-loop model for comparison. Like the chemical work in figure 4.6(b), the time per net step diverges as $\Delta G_{\text{pol}} \rightarrow \Gamma$, since each monomer addition/removal step will take finite time, but the number of such steps required for a net forwards step diverges. Unsurprisingly, the time taken for a given driving for the Hopfield model is longer than that of the simple model, due to the proofreading cycle.

4.4.5 A multi-loop generalisation of Hopfield's kinetic proofreading

Hopfield's model for proofreading may be naturally extended to include N activation stages instead of just on [121, 122]. I shall call these extensions the N -loop Hopfield models. These models can be solved recursively to write down expressions for the sums over spanning trees, $\Lambda_N^\pm(y, x)$, $Q_N(y, x)$, as a function of the number of loops, N . I shall consider the model as in figure 4.5(b).

One first needs to calculate the sums over spanning trees for each of these models. The sum over spanning trees of the N -loop model can be written in terms of sums over spanning trees of the lower loop number models. Thus, a set of recurrence relations can be set up for said sums over spanning trees. Being able to set up these recurrence relations easily is a great advantage which falls out naturally from the graphical methods. I label the reaction rates for the N -loop process as shown in figure 4.5(b). The N -loop model has one more node and two more edges than the $(N - 1)$ -loop model. Let a subscript, N , denote the sums over spanning trees for the N -loop models. Careful tracking of the

spanning trees gives

$$\Lambda_N^+ = R_{\text{act}}^+(N)\Lambda_{N-1}^+ + R_{KP}^+(N)\frac{R_{\text{pol}}^+}{R_{\text{pol}}^-}\sum_{i=0}^N\left[\prod_{j=0}^{i-1}R_{\text{act}}^+(N-j)\right]\Lambda_{N-1-i}^-, \quad (4.68)$$

$$\Lambda_N^- = R_{\text{act}}^-(N)\Lambda_{N-1}^- + R_{KP}^-(N)\sum_{i=0}^N\left[\prod_{j=0}^{i-1}R_{\text{act}}^+(N-j)\right]\Lambda_{N-1-i}^-, \quad (4.69)$$

$$Q_N = \frac{1}{R_{\text{pol}}^-}\left(\Lambda_N^- + R_{\text{pol}}^+\sum_{i=0}^N\left[\prod_{j=0}^{i-1}R_{\text{act}}^+(N-j)\right]\Lambda_{N-1-i}^-\right), \quad (4.70)$$

with initial conditions

$$\begin{aligned} \Lambda_{-1}^\pm &= R_{\text{pol}}^\pm, \\ \Lambda_0^\pm &= R_{\text{pol}}^\pm R_{\text{in}}^\pm, \\ Q_0 &= R_{\text{pol}}^+ + R_{\text{in}}^-. \end{aligned} \quad (4.71)$$

The sum-product can be eliminated by subtracting terms proportional to $\Lambda_{N-1}^\pm, Q_{N-1}$, leaving just:

$$\begin{aligned} \Lambda_N^+ &= \left(R_{\text{act}}^+(N) + \frac{R_{KP}^+R_{\text{act}}^+(N)}{R_{KP}^+(N-1)}\right)\Lambda_{N-1}^+ - \frac{R_{KP}^+(N)R_{\text{act}}^+(N)R_{\text{act}}^+(N-1)}{R_{KP}^+(N-1)}\Lambda_{N-2}^+ \\ &+ R_{KP}^+(N)\frac{R_{\text{pol}}^+}{R_{\text{pol}}^-}\Lambda_{N-1}^-, \end{aligned} \quad (4.72)$$

$$\begin{aligned} \Lambda_N^- &= \left(R_{\text{act}}^-(N) + R_{KP}^-(N) + \frac{R_{KP}^-(N)R_{\text{act}}^+(N)}{R_{KP}^-(N)}\right)\Lambda_{N-1}^- \\ &- \frac{R_{KP}^-(N)R_{\text{act}}^+(N)R_{\text{act}}^-(N-1)}{R_{KP}^-(N-1)}\Lambda_{N-2}^-, \end{aligned} \quad (4.73)$$

$$Q_N = R_{\text{act}}^+(N)Q_{N-1} + \frac{\Lambda_N^-}{R_{\text{pol}}^-} + (R_{\text{pol}}^+ - R_{\text{act}}^+(N))\frac{\Lambda_{N-1}^-}{R_{\text{pol}}^-}, \quad (4.74)$$

with the same initial conditions as above. This system of recursion relations may be used to generate the terms of the spanning tree sums quickly.

In certain simple cases eqns. 4.72, 4.73, and 4.74 can be solved as a function of N .

For example, when the reaction rates are not a function of N , such as:

$$\begin{aligned}
R_{\text{in}}^+ &= k_1 M_{\text{in}}, \\
R_{\text{in}}^- &= k_1 e^{-\Delta G_y}, \\
R_{\text{act}}^+(n) &= k_{\text{act}}, \\
R_{\text{act}}^-(n) &= k_{\text{act}} e^{\Delta G_{\text{act}}}, \\
R_{KP}^+(n) &= k_{KP} M_{\text{act}}, \\
R_{KP}^-(n) &= k_{KP} e^{-\Delta G_y},
\end{aligned} \tag{4.75}$$

for $n \in \{1, \dots, N\}$. For the step-wise process with monomer tip & xy , where k_i are overall rate constants, M_{in} , M_{act} represent the concentrations of inactive or active monomers, ΔG_{act} represents the chemical work upon moving a monomer up one activation stage. The rates in eqn. 4.75 are used for the numeric results in figures 4.7 and 4.8. In this case, the sums over spanning trees can be written in the surprisingly simple form:

$$\begin{aligned}
\Lambda_N^+(y, x) &= k_{\text{pol}} e^{-\Delta G_x} \left(k_{\text{act}}^N \left(k_1 M_{\text{in}} - k_1 M_{\text{act}} + k_{\text{act}} e^{\Delta G_y} M_{\text{act}} (e^{\Delta G_{\text{act}}} - 1) \right) \right. \\
&\quad \left. + \frac{k_{KP} M_{\text{act}} e^{-\Delta G_y}}{\Delta} \left[\frac{(k_1 \lambda_+ + k_{\text{act}}(k_{KP} - k_1))}{(\lambda_+ - k_{\text{act}})^2} \lambda_+^{N+1} \right. \right. \\
&\quad \left. \left. - \frac{(k_1 \lambda_- + k_{\text{act}}(k_{KP} - k_1))}{(\lambda_- - k_{\text{act}})^2} \lambda_-^{N+1} \right] \right), \tag{4.76}
\end{aligned}$$

$$\begin{aligned}
\Lambda_N^-(y, x) &= \frac{k_{\text{pol}} e^{-\Delta G_{\text{pol}}} e^{-\Delta G_y}}{\Delta} \left[(k_1 \lambda_+ + k_{\text{act}}(k_{KP} - k_1)) \lambda_+^N \right. \\
&\quad \left. - (k_1 \lambda_- + k_{\text{act}}(k_{KP} - k_1)) \lambda_-^N \right], \tag{4.77}
\end{aligned}$$

$$\begin{aligned}
Q_N(y, x) &= \frac{k_{\text{pol}} e^{-\Delta G_x}}{\Delta} \left[(k_1 e^{-\Delta G_y} + k_{\text{act}} - \lambda_-) \lambda_+^N - (k_1 e^{-\Delta G_y} + k_{\text{act}} - \lambda_+) \lambda_-^N \right] \\
&\quad + \frac{e^{-\Delta G_y}}{\Delta} \left[(k_1 \lambda_+ + k_{\text{act}}(k_{KP} - k_1)) \lambda_+^N - (k_1 \lambda_- + k_{\text{act}}(k_{KP} - k_1)) \lambda_-^N \right], \tag{4.78}
\end{aligned}$$

where

$$\lambda_{\pm} = \frac{1}{2} (k_{\text{act}} + k_{KP}e^{-\Delta G_y} + k_{\text{act}}e^{\Delta G_{\text{act}}} \pm \Delta), \quad (4.79)$$

$$\Delta = \sqrt{(k_{\text{act}} + k_{KP}e^{-\Delta G_y} + k_{\text{act}}e^{\Delta G_{\text{act}}})^2 - 4(k_{\text{act}})^2e^{\Delta G_{\text{act}}}}. \quad (4.80)$$

From eqns. 4.76, and 4.77 one may write the stall condition. Noting that, $\Lambda^+(y, x)$ is independent of ΔG_{pol} and $\Lambda^-(y, x)$ is proportional to $e^{-\Delta G_{\text{pol}}}$, one may write the stall point as

$$\Gamma_N = -\ln \left(\frac{\Lambda_N^+(r, r)}{e^{\Delta G_{\text{pol}}} \Lambda_N^-(r, r)} + \frac{\Lambda_N^+(w, w)}{e^{\Delta G_{\text{pol}}} \Lambda_N^-(w, w)} \right), \quad (4.81)$$

such that the dependence on ΔG_{pol} in the logarithm is cancelled out.

From these sums over spanning trees, the bulk frequencies can be calculated. The time taken per net step and the chemical work done per net step may be calculated using similar recursive relations.

For simplicity, I shall discuss the case where the monomer binding free energy is only dependent on monomer type, not on activation stage; each activation stage is associated with a free energy change of ΔG_{act} ; each active monomer is present in the environment at a concentration M_{act} except the inactive monomers at concentration M_{in} ; and the overall rate constants are k_1 for binding of inactive monomer, k_{KP} for binding of active monomers, k_{act} for activation of monomers. Under these assumptions, the corresponding rates are as in eqn. 4.75.

To reduce the frequency of incorrect monomers in the product, one wishes to have a low concentration M_{act} of active monomers in solution to force the system into utilising the proofreading cycles. Indeed, the bulk error probability in the irreversible limit (calculated using eqn. 4.29 and plotted in figure 4.7(a) shows a strong improvement with loop number for low M_{act} , but larger values of M_{act} lead to much worse performance and limited (or negative) returns to increasing the number of loops.

However, for finite driving strength ΔG_{pol} , this concentration cannot be allowed to

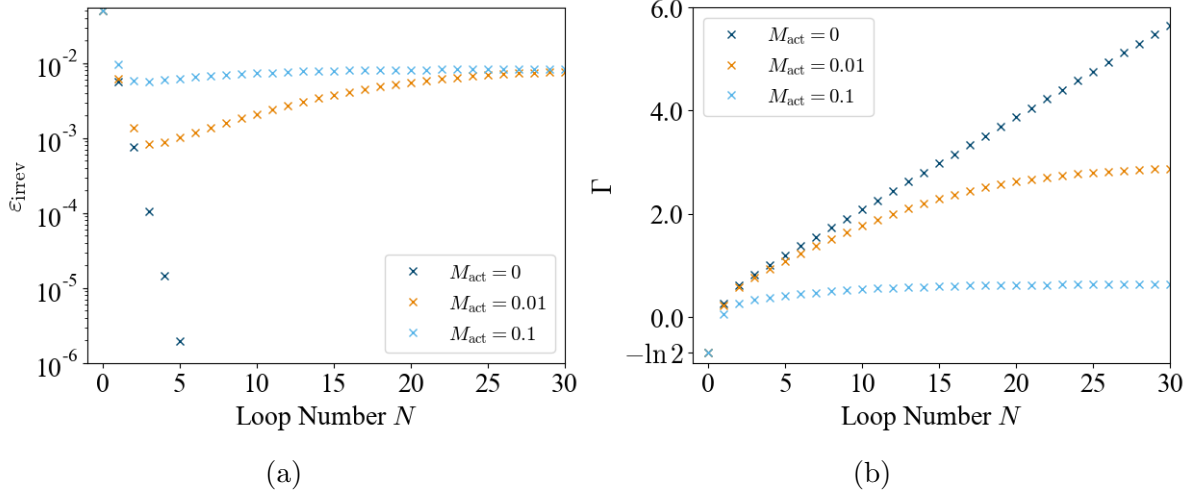


Figure 4.7: Many-loop models have limited efficacy for finite M_{act} for the proofreading model introduced in figure 4.5. Plots of (a) the error in the irreversible limit, $\epsilon_{\text{irrev}}(w)$ and (b) the stall point driving, Γ , for different values of the active monomer concentrations, M_{act} , and other parameters: $\Delta G_r = 2$, $\Delta G_w = -2$, $M_{\text{in}} = 1$, $\Delta G_{\text{act}} = -1$, k' s = 1.

be arbitrarily small. To see why, consider the stall point Γ_N (eqn. 4.81), which is plotted for a certain set of parameters in figure 4.7(b). It is observed that the stall point driving increases monotonically with N , and that this increase is faster and tends to a higher limit for smaller M_{act} . One finds that the limiting Γ scales approximately linearly with $-\ln(M_{\text{act}})$. Intuitively, introducing more monomer states at low concentration in the environment destabilises the polymer. For small M_{act} and driving ΔG_{pol} , the depolymerisation of the polymer into these activated states competes with its tendency to grow by binding to and activating the inactive monomers.

One drawback of proofreading with a large number of loops is therefore that the tendency to disassemble the growing polymer increases. A second effect is a tendency to introduce errors by alternate pathways if M_{act} is non-zero. Specifically, for $M_{\text{act}} \neq 0$, a minimum in $\epsilon_{\text{irrev}}(w)$ for a relatively small value of N can be observed in figure 4.7(a). This minimum can be explained by splitting the pathways by which a monomer can go from solution to being incorporated into the polymer into two, either starting from a fully inactive monomer or from a partially activated one. The pathway starting with an

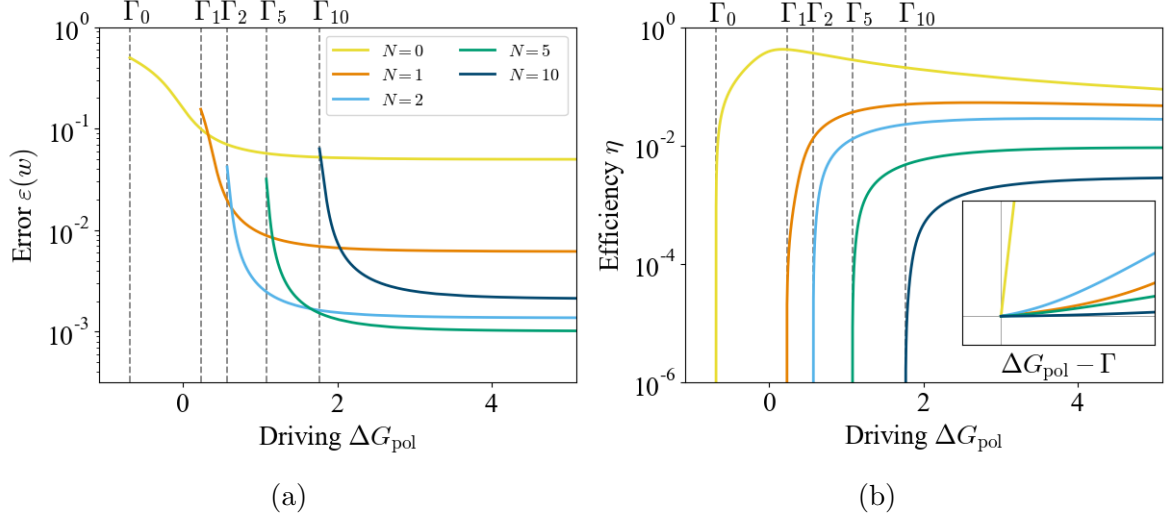


Figure 4.8: Plots of (a) the error and (b) efficiency of the N -loop proofreading model (figure 4.5) for a range of N , with the same parameters as in the one loop Hopfield case, figure 4.6, as a function of driving ΔG_{pol} . Proofreading is observed to generally increase accuracy when ΔG_{pol} exceeds the stall point, but in a thermodynamically inefficient way. The enhanced plot in the second graph shows the efficiencies near the stall point for each of the loop numbers on a non-logarithmic scale to emphasise the decreasing gradient at stall.

inactive monomer will have the highest discrimination between right and wrong monomers and will improve exponentially with more loops, as demonstrated by the exponential decrease in error for $M_{\text{act}} = 0$. However, the probability that a monomer, taking this path, will reach polymerisation falls exponentially with loop number at the same time. On the other hand, the pathway from partially active monomers will give an error that reaches some non-zero limit as the number of loops N increases. Further, the rate with which activated monomers bind to an available template site and subsequently get incorporated into the polymer will also tend to a constant. As such, the error will initially decrease exponentially with N , but for non-zero M_{act} will eventually become dominated by the less discriminating, partially active monomer pathways through which monomers are more likely to be incorporated into the polymer.

Having calculated the error probability $\epsilon(w)$ at finite driving, plotted in figure 4.8 (a); used $\epsilon(x, y)$ to calculate the entropy rate; and calculated $\Delta \mathcal{G}$; the efficiency η can be

evaluated as in eqn. 4.14. This efficiency is plotted in figure 4.8 for $N = 0, 1, 5, 10$ and a certain set of parameters. Although accuracy is generally increased above the stall point, it can be seen that, in this particular model, kinetic proofreading requires much more work than the minimum required to generate information and as such is inefficient. Additionally, the gradient of the efficiency at minimum driving, Γ_N , is zero for $N > 0$, reflecting how at minimum driving, the number of monomer addition/removal steps diverges, but the chemical work done per such step remains finite.

4.5 Conclusion

In this chapter, I have presented a method for analysing thermodynamically self-consistent copolymerisation models with complex networks of reactions leading to the incorporation or removal of monomers. By coarse-graining, a model may be transformed into a simpler model which may be solved and then afterwards, information from the fine-grained process may be put back into the model to extract thermodynamic or kinetic quantities such as chemical work done, molecule exchange or time taken. The approach allows for complex incorporation motifs to be considered alongside nearest neighbour interactions in a thermodynamically well-defined model of polymerisation with microscopic reversibility. I note that all of these features were present in the kinetic proofreading example in section 4.4.4. Moreover, phenomena such as the shift in stall point with loop number and the non-monotonicity of error rate with loop number rely on these features being present in the model.

In general, this method provides a way to extract model predictions numerically quickly and without the need for simulations. Doing so is particularly useful when simulating polymer growth is slow, either due to the details of the incorporation process or because the polymer is near its stall point. Additionally, the approach makes screening of a large parameter space for a given model topology feasible.

In addition to the numerical performance, the approach allows for analytic results in simpler models or those with helpful symmetries, as well as in certain limits for more

complex models. The process of summing over spanning trees is particularly well suited to identifying the structure of the process and providing simplified results.

Moving forwards, it is an open question as to whether components of the techniques developed here can be applied outside of the context of infinitely long polymers whose tips have reached steady state, although I address progress in this regard in chapter 6. An obvious goal would be a simplified way to analyse finite-length “oligomers” [114]. More generally, I believe one of the key equations of this thesis, eqn. 4.5, may be applied more generally for the coarse-graining of Markov processes. Specifically, that if a set of states are enclosed between two boundary states, in the sense that any path from one of the trapped states to outside must pass through one of the boundary states, then this set of states may be replaced by a pair of edges analogously to eqn. 4.5 which shall preserve steady state properties of the Markov process.

This framework could be applied to explore models of copolymerisation processes, such as those presented in [15, 91, 92, 99–104, 109–113, 115, 120–127], more straightforwardly or more thoroughly. Alternatively, the method would allow for more complex reaction steps to be included in such models. The framework presented here is particularly useful when backwards steps are relevant, either when the system is weakly driven and thus operating near stall, or when thermodynamics is of importance or interest. I also predict that it will be useful to guide design principles for synthetic copolymerisation systems, which are often particularly well described by the class of models studied here

Chapter 5

Bounds on the accuracy of catalytic templating networks with multiple pathways

The following chapter has been adapted from the preprint “Information propagation in far-from-equilibrium molecular templating networks is optimised by pseudo-equilibrium systems with negligible dissipation” [136] to fit this thesis. The content included in this this chapter is the work of myself, Benjamin Qureshi.

5.1 Introduction

As has been seen in section 1.3, biochemical systems use catalysts to selectively produce multiple distinct products from a shared set of ingredients. For example, in protein production [39], many polypeptide sequences are created from the same set of amino acids, using a set of template mRNAs and other molecules acting as selective catalysts [40]. The result is a distribution of polypeptide sequences in the cell, sharply peaked around the target sequences specified by the mRNA.

