Imperial College London

Minimal molecular mechanisms for autonomous template copying

Jordan Oliver Faure Juritz

Department of Bioengineering
Faculty of Engineering
Imperial College London

Submitted in partial fulfilment of the requirements
for the degree of PhD in Bioengineering at
Bioengineering, Imperial College London, April 2023

Statement of originality

The work reported within this thesis is an original product of my own efforts. All information or ideas that were derived from published or unpublished work have been acknowledged and referenced accordingly.

Jordan Juritz

Copyright

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.

Abstract

In this thesis, I develop models of chemical reactions in which a template copolymer directs the synthesis of new copolymers (copies). I use the models to inform the design of synthetic analogues of the information-copying molecular machines found in nature.

Template copying processes, like DNA replication, transcription, and translation, are the basis of living systems' ability to produce, maintain and evolve a diverse set of complex biomolecules using a small set of molecular building blocks. Biological copying processes can operate in constant environments driven by chemical fuel molecules and typically use each template many times over, as the copies eventually separate. To date, no synthetic systems have been created that can achieve the same feat under the same "life-like" conditions. Experimental attempts have been hindered by product inhibition, the tendency for long copies to bind to the template instead of spontaneously separating. Furthermore, few theoretical models have considered the effects of the separation of copies from the templates or the effects of product inhibition on the copy-length distributions.

In Chapter 2, I identify a lack of suitable metrics to compare template copying reactions and argue that, if we are to build them, the time-averaged turnover frequency (TATF) - a measure of the average copy production rate per unit template - should serve this purpose.

The patterns of product inhibition are more complex on longer templates. In Chapter 3, I present a coarse-grained stochastic simulation of copolymerisation on a finite-length template and use it to identify reaction mechanisms that could overcome product inhibition.

In Chapter 4, I explore how the mechanisms identified in Chapter 3 should be implemented in a model system, a DNA strand-displacement reaction network. Through simulation, I identify optimal designs to characterise experimentally.

For June and Bea

Acknowledgements

I am grateful for the many friends and colleagues who have offered their encouragement throughout this project, of whom just a few are named here. Directly or indirectly, they all played a part in creating this piece of work. I am particularly grateful to my supervisor, Thomas Ouldridge, for his support, guidance, and dedication at every stage of this project – from sketching early models together to the editing of this thesis. I would also like to thank Jenny Poulton, whose friendship and supervision during the initial years of my PhD laid a solid foundation that enabled me to navigate this project successfully. I am lucky to have spent my time at Imperial with a group of delightful and talented people in the Ouldridge Lab, including Ismael, Javier, Ben, Alicia, Jeremy, Francesca, Kate, Alex, Križan, Lisa, Wooli, Rakesh, and Aditya, and the members of the Bioengineering department. Each of you has played an integral part in creating a welcoming, inspiring, and intellectually stimulating environment. Thank you for countless enjoyable conversations.

To Mum, Dad, Jocie, my family, Bea's family, and dear friends, thank you for your love, company, and humour, which makes everything all the more worthwhile. To my grandmother, June, an excited conversation we shared on a sunny afternoon in Cape Town in April 2018 helped me realise that the questions that lie at the heart of this thesis were the ones to go after. Thank you for sharing my joy. Bea, your energy, patience, and love have given me the strength to persevere through the most challenging moments of the past few years. Thank you for believing in me and for celebrating the big and small moments along the way.

Contents

Abstract 1

1 Introduction 9
- 1.0.1 The limitations of self-assembly . . . 11
- 1.0.2 The potential of autonomous template copying systems . . . 16
- 1.0.3 Product inhibition and experimental approaches to templated reactions . . . 18
- 1.0.4 Handhold-mediated strand displacement for autonomous template copying reactions . . . 20
- 1.0.5 Length distribution control and the modelling of templated copying systems . . . 24
- 1.0.6 Aims . . . 28

1.1 Summary of sections . . . 28

1.2 Methods: Construction of thermodynamically self-consistent coarse-grained models . . . 30
- 1.2.1 Defining States . . . 30
- 1.2.2 Principle of detailed balance . . . 32
- 1.2.3 States, Free-energies, Transitions, and Rate constants . . . 32
- 1.2.4 The small number limit: Discrete-state continuous-time Markov chain . . . 33
- 1.2.5 The large number limit: Continuous-state continuous-time Markov chain and the chemical kinetic equation . . . 34

2 Metrics for comparing templated ligation reactions 35

2.1 Introduction . . . 35
- 2.1.1 Nomenclature . . . 35
- 2.1.2 Review of metrics . . . 38

2.2 Ordered bi-uni with competitive product inhibition . . . 44
- 2.2.1 Simplifying the steady-state rate law . . . 46
- 2.2.2 Estimating TATF in the irreversible limit . . . 48

2.3 Ordered bi-uni with mixed product inhibition . . . 52
- 2.3.1 Approximating TATF in the irreversible limit . . . 54

2.4 Random order bi-uni reaction . . . 57
- 2.4.1 Approximating the TATF in the irreversible limit . . . 60

2.5 Conclusions . . . 64

3 A coarse-grained model of whole-template copolymerisation dynamics 67

3.1 Introduction . . . 67

3.2 Model and methods . . . 69
- 3.2.1 Model . . . 69
- 3.2.2 Methods . . . 76

3.3 Results . . . 77
- 3.3.1 A simple model of copy-template binding and copy-copy polymerisation cannot reliably produce long polymers . . . 77
- 3.3.2 A polymerisation mechanism that disrupts copy-template binding allows for long copies 81
- 3.3.3 Weakening the final site on the template biases the production of complete polymers . . . 88

3.3.4 On-rate discrimination enables the reliable production of copies with the correct length and accurate sequences . . . 95
3.4 Conclusion . . . 99
3.5 Data Availability . . . 103

4 Modelling a DNA strand displacement implementation of templated copying 104
4.1 Introduction . . . 104
4.2 Methods and Model . . . 105
4.2.1 Methods: Construction of a DNA strand displacement reaction network model . . . 105
4.2.2 Chemical reaction network model . . . 111
4.2.3 Computational construction of a chemical reaction network model . . . 114
4.2.4 Transitions between conformations . . . 116
4.2.5 Model parameterisation: Reducing the number of free rate constants . . . 118
4.2.6 Model parameterisation: Calculating standard free energies . . . 119
4.2.7 Model parameterisation: Rate constant estimation . . . 123
4.2.8 Parameter sets span the design space . . . 126
4.2.9 Solving the system of ODEs . . . 133
4.2.10 Performance . . . 134
4.2.11 Measureables . . . 135
4.2.12 Scoring metrics based upon the measures . . . 138
4.3 Results: Optimising for independent metrics . . . 141
4.3.1 Optimising Yield . . . 141
4.3.2 Optimising reaction speed . . . 147
4.3.3 Optimising for minimal leak reaction . . . 156
4.3.4 Minimisation of dimer release . . . 161
4.3.5 Optimising across metrics . . . 166
4.4 Conclusions . . . 172

5 Conclusions 178
5.1 Summaries . . . 179
5.1.1 Chapter 2 . . . 179
5.1.2 Chapter 3 . . . 180
5.1.3 Chapter 4 . . . 181
5.2 Outlook . . . 182
5.2.1 Trimers and self-replicators . . . 182
5.2.2 Scaffolding the production of non-nucleic products . . . 184
5.2.3 Autonomous templated computation . . . 184
5.2.4 Stochastic thermodynamics of length-control . . . 186
5.3 Reflection . . . 187

A Additional material supporting chapter 2 200
A.1 Extreme parameter values create long embedded processes . . . 200
A.2 Full product length distributions accompanying main text Fig. 3.12 . . . 202
A.3 The asymmetric destabilisation and weak end mechanism is more reliable for polymer tails with higher template affinity . . . 206

List of Tables

4.1 Scaling of system size with template length 111
4.2 Nupack standard free energy calculation 120
4.3 Estimating loop free energies 121

List of Figures

1.1 Biological copying enzymes . . . 10
1.2 Comparing self-assembly and template copying . . . 12
1.3 Self-assembling systems in biology . . . 14
1.4 DNA self-assembly . . . 15
1.5 "What I cannot create, I do not understand" . . . 17
1.6 Handhold-mediated strand displacement dimerisation reaction . . . 21
1.7 HMSD based monomer sets . . . 23
1.8 Illustrations of template copying processes . . . 25
1.9 Model construction . . . 31

3.1 Data structure of the model . . . 69
3.2 Basic Model of Templated Polymerisation . . . 72
3.3 Transitions occur at any site on the template . . . 73
3.4 Simple model - parameter sweep . . . 77
3.5 Simple model - $\langle L \rangle$ against ΔG_{BB} . . . 78
3.6 Simple model - Dimer cycle . . . 79
3.7 Generic bond model . . . 80
3.8 Generic bond model - Heatmap plots . . . 82
3.9 Generic bond model - Stepping model . . . 85
3.10 Weak end model - Free energy landscape . . . 87
3.11 Weak end model . . . 88
3.12 Weak end model - Parameter sweep . . . 89
3.13 Weak end model - Length distribution surface . . . 90
3.14 Weak end model - Dimer cycles . . . 93
3.15 Binary alphabet model . . . 96
3.16 Binary alphabet model - On vs Off rate discrimination . . . 97

4.1 Proposed mechanism for templated trimerisation . . . 106
4.2 Domain level depiction of permitted reactions . . . 107
4.3 A simplified depiction of permitted reactions . . . 109
4.4 Stoichiometry matrices . . . 113
a Adjacency matrix 2 . . . 113
b Stoichiometry matrix 2 . . . 113
c Adjacency matrix 2 . . . 113
d Stoichiometry matrix 2 . . . 113
4.5 Correcting the scaling of zipping standard free energy changes with loop length . . . 122
4.6 Determining the second-order binding rate constants . . . 123
a Fitting diffusion-limited binding rates . . . 123
b Validating binding rates by comparison to dissociation rates . . . 123
4.7 Determining the leak and polymerisation rate . . . 124
a Fitting branch migration rate . . . 124

4.8 Second order rate constant . . . 125
4.9 Variation of free energies with input parameters - 1 . . . 127
4.10 Variation of free energies with input parameters - 2 . . . 128
4.11 Variation of free energies with input parameters - 3 . . . 129
4.12 Variation of rates with input parameters - 1 . . . 130
4.13 Variation of rates with input parameters - 3 . . . 131
4.14 Variation of rates with input parameters - 4 . . . 132
4.15 Yield and speed measures . . . 136
a Comparison of yield measures without leak . . . 136
b Comparison of yield measures with leak . . . 136
c Comparison of timescale measures . . . 136
4.16 Measures of the ratio of kinetic measures with and without leak reaction . . . 137
a Comparison of yield measures without and with leak . . . 137
b Comparison of yield measures without and with leak . . . 137
c Reaction order measures without and with leak . . . 137
4.17 Metric distributions . . . 140
4.18 Input parameter distributions . . . 142
4.19 Standard free energy distributions . . . 143
4.20 Optimising trimer yield . . . 144
a Free energy of polymerisation constrains the trimer yield . . . 144
b Trimer template retention increases with domain length . . . 144
4.21 Timeseries of highest yield . . . 146
4.22 Input parameter distributions that optimise speed . . . 148
4.23 Rate constant distributions that optimise speed . . . 149
4.24 Timeseries of parameters that maximise speed . . . 150
4.25 Reaction rate model . . . 150
4.26 Variation of speed with information domain length . . . 151
4.27 Input parameter distributions that minimise leak . . . 154
4.28 Rate constant distributions that minimise leak . . . 155
4.29 Flux cycle analysis of a leak-prone parameter set . . . 157
4.30 An illustration of the process in which the tail of a polymer unzips from the template and is prevented from re-zipping by the binding of a monomer. . . . 158
4.32 Measure distributions that prevent early dimer release . . . 163
4.33 Input parameter distributions that prevent dimer release . . . 164
4.34 Timeseries of best scores with weighting $w = (0,0,0,0,0,0,1)$ . . . 165
4.35 Measure distributions that were optimal overall . . . 167
4.36 Input parameter distributions . . . 168
4.37 Timeseries of best scores with weighting $w = (11010010)$ . . . 169
4.38 Flux cycle of an optimal parameter set . . . 171
4.39 Estimation of the reliability limit of information transfer . . . 176

5.1 Scaffolding and self replication . . . 183
5.2 From copying to computing . . . 185

A.1 Simple model - Low samples . . . 201
A.2 Weak end model - Length distribution surfaces 1 . . . 203
A.3 Weak end model - Length distribution surfaces 2 . . . 204
A.4 Weak end model - Length distribution surfaces 3 . . . 205
A.5 Weak end model - Strengthening ΔG_{spec} increases range of ΔG_{end} over which p_{complete} is increased 207

Chapter 1

Introduction

Parts of this introduction were developed from an early draft of the introduction that I wrote in [1].

Life relies on the storage, copying, replication, folding and catalytic properties of copolymers– polymers that consist of more than one type of monomeric unit– such as DNA, RNA, and proteins. Proteins are the functional molecules of life and are comprised of one or more polypeptide chains, copolymers consisting of amino acids linked together by peptide bonds. The sequence of amino acids along polypeptide chains (a 1D structure) directs the folding of the molecule into a three-dimensional structure, a transformation that can create binding sites for substrates and confer catalytic properties to the protein [2]. The sheer variety of copolymer sequences found in biological systems is remarkable; the human proteome alone consists of approximately 20,000 distinct proteins with a median length of 375 amino acid residues [3], each formed from an "alphabet" of just 20 amino acids. Each amino acid can be classified into one of four primary groups based upon the properties of their side-chains (hydrophobic, hydrophilic, negatively charged and positively charged)[4, 2]. As of the time of writing, over 240 million protein sequences sampled across life have been recorded on the UniProt database [5]. Furthermore, the scale of the chemical space which biology could sample from is many fold vaster. A rough upper bound on the scale of the accessible chemical space can be set by the number of permutations of different median length protein sequences, the size of the sequence space, which lies somewhere between $4^{375} \approx 10^{226}$ and $20^{375} \approx 10^{488}$. It is remarkable that, with an arsenal of just 20 amino acids, living systems can selectively produce 20,000 distinct, functional proteins in a vast chemical space of up to 10^{488} sequences.

Life's capacity to create such diverse copolymers with such a small set of "building-blocks" can be attributed to templating. Proteins, RNA and DNA are all produced through templated copying processes [2]. In these processes, the template copolymer guides an ensemble of sophisticated molecular machinery in synthesising a new copolymer by facilitating complementary interactions between monomers and sites on the template. As

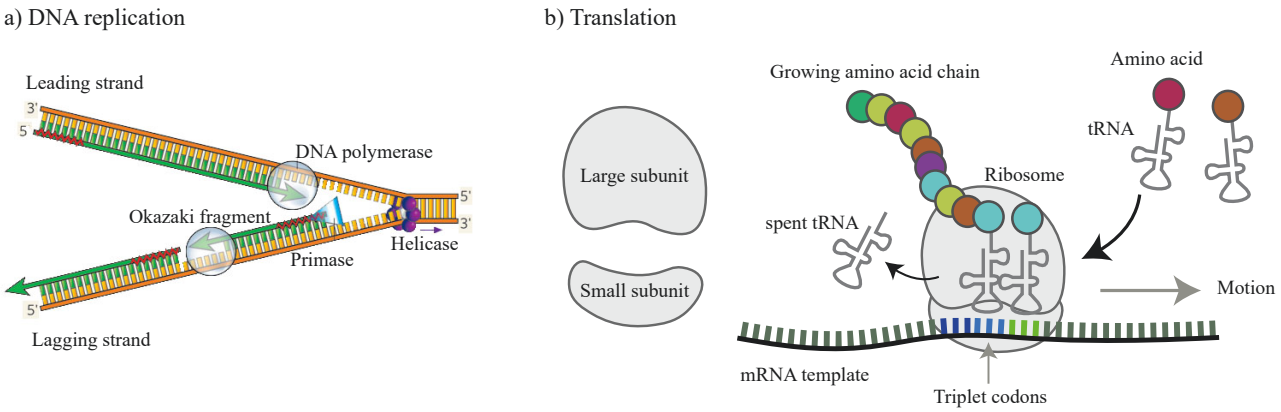


Figure 1.1: a) DNA replication. In the DNA replication cycle, a DNA duplex is separated by helicases prior to the synthesis of daughter strands. Polymerases synthesise new DNA strands under the guide of the leading and lagging strand templates. Image source: [6]. **b)** Translation of mRNA to proteins by the ribosome. The ribosome is comprised of a large and small subunit. A messenger RNA (mRNA) strand encodes the sequence of an amino acid chain (protein) in triplet codons. The ribosome progresses along the mRNA chain. Amino acids tagged with transfer RNAs (tRNAs) are prepared by aminoacyl-tRNA synthetases. The tRNAs are have regions that are complementary to the mRNA codons, and guide the amino acids into the ribsome, where they are polymerised to the end of a growing amino acid chain. The amino acid chain folds into a specific conformation as it exits and eventually separates entirely, enabling it to function elsewhere in the cell and for the mRNA template to be reused. Image based upon [7] .

illustrated in Fig. 1.1a, in DNA replication, DNA polymerases synthesise a new DNA strand complementary to a DNA template [2, 8, 9, 10]. In transcription, RNA polymerases produce a strand of RNA from a DNA template [11, 12, 13], and in translation, ribosomes produce polypeptides encoded in mRNA templates, as illustrated in Fig. 1.1b [14, 15, 16, 17, 18, 19]. Here, triplet codons in the mRNA sequence encode for specific amino acids with some redundancy, and the amino acids are delivered to the ribosome tethered to tRNAs which serve as an intermediary between the ribonucleic and peptide chemistries. Each of these fundamental biological processes can be understood as an information copying process, and hence the biophysical principles governing templated copolymer copying are fundamental to biology [20].

Two important observations can be made about these systems, which impact their physical properties and distinguish them from current approaches to nanostructure engineering like self-assembly. Firstly, these copying processes *in isolation* can occur in constant environments with no exploitable temporal or spatial gradients and are driven by chemical fuel molecules, such as ATP [1]; They are not directly driven by external processes. Our research group have referred to processes with this quality as autonomous [21, 22, 23, 24, 1]. It should be noted that preparing the chemical fuel source often exploits spatial gradients, such as the transmembrane proton gradients that drive the production of ATP by F-type ATP synthases [25]. Secondly, biological copying processes can generate copies (copolymers that share a high degree of correlation with the template sequence) that separate from the template at some stage. In transcription and translation, the copies detach from the template during

synthesis, while in DNA replication, the copy is separated from the template that it was produced from by helicases as the next round of replication begins, as shown in Fig. 1.1a. If copies were to remain permanently attached to the template, no further copies could be produced, and no net generation of complex molecules would be achieved.

The fact that the copies eventually separate from their template has significant consequences for the thermodynamics of the process, as first articulated in Ref. [26]. Assuming no sequence-specific intra-copy interactions between monomers, all bonds formed with the template during polymer growth, whether complementary or not, are eventually broken as the copies must eventually separate from the template. The free energy of the separated copolymer, which shares no permanent bonds with the template, is independent of the template's sequence, as shown in Fig. 1.2b. Therefore, the minimal work required to produce a given copolymer is not a function of the template's sequence, and, hence, an equilibrium copier is inaccurate as all copy sequences are equally likely in equilibrium. An illustrative free energy landscape is shown in the lower portion of Fig. 1.2b. Consequently, producing an accurate copy (one sequence in the space) therefore implies the creation of a far-from-equilibrium state (a very low entropy state of the system), which necessitates kinetic discrimination driven by a larger compensatory production of entropy in the environment to satisfy the second law of thermodynamics. Templated copying with autonomous separation is inherently a far-from-equilibrium process [26].

1.0.1 The limitations of self-assembly

In contrast to copying, self-assembly is a near-equilibrium process with limited potential for generating diverse ensembles of specific structures from small sets of monomers or building blocks [21, 1, 27, 28]. For a system to self-assemble monomers into a single finite or repeating structure, the interactions among the monomers must define a free-energy landscape in which the target structure represents a global minimum [28], as depicted in Fig. 1.2a. Unlike nature's template copying machinery, self-assembling systems can operate by relaxing toward equilibrium, as shown in the lower half of Fig. 1.2a.

In cases where the desired structure exhibits symmetries, these can be encoded within the monomers, reducing the number of distinct monomer types needed to form the target structure. For example, biological rotary motors, some of nature's most intricate self-assembling structures, possess parts with cylindrical symmetries. The bacterial flagellar motor, which drives bacterial locomotion, [29] and F-type ATP synthase, which facilitates the phosphorylation of ADP to ATP, [25] contain multiple ring-like structures with cylindrical symmetries formed by the repetition of just a few subunit proteins, depicted in Fig. 1.3 a and b. Similarly, virus capsids often exhibit rotational and icosahedral symmetries and are formed through the self-assembly of a small set of symmetric

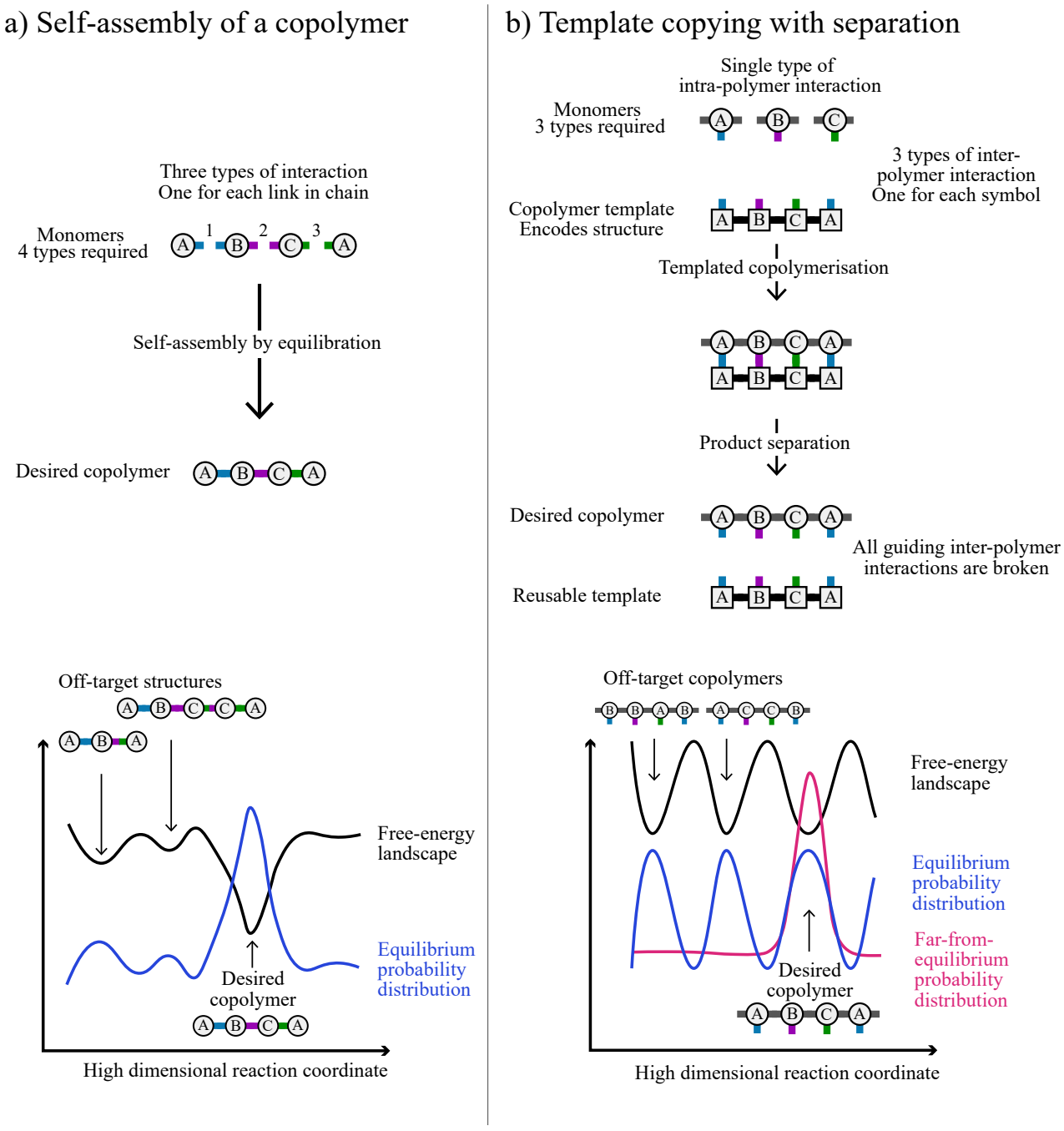


Figure 1.2: Self-assembly works close to equilibrium, whereas template copying with separation is a far-from-equilibrium process. **a)** Self-assembly of a copolymer. In self-assembly, the desired structure (the copolymer A-B-C-A) is encoded in the interactions between the building blocks. The colours of the intra-polymer bonds denote complementarity. The location of each monomer in the chain must be exactly specified; therefore, the number of building blocks required is equal to the size of the structure, not the size of the informational alphabet (A,B,C). The desired structure is the minimum of a free energy landscape and these structures have a high chance of forming correctly in equilibrium. b) Template copying with separation. The sequence of the copolymer copy is encoded in a template molecule. Therefore, intra-monomer interactions can be non-sequence specific (generic), unlike self-assembly. Consequently, the number of monomers required to form the structure is set by the size of the information alphabet (A,B,C), not the length of the product. Complementary interactions that direct the formation of the product are eventually all broken [26]. Therefore the free energy landscape of copolymer sequences is even; there can be no exploitable energetic bias toward any particular copolymer. Therefore, producing a pool of accurate copies (far-from-equilibrium distribution) of the template requires kinetic discrimination, which requires entropy dissipation into the environment [26, 24, 27].

protein subunits (capsid and scaffold proteins), shown in Fig. 1.3c. The symmetries of the self-assembled capsid structures allow the structure to be constructed from a minimal assortment of building blocks, which can be encoded in a smaller viral genome, potentially reducing the burden in the viral replication cycle and conferring an evolutionary advantage. Interestingly, there can be abundant heterogeneity in the capsid structures formed by self-assembly. A recent study found that the vast majority of capsids observed in samples had completely unique structures [30], and thus capsid assembly, the classic example of self-assembly in nature, demonstrates one of self-assembly's pitfalls; its limited ability to consistently produce homogeneous pools of structures due to the reliance on equilibrium conditions and the stochastic nature of self-assembly pathways.

The self-assembly of finite structures without symmetries is even more challenging. If the desired structure lacks symmetries, the interactions between the building blocks must encode the exact position of each block within the desired structure, and therefore the number of unique building blocks required to form a particular structure must increase. Consider the challenge of consistently self-assembling a particular copolymer with a sequence denoted ABCA from a set of building blocks with bonding domains denoted by numbers as depicted in Fig. 1.2a. This sequence would require four unique building blocks to form (A-1,1-B-2,2-C-3,3-A), despite only having three letters in its sequence alphabet (A,B,C). In general, the self-assembly of a copolymer of length N, requires $m = N$ distinct building blocks, each with the identity of its neighbours orthogonally encoded in its edges (in this case the backbone of the copolymer). This poses two challenges as the size of the system increases; a proportional increase in the number of required building blocks and an increasing difficulty in ensuring the orthogonality of all of the interactions [35]. If the interactions are not orthogonal, then off-target structures may be liable to form, depicted as local minima in the free energy landscape in Fig. 1.2a. This starkly contrasts the templating capabilities seen of biology as previously stated, in which a copolymer with any length sequence can be formed from a common set of building blocks. Self-assembly could not form 10^{488} different 375-unit long copolymers from just 20 building blocks. Rather it would require approximately $\sim 375 \times 10^{488}$ distinct building blocks, for which there are far too few particles in the universe $\sim 10^{80}$.

Furthermore, with self-assembly, it is even more challenging to create two or more distinct structures in the same environment using the same pool of building blocks without creating chimeric, misformed structures [35]. Although multifarious systems exist in which the same group of units can self-assemble into multiple structures, the required number of distinct building blocks increases with the number of distinct encoded structures, making the system prone to producing "off-target" structures [28], like those shown in Fig.1.2a.

Nonetheless, self-assembly is a powerful tool for creating structures on the macromolecular scale. It is often successful when the number of distinct structures that need to be formed is small, or where a large number of distinct building blocks with orthogonal interactions can be designed. Indeed, the formation of quaternary

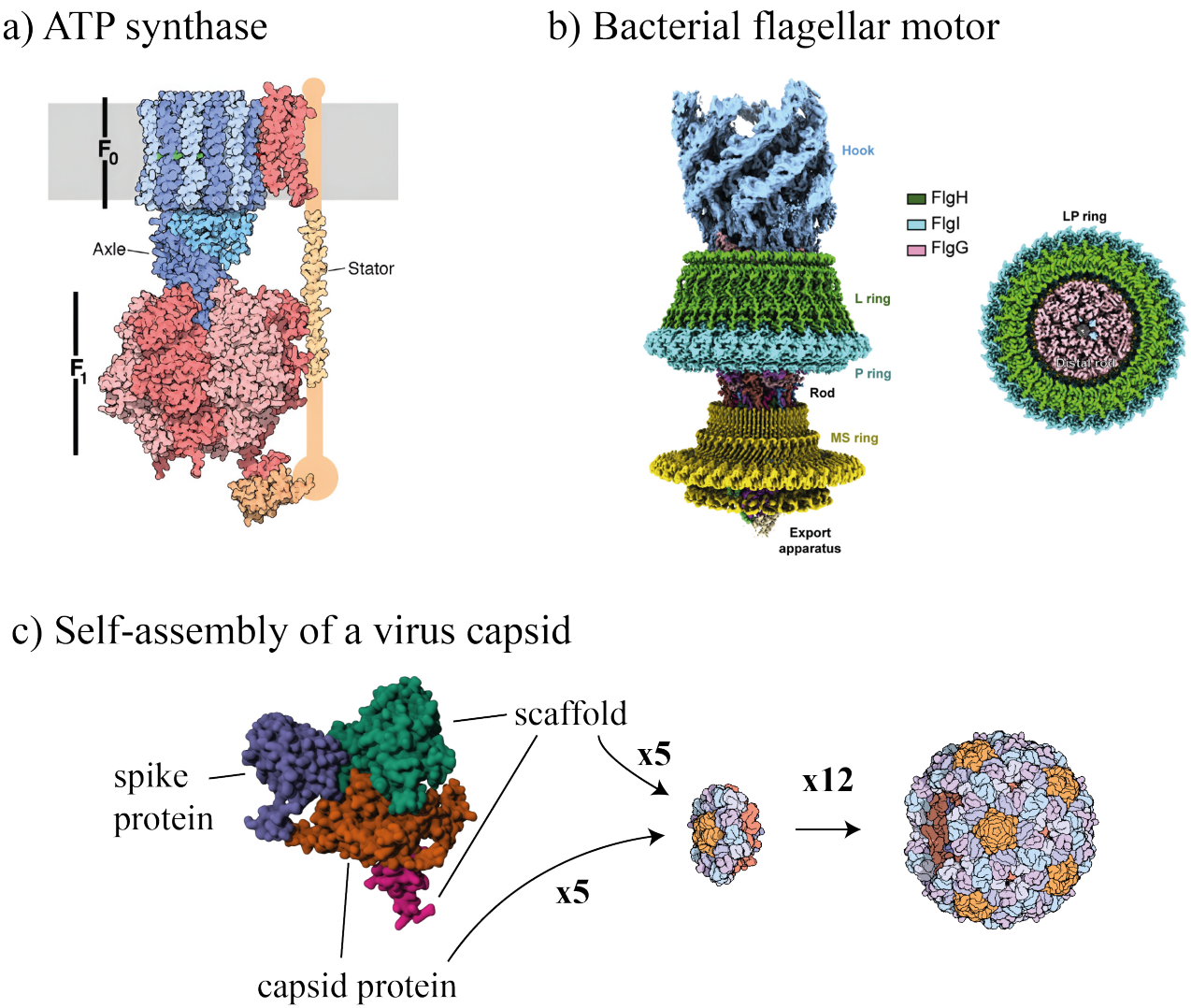


Figure 1.3: Self-assembling systems in biology. **a)** ATP synthase is a rotary motor embedded in cell membranes (shown in grey) that is driven by the motion of protons down trans-membrane potential gradients to synthesise ATP. Both the F0 and F1 subunits of the complex are comprised of repeated proteins which self-assemble into a shape with cylindrical symmetry. Source [31]. **b)** The bacterial flagellar motor drives locomotion in bacteria [32]. It comprises cylindrically symmetrical, ring-like structures built from a small set of repeated proteins, allowing it to rotate embedded in a membrane. Source [32]. **c)** Viruses assemble shell-like structures that enable their proliferation to new cells. To minimise their genome size, they exploit symmetries to assemble large structures from a few repeated building blocks (scaffold and capsid proteins). The viral subunits self assemble into pentameric groups, which then bind together to form a large sphere, with icosahedral symmetries. Adapted from sources [33] and [34].

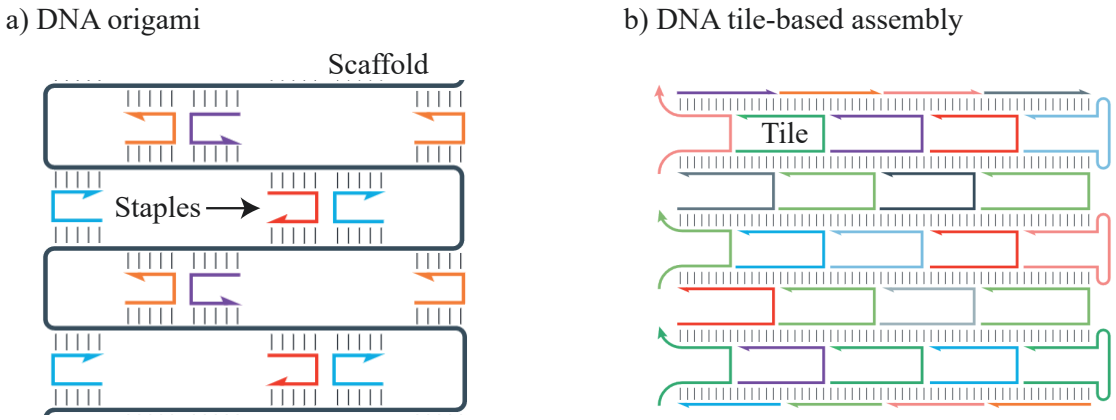


Figure 1.4: **a)** DNA origami involves the folding of a long DNA scaffold strand by short complementary staple strands. **b)** DNA tile-based assembly involves the formation of a large structure by engineering the complementarity of DNA tiles. In both cases, the reduction in the entropy of the system upon forming the desired structure is compensated by the stabilising complementarity interactions. Both images were adapted from [36].

protein complexes (self-assembled complexes of proteins) is possible because of the diversity of their amino acid sequences that can create ternary structures with sufficient complementarity and orthogonality [37, 28]. The ease with which the complementarity of DNA-DNA hybridisation reactions can be designed, as pioneered by the late Ned Seeman [38, 36], has led to the extensive use of DNA as a tool for nanotechnology. Techniques such as DNA origami, shown in Fig.1.4a, involve the folding of a "scaffold", a long single-strand of DNA, by short complementary DNA "staples" to pin the structure into complex shapes [39, 40, 41], whereas the DNA tile-based approaches, shown in Fig.1.4b , involve engineering the edges of "tiles" such that they bind together into a well-defined overall structure, often seeded by the introduction a nucleating or template structure [42, 43, 44]. Highly complex systems and algorithms have been implemented with such systems [45]. Recent advances in the predictive and generative modelling of proteins [46, 47] have opened up the doors to self-assembling systems comprised of proteins [48, 49, 50], combinations of DNA and proteins [51], and carbohydrates and proteins [52].

1.0.2 The potential of autonomous template copying systems

In the light of self-assembly, the advantages of templated copying as an architecture for creating complex structures on the molecular scale are clear:

1. the combinatorics of templating enables a vast chemical space to be explored,
2. individual sequences within this chemical space can be easily sampled by copying a template,
3. many sequences within the space can be sampled simultaneously and independently with minimal potential to form chimeric structures by copying multiple templates in the same environment,
4. and all of this can be achieved with a small alphabet of shared building blocks.

Furthermore, systems based upon templated copying are more evolvable than those that self-assemble, as the instructions to create a functional molecule are not encoded in the functional molecule itself. Hence, mutation of the functional molecule can be achieved without changing the monomer set. However, with non-templated self-assembly, evolving the structure requires changing the monomer set and synthesising a new set of molecules. The under-utilisation of template copying as a tool for engineering matter on the molecular scale represents a missed opportunity to benefit from life's fundamental mechanism for generating chemical complexity. As I will describe in the following pages, there is much to be explored in our experimental and theoretical understanding of autonomous template copying systems.

In this thesis, I describe my contributions to the theoretical understanding and engineering of a non-enzymatic molecular system capable of producing a copolymer by autonomously replicating a template reminiscent of, but not dependent on, biological copying machinery. Such a system would generate accurate copies of a template copolymer from a universal set of building blocks, powered solely by a chemical fuel source in a spatially and temporally constant environment. Additionally, the template would serve as a catalyst, guiding the formation of the copy but, crucially, enabling its separation, allowing the template to be reused multiple times.

Before delving into the literature on templated reactions and pinpointing the challenges that must be addressed to develop such a system, I will touch briefly on my motivations.

Inspired by Feynman's maxim, "What I cannot create, I do not understand," my colleagues and I in the Ouldridge lab aim to design and construct chemical systems that emulate the template copying processes observed in living systems. The copying machinery of life today is highly complex, intertwined with many intercellular processes that it has co-evolved with from its ancient origins [2, 15, 16, 54, 19]. The approach taken in this thesis is to create minimal models of just the essential features that describe the conditions and results of the process – that the template copying machines in biology use chemical fuel to produce a copolymers

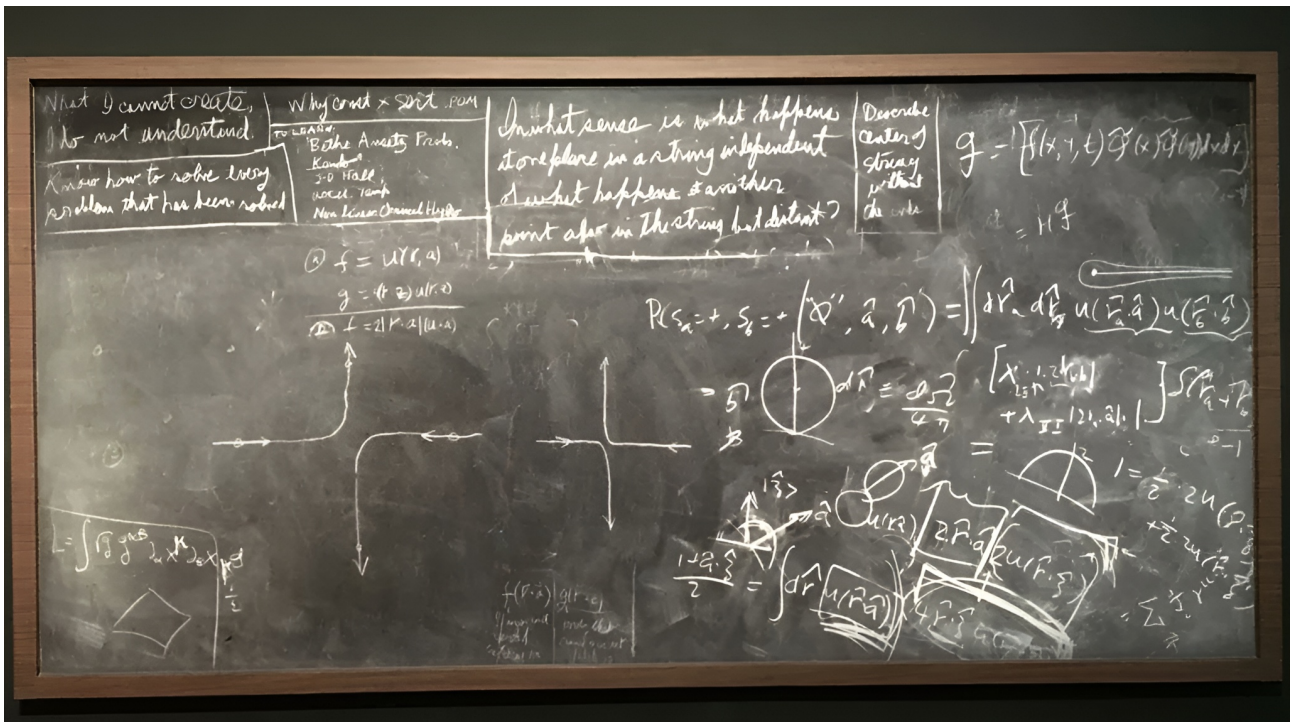


Figure 1.5: Written in the upper left-hand corner of Richard Feynman's blackboard at the time of his death were words that have inspired many, "What I cannot create, I do not understand". In that spirit, we work towards creating an autonomous template copying system. Written just below it and perhaps a few seconds after, "Know how to solve every problem that has been solved". This is left as an exercise for the reader. Source [53]

from a pool of monomers and a template copolymer that eventually separate from the template in an otherwise constant environment. In omitting the specific details of extant copying machines but preserving the fundamental requirements, we hope to articulate the physical laws these systems operate by, such as constraints on their thermodynamic and kinetic properties [26, 23, 24, 21, 22, 1]. We can observe evolution's mechanisms that solve the challenge of copying templates, and minimal models can help us to understand them in a wider context. Are there other solutions to the challenge of template copying that are yet to be discovered? If so, we should build them, not only to test our understanding of the principles but also for the advancement of our nanotechnological capabilities.

An autonomous template copying system could serve as a foundation for producing sequence-defined synthetic copolymers [55] or generating molecules encoded in templates [56, 57]. Lutz referred to a process which grants precise control over the sequence of a copolymer as a "holy grail" in the field of polymer science, inspiring gallant chemists toward this goal [55]. Creating synthetic copolymers with precise length and sequence distributions can affect their thermal, electrical, degradative, and mechanical properties [58, 59, 60]. Perhaps more exciting still, such control could open avenues toward novel catalytic properties just as the sequence of amino acids in a protein affects its function [58, 59, 60]. In the absence of templates, non-autonomous cycling

protocols can direct copolymer synthesis, but the precision with which monomers can be positioned in the copolymer is inherently limited [61]. Processes that use templates offer more control over the polymer sequence, but as highlighted in a recent review, most implementations far under-perform biological copying systems, limited by low accuracies, low yields, low abilities to reuse templates, and short templates with low information contents [62], and, I would add, rely on protocols that are non-autonomous (e.g. cannot run independently in a constant environment). Therefore, understanding the principles of autonomous template copying could provide routes toward efficient, highly sequence-selective and, fundamentally, simpler alternatives to multi-step reaction protocols [63]. Furthermore, understanding the dynamics of copying systems could pave the way to creating truly autonomous self-replicating and computing systems, which could serve as test-beds for developing our understanding of the dynamics of evolution and molecular computation in simple model systems. It is not infeasible that these systems could be coupled to create sequence-controlled products that result from computations with a self- replicator's growth dynamics.