However, biological systems often continuously turn over the products, recycling the

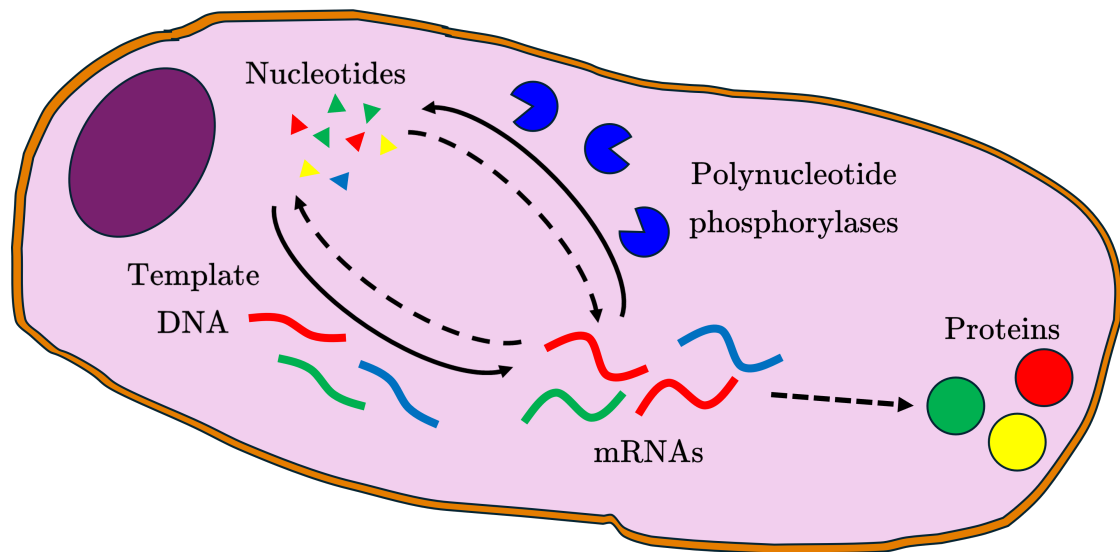


Figure 5.1: Cartoon of the multiple pathways present in the production/breaking down of mRNAs in the cell. In this cartoon, nucleotides, represented by triangles are converted into mRNAs via a pathway catalysed by template DNA (leftmost solid arrow). The mRNAs can be degraded back into nucleotides via a pathway catalysed by polynucleotide phosphorylases (right solid arrow). Each of these reactions (template DNA and polynucleotide phosphorylases) can occur in the reverse direction (dashed arrows), although the rates of these reversed reactions will be much slower than the forwards direction. The mRNAs will eventually be used to catalyse the production of proteins which are the functional molecules the cell requires.

basic ingredients for subsequent use. For example, polynucleotide phosphorylase enzymes actively degrade the pool of selectively transcribed RNA [137] (illustrated cartoonishly in figure 5.1). Degradation allows the cell to set up a finite steady-state population that is biased towards specific products, and continuous turnover and recycling of components is necessary for systems with finite resources that need to change their state in response to dynamic environments. The underlying processes of production or degradation, interacting with a given catalyst may also be complex and involve many possible routes of assembling and disassembling the products. For example, kinetic proofreading motifs (as in [92, 118, 119] and section 3.4) introduce fuel-consuming loops into the assembly process, thereby providing multiple pathways to forming each product.

So far, I have only discussed models whereby the products are produced by a single mechanism; for example, the template-based mechanisms of chapter 4. These models, additionally, only track the growth and statistics of a single product, instead of the whole distribution of products. Further, these mechanisms discussed do not model the eventual final detachment of polymer and template. As such, they do not account for the possibility of products rebinding to the template and being broken down. If one wishes to model systems with explicit destructive pathways such as via polynucleotide phosphorylase enzymes (figure 5.1), a different framework is needed. In this chapter I shall provide such a framework. This chapter is concerned more with the bigger picture of the whole polymerisation and subsequent degradation process, compared to chapter 4, where the focus was more on detailed models of polymerisation.

Previous studies [89, 91, 92, 97–104, 109–115, 120–123, 126, 134, 138] have generally focused on such models of forming a single product, in particular considering how templated copolymerisation in isolation can produce an accurate sequence. However, to the best of my knowledge, only refs. [40, 139, 140] have considered the implications of two or more pathways for production and destruction of an ensemble of products, in each case without considering a detailed model of dynamical pathways. These works hypothesise a minimal thermodynamic cost or entropy production per product made, which is related to the information accurately transferred from template to products [40, 139, 140]. However, en-

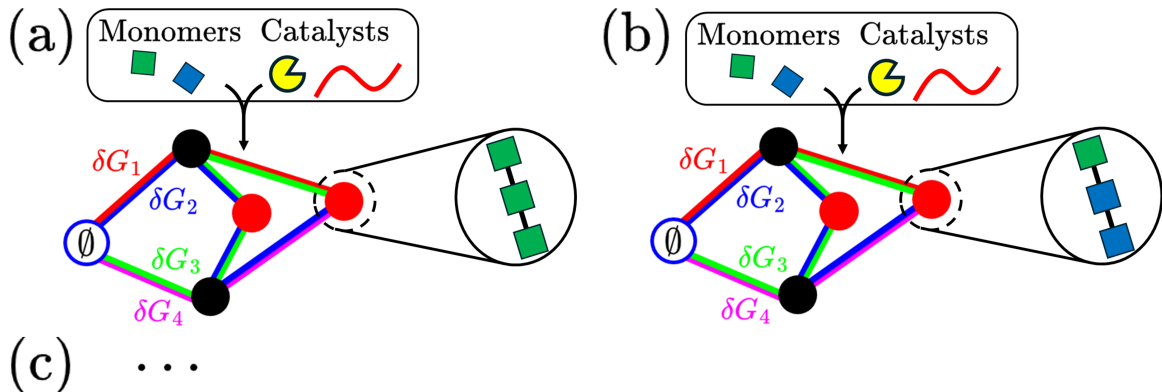


Figure 5.2: Model systems in which products are selectively produced by catalysts from a pool of monomers. A very simple example network illustrating the class of systems considered herein. Any one of a number of (polymer) products can be produced via equivalent pathways that rely on catalysed addition of input monomers. Both catalysts and monomers are assumed to be held at constant concentration by a chemostat, resulting in a linear CRN that can be represented as a graph. In (a), I illustrate a simple example of such a graph, including a null, polymer-free state; intermediate states represented by black circles and products represented by red circles. Each transition is reversible (arrowheads and rates are omitted for clarity). Each distinct pathway from the null state to the product is coloured a unique colour, and has a well-defined associated free-energy change δG . In (b) I take an alternative product as the root of the graph; the network connecting it to the null state is topologically identical to (a). The equivalent pathways for each product have the same δG , but selective catalysts ensure that the rates associated with each pathway are product-dependent. (c) indicates that if there were more products, there would be an equivalent representation of the graph for each such product node.

trophy production in stochastic thermodynamics is related to the relative rate of forwards and backwards transitions (section 2.4), rather than the relative rate of two different transitions. It therefore places no minimal cost on the specificity of a catalyst [114]. Exactly how these two ideas are reconciled is unclear.

The limiting thermodynamics of this crucial class of processes therefore remains underexplored. In this chapter, I consider the thermodynamics of a general system in which a distribution of products is created and maintained by a set of catalysts that interconvert molecules between pools of monomer inputs and pools of assembled products. These inter-conversion processes can be arbitrarily complex, but I shall assume that all products have the same thermodynamic stability [40, 113, 114, 138, 140], consistent with the fact that catalysts cannot adjust the thermodynamic stability of the products that they select. Moreover, all products can be formed by a topologically equivalent set of processes (figure 5.2). By analogy with the processes of the central dogma, I shall loosely describe the catalysts that tend to generate products as “templates”, but I do not make any particular assumptions about the detailed mechanisms.

Each pathway for creating or destroying a given product can incur a different free-energy change (figure 5.2), due to coupling with ancillary fuel molecules. These free-energy changes specify the equilibrium that would result in the long time limit if that pathway operated in isolation. I shall show that this set of equilibrium free energies sets a lower bound on the entropy of the product distribution in a non-equilibrium steady state, for any fuel-coupled, catalyst-induced kinetics. If the system is interpreted as an information channel between the input catalysts and the output products, this lower bound on entropy is equivalent to an upper bound on channel capacity (eqn. 2.5). Similarly to section 4.3, under symmetric assumptions on the relationship between input catalysts and output products, the channel capacity looks like the negative of the entropy (eqn. 2.6).

The bound on product distribution entropy also sets a bound on the accuracy with which a single product can dominate the resultant ensemble. But surprisingly, the goals of producing a distribution maximally specific to a single chemical species and producing the distribution of products with minimal entropy are generally distinct. Moreover,

contrary to my expectations, the optimal system that minimises the entropy of the product distribution – pushing the distribution as far from equilibrium as possible – is not one in which desired products are systematically assembled on a template and degraded by another catalyst [139]. Rather, the optimal system is a pseudo-equilibrium one in which kinetic specificity ensures that each product is coupled overwhelmingly to only a single pathway, with each product’s yield specified by the equilibrium properties of that pathway.

Let M be the number of possible products (*e.g.* number of sequences of monomers of a given length, in which case $M = A^L$ for polymers of length L with A different monomer types). Exploring the bound, one will observe that, for $M \rightarrow \infty$ possible products, it is possible to selectively produce a single product with perfect accuracy in steady state provided two pathways exist that are separated by ΔG in their free-energy change such that $\Delta G / \ln M > k_B T$, irrespective of other pathways. However, the probability of producing a single correct product as $M \rightarrow \infty$ is zero for $\Delta G / \ln M < k_B T$. Further, the convergence to the infinite M behaviour is slow. Similarly, zero entropy product distributions can never be achieved with finite ΔG , and even low entropy requires $\Delta G > k_B T (\ln M + \ln \ln M)$.

These results appear consistent with the notion that accurately maintaining a specific pool of products has a thermodynamic cost related to the information stored ($\ln M$ if a single template selects for a unique product) [40, 139, 140]. However, this simple picture is incomplete. Since the bound is saturated for pseudo-equilibrium systems in which each product couples overwhelmingly to a single pathway, there is no minimal entropy production per product molecule created in these steady states. Moreover, for $M \rightarrow \infty$, although a single product cannot be maintained with any accuracy for $\Delta G / \ln M < k_B T$, it is possible to maintain a vanishingly small fraction of the total possible products with perfect accuracy for any $\Delta G \gg k_B T$, even if $\Delta G \ll k_B T \ln M$. Finally, the bounds only apply to steady states, and it is possible to produce distributions of arbitrary precision at finite time with no minimal ΔG . Throughout this chapter, I shall make use of the symbol “ $\lesssim$ ” to mean “less than something which is approximately equal to” and similarly

for “ $\succsim$ ”.

5.2 Models and methods

5.2.1 Physical model

I shall now define the class of models considered, illustrated in figure 5.2. Although this figure illustrates short polymer (or oligomer) products of templating for ease of understanding, the approach here is not specific to any particular mechanistic model of template-based growth. Indeed, formally, the products need not even be polymers or oligomers.

Consider a set of input chemical species (monomers) that may be assembled into a number of different possible products. For simplicity, and to focus on information propagation, I shall assume that all input species are chemostatted at the same concentration by the cell or the environment. In the context of the central dogma of molecular biology, these inputs would be the nucleotides or amino acids. There are M possible different products that these inputs may form, corresponding to different copolymer sequences (*e.g.* the different base sequences of RNA). Finally, there are catalytic species (*e.g.* templates, polymerases, or nucleases) that facilitate the inter-conversion of inputs and products via intermediates. I assume that these catalysts are also chemostatted, and neglect any reaction involving more than one product and/or intermediate. The set of products formed is now the object of interest.

Such a model may be expressed as a deterministic CRN (section 2.3), with chemical species partitioned into products, intermediates, and inputs and catalysts. Given the assumptions about their chemostatting, the concentrations of input and catalyst species can be treated as parameters, rather than variables. Under such assumptions, the resultant CRNs are linear (section 2.3): the set of complexes contains only the null complex (containing no non-chemostatted species), and singletons (containing a single non-chemostatted species, *i.e.* an intermediate or a product). All reaction rates there-

fore depend linearly on the concentration of the non-chemostatted species. Transitions from the null complex correspond to the first association between inputs or inputs and catalysts to produce an intermediate.

This linear CRN can be drawn as a graph (figure 5.2(a)), with nodes as species, edges as reactions and weights as reaction rates, absorbing the concentration of any chemostatted species involved into the reaction rates. I shall assume the graph produced is connected, as it must be if all species can be produced from the null state. This graph defines a set of SAWs from the null state to each product; each SAW together with its inverse defines a “pathway” between the null state and a product.

The transitions are allowed to be coupled to implicit fuel molecules, meaning that different SAWs are associated with different standard free-energy changes δG . However, each edge must correspond to the turnover of a fixed number of fuel molecules; branching pathways allow for variable fuel consumption when coupled to a given catalyst. Each SAW from the null complex to a product therefore has a well-defined δG .

Products are all assumed to be connected to the null complex by topologically equivalent graphs (compare figure 5.2(a) and (b)). Heuristically, every pathway from null to a given product has a corresponding equivalent in each of the sets of pathways from null to each other product. More formally, let $\mathcal{G}$ be the vertex-labelled graph describing the linear CRN containing a vertex corresponding to the null complex $\emptyset$ and a set of vertices corresponding to products, ignoring the edge-weights. Then for any pair of products, X and Y , consider a relabelling of the nodes such that the nodes for X and Y swap labels, $\emptyset$ ’s label is not changed, but other node labels may be permuted. Suppose, there exists such a relabelling, transforming the graph to a new vertex-labelled graph $\mathcal{G}'$ such that $\mathcal{G}'$ is isomorphic as labelled graph to $\mathcal{G}$. Then the set of SAWs from $\emptyset$ to X will be topologically equivalent to those SAWs from $\emptyset$ to Y . Further, a given SAW from $\emptyset$ to X will then be topologically equivalent to a specific SAW from $\emptyset$ to Y . If these equivalent SAWs also incur the same free-energy change, then the model falls into the framework of models I discuss in the chapter.

Such an assumption on the topological equivalence of graphs makes sense for a set of

products corresponding to, for example, RNA molecules of a fixed length, which can all be grown via an equivalent set of steps. Further, all products are assumed to be equally thermodynamically stable. This assumption is consistent with the fact that catalysts cannot thermodynamically favour one product over another [40, 113, 114, 138, 140], and allows the focus to be on the question of how information is propagated from catalysts to products.

As a consequence, the equilibrium, defined by the free energy of formation of products directly from their monomers, would contain a uniform distribution over products [40], and a narrower distribution over products can only be achieved via kinetic selection of a non-equilibrium steady state by the catalysts. Moreover, there exists a finite number of SAWs from the null complex to any given product, each with a free-energy change δG_i where i indexes the given SAW. The set of SAWs from the null complex to any product will be topologically equivalent and incur the same set of total free-energy changes (figure 5.2).

These constraints on free energy only apply to full paths from the null state to completed products. Intermediate states, involving partial, catalyst-bound products, can have sequence-specific free energies. Moreover, the transition rates between states can be different, subject to the overall constraint that, along a given complete SAW from null to product, the total free-energy change is the same as its equivalent for all other products. I shall explore how relative yields of the products can be optimised by varying these strictly positive rates arbitrarily under this constraint.

5.2.2 Graphical methods for linear CRNs with a null complex

I shall now expand on the features of linear CRNs, as promised in section 2.3. The steady state concentration of species X in a linear, deterministic CRN with a null species can be found directly from the complexes graph, making use of the matrix tree theorem (eqn. 2.8). The absolute concentration is given by the ratio of the sum of spanning trees of the graph rooted at the species X to the sum over spanning trees rooted at the null

species. Explicitly,

$$c_X = \frac{\sum_{T \in \mathcal{T}(X)} \prod_{e \in T} k(e)}{\sum_{T \in \mathcal{T}(\emptyset)} \prod_{e \in T} k(e)}, \quad (5.1)$$

where $\mathcal{T}(X)$ is the set of spanning trees rooted at node X , $e \in T$ is an edge in graph T , and $k(e)$ means the weight of edge e . Although, this result can be seen in *e.g.* [141, 142], I shall prove this result here with particular reference to the effect of a null species.

Consider a linear connected CRN under mass action kinetics. Let X_i be the chemical species with concentration c_i . Assume the CRN contains some reactions of the form $\emptyset \rightleftharpoons X$. Without loss of generality, assume that there only exists at most one reaction of the form $X_i \rightarrow X_j$ (for the cases where there are multiple such reactions, replace their reaction rate with the sum over all reaction rates of reactions of that form). The steady state equation for the vector of steady state concentrations of the chemical species, $\mathbf{c}$, may be cast as the linear equation:

$$A\mathbf{c} + \mathbf{b} = \mathbf{0}, \quad (5.2)$$

where $\mathbf{b}$ is a vector such that entry b_i is the rate constant of the reaction $\emptyset \rightarrow X_i$, and A is a matrix with off diagonal entries A_{ij} equal to the rate constant of the reaction $X_j \rightarrow X_i$ and diagonal elements A_{ii} equal to minus the sum of rate constants of reactions $X_i \rightarrow X_j$ over all chemical species X_j for $j \neq i$ as well as $X_i \rightarrow \emptyset$. Note that the sum of column i of matrix A is equal to minus the rate constant of reaction $X_i \rightarrow \emptyset$. Thus, we may create the new matrix:

$$K = \left(\begin{array}{c|c} A & \mathbf{b} \\ \hline \mathbf{d}^T & -\sum_i b_i \end{array} \right), \quad (5.3)$$

where $\mathbf{d}$ is a vector such that entry d_i is equal to the rate constant of reaction $X_i \rightarrow \emptyset$. The columns of the matrix, K , now sum to zero and can be recognised as the Laplacian matrix of a certain graph. Represent K as a graph with nodes corresponding to chemical

species X_i and an additional node corresponding to $\emptyset$, and edges e corresponding to reactions between species with weights equal to their rate constant $k(e)$. Then, by the matrix tree theorem (eqn. 2.8), the determinant (up to a sign) of the sub matrix formed by deleting row and column i from matrix K is given by the sum over the set of spanning trees rooted at node i , $\mathcal{T}(X_i)$,

$$\det(K/i) = \sum_{T \in \mathcal{T}(X_i)} \prod_{e \in T} k(e), \quad (5.4)$$

where (K/i) represents matrix K with row and column i deleted. In particular,

$$\det(A) = \sum_{T \in \mathcal{T}(\emptyset)} \prod_{e \in T} k(e). \quad (5.5)$$

Thus, by Cramers rule [143], and the sign of the determinant under swapping of rows, eqn. 5.1 gives the solution to eqn. 5.2 and proves the desired result.

These sums over spanning trees may be usefully rewritten as a sum over the SAWs from the null species to the given product species X (section 2.2.1). Having done so, the results are expressed in terms of the free-energy change along a given SAW [81] S ,

$$\delta G_S = -\ln \left(\prod_{e \in S} \frac{k(e)}{k(\bar{e})} \right), \quad (5.6)$$

where $\bar{e}$ is the reverse reaction of edge e . Note that SAWs with positive δG_S correspond to pathways that tend to disassemble products. These pathway free-energy changes are the main input to the bounds to be derived in this chapter.

5.2.3 Information theoretic measures of product distributions

Using these graph theoretical methods, one may calculate the steady state concentrations for all of the product species. Having done so, and denoting the products by Z_i for $i = 1, \dots, M$, define the product distribution as the probability that a randomly chosen

product from the resultant ensemble is Z_i ,

$$\mathbb{P}(Z_i) = p_i = \frac{c_{Z_i}}{\sum_{j=1}^M c_{Z_j}} = \frac{c_{Z_i}}{c_T}, \quad (5.7)$$

where $c_T = \sum_{j=1}^M c_{Z_j}$ is the total concentration. The Shannon entropy (section 2.1) of this distribution may be written as

$$H[p_i] = - \sum_{i=1}^M p_i \ln p_i = \ln c_T - \frac{1}{c_T} \sum_{i=1}^M c_{Z_i} \ln c_{Z_i}. \quad (5.8)$$

As I have discussed before, the deterministic CRNs in this thesis can be considered to be information channels. With such a consideration in mind, the minimisation of the Shannon entropy here maximises the channel capacity of this system under certain assumptions, if the system is treated as an information channel from catalysts to products. Informally, the lower the entropy of the product distribution, the more information the samples from the distribution contain about the templates which formed them. More formally, let us assume the underlying chemistry, which determines the rate constants appearing in the un-linearised model, and the concentration of monomers, are fixed. The effective rates of the linearised network would then vary with the concentrations of catalysts only. As a specific (simple) example of how the entropy bound defines the channel capacity, let us assume that all variability is due to M sequence-specific template catalysts, one for each product, and that of these templates exactly one is present at any given time at a fixed concentration.

One can define the input to the information channel as the template that is present at high concentration; the output would then be a product sampled from the steady state product distribution for that input state. The mutual information between the template input and output products can then be calculated. If the templates are symmetric, acting equivalently relative to their ideal sequence, then each output distribution would merely be a permutation of the set of product probabilities.

A system of this kind would define a symmetric channel (section 2.1). The channel capacity of such a symmetric channel (eqn. 2.6) is given by $C = \ln M - H[p_i]$ in the notation of this chapter, where M is the number of products/sequence-specific templates, and $H([p_i])$ the Shannon entropy of the output distribution for any single template. This entropy will be the same for any input state in this symmetrised description. Hence, minimising the entropy maximises the channel capacity.

5.2.4 Simulation of specific networks

Although calculations of the bounds I shall derive (section 5.4) is often straightforward, evaluation of the actual performance of a system realisation can be more challenging. For the examples in section 5.8, steady state distributions are found by numerical solution of the underlying ordinary differential equations (ODEs). The CRNs are linear, connected and contain only one stoichiometric compatibility class. Hence, there exists a single positive steady state to the ODEs induced by mass action kinetics [81]. Initially, all concentrations are set to zero, and the ODEs are simulated for a large time until no change to the distribution is observed. For section 5.8.2, to speed up simulation, I make use of the coarse-graining of eqn. 4.5. Using these coarse-grained rates does not effect the steady state by design.

5.3 Free-energy differences between pathways bound accuracy for a simple network with two catalysts

To build intuition, I shall first consider the simplest non-trivial system in which inputs and products are inter-converted by catalysts (figure 5.3). The reaction network has two product species that I denote “right” (R) for the target species and “wrong” for the other species (W). These products are formed by catalytic activation of their inactive states $\bar{R}$ and $\bar{W}$, which are the input monomers, present at equal concentrations. Absorbing the concentrations of $\bar{R}$ and $\bar{W}$ and the catalysts into the rate constants, the system may be

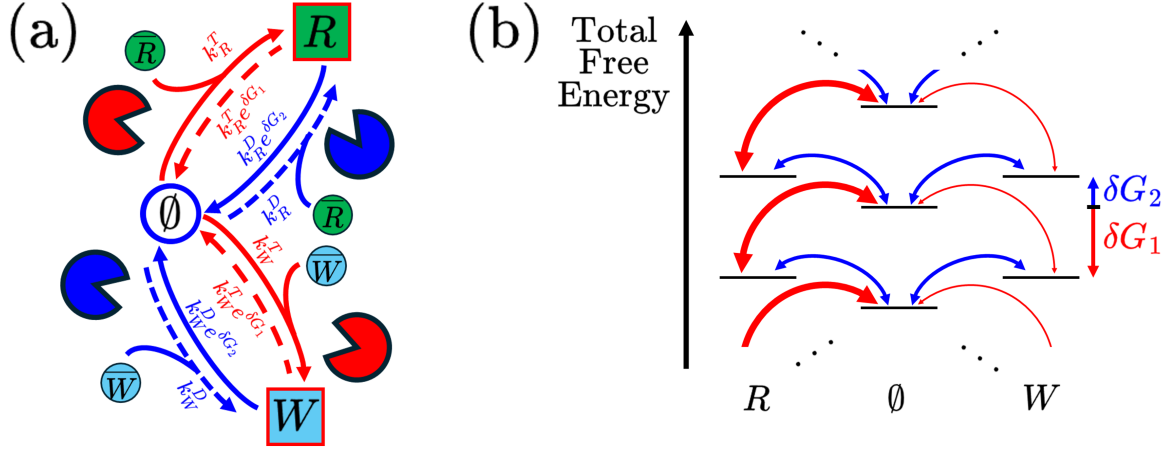
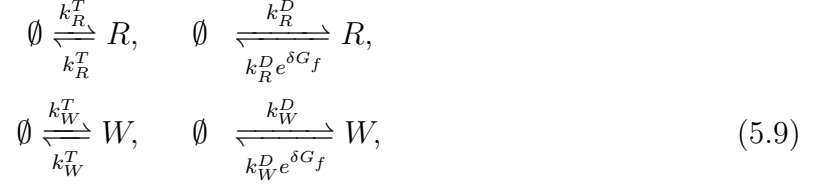


Figure 5.3: (a) The simplest non-trivial system in which monomers ($\bar{R}$ and $\bar{W}$) at equal concentration are converted to and from products R and W by two catalysts. R and W are connected to $\emptyset$ via topologically equivalent networks, involving a pathway via the red T catalyst and a pathway via the blue D catalyst. The free-energy change along each pathway is the same for each product; only the transition rates distinguish R and W . For simplicity, the short-lived catalyst-bound intermediates are not explicitly represented. $\bar{R}$ and $\bar{W}$ are each assumed to be chemostatted at the same concentration, which have been absorbed into the reaction rates. (b) Free energy landscape for the model in (a). The possibility of turning over fuel molecules introduces a ladder of R , W and null states, connected by the transitions involving the two catalysts (red and blue arrows, respectively). Overall δG from a given null state to given pair of R and W states is equal, but relative rates of downhill transitions (indicated by the thickness of the arrows) can be very different between R and W . In this case, the destructive (blue) catalyst is non-specific, with the same rates for R and W .

defined via the following CRN:



where the effective reaction rate constants are written above/below the arrows. For each product there are two pathways from the null state leading to the product, one via a catalyst that I call the template (reactions labelled by T in equation 5.9) and one via a “destructive” catalyst (reactions labelled by D in equation 5.9). For simplicity, the catalyst-bound intermediates are assumed to be short-lived, and, as such, are not explicitly represented.

The pathways via the destructive catalyst are both biased with free energy $\delta G_2 = \delta G_f > 0$ against creating the product, relative to the pathways via the template ($\delta G_1 = 0$). This bias towards destroying the product is why I, informally, call D a “destructive” catalyst. A non-zero δG_f implies that the D (or T) pathway is coupled to the turnover of implicit molecular fuel, allowing the apparent disruption of detailed balance [77]. Each individual pathway, however, has a well-defined free-energy change associated with it; in the presence of only that pathway, the system would relax to an equilibrium defined by the free-energy change of that pathway.

The two products R and W can be produced by topologically equivalent pathways, by processes with the same overall free energy change. R and W are therefore both equally thermodynamically stable and a non-equilibrium excess of R over W is only possible if the template is kinetically selective for production of R over W . This setup is a minimal analogue of a transcription-like system in which all products are thermodynamically viable, but those selected by the presence of their templates are preferentially produced.

In a conventional analysis of a system in which a template selectively produces a specific daughter molecule [89, 91, 97–104, 109–111, 113–115, 121–123, 126, 134], it is typical simply to consider how selective the template is when producing the intended product.

The net current of product, R , along pathway T , in steady state can be calculated to be

$$J^T(R) = k_R^T(1 - c_R) = \frac{k_R^T k_R^D (e^{\delta G_f} - 1)}{k_R^T + k_R^D e^{\delta G_f}}, \quad (5.10)$$

where c_R is the concentration of R . Equivalent equations have the same form for $J^D(R)$, $J^T(W)$, and $J^D(W)$. At the steady state, $J^T(s) = -J^D(s)$ for $s = R, W$.

By varying the underlying rates at fixed δG_f , the net current ratio relating the net production rate of R and W by T can be straightforwardly varied. A quick calculation shows that this ratio is

$$\frac{J^T(R)}{J^T(W)} = \left(\frac{k_R^T k_R^D}{k_R^T + k_R^D e^{\delta G_f}} \right) \bigg/ \left(\frac{k_W^T k_W^D}{k_W^T + k_W^D e^{\delta G_f}} \right). \quad (5.11)$$

This ratio can be made arbitrarily large for any δG_f , simply by picking appropriate rates $k_R^T \gg k_W^T$, forcing the template to produce output R at a rate arbitrarily higher than the rate it produces output W . The catalyst can in theory be arbitrarily kinetically selective for a given δG_f along each pathway; kinetic selectivity of catalysts is not – in principle – constrained by thermodynamics.

Molecules generated by catalytic templating networks – such as RNA and proteins – typically have their effect by being present in the cell, rather than through the process of their creation. The ratio of net currents for templating is therefore only important in so far as it enables a large ratio of product concentrations. Given the lack of a constraint on the net current ratio, one might expect to be able to produce an output distribution that contains an arbitrarily higher concentration of R than W by varying the reaction rates at fixed δG_f . However, this is not the case. The steady-state concentrations of R and W are

$$c_R = \frac{k_R^T + k_R^D}{k_R^T + k_R^D e^{\delta G_f}}, c_W = \frac{k_W^T + k_W^D}{k_W^T + k_W^D e^{\delta G_f}}. \quad (5.12)$$

Both these concentrations are bounded by $e^{-\delta G_f} \leq c_{R/W} \leq 1$. Hence, an arbitrary ratio of concentrations c_R/c_W cannot be achieved. Moreover, since ratios $J^T(R)/J^T(W)$ and c_R/c_W can be set independently, an arbitrarily high ratio of net currents does not necessarily

correspond to a high ratio of concentrations. For example, if the rates obey $k_R^T \gg k_R^D, k_W^D \gg k_W^T, k_R^D \gg k_W^T$ (figure 5.3(a)) then $J^T(R)/J^T(W)$ will be large, and c_R/c_W will be close to $e^{\delta G_f}$. However, if instead $k_R^T \gg k_R^D, k_W^D \gg k_W^T, k_W^T \gg k_R^D$, then c_R/c_W will still be close to $e^{\delta G_f}$, but $J^T(R)/J^T(W)$ will be small.

Given fixed free-energy changes, the distribution can be optimised to give the highest proportion of the right product R by varying the rate constants, yielding $k_W^T \rightarrow 0$ and $k_R^D \rightarrow 0$. In this limit, R ignores D ; the steady state yield is determined by balancing the relatively rapid production and destruction reactions on the template. For R , it is as if the yield is governed by the equilibrium of a single red transition in figure 5.3(a). W , by contrast, is overwhelmingly produced and destroyed via D ; its yield is equivalent to the equilibrium of a single blue transition in figure 5.3(a). This “pseudo-equilibrium” limit gives a concentration of $c_R = 1$ for R and a concentration of $c_W = e^{-\delta G_f}$ for W . It is worth noting that although the individual products appear to equilibrate, the product ensemble remains far from equilibrium, because different product species are coupled to different equilibria.

In conclusion, even with arbitrary relative net currents of different products along different pathways, and a far-from-equilibrium distribution of products, the relative steady-state concentrations of products are bound by the differences in the equilibrium properties of individual pathways, and the ideal behaviour involves vanishing net current around the template/destroyer cycle. In the remainder of this chapter, I shall generalise this result and explore its consequences.

5.4 General bounds on the properties of product distributions

I shall now present the main mathematical results of this chapter: an upper bound on the maximal probability of a single product $p_{\max}$, and a lower bound on the entropy $H[p_i]$ of the product distribution.

5.4.1 Bounds on the concentrations of products

To prove the bounds on probability and entropy, one must first note that, for the class of linearised CRNs considered here, the steady-state concentration of any species is bound by the set of free-energy changes along the pathways that create or destroy it. Generalising the result in section 5.3, the steady state concentration is upper-bounded by the equilibrium implied by the single pathway with the most negative free-energy change of species formation, and lower-bounded by the equilibrium implied by the pathway with the most positive free-energy change of species formation. These bounds are achieved when the rates along all other pathways tend to zero.

Mathematically, one can define $\delta G_L^{Z_i} = \max_{\text{SAWs}, S} \delta G_S^{Z_i}$ and $\delta G_U^{Z_i} = \min_{\text{SAWs}, S} \delta G_S^{Z_i}$ as the maximal and minimal free energy changes along self-avoiding walks that lead to the creation of Z_i from the null state, respectively. Then

$$e^{-\delta G_L^{Z_i}} \leq c_{Z_i} \leq e^{-\delta G_U^{Z_i}}. \quad (5.13)$$

Although similar bounds are shown in [142, 144–146], I shall prove the result here for completeness and to match my specific conventions. Further, in the cases considered in this chapter, the fact that all products are connected to the null state means that the result bounds the absolute, rather than relative, concentrations.

The proof is as follows. Consider a linear connected CRN with chemical species X_i and some reactions of the form $\emptyset \rightleftharpoons X_i$. The concentration of species X_i may be written as in eqn. 5.1. The numerator and denominator in this fraction are sums over spanning trees rooted at a given node. A sum over spanning trees rooted at node X may be factored into a sum over SAWs from some arbitrary other node to node X (eqn. 2.16). Concretely, letting Y be the other, arbitrary node, and $S(Y \rightarrow X)$ be the set of SAWs from Y to X ,

$$\sum_{T \in \mathcal{T}(X)} \prod_{e \in T} k(e) = \sum_{S \in S(Y \rightarrow X)} A(S) \prod_{e \in S} k(e), \quad (5.14)$$

where $k(e)$ is the rate constant of reaction/edge e , and $A(S)$ is a factor that, crucially, is

the same for the equivalent (reversed) SAW in $S(X \rightarrow Y)$, in which all edges are reversed compared to $S(Y \rightarrow X)$. That is to say, if one now wishes to find the sum over spanning trees rooted at Y , X may be chosen as the arbitrary other state and find:

$$\sum_{T \in \mathcal{T}(Y)} \prod_{e \in T} k(e) = \sum_{S \in S(Y \rightarrow X)} A(S) \prod_{e \in S} k(\bar{e}), \quad (5.15)$$

where $\bar{e}$ is the reverse of edge e . For the linear CRNs, utilising eq 5.6, the concentration may thus be written:

$$\begin{aligned} c_X &= \frac{\sum_{S \in S(\emptyset \rightarrow X)} A(S) \prod_{e \in S} k(e)}{\sum_{S \in S(\emptyset \rightarrow X)} A(S) \prod_{e \in S} k(\bar{e})} \\ &= \frac{\sum_{S \in S(\emptyset \rightarrow X)} A(S) \left[\prod_{e \in S} k(\bar{e}) \right] e^{-\delta G_S}}{\sum_{S \in S(\emptyset \rightarrow X)} A(S) \prod_{e \in S} k(\bar{e})}. \end{aligned} \quad (5.16)$$

Hence,

$$c_X \in \left[e^{-\max_{S \in S(\emptyset \rightarrow X)} (\delta G_S)}, e^{-\min_{S \in S(\emptyset \rightarrow X)} (-\delta G_S)} \right], \quad (5.17)$$

proving the result as required.

5.4.2 Bounds on the maximum probability and minimum entropy of the product distributions

The systems considered here have the particular property that the concentrations of all product species have the same upper and lower bounds ($e^{-\delta G_U}, e^{-\delta G_L}$). Thus the properties of the distribution defined by eqn. 5.7 are constrained. Specifically, the maximal probability of a single product has a non-trivial upper bound, while the entropy has a non-zero lower bound. I now calculate the optimal distribution for maximising one desired product and the distribution for minimising total product distribution entropy, showing that they are different.

When maximising the probability of a single product species, hereafter referred to as *specificity maximisation*, the probability distribution is clearly one product with probability $p_{\max}$, and all others with p_{low} , with

$$\begin{aligned} p_{\max} &= (1 + (M - 1)e^{-\Delta G})^{-1}, \\ p_{\text{low}} &= e^{-\Delta G} (1 + (M - 1)e^{-\Delta G})^{-1}, \end{aligned} \quad (5.18)$$

where $\Delta G = \delta G_L - \delta G_U$ and M is the number of possible product species. Note that in this chapter I use δG to describe free-energy changes along pathways, whereas I use ΔG to describe the maximum difference between these path free-energy changes.

Alternatively, one can consider *entropy minimisation*, or finding the minimal entropy of a distribution under the constraint that all species are restricted by $e^{-\delta G_L} \leq c_i \leq e^{-\delta G_U}$. Under this constraint, the entropy is bounded from below such that $H[p_i] \geq H_{\min}$, with

$$H_{\min} = \frac{(M - m)\Delta G e^{-\Delta G}}{m + (M - m)e^{-\Delta G}} + \ln(m + (M - m)e^{-\Delta G}). \quad (5.19)$$

A proof of this statement shall be given shortly. $H_{\min}$ arises from a distribution in which m species are at the upper bound concentration, $\exp(-\delta G_U)$, and $M - m$ species are at the lower bound, $\exp(-\delta G_L)$. The m that minimises entropy is given by either:

$$\left\lceil \frac{e^{-\Delta G} [\Delta G - (1 - e^{-\Delta G})]}{(1 - e^{-\Delta G})^2} M \right\rceil \quad \text{or} \quad \left\lceil \frac{e^{-\Delta G} [\Delta G - (1 - e^{-\Delta G})]}{(1 - e^{-\Delta G})^2} M \right\rceil - 1, \quad (5.20)$$

whichever gives the lower H . The ceiling function is used here because m is integer; one is required to check two values on either side of the minimum of the continuous function.

A reasonable approximation to the optimal value of m is given by

$$m_{\min} = \max \left(\frac{e^{-\Delta G} [\Delta G - (1 - e^{-\Delta G})]}{(1 - e^{-\Delta G})^2} M, 1 \right), \quad (5.21)$$

since for $m_{\min} > 1$, the difference between using the integer value of $m_{\min}$ (eqn. 5.20)

and the continuous value (eqn. 5.21) is small. In practice, using equation 5.21 produces a very slightly looser (yet still accurate) bound on $H[p_i]$. Further, using eqn. 5.21, for $m_{\min} > 1$, the minimum entropy may be simplified to:

$$H_{\min} = \ln M - \Delta G \left(1 + \frac{e^{-\Delta G}}{1 - e^{-\Delta G}} \right) + \ln \left(\frac{\Delta G}{1 + e^{-\Delta G}} \right) + 1. \quad (5.22)$$

Note that in the context of copolymerisation, $\ln M$ is proportional to the length of polymer necessary to produce M different possible product sequences.

I shall prove the entropy bound here. There is a set of M concentrations $\{c_1, \dots, c_M\}$. Denote the total concentration $c_T = \sum_{i=1}^M c_i$. From these concentrations one can define a distribution $\{p_1, \dots, p_M\}$ where $p_i = \frac{c_i}{c_T}$. The Shannon entropy of this distribution is

$$H[p_i] = - \sum_{i=1}^M p_i \ln p_i = - \frac{1}{c_T} \sum_{i=1}^M c_i \ln c_i + \ln c_T. \quad (5.23)$$

Suppose that each concentration is bounded by the same upper and lower bounds, $c_i \in [c_L, c_U]$. I now propose that the distribution of concentrations that minimises the entropy, $H[p_i]$, is that with $m_{\min}$ of the species at concentration c_U and $M - m_{\min}$ at concentration c_L , where $m_{\min}$ is given by eqn. 5.20.

To prove this claim, calculate the derivative of $H = H[p_i]$ with respect to a concentration c_α , holding all other concentrations fixed and remembering that c_T is linear in c_α ,

$$\frac{\partial H}{\partial c_\alpha} = \frac{1}{c_T} \left(- \ln \left(\frac{c_\alpha}{c_T} \right) - H \right). \quad (5.24)$$

Additionally, the second derivative,

$$\frac{\partial^2 H}{\partial c_\alpha^2} = - \frac{1}{c_T^2} \left(- \ln \left(\frac{c_\alpha}{c_T} \right) - H \right) - \frac{1}{c_T} \frac{\partial H}{\partial c_\alpha} - \frac{1}{c_T} \frac{1}{c_\alpha} \left(1 - \frac{c_\alpha}{c_T} \right), \quad (5.25)$$

is required. Evaluating the second derivative at $-\ln \left(\frac{c_\alpha}{c_T} \right) = H$, the first two terms are

zero and the third is necessarily negative for non-zero c_L . Thus

$$\left. \frac{\partial^2 H}{\partial c_\alpha^2} \right|_{-\ln\left(\frac{c_\alpha}{c_T}\right)=H} < 0. \quad (5.26)$$

And so, for any distribution, the H can be decreased by increasing the concentrations of any species i for which $-\ln\left(\frac{c_i}{c_T}\right) < H$ and decreasing the concentrations for any species whose concentration has $-\ln\left(\frac{c_i}{c_T}\right) > H$. For species for which $-\ln\left(\frac{c_i}{c_T}\right) = H$, changing the concentration in either direction will decrease H . Consequently, minimising the entropy of the distribution necessarily requires all species to be at one bound or the other: one needs m species at concentration c_U and $M - m$ at concentration c_L . Hence, the problem has been transformed into a one dimensional one of minimising H as a function of m . This entropy can be rewritten as

$$H(m) = -\frac{(M - m)\frac{c_L}{c_U} \ln\left(\frac{c_L}{c_U}\right)}{(M - m)\frac{c_L}{c_U} + m} + \ln\left((M - m)\frac{c_L}{c_U} + m\right). \quad (5.27)$$

Taking the derivative with respect to m and setting it to zero shows that

$$m_{\min} = \frac{\frac{c_L}{c_U} \left[-\ln\left(\frac{c_L}{c_U}\right) - \left(1 - \frac{c_L}{c_U}\right) \right]}{\left(1 - \frac{c_L}{c_U}\right)^2} M \quad (5.28)$$

gives a turning point for $H(m)$, which is clearly a minimum since $H(0) = H(M) = \ln M$ is maximal entropy. Since $H(m)$ has only a single turning point as a function of m , the integer that minimises $H(m)$ will be either the floor or ceiling of the above expression, and $H_{\min}$ is given by eqns. 5.19 and 5.20. Thus proving the required results.

5.4.3 The bounds in the presence of kinetic proofreading

Surprisingly, internal cycles in the reaction network, a requirement for kinetic proofreading motifs [92, 118, 119] (section 3.4), do not directly feature in the bounds derived since SAWs cannot contain cycles by definition. Adding a proofreading loop to a process may

affect ΔG , since it may provide a new pathway with a maximal or minimal free-energy change, but the SAWs identified would always be loop-free and correspond to a fixed free-energy change for product formation. The possibility of repeatedly undergoing a single dissipative cycle in an actual dynamic trajectory, consuming an arbitrary amount of molecular fuel, does not translate into an arbitrary δG along a pathway. However, the existence of loops within the network may still help to achieve a better product distribution than a network without the loops, if transition rates are constrained beyond the free-energy changes along SAWs being fixed.