1.0.3 Product inhibition and experimental approaches to templated reactions

Avoiding product inhibition is essential when creating systems that autonomously copy templates or undergo templated self-replication [64, 65, 66, 67]. Product inhibition describes the tendency of the product of a reaction to interact with the catalyst or template, causing the reaction to slow down. In a templated copolymerisation reaction, as neighbouring monomers polymerise, the number of shared bonds between the copy and template increases and the affinity of the growing copolymer for the template grows. This is a cooperative effect; the increasing affinity between the copy and template can exponentially suppress the rate of detachment. Product inhibition poses challenges for the sequence and length control of template copying reactions, as an accurate copy will have a higher affinity for the template, and hence will be likely to detach more slowly. Furthermore, given that longer polymers will detach more slowly than shorter ones, it may be challenging to prevent the early release of partial copies (posing challenges for length control), and it may be harder still to ensure the separation of long copies such that the template can be reused.

Recent work demonstrated that the inhibiting effects of a DNA polymerisation reaction could be overcome if a helicase was used to drive the product off of the template [68]. This work clearly demonstrates that some degree of energy input is required to overcome the effects of product inhibition. Various strategies have been employed to mitigate the effects of product inhibition in templated polymerisation reactions. Outside of the field of template copying, templated reactions have been used to detect and respond to the presence of certain nucleic acid biomarkers [64]. In the presence of a biomarker (the template or catalyst for the reaction), the

reaction between chemical groups with template-complementary tags is accelerated, indicating the presence of the biomarker in solution. Here the turnover, the number of successful completions of the reaction, is an important measure of sensitivity and is strongly suppressed by product inhibition. Nucleic acid templates and tags have been widely used [69, 70], but other molecules, such as proteins and antibodies, can also be used as templates and tags [71, 72]. Ligation or polymerisation reactions result in a conjoined product that necessarily has a higher affinity for the template than the substrates, and hence these reactions usually suffer from product inhibition under isothermal conditions [73, 64]. Even modifying the product post-ligation to reduce template affinity, by cyclization for instance, has proven somewhat ineffective [74, 64]. In seeking to increase the sensitivity of templated reactions to the presence of biomarkers, researchers turned to templated reactions that avoid ligation altogether and demonstrated that higher turnovers could be achieved with transfer or cleave reactions where the products have a similar or reduced affinity for the template [75, 76]. Enforcing separation of the product without compromising template reuse remains challenging; in one example, researchers sought to remove the template-complementary tags from the product as the ligation reaction occurred [77]. While the product was no longer complementary to the template, the binding sites on the template retained the tags; hence, the template was unusable for further reactions [77]. Product inhibition can have detrimental effects on the reaction in isothermal settings, even on the short templates used in the works described here.

Time-varying, non-autonomous environmental conditions have been used to drive cyclic templating and product release in non-enzymatic systems. A similar strategy is used in polymerase chain reaction (PCR), where the amplification of a DNA sequence is achieved through cycles of heating, to dissociate the product from the template, and cooling, to enable primer rebinding and subsequent growth [78]. Under time-varying heat and UV cycles, He et al. were able to drive the self-replication of dimeric DNA-origami raft structures and modulate replication rates of two subspecies by controlling the system's pH [79]. Interestingly, this system was extended such that each template could generate a "litter" of offspring at each replication round, rather than a single product per template per generation, as is usually the case in template copying reactions [80]. Cyclically driven, enzyme-free systems that can copy or replicate longer templates have also been developed [81, 82, 83, 84, 85]. In an early work, the self-replication of a system of palindromic DNA duplexes was achieved by cycling the pH and other conditions in the environment [86]. More recently, Núñez-Villanueva *et al.* developed a template-directed synthesis method based on covalent base-pairing that achieved robust information transfer through externally controlled processes [81, 82, 83, 84]. This method utilised the hydrolytic cleavage of ester bonds between the copy and the template to achieve separation of the products from the template [81].

Non-chemical sources of free energy and exploitable spatial gradients have also been used to drive the cyclical copying of longer templates [87, 78, 88, 89, 90, 91]. Otto and colleagues created a self-assembling

system that formed long, homegenous fibers, that could undergo exponential self-replication if they were broken by mechanical forces [92, 91]. Schulman et al. exploited shear flows to separate the layers of information-bearing DNA tile-based crystal, exposing more templating surfaces and enabling exponential self-replication of the crystal [87]. Inspired by the thermal gradients created in the fissures in the rocks around deep-sea hydrothermal vents [93], Braun and colleagues have harnessed the thermally-generated convective currents to design systems where templating and polymerization occur in cooler regions, while separation takes place in hotter regions and products accumulate in regions of low flow velocity [78, 88, 89, 94].

Spatially and/or temporally varying environmental conditions are believed to have played a crucial role in driving the earliest self-replicators in the emergence of life [95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106] however, it is important to note that modern biological copying enzymes do not rely on these conditions for their operation and can operate driven by chemical fuel molecules in constant environments [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19].

In these constant environments, product inhibition is known to prevent exponential growth of replicators [65, 66, 67], and, hence, open-ended molecular evolution [107]. A famous example of this was created by von Kiedrowski and colleagues, in which an oligonucleotide-based replicator suffered from product inhibition and grew with a parabolic, not exponential, growth curve [65]. In another famous work, Lincoln and Joyce developed an RNA-based system in which RNA substrate strands undergo a ribozyme-template mediated ligation to form another riboligase, which was capable of exponential replication [108]. Variants of the system with lower capacities for the product and substrate-mediated inhibition of the RNA-zymes were more effective replicators.

1.0.4 Handhold-mediated strand displacement for autonomous template copying reactions

A true minimal analogue of a biological template copying system would be capable of producing and releasing many copies of a long, information-bearing template copolymer in a constant environment, driven by chemical fuel molecules alone, unlike the vast majority of the experimental approaches described in the previous section. Our lab members have been working toward creating such a system in the past few years. Cabello-Garcia and Ouldridge developed a reaction motif based upon DNA strand displacement reactions, handhold-mediated strand displacement, that can copy dimeric templates under the conditions prescribed above, shown in Fig. 1.6 [22, 21].

The dimer copying system consists of two monomers and a template. Monomer 1 contains three domains: a short information domain, a longer backbone domain, and a very short toehold domain. The backbone and toehold domains are initially blocked by a "lock" strand, creating a dsDNA region containing some mismatches. The information domain is complementary to an information domain on the template. The second monomer

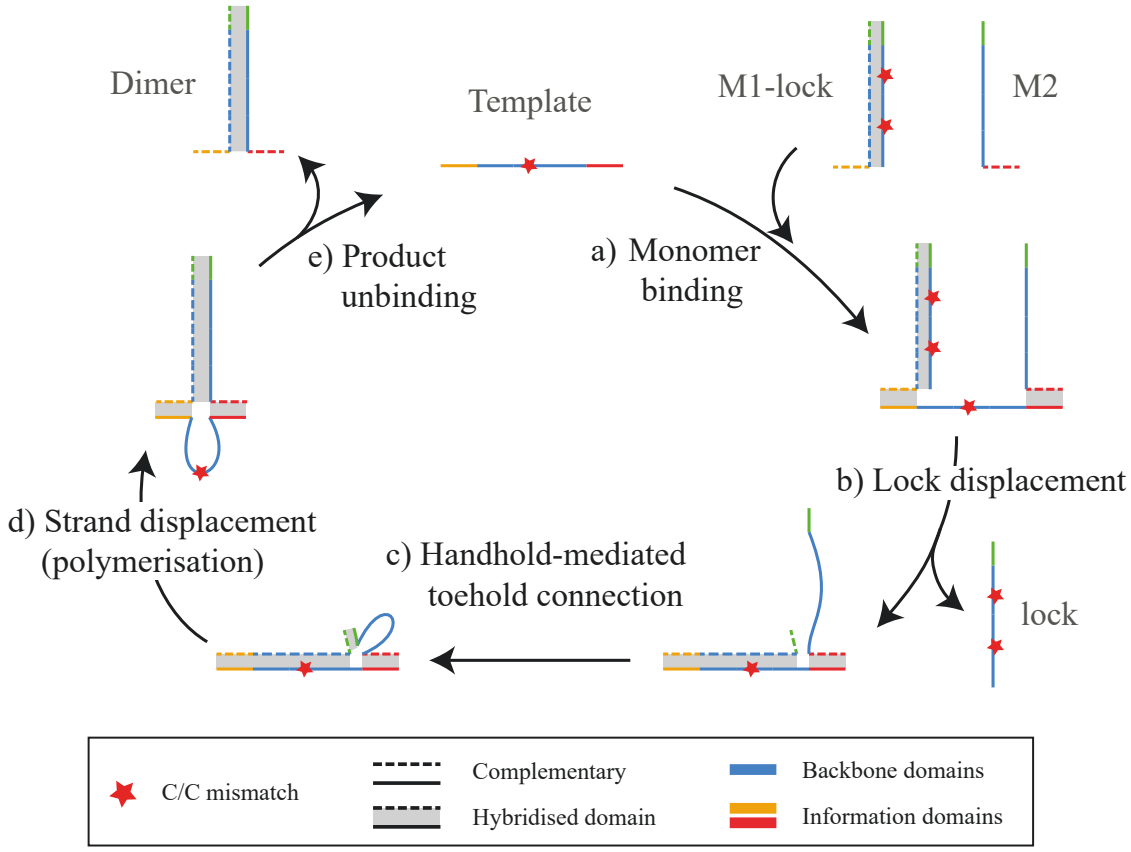


Figure 1.6: A DNA strand displacement reaction cycle that produces dimeric copies of a template using the "handhold mediated strand displacement" mechanism [22, 21]. The template has three domains; two information-bearing domains (yellow and red), and a "backbone" domain (blue). Monomer 1 is initially hybridised with a lock strand that prevents polymerisation in solution. Monomer 1 has a complementary information domain (yellow), a complementary backbone domain and a toehold domain. Monomer 2, has a complementary toehold, a backbone domain, and another complementary information domain. **a)** The monomers bind to the template from solution. **b)** Monomer 1 and the template bind completely, displacing the lock, and revealing the toehold. **c)** Due to their spatial proximity, the toeholds connect quickly. **d)** Formation of the toehold triggers another strand displacement that "polymerises" the monomers, forming a dimer with the same information domain sequence as the template. A mismatch is healed in the process, increasing the irreversibility. **e)** The dimer has few connections with the template, and detaches, enabling the next round of copying.

consists of complementary toehold and backbone domains, and a distinct information domain. The template consists of information domains complementary to each monomer and a backbone domain complementary to the domain in the first monomer. In the absence of the template, the rate of polymerisation in solution is very slow as the toehold domain of monomer 1 is occluded by the lock strand. In the presence of the template, the monomers bind to the template first by the information domains. The binding of monomer 1 displaces the lock strand, healing some mismatches and revealing the toehold. The template thus increases the polymerisation rate between the monomers by increasing their spatial proximity and revealing complementary domains. The monomers polymerise together by strand displacement, which heals another mismatch. In the polymerisation reaction, the number of connections that monomer 1 shares with the template decreases, reducing the affinity of the dimer product for the template. This leaves the dimer bound to the template by only the short informational domains, which expedites its separation. The system converts high chemical free energy monomers into sequence-defined dimers that suffer less from product inhibition. The concentrations of the monomers in solution, the dimers in solution and the healing of a number of mismatches in the process of forming the dimer drive the reaction cycle.

The system can discriminate between monomers and hence can produce accurate copies of the template. An incorrect monomer's information domains are not complementary to the template and therefore they can only bind transiently to the template, decreasing the chance that the backbone domain binds with the template by displacement of the lock. Therefore, the system is also fairly sequence selective and can accurately copy dimeric templates in a solution of mixed monomers, as shown in Cabello-Garcia's thesis [21].

Another benefit of the system is that, like the copying architectures found in life and unlike the bonds between components in self-assembling systems, the links between polymerised monomers are independent of the information-bearing part of the molecule. The backbone domains are common amongst different monomers. As illustrated in Fig. 1.7, on a template with three information-bearing domains, a second distinct type of backbone domain and toehold domain is required to avoid self-complementarity of the second monomer (M2). For templates with four or more information-bearing domains, a maximum of four types of monomer are required (of which each can have multiple variations with distinct information-bearing domains); two special monomer types are required for the first and last monomers in the sequence, and two alternating monomer types comprise the rest of the product. Therefore this system has the potential to exploit the combinatorial nature of the template to explore large, template-encoded chemical spaces with a small set of common monomers. For instance, all possible permutations of templates of length N with m types of information-bearing domains (that's $N \log_2 m$ bits of information) could be copied with just $4m$ monomer types, as illustrated in Fig. 1.7. The monomer set requires just a factor of 4 more monomer types than biological systems, two of which constitute the majority of the mass of the product (the intermediate alternating monomer types), especially for longer templates.

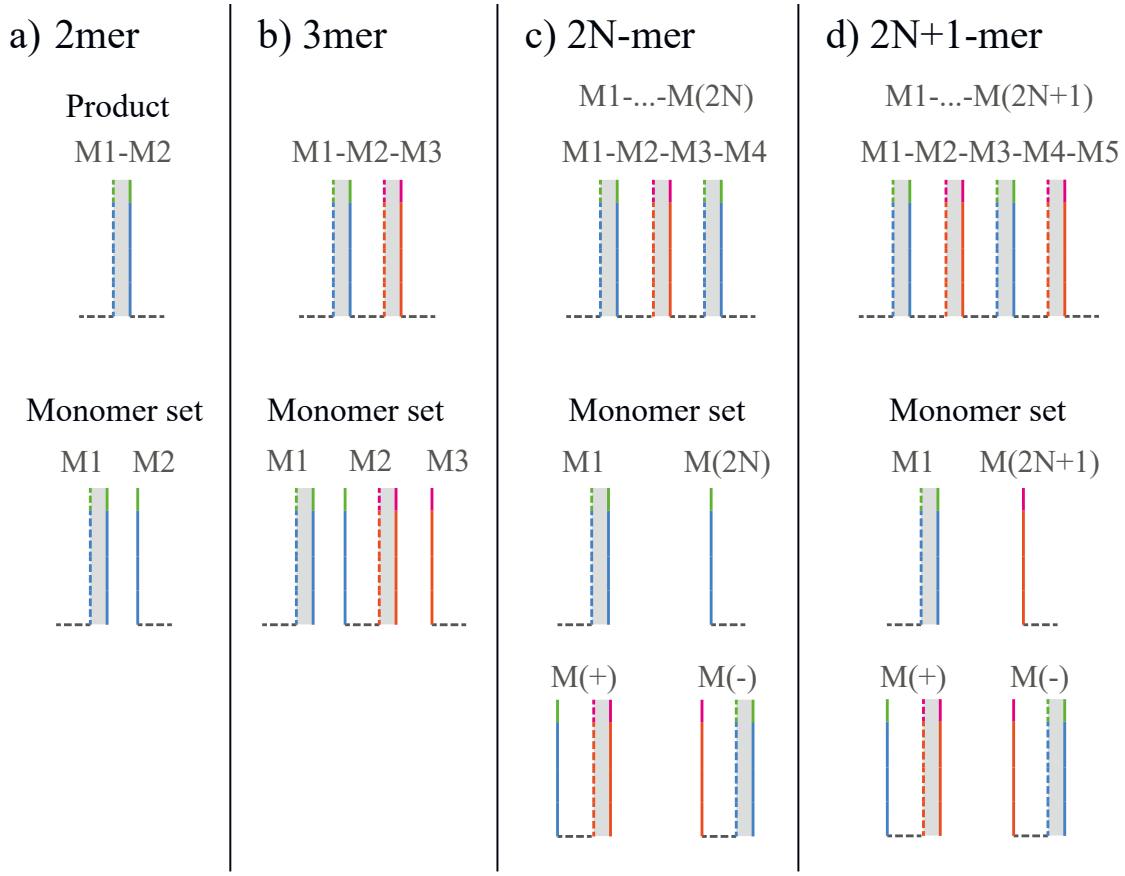


Figure 1.7: Handhold mediated strand displacement-based templating systems require at most four types of monomeric unit (of which each can have varying information domains – shown in grey here). Colours indicate complementarity. Shaded sections indicate a hybridised domain. **a)** On a template with two information domains, 2 monomer types are required. **b)** Three types of monomers are required on a template with three information domains. An alternative backbone and toehold domains shown in orange are required to prevent self-complementarity of monomer 2. **c,d)** A maximum of four types of monomers are required on a template with four or more information domains. Two monomers, M(+) and M(-), would comprise the majority of the mass of the product, and monomers M1 and M(2N) or M(2N+1) would cap the start and end of the polymer. Different end monomers M(2N) or M(2N+1) are required for templates with an even or odd number of information domains. Copying a template of length N, that contains an information domain alphabet of size m would require at most $4m$ types of monomer in this system.

A bond-weakening mechanism is central to this reaction motif; as the monomers polymerise, the number of connections they share with the template decreases, favouring product release. These mechanisms could theoretically present a route toward minimising product inhibition and achieving autonomous templated copying. The details of how such a system scales, questions related to the dynamics of these processes on longer templates and how that affects length and sequence control remain unanswered. Addressing some of these questions through coarse-grained modelling is the focus of Chapters 3 and 4 of this thesis.

1.0.5 Length distribution control and the modelling of templated copying systems

Approaches to modelling templated copying processes have, until recently, largely focused on the accuracy with which copies grow while remaining bound to templates, and have neglected to consider the effects of subsequent copy separation and length-control of the products [109, 110, 111, 112, 113, 114, 115, 116], as depicted in Fig. 1.8a. Such works more closely resemble DNA replication; however, the thermodynamics are altered by neglecting the full cycle of replication, which involves the inevitable separation of the DNA from the parent strand by helicases. As discussed in section 1.0.0, the requirement for copies to eventually separate from their template means that complementary interactions between monomers and the template can only be temporary [26], and therefore cannot contribute to the overall thermodynamics of polymer growth. Therefore, all accuracy generation must be kinetic, in stark contrast to models in which the interactions are permanent and accurate copying can be achieved in (non-kinetic) energetic regimes [114].

In a system with a binary monomer alphabet, the entropic cost per step of growth of incorporating an accurate monomer into the chain (as opposed to a random monomer), is $\ln 2$ (assuming the thermal energy unit is 1) [23]. Therefore, if the copolymer remains bound to the template and the free energy change per net step of growth of a random copolymer is ΔG_{pol}, and the free energy difference between a complementary and non-complementary copy-template interaction is ΔG_{disc}, then the free energy change per step for an accurate copolymer is $\Delta G_{\text{pol}} + \ln 2 - \Delta G_{\text{disc}}/2$ [23]. If $\ln 2 < \Delta G_{\text{disc}}/2$, then accurate copying can be more energetically favourable than growing a random copolymer if separation does not occur.

Poulton was the first to study a dynamical model of autonomous template copying with separation on an infinite length template [23], building upon the mathematical formalism described in [113]. In Poulton's model, the copolymer grows in a step-wise manner on an infinite template, the final release of the copolymer into the solution is not considered, and the precise control over the length of the product is implicit. In this model, illustrated in Fig. 1.8b, each step of growth involves the binding of a monomer and the detachment of the previous monomer in the chain. Therefore, the free energy change of incorporating a correct monomer after a

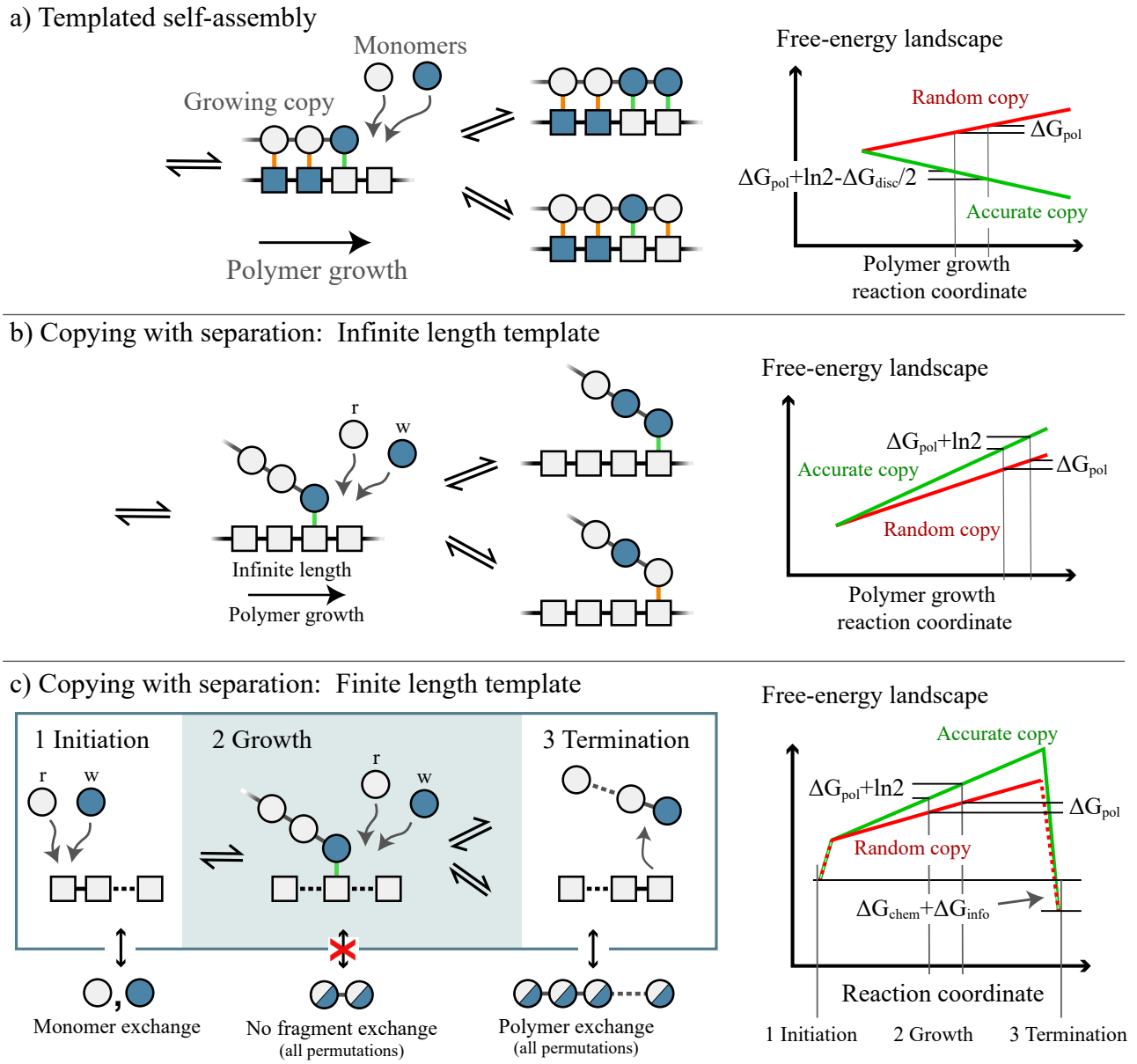


Figure 1.8: Illustrations of template copying processes. Monomers are white and blue circles. The template is a sequence of white and blue squares. Copy-template bonds are shown in orange and green depending on their complementarity. Illustrative free energy diagrams are shown, with the free energy change per step along the reaction coordinate indicated. **a)** Templated self-assembly, in which a copolymer grows while remaining bound to a template, has been studied extensively [109, 110, 111, 112, 113, 114, 115, 116]. As shown in the free energy landscape, growing an accurate copolymer can be exergonic if the discrimination energy is sufficiently strong. **b)** The dynamics of templated copying with separation was first studied in Poulton, ten Wolde, and Ouldridge [23]. The model assumes a homogeneous, template of infinite length, and considers step-wise growth in the presence of right and wrong monomers. The growing chain maintains one bond with the template. A generalised model was recently developed in Qureshi et al. [117] with arbitrary reaction networks between each step. **c)** Illustration of the model considered in Poulton and Ouldridge [24], featuring initiation, growth and termination or detachment of the copy. On a finite template, the exchange of monomers and completed polymers (but not partial polymers) sets the overall thermodynamics of the process, $\Delta G_{\mathrm{chem}} + \Delta G_{\mathrm{info}}$, with chemical and informational components which are a function of the difference in monomer and polymer bath sequence distributions[24]. Thermodynamic constraints for an infinite-length template (b) become kinetic barriers for a finite-length system.

correct monomer, or an incorrect monomer after an incorrect monomer, is just ΔG_{pol}. The free energy change of a wrong monomer following a right monomer is $\Delta G_{\text{pol}} + \Delta G_{\text{disc}}$ and the free energy change of a right monomer following a wrong monomer is $\Delta G_{\text{pol}} - \Delta G_{\text{disc}}$. Thus on average, the free energy per step of growing a random copolymer is ΔG_{pol} and is $\Delta G_{\text{pol}} + \ln 2$ for an accurate copy. Free energy above that which is required to grow a random copolymer must be used to compensate for the low entropy of the accurate copy sequence. Further, in the entropic growth regime, near the minimal required driving $\Delta G_{\text{pol}} = \ln 2$, intra-copy correlations, created by the separation between the growing copy and the template, reduce the sequence entropy and must also be paid for [23].

In subsequent work [24], illustrated in Fig. 1.8c, Poulton demonstrated that copying with autonomous separation on a finite template, where exchanges between the monomer pools and product pools in solution are considered, the thermodynamic constraints that hold for an infinite length template become kinetic constraints. Again, the release of partially grown copies was not considered. As the copies are released and mixed into the solution, the sequence entropy of the individual copolymer is not exploitable, and instead, the overall thermodynamics of the process is set by the sum of two terms, ΔG_{chem}, the average chemical free-energy change of polymerisation irrespective of the copolymer sequence, and ΔG_{info}, an informational term, the consequences of which are that copying a template to produce a copolymer that is like (unlike) the distribution of copolymers and unlike (like) the distribution of monomers in solution is more energetically costly (less costly). In the final chapter of their PhD thesis, Poulton analysed a system consisting of a template copier and template destroyer working in tandem and explored how dissipative degradation pathways can contribute to increasing the accuracy of the copier [27].

Qureshi, in collaboration with Poulton, Ouldridge, Beersing-Vasquez, and me, recently led this investigation further, by considering the effects of kinetic proofreading on the growth process [117]. Kinetic proofreading is a type of reaction motif used widely in biology, in which superfluous, dissipative cycles in the reaction network increase the ability of the system to discriminate between substrates [118, 119, 110]. Qureshi developed a method to solve a wide class of copolymerisation models, and demonstrated, among other results, that kinetic proofreading can affect the thermodynamic thresholds required for growth on infinite-length templates. Again, this model implicitly assumed that length control of the copolymers was possible.

The length distribution of products in templated reactions and the effects of product inhibition have been more directly examined by researchers exploring the RNA-world hypothesis. The RNA world hypothesis proposes that RNA or RNA-like molecules played a central role in the emergence of the earliest living systems as these molecules could act as both a store of genetic information and a catalyst of reactions, sustaining the replication of proto-biotic systems [95, 96, 97, 98, 100, 120, 121, 103, 99, 104, 105, 106]. The processes by

which an RNA world, consisting of long ($>$ 100 bases) oligonucleotides with specific sequences that confer catalytic properties, could have emerged from a pre-RNA world soup of short oligonucleotides [122, 123, 104, 105, 106, 124, 125] and that subsequently stabilised their replication [126] have recently been studied. What were the dynamics of Life's first exploration of the vast chemical spaces encoded in oligonucleotide templates?

Template-directed synthesis and templated ligation reactions have been proposed as one such route toward the elongation of oligonucleotides [122, 123, 126, 104, 105, 106, 124, 125]. In these models, the gradual lengthening of oligos is usually observed, and, in sequence-dependent models, the complementarity of interactions between oligos, sequence-dependent ligation rates, and stacking energies have all been implicated in observations of spontaneous symmtery breaking toward the formation of pools of oligos with greatly reduced sequence entropies [122, 123, 126, 104, 105, 106]. It is thought that these sequence-controlled oligo pools could have served as an enriched basis for the emergence of the first ribozymes. Indeed, a recent preprint is exploring this crucial next step, the transition from informational systems to those with additional catalytic properties [127]. These models typically invoke environmental cycling to drive these processes to overcome product inhibition, assume that secondary structures in the oligonucleotides do not form and that long polymers dehybridise in a single step when the environmental conditions are conducive, rather than modelling the dynamics of the tails of polymers [122, 123, 126, 104, 105, 106, 125]. Thus, these fascinating models, shed light on the ability of templates, driven by external cycles, to generate pools of molecules with non-random length and sequence distributions. However, the environmental conditions simulated in these models are different to the constant spatio-temporal conditions in which extant biological copying machines operate, and therefore, in asking fundamentally different questions, they shed little light on the mechanisms that might give rise to length control and the prevention of product inhibition for autonomous copying systems.

Taking inspiration from the replication cycles of viruses, Tupper and Higgs argued that non-enzymatic, rolling-circle RNA synthesis could have sustained the replication of RNA molecules in an RNA-world [124]. Here the dynamics of the tails were considered in a simple model. Short RNA strands can be displaced by longer strands on linear templates, leading to product inhibition at constant temperatures. However, oligos that grow on circular templates by extending either end of the same strand, suffer less the effects of product inhibition. If the oligo sequence contains a self-cleaving ribozyme sequence, then long overhanging tails would spontaneously sever themselves, potentially producing linear oligos with the same length as the circular template. A wider point that can be gleaned from this work is that if a polymer can remain attached to the template for long enough (here by displacing itself through synthesis, and thus not changing the number of bonds it makes with the template), but can still have some mechanism for release (here, cleavage), then faithful copying of the template sequence is possible.

1.0.6 Aims

Autonomous templated copying systems are the foundation of the generation of the chemical complexity of extant living systems and could serve as a basis for the creation of sequence-defined copolymers (a "holy-grail" of polymer science [55]), and self-replicating and information-processing molecular systems, among other applications. The work presented in this thesis, is a contribution toward the goal of creating a synthetic templated copying system that functions autonomously, as extant living systems do. The conditions for this system are that it operates in a constant environment, driven by chemical fuel molecules, and can produce exact copies of a template copolymer that spontaneously detach once formed. As I have discussed, product inhibition, the tendency for the products of a templated reaction to rebind to the template that formed it, has stymied efforts to create such systems in the past. Product inhibition increases with the length (and hence, informational content) of the template, and, thus, most copying systems created to date have been short and driven by enzymatic or external processes. Our group has recently developed an experimental system that can copy dimeric templates [22, 21], though the dynamics of the system on longer templates have yet to be demonstrated. Theoretical models of templated copying systems have largely neglected the consequences of autonomously separating the copies from the template [109, 110, 111, 112, 113, 114, 115, 116]. More recent models have accounted for product separation and assumed that the length control of products was possible under the proposed reaction mechanisms, instead focusing on the kinetic and thermodynamic factors that affect sequence-control [26, 23, 24, 27, 117]. In this thesis, I investigate the crucial missing aspect of these models, the mechanisms by which precise control over the length of copies can be achieved.

1.1 Summary of sections

In Chapter 2, I investigate a templated dimerisation reaction as if it were a bi-uni enzymatic reaction and explore suitable metrics for quantifying the performance of such a reaction, in which substrate concentration depletes over time and the rising concentration of product can cause product inhibition, further slowing the reaction. I propose the use of the time-averaged turnover frequency (TATF), which can also be expressed as the average rate of the reaction per unit concentration of the template. I examine ordered reaction mechanisms with competitive and mixed product inhibition patterns, as well as a random order mechanism with competitive inhibition. By assuming that the system evolves under a pseudo-steady state approximation, I calculate their corresponding TATF. By approximating the TATF in low, medium, and high initial substrate concentration regimes, I derive simple ratios of kinetic constants that capture the performance of the system. Interestingly, while calculating the TATF for the random order mechanism, the complexity of the system is significantly reduced, and the behaviour

of the system can be described using just four effective kinetic parameters (as opposed to the original 16). In this case, the effects of substrate depletion and product inhibition are made explicit and accounted for in the metric, which serves as a useful guide in experimental settings and the design process.

Increasing the length of the template enhances the amount of information that can be stored and copied, but also amplifies the effect of product inhibition. Moreover, the number of states that must be modelled to capture the effects of inhibition grows rapidly, which challenges the analytical approach of Chapter 2. Therefore, in Chapter 3, I develop a bottom-up model of a templated copying process that can simulate the growth of multiple polymers on an isolated template with an arbitrary length. This work was published in Ref. [1]. I disregard the product inhibition arising from the rebinding of products from solution and instead concentrate on the inhibiting effects of cooperative binding of long polymers growing on the template. The model is stochastically simulated using the Gillespie algorithm [128]. I search for mechanisms which, firstly, control the length distribution of products and, secondly, ensure the accuracy (length and sequence) of the products. Naïve mechanisms in which longer polymers bind cooperatively to the template cannot produce length-controlled polymers due to product inhibition. By coupling the energy released during polymerisation to the disruption of copy-template interactions, certain conditions can induce stable conformations in which multiple polymers are attached to the template at one of their ends in an orderly brush, which may be further extended by the binding of monomers. A directionality in the underlying mechanism induces a global directionality in the conformation of polymers on the template. By weakening the terminal site of the end of the template, the unbinding rate of full-length copies can be increased relative to incomplete polymers, enabling direct control over the length distribution of products. Lastly, I investigate whether this mechanism for length control can also generate accurate copies of the template by introducing sequence-specificity to the process. I discover that mechanisms that modulate the binding rates of correct and incorrect monomers are more effective than those which modulate the monomer unbinding rates. This body of work supports the notion that these mechanisms could successfully implement an autonomous templated copying reaction in a constant environment, generating numerous copolymers that share the exact sequence of the template.

In Chapter 4, I ground these mechanisms in an explicit model of a DNA strand displacement circuit that relies on HMSD and examine how design parameters (such as domain lengths) affect the expected performance of the system. I aim to identify optimal parameters for implementation in experimental settings. In this analysis, I allow for the rebinding of products to the template and incorporate some of the most likely off-target or leak reaction pathways into the model. Optimal performance entails maximising the yield of products, minimising the time taken to produce said products, maximising the ratio of full-length to partial-length products, and minimising sequence errors. I deterministically simulate the system for a trimeric template, which is our current

experimental target. I then analyse the factors that each independently contribute to optimal performance and, ultimately, explore globally optimal regions of the parameter space.

1.2 Methods: Construction of thermodynamically self-consistent coarse-grained models

Throughout this thesis, I used an approach to creating coarse-grained models that ensures thermodynamic self-consistency, which is described in an approachable, pedagogical review by Ouldridge [129]. I follow a procedure in which I define the system's states, the energies of those states, and then the transition rates between those states. I provide details on how these models can be treated deterministically or stochastically.

1.2.1 Defining States

The first step in constructing coarse-grained models is to decide on the level of detail to which the system will be described. Biophysical systems are composed of many atoms and molecules, each with physical states, microstates, that collectively give rise to a phenomenon of interest. These microstates are experimentally inaccessible [26]. Usually, we are interested in modelling processes that play out on larger scales, such as the concentration of interacting macromolecular species. We define macroscopic observables to describe these higher-level features of the system. These macroscopic observables define the macrostates of the system. A macrostate is a set of many microstates (molecular orientations, positions, energies) for which the macroscopic observable takes a particular value.

In this thesis, I investigate processes by which a copolymer grows on a template and study the sequences of the copolymer copies produced. Therefore, the macroscopic observables of interest are the number of monomers and polymers, whether monomeric links in these chains and between monomers and the template are formed; these define the macrostates. The microscopic features, such as solvent molecules, the orientation and position of molecules within each monomer, are swept up in the macrostates.

In principle, any choice of macrostates is valid, but often they are chosen such that the dynamics of the system on the macrostate level are Markov [129]. In this context, the Markov property means that the probability of transitioning between macrostates at any infinitesimal moment depends only on the system's current state and not the states that have been visited in the past. This condition can be met if the system is ergodic (all states are reachable given sufficient time), and if the transition between macrostates involves passing through a high free energy, and thus rarely visited, set of microstates on the barrier – transition states [129]. An illustrative free energy landscape showing two free energy basins corresponding to macrostates i and j and a transition

a) Free energy of macrostates

b) Tajectory over microstates between macrostates

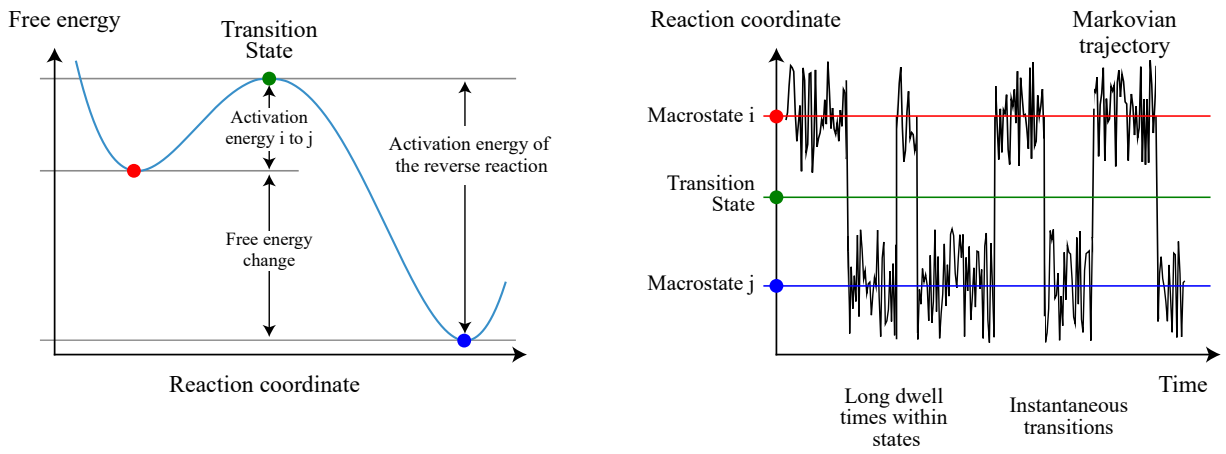


Figure 1.9: a) A hypothetical free energy landscape for a system that could be properly coarse-grained into macrostates i and j. Each macrostate is a collection of microstates for which an observable (such as the abundance of a chemical species) takes a certain value. **b)** A trajectory over microstates showing transitions between macrostates. Due to the high free energy barrier between states i and j, the duration of a transition between macrostates is typically much shorter than the dwell time between transitions. This means the system approximately equilibrates within a macrostate before making a transition. Together, these properties mean that the system can be described as a Markov process with instantaneous transitions with dwell times between transitions that are exponentially distributed [129].

state is shown in Fig. 1.9. If the timescale for transitioning between macrostates is shorter than the timescale over which the system equilibrates over the microstates within a macrostate, then the transition probabilities lose dependence on the sequence in which the macrostates were visited. A further consequence of this is that we arrive at transitions between states for which the time between transition events is much longer than the duration of a transition event itself, and thus we approximate the statistics of the transitions with waiting times that are exponentially distributed and with the transition events being instantaneous. The equilibration within macrostates and the instantaneous transition durations give rise to the Markov property. An illustration of a trajectory over the microstates of the system that satisfies these macrostate conditions is shown in Fig.1.9b. A more rigorous description of when this sort of timescale separation applies is given in Bo and Celani [130]. If the Markov property holds and the transitions between macrostates are instantaneous with exponentially distributed waiting times, then each transition from macrostate i to j can be associated with a rate $k_{j,i} \geq 0$. Thermodynamics constrains the values these transition rates can take, as I explain in the following subsection [129].

1.2.2 Principle of detailed balance

The equilibrium occupancy of a macrostate i is

$$P_\sigma^{\text{eq}}(i) = \frac{Z_\sigma(i)}{Z_\sigma} = e^{-(F_\sigma(i) - F_\sigma^{\text{eq}})/k_B T} \tag{1.1}$$

where Z_σ is the partition function of the system σ, $Z_\sigma(i)$ is the partition function of the macrostate i, $F_\sigma^{\text{eq}} = U_\sigma^{\text{eq}} - T\,S_\sigma^{\text{eq}}$ is the equilibrium free energy of the system, with internal energy U_σ^{eq} and entropy S_σ^{eq}, and temperature T. k_B is the Boltzmann constant, and $F_\sigma(i)$ is the free-energy of the macrostate i.

In equilibrium, the probability distribution over the system is stationary and, given rate constants $k_{i,j}$, all probability fluxes must balance, therefore

$$k_{j,i} P_\sigma^{\text{eq}}(i) = k_{i,j} P_\sigma^{\text{eq}}(j)\,. \tag{1.2}$$

This condition is referred to as the principle of detailed balance and it allows us to constrain the ratio of the rate constants as

$$\frac{k_{j,i}}{k_{i,j}} = \frac{P_\sigma^{\text{eq}}(j)}{P_\sigma^{\text{eq}}(i)} = \exp\left(-(F_\sigma(j) - F_\sigma(i))/k_B T\right)\,. \tag{1.3}$$

If the rate constants are chosen to satisfy the condition in Equation 1.3, then the model will have a stationary distribution equal to the true equilibrium distribution of the system. The following method ensures this is the case.

1.2.3 States, Free-energies, Transitions, and Rate constants

After having defined the system's macrostates, the free energies of each macrostate, $F_\sigma(i)$, should be calculated. Next, the valid transitions between pairs of states i and j, $k_{j,i}, k_{j,i} > 0$ should be identified. The rate constant for a reaction in one of the directions should be set, and equation 1.3 should be used to set the rate of the reverse rate constant. In any particular model, there may be many similar types of transition between different macrostates, such as unimolecular reactions (template mediated (de)polymerisation) or bimolecular reactions (binding and unbinding of monomers to/from the template). In this thesis, I have on occasion assumed that the polymerisation rate constant and the bimolecular binding rate constant are fixed quantities, and as a consequence, the depolymerisation and unbinding rate contestants are dependent on the free energy change of the reactions. A more subtle parameterisation might explore a balance of the rate constants' dependence on the free energy changes [129].

Another compatible approach, using transition state theory, assumes that the rate constant for a reaction is proportional to the occupancy of a short-lived transition state at the free-energy barrier between states

$$k_{j,i} \propto \exp\left(-E^A_{j,i}/k_B T\right) . \tag{1.4}$$

1.2.4 The small number limit: Discrete-state continuous-time Markov chain

In the limit that the number of particles in the system is small, the system can be treated as a discrete-state continuous-time Markov process. The state of the system is given by a column vector of macrostate occupation probabilities, $\mathbf{P}(t)$, where the i'th element is $P_i(t)$, the occupation probability of macrostate i at time t. The time evolution of the system obeys the following master equation

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{dt}} = \mathbf{K}\mathbf{P}, \tag{1.5}$$

where $\mathbf{K}$ is the rate matrix, with the diagonals equal to the negative sum of all rate constants of the transitions out of the state $\mathbf{K}_{i,i} = -\sum_i k_{j,i}$, and off-diagonal entries of the rows set by the rate constants for transitions into the state. The solution of the system is given by a matrix exponential, $\mathbf{P}(t) = \exp(\mathbf{K}t)\,\mathbf{P}(t=0)$, which can be solved by diagonalisation.