5.5 Physical significance of the bounds

I shall split the analysis into three main parts. In the first, I discuss the difference between specificity maximisation and entropy minimisation, looking at the properties of both of these distributions in detail. In the second I compare these results to the previous arguments from Bennett on the minimal cost of systems that create product polymers via templates and destroy them via other pathways [139]. Finally, I discuss whether and how systems can, even in principle, reach the bounds derived here.

5.5.1 Entropy minimisation versus specificity maximisation

One might naively expect that the distribution of product concentrations that minimises the entropy would be the same as the distribution that optimises the probability of a single product. However, this equivalence does not hold in general, as shown in figure 5.4, where I plot the entropy of the specificity maximisation and entropy minimisation distributions for different system sizes. The concentration of each species is constrained between $e^{-\delta G_L}$ and $e^{-\delta G_U}$, with the concentration of each product constrained to be exactly $e^{-\delta G_L}$ or $e^{-\delta G_U}$ when minimising entropy or maximising specificity. The lower concentration cannot be arbitrarily close to zero, and the higher concentration cannot be arbitrarily high. Since having multiple products at the high concentration increases the total concentration of products, suppressing the relative probabilities of sampling

products at the low concentration, the probability distribution may be sharper for systems with multiple products at a higher concentration. In this subsection, I first discuss when specificity maximisation and entropy minimisation are equivalent; then proceed to analyse specificity maximisation and finally entropy minimisation.

If $m = 1$ minimises eqn. 5.19, entropy minimisation and specificity maximisation are equivalent. The lowest entropy state then contains a single product at a high concentration, and all others at a low concentration. One finds $m_{\min} = 1$ is optimal for $\Delta G > \ln M + \ln \ln M + \mathcal{O}\left(\frac{\ln \ln M}{\ln M}\right)$ (figure 5.4). For ΔG smaller than this value, the entropy of the distribution that maximises specificity differs quite drastically from that minimising the entropy. Surprisingly, the distribution for maximising specificity has an entropy consistent with an unbiased distribution ($H = \ln M$) for $\Delta G / \ln M \leq 1$ in the limit $M \rightarrow \infty$. In other words, the concentration of the desired product tends to zero relative to the total concentration of products, even though it remains larger than all other concentrations individually.

To explore this observation further, consider the maximal product probability $p_{\max}$, plotted against ΔG in figure 5.5. The probability $p_{\max}$ is a sigmoidal function centred on $\Delta G = \ln M$, with $p_{\max} = M^{-1}$ at $\Delta G = 0$ and $p_{\max} \rightarrow 1$ as $\Delta G \rightarrow \infty$. Interestingly, the sigmoidal function does not approach a step function here, but has fixed width of switching from $p_{\max} \approx 0$ to $p_{\max} \approx 1$ around $\ln M$ (figure 5.5(a)). Instead $p_{\max}$ has the exact form $p_{\max} = (1 + e^{-(\Delta G - \ln M)})^{-1}$ as $M \rightarrow \infty$. Thus, if one were to plot $p_{\max}$ as a function of $\Delta G - \ln M$, the curves for different, but large enough, M would collapse onto each other. Therefore, to achieve a $p_{\max} \approx 1 - e^{-A}$ for small e^{-A} , one needs a $\Delta G \approx \ln M + A$. If one instead scales ΔG by system size, one finds that $p_{\max}$ tends to a step function when $M \rightarrow \infty$ (figure 5.5(b)), reiterating that the relative yield of a single product is necessarily zero below $\Delta G / \ln M = 1$ as $M \rightarrow \infty$. Although the intended product can be far more probable than any single alternative, the large number of possible alternatives dominates.

To put the bound on $p_{\max}$ into context, one can consider the implications for the charging of a single tRNA with the correct amino acid. Here, there are approximately

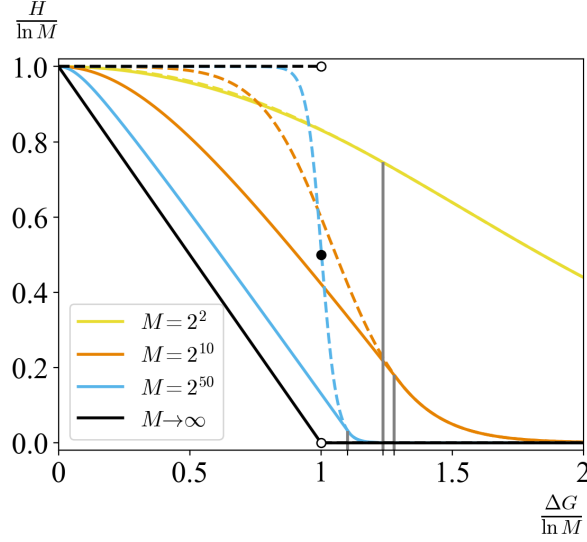
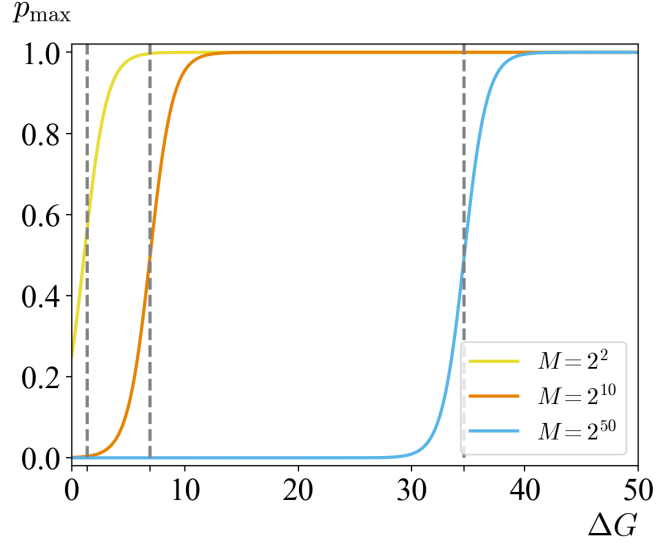


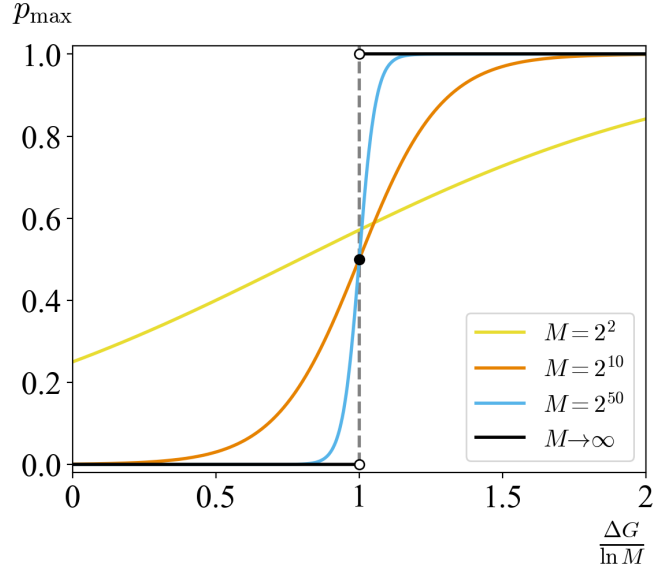
Figure 5.4: Entropy minimisation is not equivalent to specificity maximisation when $\Delta G \lesssim \ln M + \ln \ln M$. $H_{\min}$ (solid lines) and the entropy of the distribution for specificity maximisation (dotted lines) are plotted as functions of ΔG for various M . Both quantities are scaled by $\ln M$. The dashed line is discontinuous at $\Delta G = \ln M$ for $M \rightarrow \infty$. The vertical (grey) lines show $\Delta G = \ln M + \ln \ln M$. For $\Delta G \lesssim \ln M + \ln \ln M$, the minimum entropy distribution has a significantly lower entropy than the maximum specificity distribution. For $\Delta G \gtrsim \ln M + \ln \ln M$, the dotted and solid lines merge, as the minimum entropy distribution is the maximal specificity distribution. As M increases, the solid lines tend to the solid black line, albeit very slowly.

$M = 400$ different combinations of codon and amino acid, having grouped all codons that correspond to the same amino acid into one. Using $M = 400$, the bound can be calculated for the specificity with which an ensemble dominated by a single charged tRNA could be maintained in steady state. Plugging $M = 400$ into eqn. 5.18, one finds, for an error rate of 10^{-4} , *i.e.* $p_{\max} = 1 - 10^{-4}$, a free-energy difference in the pathways connecting inputs and products of approximately $(2.3A + 6)k_B T$ is needed. Thus, for an error rate of 10^{-5} , one would need a free-energy difference of around $17k_B T$ or 1 ATP [147] at $37^\circ C$, substantially higher than $kT \ln M = k_B T \ln 400 \approx 6k_B T$, the value implied by the large M limit.

For entropy minimisation, one finds that the limits $\Delta G = 0$ and $\Delta G \rightarrow \infty$ are identical to those found by specificity maximisation (figure 5.4). At $\Delta G = 0$, all pathways carry the



(a)



(b)

Figure 5.5: Specificity maximisation results in a $p_{\max}$ that is a sigmoidal function of ΔG centred at $\Delta G = \ln M$. $p_{\max}$ for a specificity-maximised system (a) as a function of ΔG and (b) as a function of scaled free energy $\Delta G / \ln M$. For (a), it can be seen that the shape of the curves quickly approaches the same functional form, but shifted by $\ln M$ (dashed lines). For (b), as M increases, $p_{\max}$ tends to a step function at $\Delta G / \ln M = 1$.

same free-energy change and so the system is a true equilibrium one and no specificity is possible, $H_{\min} = \ln M$. For $\Delta G \rightarrow \infty$, $H_{\min} \rightarrow 0$ for all M and a single product dominates.

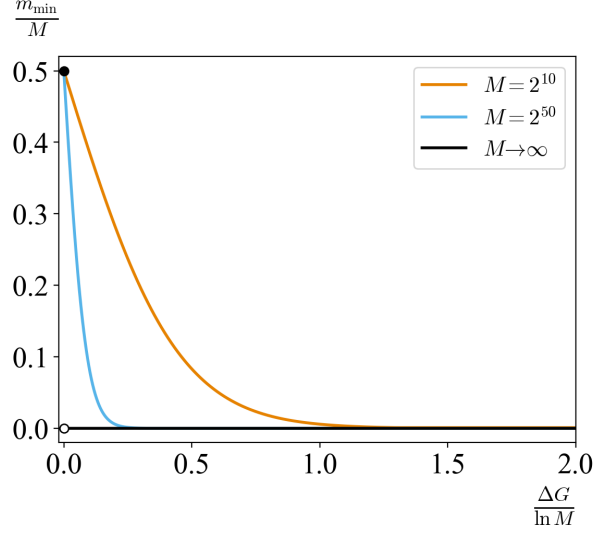
As can be seen from figure 5.4, the approach to the $M \rightarrow \infty$ limit is very slow. Even for $M = 2^{50}$, there is a significant discrepancy between $H_{\min}$ and the minimal entropy for the $M \rightarrow \infty$ limit. This difference, when normalised by $\ln M$, is approximately $\ln \ln M / \ln M$ (which is ≈ 0.1 for $M = 2^{50}$).

Now consider the properties of the entropy-minimising distribution for $m_{\min} > 1$. Here, $m_{\min}$ species are at the same high concentration, and the remaining species are at the low concentration. The fraction $m_{\min}/M$ given by eqn. 5.21 is independent of M , and monotonically decreases to 0 with increasing ΔG . Thus, if one takes the limit $\Delta G \rightarrow \infty$ at fixed $\Delta G / \ln M$, the fraction of products with high yields will tend to zero even for $\Delta G \ll \ln M$, well inside the region where specificity for a single product is impossible (figure 5.6(a)).

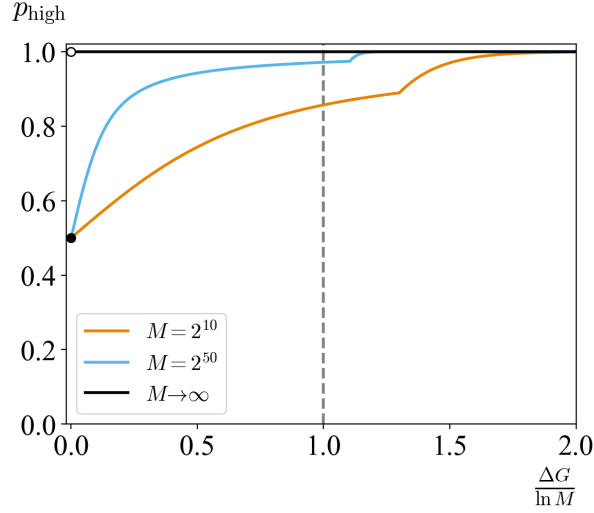
Furthermore, the probability of a product randomly chosen from this distribution being one of the $m_{\min}$ with a high concentration is given by $p_{\text{high}} = 1 - \frac{1}{\Delta G} + \frac{1}{e^{\Delta G} - 1}$. This quantity is also independent of M and increases monotonically with ΔG to 1 (figure 5.6(b)). Thus, as $M \rightarrow \infty$ at fixed $\Delta G / \ln M$, a vanishingly small proportion of the total number of products can contain an overwhelming majority of the probability, even for $\Delta G \ll \ln M$ and $m_{\min} > 1$. The black lines in figure 5.6 illustrate this surprising phenomenon in the limit of $M \rightarrow \infty$.

Turning now to $\Delta G / \ln M > 1$, consider how $H_{\min}$ behaves for large ΔG and large, but finite, M . Below $\Delta G \approx \ln M + \ln \ln M$, from eqn. 5.22, $H_{\min} \approx \ln M - \Delta G + \ln \Delta G + 1$, which is approximately linear in ΔG (figure 5.7(a)). Above $\Delta G \approx \ln M + \ln \ln M$, $H_{\min}$ decreases exponentially as $H_{\min} \approx \exp(-(\Delta G - \ln M - \ln \ln M))$. Further, at $\Delta G \approx \ln M + \ln \ln M$, $H_{\min} \approx 1$. Since the maximum value of H is $\ln M$, this value for ΔG gives the approximate scale for $H_{\min}$ to be considered small.

In figure 5.7(b), I show how the limits on the un-normalised entropy $H_{\min}$ vary with $\Delta G / \ln M$. In typical models of templating, in which each site has the same error probabil-



(a)



(b)

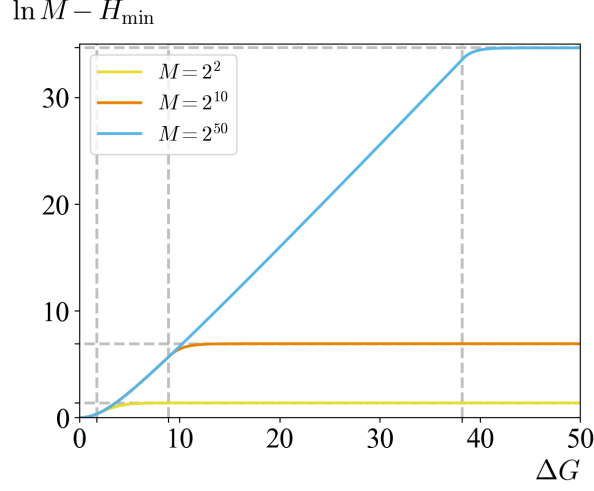
Figure 5.6: The fraction $m_{\min}/M$ of species at the high concentration bound in the entropy-minimising distribution tends to zero with scaled free energy $\Delta G/\ln M$, while the probability of picking one of these high concentration species from an ensemble, p_{high} , tends to 1. (a) The fraction of products at high concentration in the entropy-minimising distribution, approximated using eqn. 5.21, against $\Delta G/\ln M$ for two values of M and the limit as $M \rightarrow \infty$. (b) The total probability of products at the high concentration, approximated using eqn. 5.21. Notice how, even below $\Delta G/\ln M = 1$, p_{high} can be arbitrarily close to 1 and $m_{\min}/M$ arbitrarily close to 0 for large enough M . The cusp in (b) occurs when, in this approximation (eqn. 5.21), $m_{\min} = 1$ is reached. Dropping the continuous approximation for $m_{\min} > 1$ would result in more cusps, although the difference between the continuous and discrete curves is small for large M .

ity, the entropy of the distribution scales with the number of monomers in the product, $\sim \ln M$. Were the cost to reduce entropy fixed, one would expect $H_{\min}$ to be constant when M is varied at fixed $\Delta G/\ln M$. In fact, for sufficiently large $\Delta G \gtrsim 1.3 \ln M$, increasing M leads to a nominal decrease in $H_{\min}$ at fixed $\Delta G/\ln M$. Similarly for specificity maximisation (figure 5.5(a)), for sufficiently large $\Delta G \gtrsim 1.1 \ln M$, $p_{\max}$ increases with increasing M . Longer polymers therefore have much more scope within the bounds derived to use a given ΔG per monomer to suppress the absolute entropy of the resultant product distribution.

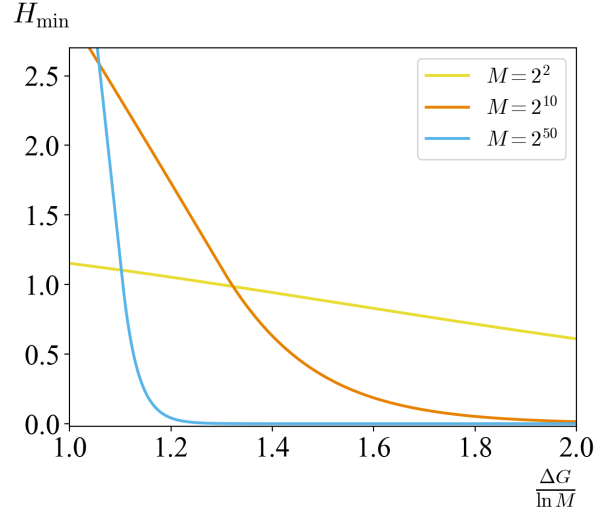
5.6 Relationship of the bounds to previously postulated costs

In his seminal paper [139], Bennett briefly discussed the production of a single product RNA copolymer using a specific template, and its subsequent destruction via a non-specific destructive enzyme. Without considering a specific model, he stated that since the enzyme is non-specific, one requires four times the tri-phosphate concentration driving the non-specific destructive enzyme than one would if it was a sequence-specific destructive enzyme. This difference would prevent the non-specific destructive enzyme from creating non-target, random sequences, instead encouraging it to indiscriminately destroy everything, including the target sequence. He therefore states that a cycle of create via template followed by destruction via destructive catalyst must use at least $k_B T \ln 4$ per nucleotide of free energy, even in steady state, corresponding to $\Delta G/\ln M > 1$ in the formalism of this chapter.

The bounds presented in this chapter update this understanding in several ways. Firstly, I point out that although two pathways with $\Delta G/\ln M > 1$ are required to generate a product distribution consisting of a single sequence in the steady state, the optimal system in terms of maintaining a specific product pool would not typically push polymers around a cycle, producing ΔG of entropy each time a polymer was created then destroyed. Instead, although the distribution over products as a whole is far from equilib-



(a)



(b)

Figure 5.7: $H_{\min}$ decreases roughly linearly with ΔG until $\Delta G \approx \ln M + \ln \ln M$, after which it decreases exponentially. I plot (a) the entropy drop $\ln M - H_{\min}$ against the free energy difference ΔG for various M , and (b) the entropy $H_{\min}$ as a function of the free-energy difference scaled by system size $\Delta G/\ln M$. In (a), for large enough ΔG , $\ln M - H_{\min} \approx \Delta G - \ln \Delta G - 1$ (eqn. 5.22), which is roughly linear and independent of M , until $\Delta G \approx \ln M + \ln \ln M$ (vertical lines) whereupon the entropy drop ceases to be linear and saturates exponentially to $H_{\min} = 0$ (horizontal lines). $\Delta G \approx \ln M + \ln \ln M$ also correspond to the switch from $m_{\min} > 1$ to $m_{\min} = 1$ (eqn. 5.21). In (b), for $\Delta G/\ln M \gtrsim 1.3$ the un-normalised entropy $H_{\min}$ is smaller for larger M at fixed $\Delta G/\ln M$.

rium, each product is effectively in its own kinetically-selected pseudo-equilibrium state. Each product is produced and destroyed by inverse pathways coupled to the same type of catalyst.

In a system in which each product was truly only coupled to a single pathway, any fuel consumption in the forwards direction would necessarily be compensated by fuel regeneration for the inverse process, which would occur at the same rate in steady state. Such a system would therefore have zero entropy production per creation event in the steady state. To see this result formally, assume that, by taking certain reaction rates to zero, a templating network is able to approach a limit in which only one pathway has forwards and backwards transition rates that are all non-vanishing for each product. Then the number of creation and destruction events per unit time will remain finite while the number of completed cycles involving distinct pathways for creation and destruction per unit time will tend to zero. Since entropy production rate scales linearly with cycle flux [148], these systems will tend towards a limit with zero entropy production rate and zero entropy production per product creation event.