Stochastic trajectories through the state space can be sampled with Monte-Carlo simulation methods [129]. A common technique in biophysical settings to use the Gillespie algorithm [128]. In this algorithm, an initial state is set. The total rate of transitions out of the current state is calculated. A transition is selected by randomly selecting one of the transitions with a probability proportional to the rate of the pathway. The time increment is increased by an amount calculated by sampling an exponential distribution of waiting times, with a rate proportional to the total rate out of the current state. The process is iterated until a sufficiently long trajectory has been generated. The statistics of the process can be inferred by analysing ensembles of trajectories. This method is used in Chapter 3.

1.2.5 The large number limit: Continuous-state continuous-time Markov chain and the chemical kinetic equation

In the limit of large numbers of particles, we may be more interested in modelling the concentration of chemical species as continuous variables that obey a deterministic ODE model. The rate of each reaction in the forward and reverse directions $\pm\rho$ occurs with mass action kinetics [131, 132], rate constants $k_{\pm\rho}$ with fluxes

$$J_{\pm\rho} \equiv k_{\pm\rho} \prod_i C_i^{\nabla^i_{\pm\rho}} \tag{1.6}$$

where $\prod_i C_i^{\nabla^i_{\pm\rho}}$ is the product of the concentrations of species i that are involved in the reaction in either direction for which $\nabla^i_{\pm\rho} \neq 0$. Again, the principle of detailed balance holds at equilibrium and constrains the rate constants

$$k_{+\rho} \prod_i C_{i,\mathrm{eq}}^{\nabla^i_{+\rho}} = k_{-\rho} \prod_i C_{i,\mathrm{eq}}^{\nabla^i_{-\rho}} \tag{1.7}$$

$$\frac{k_{+\rho}}{k_{-\rho}} = \prod_i C_{i,\mathrm{eq}}^{\nabla^i_{+\rho} - \nabla^i_{-\rho}} = \exp(-\Delta F^0 / k_B T) \tag{1.8}$$

where ΔF^0 is the free energy change of the reaction measured at standard concentrations, $C_i = C^0$ The net flux for each reaction ρ is

$$J_\rho \equiv J_{+\rho} - J_{-\rho} = k_{+\rho} \prod_i C_i^{\nabla^i_{+\rho}} - k_{-\rho} \prod_i C_i^{\nabla^i_{-\rho}} . \tag{1.9}$$

By summing over all reactions, ρ, we find that the concentration of each chemical species, C_i, evolves in time according to the following chemical kinetic equation

$$\frac{\partial C_i}{\partial t} = \sum_\rho (\nabla^i_\rho) J_\rho = \sum_\rho \nabla^i_\rho \left(k_{+\rho} C_i^{\nabla^i_{+\rho}} - k_{-\rho} C_i^{\sigma \nabla^i_{-\rho}} \right) \tag{1.10}$$

where $\nabla^i_\rho = \nabla^i_{+\rho} - \nabla^i_{-\rho}$ is an element of the stoichiometry matrix for chemical species i and reaction ρ.

The systems of differential equations in Eqn. 1.10 can be numerically challenging to solve by standard numerical methods because they can be "stiff," exhibiting timescales that vary by many orders of magnitude due to the exponential dependence of the free energies on the rate constants. Fortunately, stiff differential equation solvers, like the Radau method, are accessible within standard programming libraries such as SciPy[133].

Chapter 2

Metrics for comparing templated ligation reactions

2.1 Introduction

Templated polymerisation reactions involve the template-catalysed polymerisation of substrates. At present, there is a lack of metrics for comparing the performance of these reactions that are suitable in the context of template copying chemical reactions. In this chapter, I argue for the use of the "time-averaged turnover frequency" (TATF) (a measure of the average rate of product formation per unit catalyst over a defined period of time) in the context of templated copying, and I give analytic approximations of the metric for three mechanisms describing copy growth on a dimeric template.

2.1.1 Nomenclature

Before delving into the various metrics that have previously been used to compare the performance of templated polymerisation reactions to date, I will give a summary of the relevant terms. Using Cleland's nomenclature, templated polymerisation reactions that involve the polymerisation of two, three, four, or "n" substrates into a single product can be described as "bi/ter/quad/n-uni" reactions [134, 135, 136]. Conventionally, the concentrations of reactants are labelled [A], [B], [C] and so on; the concentrations of products are labelled [P], [Q] and so on; the concentration of free catalyst is labelled [E], and the concentration of intermediate complexes of reactants and products are labelled as [EA], [EB], [EAB], [EP], [EQ] and so on. The reaction mechanism can be classified by the reactants' binding and unbinding pathways. If the reactants bind in a specific order (e.g. A then B), the mechanism is "ordered"; if the reactants can bind in any order, the reaction is "random"; "ping-pong"

reactions involve the release of a product molecule before all reactant molecules have bound.

The rate of reactions are typically analysed at "steady-state" [134, 135, 136]. Here the "steady-state" is one in which the substrate and product concentrations are fixed (chemostatted) and the concentration of the intermediate complexes (EA, EAB, EP etc.) have reached constant values, $\frac{\mathrm{d[EA]}}{\mathrm{dt}} = \frac{\mathrm{d[EAB]}}{\mathrm{dt}} = \frac{\mathrm{d[EP]}}{\mathrm{dt}} = \ldots = 0$. This is also sometimes referred to as an "intermediate steady-state" condition. I will give a simple example of the derivation of a steady state rate law for a uni-uni reaction with competitive product inhibition.

Consider the following simple uni-uni reaction scheme that converts one substrate molecule to one product molecule

$$\begin{aligned}
\mathrm{A+E} &\underset{k_{\text{-}1}[\mathrm{EA}]}{\overset{k_1[\mathrm{A}][\mathrm{E}]}{\rightleftharpoons}} \mathrm{EA} \\
\mathrm{EA} &\underset{k_{\text{-}2}[\mathrm{EP}]}{\overset{k_2[\mathrm{EA}]}{\rightleftharpoons}} \mathrm{EP} \\
\mathrm{EP} &\underset{k_{\text{-}3}[\mathrm{P}][\mathrm{E}]}{\overset{k_3[\mathrm{EP}]}{\rightleftharpoons}} \mathrm{P+E}
\end{aligned}$$

This set of reactions can be modelled by the following set of differential equations.

$$\frac{\mathrm{d[A]}}{\mathrm{dt}} = k_{\text{-}1}[\mathrm{EA}] - k_1[\mathrm{A}][\mathrm{E}]\,, \tag{2.1}$$

$$\frac{\mathrm{d[EA]}}{\mathrm{dt}} = k_1[\mathrm{A}][\mathrm{E}] + k_{\text{-}2}[\mathrm{EP}] - (k_2 + k_{\text{-}1})[\mathrm{EA}]\,, \tag{2.2}$$

$$\frac{\mathrm{d[EP]}}{\mathrm{dt}} = k_2[\mathrm{EA}] + k_{\text{-}3}[\mathrm{P}][\mathrm{E}] - (k_3 + k_{\text{-}2})[\mathrm{EP}]\,, \tag{2.3}$$

$$\frac{\mathrm{d[P]}}{\mathrm{dt}} = k_3[\mathrm{EP}] - k_{\text{-}3}[\mathrm{P}][\mathrm{E}]\,. \tag{2.4}$$

The concentration of the template or enzyme satisfies the conservation law $[\mathrm{E}_0] = [\mathrm{E}] + [\mathrm{EA}] + [\mathrm{EP}]$, where $[\mathrm{E}_0]$ is the total concentration. In the intermediate steady-state, the differential equations satisfy

$$\frac{\mathrm{d[EA]}}{\mathrm{dt}} = 0 = k_1[\mathrm{A}][\mathrm{E}] + k_{\text{-}2}[\mathrm{EP}] - (k_2 + k_{\text{-}1})[\mathrm{EA}]\,, \tag{2.5}$$

$$\frac{\mathrm{d[EP]}}{\mathrm{dt}} = 0 = k_2[\mathrm{EA}] + k_{\text{-}3}[\mathrm{P}][\mathrm{E}] - (k_3 + k_{\text{-}2})[\mathrm{EP}]\,, \tag{2.6}$$

$$\tag{2.7}$$

which can be used, in conjunction with the enzyme conservation law $[\mathrm{E}_0]$, to calculate the rate of production of product,

$$\frac{\mathrm{d[P]}}{\mathrm{dt}}\Big|_{\mathrm{iss}} = \frac{[\mathrm{E}_0](k_1 k_2 k_3[\mathrm{A}] - k_{\text{-}1}k_{\text{-}2}k_{\text{-}3}[\mathrm{P}])}{k_2k_3 + k_3k_{\text{-}1} + k_{\text{-}1}k_{\text{-}2} + k_1(k_2 + k_3 + k_{\text{-}2})[\mathrm{A}] + k_{\text{-}3}(k_2 + k_{\text{-}1} + k_{\text{-}2})[\mathrm{P}]}\,. \tag{2.8}$$

Equation 2.8, and other rate laws, are typically expressed in terms of kinetic constants

$$\frac{\mathrm{d[P]}}{\mathrm{dt}}|_{\mathrm{iSS}} = \frac{[\mathrm{E_0}]k_{\mathrm{cat}}\left([\mathrm{A}] - \frac{[\mathrm{P}]}{K_{\mathrm{eq}}}\right)}{[\mathrm{A}] + \frac{K_{\mathrm{A}}[\mathrm{P}]}{K_{\mathrm{P}}} + K_{\mathrm{A}}}, \tag{2.9}$$

where the maximum rate per unit enzyme of the reaction in the A to P direction that is achieved under saturating substrate conditions is

$$k_{\mathrm{cat}} = \frac{k_2 k_3}{k_2 + k_3 + k_{\text{-}2}}, \tag{2.10}$$

the equilibrium constant of the reaction is

$$K_{\mathrm{eq}} = \frac{k_1 k_2 k_3}{k_{\text{-}1} k_{\text{-}2} k_{\text{-}3}}, \tag{2.11}$$

the Michaelis constant for the substrate A (that corresponds to the concentration of substrate A in the absence of product P that would give half the maximum reaction speed) is

$$K_{\mathrm{A}} = \frac{k_2 k_3 + k_3 k_{\text{-}1} + k_{\text{-}1} k_{\text{-}2}}{k_1 (k_2 + k_3 + k_{\text{-}2})}, \tag{2.12}$$

and the Michaelis constant of the product (that corresponds to the concentration of product P in the absence of substrate that would give half the maximum reaction speed in the reverse direction P to A) is

$$K_{\mathrm{P}} = \frac{k_2 k_3 + k_3 k_{\text{-}1} + k_{\text{-}1} k_{\text{-}2}}{k_{\text{-}3} (k_2 + k_{\text{-}1} + k_{\text{-}2})}. \tag{2.13}$$

As we can see in Eqn. 2.9, even in the limit that the reaction is irrersibly driven ($K_{\mathrm{eq}} \to \infty$), the formation of product can slow down the rate of the production as the product sequesters the free template E in the complex EP, and prevents the binding of substrate to form the complex EA. This is a type of product inhibition, and the effect on the rate of the reaction is captured by the term $\frac{K_{\mathrm{A}}}{K_{\mathrm{P}}}[\mathrm{P}]$ in the denominator of the the rate law in Eqn. 2.9. Typically, templated polymerisation reactions suffer from product inhibition, a tendency for products to interact with the catalyst, slowing the rate of product formation [65, 66, 67, 73, 137, 69, 75, 74, 64, 70, 76]. Next, I will explain the nomenclature used to describe different types of product inhibition.

Cleland classified several ways in which a product can inhibit the reaction [134, 135, 136]. If the presence of the inhibitor in solution competes with the reactants in binding to the unbound catalyst, thereby changing the apparent Michaelis constant of the substrate, but does not affect the maximum speed of the reaction, the pattern of inhibition is said to be competitive. If the inhibitor only binds to an intermediate substrate and catalyst complex, thereby reducing both the maximum speed of the reaction and the apparent Michaelis constant, the inhibition

pattern is “uncompetitive”. If the inhibitor can bind to the catalyst whether or not the substrate is bound, the inhibition pattern is described as mixed, the maximum speed will change, and the apparent Michaelis constant may change. If the inhibition pattern is mixed and the inhibitor can bind to the catalyst with the same affinity whether or not the substrate is bound, then the apparent Michaelis constant will not change, and the pattern of inhibition is described as “non-competitive”. Inhibition patterns provide signatures of the reaction mechanism that can be measured experimentally, and therefore, they have long been standard tools for enzymologists elucidating the reaction pathways in enzymes [134, 135, 136, 138, 139]. Also, many therapeutic molecules act by inhibiting enzymes, and hence understanding inhibition patterns and the kinetic effects of inhibition are important in medecine [140].

2.1.2 Review of metrics

Little attention has been given to bi-uni reactions in the literature [141], despite the many decades since the first formulation of the kinetic rate laws [134, 135, 136, 138, 139] and their relative algebraic simplicity compared to higher order product mechanisms, such as bi-bi reactions [140, 142]. Indeed, to aid enzymologists in the process of identifying the kinetics mechanisms of bi-substrate enzymes, some works have near-exhaustively listed all the rapid equilibrium and steady-state rate laws for bi-bi reactions of a wide array of mechanisms (ordered, ping-pong, random, and so on), without product inhibition [140] and under product inhibition [142]. Ohs et al. [143] derived the steady-state rate laws for a branched bi-uni reaction mechanism, in which two substrates competed for incorporation into the product. This model treated the ligation and product release as a single-step reaction which would not be an accurate description of a templated ligation reaction in which the product may remain firmly bound to the template, and the polymerisation reaction may be faster than the rate of detachment. The rate laws for bi-uni reactions will generally have fewer terms than the bi-bi mechanisms.

In the Ouldridge lab, we are developing autonomous template copying systems. As we build and test these systems, we will need suitable metrics for measuring and comparing their performance. We require that these systems convert a maximal amount of substrates to product, do so rapidly and that each template can be reused many times. However, in creating a long polymeric copy, the copy may have a higher affinity for the template than the substrates and thus may suffer from product inhibition [65, 66, 67, 73, 137, 69, 75, 74, 64, 70, 76]. Therefore the metric that we use to compare the systems should measure the speed of the reaction and should account for the slowing effects of substrate depletion and product inhibition. Few metrics that have been used previously are well-suited in this regard, as I will discuss in the following paragraphs. For reference, an ordered bi-uni reaction comprised of the following reaction steps,

$$\begin{aligned}
\text{A+E} &\rightleftharpoons \text{EA} \\
\text{EA+B} &\rightleftharpoons \text{EAB} \\
\text{EAB} &\longrightarrow \text{EP} \\
\text{EP} &\rightleftharpoons \text{P+E} \\
\text{EA+P} &\rightleftharpoons \text{EAP}
\end{aligned}$$

with two reversible substrate binding steps, an irreversible polymerisation step, and mixed product inhibition, has the following production rate $\frac{\mathrm{d[P]}}{\mathrm{dt}} = v$ when the intermediate complexes (EA, EAB, EP, EAP) are in an intermediate steady-state ($\frac{\mathrm{d[EA]}}{\mathrm{dt}} = \frac{\mathrm{d[EA]}}{\mathrm{dt}} = \frac{\mathrm{d[EA]}}{\mathrm{dt}} = \frac{\mathrm{d[EA]}}{\mathrm{dt}} = 0$,

$$v = \frac{k_{\mathrm{cat}}[\mathrm{A}][\mathrm{B}][\mathrm{E}_0]}{(K_{\mathrm{iA}}K_{\mathrm{B}} + K_{\mathrm{A}}[\mathrm{B}])\left(1 + \frac{[\mathrm{P}]}{K_{\mathrm{icP}}}\right) + K_{\mathrm{B}}[\mathrm{A}]\left(1 + \frac{[\mathrm{P}]}{K_{\mathrm{incP}}}\right) + [\mathrm{A}][\mathrm{B}]}, \tag{2.14}$$

where $[\mathrm{E}_0]$ is the total amount of catalyst; [A], [B] and [P] are the concentrations of substrates and products; $[\mathrm{A}][\mathrm{B}][\mathrm{E}_0]k_{\mathrm{cat}}$ is the maximum reaction rate; K_{A} is the Michaelis constant of the reaction in the absence of product and under saturating concentrations of substrate B; K_{B} is the Michaelis constant of the reaction in the absence of product and under saturating concentrations of substrate A; K_{iA} is the inhibition constant of substrate A when substrate B is varied and, in this case, is also equal to the dissociation constant of substrate A and the unbound catalyst; K_{icP} is the competitive inhibition constant and, as the polymerisation reaction is irreversible, is also equal to the dissociation constant of the product for the unbound catalyst; and K_{incP} is the uncompetitive product inhibition constant which is also equal to the dissociation constant of the product and the substrate bound catalyst EA. A detailed derivation of Eqn. 2.14, is given in section 2.3.

The turnover, the turnover frequency, the turnover number and the total turnover number are all quantities based upon the "turnover" of the reaction, a measure of the number of completed cycles of the templated reaction, have been used to compare templated and enzymatic reactions [137, 69, 75, 74, 64, 70, 76]. These turnover-related quantities have different units and are often mistakenly equated as argued by Kozuch and Martin [144]. To disambiguate the terms and to set a standard for their use, Kozuch and Martin provided "catalyst-focused" (minimally affected by reactants and hence measured under saturating reactant conditions) definitions of the Turnover Frequency (TOF) and the Turnover Number (TON). The TOF was defined as the instantaneous time derivative of the number of turnovers of the reaction cycle per active enzyme site when the catalyst has such a low concentration to be infinitely dilute, the reactants and products are in saturation (if the saturating conditions can be met), and accompanied by a statement of the concentrations and temperature used. This definition of TOF reduces to the k_{cat} of the catalyst for a uni-uni enzymatic reaction unaffected by inactivation (degradation

of the template or enzyme) or inhibition. As stated by Kozuch and Martin, this definition is problematic when considering reversible reactions that suffer from product inhibition, as the reaction may even run in reverse under saturating conditions of both substrate and product, giving a negative TOF. Furthermore, for inhibited reactions, saturation may not be reachable under typical conditions, and therefore, TOF may be a poor indicator of the catalyst's performance in the setting of a templated reaction.

Kozuch and Martin defined the Turnover Number (TON) as the maximum number of turnovers of the catalyst throughout its usable lifetime, which, theoretically, is the integral of their definition of the TOF over an infinite period during which the reactant and product concentrations are held constant. While such a metric may be useful in industrial-scale, heterogeneous catalysis, it is less practically useful for comparing templated reactions. The primary reason for this is the aforementioned issues with the TOF not being a good measure of the performance of templated reactions. Secondly, precisely measuring and controlling the concentrations of reactants and products may be difficult in practice. Therefore, the "catalyst-focused" metrics that Kozuch and Martin propose are not well-suited for comparing templated reactions and have not been widely used in this field.

In templated reactions, where the effects of product inhibition can be minimised, for instance in templated reactions that result in the cleaving of substrate [76], and the maximum speed, k_{cat} can be reached under saturating substrate conditions, the turnover frequency, TOF, is a suitable metric for comparing the performance of reactions. However, when the templated reaction does suffer from product inhibition or saturating conditions cannot be met, the TOF as defined by Kozuch and Martin [144] is a less informative measure, and the total turnover number (TTN), measured as the number of turnovers of the catalyst per unit catalyst after a period given a certain initial substrate concentration and temperature, is used [137, 69, 75, 74, 64, 70, 76]. If the concentration of product produced during after a time t is $[\text{P}]_t$, the TTN can be expressed as

$$\text{TTN} = \frac{[\text{P}]_t}{[\text{E}_0]}\,. \tag{2.15}$$

The TTN can be pushed higher by reducing the catalyst concentration to a level where the reaction rate is just distinguishable over the background reaction rate [75]. Hence, the maximum TTN is also a function of the non-templated reaction. TTNs are often quoted at varying reactant and catalyst concentrations, temperatures and over different lengths of time making comparisons between TTNs challenging. Moreover, the TTN is not a mechanistic metric, it does not provide insight into the mechanism by which good performance is achieved. I note that, in a setting where the detection of a very low signal (template or enzyme) is the goal (not the rapid production of products), the TTN is a suitable metric, as it measures the sensitivity of the system [137, 69, 75, 74, 64, 70, 76].

Banta and Wheeldon used steady-state rate laws to derive metrics to compare the performance of enzyme mutants in Ref. [141]. However, their direct suitability for comparing reactions that slow over time due to product inhibition is limited. They identified ratios of the kinetic parameters that approximate the reaction rate in different substrate concentration regimes and cautioned against the use of inappropriate metrics such as $\frac{k_{cat}}{K_A}$ and $\frac{k_{cat}}{K_B}$. For instance, $\frac{k_{cat}}{K_A}$ reduces to the rate constant of binding of substrate A, and hence is not a good measure of the rate of the overall reaction in a standard bi uni reaction without product inhibition. The ratios $\frac{k_{cat}}{K_A}$ and $\frac{k_{cat}}{K_B}$ for bi-bi reactions are far less informative than the commonly used metric $\frac{k_{cat}}{K_M}$ for uni-uni reactions.

Instead, Banta and Wheeldon advocated for the use of the following kinetic metrics [141]. In the high substrate regime, the reaction is first-order and proceeds with a rate constant k_{cat}. At low substrate concentrations, the reaction is third-order and proceeds with a rate constant $\frac{k_{cat}}{K_{iA}K_B}$. For higher-order substrate reactions with ordered substrate binding, they argued for the use of metric $\frac{k_{cat}}{K_{iA}K_{iB}...K_Z}$ at low substrate concentration, where K_{iX} is the substrate inhibition constant of substrate X and K_Z is the Michaelis constant of the final substrate to bind Z. For reaction mechanisms where a steady-state rate law cannot be written in terms of simple kinetic parameters, such as a random bi-bi reaction, at least in the limit of low concentrations the sum over the two different binding pathways $\frac{k_{cat,A}}{K_{iA}K_B} + \frac{k_{cat,B}}{K_{iB}K_A}$ approximates the third order reaction rate constant, and hence could be used as a metric for comparison. While these metrics are useful for comparing performance in the steady-state in the absence of product inhibition, they are less useful for inhibited reactions. Product inhibition will tend to reduce the reaction rate over the course of the reaction and hence reactions with a third-order rate constant $\frac{k_{cat}}{K_{iA}K_B}$ and lower product inhibition constants K_P may be slower on average than reactions with lower $\frac{k_{cat}}{K_{iA}K_B}$ and higher product inhibition constants K_P. Nonetheless, analysis of steady-state rate laws can serve as a source of useful metrics for comparing reactions.

Fox and Clay devised the "catalytic effectiveness", $\bar{v} \equiv \frac{[P]_t}{t}$, as a measure of the average rate of production of the product that accounts for the slowing effects of product inhibition over the course of the reaction, but only applied it to uni-reactant reactions [145]. The TATF is closely related to the TTN $\equiv \frac{[P]_t}{[E_0]} = \frac{\bar{v}t}{[E_0]}$. Under the assumption that the evolution of the concentration of the external species is slow compared to the dynamics of the product-bound and substrate-bound catalyst (intermediate species), the system evolves as though the intermediate species are in steady-state with one another at each infintesimal moment in time. Thus, the time derivative of the product concentration follows an intermediate steady-state rate law. An algebraic form for the catalytic effectiveness of a bi-uni reaction can be calculated as follows by integrating the reciprocal of the

steady-state rate law,

$$\left.\frac{d[\mathrm{P}]}{dt}\right|_{SS} = v([\mathrm{A}],[\mathrm{B}],[\mathrm{P}])\,, \tag{2.16}$$

$$\int_0^{[\mathrm{P}]_t} \frac{1}{v([\mathrm{A}]_0-[\mathrm{P}]',[\mathrm{B}]_0-[\mathrm{P}]',[\mathrm{P}]')} d[\mathrm{P}]' = t\,, \tag{2.17}$$

$$\bar{v} \equiv \frac{[\mathrm{P}]_t}{t} = \frac{[\mathrm{P}]_t}{\int_0^{[\mathrm{P}]_t} v([\mathrm{A}]_0-[\mathrm{P}]',[\mathrm{B}]_0-[\mathrm{P}]',[\mathrm{P}]')^{-1}\, d[\mathrm{P}]'} \tag{2.18}$$

where $[\mathrm{A}]_0$ and $[\mathrm{B}]_0$ are the initial concentrations of substrates A and B, and I have assumed that the concentration of the template or catalyst is low enough that it sequesters a negligible portion of the substrate or product in solution.

The catalytic effectiveness is a powerful metric in that it is a direct measure of the rate of the reaction that accounts for the effects of inhibitors and the depletion of substrate over time. However, the metric scales proportionally with the catalyst concentration. Therefore, in this chapter I argue that the catalytic effectiveness per unit catalyst (or template/enzyme) would be a more suitable metric for our purposes, given that we require that each template generates many copies. I refer to the catalytic effectiveness per unit catalyst (or template/enzyme) as the time-averaged total turnover frequency (TATF), which I define as follows:

$$\text{TATF: } r(t) \equiv \frac{[\mathrm{P}]_t}{t\,[\mathrm{E}_0]} \approx \frac{[\mathrm{P}]_t}{[\mathrm{E}_0]\int_0^{[\mathrm{P}]_t} v([\mathrm{A}]_0-[\mathrm{P}]',[\mathrm{B}]_0-[\mathrm{P}]',[\mathrm{P}]')^{-1}\, d[\mathrm{P}]'} \tag{2.19}$$

To obtain accurate measurements of TATF, it should be measured with a sufficiently low catalyst concentration to ensure that the catalyst's sequestration of substrates and products does not significantly affect their solution concentration. Initially, the concentrations of substrates in solution should be equal $[\mathrm{A}]_0 = [\mathrm{B}]_0$ and the concentration of product should be $[\mathrm{P}]_0 = 0$. TATF is a time-averaged measure of the reaction rate that takes into account the changes in kinetics as the substrate concentrations decrease and the product concentrations increase during the reaction. The TATF quoted in Eqn.2.19 is a function of time, but it should be reported after a specified period sufficient for the product concentration to achieve a defined fraction of the ultimate yield of the reaction. For instance, r_{99} is the TATF measured over the time required for the product to attain 99% of the final concentration.

The TATF is a frequency and is measured in Hertz. The TATF should be interpreted as the frequency at which a single catalyst produces a product molecule, averaged over the course of a reaction until close to completion. If the concentration of substrates remains saturating over the period across which the TATF is measured, then the TATF will be independent of the substrate concentration. Otherwise, the TATF will be a function of the initial substrate concentration. If the initial substrate concentration is very low, such that the

reaction displays second-order substrate kinetics in the case of a bi-uni reaction, then the TATF can be interpreted as the time-averaged third-order rate constant for the reaction. Therefore, the conditions under which the TATF was measured should be quoted, or, failing this, an indication of the reaction's operating regime should be considered.

I now proceed to calculate the TATF for a variety of idealised dimerisation mechanisms.

2.2 Ordered bi-uni with competitive product inhibition

In this section, I analyse the following bi-uni ordered enzymatic reaction:

$$\begin{aligned}
\text{A+E} &\underset{k_{-1}[\text{EA}]}{\overset{k_1[\text{A}][\text{E}]}{\rightleftharpoons}} \text{EA} \\
\text{EA+B} &\underset{k_{-2}[\text{EAB}]}{\overset{k_2[\text{B}][\text{EA}]}{\rightleftharpoons}} \text{EAB} \\
\text{EAB} &\underset{k_{-3}[\text{EP}]}{\overset{k_3[\text{EAB}]}{\rightleftharpoons}} \text{EP} \\
\text{EP} &\underset{k_{-4}[\text{P}][\text{E}]}{\overset{k_4[\text{EP}]}{\rightleftharpoons}} \text{P+E}
\end{aligned}$$

This set of reactions can be modelled by the following set of differential equations.

$$\frac{\mathrm{d}[\text{EA}]}{\mathrm{dt}} = k_1[\text{A}][\text{E}] + k_{-2}[\text{EAB}] - k_{-1}[\text{EA}] - k_2[\text{EA}][\text{B}]\,, \tag{2.20}$$

$$\frac{\mathrm{d}[\text{EAB}]}{\mathrm{dt}} = k_2[\text{EA}][\text{B}] + k_{-3}[\text{EP}] - k_{-2}[\text{EAB}] - k_3[\text{EAB}]\,, \tag{2.21}$$

$$\frac{\mathrm{d}[\text{EP}]}{\mathrm{dt}} = k_3[\text{EAB}] + k_{-4}[\text{P}][\text{E}] - k_{-3}[\text{EP}] - k_4[\text{EP}]\,, \tag{2.22}$$

$$\frac{\mathrm{d}[\text{A}]}{\mathrm{dt}} = k_{-1}[\text{EA}] - k_1[\text{A}][\text{E}]\,, \tag{2.23}$$

$$\frac{\mathrm{d}[\text{B}]}{\mathrm{dt}} = k_{-2}[\text{EAB}] - k_2[\text{EA}][\text{B}]\,, \tag{2.24}$$

$$\frac{\mathrm{d}[\text{P}]}{\mathrm{dt}} = k_4[\text{EP}] - k_{-4}[\text{P}][\text{E}]\,. \tag{2.25}$$

Here [E] is the concentration of free enzyme, [A] and [B] are the concentrations of free substrate, and [P] is the concentration of free product. I assume that the total enzyme concentration is fixed, $[\text{E}_0] = [\text{E}] + [\text{EA}] + [\text{EAB}] + [\text{EP}]$. In the intermediate steady-state, the derivatives of the concentrations satisfy

$$\frac{\mathrm{d}[\text{EA}]}{\mathrm{dt}} = \frac{\mathrm{d}[\text{EAB}]}{\mathrm{dt}} = \frac{\mathrm{d}[\text{EP}]}{\mathrm{dt}} = \frac{\mathrm{d}[\text{E}]}{\mathrm{dt}} = 0\,, \tag{2.26}$$

and, therefore, in the intermediate steady-state, the production rate of product, $\frac{\mathrm{d}[\text{P}]}{\mathrm{dt}}$, is

$$\frac{\mathrm{d}[\text{P}]}{\mathrm{dt}} = [\text{E}_0]\frac{k_1k_2k_3k_4[\text{A}][\text{B}] - k_{-1}k_{-2}k_{-3}k_{-4}[\text{P}]}{\aleph}\,, \tag{2.27}$$

where

$$\begin{aligned}\aleph =& k_4k_{\text{-}1}(k_3+k_{\text{-}2})+k_{\text{-}1}k_{\text{-}2}k_{\text{-}3}+k_1k_4(k_3+k_{\text{-}2})\\ &+k_1k_{\text{-}2}k_{\text{-}3}[\mathrm{A}]+k_2k_3k_4[\mathrm{B}]\\ &+k_1k_2(k_3+k_4+k_{\text{-}3})[\mathrm{A}][\mathrm{B}]\\ &+k_{\text{-}4}(k_3k_{\text{-}1}+k_{\text{-}1}k_{\text{-}2}+k_{\text{-}1}k_{\text{-}3}+k_{\text{-}2}k_{\text{-}3})[\mathrm{P}]\\ &+k_2k_{\text{-}4}(k_3+k_{\text{-}3})[\mathrm{P}][\mathrm{B}]\,.\end{aligned}$$

The steady-state rate of product formation per unit catalyst can be recast in terms of kinetic constants as follows,

$$R([\mathrm{A}],[\mathrm{B}],[\mathrm{P}]) = \frac{1}{[\mathrm{E}_0]}\frac{\mathrm{d}[\mathrm{P}]}{\mathrm{dt}} \tag{2.28}$$

$$= \frac{k_{\text{cat}}[\mathrm{E}_0]\left([\mathrm{A}][\mathrm{B}]-\frac{[\mathrm{P}]}{K_{\text{eq}}}\right)}{K_{\mathrm{B}}K_{\mathrm{iA}}\left(1+\frac{[\mathrm{P}]}{K_{\mathrm{P}}}\right)+[\mathrm{A}]K_{\mathrm{B}}+[\mathrm{B}]K_{\mathrm{A}}\left(1+\frac{[\mathrm{P}]}{K_{\mathrm{iP}}}\right)+[\mathrm{A}][\mathrm{B}]}\,, \tag{2.29}$$

where the equilibrium constant is

$$K_{\text{eq}} = \frac{k_1k_2k_3k_4}{k_{\text{-}1}k_{\text{-}2}k_{\text{-}3}k_{\text{-}4}}\,, \tag{2.30}$$

the limiting rate (under saturating substrate concentrations) of the reaction in the forward direction per unit catalyst is

$$k_{\text{cat}} = \frac{k_3k_4}{k_3+k_4+k_{\text{-}3}}\,, \tag{2.31}$$

the Michaelis constant for substrate A under saturating concentration of B is

$$K_{\mathrm{A}} = \frac{k_3k_4}{k_1(k_3+k_4+k_{\text{-}3})}\,, \tag{2.32}$$

the Michaelis constant for substrate B under saturating concentration of A is

$$K_{\mathrm{B}} = \frac{k_3k_4+k_{\text{-}2}(k_4+k_{\text{-}3})}{k_2(k_3+k_4+k_{\text{-}3})}\,, \tag{2.33}$$

the inhibition constant of substrate A for substrate B is

$$K_{\mathrm{iA}} = \frac{k_{\text{-}1}}{k_1}\,, \tag{2.34}$$

the Michaelis constant for product P is

$$K_{\mathrm{P}} = \frac{k_{\text{-}1}(k_3 k_4 + k_{\text{-}2}(k_4 + k_{\text{-}3}))}{k_{\text{-}4}(k_{\text{-}1}(k_3 + k_{\text{-}2}) + k_{\text{-}3}(k_{\text{-}1} + k_{\text{-}2}))}, \tag{2.35}$$

and the inhibition constant of product is

$$K_{\mathrm{iP}} = \frac{k_3 k_4}{k_{\text{-}4}(k_3 + k_{\text{-}3})}. \tag{2.36}$$

2.2.1 Simplifying the steady-state rate law

I use the steady-state rate law in Eqn. 2.27 for this ordered bi-uni reaction to approximate the TATF as defined in Eqn. 2.19. Before attempting the integral in Eqn. 2.19, I make the assumption that the reaction is sufficiently strongly driven to be effectively irreversible. Let the initial substrate concentrations be $[\mathrm{A}]_0 = [\mathrm{B}]_0 = [\mathrm{C}_0]$. Assuming that the template concentration is low enough that, at equilibrium, the template sequesters a minimal amount of substrate and product, the equilibrium product concentration, $[\mathrm{P}]_{EQ}$ that satisfies $R([\mathrm{C}_0] - [\mathrm{P}]_{EQ}, [\mathrm{C}_0] - [\mathrm{P}]_{EQ}, [\mathrm{P}]_{EQ}) = 0$ (the rate of the reaction is 0 in equilibrium) is

$$[\mathrm{P}]_{EQ} = [\mathrm{C}_0] - \frac{\sqrt{1 + 4[\mathrm{C}_0]K_{\mathrm{eq}}} - 1}{2K_{\mathrm{eq}}}. \tag{2.37}$$

The reaction will convert the majority of substrate to product $[\mathrm{P}]_{EQ} \approx [\mathrm{C}_0]$ if $\sqrt{[\mathrm{C}_0]K_{\mathrm{eq}}} \gg 1$. Hence, a high equilibrium constant and a sufficiently high intial substrate concentration are required to ensure complete conversion. At low substrate concentrations, care should be taken to ensure that this condition of strong irreversibility holds when using the following simplified form of the rate equation, otherwise a reversible model may be more appropriate. If the polymerisation step is effectively irreversible, $k_{\text{-}3} \approx 0$, then $K_{\mathrm{eq}} \to \infty$, and the steady-state rate can be reduced to

$$R([\mathrm{A}], [\mathrm{B}], [\mathrm{P}])_{irr} = \lim_{k_{\text{-}3} \to 0} R([\mathrm{A}], [\mathrm{B}], [\mathrm{P}]) \tag{2.38}$$

$$= \frac{k_{\mathrm{cat}}[\mathrm{E}_0][\mathrm{A}][\mathrm{B}]}{K_{\mathrm{B}}K_{\mathrm{iA}}\left(1 + \frac{[\mathrm{P}]}{K_{\mathrm{P}}}\right) + [\mathrm{A}]K_{\mathrm{B}} + [\mathrm{B}]K_{\mathrm{A}}\left(1 + \frac{[\mathrm{P}]}{K_{\mathrm{iP}}}\right) + [\mathrm{A}][\mathrm{B}]}, \tag{2.39}$$

where the maximum rate of reaction in the forward direction per unit catalyst in the saturating substrates limit reduces to

$$\lim_{k_{\text{-}3} \to 0} k_{\mathrm{cat}} = \frac{k_3 k_4}{k_3 + k_4}, \tag{2.40}$$

the Michaelis constant for substrate A under saturating concentration of B reduces to

$$\lim_{k_{-3}\to 0} K_{\mathrm{A}} = \frac{k_3 k_4}{k_1(k_3+k_4)}\,, \tag{2.41}$$

the Michaelis constant for substrate B under saturating concentration of A reduces to

$$\lim_{k_{-3}\to 0} K_{\mathrm{B}} = \frac{k_4(k_3+k_{-2})}{k_2(k_3+k_4)}\,, \tag{2.42}$$

the inhibition constant of substrate A for substrate B reduces to

$$\lim_{k_{-3}\to 0} K_{\mathrm{iA}} = \frac{k_{-1}}{k_1}\,, \tag{2.43}$$

and the Michaelis constant and inhibition constant associated with the product P reduce to the dissociation constant of the product

$$\lim_{k_{-3}\to 0} K_{\mathrm{P}} = \lim_{k_{-3}\to 0} K_{\mathrm{iP}} = \frac{k_4}{k_{-4}}\,. \tag{2.44}$$

The fractional error in neglecting the numerator term $[\mathrm{A}][\mathrm{B}] - [\mathrm{P}]/K_{\mathrm{eq}} \approx [\mathrm{A}][\mathrm{B}]$ goes as

$$\mathrm{err} = \frac{[\mathrm{P}]/K_{\mathrm{eq}}}{[\mathrm{A}][\mathrm{B}] - [\mathrm{P}]/K_{\mathrm{eq}}} = \frac{[\mathrm{P}]/K_{\mathrm{eq}}}{([\mathrm{C}_0]-[\mathrm{P}])([\mathrm{C}_0]-[\mathrm{P}]) - [\mathrm{P}]/K_{\mathrm{eq}}}\,. \tag{2.45}$$

The error in the approximation increases over the course of the reaction as the product concentration rises. Assuming $\sqrt{[\mathrm{C}_0]K_{\mathrm{eq}}} \gg 1$, let $[\mathrm{P}] = (1-\varepsilon)[\mathrm{C}_0]$ be the concentration of product close to the end of the reaction, where $\varepsilon > 1/\sqrt{[\mathrm{C}_0]K_{\mathrm{eq}}}$. The error of the approximation is

$$\mathrm{err} = \frac{1}{\frac{\varepsilon^2[\mathrm{C}_0]K_{\mathrm{eq}}}{1-\varepsilon} - 1} \approx \frac{1}{\varepsilon^2[\mathrm{C}_0]K_{\mathrm{eq}}}\,, \tag{2.46}$$

a quadratic in $\frac{1}{\varepsilon\sqrt{[\mathrm{C}_0]K_{\mathrm{eq}}}} < 1$. Hence, the product concentration dependent term in the numerator can be neglected if the substrate concentration at the end of the averaging period is sufficiently high $\varepsilon^2[\mathrm{C}_0] > 1/K_{\mathrm{eq}}$. If the substrate concentration at the end of the period is not high enough, then the retroactivity of the reaction will have a strong effect on the observed rate, and the approximation may not hold.

2.2.2 Estimating TATF in the irreversible limit

Using the steady-state rate law in the irreversible limit, I approximate the integral I as

$$I = [\mathrm{E}_0] \int_0^{[\mathrm{P}]_t} v\left([\mathrm{A}]_0 - [\mathrm{P}]', [\mathrm{B}]_0 - [\mathrm{P}]', [\mathrm{P}]'\right)^{-1} d[\mathrm{P}]' \tag{2.47}$$

$$\approx \int_0^{[\mathrm{P}]_t} R\left([\mathrm{A}]_0 - [\mathrm{P}]', [\mathrm{B}]_0 - [\mathrm{P}]', [\mathrm{P}]'\right)_{\mathrm{irr}}^{-1} d[\mathrm{P}]' \tag{2.48}$$

$$= \frac{1}{k_{\mathrm{cat}}} \Bigg[-\log([\mathrm{C}_0] - [\mathrm{P}])\left(\frac{K_{\mathrm{A}}}{K_{\mathrm{iP}}}([\mathrm{C}_0] + K_{\mathrm{iP}}) + \frac{K_{\mathrm{B}}}{K_{\mathrm{P}}}(K_{\mathrm{P}} - K_{\mathrm{iA}})\right) \tag{2.49}$$

$$+ \frac{([\mathrm{C}_0] - [\mathrm{P}])(K_{\mathrm{A}} - K_{\mathrm{iP}})}{K_{\mathrm{iP}}} + \frac{K_{\mathrm{B}} K_{\mathrm{iA}}([\mathrm{C}_0] + K_{\mathrm{P}})}{K_{\mathrm{P}}([\mathrm{C}_0] - [\mathrm{P}])} \Bigg]_{[\mathrm{P}]=0}^{[\mathrm{P}]=[\mathrm{P}]_t},$$

which can be used to approximate the TATF averaged over a time period in which the product concentration reached $[\mathrm{P}] = (1-\varepsilon)[\mathrm{C}_0]$

$$\mathrm{TATF}_{1-\varepsilon} = \frac{[\mathrm{P}]}{I} \tag{2.50}$$

$$\approx \frac{k_{\mathrm{cat}}[\mathrm{C}_0]^2}{\frac{K_{\mathrm{iA}}K_{\mathrm{B}}}{\varepsilon}\left(1 + \frac{[\mathrm{C}_0]}{K_{\mathrm{P}}}\left(1 - \frac{\varepsilon \log\left(\frac{1}{\varepsilon}\right)}{1-\varepsilon}\right)\right) + [\mathrm{C}_0](K_{\mathrm{A}} + K_{\mathrm{B}})\frac{\log\left(\frac{1}{\varepsilon}\right)}{1-\varepsilon} + [\mathrm{C}_0]^2\left(\frac{K_{\mathrm{A}}}{K_{\mathrm{iP}}}\left(\frac{\log\left(\frac{1}{\varepsilon}\right)}{1-\varepsilon} - 1\right) + 1\right)}. \tag{2.51}$$

For $\varepsilon = 0.01$, the TATF is averaged over the time until the product concentration reaches 99% of the initial substrate concentration, for which we get

$$\mathrm{TATF}_{0.99} \approx \frac{k_{\mathrm{cat}}[\mathrm{C}_0]^2}{100K_{\mathrm{B}}K_{\mathrm{iA}}\left(1 + 0.95\frac{[\mathrm{C}_0]}{K_{\mathrm{P}}}\right) + 4.7[\mathrm{C}_0](K_{\mathrm{A}} + K_{\mathrm{B}}) + [\mathrm{C}_0]^2\left(1 + \frac{3.7K_{\mathrm{A}}}{K_{\mathrm{iP}}}\right)}. \tag{2.52}$$

Assuming that the polymerisation step is irreversible $k_{-3} \to 0$ and that the binding rates of reactants and products are roughly similar $k_1 \approx k_2 \approx k_{-4}$, the $\mathrm{TATF}_{0.99}$ can be simplified further, owing to the approximation that

$$\lim_{k_{-3}\to 0} \frac{K_{\mathrm{B}}K_{\mathrm{iA}}}{K_{\mathrm{P}}(K_{\mathrm{A}} + K_{\mathrm{B}})} = \lim_{k_{-3}\to 0} \frac{k_{-4}(k_{-1}(k_3 + k_{-2}) + k_{-3}(k_{-1} + k_{-2}))}{k_1 k_4(k_3 + k_{-2}) + k_1 k_{-2} k_{-3} + k_2 k_3 k_4} \tag{2.53}$$

$$= \frac{k_{-1}k_{-4}(k_3 + k_{-2})}{k_4(k_1(k_3 + k_{-2}) + k_2 k_3)} \tag{2.54}$$

$$\approx \frac{k_{-1}}{k_4} \geq 1, \tag{2.55}$$

which is much greater than unity if the first substrate unbinds the template faster than the product and is near unity if their unbinding rates are similar. Therefore we can reduce the $\mathrm{TATF}_{0.99}$ to

$$\mathrm{TATF}_{0.99} \approx \frac{[\mathrm{C}_0]^2 k_{\mathrm{cat}}}{100 K_{\mathrm{B}} K_{\mathrm{iA}} \left(1 + \frac{[\mathrm{C}_0]}{K_{\mathrm{P}}}\right) + [\mathrm{C}_0]^2 \left(1 + 3.7 \frac{K_{\mathrm{A}}}{K_{\mathrm{iP}}}\right)}. \tag{2.56}$$

Depending on the intial concentration at which it is measured, the TATF may be a quadratic, linear or constant function of $[\mathrm{C}_0]$.

High initial substrate concentrations

Over the course of the reaction, the kinetics will be saturated and the $\mathrm{TATF}_{0.99}$ will be independent of the initial concentration if $[\mathrm{C}_0] > \frac{100 K_{\mathrm{B}} K_{\mathrm{iA}}}{K_{\mathrm{P}}\left(1 + \frac{3.7 K_{\mathrm{A}}}{K_{\mathrm{iP}}}\right)} > \sqrt{\frac{100 K_{\mathrm{B}} K_{\mathrm{iA}}}{1 + \frac{3.7 K_{\mathrm{A}}}{K_{\mathrm{iP}}}}}$. It can be seen that this inequality holds quite generally, as with a little rearrangement

$$\frac{100 K_{\mathrm{B}} K_{\mathrm{iA}}}{K_{\mathrm{P}} \left(1 + \frac{3.7 K_{\mathrm{A}}}{K_{\mathrm{iP}}}\right)} > \sqrt{\frac{100 K_{\mathrm{B}} K_{\mathrm{iA}}}{1 + \frac{3.7 K_{\mathrm{A}}}{K_{\mathrm{iP}}}}}, \tag{2.57}$$

can be transformed to

$$10 > \frac{2.2 K_{\mathrm{P}}}{\sqrt{K_{\mathrm{B}} K_{\mathrm{iA}}}} \geq \frac{K_{\mathrm{P}} \sqrt{1 + \frac{3.7 K_{\mathrm{A}}}{K_{\mathrm{iP}}}}}{\sqrt{K_{\mathrm{B}} K_{\mathrm{iA}}}}, \tag{2.58}$$

as the term $\frac{K_{\mathrm{A}}}{K_{\mathrm{iP}}} = \frac{k_{\text{-}4} k_3}{k_1 (k_3 + k_4)} \leq 1$. Substituting in the rate constants, the inequalities below,

$$10 > \frac{2.2 K_{\mathrm{P}}}{\sqrt{K_{\mathrm{B}} K_{\mathrm{iA}}}} \tag{2.59}$$

$$4.5 > 1 > \frac{\sqrt{k_1 k_2}}{k_{\text{-}4}} \sqrt{\frac{k_4}{k_{\text{-}1}}} \frac{\sqrt{k_3 + k_4}}{\sqrt{k_3 + k_{\text{-}2}}}, \tag{2.60}$$

will usually condition hold, as the rate of product detachment will usually be comparable or much less than the rate of substrate detachment $k_4 < k_{\text{-}1}, k_{\text{-}2}$ and the binding rates are approximately similar $k_{\text{-}4} \lesssim k_1 \approx k_2$. The binding rate constant of the ligated product may be less than the binding rate constants of the substrates if, due to the molecule's increased molecular weight, it has a lower diffusion constant.

In the limit that the concentration is sufficiently high to remain saturating throughout the course of the reaction ($[\mathrm{C}_0] > \frac{100 K_{\mathrm{B}} K_{\mathrm{iA}}}{K_{\mathrm{P}}\left(1 + \frac{3.7 K_{\mathrm{A}}}{K_{\mathrm{iP}}}\right)}$), the TATF reduces to

$$\lim_{[\mathrm{C}_0] \to \infty} \mathrm{TATF}_{0.99} = \frac{k_{\mathrm{cat}}}{1 + 3.7 \frac{K_{\mathrm{A}}}{K_{\mathrm{iP}}}} \approx \frac{k_3 k_4}{k_3 (1 + 3.7 \frac{k_{\text{-}4}}{k_1}) + k_4}. \tag{2.61}$$

Eqn. 2.61 indicates that competitive inhibition slows down the average speed of the reaction relative to an

uninhibited reaction unless the reaction is rate limited by the polymerisation step ($k_3 \ll k_4$). In this limit the rate of the reaction is independent of the substrate and product binding and unbinding rates. When the reaction isn't rate limited by polymerisation, the speed of the reaction is dependent on the formation of the EA and EP states, and hence on the competition between the binding of substrate A and the product P to the enzyme.

Intermediate initial substrate concentrations

When the initial concentration lies between $K_P < [C_0] < \frac{100 K_B K_{iA}}{K_P\left(1+\frac{3.7 K_A}{K_{iP}}\right)}$, over the course of the reaction the substrate concentration drops to an extent that it is no longer saturating and the rising product concentration has an inhibiting effect on the average speed of the reaction. Here the TATF is approximately linearly dependent on concentration. We can approximate the $TATF_{0.99}$ under these conditions as

$$TATF_{0.99} \approx \frac{[C_0] K_P k_{cat}}{100 K_B K_{iA}} = \frac{[C_0] k_1 k_2 k_3 k_4}{100 (k_3 + k_{-2}) k_{-1} k_{-4}}. \tag{2.62}$$

In this semi-saturated regime, Eqn. 2.62 would serve as a useful metric for comparing reactions. In the irreversible polymerisation limit ($k_{-3} \to 0$), the TATF is proportional to the rate constant of binding the second substrate k_2, the probability of successfully polymerising the substrate $\frac{k_3}{k_3+k_{-2}}$, and the ratio of the dissociation constants of the first substrate $\frac{k_{-1}}{k_1}$ and the product $\frac{k_4}{k_{-4}}$. Variants with higher k_{cat}, lower thresholds for substrate saturation $\sqrt{K_B K_{iA}}$ and higher thresholds for product inhibition K_P would score more highly against this metric. Such reactions would convert a large portion of substrates to product before slowing significantly due to the depleted, non-saturating substrate concentrations and product inhibition.

Low initial substrate concentrations

In the limit that the initial substrate concentrations are so low that the reaction is not in a substrate or product saturated regime for the majority or the entirety of the reaction, $[C_0] < K_P < \sqrt{\frac{100 K_B K_{iA}}{1+\frac{3.7 K_A}{K_{iP}}}}$, the $TATF_{0.99}$ is approximately

$$\lim_{[C_0]\to 0} TATF_{0.99} = \frac{[C_0]^2 k_{cat}}{100 K_B K_{iA}} = \frac{[C_0]^2 k_1 k_2 k_3}{100 k_{-1} (k_3 + k_{-2})}, \tag{2.63}$$

which is proportional to the metric derived by Banta and Wheeldon in Ref. [141] and used in Ref. [146] at low substrate concentrations and in the absence of product inhibition. In this regime, the reaction has second-order kinetics for most of its duration and slows significantly as the substrate is depleted. The TATF is proportional to the occupancy of the first substrate-bound complex ($\frac{[C_0] k_1}{k_{-1}}$) and the rate of binding and successfully polymerising the second substrate $\frac{[C_0] k_2 k_3}{k_3+k_{-2}}$. In this regime, variants with a higher k_{cat} and a lower threshold for substrate

saturation $\sqrt{K_B K_{iA}}$ would have a higher TATF. Care should be taken when measuring the TATF in this regime that the equilibrium condition $\sqrt{[C_0]K_{eq}} \gg 1$ and the condition on the irreversibility assumption $\varepsilon^2[C_0] \gg 1/K_{eq}$ hold, if the measurement is to be compared to Eqn. 2.63. If the substrate concentration is not sufficiently high to drive the near-complete conversion of substrate into product or the substrate concentration at the end of the observation period is so low that the reaction runs in reverse, then the true TATF may be far less than the TATF approximated in the irreversible limit.

2.3 Ordered bi-uni with mixed product inhibition

Here I expand upon the model used in Section 2.2 by introducing a pathway for non-competitive product inhibition. Non-competitive product inhibition occurs when the product can bind to a substrate-bound catalyst complex, such as EA.

$$\begin{aligned}
\text{A+E} &\underset{k_{-1}[\text{EA}]}{\overset{k_1[\text{A}][\text{E}]}{\rightleftharpoons}} \text{EA} \\
\text{EA+B} &\underset{k_{-2}[\text{EAB}]}{\overset{k_2[\text{B}][\text{EA}]}{\rightleftharpoons}} \text{EAB} \\
\text{EAB} &\underset{k_{-3}[\text{EP}]}{\overset{k_3[\text{EAB}]}{\rightleftharpoons}} \text{EP} \\
\text{EP} &\underset{k_{-4}[\text{P}][\text{E}]}{\overset{k_4[\text{EP}]}{\rightleftharpoons}} \text{P+E} \\
\text{EA+P} &\underset{k_{-5}[EAP]}{\overset{k_5[\text{EA}][\text{P}]}{\rightleftharpoons}} \text{EAP}
\end{aligned}$$

This set of reactions can be modelled by the following set of differential equations.

$$\frac{\text{d[EA]}}{\text{dt}} = k_1[\text{A}][\text{E}] + k_{-2}[\text{EAB}] + k_{-5}[\text{EAP}] - (k_2[\text{B}] + k_5[\text{P}] + k_{-1})[\text{EA}], \tag{2.64}$$

$$\frac{\text{d[EAB]}}{\text{dt}} = k_2[\text{B}][\text{EA}] + k_{-3}[\text{EP}] - (k_3 + k_{-2})[\text{EAB}], \tag{2.65}$$

$$\frac{\text{d[EP]}}{\text{dt}} = k_{-4}[\text{E}][\text{P}] + k_3[\text{EAB}] - (k_4 + k_{-3})[\text{EP}], \tag{2.66}$$

$$\frac{\text{d[EAP]}}{\text{dt}} = k_5[\text{EA}][\text{P}] - k_{-5}[\text{EAP}], \tag{2.67}$$

$$\frac{\text{d[P]}}{\text{dt}} = k_{-5}[\text{EAP}] + k_4[\text{EP}] - (k_5[\text{EA}] + k_{-4}[\text{E}])[\text{P}]. \tag{2.68}$$

As before, [E] is the concentration of free enzyme, [A] and [B] are the concentrations of free substrate, and [P] is the concentration of free product. I assume that the total enzyme concentration is fixed, $[\text{E}_0] = [\text{E}] + [\text{EA}] + [\text{EAB}] + [\text{EP}] + [\text{EAP}]$.

In the intermediate steady-state, the derivatives of the following concentrations satisfy

$$\frac{\text{d[EA]}}{\text{dt}} = \frac{\text{d[EAB]}}{\text{dt}} = \frac{\text{d[EP]}}{\text{dt}} = \frac{\text{d[EAP]}}{\text{dt}} = \frac{\text{d[E]}}{\text{dt}} = 0, \tag{2.69}$$

and the rate of production, $\frac{\text{d[P]}}{\text{dt}}$, is given by

$$\frac{\text{d[P]}}{\text{dt}} = \frac{k_{\text{cat}}[\text{E}_0]\left([\text{A}][\text{B}] - [\text{P}]/K_{\text{eq}}\right)}{K_{\text{B}}K_{\text{iA}}\left(\frac{[\text{P}]}{K_{\text{P}}}\left(\frac{[\text{P}]}{K_{\text{inc2P}}}+1\right)+1\right) + [\text{A}]K_{\text{B}}\left(\frac{[\text{P}]}{K_{\text{incP}}}+1\right) + [\text{B}]K_{\text{A}}\left(\frac{[\text{P}]}{K_{\text{icP}}}+1\right) + [\text{A}][\text{B}]}, \tag{2.70}$$

where the equilibrium constant is

$$K_{\text{eq}} = \frac{k_1 k_2 k_3 k_4}{k_{\text{-}1} k_{\text{-}2} k_{\text{-}3} k_{\text{-}4}}, \tag{2.71}$$

the maximum rate of reaction in the forward direction per unit catalyst under saturating substrate concentrations is

$$k_{\text{cat}} = \frac{k_3 k_4}{k_3 + k_4 + k_{\text{-}3}}, \tag{2.72}$$

the Michaelis constant for substrate A under saturating concentration of B is

$$K_{\text{A}} = \frac{k_3 k_4}{k_1 (k_3 + k_4 + k_{\text{-}3})}, \tag{2.73}$$

the Michaelis constant for substrate B under saturating concentration of A is

$$K_{\text{B}} = \frac{k_3 k_4 + k_{\text{-}2}(k_4 + k_{\text{-}3})}{k_2 (k_3 + k_4 + k_{\text{-}3})}, \tag{2.74}$$

the inhibition constant of substrate A for substrate B is

$$K_{\text{iA}} = \frac{k_{\text{-}1}}{k_1}, \tag{2.75}$$

the Michaelis constant for product P is

$$K_{\text{P}} = \frac{k_{\text{-}1}(k_3 k_4 + k_{\text{-}2}(k_4 + k_{\text{-}3}))}{k_{\text{-}4}(k_{\text{-}1}(k_3 + k_{\text{-}2}) + k_{\text{-}3}(k_{\text{-}1} + k_{\text{-}2}))}, \tag{2.76}$$

the competitive inhibition constant of product is

$$K_{\text{icP}} = \frac{k_3 k_4}{k_{\text{-}4}(k_3 + k_{\text{-}3})}, \tag{2.77}$$

the linear non-competitive inhibition constant of product is

$$K_{\text{incP}} = \frac{k_{\text{-}5}}{k_5}, \tag{2.78}$$

and the quadratic non-competitive inhibition constant of product is

$$K_{\text{inc2P}} = \frac{k_{\text{-}5}(k_{\text{-}1}(k_3 + k_{\text{-}2}) + k_{\text{-}3}(k_{\text{-}1} + k_{\text{-}2}))}{k_5 k_{\text{-}2} k_{\text{-}3}}. \tag{2.79}$$

K_{inc2P} tends to infinity, and hence the term containing K_{inc2P} in Eqn.2.70 has minimal effect on the TATF when the polymerisation reaction is irreversible, $k_{\text{-3}} \to 0$. However, if the reaction is not irreversible, this term can cause a significant reduction in the speed of the reaction at high product concentration.

2.3.1 Approximating TATF in the irreversible limit

As in Section 2.2, I approximate the TATF in the limit of irreversible polymerisation ($k_{\text{-3}} \to 0$) as

$$\text{TATF}_{1-\varepsilon} \approx \frac{k_{\text{cat}}[\text{C}_0]^2}{\frac{K_\text{B}K_\text{iA}}{\varepsilon}\left(\frac{\eta_1[\text{C}_0]}{K_\text{P}}+1\right)+\eta_2[\text{C}_0](K_\text{A}+K_\text{B})+[\text{C}_0]^2\left(1+\eta_3\left(\frac{K_\text{A}}{K_\text{icP}}+\frac{K_\text{B}}{K_\text{incP}}\right)\right)}, \tag{2.80}$$

where

$$\eta_1 = 1-\frac{\varepsilon\log\left(\frac{1}{\varepsilon}\right)}{1-\varepsilon}, \tag{2.81}$$

$$\eta_2 = \frac{\log\left(\frac{1}{\varepsilon}\right)}{1-\varepsilon}, \tag{2.82}$$

$$\eta_3 = \frac{\log\left(\frac{1}{\varepsilon}\right)}{1-\varepsilon}-1, \tag{2.83}$$

and the kinetic constants reduce to the same values as in Section 2.2, with the Michaelis constant of the product and the competitive inhibition constant both tending to the dissociation constant of the product for the free catalyst

$$\lim_{k_{\text{-3}}\to 0} K_\text{P} = \lim_{k_{\text{-3}}\to 0} K_\text{icP} = \frac{k_4}{k_{\text{-4}}}. \tag{2.84}$$

For $\varepsilon = 0.01$, the TATF is approximately

$$\text{TATF}_{0.99} \approx \frac{k_{\text{cat}}[\text{C}_0]^2}{100K_\text{B}K_\text{iA}\left(1+\frac{0.95[\text{C}_0]}{K_\text{P}}\right)+4.7[\text{C}_0](K_\text{A}+K_\text{B})+[\text{C}_0]^2\left(1+3.7\left(\frac{K_\text{A}}{K_\text{icP}}+\frac{K_\text{B}}{K_\text{incP}}\right)\right)} \tag{2.85}$$

$$\approx \frac{k_{\text{cat}}[\text{C}_0]^2}{100K_\text{B}K_\text{iA}\left(1+\frac{[\text{C}_0]}{K_\text{P}}\right)+[\text{C}_0]^2\left(1+3.7\left(\frac{K_\text{A}}{K_\text{icP}}+\frac{K_\text{B}}{K_\text{incP}}\right)\right)}, \tag{2.86}$$

where again I have relied on the approximation that $\frac{K_\text{B}K_\text{iA}}{K_\text{P}(K_\text{A}+K_\text{B})} \approx \frac{k_4}{k_{\text{-1}}} \lesssim 1$.

High initial substrate concentrations

In the limit that the initial substrate concentration is high

$$[\text{C}_0] > \frac{100K_\text{B}K_\text{iA}}{K_\text{P}\left(1+3.7\left(\frac{K_\text{A}}{K_\text{icP}}+\frac{K_\text{B}}{K_\text{incP}}\right)\right)} > \sqrt{\frac{100K_\text{B}K_\text{iA}}{1+3.7\left(\frac{K_\text{A}}{K_\text{icP}}+\frac{K_\text{B}}{K_\text{incP}}\right)}}, \tag{2.87}$$

the substrates and product remain at saturating concentrations throughout the course of the reaction, and the $\mathrm{TATF}_{0.99}$ becomes

$$\lim_{[\mathrm{C}_0]\to\infty} \mathrm{TATF}_{0.99} = \frac{k_{\mathrm{cat}}}{1+3.7\left(\frac{K_{\mathrm{A}}}{K_{\mathrm{icP}}}+\frac{K_{\mathrm{B}}}{K_{\mathrm{incP}}}\right)} = \frac{k_3k_4}{k_3\left(\frac{3.7k_{\text{-}4}}{k_1}+1\right)+\frac{3.7k_3k_4k_5}{k_2k_{\text{-}5}}+k_4\left(\frac{3.7k_5k_{\text{-}2}}{k_2k_{\text{-}5}}+1\right)}, \tag{2.88}$$

where, in the irreversible limit $k_{\text{-}3}=0$, the ratios of kinetic parameters are given by

$$\frac{K_{\mathrm{A}}}{K_{\mathrm{icP}}} = \frac{k_3k_{\text{-}4}}{k_1(k_3+k_4)}, \tag{2.89}$$

$$\frac{K_{\mathrm{B}}}{K_{\mathrm{incP}}} = \frac{k_4k_5(k_3+k_{\text{-}2})}{k_2k_{\text{-}5}(k_3+k_4)}. \tag{2.90}$$

In the limit that the reaction isn't rate limited by polymerisation ($k_3 \gg k_{\text{-}4}$), the TATF becomes

$$\lim_{k_3\gg k_{\text{-}4}}\left(\lim_{[\mathrm{C}_0]\to\infty}\mathrm{TATF}_{0.99}\right) = \frac{k_4}{\frac{3.7k_{\text{-}4}}{k_1}+\frac{3.7k_4k_5}{k_2k_{\text{-}5}}+1}, \tag{2.91}$$

the reaction is slowed by events in which the product rather than the substrate binds to the free template, corresponding to the term $\frac{k_{\text{-}4}}{k_1}$, and the non-competitive inhibition of the product sequestering the substrate-bound enzyme EAP, corresponding to the term $\frac{k_4k_5}{k_2k_{\text{-}5}}$. The non-competitive effect is weak if the product has a lower affinity for the substrate-bound catalyst ($k_{\text{-}5} \gg k_4$) and is stronger if the affinity is unchanged by the presence of substrate ($k_{\text{-}5} \approx k_4$). Alternatively, if the reaction isn't rate limited by the unbinding of the product ($k_{\text{-}1}, k_{\text{-}2} \gtrsim k_{\text{-}5} \gtrsim k_{\text{-}4} \gg k_3$), then the TATF reduces to

$$\lim_{k_{\text{-}5},k_{\text{-}4}\gg k_3}\left(\lim_{[\mathrm{C}_0]\to\infty}\mathrm{TATF}_{0.99}\right) = \frac{k_3}{1+\frac{3.7k_5k_{\text{-}2}}{k_2k_{\text{-}5}}}. \tag{2.92}$$

In this limit, the reaction is strongly affected by the balance between the affinities of the second substrate B and the product for the substrate-bound enzyme EA. If the dissociation constant of the product and the complex EA is lower than that of the substrate B for the complex EA, ($\frac{k_{\text{-}5}}{k_5} < \frac{k_{\text{-}2}}{k_2}$), then the EAP complex will be favoured over EAB sooner in the course of the reaction, and therefore the inhibitory effect will greatly reduce the average rate. In contrast to a system that only suffers from competitive product inhibition, as presented in section 2.2, the effects of non-competitive inhibition may significantly reduce the maximum speed of the reaction under saturating initial conditions.

Intermediate and low initial substrate concentrations

Relative to a system without non-competitive inhibition, introducing non-competitive inhibition changes the concentration boundaries across which the TATF is a quadratic, linear or constant function of the initial substrate concentration, though it has no effect on the functional forms of the TATF within these regimes. At intermediate concentrations, $K_{\mathrm{P}} < [\mathrm{C}_0] < \frac{100 K_{\mathrm{B}} K_{\mathrm{iA}}}{K_{\mathrm{P}}\left(1+3.7\left(\frac{K_{\mathrm{A}}}{K_{\mathrm{icP}}}+\frac{K_{\mathrm{B}}}{K_{\mathrm{incP}}}\right)\right)} < \frac{100 K_{\mathrm{B}} K_{\mathrm{iA}}}{K_{\mathrm{P}}\left(1+3.7\frac{K_{\mathrm{A}}}{K_{\mathrm{icP}}}\right)}$, the TATF is approximately linearly proportional to the initial concentration and has an identical form to Eqn. 2.62,

$$\mathrm{TATF}_{0.99} \approx \frac{[\mathrm{C}_0] k_{\mathrm{cat}} K_{\mathrm{P}}}{100 K_{\mathrm{B}} K_{\mathrm{iA}}}, \tag{2.93}$$

and in the limit of very low concentrations $[\mathrm{C}_0] < K_{\mathrm{P}} < \sqrt{\frac{100 K_{\mathrm{B}} K_{\mathrm{iA}}}{1+3.7\left(\frac{K_{\mathrm{A}}}{K_{\mathrm{icP}}}+\frac{K_{\mathrm{B}}}{K_{\mathrm{incP}}}\right)}}$, the $\mathrm{TATF}_{0.99}$ is similar to Eqn. 2.63

$$\lim_{[\mathrm{C}_0]\to 0} \mathrm{TATF}_{0.99} = \frac{[\mathrm{C}_0]^2 k_{\mathrm{cat}}}{100 K_{\mathrm{B}} K_{\mathrm{iA}}}. \tag{2.94}$$

2.4 Random order bi-uni reaction

In this section, I calculate the TATF for a random order bi-uni reaction with competitive product inhibition. A random bi-uni reaction can be described by the following set of reactions:

$$\begin{aligned}
\text{E+A} &\underset{k_{\text{-1A}}[\text{EA}]}{\overset{k_{\text{1A}}[\text{A}][\text{E}]}{\rightleftharpoons}} \text{EA} \\
\text{E+B} &\underset{k_{\text{-1B}}[\text{EB}]}{\overset{k_{\text{1B}}[\text{B}][\text{E}]}{\rightleftharpoons}} \text{EB} \\
\text{EA+B} &\underset{k_{\text{-2A}}[\text{EAB}]}{\overset{k_{\text{2A}}[\text{B}][\text{EA}]}{\rightleftharpoons}} \text{EAB} \\
\text{EB+A} &\underset{k_{\text{-2B}}[\text{EAB}]}{\overset{k_{\text{2B}}[\text{A}][\text{EB}]}{\rightleftharpoons}} \text{EAB} \\
\text{EAB} &\underset{k_{\text{-3}}[\text{EP}]}{\overset{k_{3}[\text{EAB}]}{\rightleftharpoons}} \text{EP} \\
\text{EP} &\underset{k_{\text{-4}}[\text{P}][\text{E}]}{\overset{k_{4}[\text{EP}]}{\rightleftharpoons}} \text{P+E}
\end{aligned}$$

$k_{1\text{A}}$, $k_{\text{-1A}}$, $k_{2\text{A}}$, $k_{\text{-2A}}$ are the rate constants associated with the pathway in which substrate A binds to the enzyme first, and $k_{1\text{B}}$, $k_{\text{-1B}}$, $k_{2\text{B}}$, $k_{\text{-2B}}$ are the rate constants associated with the pathway in which substrate B binds to the enzyme first. The free-energetic change associated with forming the transition complex EAB should be independent of the pathway which formed it and hence the substrate binding rate constants are constrained as

$$\frac{k_{1\text{A}}\,k_{2\text{A}}}{k_{\text{-1A}}\,k_{\text{-2A}}} = \frac{k_{1\text{B}}\,k_{2\text{B}}}{k_{\text{-1B}}\,k_{\text{-2B}}} \tag{2.95}$$

This system can be described by the following set of differential equations:

$$\begin{aligned}
\frac{\text{d[EA]}}{\text{dt}} &= k_{1\text{A}}[\text{A}][\text{E}] + k_{\text{-2A}}[\text{EAB}] - k_{2\text{A}}[\text{B}][\text{EA}] - k_{\text{-1A}}[\text{EA}]\,, \\
\frac{\text{d[EB]}}{\text{dt}} &= k_{1\text{B}}[\text{B}][\text{E}] + k_{\text{-2B}}[\text{EAB}] - k_{\text{-1B}}[\text{EB}] - k_{2\text{B}}[\text{A}][\text{EB}]\,, \\
\frac{\text{d[EAB]}}{\text{dt}} &= k_{2\text{B}}[\text{A}][\text{EB}] + k_{2\text{A}}[\text{B}][\text{EA}] + k_{\text{-3}}[\text{EP}] - (k_3 + k_{\text{-2A}} + k_{\text{-2B}})[\text{EAB}]\,, \\
\frac{\text{d[EP]}}{\text{dt}} &= k_{\text{-4}}[\text{E}][\text{P}] + k_3[\text{EAB}] - k_4[\text{EP}] - k_{\text{-3}}[\text{EP}]\,, \\
\frac{\text{d[A]}}{\text{dt}} &= k_{\text{-1A}}[\text{EA}] + k_{\text{-2B}}[\text{EAB}] - k_{1\text{A}}[\text{A}][\text{E}] - k_{2\text{B}}[\text{A}][\text{EB}]\,, \\
\frac{\text{d[B]}}{\text{dt}} &= k_{\text{-1B}}[\text{EB}] + k_{\text{-2A}}[\text{EAB}] - k_{1\text{B}}[\text{B}][\text{E}] - k_{2\text{A}}[\text{B}][\text{EA}]\,, \\
\frac{\text{d[P]}}{\text{dt}} &= k_4[\text{EP}] - k_{\text{-4}}[\text{E}][\text{P}]\,.
\end{aligned}$$

Solving the system when all intermediate states are in the steady-state, $\frac{\mathrm{d[EA]}}{\mathrm{dt}} = \frac{\mathrm{d[EB]}}{\mathrm{dt}} = \frac{\mathrm{d[EAB]}}{\mathrm{dt}} = \frac{\mathrm{d[EP]}}{\mathrm{dt}} = 0$, I calculate the steady-state production rate as:

$$\frac{\mathrm{d[P]}}{\mathrm{dt}} = k_{\mathrm{cat}}[\mathrm{E_0}]\left([\mathrm{A}][\mathrm{B}] - \frac{[\mathrm{P}]}{K_{\mathrm{eq}}}\right)\left(\eta_1[\mathrm{A}] + \eta_2[\mathrm{B}] + \eta_3 + \eta_4\right)/\Gamma, \tag{2.96}$$

where the equilibrium constant is

$$K_{\mathrm{eq}} = \frac{k_{1\mathrm{A}}k_{2\mathrm{A}}k_3k_4}{k_{\text{-}1\mathrm{A}}k_{\text{-}2\mathrm{A}}k_{\text{-}3}k_{\text{-}4}} = \frac{k_{1\mathrm{B}}k_{2\mathrm{B}}k_3k_4}{k_{\text{-}1\mathrm{B}}k_{\text{-}2\mathrm{B}}k_{\text{-}3}k_{\text{-}4}}, \tag{2.97}$$

the maximum rate of reaction in the forward direction per unit catalyst is

$$k_{\mathrm{cat}} = \frac{k_3k_4}{k_3 + k_4 + k_{\text{-}3}}, \tag{2.98}$$

and where the terms in the numerator are

$$\eta_1 = k_{1\mathrm{A}}\,,\ \eta_2 = k_{1\mathrm{B}}\,,\ \eta_3 = \frac{k_{1\mathrm{B}}k_{\text{-}1\mathrm{A}}}{k_{2\mathrm{A}}}\,,\ \eta_4 = \frac{k_{1\mathrm{A}}k_{\text{-}1\mathrm{B}}}{k_{2\mathrm{B}}}\,. \tag{2.99}$$

The denominator Γ is given by the following polynomial of [A],[B],[P],

$$\begin{aligned}\Gamma = {} & \Gamma_{1,1} + \Gamma_{1,2}[\mathrm{B}] + \Gamma_{1,3}[\mathrm{B}]^2 + \Gamma_{2,1}[\mathrm{A}] + \Gamma_{2,2}[\mathrm{A}][\mathrm{B}] + \Gamma_{2,3}[\mathrm{A}][\mathrm{B}]^2 \\ & + \Gamma_{3,1}[\mathrm{A}]^2 + \Gamma_{3,2}[\mathrm{A}]^2[\mathrm{B}] + [\mathrm{P}]\left(\Gamma^P_{1,1} + \Gamma^P_{1,2}[\mathrm{B}] + \Gamma^P_{2,1}[\mathrm{A}] + \Gamma^P_{2,2}[\mathrm{A}][\mathrm{B}]\right),\end{aligned} \tag{2.100}$$

where the factors are given by

$$
\begin{aligned}
\Gamma_{1,1} &= \frac{k_{\text{-}1\text{A}}k_{\text{-}1\text{B}}(k_3k_4+(k_4+k_{\text{-}3})(k_{\text{-}2\text{A}}+k_{\text{-}2\text{B}}))}{k_{2\text{A}}k_{2\text{B}}(k_3+k_4+k_{\text{-}3})}\,,\\
\Gamma_{1,2} &= \frac{k_{1\text{B}}k_{\text{-}1\text{A}}(k_3k_4+(k_4+k_{\text{-}3})(k_{\text{-}2\text{A}}+k_{\text{-}2\text{B}}))}{k_{2\text{A}}k_{2\text{B}}(k_3+k_4+k_{\text{-}3})}+\frac{k_{\text{-}1\text{B}}(k_3k_4+k_{\text{-}2\text{B}}(k_4+k_{\text{-}3}))}{k_{2\text{B}}(k_3+k_4+k_{\text{-}3})}\,,\\
\Gamma_{1,3} &= \frac{k_{1\text{B}}(k_3k_4+k_{\text{-}2\text{B}}(k_4+k_{\text{-}3}))}{k_{2\text{B}}(k_3+k_4+k_{\text{-}3})} = k_{1\text{B}}K_{2\text{A}}\,,\\
\Gamma_{2,1} &= \frac{k_{1\text{A}}k_{\text{-}1\text{B}}(k_3k_4+(k_4+k_{\text{-}3})(k_{\text{-}2\text{A}}+k_{\text{-}2\text{B}}))}{k_{2\text{A}}k_{2\text{B}}(k_3+k_4+k_{\text{-}3})}+\frac{k_{\text{-}1\text{A}}(k_3k_4+k_{\text{-}2\text{A}}(k_4+k_{\text{-}3}))}{k_{2\text{A}}(k_3+k_4+k_{\text{-}3})}\,,\\
\Gamma_{2,2} &= \frac{k_{2\text{A}}(k_{1\text{A}}k_3k_{\text{-}1\text{B}}+k_{1\text{A}}(k_4+k_{\text{-}3})(k_{\text{-}1\text{B}}+k_{\text{-}2\text{B}})+k_{2\text{B}}k_3k_4)}{k_{2\text{A}}k_{2\text{B}}(k_3+k_4+k_{\text{-}3})}\\
&\quad+\frac{k_{1\text{B}}k_{2\text{B}}(k_3k_{\text{-}1\text{A}}+(k_4+k_{\text{-}3})(k_{\text{-}1\text{A}}+k_{\text{-}2\text{A}}))}{k_{2\text{A}}k_{2\text{B}}(k_3+k_4+k_{\text{-}3})}\,,\\
\Gamma_{2,3} &= k_{1\text{B}}\,,\\
\Gamma_{3,1} &= \frac{k_{1\text{A}}(k_3k_4+k_{\text{-}2\text{A}}(k_4+k_{\text{-}3}))}{k_{2\text{A}}(k_3+k_4+k_{\text{-}3})} = k_{1\text{A}}K_{2\text{B}}\,,\\
\Gamma_{3,2} &= k_{1\text{A}}\,,\\
\Gamma^P_{1,1} &= \frac{k_{\text{-}1\text{A}}k_{\text{-}1\text{B}}k_{\text{-}4}(k_3+k_{\text{-}2\text{A}}+k_{\text{-}2\text{B}})}{k_{2\text{A}}k_{2\text{B}}(k_3+k_4+k_{\text{-}3})}\\
&\quad+\frac{k_{\text{-}1\text{A}}k_{\text{-}3}k_{\text{-}4}(k_{\text{-}1\text{B}}+k_{\text{-}2\text{B}})}{k_{2\text{A}}k_{2\text{B}}(k_3+k_4+k_{\text{-}3})}+\frac{k_{\text{-}1\text{B}}k_{\text{-}2\text{A}}k_{\text{-}3}k_{\text{-}4}}{k_{2\text{A}}k_{2\text{B}}(k_3+k_4+k_{\text{-}3})}\,,\\
\Gamma^P_{1,2} &= \frac{k_{\text{-}4}(k_{\text{-}1\text{B}}(k_3+k_{\text{-}2\text{B}})+k_{\text{-}3}(k_{\text{-}1\text{B}}+k_{\text{-}2\text{B}}))}{k_{2\text{B}}(k_3+k_4+k_{\text{-}3})} = k_{1\text{B}}\frac{K_{2\text{A}}K_{\text{iB}}}{K_{\text{PB}}}\,,\\
\Gamma^P_{2,1} &= \frac{k_{\text{-}4}(k_{\text{-}1\text{A}}(k_3+k_{\text{-}2\text{A}})+k_{\text{-}3}(k_{\text{-}1\text{A}}+k_{\text{-}2\text{A}}))}{k_{2\text{A}}(k_3+k_4+k_{\text{-}3})} = k_{1\text{A}}\frac{K_{2\text{B}}K_{\text{iA}}}{K_{\text{PA}}}\,,\\
\Gamma^P_{2,2} &= \frac{k_{\text{-}4}(k_3+k_{\text{-}3})}{k_3+k_4+k_{\text{-}3}} = k_{1\text{A}}\frac{K_{1\text{A}}}{K_{\text{iP}}} = k_{1\text{B}}\frac{K_{1\text{B}}}{K_{\text{iP}}}\,.
\end{aligned}
$$

As the numerator term $\eta_1[\text{A}]+\eta_2[\text{B}]+\eta_3+\eta_4$ doesn't factorise the denominator Γ, the steady-state rate **cannot** be expressed in the following form

$$k_{\text{cat}}[\text{E}_0]\left([\text{A}][\text{B}]-\frac{[\text{P}]}{K_{\text{eq}}}\right)/\Gamma' \tag{2.101}$$

where Γ' is a polynomial of the substrate and product concentrations with coefficients that are simple functions of the kinetic michaelis constants. However, the following kinetic constants can be written down for each independent binding pathway (A binding first or B binding first). $K_{1\text{X}}$ is the Michaelis constant of substrate X associated with the pathway in which the substrate X binds first, and $K_{2\text{X}}$ is the effective Michaelis constant for the pathway in which substrate X binds second. K_{iX} is the effective inhibition constant of the substrate X when

X binds first. The kinetic constants for each pathway are

$$
\begin{aligned}
K_{2\mathrm{B}} &= \frac{k_3k_4 + k_{\text{-}2\mathrm{A}}(k_4 + k_{\text{-}3})}{k_{2\mathrm{A}}(k_3 + k_4 + k_{\text{-}3})}, \\
K_{1\mathrm{A}} &= \frac{k_3k_4}{k_{1\mathrm{A}}(k_3 + k_4 + k_{\text{-}3})}, \\
K_{\mathrm{iA}} &= \frac{k_{\text{-}1\mathrm{A}}}{k_{1\mathrm{A}}}, \\
K_{\mathrm{PA}} &= \frac{k_{\text{-}1\mathrm{A}}(k_3k_4 + k_{\text{-}2\mathrm{A}}(k_4 + k_{\text{-}3}))}{k_{\text{-}4}(k_{\text{-}1\mathrm{A}}(k_3 + k_{\text{-}2\mathrm{A}}) + k_{\text{-}3}(k_{\text{-}1\mathrm{A}} + k_{\text{-}2\mathrm{A}}))}, \\
K_{2\mathrm{A}} &= \frac{k_3k_4 + k_{\text{-}2\mathrm{B}}(k_4 + k_{\text{-}3})}{k_{2\mathrm{B}}(k_3 + k_4 + k_{\text{-}3})}, \\
K_{1\mathrm{B}} &= \frac{k_3k_4}{k_{1\mathrm{B}}(k_3 + k_4 + k_{\text{-}3})}, \\
K_{\mathrm{iB}} &= \frac{k_{\text{-}1\mathrm{B}}}{k_{1\mathrm{B}}}, \\
K_{\mathrm{PB}} &= \frac{k_{\text{-}1\mathrm{B}}(k_3k_4 + k_{\text{-}2\mathrm{B}}(k_4 + k_{\text{-}3}))}{k_{\text{-}4}(k_{\text{-}1\mathrm{B}}(k_3 + k_{\text{-}2\mathrm{B}}) + k_{\text{-}3}(k_{\text{-}1\mathrm{B}} + k_{\text{-}2\mathrm{B}}))}, \\
K_{\mathrm{iP}} &= \frac{k_3k_4}{k_{\text{-}4}(k_3 + k_{\text{-}3})},
\end{aligned}
$$

and, further,

$$
\begin{aligned}
K_{2\mathrm{B}}K_{\mathrm{iA}} &= \frac{k_{\text{-}1\mathrm{A}}(k_3k_4 + k_{\text{-}2\mathrm{A}}(k_4 + k_{\text{-}3}))}{k_{1\mathrm{A}}k_{2\mathrm{A}}(k_3 + k_4 + k_{\text{-}3})}, \\
K_{2\mathrm{A}}K_{\mathrm{iB}} &= \frac{k_{\text{-}1\mathrm{B}}(k_3k_4 + k_{\text{-}2\mathrm{B}}(k_4 + k_{\text{-}3}))}{k_{1\mathrm{B}}k_{2\mathrm{B}}(k_3 + k_4 + k_{\text{-}3})}, \\
\frac{K_{2\mathrm{B}}K_{\mathrm{iA}}}{K_{\mathrm{PA}}} &= \frac{k_{\text{-}4}(k_{\text{-}1\mathrm{A}}(k_3 + k_{\text{-}2\mathrm{A}}) + k_{\text{-}3}(k_{\text{-}1\mathrm{A}} + k_{\text{-}2\mathrm{A}}))}{k_{1\mathrm{A}}k_{2\mathrm{A}}(k_3 + k_4 + k_{\text{-}3})}, \\
\frac{K_{2\mathrm{A}}K_{\mathrm{iB}}}{K_{\mathrm{PB}}} &= \frac{k_{\text{-}4}(k_{\text{-}1\mathrm{B}}(k_3 + k_{\text{-}2\mathrm{B}}) + k_{\text{-}3}(k_{\text{-}1\mathrm{B}} + k_{\text{-}2\mathrm{B}}))}{k_{1\mathrm{B}}k_{2\mathrm{B}}(k_3 + k_4 + k_{\text{-}3})}, \\
\frac{K_{1\mathrm{A}}}{K_{\mathrm{iP}}} &= \frac{k_{\text{-}4}(k_3 + k_{\text{-}3})}{k_{1\mathrm{A}}(k_3 + k_4 + k_{\text{-}3})}, \\
\frac{K_{1\mathrm{B}}}{K_{\mathrm{iP}}} &= \frac{k_{\text{-}4}(k_3 + k_{\text{-}3})}{k_{1\mathrm{B}}(k_3 + k_4 + k_{\text{-}3})}.
\end{aligned}
$$