Secondly, as shown in figures 5.4 and 5.5, while $\Delta G/\ln M > 1$ does serve as a necessary condition for perfect specificity in the limit of infinite polymer length, the approach to this limit is slow. A finite size correction on the order of $\ln L/L$ (where $L \propto \ln M$ would be the length of a polymer product) provides the minimum free-energy difference above $\Delta G = \ln M$ for high specificity to be achievable.

Thirdly, the bounds are fundamentally a property of distributions of products and monomers, rather than the specificity of templates [114]. Although the kinetic selectivity of the template is important in determining how close a system can get to the bound (see section 5.8), it doesn't set the bound. This observation resolves the paradox that there are thermodynamic constraints on the templated ensemble, even though there are no thermodynamic constraints on catalytic specificity.

That the bounds are unrelated to catalytic specificity is particularly emphasised by considering the region $\Delta G/\ln M < 1$. In this region, for large M , a fraction of the possible products $m_{\min}/M$ dominates the ensemble. In principle, such an ensemble could follow

from $m_{\min}/M$ distinct templates each catalysing formation of a single product sequence with high accuracy, even for $\Delta G/\ln M < 1$. Intriguingly, for $M \rightarrow \infty$, a vanishingly small fraction $m_{\min}/M$ of the possible products can constitute the full ensemble, dominating any other products, even for $\Delta G \ll \ln M$.

Fourthly, the limits on entropy and specificity with ΔG hold not only in the case of a non-specific destructive catalyst but also for a specific one, or indeed any complex set of pathways between monomers and products.

Finally, although this chapter mainly focuses on the issue of creating a steady state distribution of high specificity, I note by simple example that away from steady state, no free-energy differences are required to achieve arbitrary specificity. Consider a simple system with two products as follows:



This CRN is a simplified version of the system considered in section 5.3, with only a single catalyst and thus $\Delta G = 0$. The steady state bound on this distribution would give an unbiased equilibrium of $p_R = p_W = 1/2$. However, the concentrations of R and W can be very different at short times. Starting with initial conditions $c_R(0) = c_W(0) = 0$,

$$\lim_{t \rightarrow 0} p_R(t) = \lim_{t \rightarrow 0} \frac{c_R(t)}{c_R(t) + c_W(t)} = \frac{k_R}{k_R + k_W}, \quad (5.30)$$

which is only bounded by 0,1. Thus, the entropy can temporarily be below the steady-state bound and one can achieve arbitrarily high specificity for any set of free-energy differences between pathways.

5.7 The achievability of the bounds

The bounds I have derived are not always achievable, even if a system is allowed to have arbitrary reaction rates consistent with the overall δG of each path. Whether or not the bounds are achievable depends on the topology of the CRN. To achieve the bounds, it is

essential that it is possible to make the pathways corresponding to δG_U for m high-yield products and δG_L for $M - m$ low-yield products fast, while keeping all other pathways slow. Given that pathways to a product will typically share edges with pathways to other products, this independent scaling is not always possible.

A pleasing aspect of the bounds in this chapter is that, although finding full steady-state concentration requires knowledge of all of the SAWs from $\emptyset$ to the product, identifying the bounds only requires knowledge of the pathways with the highest and lowest free-energy-change. These free-energy changes can often be intuited or proven without knowing all of the pathways. Further, it is then often possible to check whether or not the set of pathways identified is indeed compatible with reaching the bound.

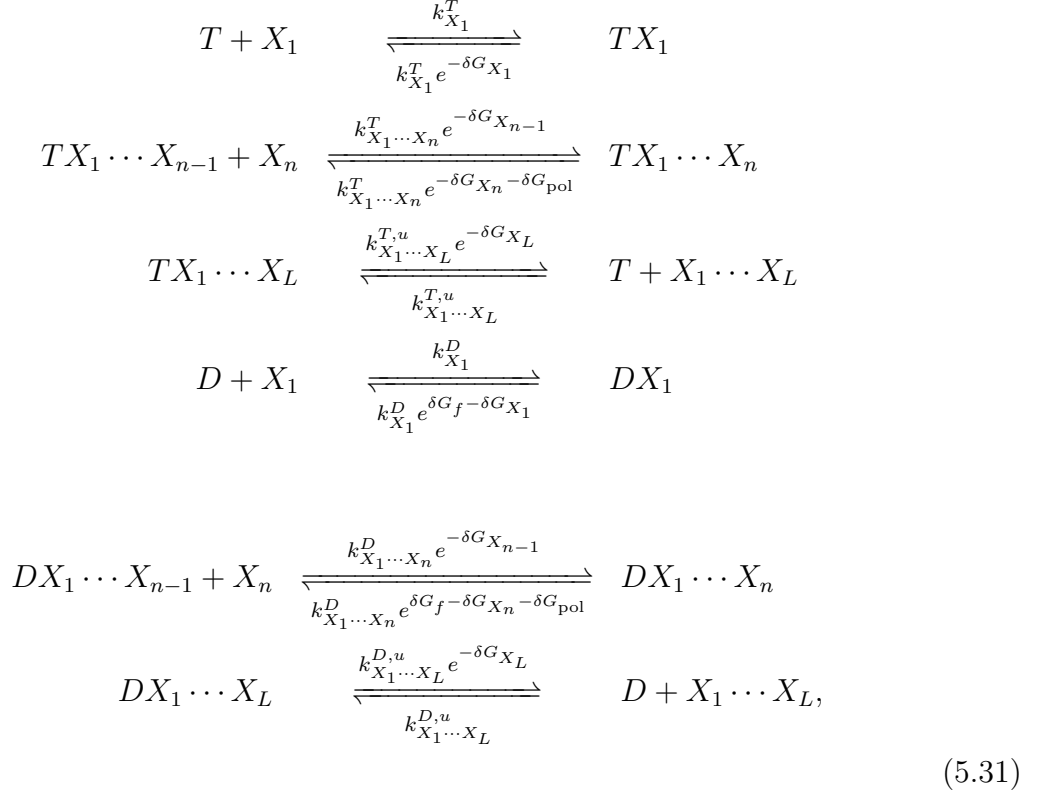
5.8 Examples

Having explored the properties of the bound itself, I shall consider two example systems to see how/whether the bound is reached as the rates along different edges are extremised.

5.8.1 Example 1: Templated polymerisation with a destructive catalyst

In chapter 4, I studied a system in which a polymer is grown on a template. Monomers of either the “right” or the “wrong” type are added one at a time to a polymer in contact with a template, while the product polymer continually unbinds from the template from behind its leading edge. I shall now extend the system to include final dissociation from the template for complete polymers, and also a “destructive template”. This destructive template participates in identical reactions to the template, except that polymerisation is driven backwards due to consumption of fuel molecules, meaning products tend to be destroyed rather than grown.

A full CRN for this model is as follows:



for $n \leq L$, $X_i \in \{R, W\}$. Dynamics is assumed to follow mass action kinetics, with rate constants given above and below the harpoons. T represents the template, D the destructive catalyst and R, W the “right” and “wrong” monomers. The products are $X_1 \cdots X_L$, representing the polymers of length L . Here, there are $M = 2^L$ different products. $TX_1 \cdots X_n$ and $DX_1 \cdots X_n$ represent partial polymers, $X_1 \cdots X_n$, bound to the template or destructive catalyst. Monomer X binds to the template or destructive catalyst with standard free-energy $-\delta G_X$ and the standard free-energy of polymerisation is $-\delta G_{\text{pol}}$ in the absence of fuel. The destructive catalyst has an additional free energy δG_f per length driving the disassembly of polymers.

I illustrate the linearised network (assuming monomers and catalysts are coupled to chemostats) for the case of dimerisation in figure 5.8. The thermodynamics of the model are characterised by the standard polymerisation free energy of the monomers ($-\delta G_{\text{pol}}$);

the standard free-energy change of binding to the template for right and wrong monomers ($-\delta G_R$ and $-\delta G_W$), and the free energy of fuel turnover $-\delta G_f$. Assume that both right and wrong monomers are held at concentration c .

For the model depicted in figure 5.8, one can calculate the free-energy change for different paths to each product state. For example, the paths from the null state to RR are:

$$\begin{aligned}
1) \quad \emptyset &\rightarrow TR \rightarrow TRR \rightarrow RR, \\
2) \quad \emptyset &\rightarrow TR \rightarrow TRW \rightarrow RW \rightarrow DRW \\
&\quad \rightarrow DR \rightarrow DRR \rightarrow RR, \\
3) \quad \emptyset &\rightarrow DR \rightarrow DRW \rightarrow RW \rightarrow TRW \\
&\quad \rightarrow TR \rightarrow TRR \rightarrow RR, \\
4) \quad \emptyset &\rightarrow DR \rightarrow DRR \rightarrow RR.
\end{aligned} \tag{5.32}$$

These incur free-energy changes:

$$\begin{aligned}
1) \quad &-(\delta G_{\text{pol}} + 2 \ln c), \\
2) \quad &-(\delta G_{\text{pol}} + 2 \ln c), \\
3) \quad &-(\delta G_{\text{pol}} + 2 \ln c) + \delta G_f, \\
4) \quad &-(\delta G_{\text{pol}} + 2 \ln c) + 2\delta G_f.
\end{aligned} \tag{5.33}$$

Each pathway contains the terms $-(\delta G_{\text{pol}} + 2 \ln c)$, corresponding to the standard free-energy change of product formation without any fuel turnover. Since the bounds only rely on the differences in free energies between pathways, one may drop these contributions, and the fuel free energy δG_f alone determines the bounds. For polymers of length L , the equivalent standard free energy of product formation is $(L - 1)\delta G_{\text{pol}} + L \ln c$, and it too may be dropped for consideration of the bounds.

The bounds are achieved when rates of the path with the most negative free-energy change are maximised for the desired product(s), and rates of the path with the most positive free-energy change are maximised for all other products. For the $L = 2$ case, for

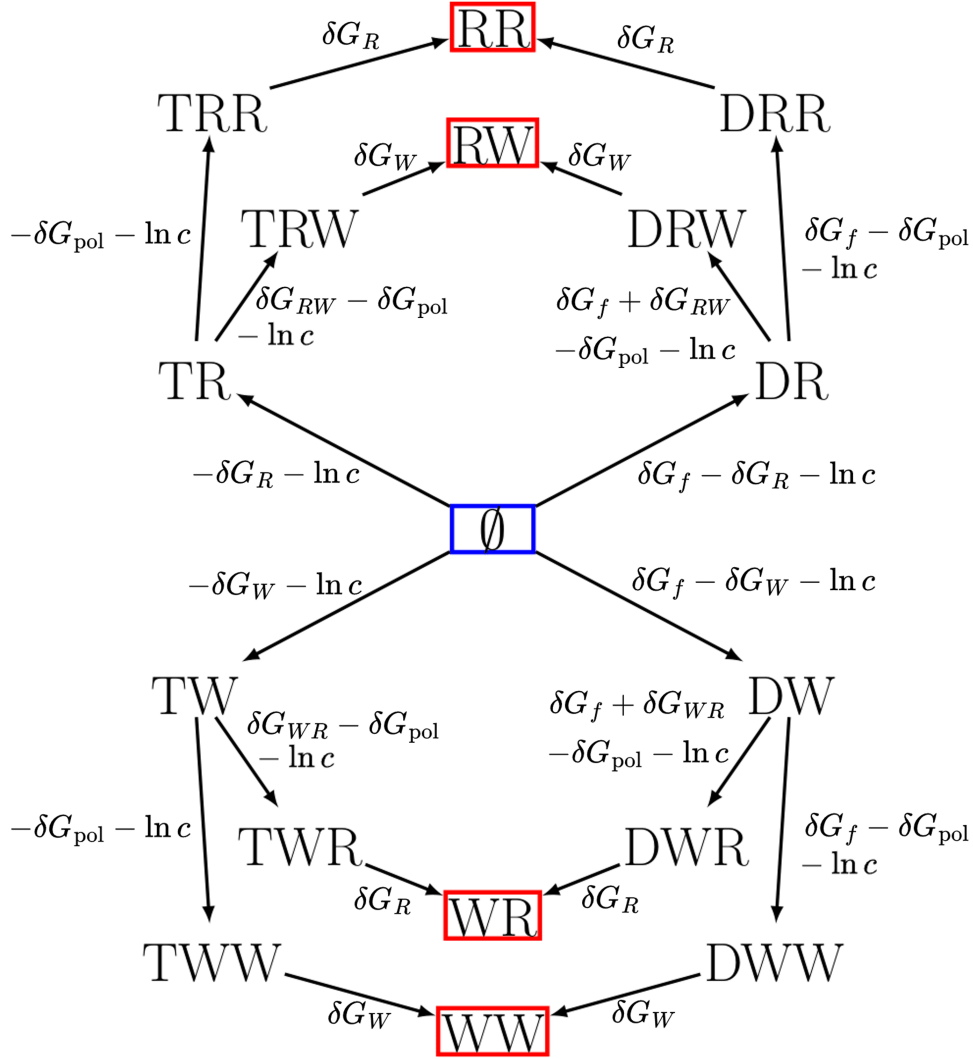


Figure 5.8: Linearised CRN for a system in which dimers are grown/destroyed via a template “ T ” and by a destructive template “ D ”. The system starts in the null state (blue) and a right or wrong (“ R ” and “ W ”) monomer attach to either the template “ T ” or destructive enzyme “ D ”. A second monomer can attach and polymerise with the first, yielding a dimer that then detaches, giving the four red output states. Each arrow represents a reversible reaction, with the free-energy change in direction of the arrow indicated. For brevity, $\delta G_{RW} = -\delta G_{WR} = \delta G_R - \delta G_W$. To reach each red product node, there are four self-avoiding walks from the blue null state.

example, one could therefore maximise the rate of path 1 for RR , and the equivalents of path 4 for other products.

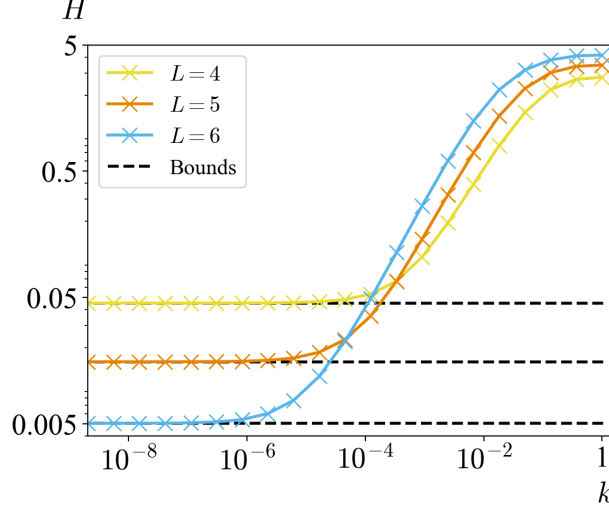


Figure 5.9: The entropy bound may be saturated by model systems in the limit that some reaction rates are much smaller than others. I plot the entropy, H , of the product distribution as a function of the slow reaction rates, k , for different template lengths, L , for the simple production and destruction model introduced in section 5.8.1. The data is obtained for a fixed $\delta G_f = 2$, $\delta G_{\text{pol}} = 0$, $\delta G_R = 2$, $\delta G_W = -2$ and $c = 1$. Note that longer templates reach a lower entropy bound for a fixed fuel turnover per unit length, δG_f .

In fact, for this system, the entropy bound is formally achievable for arbitrary L . The edges of the graph can be split into two sets; one in which the rates are ~ 1 , and one in which the rates are $\sim k$. In figure 5.9, I show $H[p_i]$ as $k \rightarrow 0$ for a particular choice of these sets, in which all the reactions leading directly to the fully correct sequence being created/destroyed on the template are set to be fast (not proportional to k), as are all the reactions leading directly to the creation/destruction of other sequences on the destroyer. More concretely, I set the rates of $\emptyset \rightarrow TR$, $TR \rightarrow TRR$, ..., $TR^{L-1} \rightarrow TR^L$, $TR^L \rightarrow R^L$ equal to a non-zero constant $k_J = 1$ (here, R^L corresponds to L copies of R). I also set all the rates of $\emptyset \rightarrow DX_1$, $DX_1 \rightarrow DX_1X_2$, ..., $DX_1...X_{L-1} \rightarrow DX_1...X_L$, $DX_1...X_L \rightarrow X_1...X_L$, where $X_i = R$ or W but excluding $X_1 \cdots X_L$ all being R , equal to 1. Conversely, I set the rates of $\emptyset \rightarrow DR$, $DR \rightarrow DRR$, ..., $DR^{L-1} \rightarrow DR^L$, $DR^L \rightarrow R^L$

and $\emptyset \rightarrow TX_1$, $TX_1 \rightarrow TX_1X_2$, ..., $TX_1\dots X_{L-1} \rightarrow TX_1\dots X_L$, $TX_1\dots X_L \rightarrow X_1\dots X_L$, where $X_i = R$ or W but excluding $X_1 \dots X_L$ all being R , equal to $k_I = k$. The reverse reactions of those listed above have a rate determined by the free-energy change of reaction. The system saturates the entropy bound found in section 5.4 as $k \rightarrow 0$. Note that for the value of δG_f used, the minimal entropy and maximal specificity distributions are the same. Further, one can see from figure 5.9, that, for this choice of rates, the bound is very tight for $k \lesssim 10^{-6}$.

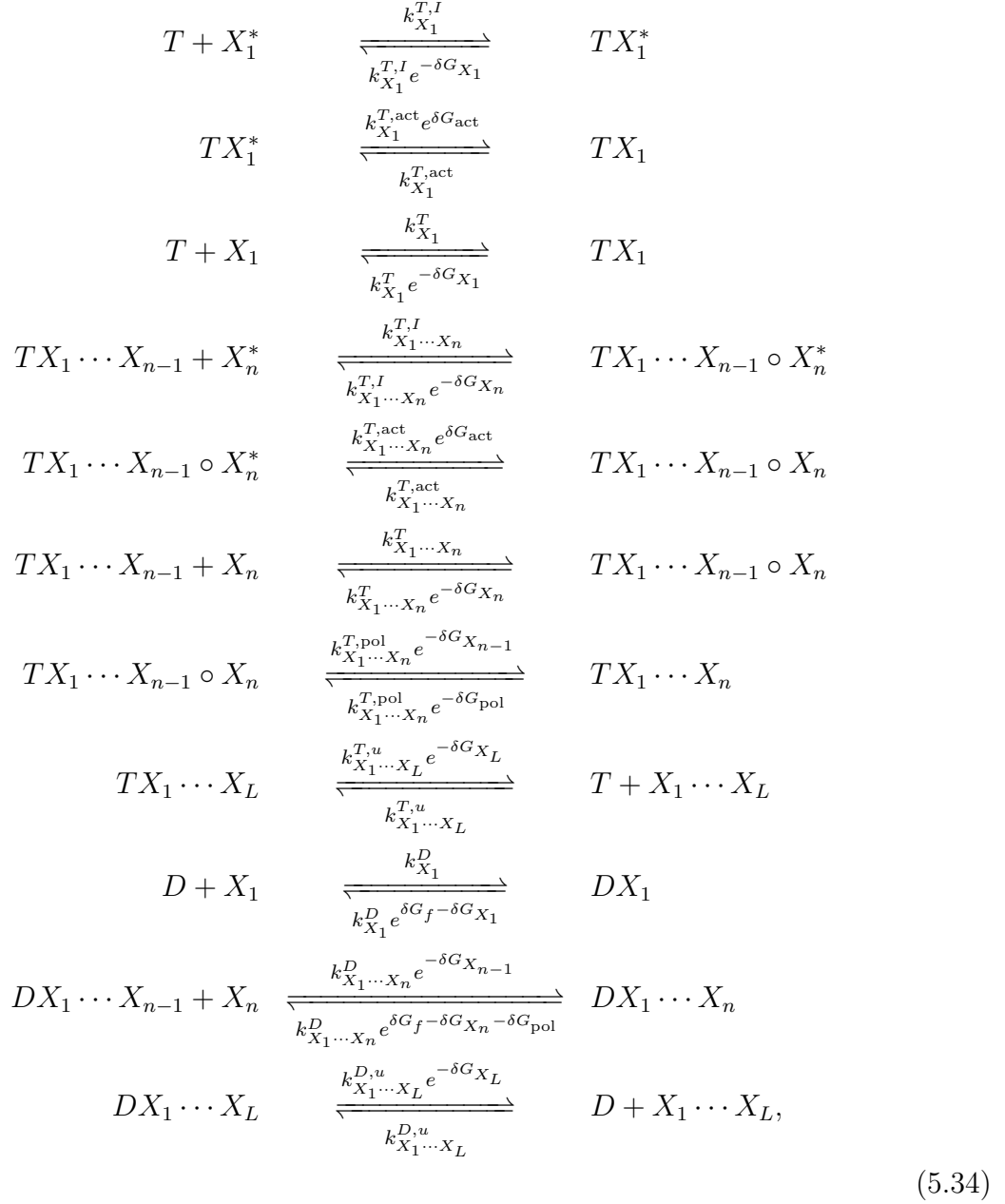
There are many possible ways to choose sets of edges that can saturate the bound in the limit $k \rightarrow 0$. Here, I have chosen a set of rates that specifically highlights a full pathway to each product for illustrative purposes. One might also wish to choose a minimum set of reactions to have rate constant k while still saturating the bound in the limit $k \rightarrow 0$. For example, letting the slow reactions be $DR^L \rightarrow R^L$, where R^L means L copies of R , and $TX_1\dots X_L \rightarrow X_1\dots X_L$, excluding $X_1 \dots X_L$ all being R , will still saturate the bound in the limit $k \rightarrow 0$. Further, note that it is possible to saturate the bound with a non-specific destructive catalyst, where the reaction rates are independent of the polymer sequence. In the examples I have identified, such a network requires at least three rate scales *e.g.* $\sim 1, k, k^2$.

It is worth stressing here that although the bound is formally attainable in this system, doing so relies on the ability to manipulate rate constants arbitrarily, subject to thermodynamic constraints. In a more realistic model of a templating system, constraints on relative rates may also be relevant; these constraints may stop the system reaching the bounds on accuracy or product entropy.

5.8.2 Example 2: Hopfield-like kinetic proofreading with a destructive catalyst

To illustrate the application of the bound to a more complex network, and to demonstrate the possibility of non-trivial pathways defining the bound, I consider an extension to the previous model, wherein the template also performs kinetic proofreading (section 3.4).

The full chemical reaction network for a proofreading template of arbitrary length is:



where $n \leq L$, $X_i \in \{R, W\}$, $X_i^* \in \{R^*, W^*\}$. As in the previous CRN, T represents the template, D the destructive catalyst, R the “right” monomers and W the “wrong” monomers. R^* and W^* are non-activated monomers. Dynamics is assumed to follow mass

action kinetics, with rate constants given above and below the harpoons. The species, $TX_1 \cdots X_{n-1} \circ X_n$ ($TX_1 \cdots X_{n-1} \circ X_n^*$) represent a complex of polymer of length $n - 1$ bound to the template as well as a (non-)activated monomer X_n (X_n^*) bound, but not yet polymerised into a single polymer of length n . $X_1 \cdots X_L$ are the products. Both non-activated (X^*) and activated monomers (X) bind to the template or destructive catalyst with standard free energy $-\delta G_X$. If a non-activated monomer is bound to the template, it may be activated, with a free-energy change δG_{act} . If an activated monomer is bound to the template, it may be polymerised into the growing copolymer; the standard free-energy change of polymerisation is $-\delta G_{\text{pol}}$. The destructive catalyst has an additional free-energy δG_f per length driving the disassembly of polymers.

Once again, the system can be linearised by assuming that monomers and catalysts are coupled to chemostats. In figure 5.10, I show part of the linearised CRN graph; this fragment should be inserted into figure 5.8 in place of the pathway $(\emptyset \rightarrow TR \rightarrow TRR \rightarrow RR)$, with similar modifications to all other template-based pathways to RW , WR , and WW . Proofreading adds complexity to the graph in the form of additional loops, and I now explicitly consider a polymerisation step independently from the binding to the template.

Monomers are now present in inactive (starred) and active forms, with $-\delta G_{\text{act}}$, representing the free-energy change of activation. Each non-activated monomer R^* , W^* is assumed to be chemostatted at the same concentration as each activated monomer R , W , c . Dropping this assumption would only cause a shift to δG_{act} . As a result, δG_f and δG_{act} control the concentration bounds. Assume $\delta G_f, \delta G_{\text{act}} \geq 0$.

The concentration bounds of this system, determined by the SAWs with maximal and minimal free-energy changes, are difficult to prove and depend on the values of both the aforementioned free-energy parameters, as well as whether L is odd or even. However, we can identify specific pathways that bound δG_U and δG_L . Specifically,

$$\delta G_U \geq \delta G_U^* = \left\{ \frac{\frac{L^2}{4} + \frac{L}{2}}{\frac{(L+1)^2}{4}} \right\} \delta G_{\text{act}}, \quad (5.35)$$

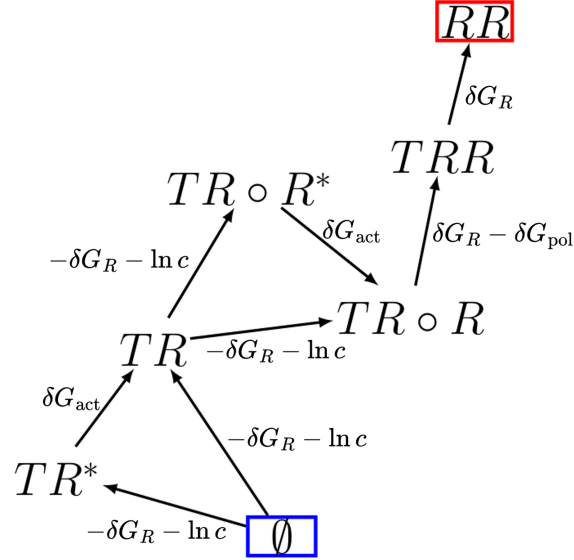


Figure 5.10: Modifying the template-based reactions of the model in section 5.8.1 to include kinetic proofreading. The template-based reactions leading directly to the dimer RR are shown here. These reactions should replace the equivalent template-based reactions in figure 5.8. As before, arrows represent reversible reactions with free-energy change in the direction of the arrow indicated. On the template, from state TR , either a non-activated (R^*) or activated (R) monomer may bind to the template. If that monomer has bound, but has not yet been polymerised into the growing polymer, it is represented by $TR \circ R^*$ or $TR \circ R$. When bound to the template, non-activated monomers may be activated, as shown by the transitions in which R^* is converted to R . When there is an activated monomer at the end of the growing polymer ($TR \circ R$), that monomer may be polymerised into the growing polymer to reach a polymerised state (TRR). After a full length polymer has grown on the template, it may detach to a product (RR for a dimer template).

and

$$\delta G_L \leq \delta G_L^* = - \left\{ \begin{array}{c} \frac{L^2}{4} \\ \frac{L^2-1}{4} \end{array} \right\} \delta G_{\text{act}} - L\delta G_f \quad (5.36)$$

where the top value in each brace is for even L and the bottom value for odd L . Unlike the simple system in section 5.8.1, the SAWs that exhibit δG_U^* and δG_L^* are not the intuitively simple pathways that go via the template and the destroyer, respectively. Instead, the extremal pathways correspond to snaking through the CRN, alternately using both the template and destructive catalyst to first create a sequence, and then convert it into another sequence. I show these pathways for $L = 4$ in figures 5.11, and 5.12. From figure 5.11, the pathway depicted picks up a free-energy change of δG_{act} for each reaction on the template (left-hand side), that is depicted with a solid line, but excluding the detachment reactions (*i.e.* reactions of the form $TX_1 \cdots X_L \rightleftharpoons X_1 \cdots X_L$). There is no contribution to the free-energy due to the reactions with the destructive enzyme (right-hand side), since gain in free energy due to reactions removing a monomer on the destructive enzyme is cancelled by an equal loss in free energy due to reactions adding a monomer on the destructive enzyme. Thus, the total free-energy change for this path is

$$(L + (L - 2) + \cdots) \delta G_{\text{act}}. \quad (5.37)$$

evaluating this sum for even and odd L gives eqn. 5.35. Similarly, from figure 5.12, there is a contribution of $-L\delta G_f$ only due to the initial reactions on the destructive enzyme ($\emptyset \rightarrow DW \rightarrow \cdots \rightarrow DWRR \cdots R$), with other reactions on the destructive enzyme cancelling each other. Following this initial contribution, the solid lines on the template, excluding the detachment reaction, each contribute $-\delta G_{\text{act}}$, leading to a total contribution of

$$-L\delta G_f - ((L - 1) + (L - 3) + \cdots) \delta G_{\text{act}}. \quad (5.38)$$

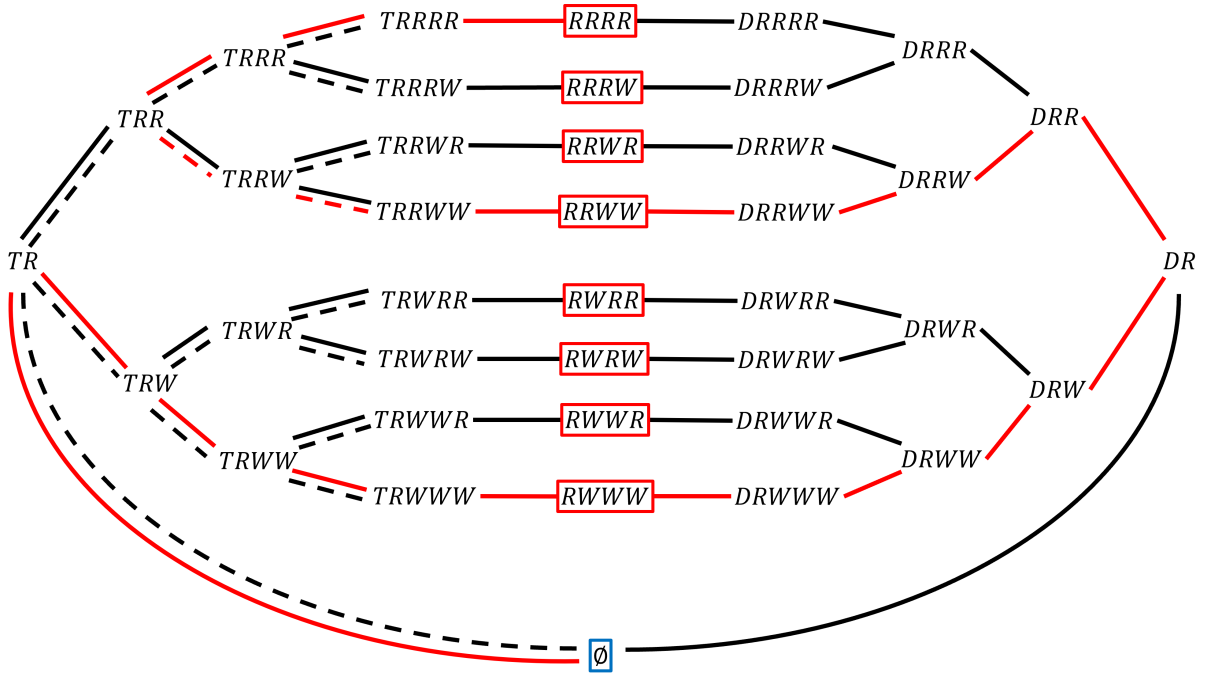
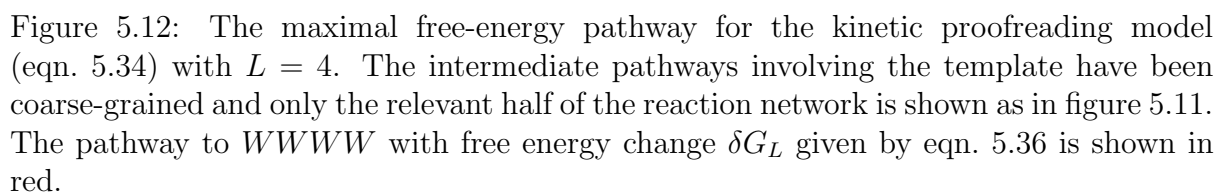


Figure 5.11: The minimal free-energy pathway for the kinetic proofreading model (eqn. 5.34) with $L = 4$. For simplicity, the intermediate states for the reactions involving templates have been coarse-grained, and these processes have been presented by using two lines for the pathways through the kinetic proofreading motifs. The solid line represents the inactivated monomer pathway ($TX_1 \cdots X_{n-1} \rightarrow TX_1 \cdots X_{n-1} \circ X_n^* \rightarrow TX_1 \cdots X_{n-1} \circ X_n \rightarrow TX_1 \cdots X_n$) and the dashed line represents the active monomer pathway ($TX_1 \cdots X_{n-1} \rightarrow TX_1 \cdots X_{n-1} \circ X_n \rightarrow TX_1 \cdots X_n$). The pathway to $RRRR$ with free energy change δG_U given by eqn. 5.35 is shown in red. For simplicity only the relevant half of the reaction network is shown.

Again, evaluating the sum for even and odd L gives eqn. 5.36.

These SAWs would not only be absurd as the dominant pathways in a real system, they actually define a bound that is unachievable, even in principle. Since the pathways exhibiting δG_U^* and δG_L^* pass through other products as intermediates, scaling these pathways to be fast would necessarily result in sub-pathways to other products being fast.

Nonetheless, the existence of these snaking pathways can provide some advantage, at least in principle. In figure 5.15, two attempts to find parameters that minimise entropy



for a system with $L = 4$ are shown. In the first “naïve” scheme (figure 5.13), plotted in blue, reactions contributing to assembly of $RRRR$ via the template or assembly of any other sequence via the destructive catalyst are assigned rates of 1 and all other rates are taken as $\sim k$. The second “best guess” scheme (figure 5.14), plotted in red, makes use of these snaking pathways. The reactions contributing to assembly of $RRRR$ via the template are still assigned rates of 1. However, for all other products, the longest snaking pathway which does not intersect with the pathway leading to $RRRR$ has rates assigned 1.

As $k \rightarrow 0$, the snaking pathways outperform the non-snaking pathways significantly, resulting in approximately half the entropy of the product distribution (figure 5.15). Notably, however, neither the best guess nor the naïve system converges on the bound as $k \rightarrow 0$; the best guess approaches $H = 1.8 \times 10^{-2}$ as compared to the bound at $H_{\min} = 1.9 \times 10^{-4}$, two orders of magnitude lower.

Although the best guess system can outperform the naïve system in principle, when absolute rates can be chosen freely, rates may in practice be mechanistically constrained. As a result, this out-performance may not be achievable in a specific model. Indeed, at moderate k , the entropies of the naïve system and the best guess converge.

5.9 Conclusion

Considering systems in which catalysts selectively produce products drawn from a large set of possibilities, I have derived general bounds on the steady-state distribution of products. The maximal difference in free-energy changes between formation paths for the products sets an upper bound on the specificity (or probability with which a single sequence can be selected from the product distribution) and a lower bound on the entropy of that product distribution. These bounds are remarkably simple, applying to arbitrarily complex networks regardless of details such as kinetic proofreading.

Several features of these bounds are, at least to me, unexpected. The bounds modify the understanding that producing sequence-specific products has a “cost” related to the

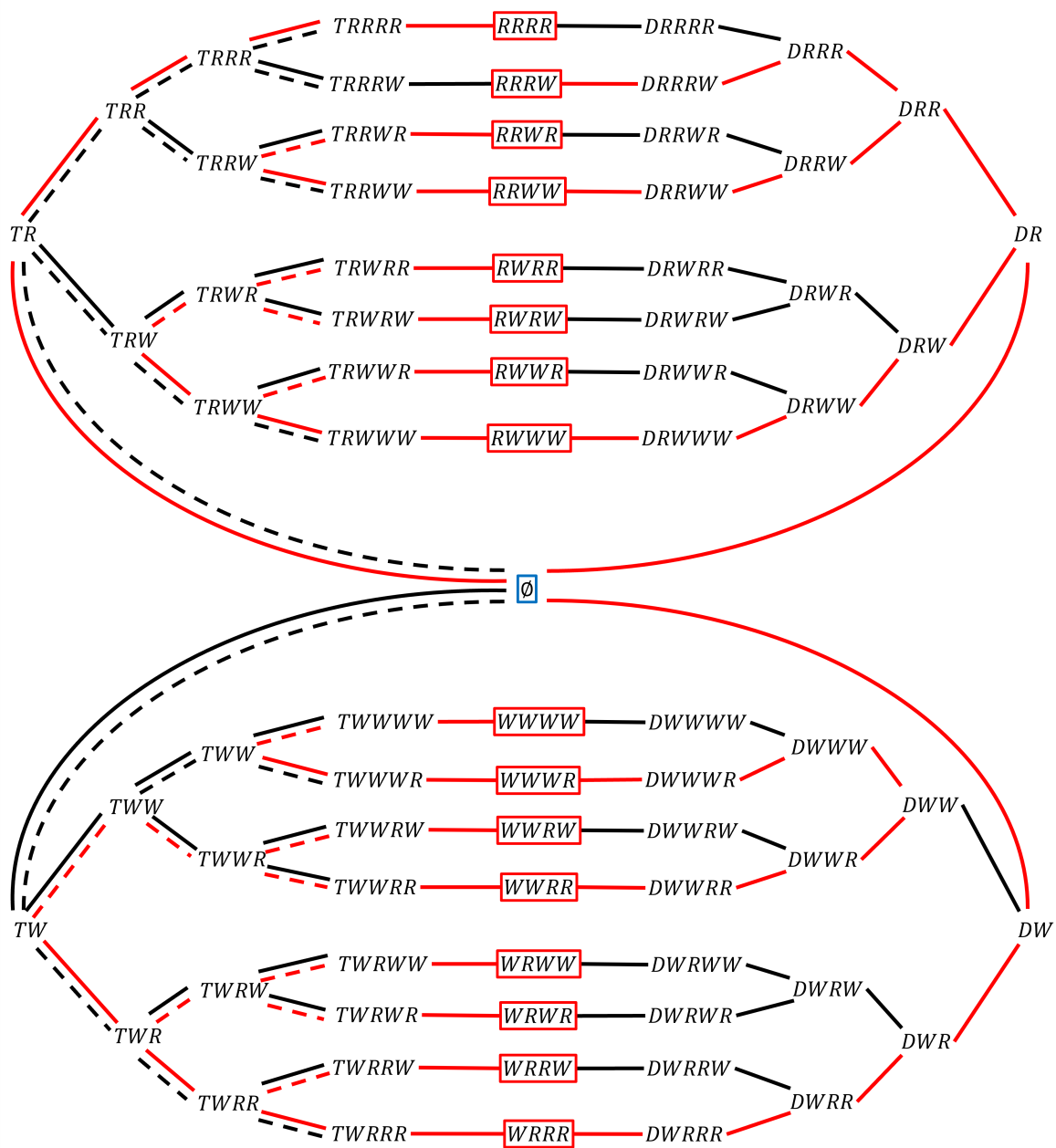


Figure 5.14: Fast reactions (red) with rate ~ 1 for a best guess at how to minimize the entropy of the product distribution.

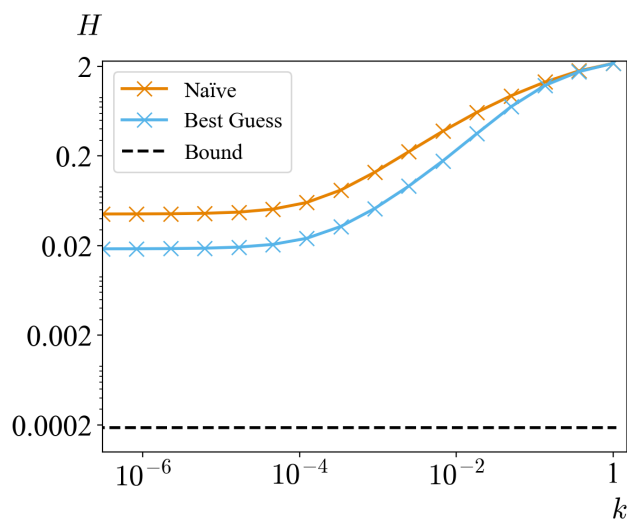


Figure 5.15: Entropy H for two attempts to optimise the entropy, plotted alongside the lower bound $H_{\min}$ implied by eqs. 5.35 and 5.36, plotted as a function of the parameter k that sets the overall scale of slow reactions relative to fast ones. In the “naïve” approach, rates favour pathways to the correct product on the template and the incorrect ones on the destroyer. The “best guess” favours the snaking pathways described in text. The data is obtained for $\delta G_f = 1$, $\delta G_{\text{act}} = 1$, $\delta G_{\text{pol}} = 0$, $\delta G_R = 2$, $\delta G_W = -2$ and $c = 1$.

information stored [40, 139, 140], and help to resolve the paradox that thermodynamics imposes limits on templating despite the fact that catalysts can be arbitrarily specific for their substrate with no minimal cost.

Most importantly, the bounds are pseudo-equilibrium in nature. In the optimal system, kinetic selectivity ensures that each product is coupled to only a single pathway, with the yield determined by the equilibrium properties of that pathway. As a result, although maintaining an ensemble with exactly one product drawn from M possibilities requires pathways differentiated by a free-energy change ΔG such that $\Delta G / \ln M > 1$, this ΔG is not a cost in the sense of entropy production. A low entropy product distribution will be far from equilibrium, with a high non-equilibrium free energy, and will consequently require more chemical work to create it in the first place than a uniform ensemble [40, 114]. But once in the dynamic steady state, a highly-specific ensemble (which is capable of dynamically responding to changing catalyst concentrations) can, in theory, be maintained with negligible thermodynamic cost; in the language of stochastic thermodynamics, there is no minimal “housekeeping” entropy production [148].