2.4.1 Approximating the TATF in the irreversible limit

As before, the TATF can be approximated in the irreversible limit ($K_{\mathrm{eq}} \to \infty$) by taking the integral in Eqn. 2.19, which yields

$$
TATF_{1-\varepsilon} = \frac{[\mathrm{C}_0]^2 k_{\mathrm{cat}}}{\mathrm{T}_1 + \mathrm{T}_2[\mathrm{C}_0] + \mathrm{T}_3[\mathrm{C}_0]^2} \tag{2.102}
$$

with

$$\mathrm{T}_1 = \frac{\mathrm{c}_1}{\varepsilon}\,, \tag{2.103}$$

$$\mathrm{T}_2 = \mathrm{c}_2\left(\frac{1}{\varepsilon} - \frac{\log\left(\frac{1}{\varepsilon}\right) - \log(f(\varepsilon))}{1-\varepsilon}\right) + \mathrm{c}_3\frac{\log\left(\frac{1}{\varepsilon}\right) - \log(f(\varepsilon))}{1-\varepsilon} + \mathrm{c}_4\frac{\log(f(\varepsilon))}{1-\varepsilon}\,, \tag{2.104}$$

$$\mathrm{T}_3 = 1 + \mathrm{c}_5\left(\frac{\log(f(\varepsilon))}{1-\varepsilon} - 1\right) + \mathrm{c}_6\frac{\log\left(\frac{1}{\varepsilon}\right) - \log(f(\varepsilon))}{1-\varepsilon}\,, \tag{2.105}$$

where

$$\begin{aligned}
\mathrm{c}_1 &= \frac{\Gamma_{1,1}}{\eta_3+\eta_4} = K_{2\mathrm{S}}K_{\mathrm{iS}}\,,\\
\mathrm{c}_2 &= \frac{\Gamma^P_{1,1}}{\eta_3+\eta_4} = \frac{K_{2\mathrm{S}}K_{\mathrm{iS}}}{K_{\mathrm{P}}}\,,\\
\mathrm{c}_3 &= \frac{(\Gamma_{1,2}+\Gamma_{2,1})(\eta_3+\eta_4) - \Gamma_{1,1}(\Gamma_{2,3}+\Gamma_{3,2})}{(\eta_3+\eta_4)^2}\,,\\
&= \frac{k_{1\mathrm{A}}(K_{2\mathrm{B}}K_{\mathrm{iA}} - K_{2\mathrm{S}}K_{\mathrm{iS}}) + k_{1\mathrm{B}}(K_{2\mathrm{A}}K_{\mathrm{iB}} - K_{2\mathrm{S}}K_{\mathrm{iS}})}{\frac{k_{1\mathrm{A}}k_{\text{-}1\mathrm{B}}}{k_{2\mathrm{B}}} + \frac{k_{1\mathrm{B}}k_{\text{-}1\mathrm{A}}}{k_{2\mathrm{A}}}} + K_{2\mathrm{S}}K_{\mathrm{iS}}\left(\frac{k_{1\mathrm{A}}}{k_{\text{-}1\mathrm{A}}} + \frac{k_{1\mathrm{B}}}{k_{\text{-}1\mathrm{B}}}\right)\\
\mathrm{c}_4 &= \frac{\Gamma_{1,3}+\Gamma_{2,2}+\Gamma_{3,1} - (\eta_3+\eta_4)}{\Gamma_{2,3}+\Gamma_{3,2}} + \frac{\Gamma^P_{2,2}(\eta_3+\eta_4)}{(\Gamma_{2,3}+\Gamma_{3,2})^2} - \frac{\Gamma^P_{1,2}+\Gamma^P_{2,1}}{\Gamma_{2,3}+\Gamma_{3,2}}\\
\mathrm{c}_5 &= \frac{\Gamma^P_{2,2}}{\Gamma_{2,3}+\Gamma_{3,2}} = \frac{K_{1\mathrm{S}}}{K_{\mathrm{iP}}}\,,\\
\mathrm{c}_6 &= \frac{(\Gamma^P_{1,2}+\Gamma^P_{2,1})(\eta_3+\eta_4) - \Gamma^P_{1,1}(\Gamma_{2,3}+\Gamma_{3,2})}{(\eta_3+\eta_4)^2}\,,\\
f(\varepsilon) &= \frac{[\mathrm{C}_0](\Gamma_{2,3}+\Gamma_{3,2}) + \eta_3 + \eta_4}{\varepsilon\,[\mathrm{C}_0](\Gamma_{2,3}+\Gamma_{3,2}) + \eta_3 + \eta_4}\,.
\end{aligned}$$

The terms in the factors c_i can be cast in terms of the effective kinetic constants that are defined as

$$\begin{aligned}
K_{2\mathrm{S}}K_{\mathrm{iS}} &= \frac{K_{\mathrm{AB}}(k_3k_4 + (k_4+k_{\text{-}3})(k_{\text{-}2\mathrm{A}}+k_{\text{-}2\mathrm{B}}))}{(k_{\text{-}2\mathrm{A}}+k_{\text{-}2\mathrm{B}})(k_3+k_4+k_{\text{-}3})} \to \frac{k_4K_{\mathrm{AB}}(k_3+k_{\text{-}2\mathrm{A}}+k_{\text{-}2\mathrm{B}})}{(k_3+k_4)(k_{\text{-}2\mathrm{A}}+k_{\text{-}2\mathrm{B}})}\,,\\
K_{\mathrm{P}} &= \frac{k_3k_4 + (k_4+k_{\text{-}3})(k_{\text{-}2\mathrm{A}}+k_{\text{-}2\mathrm{B}})}{k_{\text{-}4}\left((k_3+k_{\text{-}2\mathrm{A}}+k_{\text{-}2\mathrm{B}}) + k_{\text{-}3}\left(1+\frac{k_{\text{-}2\mathrm{A}}}{k_{\text{-}1\mathrm{A}}}+\frac{k_{\text{-}2\mathrm{B}}}{k_{\text{-}1\mathrm{B}}}\right)\right)} \to \frac{k_4}{k_{\text{-}4}}\,,\\
\frac{K_{1\mathrm{S}}}{K_{\mathrm{iP}}} &= \frac{k_{\text{-}4}(k_3+k_{\text{-}3})}{(k_{1\mathrm{A}}+k_{1\mathrm{B}})(k_3+k_4+k_{\text{-}3})} \to \frac{k_3k_{\text{-}4}}{(k_{1\mathrm{A}}+k_{1\mathrm{B}})(k_3+k_4)}\,,\\
K_{1\mathrm{S}} &= \frac{k_3k_4}{(k_{1\mathrm{A}}+k_{1\mathrm{B}})(k_3+k_4+k_{\text{-}3})} \to \frac{k_3k_4}{(k_{1\mathrm{A}}+k_{1\mathrm{B}})(k_3+k_4)}\,,\\
K_{\mathrm{iP}} &= \frac{k_3k_4}{k_{\text{-}4}(k_3+k_{\text{-}3})} \to \frac{k_4}{k_{\text{-}4}}\,,\\
K_{\mathrm{AB}} &= \frac{k_{\text{-}1\mathrm{A}}\,k_{\text{-}2\mathrm{A}}}{k_{1\mathrm{A}}\,k_{2\mathrm{A}}} = \frac{k_{\text{-}1\mathrm{B}}\,k_{\text{-}2\mathrm{B}}}{k_{1\mathrm{B}}\,k_{2\mathrm{B}}}\,,
\end{aligned} \tag{2.106}$$

where K_{1S} is an effective Michaelis constant for the first substrate binding steps, K_{2S} is the effective second substrate Michaelis constant for the second substrate binding steps, both of which are close in form to their ordered mechanism counterparts. K_P is the Michaelis constant of the product, K_{iP} is the competitive product inhibition constant, and K_{AB} is the dissociation constant of the E to EAB transition.

High initial substrate concentrations

In the limit that the initial substrate concentration is high enough to remain saturating throughout the course of the reaction (and thus the product quickly reaches a saturating concentration too), $[C_0] \gg \frac{\eta_3+\eta_4}{\varepsilon(\Gamma_{2,3}+\Gamma_{3,2})}$, the term $f(\varepsilon) \to \frac{1}{\varepsilon}$, and the TATF becomes

$$\lim_{[C_0]\to\infty} TATF_{1-\varepsilon} = \frac{k_{cat}}{1-c_5\left(1-\frac{\log\left(\frac{1}{\varepsilon}\right)}{1-\varepsilon}\right)}, \tag{2.107}$$

$$\lim_{[C_0]\to\infty} TATF_{0.99} = \frac{k_{cat}}{1+3.7\frac{K_{1S}}{K_{iP}}} \approx \frac{k_3k_4}{k_3(1+3.7\frac{k_{-4}}{k_{1A}+k_{1B}})+k_4}, \tag{2.108}$$

which is identical in form to the TATF for an ordered reaction with competitive product inhibition, as in Eqn. 2.61, with an effective kinetic constant K_{1S}. K_{1S} is a function of the sum of the first binding rates for both pathways $k_{1S} = k_{1A} + k_{1B}$. In this limit, the same intuition applies, that in the limit that the reaction isn't rate limited by the polymerisation step ($k_3 \gg k_{-4}$), the reaction is dependent on the formation of the complex EP. Hence, the competition between substrates and the product for binding to the small amount of available enzyme slows the reaction. However, if the polymerisation step is rate limiting, ($k_3 \ll k_{-4}$), then the rate of the reaction is dependent on the formation of the complex EAB and loses dependence on the binding rates of the product and substrates for the free catalyst.

Low initial substrate concentrations

In the limit that the concentration is far below the saturating substrate and product concentrations throughout the reaction, while remaining sufficiently driven to convert the majority of the substrate to the product $\varepsilon^2[C_0] > 1/K_{eq}$, the TATF has a quadratic dependence on $[C_0]$

$$TATF_{1-\varepsilon} = \frac{\varepsilon[C_0]^2k_{cat}}{c_1}, \tag{2.109}$$

$$TATF_{0.99} = \frac{[C_0]^2k_{cat}}{100K_{2S}K_{iS}} = \frac{[C_0]^2k_{1A}k_{2A}k_3}{100k_{-1A}(k_3+k_{-2A}+k_{-2B})} + \frac{[C_0]^2k_{1B}k_{2B}k_3}{100k_{-1B}(k_3+k_{-2A}+k_{-2B})}. \tag{2.110}$$

Here the TATF is given by the sum for each pathway of the product of the occupancy of the first substrate-bound complex ($[EA] \propto [C_0]\frac{k_{1A}}{k_{-1A}}$, $[EB] \propto [C_0]\frac{k_{1B}}{k_{-1B}}$) and the rate of successful binding and polymerisation of the second substrate ($\frac{k_{2B}k_3[C_0]}{k_3+k_{-2A}+k_{-2B}}$,$\frac{k_{2A}k_3[C_0]}{k_3+k_{-2A}+k_{-2B}}$). This form of the TATF closely resembles that given in Eqn. 2.63, for an ordered bi-uni reaction in the non-saturating limit.

Intermediate initial substrate concentrations

At intermediate initial substrate concentrations, the TATF is approximately linearly dependent on $[C_0]$. If $[C_0] > \frac{\eta_3+\eta_4}{\varepsilon(\Gamma_{2,3}+\Gamma_{3,2})}$, the TATF can be approximated as

$$TATF_{1-\varepsilon} = \frac{[C_0]k_{cat}}{\frac{K_{2S}K_{iS}}{\varepsilon K_P} + \frac{c_4 \log\left(\frac{1}{\varepsilon}\right)}{1-\varepsilon}} \tag{2.111}$$

$$TATF_{0.99} = \frac{[C_0]k_{cat}}{100\frac{K_{2S}K_{iS}}{K_P} + 4.7c_4} \tag{2.112}$$

and if $[C_0] < \frac{\eta_3+\eta_4}{\varepsilon(\Gamma_{2,3}+\Gamma_{3,2})}$, the TATF is approximately

$$TATF_{1-\varepsilon} = \frac{[C_0]k_{cat}}{\frac{K_{2S}K_{iS}}{K_P}\left(\frac{1}{\varepsilon} - \frac{\log\left(\frac{1}{\varepsilon}\right)}{1-\varepsilon}\right) + \frac{c_3 \log\left(\frac{1}{\varepsilon}\right)}{1-\varepsilon}}, \tag{2.113}$$

$$TATF_{0.99} = \frac{[C_0]k_{cat}}{95.3\frac{K_{2S}K_{iS}}{K_P} + 4.7c_3} \tag{2.114}$$

Under the simplifying assumption that $100\frac{K_{2S}K_{iS}}{K_P} > 4.7c_4$ and $95.3\frac{K_{2S}K_{iS}}{K_P} > 4.7c_3$, the TATF may be approximated as

$$TATF_{1-\varepsilon} = \frac{[C_0]K_P k_{cat}}{100K_{2S}K_{iS}} = \frac{(k_{1A}k_4)([C_0]k_{2A}k_3)}{100(k_{-1A}k_{-4})(k_3+k_{-2A}+k_{-2B})} + \frac{(k_{1B}k_4)([C_0]k_{2B}k_3)}{100(k_{-1B}k_{-4})(k_3+k_{-2A}+k_{-2B})} \tag{2.115}$$

which also bears a close resemblance to the TATF for the ordered reaction mechanism with competitive inhibition in the same regime in Eqn. 2.62. Here the TATF also separates into a sum over terms for each pathway proportional to the ratio of dissociation constants of the product and each substrate ($K_{iP}/K_{iA} = \frac{k_{1A}k_4}{k_{-1A}k_{-4}}$,$K_{iP}/K_{iB} = \frac{k_{1B}k_4}{k_{-1B}k_{-4}}$) and the rate of binding and successful polymerisation of the second substrate ($\frac{k_{2B}k_3[C_0]}{k_3+k_{-2A}+k_{-2B}}$,$\frac{k_{2A}k_3[C_0]}{k_3+k_{-2A}+k_{-2B}}$).

2.5 Conclusions

In moving toward the goal of engineering a chemical system capable of producing product molecules encoded in catalytic, information-bearing template molecules, we require metrics suited to quantifying their performance. These types of reactions involve more than one substrate, and suffer from various types of product inhibition and a reduction in speed due to the conversion of substrate to product. While there are metrics which capture reactions with some of these qualities [146, 141, 144], they are not entirely suitable. Therefore, in this chapter, I have focused on the development of metrics that can be used to compare templated dimerization reactions (bi-uni enzymatic reactions) that take into account the main factors that could hinder a chemical copying reaction; competitive and non-competitive product inhibition, and the depletion of substrate over the course of a reaction. These qualities are captured by the time-averaged turnover frequency (TATF), which is equal to the catalytic effectiveness per unit catalyst first defined in Ref. [145]. The TATF should be measured at a constant temperature, with an equal initial concentration of each substrate and an absence of product, and with a catalyst concentration that is sufficiently low so that the sequestration of the substrate in solution is negligible. For comparison, the TATF should be measured until a defined fraction of the ultimate yield of the reaction has been reached. Measuring the TATF until a greater fraction of conversion has been reached will exacerbate the slowing effects of substrate depletion and product inhibition.

I calculated the TATF for an ordered bi-uni reaction with competitive product inhibition, an ordered bi-uni reaction with competitive and non-competitive (mixed) product inhibition, and a random order bi-uni reaction with competitive inhibition. In each case, I assumed that the evolution of the concentration of the external species is slow compared to the dynamics of the product-bound and substrate-bound catalyst (intermediate species), and therefore the system evolves as though, at each moment, the intermediate species are in steady-state with one another. Under this condition, the time evolution of the product concentration follows an intermediate steady-state rate law. I took the limit that the ligation or polymerisation step was irreversible and used the corresponding steady-state rate law to calculate the TATF. In each case, I analysed the TATF in the limit that the initial substrate concentration was high enough such that the substrate and product are effectively saturated throughout the course of the reaction, in the limit that the initial concentration is so low that both substrate and product are non-saturating throughout, and at intermediate concentrations in which the substrate concentration starts in a saturated regime and is quickly depleted below the level of saturation, while the product concentration rises to inhibit the reaction.

In the low concentration regime, both the competitively product-inhibited and mixed product-inhibited ordered bi-uni reactions have a TATF that is approximately $\mathrm{TATF}_{0.99} = \frac{[\mathrm{C}_0]^2 k_{\mathrm{cat}}}{100 K_{\mathrm{B}} K_{\mathrm{iA}}}$. In the intermediate concen-

tration regime, both have a TATF that is approximately $\text{TATF}_{0.99} = \frac{[\text{C}_0]K_\text{P}k_\text{cat}}{100K_\text{B}K_\text{iA}}$. In the high concentration regime, the competitively inhibited reaction has a TATF that is approximately $\text{TATF}_{0.99} = \frac{k_\text{cat}}{1+3.7\frac{K_\text{A}}{K_\text{icP}}}$, while the mixed inhibited reaction has a TATF that is approximately $\text{TATF}_{0.99} = \frac{k_\text{cat}}{1+3.7(\frac{K_\text{A}}{K_\text{icP}}+\frac{K_\text{B}}{K_\text{incP}})}$, where K_icP is the competitive inhibition constant and K_incP is the non-competitive inhibition constant. While the maximum TATF of both reactions is less than k_cat, the reaction with mixed inhibition can be far more strongly inhibited by the product, especially in the regime in which the reaction is rate limited by the polymerisation step.

In calculating the TATF of a random order bi-uni reaction, I enforced that the stability of the transition complex EAB was independent of the pathway by which it is reached, which requires that the product of the dissociation constants for each binding pathway must be equal, thereby constraining the binding and unbinding rates. The TATF can be expressed in the low, intermediate and high concentration regimes in a form that is analogous to the ordered mechanism with effective kinetic constants ($K_\text{2S}, K_\text{1S}, K_\text{iS}$) that feature terms which sum over the rates of both binding pathways. In the low concentration regime the $\text{TATF}_{0.99} = \frac{[\text{C}_0]^2k_\text{cat}}{100K_\text{2S}K_\text{iS}}$, in the intermediate concentration regime, the $\text{TATF}_{0.99} = \frac{[\text{C}_0]K_\text{P}k_\text{cat}}{100K_\text{2S}K_\text{iS}}$, and in the high concentration regime the TATF is approximately $\text{TATF}_{0.99} = \frac{k_\text{cat}}{1+3.7\frac{K_\text{1S}}{K_\text{iP}}}$.

Part of the utility of the approach presented in this chapter is that it provides a method to reduce the complexity of the systems (such as a random order reaction) down to a few informative and, crucially, comparable ratios of kinetic constants. The steady-state rate of a random bi-uni reaction with competitive product inhibition only cannot be expressed in the standard form (as a ratio with the typical numerator $k_\text{cat}([\text{A}][\text{B}]-[\text{P}]/K_\text{eq})$ and denominator given by a polynomial of the substrate and product concentrations) and is a function of 17 kinetic parameters. In calculating the TATF, assuming irreversibility, setting the substrate concentrations to be equal, and integrating over the course of the reaction until a certain fraction of the yield has been achieved, the behaviour of the system can be approximated across high, intermediate and low concentration regimes with just four parameters k_cat, $K_\text{2S}K_\text{iS}$, K_1S, and $K_\text{P} \approx K_\text{iP}$, and in functional forms that are near identical to, and hence can readily be compared with, other more tractable reaction mechanisms. I predict that a random order bi-uni mechanism would yield a $TATF_{0.99}$ given by $\frac{[\text{C}_0]^2k_\text{cat}}{100K_\text{2S}K_\text{iS}}$, $\frac{[\text{C}_0]K_\text{P}k_\text{cat}}{100K_\text{2S}K_\text{iS}}$, and $\frac{k_\text{cat}}{1+3.7(\frac{K_\text{1S}}{K_\text{icP}}+\frac{K_\text{2S}}{K_\text{incP}})}$, in the low, intermediate and high concentration regimes respectively, where K_incP is the aggregated non-competitive inhibition constant across both arms of the binding pathway.

While it was out of the scope of this chapter, one could study the TATF where the reaction isn't assumed to be irreversible and could derive metrics based on the incomplete conversion of substrate to product, and could apply the method to the study of higher-order mechanisms with product and non-product inhibitors. Further, systems in which there are multiple types of substrate in solution, such is the case in templated copying reactions, could be analysed. Here, one would want to optimise for the specific production of one product encoded by

the catalyst or template over the course of a reaction. In this setting, if the correct product is formed quickly, the correct monomers will be depleted, the relative concentration of the other incorrect monomer types will be enhanced, and product inhibition will affect the system. This will promote the production of undesirable products and also slow the overall rate of the reaction. Therefore, metrics that capture the time-averaged specificity of the reaction and that can be approximated by simple kinetic constants, and hence can provide mechanistic insights, would be desirable.

In experimental settings, measuring the TATF can provide a simple route to the calculation of useful kinetic constants, and ratios thereof, as noted in Ref. [141]. Measuring the average speed of the reaction over the majority of the time course can offer an alternative to or complement approaches based on initial rate analysis. Serial dilutions of the substrate could be used to characterise the TATF over a range of concentrations, without the need for careful preparation of substrate and product concentrations. In systems that obey random binding patterns, the effective Michaelis constants derived here may prove useful for comparison to ordered mechanisms. If these metrics are used, care should be taken that the substrate concentrations are sufficiently high that a near-complete conversion of substrate to product is achieved. In the context of the theory-led design of chemical copying systems, the numerical calculation of the metrics derived here based upon the estimation of kinetic rate constants could aid in predicting the performance of the systems, and hence aid in identifying optimal reaction designs.

While the metrics developed in this chapter provide a useful framework for analyzing and comparing templated dimerisation reactions, their applicability may be limited to systems with few substrates and products. For longer templates or catalysts with more binding sites, the number of configurations of substrates and partial products bound to the catalyst grows rapidly, and there may be numerous, more complex mechanisms of inhibition (e.g. there are multiple ways a product molecule could bind to each site on the catalyst). Writing down the differential equations that govern such systems, and calculating the steady-state rates and the time-averaged turnover frequencies would become challenging (a computational approach to this problem is used in Chapter 4). Therefore, in Chapter 3, I focus on constructing a bottom-up computational model of a templated copying system that can simulate a system with an arbitrary template length. In the model, I specify the rules (fundamental reactions that can apply in local conformations) and let the system evolve stochastically, according to the rules, sampling the vast space of configurations of monomers and polymers on a template. By gradually introducing more complex interactions (context-dependent rules and rate parameterisations), I search for mechanisms that can achieve length and sequence control in an isothermal environment on arbitrarily long templates.

Chapter 3

A coarse-grained model of whole-template copolymerisation dynamics

The contents of this chapter are adapted from work I published in Ref. [1] with supervision from Thomas Ouldridge and Jenny Poulton.

3.1 Introduction

In the previous chapter of this thesis, I derived metrics to compare the performance of templated dimerisation reactions under various reaction pathways. Product inhibition poses challenges for copying even short templates and presents an even greater challenge for systems that copy longer templates. In this chapter, I focus on identifying molecular mechanisms that could enable faithful copying of very long templates, with the goal of engineering autonomous templated information copiers that function without external intervention, driven solely by chemical free energy.

Unlike the previous chapter, where systems were analytically tractable due to their small state spaces, systems operating on long templates have vast state spaces. Therefore, I do not use the approach of solving for steady-state rate laws here. Instead, I developed a bottom-up model that allows all possible pathways toward product formation on a template of a chosen length. The model is stochastic and considers the mutual inhibitory effects of multiple polymers growing on the same template. However, it does not account for the inhibitory effects of products rebinding from the solution. The inhibitory effects of products rebinding from the solution are secondary to the cooperative effects that prevent product release in the first place; there will be no polymers in solution to inhibit the copying reaction if the system is incapable of producing them from scratch.

In this chapter, I argue that energetically decoupling the length of a polymer from its affinity for the template,

by channelling the free energy released in polymerisation to the disruption of copy-template bonds as achieved by HMSD [22], would be sufficient to overcome product inhibition even on long templates in an isothermal environment. Furthermore, I hypothesize that additionally weakening the copy-template connection at the final template site would result in the selective release of complete copies rather than shorter fragments. I explore these arguments through simulations of coarse-grained models.

In Section 3.2.1, I introduce a basic model of isothermal templated copolymerisation with separation. In Section 3.3.1, I demonstrate that cooperative binding to the template prohibits the release of long polymers under a basic growth mechanism, showing that it is necessary to have more complex interactions between copy and template to promote reliable separation in a constant environment. In Section 3.3.2, I propose and investigate a mechanism that could alleviate the cooperative effect of product inhibition, resulting in an increase in mean polymer length. Extending this mechanism in Section 3.3.3, I bias the production of complete polymers by selectively weakening the final site on the template. Finally, in Section 3.3.4, I demonstrate that long and accurate copies of the template can be generated if the correct and incorrect monomers have different binding rates.

3.2 Model and methods

3.2.1 Model

State space

As shown in Fig. 3.1, I consider a copolymer template $T = t_1 t_2 t_3 ... t_{L_T}$, where t_i is an integer indicating the monomer type that can take values $0, 1, ..., \alpha_T - 1$, where α_T is the number of distinct types of monomeric unit in the template - the size of the monomer alphabet. A single template copolymer is suspended in a large-volume, well-mixed bath of a second distinct type of monomers that interact with the template. The identity of these monomers is labelled by an integer that can take values $0, 1, ..., \alpha_C - 1$. There are α_C distinct types of copy monomer, the i^{th} of which has a concentration $[\mathrm{M}]_i$. These monomers can bind to the template at any unoccupied site on the template. In this chapter, I consider both homogeneous copy-template systems in which there is a single monomer type, $\alpha_T = \alpha_C = 1$, and also information-bearing binary systems in which there are two monomer types for each of the copy and template monomers $\alpha_T = \alpha_C = 2$. Following Ref. [23], I assume that copy/template interactions are symmetric with respect to interchange of 0 and 1 and that the concentrations of monomers 0 and 1 are equal throughout, giving identical dynamics for all template sequences. The template sequence $T = 000...0$ is uniform in all simulations. When the copy-template alphabets are binary, $\alpha_T = \alpha_C = 2$, monomers of type 0 and 1 in the copy copolymer can then simply be interpreted as "correct" and "incorrect" matches, respectively. I refer to all molecules as polymers unless their nature as copolymers is important.

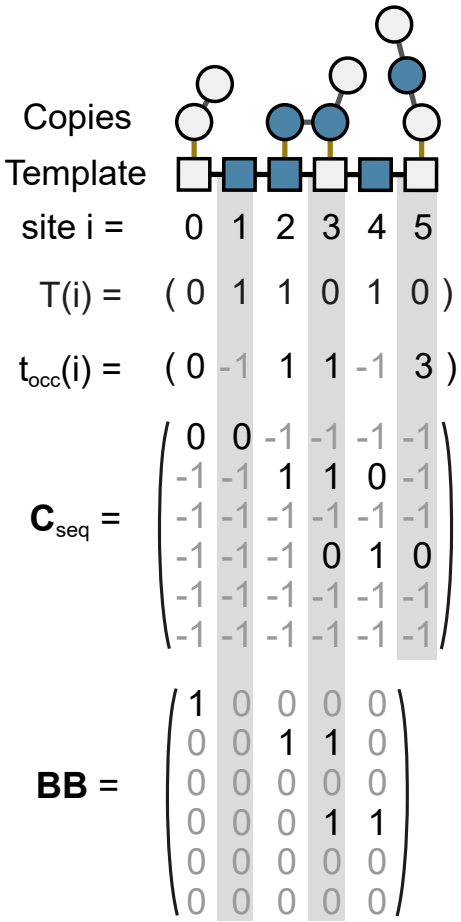


Figure 3.1: The state of the system is described by t_{occ}, $\mathbf{C}_{\mathrm{seq}}$ and $\mathbf{BB}$. Sites on the template are indexed with i. The array $T(i)$ defines the template sequence. t_{occ} is an array containing the polymer label of each copy unit attached to the template. The label defines the row of $\mathbf{C}_{\mathrm{seq}}$ that corresponds to the copolymer's sequence. $\mathbf{BB}$ defines the backbone bonds. In this example, three polymers (labelled 0, 1 and 3) are attached by one, two and one bond/s to the template, respectively. Figure adapted with permission from Ref. [1] under a CC BY license.

A copy unit with no neighbouring backbone bonds is a monomer. As shown in Fig. 3.1, copy polymers are chains of copy units that share a 'backbone' linking one unit to the next. Each copy monomer may form

up to two backbone bonds, one with a copy unit ahead and one with a copy unit behind with the boundary condition that backbone bonds can only be formed ahead at site 0, and behind at site $L_T - 1$. As depicted in Fig. 3.1, copy polymers can occupy contiguous stretches of template sites and may have 'tails', portions of their length that are not directly bound to the template, but are indirectly tethered to the template by bound units with which they share a backbone bond. These tails may be present at either end of the polymer. Copy polymer 'bridges', unbound stretches between bound units within the same copy polymer, are neglected. I assume that such bridging behaviour is strongly suppressed due to the constrained close proximity enforced by the bound units, just as internal bubbles are strongly suppressed between complementary nucleic acid duplexes [147].

Assuming that the concentration of monomer species $[M_i]$ in the bath is constant, the simulations are initiated with no polymers in solution and an empty template (all sites unoccupied). Given the large volume of the bath, the concentration of copy polymers is assumed to remain negligible compared to copy monomer concentrations. After polymers unbind from the template, they diffuse into the bath and do not re-bind to the template. Hence, the problem of product inhibition due to very high concentrations of products is neglected since it is secondary to the inhibition caused by products that fail to detach from the template. The models presented here do not allow polymers to bind to the template from the solution, and bridges cannot form loops causing the contour length of a copy polymer to exceed the length of the template, L_T. Therefore, only copy polymers with lengths $L \leq L_T$ can be produced by the system.

The state of the system is uniquely specified by 1) the sequence of monomers in all growing copy polymers, including those with only one unit (monomers), and 2) specification of the site each monomer is attached to on the template (if it is not attached directly to a template site, the monomer exists in a tail) as outlined in Fig. 3.1. I omit consideration of the possible physical conformations of the system within each of these macrostates.

As I indicate in Fig. 3.1, sites in the system are indexed with i running from 0 to $L_T - 1$. The i^{th} site on the template can be unoccupied or occupied by a copy unit, either a monomer or part of a polymer. Each isolated monomer or polymer on the template is given a unique label. If two copy units share a backbone, even indirectly via other monomers, they must be part of the same polymer, and therefore they share the same label. At most the template can be occupied by L_T distinct monomers or polymers at any moment, and hence the polymer labels run from 0 to $L_T - 1$. I introduce this label to enable easy accounting when extending, fragmenting or recombining polymers. The template occupation labels are defined by a vector t_{occ}, where the i^{th} component, $t_{\text{occ}}(i)$, is the label l of the copy unit bound to the template at site i, or -1 if the site is empty.

The sequence of all polymers - including units in tails- and monomers in the system is are defined in an $L_T \times L_T$ matrix, $\mathbf{C}_{\text{seq}}$, in which the l^{th} row is the sequence of the copy unit with label l, buffered with -1s before and after the sequence. Therefore, $\mathbf{C}_{\text{seq}}(t_{\text{occ}}(i), i)$ is the monomer type of the individual copy unit occupying the

template at site i.

Each copy unit may form up to two backbone bonds, one with a copy unit ahead and one with a copy unit behind, subject to the condition that backbone bonds can only be formed ahead at site 0 and behind at site $L_T - 1$. The $L_T \times L_T - 1$ matrix **BB** defines the state of all backbone bonds in the system. The i^{th} column of the l^{th} row of **BB** is 1 if there is a backbone bond between copy units with label l at sites i and $i+1$ (not necessarily connected to the template), and 0 otherwise.

Transition rules

The transition rules that define the permitted state changes are listed in Fig. 3.2 and illustrated in in Fig. 3.3. All transitions in this model are microscopically reversible, apart from the final detachment of copolymers, since the concentration of products in solution, and hence the rate of polymer rebinding, is assumed to be zero. The transitions that feature in this basic model are as follows:

1. Monomer binding: A copy monomer in solution may diffuse close enough to an unoccupied site i on the template for an attractive interaction to cause it to bind there.

2. Monomer unbinding: The reverse reaction may also occur, in which a monomer bound at site i on the template may unbind and freely diffuse away from the template.

3. Tail unzipping: A unit of a copy polymer that is either bound at the end of a polymer or is the last bound unit before a tail may unbind from the template. This increases the length of the tail by 1.

4. Tail zipping: In reverse, the first unit in the tail of a polymer hovering over site i (which must have a template bound and backbone-linked neighbour at either site $i-1$ or site $i+1$) may bind to site i if the site is not already occupied.

5. Polymer release: If the last bond between a polymer and the template breaks, the polymer is released. This liberated polymer mixes with the large volume and never returns to bind to the template.

6. Polymerisation: Polymerisation can only occur between available, template-bound neighbours. Specifically, if a bound copy unit at site i can make a forward backbone bond to site $i+1$, and a bound copy unit at site $i+1$ is capable of receiving a backbone bond from site i, then a backbone bond may form.

7. Depolymerisation: breaking of backbone bonds can only occur between template-bound copy units that share a backbone bond, a consequence of this being that units in copolymer tails cannot spontaneously depolymerise.

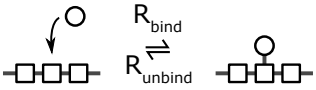
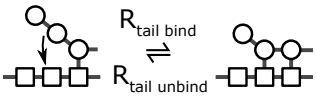
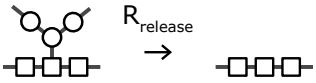
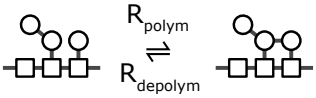
#	Transition	Representation	Free energy change, ΔG	Rates
1 2	Monomer binding Monomer unbinding		ΔG_{spec} - ln[M] = ln(R_{unbind} / R_{bind})	$R_{bind} = k_0 [M]$ $R_{unbind} = k_0 e^{\Delta G spec}$
3 4	Tail binding Tail unbinding		ΔG_{spec} - ln[M_{eff}] = ln($R_{tail\ unbind}$ / $R_{tail\ bind}$)	$R_{tail\ bind} = k_0 [M_{eff}]$ $R_{tail\ unbind} = k_0 e^{\Delta G spec}$
5	Release		Irreversible, ΔG = -∞	$R_{release} = k_0 e^{\Delta G spec}$
6 7	Polymerisation Depolymerisation		ΔG_{BB} = ln($R_{depolym}$ / R_{polym})	$R_{polym} = k$ $R_{depolym} = k e^{\Delta G BB}$

Figure 3.2: The basic model of templated polymerisation includes seven types of transitions, each of which alters the system at a single site or between neighbouring sites during polymerisation. The free energy change linked with each transition restricts the log of the ratio of backward-to-forward transitions based on the principle of local detailed balance. This principle minimises the number of free parameters in the model and ensures that it is thermodynamically self-consistent. The release of a polymer into solution is considered to be irreversible and is associated with an infinite free energy change. Figure reproduced with permission from Ref. [1] under a CC BY license.

I assume that the polymerisation or depolymerisation of molecules not attached to the template in the baths or on the end of free copolymer tails occurs at such a slow rate that it can be neglected.

Concentrations and transition energies

I set the concentration of all types, i, of monomer units in solution to be equal, $[\mathrm{M}]_i = [\mathrm{M}]$. Here, $[\mathrm{M}]$ represents a dimensionless quantity defined relative to an arbitrary reference concentration for all monomer types i. Templates with an over-abundance of a certain monomer type could be accurately copied more easily if the corresponding copy monomer was over-represented in the pool too [24]. However, a biased monomer pool would, on average, provide no advantage for copying an arbitrary template sequence. My focus is on designing copying systems in which any template can be copied with the same efficiency and accuracy as any other template using the same set of monomers, hence setting the monomer concentrations equal is a fair reflection of the conditions that may be used experimentally.

The monomeric units in the tails of the polymers are constrained to a small volume in closer proximity to the template and neighbouring monomers than the free monomers in the baths. Therefore, I define an effective concentration of the first monomer in the tail, $[\mathrm{M_{eff}}] \gg [\mathrm{M}]$, which captures this close proximity.

Bond formation is parameterized through the standard free-energy change of reactions at the reference

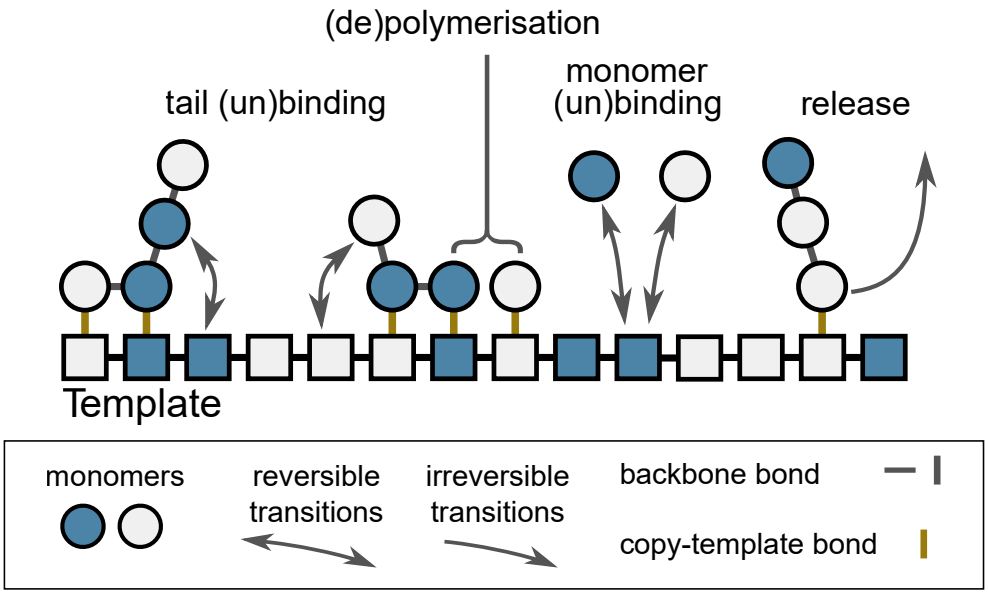


Figure 3.3: Simple transition rules between coarse grained states of a templated polymerisation process. Each site on the template can be occupied by one monomeric unit. Monomers may bind to and unbind from the template from solution and from the tails of copolymers. Neighbouring units on the template may polymerise or depolymerise. All transitions, aside from the final release of a complete polymer from the template, are reversible, and can occur at any site on the template. Figure reproduced with permission from Ref. [1] under a CC BY license.

concentration ($[\mathrm{M}] = 1$). In this chapter, I define all free energies as dimensionless quantities relative to a thermal energy unit of unity $kT = 1$. The formation of the backbone bond between monomers in isolation is associated with a standard free-energy change of $\Delta G_{\mathrm{BB}} + \ln[\mathrm{M_{eff}}]$; ΔG_{BB} reflects the chemical bond strength and $\ln[\mathrm{M_{eff}}]$ the loss of entropy associated with binding from solution. More negative values of ΔG_{BB} favour polymer growth.

Initially, in Section 3.3.1, I consider a simple model of cooperative copy-template interactions that are depicted in Fig. 3.2. In this picture, individual monomers bind to the template with a standard free energy change of ΔG_{spec} – called "specific" because it depends on whether the copy and template monomers are complementary. The standard free-energy change for binding of a copy polymer of length l, forming exactly l bonds with the template is

$$\sum_{i=0}^{l-1} \Delta G^{i}_{\mathrm{spec}} - (l-1)\ln\left([\mathrm{M_{eff}}]\right), \tag{3.1}$$

where $\Delta G^{i}_{\mathrm{spec}}$ is the specific bond free energy of the i^{th} copy-template pair. Here the factor $(l-1)\ln[\mathrm{M_{eff}}]$ captures the cooperative nature of the bond between a long template and a long copy that is responsible for product inhibition. This formulation is a simple model of cooperativity: a copolymer of length l gains l favourable bonds of the same strength but, for $l-1$ of those bonds, the entropic cost is reduced enormously relative to binding from solution. In Section 3.3.2 I extend the model to allow for less cooperative behaviour, and in 3.3.3 I consider a model in which monomers bind less strongly at the end of the template.

In the majority of this chapter, I consider a uniform template where there is one type of copy and template monomer, and hence the specific bond strength, ΔG_{spec} takes one value. In the final section of this chapter, Section 3.3.4, I consider a model with a binary copy monomer alphabet with $\alpha_C = 2$.

Parameterisation of transition rates

I assume that each transition described in Section 3.2.1 and depicted in Fig. 3.2 is well described as an instantaneous process with an average rate. Since the thermodynamics of copy production is an important feature underlying its physics [26], I ensure that the thermodynamics of the system is self-consistent by applying the principle of local detailed balance to set the transition rates [129]. The principle of local detailed balance states that the ratio of forwards to backwards rates between any pair of states is constrained by the chemical free-energy change associated with the forwards transition ΔG,

$$\frac{R_{\text{forward}}}{R_{\text{reverse}}} = e^{-\Delta G}. \tag{3.2}$$

I use this equation to parameterise the transition rates between states in the system, as shown in Fig. 3.2.

Monomer binding and unbinding is associated with a free-energy change of $\Delta G = \Delta G_{\text{spec}} - \ln[\text{M}]$, since one specific bond is formed in the process of monomer binding. The ratio of the monomer binding rate R_{bind} to the monomer unbinding rate R_{unbind} is given by

$$\frac{R_{\text{bind}}}{R_{\text{unbind}}} = [\text{M}] e^{-\Delta G_{\text{spec}}}. \tag{3.3}$$

I begin by assuming that monomers bind to the template sites with mass action kinetics, giving $R_{\text{bind}} \propto [\text{M}]$, and that the monomers unbind from the template with a rate that is exponentially dependent on the strength of the specific bond ΔG_{spec}, and hence I arrive at

$$R_{\text{bind}} = k_0 [\text{M}] \text{ and } R_{\text{unbind}} = k_0 e^{\Delta G_{\text{spec}}}. \tag{3.4}$$

In Section 3.3.4, I consider the competition between two types of monomer and compare the effects of two different ways of parameterising the unbinding and binding rates, in which either unbinding rate or the binding rate is exponentially dependent on ΔG_{spec}. Prior to Section 3.3.4, I maintain the parameterisation given above for simplicity.

The free-energy change associated with copy units in the tails at either end of polymers binding to a template

site is $\Delta G = \Delta G_{\text{spec}} - \ln[\text{M}_{\text{eff}}]$. Therefore, the rates of tail binding and unbinding are constrained by

$$\frac{R_{\text{tail bind}}}{R_{\text{tail unbind}}} = [\text{M}_{\text{eff}}] e^{-\Delta G_{\text{spec}}}. \tag{3.5}$$

Following a similar argument to that for monomers binding from solution, I set the rate that a unit in a copolymer tail rebinds to an available site on the template to be

$$R_{\text{tail bind}} = k_0 \, [\text{M}_{\text{eff}}]. \tag{3.6}$$

The rate at which polymer units unbind from a template site to become part of a tail is then identical to the rate at which a monomer in the same position would detach,

$$R_{\text{tail unbind}} = k_0 \, e^{\Delta G_{\text{spec}}}. \tag{3.7}$$

In this model, the final detachment and release of a copolymer into solution is irreversible. Release occurs when the last copy-template bond is broken, and therefore I set the rate of release to equal the rate at which monomers unbind from the template, giving

$$R_{\text{release}} = k_0 \, e^{\Delta G_{\text{spec}}}. \tag{3.8}$$

Polymerisation between units attached to the template at site i and $i+1$ is associated with a free energy change of ΔG_{BB}. Therefore

$$\frac{R_{\text{polym}}}{R_{\text{depolym}}} = e^{-\Delta G_{\text{BB}}}. \tag{3.9}$$

I choose $R_{\text{polym}} = k$, and therefore $R_{\text{depolym}} = k e^{\Delta G_{\text{BB}}}$. Under this parameterisation, increasing the backbone bond strength increases the average lifetime of a backbone.

With these assignations for the free-energy change of each reaction, the overall free-energy change of incorporating a single monomer into a polymer in solution is

$$\Delta G_{\text{pol}} = \Delta G_{\text{BB}} - \ln \frac{[\text{M}]}{[\text{M}_{\text{eff}}]}, \tag{3.10}$$

as required.

3.2.2 Methods

Gillespie simulation

I simulate the dynamics of this system using Gillespie's kinetic Monte Carlo simulation method [128] with two rounds of stochastic sampling. The first round of sampling selects the template site to update, with a probability proportional to the sum of transition rates at that site. The second round selects the transition rule to apply at that site with a probability proportional to its individual transition rate. This two-stage sampling method is advantageous because the valid transitions of the system only change the rates for subsequent transitions in a localised region of the system. For instance, a monomer binding the template at site 3 will not change the conformation of the system at site 10 or 100 and hence doesn't change the valid transitions around site 10 or 100. Given the locality of the model, I need only to recalculate the valid transitions and sum over the valid rates at the few sites on either side of the site which was last updated at each step.

Parameters, initial conditions and stopping criteria

I use dimensionless units, setting $k_B T = 1$ and $k_0 = 1$. I also use $[\mathrm{M_{eff}}] = 100$ throughout. I vary the free parameters k, $[\mathrm{M}]$, ΔG_{spec}, ΔG_{BB}, and ΔG_{gen} and ΔG_{end} that will be introduced later. Unless stated otherwise, the system's initial condition (conformation) in each simulation is an empty template. I simulate systems with templates of length $L_T = 10, 30$ and 100. I calculate statistics of the copying process, such as the mean polymer length and, later when considering systems with two distinct monomer types, the error rate (defined as the number of mistakes contained within a polymer divided by its length), over the first fixed number of polymers (polymers with length ≥ 2) that form upon and are released from the template. I take averages of quantities such as the mean polymer length by running multiple simulations with the same parameters, but different random number seeds. Where the number of polymers produced by the system (the sample size) generated within the allotted compute-time for each simulation is less than 10, the data are excluded as they are unrepresentative of steady state. In each case, I run around 10 independent repeats for each set of parameter values, with each independent simulation stopping after producing on the order of 1000 polymers, though this varied from case to case. The code and input files required to reproduce the data presented in this chapter are freely accessible through the link provided under Data Availability at the end of this chapter.

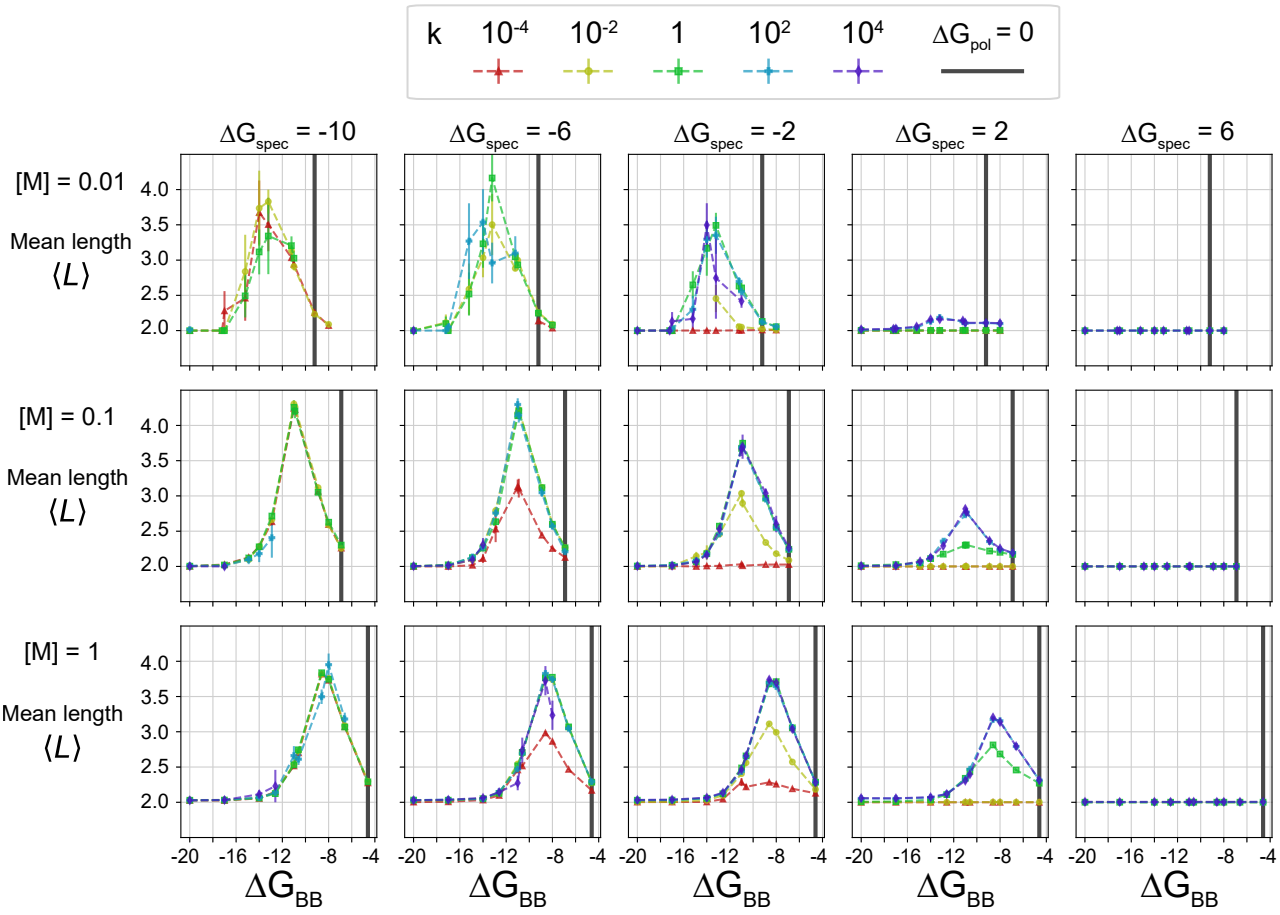


Figure 3.4: For the basic model of Fig. 3.2, a wide search over multiple axes of the parameter space shows that the mean polymer length is always low. Each graph shows an average of the mean polymer length produced by the system, $\langle L \rangle$, with SEM error bars, on a template of length $L_T = 30$, against ΔG_{BB}, with $-\Delta G_{\mathrm{BB}}$ being the driving force behind backbone bond formation. The monomer concentration [M] is varied between rows, and the specific bond strength, ΔG_{spec}, between the columns, while the polymerisation rate, k, is varied and represented with different colours and marker shapes. Error-bars are smaller than line width where not visible. A dark grey line gives the value of ΔG_{BB} at which $\Delta G_{\mathrm{pol}} = 0$. In all cases, $\langle L \rangle \ll L_T$, although a small peak in $\langle L \rangle$ is observed at moderate values of ΔG_{BB}. Figure reproduced with permission from Ref. [1] under a CC BY license.

3.3 Results

3.3.1 A simple model of copy-template binding and copy-copy polymerisation cannot reliably produce long polymers

I begin by asking whether the simple model introduced in Section 3.2.1, with transitions parameterised as in Fig. 3.2, is capable of producing long polymers consistently. In Fig. 3.4, I analyse the polymers produced by the system on a template with length $L_T = 30$. The mean polymer length $\langle L \rangle$ is shown against the free energy driving backbone formation on the template, ΔG_{BB} as I vary the polymerisation rate $k = [10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}]$, the monomer concentration $[\mathrm{M}] = [0.01, 0.1, 1]$, the specific copy-template bond

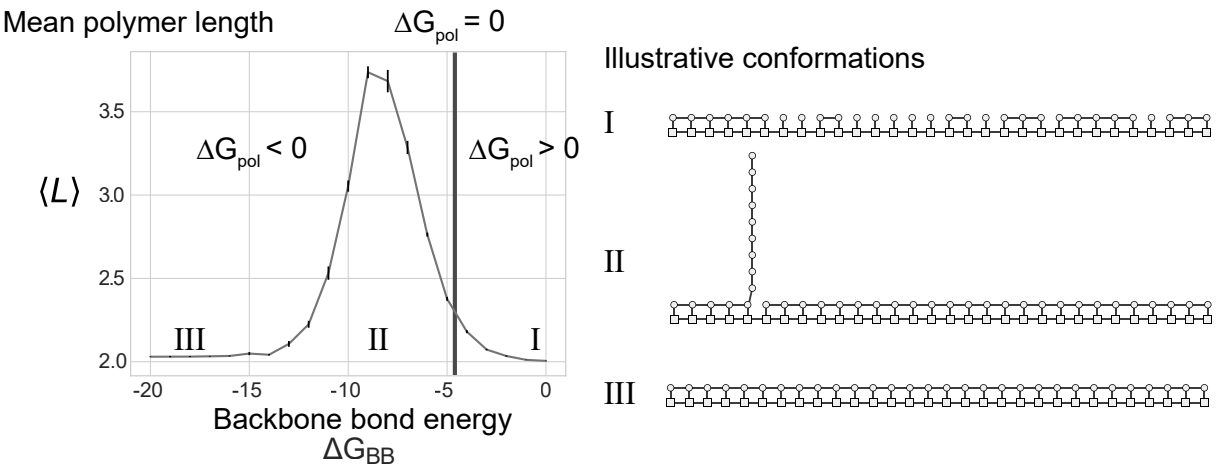


Figure 3.5: Typical configurations help to explain the observed mean length $\langle L \rangle$ of products in the simple model of templated polymer growth. $\langle L \rangle$ is plotted against ΔG_{BB} for a template of length 30, $\Delta G_{\mathrm{spec}} = -4$, $[\mathrm{M}] = 1$ and $k = 1$, showing a moderate peak surrounded by regions of $\langle L \rangle = 2$. I extract typical configurations at the labelled values of ΔG_{BB}: **(I)** Weak backbone, many small, fragmented polymers; **(II)** Moderate backbone strength, a small number of longer fragments; **(III)** Strong backbone, a single long polymer covers the whole template. Figure reproduced with permission from Ref. [1] under a CC BY license.

strength $\Delta G_{\mathrm{spec}} = [-10, -6, -2, 2, 6]$, and backbone strength ΔG_{BB} ranging from -4.6 to -20. Five independent simulations were run for each parameter value, with a target of producing 2000 polymers. When parameter values were extreme, long embedded Markov processes were observed that were time-consuming to simulate. Refer to Section A.1 for a discussion of these parameter values.

Fig.3.4 shows that the system produces polymers with a low mean length, $2 \lesssim \langle L \rangle \lesssim 4$, across the entire parameter space. In order to understand why, I examine regions of the parameter space and analyse the conformation of the growing polymers on the template. To aid in the analysis, I consider a slice through the parameter space and focus on the specific case where $\Delta G_{\mathrm{spec}} = -4$, $[\mathrm{M}] = 1$, and $k = 1$, while keeping other parameters the same as in Fig.3.4. I average over 5 independent repeats of simulations that each generated 1000 polymers at each data point as the backbone bond strength ΔG_{BB} is varied. Fig. 3.5 presents the results and includes characteristic snapshots from the simulations.

In some parameter regions, a template-bound polymer is an unstable conformation and hence only short-lived dimers are observed. The instability may be due to either the dissociation of polymers into monomers free in solution ($\Delta G_{\mathrm{BB}} + \Delta G_{\mathrm{spec}} - \ln[\mathrm{M}] > 0$) or depolymerisation of template bound polymers ($\Delta G_{\mathrm{BB}} > 0$). These regions don't feature heavily in Fig. 3.4 and Fig. 3.5, but are consistent with $\langle L \rangle \to 2$ as ΔG_{BB} gets more positive as the only "polymers" produced are transiently bound dimers.

As ΔG_{BB} becomes stronger (more negative), longer polymers form on the template. An example configuration is shown in Fig. 3.5 I. The products still tend to have a mean length $\langle L \rangle \approx 2$, because although longer copies

are present on the template in significant numbers, they tend to detach more slowly than dimers due to the cooperative bond with the template. Thus dimers dominate the distribution of products.

At extremely negative values of ΔG_{BB}, configurations such as Fig. 3.5 **III** are obtained. In this limit, a single long polymer covers the template. Strong backbone bonds and the cooperativity with which the long copy binds to the template contribute to the stability of this conformation. However, this cooperativity does not prevent the ends of the long polymer from "fraying" (undergoing transitory detachment from the template) which allows dimers to form under the raised tails (see Fig. 3.6); these dimers are formed and released into solution much faster than the longer polymer, and are therefore dominant in the product length distribution.

For some sets of parameters, a moderate peak in $\langle L \rangle$ is observed between these limits. In this case, although ΔG_{BB} is negative, backbone bonds within the copies are broken at an appreciable rate, forming fragmented polymers of moderate length on the template. One fragment can displace the other from the template in a step-by-step fashion, via a mechanism of fraying and polymerisation at the junction between fragments. An intermediate state in this process is shown in Fig. 3.5 **II**. This peak occurs at slightly negative values of $\Delta G_{\mathrm{pol}} < 0$ (left of the dark grey vertical line as shown in Fig. 3.5) , which can be rationalised by noting that adding a single monomer to the tail of template-attached polymer is associated with a free-energy change of ΔG_{pol}, and so states with long tails are thermodynamically unfavourable for $\Delta G_{\mathrm{pol}} > 0$.

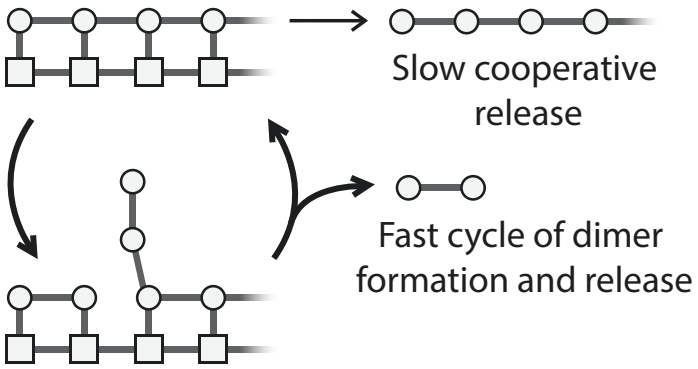


Figure 3.6: Production of dimers in the limit of strong backbone bonds. Long polymers form on the template when the backbone bond ΔG_{BB} is sufficiently strong. Cooperative binding slows the complete release of the long polymer, while dimers are formed and released under its fraying ends. Figure reproduced with permission from Ref. [1] under a CC BY license.

Although this mechanism can provide some increase in $\langle L \rangle$, it does not offer a route to reliable template copying. Here a longer polymer is more likely to push a shorter one off the template and no complete polymers are produced. Importantly, even the systems that give $\langle L \rangle > 2$ still produce outputs that are much shorter than the template L_T. Either template-bound polymer formation is too unfavourable or cooperative template binding prevents the release of long stable polymers (product inhibition). Putative displacement mechanisms can be initiated at any location and produce fragmented polymers with a bias towards shorter products. These results suggest that, for systems that are well-described by the basic model of Section 3.2.1, the basic templating mechanism is incapable of generating and, crucially, releasing long polymers reliably. Indeed, in Ref. [105], in a

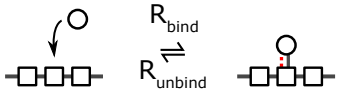
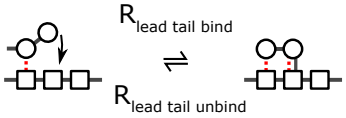
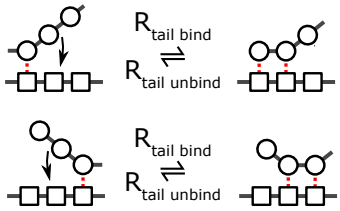
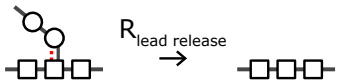
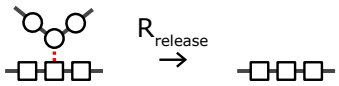
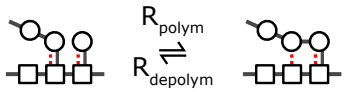
#	Transition	Representation	Free energy change, ΔG	Rates
1 2	Monomer binding Monomer unbinding		$\Delta G_{spec} + \Delta G_{gen} - \ln[M]$ $= \ln(R_{unbind} / R_{bind})$	$R_{bind} = k_0 [M]$ $R_{unbind} = k_0 e^{\Delta G_{gen} + \Delta G_{spec}}$
3 4	Leading tail binding Leading tail unbinding		$\Delta G_{spec} + \Delta G_{gen} - \ln[M_{eff}]$ $= \ln(R_{lead\ tail\ unbind} / R_{lead\ tail\ bind})$	$R_{lead\ tail\ bind} = k_0 [M_{eff}]$ $R_{lead\ tail\ unbind} = k_0 e^{\Delta G_{gen} + \Delta G_{spec}}$
3′ 4′	Tail binding Tail unbinding		$\Delta G_{spec} - \ln[M_{eff}]$ $= \ln(R_{tail\ unbind} / R_{tail\ bind})$	$R_{tail\ bind} = k_0 [M_{eff}]$ $R_{tail\ unbind} = k_0 e^{\Delta G_{spec}}$
5	Leading release		Irreversible, $\Delta G = -\infty$	$R_{lead\ release} = k_0 e^{\Delta G_{gen} + \Delta G_{spec}}$
5′	Release		Irreversible, $\Delta G = -\infty$	$R_{release} = k_0 e^{\Delta G_{spec}}$
6 7	Polymerisation Depolymerisation		$\Delta G_{BB} - \Delta G_{gen}$ $= \ln(R_{depolym} / R_{polym})$	$R_{polym} = k$ $R_{depolym} = k e^{\Delta G_{BB} - \Delta G_{gen}}$

Figure 3.7: The generic bond model is an extension of the basic model presented in Fig. 3.2. Here there are two kinds of copy-template bond, the specific bond (dashed red line) that is potentially sensitive to the match between copy and template units, and the generic bond (solid black line), that is not sensitive to the match. The generic bond is formed during monomer binding and breaks upon polymerisation, and *vice versa* for the reverse reactions. Units at the leading edge of copy polymers (at the right-hand edge) bind to the template more strongly than other monomers in copies due to the modified polymerisation mechanism. The generic bond free energy, ΔG_{gen}, like the specific bond free energy, ΔG_{spec}, doesn't contribute to the overall free energy change of polymer extension ΔG_{pol}. Figure reproduced with permission from Ref. [1] under a CC BY license.

bottom-up, coarse-grained model of a replicating system that obeyed similar dynamics to that described by the model in Fig. 3.2, the authors observed the formation of only long polymers that remained bound in complexes under isothermal conditions.

In the next section, I extend the model to allow for a more complex polymerisation mechanism. This scheme can be viewed simultaneously as a way to avoid product inhibition and a way to favour effective displacement mechanisms, by disrupting the cooperativity of copy-template interactions in a directional manner.

3.3.2 A polymerisation mechanism that disrupts copy-template binding allows for long copies

In the context of uni-uni catalysts, Ref. [148] argued that product inhibition can be reduced if part of the free energy of product formation is diverted into destabilizing the interaction of catalyst and product. With that idea in mind, I develop a model in which backbone bonds in the copy can only be formed at the expense of breaking a bond between copy and template. The scheme, illustrated in Fig. 3.7, could describe a range of chemical systems with competitive bond formation, but in particular, resembles the handhold-mediated strand displacement mechanism introduced by Cabello-Garcia *et al.* [22].

Formally, I implement this mechanism by splitting the copy-template bond into two parts, one of which must be broken for a backbone bond to form. The free-energy change of monomers binding to the template from solution is now given by $\Delta G = \Delta G_{\text{spec}} + \Delta G_{\text{gen}} - \ln[\text{M}]$, where the generic bond ΔG_{gen} is not dependent on the match between copy and template. I assume that the polymers have a directional asymmetry (as is typical in macromolecular polymers like DNA and RNA [2]), and represent this in diagrams such as Fig. 3.7 by drawing the specific bond (dashed) on the left and the generic bond (solid) on the right. This asymmetry will be retained throughout the chapter.

Reparameterisation

As shown in reactions 6 and 7 of Fig. 3.7, the generic bond in the leftmost monomer breaks when polymerisation occurs between neighbouring monomers, leaving the leftmost monomer less tightly bound to the template. In visual representations, polymerisation will always disrupt the bond of the left or 'lagging' monomer; the generic bond of the right or 'leading' monomer is unaffected. The total free-energy change of the template-attached polymerisation step is then $\Delta G = \Delta G_{\text{BB}} - \Delta G_{\text{gen}}$. ΔG_{gen}, like ΔG_{spec}, does not contribute to the the overall free energy change of extending a polymer tail (or a polymer in solution), $\Delta G_{\text{pol}} = \Delta G_{\text{BB}} - \ln[\text{M}]/[\text{M}_{\text{eff}}]$. I assume that the disruption of the generic bond during polymerisation happens at the same time as the polymerisation reaction; *i.e.,* the bond is transferred from template to polymer via an "invasion" mechanism, as occurs in Ref. [22].

Having modified the thermodynamics of the model, I modify the kinetics as follows. As shown in Fig. 3.7, generic bonds are only formed during the binding of monomers from solution, when the leading monomer in a copy polymer rebinds to the template from a tail state, or when the backbone between two template-attached monomers is disrupted. For simplicity, I assume that all unbinding reactions that break a generic bond have rates that scale as $R_{\text{unbind}} \propto e^{\Delta G_{\text{gen}}}$ (so ΔG_{gen} does not appear in the binding rates), and that $R_{\text{depolym}} = k e^{\Delta G_{\text{BB}} - \Delta G_{\text{gen}}}$, $R_{\text{polym}} = k$.

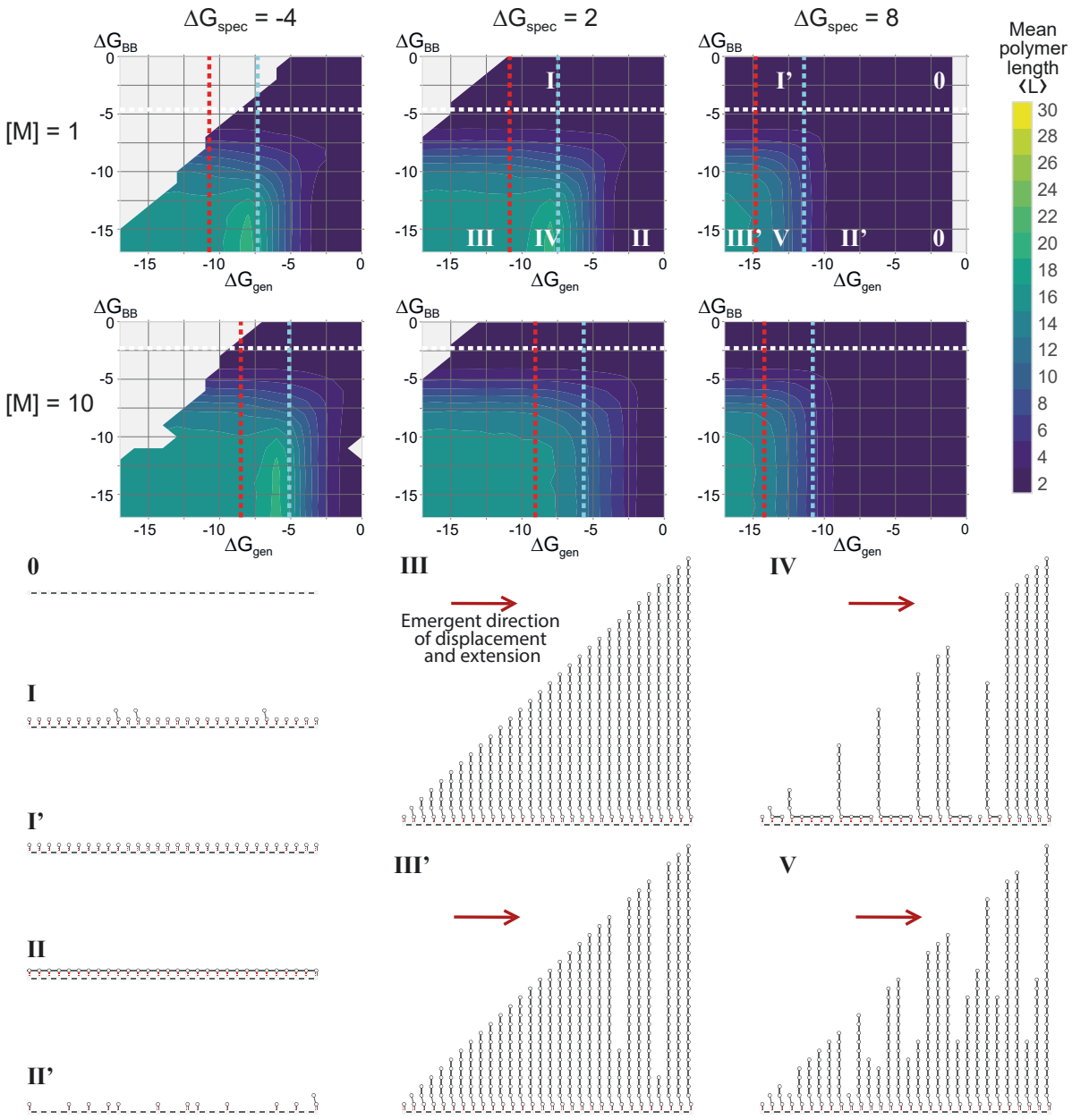


Figure 3.8: Coupling polymerisation to copy-template bond disruption produces copies with $\langle L\rangle \sim L_T/2$. Surface plots of $\langle L\rangle$ versus ΔG_{BB} and ΔG_{gen} for $[\mathrm{M}] = 1, 10$ and $\Delta G_{\mathrm{spec}} = -4, 2, 8$ on a $L_T = 30$ template are shown. Snapshots of the system at $t = 200000$ for different parameter points are also provided. **(White dashed line constraint)**: $\Delta G_{\mathrm{BB}} < \ln([\mathrm{M}]/[\mathrm{M_{eff}}])$. If the constraint is not satisfied, the template is either occupied by monomers and dimers (**I** and **I'**) or is empty **0** because the backbone bond is too weak, hence, the system produces only dimers $\langle L\rangle \approx 2$. **(Dashed cyan line constraint)**: $\Delta G_{\mathrm{gen}} < -\Delta G_{\mathrm{spec}} + \ln \frac{2k_0}{L_T \langle \tau_{\mathrm{disp}}\rangle}$. If satisfied, the generic bond is strong enough that a lagging polymer tends to step forward sooner than it detaches. If not, either the template is empty at **0** since everything detaches rapidly or, when the specific bonds are stronger, the template is occupied by a long polymer, since invading copies cannot force it off the template (configuration **II**). Here the system still produces dimers $\langle L\rangle \approx 2$. **(Dashed red line constraint)**: $\Delta G_{\mathrm{gen}} < -\Delta G_{\mathrm{spec}} + \ln \frac{2k_0}{L_T^2 \langle \tau_{\mathrm{disp}}\rangle}$. To the left of this constraint, lagging polymers step forwards fast enough to create a dense, orderly polymer brush, as in **III** and **III'**. A uniform product length distribution is observed with a mean length $\langle L\rangle = L_T/2$. Between the red and cyan constraints, at **IV**, slower advancement of the lagging polymers leads either to a less dense brush with some tails occupying the template **IV**, or the appearance of gaps that initiate shorter polymers within the brush **V**. In configurations such as **IV**, $\langle L\rangle \gtrsim L_T/2$ is observed, whereas for parameters that lead to configurations like **V**, the mean length $\langle L\rangle < L_T/2$ but is still relatively large. Figure reproduced with permission from Ref. [1] under a CC BY license.

Results

In Fig. 3.8, I present multiple contour plots of the averaged mean length of polymers produced from a template of length $L_T = 30$ as the backbone bond strength ΔG_{BB} and the generic bond strength ΔG_{gen} take values in the range $[0, -1, ..., -17]$. For each point sampled, I ran 5 simulations each with a different random number seed. Each independent simulation generated 1000 polymers. I calculated the average of the mean length of the 1000 polymers across the 5 independent simulations. Data are excluded for simulations that didn't produce 1000 polymers within the allowed window. I set $k = 1$, and consider $[\mathrm{M}] = [1, 10]$, $\Delta G_{\mathrm{spec}} = [-4, 2, 8]$. Additionally, I provide snapshots showing the wide variety of conformations reached by the system at selected points in the parameter space when the simulation time passed $t = 200,000$ units.

In Fig. 3.8, a high mean length of around $L_T/2 = 15$ is achieved over large regions of parameter space. This behaviour corresponds to a fully-occupied template, with a "brush" of polymers attached to it (regions exemplified by **III**, **III'**, **IV** and **V**). In all of these regimes, coupling polymerisation to the disruption of generic copy-template bonds has two important effects.

Firstly, provided the disrupted bond is strong enough, the mechanism can overcome the cooperative binding of the tails of polymers to the template. The leading monomer (rightmost) of a copy binds strongly to the template, but the monomers in the tails bind much more weakly. Undesirable conformations in which a single, long polymer is bound to the template are prevalent (see configuration **II** in 3.8) when the generic bond is weak and the specific bond is strong, as the bond weakening mechanism isn't sufficiently strong to overcome the cooperativity of tail binding.

Secondly, the asymmetry introduced by the generic bond solves the problems associated with the unbiased strand displacement mechanisms mention in Section 3.3.1. The leading edge of a lagging polymer out-competes the tail of the polymer ahead by binding to the template with greater strength at their junction. This enables the system to act as a ratchet in which biased strand displacement and extension drives shorter polymers to push the tails of longer polymers ahead of them off of the template. The emergent direction of displacement and extension is indicated in Fig. 3.8 by a red arrow. As shown in configuration **III** of 3.8, the result is a brush of polymers across the length of the template, each connected to the template by a single, strong bond at their leading edge. The random detachment of polymers from this brush produces a distribution with an average polymer length $\langle L \rangle \sim L_T/2 =\sim 15$.

Some simple physical reasoning can explain the regions of parameter space in which various behaviours are observed. For long polymer copies to form, the incorporation of monomers into copy polymers must be thermodynamically favourable [24]. Thus $\Delta G_{\mathrm{pol}} = \Delta G_{\mathrm{BB}} - \ln([\mathrm{M}]/[\mathrm{M}_{\mathrm{eff}}]) < 0$, which requires a sufficiently

strong backbone:

$$\Delta G_{\mathrm{BB}} \ll \ln([\mathrm{M}]/[\mathrm{M}_{\mathrm{eff}}]). \tag{3.11}$$

The line representing this constraint is shown in white in Fig. 3.8. Above this line, the system exhibits either an empty template as in configuration **0** if the copy-template interaction is weak ($\Delta G_{\mathrm{spec}} + \Delta G_{\mathrm{gen}} > \ln[\mathrm{M}_{\mathrm{eff}}]$); or a template largely covered in monomers, as in **I** and **I'**, if the copy-template interaction is strong ($\Delta G_{\mathrm{spec}} + \Delta G_{\mathrm{gen}} < \ln[\mathrm{M}_{\mathrm{eff}}]$). When the constraint of Eq. 3.11 is not satisfied, the system is only capable of producing short polymers, as in Section 3.3.1 and Fig. 3.5, because long polymers are unstable structures.

Given a sufficiently stable backbone, long polymers can form. However, the strength of the generic bond ΔG_{gen} can drastically change the conformations that these long polymers settle into, thereby changing the typical length of polymers that are produced by the system. When the generic bond is weak, as in **II** where $\Delta G_{\mathrm{gen}} = -2$, single, long polymers are stuck on the template, and dimers are created and released under the fraying end of the polymer as in 3.3.1. Note that when $\Delta G_{\mathrm{gen}} = 0$, the system is identical to the model presented in Section 3.3.1. When the generic bond is strong and takes very negative values, as in **III** and **III'**, I observe conformations where the polymer tails form a dense brush, and the mean length is roughly half the template length $\langle L \rangle = L_T/2$. I also observe regions with configurations like **IV** and **V**, in which the copy polymers form a less dense, less regular brush.

Two key timescales determine whether a brush of long polymers will form. The first is the time taken for a polymer to extend forward into a space that is occupied by the lagging edge of the polymer in front. I call this time $\langle \tau_{\mathrm{disp}} \rangle$, since the lagging tail of the polymer in front is effectively displaced by the growth of the polymer behind (although in many cases its binding is weak to begin with). The second is the timescale on which the leading edge of a polymer detaches, $\langle \tau_{\mathrm{unbind}} \rangle$. I estimate $\langle \tau_{\mathrm{disp}} \rangle$ with the simple discrete-state, continuous-time model in Fig. 3.9. I depict the four states involved in a single forward step alongside their transition rates. Treating this subprocess in isolation, I obtain a rate matrix **K** with entries:

$$\mathbf{K} = \begin{pmatrix} -k_0 e^{\Delta G_{\mathrm{spec}}} & k_0[\mathrm{M}_{\mathrm{eff}}] & 0 & 0 \\ k_0 e^{\Delta G_{\mathrm{spec}}} & -k_0[\mathrm{M}_{\mathrm{eff}}] - k_0[\mathrm{M}] & k_0 e^{\Delta G_{\mathrm{gen}} + \Delta G_{\mathrm{spec}}} & 0 \\ 0 & k_0[\mathrm{M}] & -k_0 e^{\Delta G_{\mathrm{gen}} + \Delta G_{\mathrm{spec}}} - k & 0 \\ 0 & 0 & k & 0 \end{pmatrix}$$

The mean first-passage time $\langle \tau_{\mathrm{disp}} \rangle$ from state 0 to state 3 is given by the $(3,0)$ element of the Drazin inverse of **K**, which takes a value

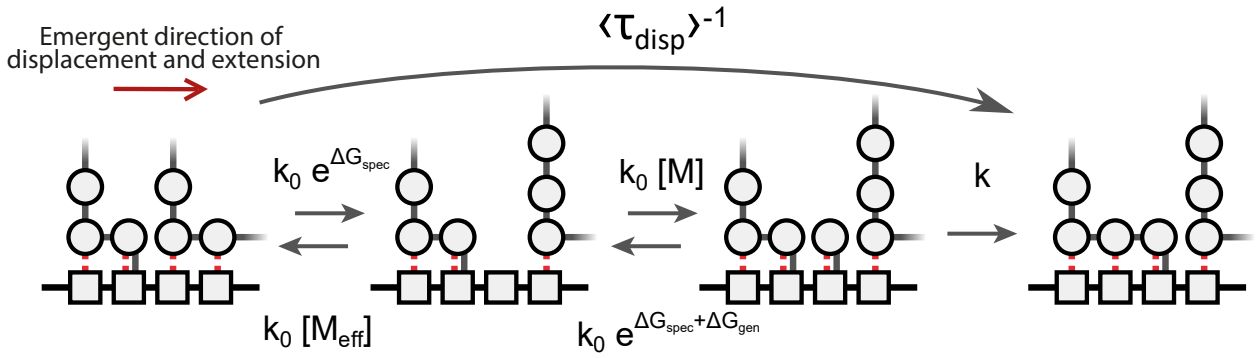


Figure 3.9: Model to estimate the time, $\langle\tau_{\text{disp}}\rangle$, taken for a lagging polymer to extend by one unit, displacing the tail of the polymer ahead. Figure reproduced with permission from Ref. [1] under a CC BY license.

$$\begin{aligned}\langle\tau_{\text{disp}}\rangle =&(k(1+([\text{M}]+[\text{M}_{\text{eff}}])e^{-\Delta G_{\text{spec}}}) \\ &+k_0(e^{\Delta G_{\text{gen}}+\Delta G_{\text{spec}}}+[\text{M}]+e^{\Delta G_{\text{gen}}}[\text{M}_{\text{eff}}]))/(kk_0[\text{M}]),\end{aligned} \tag{3.12}$$

$\langle\tau_{\text{unbind}}\rangle$ is simply given by the inverse of the monomer unbinding rate, $\langle\tau_{\text{unbind}}\rangle = (R_{\text{unbind}})^{-1} = \frac{1}{k_0}e^{-\Delta G_{\text{gen}}-\Delta G_{\text{spec}}}$. To produce copies of length $\sim L_T/2$, it is necessary that a growing polymer can make at least $L_T/2$ displacement steps before it detaches at its leading edge, $\langle\tau_{\text{disp}}\rangle L_T/2 < \langle\tau_{\text{unbind}}\rangle$. Otherwise, the growing polymer will detach sooner than it grows, limiting its length. By rearranging this inequality, I arrive at

$$\Delta G_{\text{gen}} < -\Delta G_{\text{spec}} + \ln\frac{2k_0}{L_T\langle\tau_{\text{disp}}\rangle} \tag{3.13}$$

which can be implicitly solved for ΔG_{gen}. This constraint is represented by the cyan dashed line in Fig. 3.8. When this constraint isn't satisfied, newly-initiated polymers tend to detach before they extend. When the total copy-template affinity $\Delta G_{\text{spec}} + \Delta G_{\text{gen}}$ is strong, as at **II**, I observe a single, long polymer stuck to the template, and when the affinity is weak, at **0**, I observe empty templates. Polymer brush configurations **III**, **III'**, **IV** and **V**, which all produce relatively long distributions, all lie to the left of this constraint.

For a dense, ordered brush conformation as in **III** and **III'** to be stable, $\sim L_T/2$ sequential extension and displacement steps must occur on a timescale shorter than that at which any of the L_T polymers fall off, enabling the system to "heal" the brush after detachment events. Therefore, a dense, brush-like conformation will be reached when $\langle\tau_{\text{disp}}\rangle L_T/2 < \langle\tau_{\text{unbind}}\rangle/L_T$. Rearranging, I obtain

$$\Delta G_{\text{gen}} < -\Delta G_{\text{spec}} + \ln\frac{2k_0}{L_T^2\langle\tau_{\text{disp}}\rangle}, \tag{3.14}$$

represented by the red dashed line in Fig. 3.8. In a maximally-displaced, brush-like conformation, all polymers are only attached by their leading edge. Hence all polymers are equally likely to detach at any moment and therefore the product length distribution is also uniform, with a mean length very close to $\langle L \rangle = L_T/2$.

When ΔG_{gen} lies between the constraints Eqs. 3.13 and 3.14, the generic bond is strong enough to enable long polymers to form but insufficient to reach a maximally-displaced, brush-like conformation. If the specific bond strength is strong enough that lagging tails tend to bind back to the template, and are only removed via displacement, conformations similar to **IV** are seen. Here, gaps appear between the leading edges of polymers, with weakly-bound tails occupying the intervening sites. Some shorter polymers occupy large swathes of the template, while some longer polymers have been pushed to the end of the template where they can advance no further. This constitutes an inversion of the cooperative effect; longer polymers which bunch at the end of the template tend to have fewer template connections than shorter polymers, which skews the product length distribution slightly toward long polymers giving $\langle L \rangle > L_T/2$.

This effect is not observed when the specific bond is weak enough that polymer tails can spontaneously detach from the template without the need for displacement, as seen at **V** where $\Delta G_{\text{spec}} = 8 > \ln[\text{M}_{\text{eff}}]$. Here gaps appear quicker than they can be filled by extension of polymers already on the template, and monomers that bind within these gaps can sometimes form dimers. The tails of the dimers may spontaneously detach from the template, preventing the incorporation of the dimer into the polymer behind, and initiating a new copy from the centre of the template. As a consequence, the ordered polymer brush conformation of **III** or **III'** is disrupted by shorter polymers that have incorrectly initiated in the middle of the template, as seen in **V**. Consequently, the system produces polymers with a mean length $\langle L \rangle < L_T/2$.

I have not observed any parameter values in which the template sites are typically unoccupied but in which long copies are frequently produced. To get long polymers, monomer binding in isolation must be stable ($R_{\text{bind}}/R_{\text{unbind}} > 1$) since cooperative binding is disrupted. Therefore the only way that the template sites can typically be available is if both (a) lagging tails tend to unbind, and (b) if stepping forward is much faster than the rate at which monomers fill in behind. But a monomer filling in a gap happens at least as quickly as a polymer can step forwards, by definition, and much faster if the polymer runs into a traffic jam of other polymers. So empty sites tend to be a rare occurrence unless binding to the template is just pathologically unstable.