Such behaviour is profoundly different from apparently similar information-processing networks, such as a push-pull phosphorylation network [149]. In that case, accurate information transmission is only possible if the system undergoes cyclic phosphorylation by one pathway, and dephosphorylation by another, resulting in a high housekeeping entropy production. Fundamentally, the difference arises because the system considered here relies on catalytic rates distinguishing between different pathways to different products, which is not constrained by entropy production, whereas the push-pull network of [149] relies on catalysts selecting a forwards pathway over its reverse, which is so constrained [77].

Indeed, the thermodynamic limits on the product distribution that do exist do not reflect the actual selectivity of the underlying catalytic processes, which is the typical focus of work on templated processes [89, 91, 92, 99–104, 109–115, 120–123, 126, 134, 138]. The ability of a template to kinetically select certain products determines how close to the bound a system can come, but it does not determine the bound itself.

For example, although a single product cannot dominate the steady state for

$\Delta G/\ln M < 1$, this does not mean that catalysts cannot be arbitrarily specific; simply that the steady state achieved via the set of available catalysts cannot be dominated by one product. Indeed, remarkably, for $M \rightarrow \infty$ the lower bound on product entropy is achieved when a vanishingly small fraction $m_{\min}/M$ of the possible products dominate the ensemble, even for $\Delta G \ll \ln M$. Such a distribution could arise from a system with $m_{\min}$ templates each catalysing a single product with high accuracy, and a non-specific destructive catalyst.

The above results suggest that, in some sense, catalytic templating ensembles do not have an inherent minimum maintenance “cost” related to the accuracy of information propagation from a single template to its product. From another perspective, however, the bounds show that maintaining a product distribution that is highly-specific for a single product is hard relative to previous claims [40, 139, 140]. The minimal entropy distribution is only dominated by a single product for $\Delta G \gtrsim \ln M + \ln \ln M$, a condition that converges to the more familiar $\Delta G/\ln M > 1$ achingly slowly with M . Moreover, the minimal entropy of a product distribution with finite M is larger than that for $M \rightarrow \infty$ for all values of $\Delta G/\ln M$. For any finite M , perfect accuracy is only possible in the limit $(\Delta G - \ln M) \rightarrow \infty$, although good accuracy is possible for $\Delta G \gtrsim \ln M + \ln \ln M$.

This bound opens up several questions. Firstly: why do cells not operate in the manner suggested by the optimal behaviour? Although it may be practically impossible to obtain the total kinetic selectivity required for optimally efficient behaviour, cells do not even pursue a strategy that approaches this limit. RNA molecules are not destroyed by the reverse of their creation pathway, and the use of kinetic proofreading during templating means that multiple template-based pathways with different δG are relevant. I expect that practical constraints that would limit the speed and specificity of catalysts operating in this reversible regime must be a factor. Indeed, exploring realistic models, that are constrained by the actual chemistry, and by the biological context, would be highly informative. It may also be true that, in practical contexts, the “excess” entropy production due to responding to changing conditions [148] dominates the housekeeping entropy production for the steady state.

It is also worth noting that templating in the thermodynamically optimal context leads to binary outputs. The steady-state concentration of a given product is high if the relevant template is present, and low otherwise, regardless of the precise concentration of the template. Biological systems likely benefit from the functionality of analogue control, in which the product concentration varies gradually with the template concentration, away from this limit.

Synthetic information-processing systems, however, may operate quite differently. In synthetic systems, a binary output is often preferable, and products may be much simpler than long protein and RNA sequences. Said synthetic systems may therefore be engineered to operate in the low-cost fashion identified here. Such a system would be reminiscent of the behaviour of dynamic combinatorial chemistry ensembles [150], but with the system designed so that the presence/absence of specific catalysts changes the free energy of the dominant pathway to an assembled product.

Secondly, a number of simplifying assumptions have been made in this chapter. the first assumption is that the networks can be linearised. In other words, it is assumed that: (a) fuel and monomer concentrations are effectively chemostatted by the environment or cell; (b) products or partial product fragments do not interact with each other; and (c) templates and other catalysts are not depleted by the presence of products or partial products. Of these, assumption (c) is most likely to be violated. Systems may not be dilute enough to neglect sequestration of templates and other catalysts by monomers and products. It is therefore an interesting open question how the behaviour changes if the catalysts become saturated. Pathways of well-defined δG still exist, which may allow similar limits to be derived [151].

Relatedly, in this analysis the yield of partially formed products, which will be relevant if templating is not highly processive (as it is in biology), or template-bound intermediates has not been considered. If these species have downstream functionality, their presence will be relevant to the system's output, and partial products may also act to sequester templates and other catalysts, if the linear assumption breaks down.

Perhaps most significantly, however, all products are assumed to be equally thermo-

dynamically stable, and produced via equivalent pathways. In practice, although the presence of catalysts cannot change the thermodynamic stability of the products, some products will be more stable than others due to intra-product interactions or because monomers are present at different concentrations. More fundamentally, network topologies will necessarily vary for polymers of different length. As a result, δG_U and δG_L will be product-specific in general. In this case, much of the analysis would still apply directly, with individual products pushed either to their maximum ($e^{-\delta G_U}$) or minimum ($e^{-\delta G_L}$) concentrations to minimise entropy, and the optimal distribution being achieved when each product is overwhelmingly coupled to a single pathway. However, it will likely be easier to create distributions that are sharply peaked about the most stable products, and harder in other cases. Given that the aim of processes like transcription and translation is to produce functional, rather than stable, products, it is unclear whether this asymmetry would actually be beneficial.

Finally, the dynamics of these networks have not been considered in detail. I have shown that at finite times, the bounds can be violated, but have not attempted to explore for how long this is possible. Indeed, one might ask whether cells are effectively in this non-steady-state limit since some theoretically possible polymers (such as chimeric RNA/DNA sequences) are essentially never created by the cell. The question of how these networks respond to dynamical changes in the concentration of the catalysts, and the associated thermodynamic costs, is also open.

Chapter 6

Conclusion

Templating is a powerful mechanism to produce complex structures and sequences at a nanoscopic scale without the aid of a sentient observer. In particular, biological systems make use of templating for the central dogma where information, stored in DNA, is transmitted via templating to mRNAs and then proteins. Due to its ubiquity in biology, its potential for synthesis of synthetic molecules, and its interesting mathematical qualities, templating models have a rich history of being studied [92, 99–112]. Nonetheless, the detailed models hitherto present in the literature have not been able to analyse thermodynamically consistent models of templated polymerisation with complex sets of reactions leading to monomer incorporation and dependence monomer incorporation rates on previous monomers. Additionally, the class of systems, in which the templating is embedded in a complex non-equilibrium network of assembly and disassembly pathways, hasn't been studied at all in the literature. The main aims of this thesis were to fill in these holes in our understanding of templating.

After introducing the relevant biological and mathematical background in chapters 1,2, and 3, through chapter 4, I address this first deficiency by developing a framework for solving a class of polymerisation models, which includes these thermodynamically consistent, multi-step models with dependence on previous monomers. The key insights of this framework are twofold. The first is through eqn. 4.5, whereby the complex network

of reaction rates could be coarse-grained into a single pair of rates, allowing the system to be solved via previously described methods [97]. This coarse-graining preserves the statistics of the polymer produced, but does not preserve the dynamic or thermodynamic fine-grained quantities such as time or entropy production. However, the second key insight of this chapter comes in to play here. This insight was to notice that the dynamic fine-grained quantities of interest could be calculated by considering a frame shift, analysing the process as it jumps between certain reference states that I call completed states. Thus, through this analysis, the fine-grained quantities can be added back in and one can use this framework to study problems such as trade-offs between accuracy, free-energy costs and speed of polymerisation.

Two main examples demonstrate the utility of this method. The first demonstration is in section 4.4.3, whereby the graphical method allows one easily to prove how models without fuel consuming cycles behave. The second main example shows how complex polymerisation models can be studied. In particular, in section 4.4.5, I utilise the method to solve an N -loop generalisation of Hopfield’s kinetic proofreading. Figures 4.7 and 4.8 highlight key aspects of these multi-loop Hopfield networks that can now be studied with access to this new method. These features are an optimum number of loops when considering a non-zero concentration of activated monomers, and the relative inefficiency of these networks at utilising the free-energy available to them.

The framework presented in this chapter is not infallible, however. There are a few important limitations of the method (as presented) which need to be considered. One such limitation is that the method as presented only applies to long polymers, and that the sequence of the template has been lost in the mapping to right and wrong monomers. These assumptions however can be relaxed with more analysis. Indeed, a current (as of the writing of this thesis) PhD student, Jeremy Guntoro, has been working on relaxing these conditions, in an as-of-yet unpublished work. In this article, Guntoro is able to extend the method to explicitly include finite-sized polymers and break the symmetries of interactions between different monomers in the copy polymer and template polymer. This breaking of symmetries of monomer-template interactions can be seen most clearly

by example. Letting A, B be monomer types in the template and a, b be monomer types in the copy polymer, the symmetries broken are that interactions between A with a need not be the same as b with B and similarly for A with b , not being the same as B with a .

Guntoro is able to utilise eqn. 4.5 to coarse-grain an arbitrary, complex network leading to monomer incorporation and solve for the statistics of polymers produced by the coarse-grained process. Following this, Guntoro is able to adapt the philosophy of extracting the fine-grained quantities as in section 4.3. The extensions to the framework requires solving for the expected number of times the system is in a given tip state, defined now by the monomers both in the copy polymer and template at the tip of the growing copolymer. This expected number of times replaces the $1/(2P-1)$ term in, for example, eqns. 4.13 and 4.18.

Another limitation of the method presented in chapter 4 is that the dependence of reaction rates on previous monomers was limited only to the penultimate monomer in the growing copolymer. However, I believe it should be possible to drop this assumption. One should still be able to utilise the philosophy of coarse-grained and fine-grained quantities to analyse such models. In particular, one should still be able to use eqn. 4.5 to coarse-grain the network of reactions leading to monomer incorporation allowing them to utilise the methods of Gaspard and Andrieux for solving k^{th} order Markov chain polymerisation models [116], whereby the monomer incorporation rates depend on the k monomers behind the growing tip of the copolymer. Extracting the fine-grained quantities of such a model will likely be cumbersome, however, as one would need to define a tip distribution ξ over the M^{k+1} tip states, where M is the number of monomer types, and use this distribution to calculate any fine-grained quantities from the M^{k+1} possible step-wise processes.

Finally, the framework of chapter 4 relies on the models having the tree-like structure (figure 4.1) whereby any trajectory from one completed state to another is confined to a given “petal”. Breaking such an assumption would not be possible within the framework. As such, if one wishes to analyse a model, for which this assumption does not hold, they would require a different framework. I do not believe it would be possible to adapt this

framework to such a set of models.

The final comment I shall make regarding chapter 4 is that the aims of this chapter were to introduce a mathematical framework. As such only a limited number of example systems were analysed. It would be of great interest to see what other classes of models would be most enlightening to study with this framework. In particular, it would be interesting to see what physical principles and trade-offs could be derived for certain classes of models using this method.

In chapter 5, I aim to address the second deficiency in the literature by analysing catalytic templating systems with multiple, distinct pathways for polymerisation/degradation. In comparison to the method for precisely solving detailed polymerisation models in chapter 4, the aims of chapter 5 stray away from the details to ask broader questions about the limits of such systems. In this chapter, I derive bounds, eqns. 5.18 and 5.19, on the accuracy of a given polymerisation system as a function of the free-energy differences between the pathways of said polymerisation system.

The main results of this chapter, however, come from the nature of the systems which achieve these bounds. Although the bounds are set by free-energy differences between two different pathways, the systems which achieve the bounds do not utilise the cycling of creating a polymer via one pathway and destroying via another as one might initially expect. Instead, these optimally accurate systems essentially equilibrate with a single pathway for each product. As such, the optimally accurate systems have no running costs, known as housekeeping entropy production. This optimally accurate operating regime is not what is observed in biological systems, raising the question of why not. However, the optimally accurate regime might prove most useful for synthetic catalytic templating systems.

The bounds, as derived in chapter 5, apply to a relatively limited class of systems, being those which can be described by linearised chemical reaction networks. This assumption of linearised chemical reaction networks would be the most important limitation to address in subsequent analysis of these multi-pathway polymerisation systems. I believe it should indeed be possible to expand the scope of these bounds to certain other classes

of networks. For example, one might imagine networks, similar to the linearised networks except catalysts, such as templates, are not chemostatted. In such networks, there should still be well defined pathways with well defined free energies. Thus, one should still be able to bound the relative concentrations of two products by considering pathways between said products and one might be able to look to the rational parametrisation theorem [142] for more insight on how to proceed. For more and more general chemical reaction networks, however, it might not be possible to derive similar bounds. One would need to ensure that well-defined pathways, with well-defined free-energy changes can be defined for the chemical reaction networks, which is not guaranteed.

6.1 Final remarks

Templated polymerisation is an interesting problem for both the fundamental sciences and engineering. On the fundamental side, templated polymerisation is an excellent playground for the study of the intersections between computing and thermodynamics. On the engineering side, templated polymerisation promises exiting prospects for the synthesis of complex molecules. Through this thesis, I have attempted to expand our fundamental physical understanding of templated polymerisation via the introduction of a framework for analysing a large class of models, and bounds on the accuracy of another class of models. These contributions, I hope, should help both to test and develop, theoretically, some of the more general results of stochastic thermodynamics of computation as well as to aid in the design of synthetic polymerisation systems.

Bibliography

- [1] Felix Ritort. *Nonequilibrium Fluctuations in Small Systems: From Physics to Biology*, pages 31–12. John Wiley & Sons, Ltd, Hoboken, NJ, 2007.
- [2] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, and Martin Raff. *Molecular Biology of the Cell*. Garland Science, New York, NY, 2014.
- [3] Robert M Macnab. How bacteria assemble flagella. *Annual Review of Microbiology*, 57:77–100, 2003.
- [4] Kristen Johannes, Jacklyn Powers, Lisa Couper, Matt Silberglitt, and Jodi Davenport. Tangible models and haptic representations aid learning of molecular biology concepts. *Grantee Submission*, 2016.
- [5] Sarah Jane Flint, Vincent R Racaniello, Glenn F Rall, Theodora Hatziiioannou, and Anna Marie Skalka. *Principles of virology, Volume 2: pathogenesis and control*. John Wiley & Sons, Hoboken, NJ, 2020.
- [6] Robijn F Bruinsma, William M Gelbart, David Reguera, Joseph Rudnick, and Roya Zandi. Viral self-assembly as a thermodynamic process. *Physical Review Letters*, 90:248101, 2003.
- [7] James Watson and Francis Crick. Structure of small viruses. *Nature*, 177:473–5, 1956.
- [8] William B Sherman and Nadrian C Seeman. A precisely controlled dna biped walking device. *Nano Letters*, 4:1203–1207, 2004.

- [9] Jonathan Bath and Andrew J Turberfield. Dna nanomachines. *Nature Nanotechnology*, 2:275–284, 2007.
- [10] Ebbe S Andersen, Mingdong Dong, Morten M Nielsen, Kasper Jahn, Ramesh Subramani, Wael Mamdouh, Monika M Golas, Bjoern Sander, Holger Stark, Cristiano LP Oliveira, et al. Self-assembly of a nanoscale dna box with a controllable lid. *Nature*, 459:73–76, 2009.
- [11] Rahul Chhabra, Jaswinder Sharma, Yan Liu, Sherri Rinker, and Hao Yan. Dna self-assembly for nanomedicine. *Advanced Drug Delivery Reviews*, 62:617–625, 2010.
- [12] Qinqin Hu, Hua Li, Lihua Wang, Hongzhou Gu, and Chunhai Fan. Dna nanotechnology-enabled drug delivery systems. *Chemical Reviews*, 119:6459–6506, 2018.
- [13] Manas Chanda. *Introduction to polymer science and chemistry: a problem-solving approach*. CRC Press, Boca Raton, FL, 2013.
- [14] Charles G Overberger. Copolymerization: 1. general remarks; 2. selective examples of copolymerizations. *Journal of Polymer Science: Polymer Symposia*, 72:67–69, 1985.
- [15] Frank R Mayo and Frederick M Lewis. Copolymerization. i. a basis for comparing the behavior of monomers in copolymerization; the copolymerization of styrene and methyl methacrylate. *Journal of the American Chemical Society*, 66:1594–1601, 1944.
- [16] Alfred Rudin and Phillip Choi. *The Elements of Polymer Science and Engineering*. Academic press, San Diego, 1982.
- [17] Nadrian C Seeman and Hanadi F Sleiman. Dna nanotechnology. *Nature Reviews Materials*, 3:1–23, 2017.

- [18] Neville R Kallenbach, Rong-Ine Ma, and Nadrian C Seeman. An immobile nucleic acid junction constructed from oligonucleotides. *Nature*, 305:829–831, 1983.
- [19] Nadrian C Seeman. Nucleic acid junctions and lattices. *Journal of Theoretical Biology*, 99:237–247, 1982.
- [20] Tsu Ju Fu and Nadrian C Seeman. Dna double-crossover molecules. *Biochemistry*, 32:3211–3220, 1993.
- [21] Erik Winfree, Furong Liu, Lisa A Wenzler, and Nadrian C Seeman. Design and self-assembly of two-dimensional dna crystals. *Nature*, 394:539–544, 1998.
- [22] David Doty, Matthew J Patitz, and Scott M Summers. Limitations of self-assembly at temperature 1. *Theoretical Computer Science*, 412:145–158, 2011.
- [23] Paul WK Rothmund. Folding dna to create nanoscale shapes and patterns. *Nature*, 440:297–302, 2006.
- [24] Arvind Murugan, James Zou, and Michael P Brenner. Undesired usage and the robust self-assembly of heterogeneous structures. *Nature Communications*, 6:6203, 2015.
- [25] Katherine E Dunn, Frits Dannenberg, Thomas E Ouldridge, Marta Kwiatkowska, Andrew J Turberfield, and Jonathan Bath. Guiding the folding pathway of dna origami. *Nature*, 525:82–86, 2015.
- [26] David Soloveichik and Erik Winfree. Complexity of self-assembled shapes. *SIAM Journal on Computing*, 36:1544–1569, 2007.
- [27] Joakim Bohlin, Andrew J Turberfield, Ard A Louis, and Petr Sulc. Designing the self-assembly of arbitrary shapes using minimal complexity building blocks. *ACS nano*, 17:5387–5398, 2023.
- [28] Jeremy Gunter and Thomas E Ouldridge. Deterministic molecular assembly with a finite set of building blocks: Universal assembly kits for backbone-assisted and

- sequence-directed abstract tile assembly models. *arXiv preprint arXiv:2402.19225*, 2024.
- [29] Leonard M Adleman. Molecular computation of solutions to combinatorial problems. *Science*, 266:1021–1024, 1994.
 - [30] Constantine G Evans and Erik Winfree. Physical principles for dna tile self-assembly. *Chemical Society Reviews*, 46:3808–3829, 2017.
 - [31] Erik D Demaine, Martin L Demaine, Sándor P Fekete, Matthew J Patitz, Robert T Schweller, Andrew Winslow, and Damien Woods. One tile to rule them all: simulating any turing machine, tile assembly system, or tiling system with a single puzzle piece. *arXiv preprint arXiv:1212.4756*, 2012.
 - [32] Damien Woods. Intrinsic universality and the computational power of self-assembly. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373:20140214, 2015.
 - [33] Bernard Yurke, Andrew J Turberfield, Allen P Mills Jr, Friedrich C Simmel, and Jennifer L Neumann. A dna-fuelled molecular machine made of dna. *Nature*, 406:605–608, 2000.
 - [34] David Soloveichik, Georg Seelig, and Erik Winfree. Dna as a universal substrate for chemical kinetics. *Proceedings of the National Academy of Sciences*, 107:5393–5398, 2010.
 - [35] Lulu Qian, Erik Winfree, and Jehoshua Bruck. Neural network computation with dna strand displacement cascades. *Nature*, 475:368–372, 2011.
 - [36] Alexander Dack, Benjamin Qureshi, Thomas E Ouldridge, and Tomislav Plesa. Recurrent neural chemical reaction networks that approximate arbitrary dynamics. *arXiv preprint arXiv:2406.03456*, 2024.