I have found that applying the two constraints of Eqs. 3.11 and 3.13 is generally sufficient to identify the region in which long copies are produced (Fig. 3.8). I note that in the upper left triangle of Fig. 3.8, where $G_{BB} \gg G_{gen}$, the system failed to produce 1000 polymers during the allotted run-time and the data were excluded. When $G_{BB} \gg G_{gen}$, the depolymerisation step in the model becomes very fast causing repeated polymerisation and depolymerisation events. In this region, the embedded process of polymer growth becomes very long and

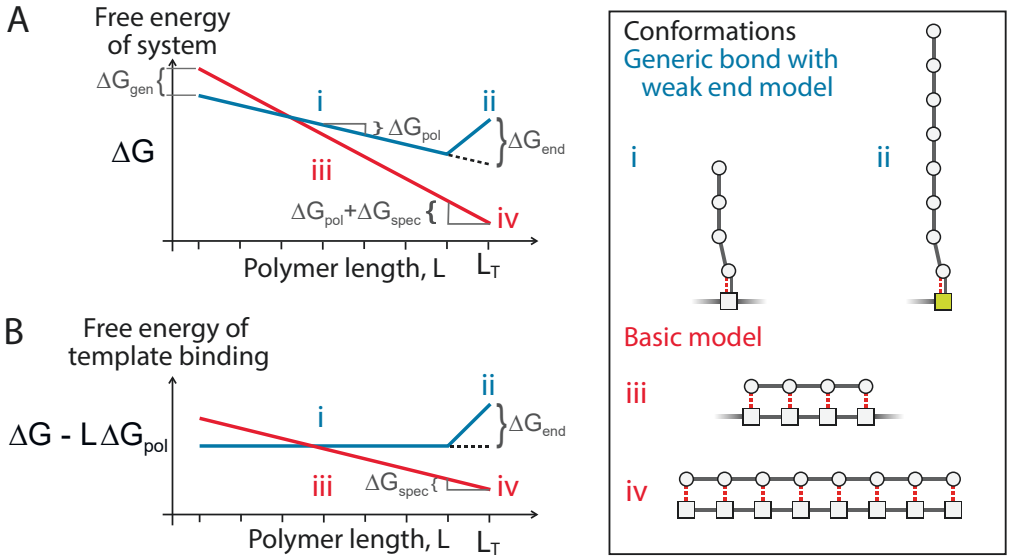


Figure 3.10: A. The free energy of a polymer, ΔG, decreases with length, L when $\Delta G_{\text{pol}} < 0$ in all models. **B.** The component of the free energy due to the copy-template bonds is given by $\Delta G - L\Delta G_{\text{pol}}$. The copy-template free energy decreases with length in the basic model (red line and conformations iii and iv) which suppresses the release of long polymers. In the maximally displaced conformations (i and ii) generated under the generic bond model (blue line), all polymers are only connected to the template at their leading edge, and hence the free energy landscape is flat. Weakening the final site on the template by an amount ΔG_{end} destabilises polymers with length $L = L_T$ (ii), leading to their selective release. Figure reproduced with permission from Ref. [1] under a CC BY license.

requires ever longer run times to simulate. In the under-sampled regions, I have no evidence to suggest that the polymer length distributions produced by the system deviate from the simple arguments presented above.

In this section, I have demonstrated that long polymers can be produced if the free energy released during polymerisation is channelled into the destabilisation of the interactions between polymer tails and the template. However, to achieve true copying, the polymers must have the same length as the template. In the following section, I demonstrate that this can be achieved by weakening the copy-template interaction strength at end of the template.

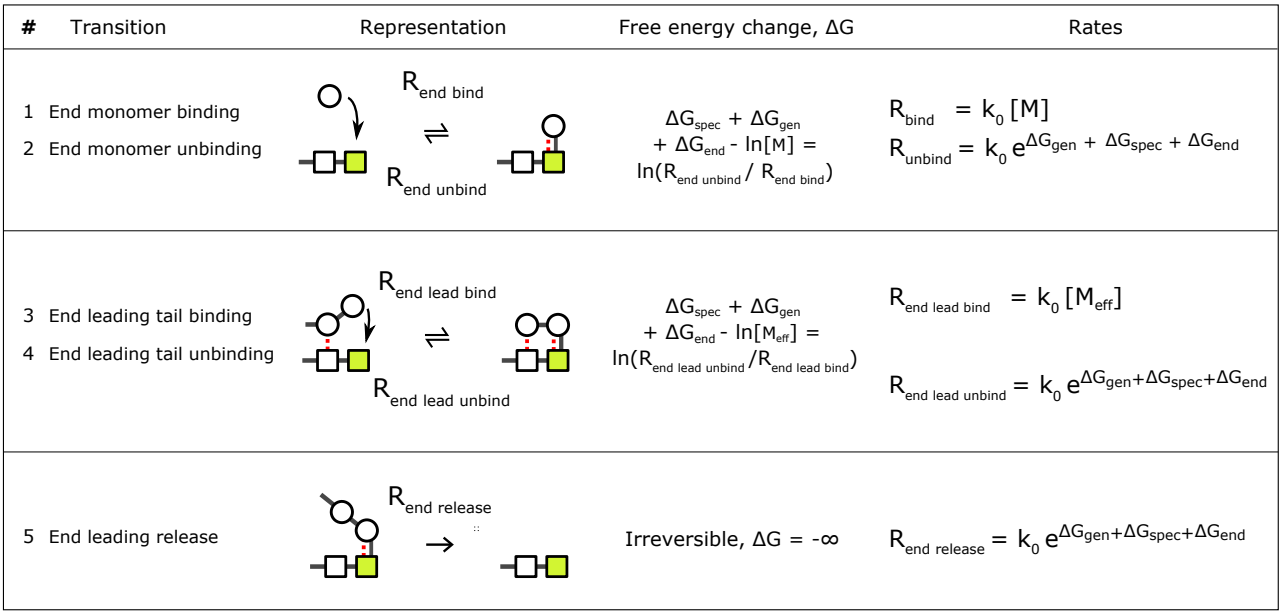


Figure 3.11: The dynamics of the system at the weakened end (green square) of the template. I extend the generic bond model presented in Fig. 3.7 and modify the copy-template interaction strength at the end of the template with a free energetic term ΔG_{end}. When $\Delta G_{\text{end}} > 0$, the bond made between copy units and the final template site is weakened. As before, the right hand edge of the copy polymer is the leading edge. Figure reproduced with permission from Ref. [1] under a CC BY license.

3.3.3 Weakening the final site on the template biases the production of complete polymers

In this section, I propose a simple adjustment to the model presented in Section 3.3.2 that allows for the systematic release of complete polymers from the template when the system is in a brush-like conformation. Here, the copy-template bond at the final template site is weakened with a free-energetic factor $\Delta G_{\text{end}} > 0$. This change is somewhat analogous to a stop codon in translation [14]. Crucially, since the end of the template is a unique site, this change can be made in a way that doesn't interfere with the copy-template interactions at other sites.

The effect of this adjustment is illustrated in the free energy profiles shown in Fig. 3.10. In the previous section, I demonstrated that disrupting the copy-template bond upon polymerisation could allow the lagging tail of polymers to detach from the template, meaning that, although polymerisation was thermodynamically "down-hill" (blue line in Fig. 3.10 A), the binding strength of the copy to the template did not increase as the copy extends (blue line in Fig. 3.10 B); the copy-template binding free energy landscape was flat and the length of a polymer was decoupled from the number of connections it maintained with the template. Here, I hypothesise that weakening the copy-template interaction strength at the end of the template, ΔG_{end}, will destabilise the binding of completed polymers and promote their release over incomplete polymers, leading to tight control over the product length distribution.

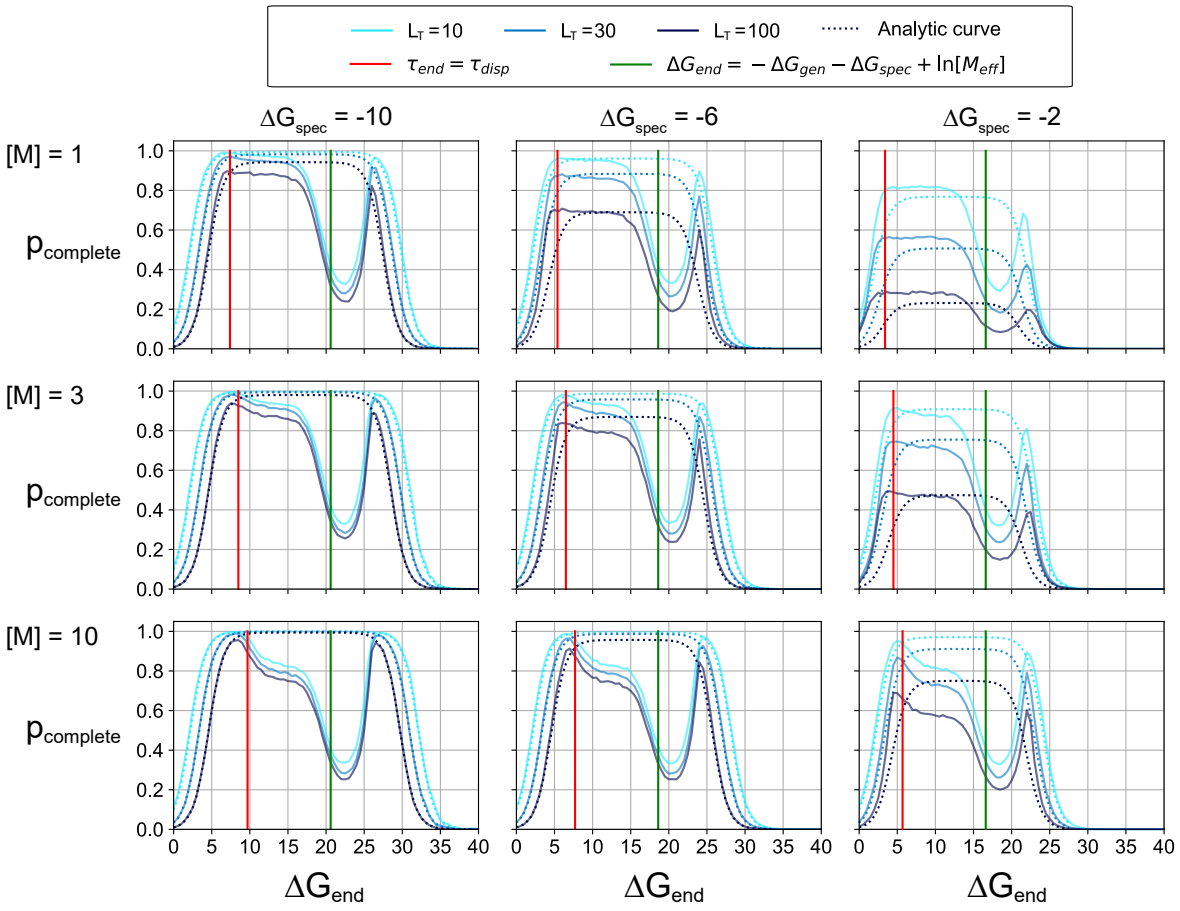


Figure 3.12: Destabilising the last site on the template with ΔG_{end} increases the probability of producing a complete polymer p_{complete} when the configuration of copies forms a dense brush as seen in conformation **III** in Fig. 3.8. I plot the averaged probability of producing a complete polymer, p_{complete}, against the destabilising energy penalty, ΔG_{end}, for a range of system parameters. I compare the data to a simple analytical model that accounts for the flux of complete polymers and uniformly distributed polymers from the template, and also indicate the points at which $\langle \tau_{\text{disp}} \rangle = \langle \tau_{\text{end}} \rangle$ (red line) and $\Delta G_{\text{end}} = -\Delta G_{\text{gen}} - \Delta G_{\text{spec}} + \ln[\text{M}_{\text{eff}}]$ (green line) to guide interpretation, as explained in the text. Figure reproduced with permission from Ref. [1] under a CC BY license.

The thermodynamics and kinetics of the modified system are identical to the model presented in Section 3.3.2 and Fig. 3.7, apart from at the very end of the template. In Fig. 3.11, I outline the thermodynamic and kinetic changes to the model. The free-energy change of binding at the end of the template is now given by $\Delta G = \Delta G_{\text{spec}} + \Delta G_{\text{gen}} + \Delta G_{\text{end}} - \ln[\text{M}]$, and the free-energy change of polymer tail binding at the end of the template is now given by $\Delta G = \Delta G_{\text{spec}} + \Delta G_{\text{gen}} + \Delta G_{\text{end}} - \ln[\text{M}_{\text{eff}}]$. The rates of monomer unbinding, leading-edge polymer unbinding and polymer termination at the end of the template all take the value $R = k_0 e^{\Delta G_{\text{spec}} + \Delta G_{\text{gen}} + \Delta G_{\text{end}}}$; binding rates are unchanged relative to Section 3.3.2.

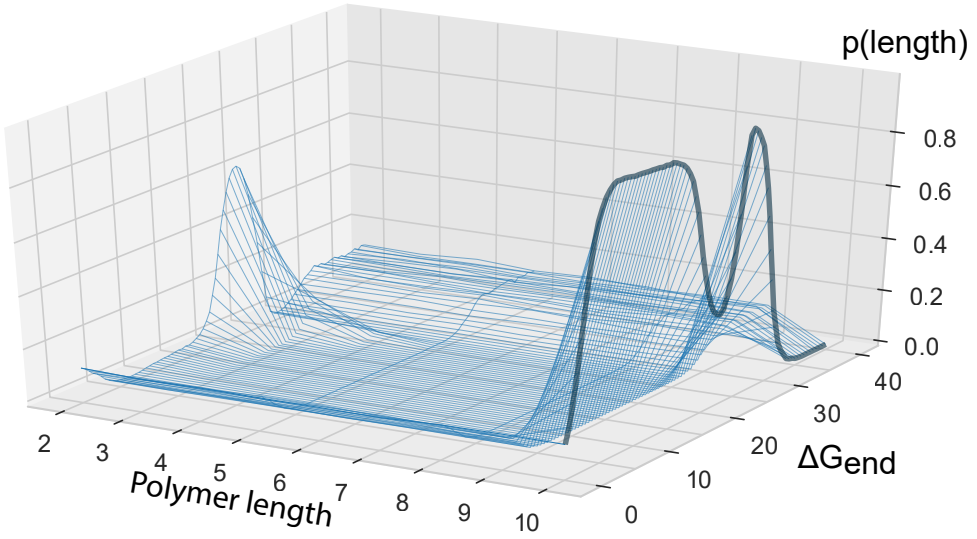


Figure 3.13: The full length distributions of the data shown in Fig. 3.12 for $[\mathrm{M}] = 1$ and $\Delta G_{\mathrm{spec}} = -4$ and a template of length $L_T = 10$, averaged over 5 independent simulations that generated up to 2000 polymers each at each value of ΔG_{end}. The line corresponding to the variation of p_{complete} with ΔG_{end} is accentuated in black. At $\Delta G_{\mathrm{end}} = 0$, the product length distribution is approximately uniform. At low ΔG_{end} there is a sharp increase in the probability of producing complete polymers with increasing ΔG_{end} before a plateau is reached. At higher ΔG_{end}, a peak in the production of dimers is observed. As ΔG_{end} is increased further, there is a final peak in the production of full-length products, after which the system produces a uniform distribution of products excluding complete polymers. Figure reproduced with permission from Ref. [1] under a CC BY license.

Results

In Fig. 3.12, the probability of producing a full-length copy polymer, p_{complete} is plotted against the destabilising free-energy penalty of the last site, ΔG_{end}, at a few values of the generic bond strength ΔG_{gen}, the monomer concentration $[\mathrm{M}]$ and the template length L_T, as a function of ΔG_{end}. I set $k = 1$, $\Delta G_{\mathrm{spec}} = -4$, and $\Delta G_{\mathrm{BB}} = -20$, which puts the system into a dense brush steady state conformation when $\Delta G_{\mathrm{end}} = 0$, as seen in Section 3.3.2. I conducted 5 independent repeated simulations for each sampled parameter point, with each simulation capable of producing up to 2000 polymers or being truncated after 4 hours of run-time. In Fig. 3.13, I show the full product length distribution for $[\mathrm{M}] = 1$, $\Delta G_{\mathrm{spec}} = -4$, $L_T = 10$.

Fig. 3.12 shows that weakening the end of the template can increase the proportion of complete products, p_{complete}, and that $p_{\mathrm{complete}} \approx 1$ is possible under certain conditions. When $\Delta G_{\mathrm{end}} = 0$, the polymer brushes encountered in Section 3.3.2 tend towards a jammed state, as depicted in Fig. 3.14 **A**, with detachment events occurring at any point on the template. Weakening the final site on the template creates a release point that allows this polymer 'traffic' to flow through, with unbinding predominantly occurring at the final site where complete polymers are attached by a weakened final copy-template bond, as seen in the binding free energy landscape in Fig. 3.10.

A simple model can explain the proportion of complete polymers in the product pool, p_{complete}, produced by the system when a dense conformation has been reached. This model considers the balance of the rate of formation and release of complete polymers and the rate of release of polymers from the rest of the template. In the densely-packed regime, as depicted in Fig. 3.14 **A**, all polymers have one bond with the template. The timescale on which one of $L_T - 1$ incomplete polymers are released from a template of length L_T is approximately,

$$\langle \tau_{\text{uniform}} \rangle = (L_T - 1)^{-1} R^{-1}_{\text{lead release}} \approx (L_T - 1)^{-1} k_0 \, e^{-\Delta G_{\text{gen}} - \Delta G_{\text{spec}}}. \tag{3.15}$$

Next, I estimate the time it takes the system to complete a growing polymer and release it from the template. To extend the polymer attached to the penultimate site, a free monomer must occupy the final site. The fraction of time that this final site is occupied (as opposed to empty), g, can be approximated by the binding equilibrium of the monomer

$$g \approx \frac{k_0 \, [\text{M}]}{k_0 \, [\text{M}] + k_0 \, e^{\Delta G_{\text{gen}} + \Delta G_{\text{spec}} + \Delta G_{\text{end}}}} \tag{3.16}$$

Given a polymerisation rate k, the average timescale on which a polymer with length $L_T - 1$ is completed by polymerising with a monomer at the end site on the template is

$$\langle \tau_{\text{form}} \rangle = 1/gk \approx \frac{[\text{M}] + e^{\Delta G_{\text{gen}} + \Delta G_{\text{spec}} + \Delta G_{\text{end}}}}{[\text{M}] \, k} \tag{3.17}$$

In this simple model, I assume that the time it takes this newly completed polymer to fall off the template is rate limited by either the timescale on which the copy-template bond at the end of the template is broken,

$$\langle \tau_{\text{end}} \rangle = 1/k_0 \, e^{\Delta G_{\text{gen}} + \Delta G_{\text{spec}} + \Delta G_{\text{end}}}, \tag{3.18}$$

or the time it takes for the polymer unit at the penultimate template site to be displaced by a polymer behind, which is given by $\langle \tau_{\text{disp}} \rangle$ as calculated in Eq. 3.12 in Section 3.3.2.

Taken together, the timescale on which complete polymers fall off the template is approximately

$$\langle \tau_{\text{fall}} \rangle \approx \langle \tau_{\text{end}} \rangle + \langle \tau_{\text{disp}} \rangle, \tag{3.19}$$

and the total timescale of producing and releasing a complete polymer is

$$\langle \tau_{\text{complete}} \rangle \approx \langle \tau_{\text{form}} \rangle + \langle \tau_{\text{fall}} \rangle. \tag{3.20}$$

Treating the inverse of each of these timescales as an effective rate, I approximate the steady state value of p_{complete} as

$$p_{\text{complete}} \approx \frac{1/\langle\tau_{\text{complete}}\rangle}{1/\langle\tau_{\text{complete}}\rangle + 1/\langle\tau_{\text{uniform}}\rangle}\,. \tag{3.21}$$

With no free fitting parameters, this simple model captures remarkably well the range of ΔG_{end} over which p_{complete} is high and the maximum value of p_{complete} over the whole range of ΔG_{end}, as seen in Fig. 3.12 and A.5. This fact suggests that the physics incorporated into this simple model explains the majority of the behaviour of the system in this regime.

Specifically, in Fig. 3.12, at $\Delta G_{\text{end}} = 0$, there is no bias for completed polymers to fall off the template and an approximately uniform product length distribution is observed, as encountered in Section 3.3.2, resulting in $p_{\text{complete}} \approx 1/L_T$. At small values of ΔG_{end}, breaking the copy-template bond at the end of the template remains the rate-limiting step in producing a full-length copy and $\langle\tau_{\text{complete}}\rangle \approx \langle\tau_{\text{end}}\rangle$. As ΔG_{end} increases, the time to break the end copy-template bond, $\langle\tau_{\text{end}}\rangle$, decreases exponentially and therefore p_{complete} increases exponentially. Eventually, however, p_{complete} saturates when $\langle\tau_{\text{end}}\rangle < \langle\tau_{\text{disp}}\rangle$; at this point, $\langle\tau_{\text{complete}}\rangle \approx \langle\tau_{\text{disp}}\rangle$ and breaking the copy-template bond at the end is no longer rate-limiting. Here the model suggests that no further increase in p_{complete} can be achieved by increasing ΔG_{end}; the value of the saturating level is determined by the relative timescales of displacement $\langle\tau_{\text{disp}}\rangle$ and spontaneous detachment from the body of the template $\langle\tau_{\text{uniform}}\rangle$, neither of which are effected by the properties of the end of the template.

At very high values of ΔG_{end}, the simple model accurately predicts the drop in p_{complete} to zero shown in Fig. 3.12. Here, the final site on the template is so unstable that the monomer occupancy, g, tends to zero. Consequently, $\langle\tau_{\text{form}}\rangle$ diverges and no full-length copies form. Instead, the system effectively operates as a uniform template of length $L_T - 1$, and generates a uniform distribution of polymers with lengths $L \leq L_T - 1$, but no polymers with length $L = L_T$. This behaviour is particularly clear in the whole length distribution shown in Fig. 3.13.

The simple model presented here assumes that the system operates in the dense, ordered brush regime. However, in Fig. 3.12, where $[\text{M}] = 1$, $\Delta G_{\text{gen}} = -8$ and $\Delta G_{\text{spec}} = -2$, the constraint given in Eq. 3.14 that predicts whether the dense conformation is reached, $\langle\tau_{\text{disp}}\rangle L_T/2 < \langle\tau_{\text{unbind}}\rangle/L_T$, is not satisfied. Here the model's estimates of the level of the plateau in p_{complete} are less accurate. The biggest failure of the simple model, however, is not predicting the dips in p_{complete} that occur at intermediate values of ΔG_{end}. In Fig. 3.12, a single dip in p_{complete} is observed for $[\text{M}] = 1$, and a dip and an intermediate plateau are observed for $[\text{M}] = 10$. Looking at the full polymer length distributions shown in Fig. 3.13 and Figs. A.2, A.3 and A.4, these dips correspond to increased dimer production. Moreover, from snapshots of simulations shown in Fig. 3.14 **E**, the

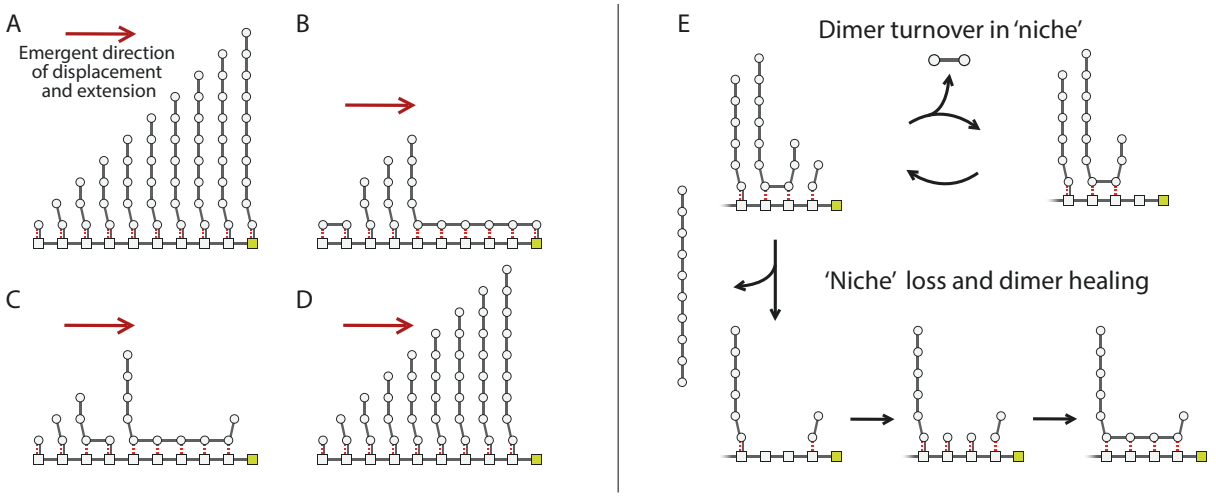


Figure 3.14: Weakening the end of the template with ΔG_{end} affects the typical conformations of the system on the templates. I illustrate typical configurations from the simulations reported in Fig. 3.13. **A** At $\Delta G_{\text{end}} = 0$, the system reaches a dense brush. **B** At $\Delta G_{\text{end}} = 10$, completed polymers are released sooner than displacement can compensate, and so the brush of copies becomes less dense. **C** At $\Delta G_{\text{end}} = 22$, the final template site has low occupancy. Here dimers form under the fraying ends of completed polymers. **D** At $\Delta G_{\text{end}} = 35$, the occupancy of the final site is so low that incomplete polymers detach sooner than they can complete by polymerising at the end of the template. **E** Niche formation. The leading edge of completed polymers may fray, creating a protective niche that promotes the cyclical turnover of dimers. If the completed polymer falls off faster than a dimer can form, then the dimer can be incorporated into the preceding polymer. Figure reproduced with permission from Ref. [1] under a CC BY license.

dimer formation occurs at the final two sites on the template underneath the fraying end of completed- but not yet detached- polymers.

Although a quantitative model that predicts the occurrence of 'dimer dips' could not be established despite significant effort, a qualitative understanding of the factors that cause dimer production has been achieved. The production of dimers occurs at fairly high values of ΔG_{end}, which are certainly not necessary for precise length control (it would be sufficient to just weaken the connection by a few units of thermal energy, as is clear in Fig. 3.12), and may be experimentally inaccessible. Nevertheless, I can offer an explanation as to how this unexpected phenomenon arises in the model.

When ΔG_{end} is weakened sufficiently that $\langle \tau_{\text{end}} \rangle < \langle \tau_{\text{disp}} \rangle$ (shown as a red line in Fig. 3.12), the leading edge of completed polymers are released faster than the tails are fully displaced, leading to less dense conformations such as Fig. 3.14 **B** and **C**. Completed polymers tend to be displaced from the template from their lagging edge, and can spontaneously fray from their leading edge at the end of the template. When $\Delta G_{\text{end}} > -\Delta G_{\text{gen}} - \Delta G_{\text{spec}} + \ln[\text{M}_{\text{eff}}]$ (equality at green line in Fig. 3.12), frayed configurations at the leading edge, such as Fig. 3.14 **C**, are more likely than un-frayed alternatives such as Fig. 3.14 **B**. This constraint on ΔG_{end} locates the "dimer

dip". If a completed polymer frays far enough to expose the penultimate site on the template, a monomer may occasionally bind there with a strong combined generic and specific bond, as shown in the upper half of Fig. 3.14 **E**. This penultimate monomer is unable to polymerise into the completed polymer behind it and instead dimerises with monomers that occasionally bind to the final template site. Completed polymers thus create a 'protective niche' for the formation of dimers. Upon dimerising, the dimer is rapidly released from the template. This cycle is shown in Fig. 3.14 **E**.

If the timescale of formation and release of dimers in the protective niche is shorter than the average timescale on which the completed polymers are released, then multiple dimers may be released for each completed polymer. When ΔG_{end} is increased further, the monomer occupancy of the final site reduces and the time to form a dimer increases. If the timescale of formation and release of dimers in the protective niche is greater than the timescale on which the completed polymers are released, then the completed polymers fall off, removing the protective niche. As depicted in Fig. 3.14 **E**, without the protective niche, then the dimer is able to polymerise into the preceding polymer, which dramatically reduces the number of dimers that are released into the product pool. Hence, after the dimer dip, a sharp increase in p_{complete} is observed again, close to the saturating value that the simple model predicts, before the eventual drop off to $p_{\text{complete}} \approx 0$ at the highest values of ΔG_{end}. At high monomer concentrations, such as $[\text{M}] = 10$ in Fig. 3.12, there is an intermediate plateau in p_{complete} as well as the dimer dip, which occurs when $\langle \tau_{\text{end}} \rangle < \langle \tau_{\text{disp}} \rangle$ shortly after the model predicts saturation. Due to the increased monomer concentration, fraying events at the end of the template result in more monomer invasions at the penultimate site causing the more frequent formation of dimers.

3.3.4 On-rate discrimination enables the reliable production of copies with the correct length and accurate sequences

In Section 3.3.3, I identified a mechanism for producing polymers of a single monomer type with a specific template-determined length in a reliable fashion. In this section, I consider whether the mechanisms that grant control over the length distribution are compatible with mechanisms control the accuracy of copies that are generated from a pool of two types of monomer. I extend the models presented in previous sections and consider the growth of binary copies. The parameterisation is given in Fig. 3.15. It is identical to that in Fig. 3.11, except that: (i) ΔG_{spec} now depends on the match between template and copy monomer; (ii) I allow the specific binding free energy to appear either in the binding or the unbinding rates.

Reparametersation

In general, the model contains four specific bond strengths, ΔG_{00}, ΔG_{01}, ΔG_{10}, and ΔG_{11}, where the first and second indices give the copy unit and template unit respectively. I make the simplifying assumption that all copy monomer concentrations are equal $[\mathrm{M}_0] = [\mathrm{M}_1] = [\mathrm{M}]$ and that all interactions are symmetric (all mismatches are equal, as are all matches) following [23]. If the correct pairings, 00 and 11, have the same free energy $\Delta G_{00} = \Delta G_{11} = \Delta G_r$, and if the incorrect pairings are also equal, $\Delta G_{01} = \Delta G_{10} = \Delta G_w$, the specific bond strengths can be described with two energy parameters, $\Delta G_{\text{spec,r}} = \Delta G_r$ or $\Delta G_{\text{spec,w}} = \Delta G_w$, the standard free energy of forming a right match or wrong match copy template bond. By making these simplifying assumptions, the copy sequence can be described based solely on whether the monomers are matched (right, r) or unmatched (wrong, w) with respect to the template, and the model becomes independent of the specific template sequence under consideration.

Again, the principle of local detailed balance is used to constrain the forward and backward transition rates between any pair of states. The ratio of the monomer binding rates $R_{\text{bind,r/w}}$ to the monomer unbinding rates $R_{\text{unbind,r/w}}$ for right or wrong template matches is proportional to the exponential of the specific bond strength (amongst other factors in Fig. 3.11):

$$\frac{R_{\text{bind,r/w}}}{R_{\text{unbind,r/w}}} \propto e^{-\Delta G_{\text{r/w}}}. \tag{3.22}$$

I compare two contrasting parameterizations of the model, as detailed in Fig.3.15. If the unbinding rates depend on the match between the copy and the template, $R_{\text{unbind,r/w}} \propto e^{\Delta G_{\text{r/w}}}$, as in the previously presented models, then the system demonstrates "off-rate" discrimination, and the monomer binding rates remain constant. Alternatively, if the monomer binding rates depend on the match between the copy and the template, $R_{\text{bind,r/w}} \propto e^{-\Delta G_{\text{r/w}}}$, then the system demonstrates "on-rate" discrimination, and the unbinding rates remain constant. The

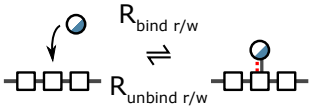
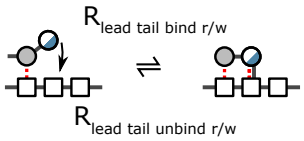
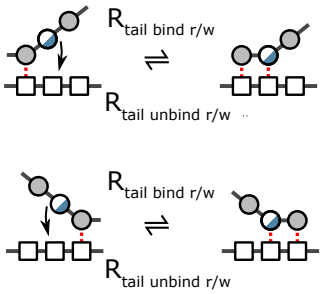
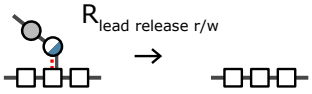
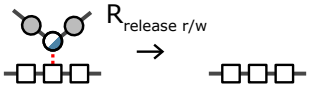
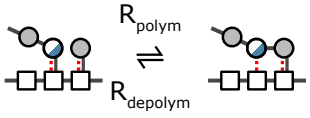
# Transition	Representation	Free energy change, ΔG	Rates **'Off-rate'** discrimination	Rates **'On-rate'** discrimination
1 Monomer binding 2 Monomer unbinding		$\Delta G_{spec,r/w} + \Delta G_{gen} - \ln[M] = \ln(R_{unbind\ r/w} / R_{bind\ r/w})$	$R_{bind\ r/w} = k_0 [M]$ $R_{unbind\ r/w} = k_0\, e^{\Delta G_{gen} + \Delta G_{spec,r/w}}$	$R_{bind\ r/w} = k_0 [M]\, e^{-\Delta G_{spec,r/w}}$ $R_{unbind\ r/w} = k_0\, e^{\Delta G_{gen}}$
3 Leading binding 4 Leading unbinding		$\Delta G_{spec,r/w} + \Delta G_{gen} - \ln[M_{eff}] = \ln(R_{lead\ tail\ unbind\ r/w}/R_{lead\ tail\ bind\ r/w})$	$R_{lead\ tail\ bind\ r/w} = k_0 [M_{eff}]$ $R_{lead\ tail\ unbind\ r/w} = k_0\, e^{\Delta G_{gen} + \Delta G_{spec,r/w}}$	$R_{lead\ tail\ bind\ r/w} = k_0 [M_{eff}]\, e^{-\Delta G_{spec,r/w}}$ $R_{lead\ tail\ unbind\ r/w} = k_0\, e^{\Delta G_{gen}}$
3′ Tail binding 4′ Tail unbinding		$\Delta G_{spec\ r/w} - \ln[M_{eff}] = \ln(R_{tail\ unbind\ r/w}/R_{tail\ bind\ r/w})$	$R_{tail\ bind\ r/w} = k_0 [M_{eff}]$ $R_{tail\ unbind\ r/w} = k_0\, e^{\Delta G_{spec,r/w}}$	$R_{tail\ bind\ r/w} = k_0 [M_{eff}]\, e^{-\Delta G_{spec,r/w}}$ $R_{tail\ unbind\ r/w} = k_0$
5 Leading release		Irreversible, ΔG = -∞	$R_{lead\ release\ r/w} = k_0\, e^{\Delta G_{gen} + \Delta G_{spec,r/w}}$	$R_{lead\ release\ r/w} = k_0\, e^{\Delta G_{gen}}$
5′ Release		Irreversible, ΔG = -∞	$R_{release\ r/w} = k_0\, e^{\Delta G_{spec,r/w}}$	$R_{release\ r/w} = k_0$
6 Polymerisation 7 Depolymerisation		$\Delta G_{BB} - \Delta G_{gen} = \ln(R_{depolym} / R_{polym})$	$R_{polym} = k$ $R_{depolym} = k\, e^{\Delta G_{BB} - \Delta G_{gen}}$	$R_{polym} = k$ $R_{depolym} = k\, e^{\Delta G_{BB} - \Delta G_{gen}}$

Figure 3.15: In a system with two competing monomer types, "right" (white) specfic copy-template bonds have a strength $\Delta G_{\mathrm{spec,r}}$ and "wrong" (blue) specific copy-template bonds have a strength $\Delta G_{\mathrm{spec,w}}$. I have half-shaded the monomers that affect the transition rates. The identity of monomers shaded gray do not affect the transition rates. I consider parameterisations in which discrimination occurs on the binding step ("on-rate" discrimination) or on the unbinding step ("off-rate" discrimination) as described in Section 3.3.4. Unbinding reactions at the end of the template are sped up by a factor $\exp(\Delta G_{\mathrm{end}})$ for both monomer types as depicted in Fig. 3.11 but are not shown here for brevity.

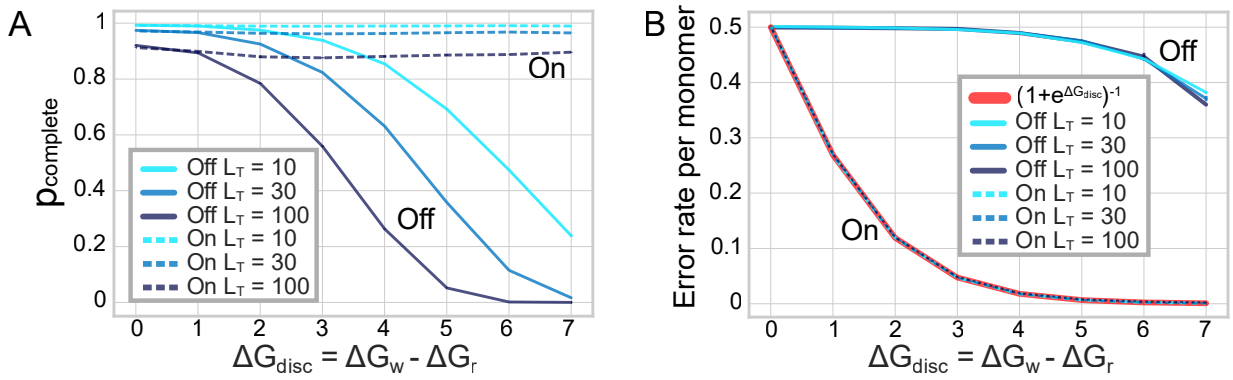


Figure 3.16: Models with on-rate discrimination can yield complete polymers with high probability, $p_{\text{complete}} \approx 1$,and low error rates per monomer, whereas models with off-rate discrimination cannot reliably generate complete polymers with low error rates per monomer. I present **A** the mean of p_{complete} and **B** the mean error rate per monomer over 5 repeats at each value of the discrimination free energy, $\Delta G_{\text{disc}} = \Delta G_{\text{w}} - \Delta G_{\text{r}}$, for simulations using the model in Fig. 3.15 with an alphabet size of two. Simulations are performed for $L_T = 10$, 30 and 100 using $k = 1$, $\Delta G_{\text{BB}} = -20$, $\Delta G_{\text{gen}} = -12$, $\Delta G_{\text{end}} = 7$, and $[M_0] = [M_1] = 1$ for models with on- and off-rate discrimination, shown with dotted and solid lines, respectively. The Standard Error of the Mean (SEM) is shown as error bars, which are generally smaller than the line width. In **B** the error rates per monomer for different template lengths overlap one another and lie very close to the limiting line $(1+\exp(\Delta G_{\text{disc}}))^{-1}$. Figure adapted reproduced with permission from Ref. [1] under a CC BY license.

ΔG_{spec} sensitivity is applied consistently across all binding and unbinding transitions for on and off-rate discrimination, as shown in Fig.3.15. Similar to Section 3.3.3, all unbinding rates for units located at the end of the template are increased by a factor of $e^{\Delta G_{\text{end}}}$.

Results

The parameter values that produced complete polymers in Section 3.3.3 were used to generate the data shown in Fig. 3.16. Specifically, $k = 1$, $\Delta G_{\text{BB}} = -20$, $\Delta G_{\text{gen}} = -12$, $\Delta G_{\text{end}} = 7$, and $[M_0] = [M_1] = 1$ were set. The value of $\Delta G_{\text{spec,w}}$ was set to $\Delta G_w = -1$ for wrong-match monomers, and $\Delta G_{\text{spec,r}}$ was set to $\Delta G_r = -1, -2, \ldots, -8$ for right-match monomers. To generate each data point in Fig. 3.16, 5 independent simulations were run, with each simulation capable of producing up to 5000 copolymers. In addition to calculating the average of p_{complete}, the average error rate per monomer (the number of wrong units in the copolymer divided by the polymer length) for complete polymers was also calculated.

Fig. 3.16 **A** clearly shows that on-rate discrimination is effective at maintaining a high proportion of complete polymers even at long lengths and high values of $\Delta G_{\text{disc}} = \Delta G_w - \Delta G_r$. In contrast, off-rate discrimination fails to produce long polymers as the discrimination free-energy difference increases. The generic bond and weak end mechanisms, discussed in previous sections, require the leading edge of polymers to remain bound to the template until the polymer has grown to reach the end of the template in order to generate and selectively release

long polymers. However, if the system discriminates on the off-rates, incorporation of incorrect monomers into a polymer can lead to its premature release, thereby reducing p_{complete}.

In Fig. 3.16 **B**, on-rate discrimination delivers low error rates per monomer, close to the ratio of binding rates implied by the discrimination free energy $1/(1+\exp(\Delta G_{\text{disc}}))$. Off-rate discrimination is inaccurate for moderate discrimination free energies as a strong ΔG_{gen} allows incorrect monomers to be incorporated into copolymers. For off-rate discrimination, there is a trade-off between accuracy and length control, p_{complete}, for the bond strengths I have considered here.

If the rate at which wrong monomers detach were higher than the rate of polymerisation, $k < k_0 \exp(\Delta G_{\text{gen}} + \Delta G_{\text{w}})$ (which implies $\Delta G_{\text{w}} > -\Delta G_{\text{gen}} = 12$), then incorrect monomers that bind to the template would be more likely to detach than incorporate into a growing chain. Hence the overall accuracy of the polymers would increase, and, as incorrect monomers are unlikely to be incorporated into growing chains, it should also prevent the early detachment of growing copies with accurate sequences, promoting length-control. Therefore, off-rate discrimination should provide length and sequence control only when the free energetic penalty of forming an incorrect monomer-template bond is very high (assuming the polymerisation rate is fixed). Hence, in most cases, on-rate discrimination is required to gain control of the length and accuracy in these minimal models of autonomous templated copying.

3.4 Conclusion

Unlike the template-copying machinery found in biological systems that can cyclically copy templates under constant environmental conditions [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19], state-of-the-art synthetic molecular copying systems require external intervention or non-chemical driving [79, 80, 81, 87, 78, 88, 89, 90] to operate cyclically. In this chapter, I described coarse-grained models of isothermal, enzyme-free, templated copolymerisation processes in which the copy copolymers can spontaneously detach from the template in a constant environment, driven by chemical free energy alone.

The central challenge of producing polymer copies is overcoming product inhibition, which is exacerbated by the natural cooperativity of the interactions between two polymers. In Sections 3.3.1 through 3.3.3, I sought to find a mechanism that would give control over the length of the copies, and then in Section 3.3.4 I identified a monomer discrimination mechanism that would additionally ensure copy accuracy. In Section 3.3.1, I demonstrated that the cooperativity of polymer template binding prevents the reliable production of full-length copies. In Section 3.3.2, I demonstrated that a surprisingly simple mechanism – using the free-energy of polymerisation to weaken the interaction between the monomers and the template behind the leading edge of the copy – could generate long polymers. In regions of the parameter space where this mechanism worked most effectively, the length of a polymer was decoupled from the number of connections that it shared with the template, and the overall product length distribution was uniform. Combining this mechanism with a weakened copy-template bond at the final site on the template, as explored in Section 3.3.3, was sufficient to bias the system toward producing only fully-grown copies, even when the template length was long ($L_T = 100$).

These proposed mechanisms require: 1) strong template-monomer bonds (so that the leading edge of growing polymers remain bound to the template for long enough to extend along its full length); 2) strong backbone bonds between monomeric units within polymers (so that formed polymers don't spontaneously fragment) that, upon formation 3) cause disruption to the copy-template bond of the 'back' monomer (so that the long cooperatively binding tails of polymers can be displaced from template by other polymers); and 4) a moderate destabilisation of the copy-template bond at the end of the template (to enable the selective release of fully-formed polymers). As I demonstrated in Section 3.3.4, to copy specific sequences accurately in addition to producing polymers of a fixed length, it is necessary to also have "on-rate" discrimination, in which matching sequences bind to the template faster.