- [37] Ruedi Aebersold, Jeffrey N Agar, I Jonathan Amster, Mark S Baker, Carolyn R Bertozzi, Emily S Boja, Catherine E Costello, Benjamin F Cravatt, Catherine Fenselau, Benjamin A Garcia, Ying Ge, Jeremy Gunawardena, Ronald C Hendrickson, Paul J Hergenrother, Christian G Huber, Alexander R Ivanov, Ole N Jensen, Michael C Jewett, Neil L Kelleher, Laura L Kiessling, Nevan J Krogan, Martin R Larsen, Joseph A Loo, Rachel R Ogorzalek Loo, Emma Lundberg, Michael J MacCoss, Parag Mallick, Vamsi K Mootha, Milan Mrksich, Tom W Muir, Steven M Patrie, James J Pesavento, Sharon J Pitteri, Henry Rodriguez, Alan Saghatelian, Wendy Sandoval, Hartmut Schlüter, Salvatore Sechi, Sarah A Slavoff, Lloyd M Smith, Michael P Snyder, Paul M Thomas, Mathias Uhlén, Jennifer E Van Eyk, Marc Vidal, David R Walt, Forest M White, Evan R Williams, Therese Wohlschlager, Vicki H Wysocki, Nathan A Yates, Nicolas L Young, and Bing Zhang. How many human proteoforms are there? *Nature Chemical Biology*, 14:206–214, 2018.
- [38] Pablo Sartori and Stanislas Leibler. Lessons from equilibrium statistical physics regarding the assembly of protein complexes. *Proceedings of the National Academy of Sciences*, 117:114–120, 2020.
- [39] Francis Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.
- [40] Thomas E Ouldridge and Pieter Rein ten Wolde. Fundamental costs in the production and destruction of persistent polymer copies. *Physical Review Letters*, 118:158103, 2017.
- [41] Jeroen De Neve, Joris J Haven, Lowie Maes, and Tanja Junkers. Sequence-definition from controlled polymerization: the next generation of materials. *Polymer Chemistry*, 9:4692–4705, 2018.
- [42] Rachel K O’Reilly, Andrew J Turberfield, and Thomas R Wilks. The evolution of dna-templated synthesis as a tool for materials discovery. *Accounts of Chemical Research*, 50:2496–2509, 2017.

- [43] Dmitry L Usanov, Alix I Chan, Juan Pablo Maianti, and David R Liu. Second-generation dna-templated macrocycle libraries for the discovery of bioactive small molecules. *Nature chemistry*, 10:704–714, 2018.
- [44] Robert B Merrifield. Solid phase peptide synthesis. i. the synthesis of a tetrapeptide. *Journal of the American Chemical Society*, 85:2149–2154, 1963.
- [45] Robert L Letsinger and V Mahadevan. Oligonucleotide synthesis on a polymer support¹, 2. *Journal of the American Chemical Society*, 87:3526–3527, 1965.
- [46] Alex Hoose, Richard Vellacott, Marko Storch, Paul S Freemont, and Maxim G Ryadnov. Dna synthesis technologies to close the gene writing gap. *Nature Reviews Chemistry*, 7:144–161, 2023.
- [47] Kevin Schiefelbein and Nina Hartrampf. Flow-based methods in chemical peptide and protein synthesis. *Chimia*, 75:480–480, 2021.
- [48] Nicole E Gregorio, Max Z Levine, and Javin P Oza. A user’s guide to cell-free protein synthesis. *Methods and Protocols*, 2:24, 2019.
- [49] Jia Niu, Ryan Hili, and David R Liu. Enzyme-free translation of dna into sequence-defined synthetic polymers structurally unrelated to nucleic acids. *Nature Chemistry*, 5:282–292, 2013.
- [50] Dehui Kong, Wayland Yeung, and Ryan Hili. Generation of synthetic copolymer libraries by combinatorial assembly on nucleic acid templates. *ACS Combinatorial Science*, 18:355–370, 2016.
- [51] Alexander E Stross, Giulia Iadevaia, Diego Núñez-Villanueva, and Christopher A Hunter. Sequence-selective formation of synthetic h-bonded duplexes. *Journal of the American Chemical Society*, 139:12655–12663, 2017.
- [52] Jean-François Lutz. *Sequence-controlled polymers*. Wiley-VCH, Weinheim, Germany, 2018.

- [53] Diego Núñez-Villanueva, Maria Ciaccia, Giulia Iadevaia, Elena Sanna, and Christopher A Hunter. Sequence information transfer using covalent template-directed synthesis. *Chemical Science*, 10:5258–5266, 2019.
- [54] Diego Núñez-Villanueva and Christopher A Hunter. Replication of sequence information in synthetic oligomers. *Accounts of Chemical Research*, 54:1298–1306, 2021.
- [55] Michael P Robertson and Gerald F Joyce. Highly efficient self-replicating rna enzymes. *Chemistry & Biology*, 21:238–245, 2014.
- [56] Tamara Kosikova and Douglas Philp. Two synthetic replicators compete to process a dynamic reagent pool. *Journal of the American Chemical Society*, 141:3059–3072, 2019.
- [57] Charles Walter. *The Use of Product Inhibition and Other Kinetic Methods in the Determination of Mechanisms of Enzyme Action*, pages 645–724. John Wiley & Sons, Ltd, Hoboken, NJ, 1964.
- [58] Javier Cabello-Garcia, Wooli Bae, Guy-Bart V Stan, and Thomas E Ouldridge. Handhold-mediated strand displacement: a nucleic acid based mechanism for generating far-from-equilibrium assemblies through templated reactions. *ACS Nano*, 15:3272–3283, 2021.
- [59] Yijun Guo, Bing Wei, Shiyan Xiao, Dongbao Yao, Hui Li, Huaguo Xu, Tingjie Song, Xiang Li, and Haojun Liang. Recent advances in molecular machines based on toehold-mediated strand displacement reaction. *Quantitative Biology*, 5:25–41, 2017.
- [60] Javier Cabello-Garcia, Rakesh Mukherjee, Wooli Bae, Guy-Bart V Stan, and Thomas E Ouldridge. Information propagation through enzyme-free catalytic templating of dna dimerization with weak product inhibition. *bioRxiv preprint*, 2023.

- [61] Jack W Szostak. The narrow road to the deep past: in search of the chemistry of the origin of life. *Angewandte Chemie International Edition*, 56:11037–11043, 2017.
- [62] Jack W Szostak, David P Bartel, and P Luigi Luisi. Synthesizing life. *Nature*, 409:387–390, 2001.
- [63] Thomas M Cover and Joy A Thomas. *Elements of information theory*. Wiley-Interscience, Hoboken, NJ, 2006.
- [64] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27:379–423, 1948.
- [65] William T Tutte. The dissection of equilateral triangles into equilateral triangles. *Mathematical Proceedings of the Cambridge Philosophical Society*, 44:463–482, 1948.
- [66] Matthew Thomson and Jeremy Gunawardena. The rational parameterisation theorem for multisite post-translational modification systems. *Journal of Theoretical Biology*, 261:626–636, 2009.
- [67] Nicolaas Van Kampen. *Stochastic processes in physics and chemistry*. Elsevier, Amsterdam, NL, 1992.
- [68] Arnold Nordsieck, Willis E Lamb Jr, and George E Uhlenbeck. On the theory of cosmic-ray showers i the furry model and the fluctuation problem. *Physica*, 7:344–360, 1940.
- [69] Terrell L Hill. Studies in irreversible thermodynamics iv. diagrammatic representation of steady state fluxes for unimolecular systems. *Journal of Theoretical Biology*, 10:442–459, 1966.
- [70] Venkat Anantharam and Pantelis Tsoucas. A proof of the markov chain tree theorem. *Statistics & Probability Letters*, 8:189–192, 1989.

- [71] Terrell L Hill. Interrelations between random walks on diagrams (graphs) with and without cycles. *Proceedings of the National Academy of Sciences*, 85:2879–2883, 1988.
- [72] Cheng Qian, Min Qian, and Minping Qian. Markov chain as a model of hill’s theory on circulation. *Scientia Sinica*, 24:1431–1448, 1981.
- [73] Minping Qian, Min Qian, and Cheng Qian. Circulations of markov chains with continuous time and the probability interpretation of some determinants. *Scientia Sinica*, 27:470–481, 1984.
- [74] John G Kemeny and James Laurie Snell. *Finite Markov chains*. Springer, New York, NY, 1983.
- [75] Hans-Helmut Kohler and Eva Vollmerhaus. The frequency of cyclic processes in biological multistate systems. *Journal of Mathematical Biology*, 9:275–290, 1980.
- [76] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22:403–434, 1976.
- [77] Thomas E Ouldridge. The importance of thermodynamics for molecular systems, and the importance of molecular systems for thermodynamics. *Natural Computing*, 17:3–29, 2018.
- [78] Riccardo Rao and Massimiliano Esposito. Nonequilibrium thermodynamics of chemical reaction networks: Wisdom from stochastic thermodynamics. *Physical Review X*, 6:041064, 2016.
- [79] Riccardo Rao and Massimiliano Esposito. Conservation laws and work fluctuation relations in chemical reaction networks. *The Journal of Chemical Physics*, 149:245101, 2018.

- [80] Matteo Polettini, Artur Wachtel, and Massimiliano Esposito. Dissipation in noisy chemical networks: The role of deficiency. *The Journal of Chemical Physics*, 143:184103, 2015.
- [81] Matteo Polettini and Massimiliano Esposito. Irreversible thermodynamics of open chemical networks. i. emergent cycles and broken conservation laws. *The Journal of Chemical Physics*, 141:024117, 2014.
- [82] David F Anderson, Gheorghe Craciun, and Thomas G Kurtz. Product-form stationary distributions for deficiency zero chemical reaction networks. *Bulletin of Mathematical Biology*, 72:1947–1970, 2010.
- [83] Shesha Gopal Marehalli Srinivas, Matteo Polettini, Massimiliano Esposito, and Francesco Avanzini. Deficiency, kinetic invertibility, and catalysis in stochastic chemical reaction networks. *The Journal of Chemical Physics*, 158:204108, 2023.
- [84] Martin Feinberg. Chemical reaction network structure and the stability of complex isothermal reactors—i. the deficiency zero and deficiency one theorems. *Chemical Engineering Science*, 42:2229–2268, 1987.
- [85] William J Heuett and Hong Qian. Grand canonical markov model: A stochastic theory for open nonequilibrium biochemical networks. *The Journal of Chemical Physics*, 124:044110, 2006.
- [86] Udo Seifert. Stochastic thermodynamics of single enzymes and molecular motors. *The European Physical Journal E*, 34:1–11, 2011.
- [87] Massimiliano Esposito and Christian Van den Broeck. Second law and landauer principle far from equilibrium. *Europhysics Letters*, 95:40004, 2011.
- [88] Artur Wachtel, Riccardo Rao, and Massimiliano Esposito. Thermodynamically consistent coarse graining of biocatalysts beyond michaelis–menten. *New Journal of Physics*, 20:042002, 2018.

- [89] Benjamin Qureshi, Jordan Juritz, Jenny M Poulton, Adrian Beersing-Vasquez, and Thomas E Ouldrige. A universal method for analyzing copolymer growth. *The Journal of Chemical Physics*, 158:104906, 2023.
- [90] Stephen Whitelam, Rebecca Schulman, and Lester Hedges. Self-assembly of multicomponent structures in and out of equilibrium. *Physical Review Letters*, 109:265506, 2012.
- [91] Michael Nguyen and Suriyanarayanan Vaikuntanathan. Design principles for nonequilibrium self-assembly. *Proceedings of the National Academy of Sciences*, 113:14231–14236, 2016.
- [92] Charles H Bennett. Dissipation-error tradeoff in proofreading. *BioSystems*, 11:85–91, 1979.
- [93] Pierre Gaspard. Iterated function systems for dna replication. *Physical Review E*, 96:042403, 2017.
- [94] David Andrieux and Pierre Gaspard. Nonequilibrium generation of information in copolymerization processes. *Proceedings of the National Academy of Sciences*, 105:9516–9521, 2008.
- [95] David Andrieux and Pierre Gaspard. Molecular information processing in nonequilibrium copolymerizations. *The Journal of Chemical Physics*, 130:014901, 2009.
- [96] David Andrieux and Pierre Gaspard. Information erasure in copolymers. *Europhysics Letters*, 103:30004, 2013.
- [97] Pierre Gaspard and David Andrieux. Kinetics and thermodynamics of first-order markov chain copolymerization. *The Journal of Chemical Physics*, 141:044908, 2014.
- [98] Pablo Sartori and Simone Pigolotti. Kinetic versus energetic discrimination in biological copying. *Physical Review Letters*, 110:188101, 2013.

- [99] Pablo Sartori and Simone Pigolotti. Thermodynamics of error correction. *Physical Review X*, 5:041039, 2015.
- [100] Mamata Sahoo, Noushad Arsha, Priya Ranjan Baral, and Stefan Klumpp. Accuracy and speed of elongation in a minimal model of dna replication. *Physical Review E*, 104:034417, 2021.
- [101] Yonghyun Song and Changbong Hyeon. Thermodynamic cost, speed, fluctuations, and error reduction of biological copy machines. *The Journal of Physical Chemistry Letters*, 11:3136–3143, 2020.
- [102] Simone Pigolotti and Pablo Sartori. Protocols for copying and proofreading in template-assisted polymerization. *Journal of Statistical Physics*, 162:1167–1182, 2016.
- [103] Felix Wong, Ariel Amir, and Jeremy Gunawardena. Energy-speed-accuracy relation in complex networks for biological discrimination. *Physical Review E*, 98:012420, 2018.
- [104] William D Piñeros and Tsvi Tlusty. Kinetic proofreading and the limits of thermodynamic uncertainty. *Physical Review E*, 101:022415, 2020.
- [105] Pierre Gaspard. Kinetics and thermodynamics of living copolymerization processes. *Philosophical Transactions of the Royal Society A*, 374:20160147, 2016.
- [106] Pierre Gaspard. Kinetics and thermodynamics of dna polymerases with exonuclease proofreading. *Physical Review E*, 93:042420, 2016.
- [107] Pierre Gaspard. Kinetics and thermodynamics of exonuclease-deficient dna polymerases. *Physical Review E*, 93:042419, 2016.
- [108] David Andrieux and Pierre Gaspard. Molecular information processing in nonequilibrium copolymerizations. *The Journal of Chemical Physics*, 130:014901, 2009.

- [109] Riccardo Rao and Luca Peliti. Thermodynamics of accuracy in kinetic proofreading: dissipation and efficiency trade-offs. *Journal of Statistical Mechanics: Theory and Experiment*, 2015:P06001, 2015.
- [110] Kinshuk Banerjee, Anatoly B Kolomeisky, and Oleg A Igoshin. Elucidating interplay of speed and accuracy in biological error correction. *Proceedings of the National Academy of Sciences*, 114:5183–5188, 2017.
- [111] Davide Chiuchiù, Yuhai Tu, and Simone Pigolotti. Error-speed correlations in biopolymer synthesis. *Physical Review Letters*, 123:038101, 2019.
- [112] Arvind Murugan, David A Huse, and Stanislas Leibler. Speed, dissipation, and error in kinetic proofreading. *Proceedings of the National Academy of Sciences*, 109:12034–12039, 2012.
- [113] Jenny M Poulton, Pieter Rein ten Wolde, and Thomas E Ouldridge. Nonequilibrium correlations in minimal dynamical models of polymer copying. *Proceedings of the National Academy of Sciences*, 116:1946–1951, 2019.
- [114] Jenny M Poulton and Thomas E Ouldridge. Edge-effects dominate copying thermodynamics for finite-length molecular oligomers. *New Journal of Physics*, 23:063061, 2021.
- [115] Mamata Sahoo and Stefan Klumpp. Backtracking dynamics of rna polymerase: pausing and error correction. *Journal of Physics: Condensed Matter*, 25:374104, 2013.
- [116] Pierre Gaspard. Growth and dissolution of macromolecular markov chains. *Journal of Statistical Physics*, 164:17–48, 2016.
- [117] Davide Chiuchiù, James Ferrare, and Simone Pigolotti. Assembly of heteropolymers via a network of reaction coordinates. *Physical Review E*, 100:062502, 2019.

- [118] John J Hopfield. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proceedings of the National Academy of Sciences*, 71:4135–4139, 1974.
- [119] Jacques Ninio. Kinetic amplification of enzyme discrimination. *Biochimie*, 57:587–595, 1975.
- [120] Arvind Murugan, David A Huse, and Stanislas Leibler. Discriminatory proofreading regimes in nonequilibrium systems. *Physical Review X*, 4:021016, 2014.
- [121] Qiwei Yu, Anatoly B Kolomeisky, and Oleg A Igoshin. The energy cost and optimal design of networks for biological discrimination. *Journal of the Royal Society Interface*, 19:20210883, 2022.
- [122] Måns Ehrenberg and Clas Blomberg. Thermodynamic constraints on kinetic proofreading in biosynthetic pathways. *Biophysical Journal*, 31:333–358, 1980.
- [123] Qiu-Shi Li, Pei-Dong Zheng, Yao-Gen Shu, Zhong-Can Ou-Yang, and Ming Li. Template-specific fidelity of dna replication with high-order neighbor effects: A first-passage approach. *Physical Review E*, 100:012131, 2019.
- [124] Frederick T Wall. The structure of vinyl copolymers. *Journal of the American Chemical Society*, 63:1862–1866, 1941.
- [125] Frederick T Wall. The structure of copolymers. ii1. *Journal of the American Chemical Society*, 66:2050–2057, 1944.
- [126] Vahe Galstyan and Rob Phillips. Allostery and kinetic proofreading. *The Journal of Physical Chemistry B*, 123:10990–11002, 2019.
- [127] Joel D Mallory, Oleg A Igoshin, and Anatoly B Kolomeisky. Do we understand the mechanisms used by biological systems to correct their errors? *The Journal of Physical Chemistry B*, 124:9289–9296, 2020.

- [128] Pierre Gaspard. Template-directed copolymerization, random walks along disordered tracks, and fractals. *Physical Review Letters*, 117:238101, 2016.
- [129] Field Cady and Hong Qian. Open-system thermodynamic analysis of DNA polymerase fidelity. *Physical Biology*, 6:036011, 2009.
- [130] Massimiliano Esposito. Stochastic thermodynamics under coarse graining. *Physical Review E*, 85:041125, 2012.
- [131] Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford University Press, Oxford, UK, 2020.
- [132] Thomas E Ouldridge, Rory Brittain, and Pieter Rein ten Wolde. The power of being explicit: Demystifying work, heat, and free energy in the physics of computation. In Chris Kempes, David H Wolpert, Peter F Stadler, and Joshua A Grochow, editors, *The Energetics of Computing in Life & Machines*, pages 307–351. SFI Press, Santa Fe, NM, 2019.
- [133] Massimiliano Esposito, Katja Lindenberg, and Christian Van den Broeck. Extracting chemical energy by growing disorder: efficiency at maximum power. *Journal of Statistical Mechanics: Theory and Experiment*, 2010:P01008, 2010.
- [134] Yong-Shun Song, Yao-Gen Shu, Xin Zhou, Zhong-Can Ou-Yang, and Ming Li. Proofreading of dna polymerase: a new kinetic model with higher-order terminal effects. *Journal of Physics: Condensed Matter*, 29:025101, 2016.
- [135] David A Harville. *Matrix Algebra From a Statistician’s Perspective*. Springer, New York, NY, 1998.
- [136] Benjamin Qureshi, Jenny M Poulton, and Thomas E Ouldridge. Information propagation in far-from-equilibrium molecular templating networks is optimised by pseudo-equilibrium systems with negligible dissipation. *arXiv preprint arXiv:2404.02791*, 2024.

- [137] James D Watson. *Molecular biology of the gene*. Benjamin-Cummings Publishing Company, Boston, MA, 2014.
- [138] Jordan Juritz, Jenny M Poulton, and Thomas E Ouldridge. Minimal mechanism for cyclic templating of length-controlled copolymers under isothermal conditions. *The Journal of Chemical Physics*, 156:074103, 2022.
- [139] Charles H Bennett. The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, 21:905–940, 1982.
- [140] Arthur Genthon, Carl D Modes, Frank Jülicher, and Stephan W Grill. Non equilibrium transitions in a polymer replication ensemble. *arXiv preprint arXiv:2403.05665*, 2024.
- [141] Jeremy Gunawardena. A linear framework for time-scale separation in nonlinear biochemical systems. *PLoS ONE*, 7:e36321, 2012.
- [142] Kee-Myoung Nam, Rosa Martinez-Corral, and Jeremy Gunawardena. The linear framework: using graph theory to reveal the algebra and thermodynamics of biomolecular systems. *Interface Focus*, 12:20220013, 2022.
- [143] Stephen M Robinson. A short proof of cramer’s rule. *Mathematics Magazine*, 43:94–95, 1970.
- [144] Christian Maes and Karel Netočný. Heat bounds and the blowtorch theorem. In *Annales Henri Poincaré*, volume 14, pages 1193–1202. Springer, 2013.
- [145] Meritxell Sáez, Elisenda Feliu, and Carsten Wiuf. *Linear Elimination in Chemical Reaction Networks*, pages 177–193. Springer International Publishing, New York, NY, 2019.
- [146] Uğur Çetiner and Jeremy Gunawardena. Reformulating nonequilibrium steady states and generalized hopfield discrimination. *Physical Review E*, 106:064128, 2022.

- [147] Philip C Nelson. *Biological physics: energy, information, life*. Chiliacon Science, Philadelphia PA, 2020.
- [148] Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75:126001, 2012.
- [149] Thomas E Ouldridge, Christopher C Govern, and Pieter Rein ten Wolde. Thermodynamics of computational copying in biochemical systems. *Physical Review X*, 7:021004, 2017.
- [150] Sylvain Ladame. Dynamic combinatorial chemistry: on the road to fulfilling the promise. *Organic and Biomolecular Chemistry*, 6:219–226, 2008.
- [151] Shiling Liang, Paolo De Los Rios, and Daniel Maria Busiello. Thermodynamic bounds on symmetry breaking in linear and catalytic biochemical systems. *Physical Review Letters*, 132:228402, 2024.