The conditions stated above can be met by DNA-based chemical reaction systems that exploit handhold-mediated DNA strand displacement (HMSD) [22, 21]. In HMSD, polymerisation between monomers on the template results in the weakening of the copy-template connection in the 'lagging' monomer through the transfer

of bonds via strand displacement. Moreover, templating via HMSD has the potential to achieve effective on-rate discrimination during monomer binding, using a mechanism in which a transient match-specific toehold forms first, with the strong generic bond only forming after an additional (toehold-mediated) strand displacement reaction [149, 150].

The models I created generate a surprisingly rich range of behaviours and their dynamics did not match my initial expectations. I expected that optimal copying would involve the growth of isolated copies on a template, as is frequently observed in transcription [11, 12, 13] and translation [14, 15, 16, 17, 18, 19], and as reflected in previous models of templated copolymerisation [23]. In such a setting, it would be necessary to prohibit copies from starting at random locations on the template and thereby generating truncated products. However, the natural tendency of brush-like configurations observed here to absorb isolated monomers into growing copies at all points on the template solves that challenge, which otherwise might require complex activation and deactivation of template sites ahead of and behind the growing copy. Models of this type were explored initially but were not investigated further, once it was determined that length control could be achieved with the simple mechanisms described herein. Moreover, the observed displacement behaviour means that copying is possible even when the tails of polymers have a moderate affinity with the template; indeed, this regime was more effective in producing longer polymers than regimes in which tail affinity was so low that they would detach without displacement, as described in Section A.3 and Fig. A.5.

Extant biological template copying processes, such as transcription [11, 12, 13] and translation [14, 15, 16, 17, 18, 19], can operate in sparse regimes in which the copies do not fully occupy the template, although reasonably dense regimes are also observed [54, 151]. The sparse regime of operation is possible because the initiation of polymer growth only occurs at specific locations on the template, at ribosome binding sites in translation for instance [14], in tightly controlled ways that involve complex enzymatic machinery. Similarly, the enzymatic machinery of transcription and translation provides a larger, moving window in which copies are bound to their template [14, 15, 16, 17, 18, 19]. This factor enables off-rate discrimination between the incoming monomers without causing partially formed polymers to fall off prematurely. It is challenging to see how to engineer this in a simple setting without the interplay of enzymes that can define a scale for this template-attached region, but such a scheme may facilitate additional mechanisms such as kinetic proofreading to increase the accuracy above the limit set by the free energetic discrimination between monomer types.

Mechanisms that rely on the displacement of copies by other copies have been considered as solutions to product inhibition in an RNA world [124, 152]. In particular, Tupper and Higgs have argued that a rolling-circle copying of a circular template can alleviate product inhibition via strand displacement [124]. This rolling circle approach eliminates the tendency of short, newly-initialised copies to be out-competed by longer polymers

on a linear template, as observed in Section 3.3.1. The mechanism assumes directional polymerisation, but polymerisation is still potentially hampered by a displacement-based competition between polymer ends for binding to the template, with no intrinsic bias favouring the desired pre-polymerisation configuration.

The asymmetric destabilisation in my model induces a directionally-biased strand displacement between copy polymers, which can overcome product inhibition even on long templates. The mechanism may also work well in conjunction with the rolling circle mechanism of Tupper and Higgs, providing a directional bias to resolve the competition for the template. However, the chemistry and bonding stability of individual nucleotides as explored by Tupper and Higgs [124] – as opposed to larger oligonucleotide monomers as used in processes like HMSD [22] – may not be well-suited to the mechanisms investigated in this chapter. Outside of DNA-based reactions, Osuna Gálvez and Bode's reaction in Ref [77] – in which polymerisation is directly coupled to the disruption of product-template connections – shares the most resemblance to the mechanisms described by the model, though the templates cannot be reused. It is an open question whether small organic molecules with the appropriate chemistry can be identified and used to build synthetic copying systems with the properties stipulated in this chapter.

I have attempted to keep the models in this Chapter as simple and general as possible so that they can be used to guide the design of synthetic copying systems in a wide array of contexts. I have also attempted to make all chemical steps explicit, rather than invoking a chemical *deus ex machina* that resolves the central challenges of the copy processes, just as highly-evolved enzymatic machinery do *in vivo*. However, a key assumption of the model is that polymerisation and depolymerisation only occur while monomers are attached to the template - the template acts as a catalyst for bond formation and breakage. A passive template that simply brings reactants into close proximity could certainly allow polymerisation reactions that are prohibitively slow in solution to proceed on the template. However, such a mechanism would not accelerate depolymerisation as well, and so unless the template has some direct chemical coupling to the bond formation mechanism, as is typical in enzymes [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19], it is hard to see how to justify this catalytic assumption. Indeed, as demonstrated in Section 3.3.1, simple templates are not effective catalysts in the model – they accelerate the polymerisation reaction by forming a stable complex with the product, preventing cyclic copying. However, under the modified mechanism presented in Section 3.3.2, the competition between the polymerisation bond and the generic bond with the template provides a direct mechanism for catalysis, justifying the assumption that (de)polymerisation only occurs while monomers are attached to the template.

I have also assumed that polymers that detach from the template and are released into the bath do not rebind to the template. If the concentration of polymers in the bath grew to significant levels, that rebinding could occur. Indeed, the rebinding and subsequent elongation of products has been presented as a plausible pathway to the

generation of longer RNA oligomers, bridging a portion of the gap between short RNA building blocks and the long RNA molecules that make up ribozymes in a series of papers on templated ligation in an RNA world [105, 104, 106]. However, polymerisation is coupled to the disruption of the copy-template bonds in my mechanism, reducing the interaction strength that most units in the polymer chain could rebind by. Only the leading edge of a grown polymer could potentially stick tightly to the template, possibly causing unwanted over-extension of polymers if this reattachment happened at an earlier site on the template that happened to be complementary. If one were to design a copying machine based upon the biased displacement mechanism introduced here, a special monomeric unit could be reserved to bind effectively only to the ends of templates, and thus cap the end of completed polymers. This approach would prevent further unwanted extension and reduce the copy-template affinity for completed products, alleviating the aforementioned issues.

Another key assumption used in this chapter is that polymer tails do not interact with one another. Our models suggest a dense brush of polymers is beneficial for length control. However, this state may introduce interactions between polymer tails, especially on long templates, in *in vitro* settings. The importance of the interaction between polymer tails and the formation of bulges on the template should be considered on a case-by-case basis in experimental settings.

Finally, in our results, I have not focused on the speed with which copies are produced, only the relative yield of longer or full-length copies. In this setting, I observed that polymer tails that have some affinity to the template are advantageous in maximising the likelihood of producing full-length copies. However, such behaviour likely comes at the cost of somewhat slowed copy production, since template sites are more likely to be blocked. The optimal balance will depend on the system in question and again will require further investigation on a case-by-case basis.

In the following chapter, I focus on exploring how the mechanisms identified in this chapter could and should be implemented in an HMSD reaction network and how the rebinding of products from the solution to the template affects the overall performance (in terms of product yield, proxies for accuracy, and overall speed) of the copying process. I frame this as an exploration of the parameter space of a proposed design of a trimeric copying system, and seek to optimise the yield, proxies for accuracy and the overall reaction speed by tweaking DNA domain lengths, in an effort to bring these intriguing, but abstract, mechanisms closer to experimental implementation.

3.5 Data Availability

The code, input files, and analysis scripts required to reproduce the data presented in this chapter are available for download from the Zenodo file repository through the following link: DOI 10.5281/zenodo.5643001.

Chapter 4

Modelling a DNA strand displacement implementation of templated copying

4.1 Introduction

In Chapter 3, I analysed a templated copying process under the assumption that the concentration of polymer species in solution remained negligible. I sought to uncover mechanisms by which product inhibition arising from the failure to release the product from the template could be mitigated and how length and sequence control could be achieved by controlling the flux of polymers that detach from the template. However, in practice, any ideal templated copying reaction would be capable of converting the majority of the monomeric pool into polymers or polymer fragments, and therefore, the interaction of the template with the products in the solution cannot be neglected. Competitive inhibition was observed in HMSD [22, 21] and is expected to get worse with increased template length, as the products will have more ways of rebinding to the template. Therefore, I set about developing a model that explicitly accounted for the dynamics of polymers attached to the template and the interaction of the template complexes with the pools of polymers in solution, to see if there were regions in the parameter space for which these and other effects could be mitigated.

In developing this model, I also wanted to aid in the design and characterisation of a trimeric templated copying system, shown in Fig. 4.1, that is under development in our lab. This system is a DNA strand displacement network that uses handhold-mediated strand displacement [21, 22]. As I will explain in this chapter, I sought to parameterise the model using the small amount of experimental data available [21], such that the predictions the model makes are grounded in an accurate underlying representation of the reaction motifs.

Experimentally testing all variations of the system design is not feasible, and thus, narrowing down the

"design-space" with a computation model will save on time and resources. Further, with a predictive model of how the design parameters affect the thermodynamics and kinetics of the system, we can explicitly model sub-processes that would be challenging to resolve experimentally using fluorophore quencher pairs [21, 22]. In the lab, we will only get indirect reporters of strand combinations (e.g. dimers and trimers), with a limited capacity to discern the pathways by which they formed.

Furthermore, with a predictive model that accounts for the main failure modes of the system, we can constrain regions of the design-space that show good or bad performance, we can identify the main reaction pathways that result in a good or bad performance, and ground the design in quantitative statements about the kinetics of the assortment of processes involved. All models have limitations, and I do not claim that the model in this system will suppress all possible leak pathways. However, by explicitly including some of the main expected failure modes of the system into our model, we can at least search for ways around them.

4.2 Methods and Model

4.2.1 Methods: Construction of a DNA strand displacement reaction network model

In figure 4.1, I show the strand displacement network design that is this chapter's focus. The design for this circuit is attributable to members of the Ouldridge lab, including Thomas Ouldridge, Javier Cabello-Garcia and Rakesh Mukherjee (unpublished). The intended purpose of this reaction is to catalytically polymerise monomers into sequence-controlled trimers with minimal product inhibition.

The template in Fig. 4.1 is comprised of three "information-bearing" domains, interspaced with "backbone" domains, and an extended domain at the 3' end of the template. The monomers are comprised of a complementary information domain, a backbone domain, a toehold domain, and, at the start of the reaction, are hybridised with a lock strand covering the backbone and toehold domain with a number of internal C/C mismatches. Monomers have alternating backbone sequences to prevent monomers, such as monomer 2, from forming a hairpin loop structure that would render it unable to partake in the reaction cycle. There may be multiple "species" of each monomer type, each with the same backbone and a different information-bearing domain that encodes a different unit of information.

The proposed reaction cycle shown in Fig. 4.1 goes as follows: a) Locked monomers bind complementary information domains on the template; b) the lock strands are displaced as the backbones of the monomers bind to the template, healing a C/C mismatch; c) the toehold domains of neighbouring strands connect; d) toehold connection triggers another strand displacement reaction, this time polymerising neighbouring monomers together; e) the trimer unzips from the template; f) the trimer finally detaches from the end domain of the

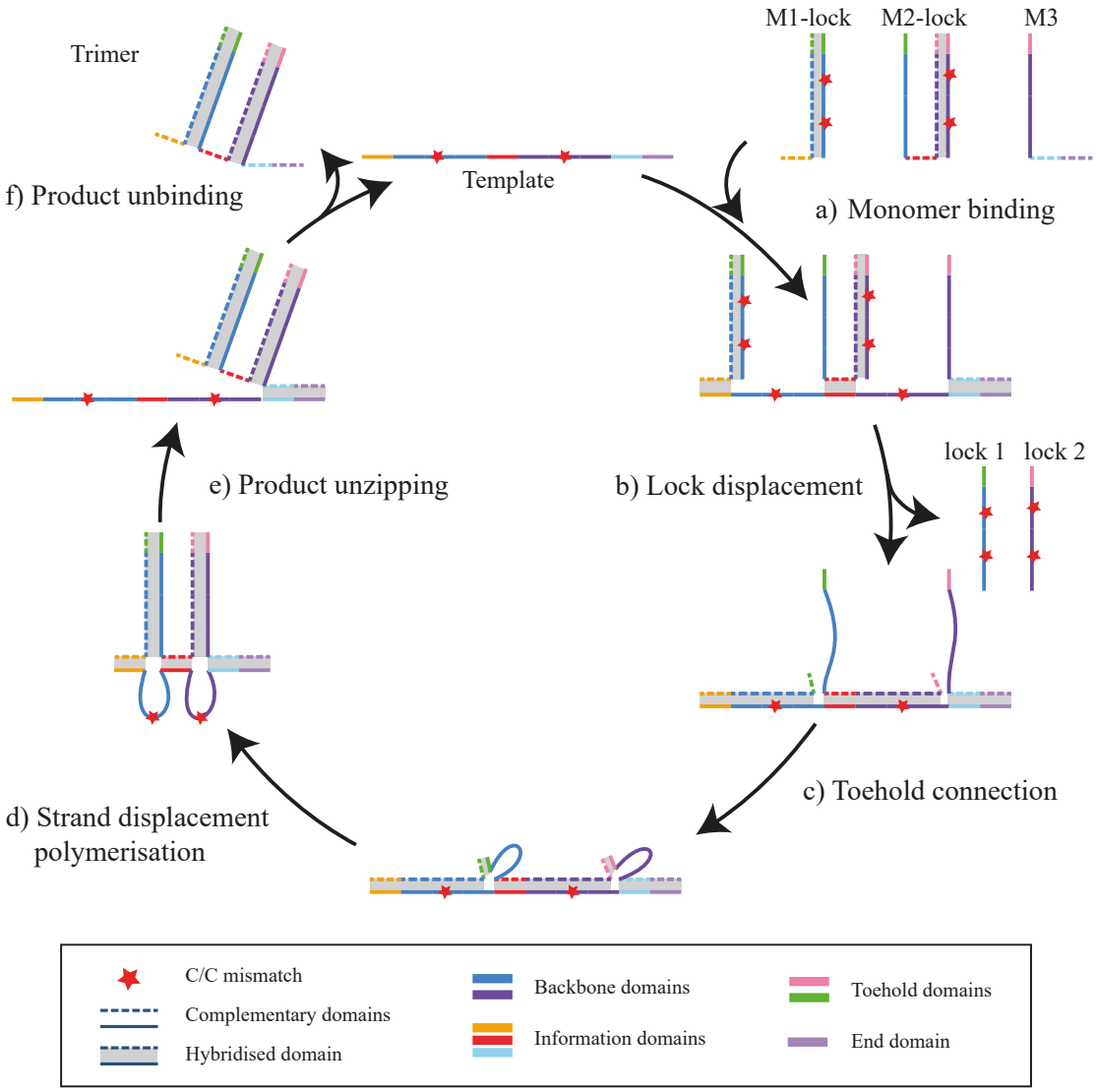


Figure 4.1: A domain-level depiction of the proposed strand-displacement reaction circuit for the templated production of trimers. The reaction cycle is illustrative only. The template catalyses the polymerisation of monomers M1, M2 and M3 to form a trimeric copolymer. Monomers with non-complementary information domains, that would not bind strongly to the template, have been omitted for visual clarity, but are also not included in the model. **a)** Monomers with locked backbone domains bind to the template from solution via the information-bearing domain. **b)** The monomers bind to the template displacing the lock strand and revealing the toehold. This step may heal one or more C/C mismatches, driving it forwards. **c)** The toeholds of neighbouring monomers connect, initiating another strand displacement. **d)** The neighbouring monomers polymerise through strand displacement. Further C/C mismatches may be healed, driving this reaction forward. Polymerisation reduces the number of bonds that monomers share with the template and creates loops in the template, reducing the affinity of the trimer and the template. **e)** The extended end domain keeps the trimer attached while other domains unzip. **f)** Finally, the trimer detaches from the template from the end domain. The template is once again available to catalyse further copolymerisation. Monomer binding may occur after or during step f. Content in this figure is redrawn and adapted from unpublished work by Rakesh Mukherjee (unpublished).

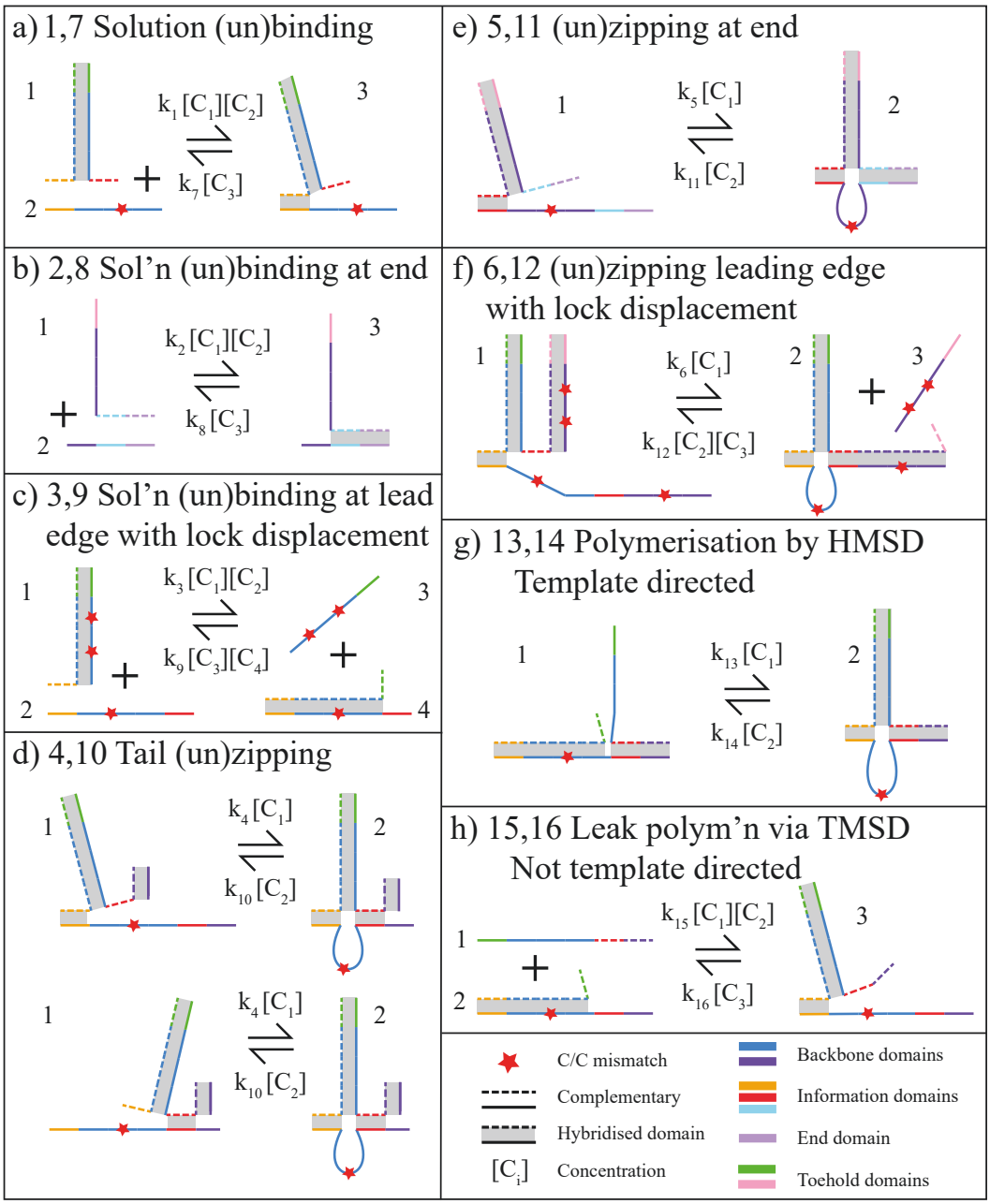


Figure 4.2: A domain-level depiction of the strand displacement reactions described in the model. Reactions shown are the minimal set of reactions that can occur in the model. Strands may be comprised of extra domains (not shown) which do not affect the rates of the reaction. **a)** Reactions 1,7 Solution binding and unbinding of the tails of polymers to and from the template. **b)** Reaction 2,8 Solution binding and unbinding of polymers or monomers at the end of the template. **c)** Reaction 3,9: Solution binding and unbinding of monomers or the leading edge of polymers from solution with lock displacement or invasion. **d)** Reaction 4,10: The zipping and unzipping of the tails of template-bound polymers. **e)** Reaction 5,11: The zipping and unzipping of template-bound polymers to the end of the template. **f)** Reaction 6,12: The zipping and unzipping of the leading edge of template-bound polymers with the displacement/invasion of a lock strand. **g)** Reaction 13,14: Polymerisation by handhold-mediated strand displacement. Toehold connection and the subsequent branch migration are treated as a single-step process. **h)** Reaction 15,16: Polymerisation by toehold-mediated strand displacement from solution. Here the polymerisation reaction is not template-directed and hence may result in inaccuracies in the copolymer sequence. Content in this figure is redrawn and adapted from unpublished work by Rakesh Mukherjee (unpublished).

template, completing the cycle. The reaction cycle consumes locked monomers and generates lock strands as waste, trimers as product, and heals a number of C/C mismatches that drive the process forwards.

The proposed mechanism is designed to achieve sequence control by precluding non-complementary monomers from binding to the template. As seen in Fig. 4.1a, monomers bind to the template first by the information-bearing domain, and subsequently form a cooperative bond with the template through strand displacement. If the information domains are not complementary, then the monomers will not form a stable connection with the template, and hence will rarely undergo a full strand displacement. Therefore, preventing the binding of incorrect monomers prevents the incorporation of the incorrect monomer into the growing chain. This mechanism was shown to be effective for achieving control over the sequence of dimers in Cabello-Garcia's thesis [21].

It is expected that the proposed mechanism in Fig. 4.1 should experience minimal product inhibition for the following reasons. While the monomers that drive the reaction (that are complementary to a site on the template) have a high affinity for the template, due to the healing of a number of C/C mismatches and the long backbone domain, the products of the reaction do not. As explored in Chapter 3 of this thesis, here the polymerisation of neighbouring monomers is coupled to a reduction in the number of bonds that the product makes with the template, as seen in step d) of Fig. 4.1. This reduction in the binding energy encourages product release and should prevent product rebinding. Further, as predicted in the previous chapter of this thesis, the binding of monomers underneath the formed trimer may further prevent product inhibition.

A priori, it is challenging to predict which design parameters must be tweaked in order for this reaction cycle to proceed as intended. A careful balance of rates will be needed to achieve tight length and sequence control while minimising the effect of leak reaction pathways. For instance, changing the length of the toehold and information-bearing domains can change the reaction rates by orders of magnitude and may also render the cycle susceptible to undesirable leak reactions, such as non-template-directed toehold-mediated strand displacement between monomers on the template and in solution [22]. Therefore, based upon measurements of the kinetics of HMSD [21], I developed a bottom-up model of the reaction cycle to explore the design space for desirable performance.

Current experimental efforts are directed toward extending the dimerisation reaction to operate on trimeric templates. Here, the ideal reaction cycle would produce a high concentration of trimers each of which have the same sequence as the template, a minimal concentration of dimers, and would do so rapidly. I explicitly chose to neglect the sequence-dependent properties of the system and model the process as a homopolymerisation, rather than a copolymerisation process, with proxy measures for the sequence-related effects that we seek to minimise. There were four reasons for this. First, we assume that the binding of incorrect monomers to the

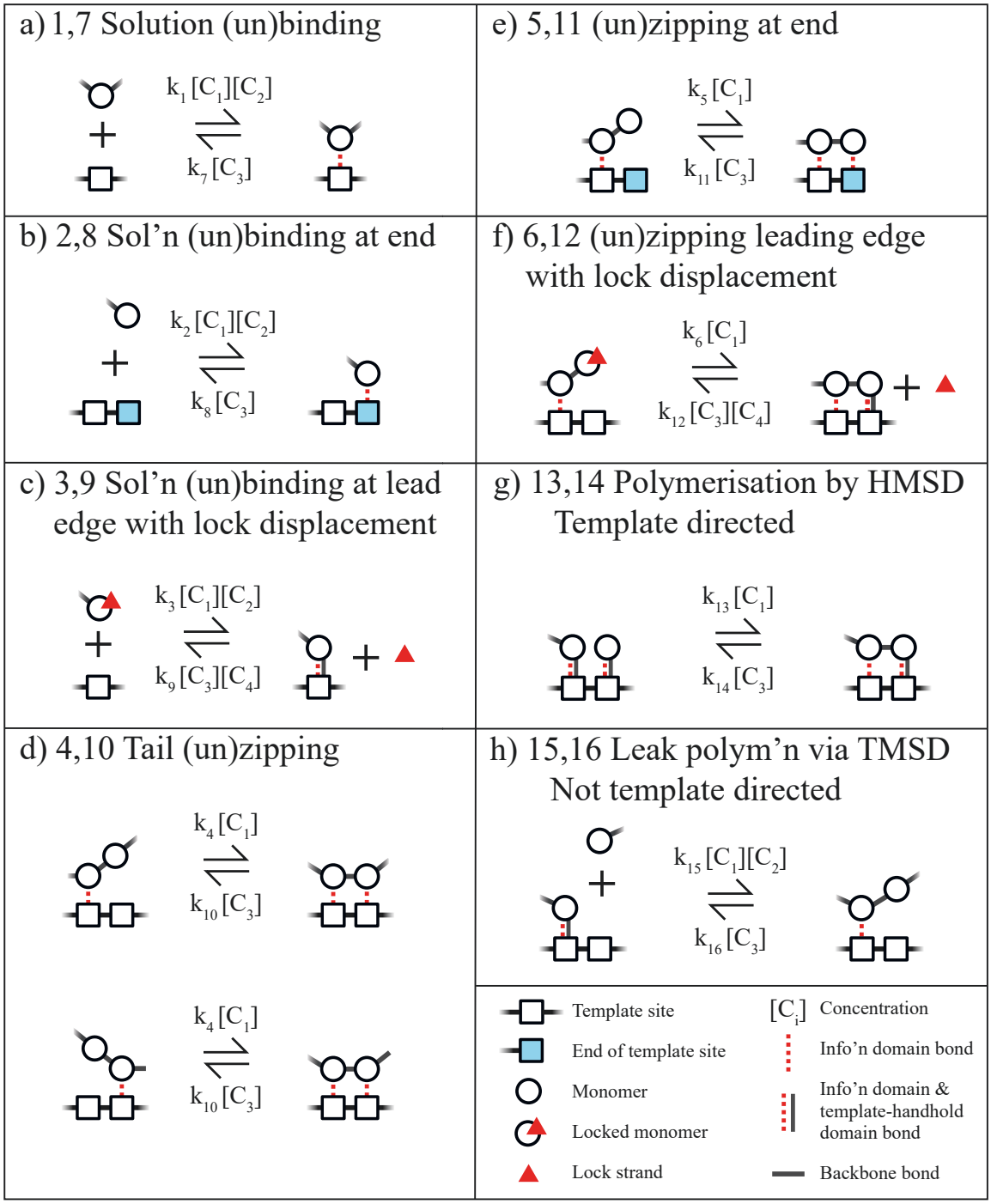


Figure 4.3: A simplified depiction of the minimal reaction motifs that comprise the model. The reactions are the same as described in Fig. 4.2, though here they are represented with a two-bond notation. Squares represent template sites and circles represent monomers. Red dashed lines represent hybridised information domains and dark grey lines represent hybridised backbone domains either between monomers and the template or bewtween monomers in a polymer. The lock strand is represented as a red triangle. Bonds with a fading gradient indicate that other units may or may be attached elsewhere on the polymer, but they do not affect the kinetics.

template and their subsequent polymerisation into polymers through the standard reaction pathway shown in Fig. 4.1, is expected to be minimal as complete monomer binding involves first forming a stable complementary bond with the template. If the monomer is not complementary to the template site, then it is unlikely to bind via strand displacement (the rate of toehold-mediated strand displacement is very low for non-complementary toeholds). Thus, we expect the main sources of sequence errors in the final polymer to arise from non-template mediated effects, which can be quantified indirectly as described in section 4.2.11. Secondly, the leak reaction (non-template directed polymerisation) can introduce sequence errors and its effect can be quantified without explicitly accounting for the monomer sequence, as will be explained in section 4.2.1. Thirdly, other desirable qualities, such as length control and overall reaction speed are independent of monomer sequence. Finally, the model is simpler without accounting for the monomer sequences, leading to reduced computational complexity and run-time.

A coarse-grained representation of the underlying reaction motifs

The model in this chapter is based upon a set of reaction motifs shown in figures 4.2 and 4.3, represented on the domain level and as "sticks and blobs" respectively. All reactions in the model are assumed to be reversible and either first or second-order in concentration with mass action kinetics [132]. The states and reactions included in the model were chosen to best represent the macrostates of the system that were determined experimentally for a dimeric template [22, 21].

The reactions are labelled 1 through 16. A brief description of each pair of reactions is given below.

Reactions 1 and 7: Solution binding and unbinding of the tails of polymers to and from the template.

Reactions 2 and 8: Solution binding and unbinding of polymers or monomers at the end of the template.

Reactions 3 and 9: Solution binding and unbinding of monomers or the leading edge of polymers from solution with lock displacement or invasion. In the forward direction, this reaction heals a number of C/C mismatches, suppressing the rate of reaction 9.

Reactions 4 and 10: The zipping and unzipping of the tails of template-bound polymers.

Reactions 5 and 11: The zipping and unzipping of template-bound polymers to the end of the template.

Reactions 6 and 12: The zipping and unzipping of the leading edge of template-bound polymers with the displacement/invasion of a lock strand. In the forward direction (6), this reaction heals a number of C/C mismatches, suppressing the rate of reaction 12.

Reactions 13 and 14: Polymerisation by handhold-mediated strand displacement. Toehold connection and the subsequent branch migration are treated as a single-step process. In the forward direction (13), this reaction heals a number of C/C mismatches, suppressing the rate of reaction 14.

Reactions 15 and 16: Leak reaction. Polymerisation by toehold-mediated strand displacement from solution. Here the polymerisation reaction is not template-directed and hence may result in inaccuracies in the copolymer sequence. In the forward direction (15), this reaction heals a number of C/C mismatches, suppressing the rate of reaction 16.

Reactions 1 and 2 and 7 and 8 are simple bimolecular binding and unimolecular unbinding reactions. Reactions 3, 9 and 15 are bimolecular toehold-mediated strand displacement reactions. Reactions 4, 5, 10, 11, 13, 14 and 16 are all unimolecular reactions.

The set of 16 fundamental reaction motifs can occur at multiple points on the template, and hence there are many more than 16 actual transitions between isolated template configurations in the complete model. For instance, polymerisation could occur between monomers bound on sites 1 and 2 on the template, without being affected by the myriad of different conformations involving site 3 or the tails of non-bound polymers, and also at sites 2 and 3 without being affected by the conformations involving site 1. Hence, the reaction could occur in many conformations for each reaction motif listed above. The scaling of the total number of pairs of reactions with the template length n is given in Table 4.1.

Template length, n	**2**	**3**	**4**	**5**	**6**
Number of row vectors	6	16	36	71	127
Number of "in solution" species	3	6	10	15	21
Number of "template bound" species	10	97	1565	38333	817597
Total number of species	13	103	1575	38348	817618
Number of reactions	18	328	7978	264219	-

Table 4.1: The scaling of the number of chemical species and the number of reactions in the system with the template length n. The number of chemical species and differential equations (Eqn. 4.4) that must be solved grows quickly with template length. Methods to reduce the number of species, such as neglecting low-concentration chemical species and discarding transitions that only affect the system on very long timescales would be required to simulate systems of length $n > 4$.

4.2.2 Chemical reaction network model

I model the system as a chemical reaction network comprised of template-bound and in-solution species and simulate the system as a set of deterministic ODEs. In the model, I assume that polymers cannot grow longer than the length of the template itself, and therefore the maximum length of a polymer is the length of the template

n. Therefore, an in-solution chemical species exists for each possible contiguous sub-sequence of a polymer of length n in solution. These chemical species are monomers, M_1, M_2, M_3, $\ldots M_n$, dimers $D_{1,2}$, $D_{2,3}$, $D_{3,4}$, $\ldots D_{n-1,n}$, trimers $T_{1,2,3}$, $T_{2,3,4}$, $\ldots$, $T_{n-2,n-1,n}$, and so on up to and including a complete polymer $N_{1,2,3,\ldots,n-1,n}$.

The lock strand is treated as an external chemical species that is chemostatted at a concentration of 10nM. Aside from the lock strand, the chemical reaction network is closed. For a template of length n, there are $\frac{n(n+1)}{2}$ "in-solution" species.

There is a template-bound (internal) chemical species for each configuration of monomers and polymers on the template that is reachable from the empty template configuration via the reaction motifs described in Fig. 4.3. I discuss the method for determining reachability and constructing the network of configurations in section 4.2.3. Here, I briefly describe how I convert the chemical reaction network to a set of differential equations that can be solved.

Using the nomenclature described in Ref. [131], a general chemical reaction network consisting of a set of chemical species with index $\sigma \in \mathscr{S}$, X_σ, with molarity N^σ, concentration $Z^\sigma \equiv N^\sigma/V$, that interact via reversible reactions (+ forward, - reverse) $\pm\rho \in \mathscr{R}$ with rate constants $k^{\pm\rho}$ can be written as

$$\sum_\sigma \nabla^\sigma_{+\rho} \mathrm{X}_\sigma \underset{k^{-\rho}}{\overset{k^{+\rho}}{\rightleftharpoons}} \sum_\sigma \nabla^\sigma_{-\rho} \mathrm{X}_\sigma, \tag{4.1}$$

where $\mp\nabla^\sigma_{\pm\rho}$ are the stoichiometric coefficients of the reactions in the forward and reverse directions describing the number of molecules X_σ that contribute to the reaction in each direction. We can construct non-negative matrices $\nabla_\pm \equiv \left[\nabla^\sigma_{\pm\rho}\right]$ in which each column corresponds to a reaction ρ and each row corresponds to the whether species σ is involved in reaction ρ in the respective direction. We construct the stoichiometric matrix by taking the difference of these matrices in the forward and reverse direction $\nabla \equiv \nabla_- - \nabla_+$. The stoichiometric matrix for our model on templates of length $n = 2$ and $n = 3$ is shown in Fig. 4.4.

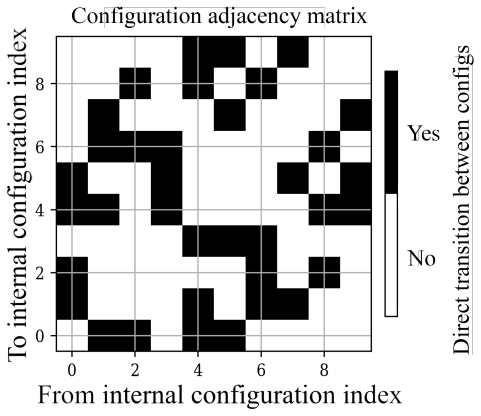
I assume that the rate of each reaction occurs with mass action kinetics [131, 132] with fluxes

$$J^{\pm\rho} \equiv k^{\pm\rho} Z^{\sigma\,\nabla^\sigma_{\pm\rho}} \tag{4.2}$$

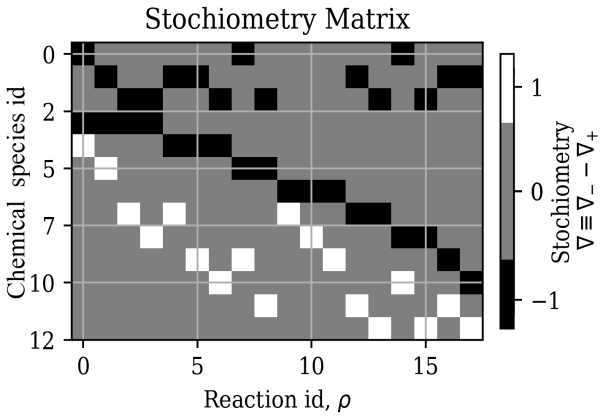
and a net flux for each reaction ρ

$$J^\rho \equiv J^{+\rho} - J^{-\rho} = k^{+\rho} Z^{\sigma\,\nabla^\sigma_{+\rho}} - k^{-\rho} Z^{\sigma\,\nabla^\sigma_{-\rho}} \tag{4.3}$$

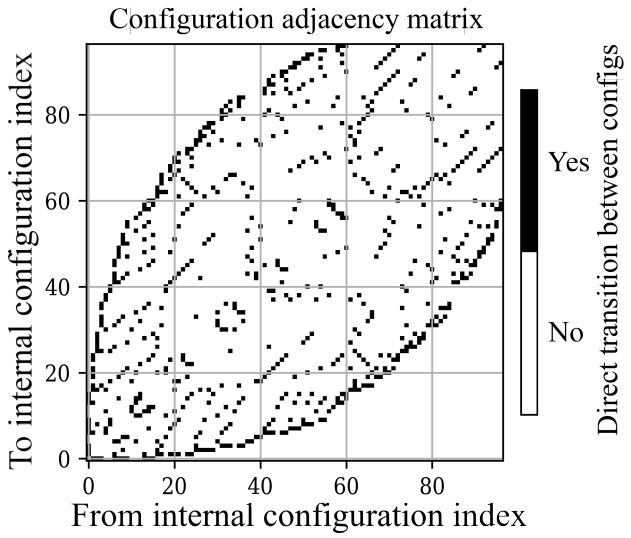
where $Z^{\sigma\,\nabla^\sigma_{\pm\rho}} = \prod_\sigma Z^{\sigma\,\nabla^\sigma_{\pm\rho}}$ is the product of concentrations over species that are involved in reaction ρ.



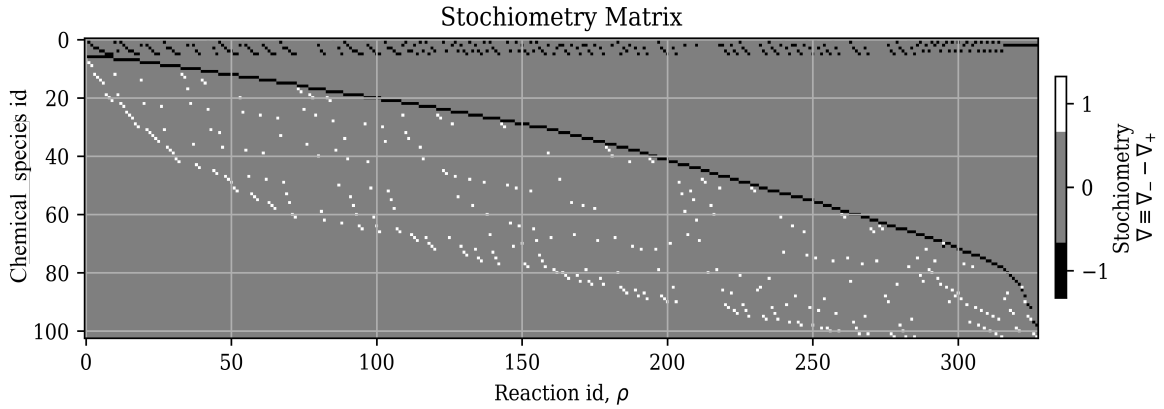
(a) Adjacency matrix for template length $n = 2$.



(b) Stoichiometry matrix for template length $n = 2$.



(c) Adjacency matrix for template length $n = 3$.



(d) Stoichiometry matrix for template length $n = 3$.

Figure 4.4: The adjacency matrices and stoichiometry matrices for template lengths of $n = 2$ and $n = 3$. The black elements in the adjacency matrices indicate that the conformations associated with the row index and column index can be reached directly from one another via an elementary reaction motif. The grey squares in the stoichiometry matrix indicate that there is no net production of the species in the reaction. The black (white) squares indicate that the chemical species is consumed (generated) by the reaction according to the conventional reaction direction.

The concentration of each chemical species, Z^σ, evolves in time as

$$\frac{\partial Z^\sigma}{\partial t} = \sum_\rho \nabla_\rho^\sigma J^\rho = \sum_\rho \nabla_\rho^\sigma \left(k^{+\rho} Z^{\sigma \nabla_{+\rho}^\sigma} - k^{-\rho} Z^{\sigma \nabla_{-\rho}^\sigma} \right) \tag{4.4}$$

where ∇_ρ^σ is the element of the stoichiometry matrix in row σ and column ρ.

Eqn. 4.4 can be solved as a set of coupled differential equations. Before I discuss the method by which these equations were solved, I describe the method by which the set of chemical species, the set of reactions and this set of differential equations were constructed computationally.

4.2.3 Computational construction of a chemical reaction network model

Here I describe the method by which I construct the model for a template of arbitrary length, n. I use a template length of $n = 3$ as the focus of the rest of this chapter.

As mentioned in section 4.2.2, each chemical species described in the model, apart from the chemostatted lock strand, corresponds to either a monomer or part of a polymer in solution (solution species) or a particular configuration of monomers and polymers bound to the template (template-bound species) or the empty template configuration.

As shown in Table 4.1, many possible conformations can be generated by applying the transitions depicted in figures 4.2 and 4.3. Therefore, there can be a large number of chemical species and transitions that convert species in the model. Constructing the set of ODEs by hand (Eqn 4.4) is intractable for templates of length 3 or greater. Instead, I used a pattern-matching computational approach to enumerate all possible conformations (each one corresponding to a chemical species), construct the set of transitions between configurations, and set up a system of ODEs (Eqn 4.4), which were then solved numerically.

First, I developed a way to uniquely describe the conformational state encoded by each template-bound chemical species in the model (internal species). The conformational state encoded by each template-bound chemical species is defined by an $n \times n$ matrix, S, whose rows can be permuted without changing the encoded conformation. On a template of length n, there can be at most n bound monomers or polymers on the template. Each polymer must be bound to the template by at least one domain. The rows of the conformation matrix S encode the conformation of each contiguous polymer or monomer on the template. The elements of each row may be either 1, 0 or -1. If element i of the row has a value 1, this indicates a template-bound monomeric unit is at location i on the template. If element i of the row has a value -1, this indicates that there is a monomeric unit "above" site i on the template. This monomeric unit must be connected to at least one template-bound unit elsewhere in the polymer, corresponding to a 1 value elsewhere in the row. If element i of the row has a value 0,

this indicates that there is no monomeric unit within this polymer conformation above or on template site i.

To construct all possible conformation matrices, S, I begin by constructing all physically possible row vectors. First, I create a list of row vectors corresponding to all possible permutations of n choices of the elements $1, 0, -1$, of which there are 3^n. I discard all row vectors that contain non-contiguous blocks of 1's or -1's e.g. [1,1,0,-1,1] would be discarded, whereas [0,0,-1,1,0] would be kept. Secondly, I discard all row vectors that do not contain at least one template bound unit (a value 1) e.g. [0 0 -1 -1 0] would be discarded. Finally, I discard any conformations that feature bulges between template bound units e.g. [1 1 -1 -1 1 1] would be discarded while [1 -1 -1 0 0] would be accepted. I discarded any vectors that encode bulges as I assumed that these conformations would be highly unstable, and thus wouldn't affect the overall dynamics [147]. The remaining row vectors are referred to as valid row vectors. They encode all physically plausible conformations of contiguous polymers and monomers bound to the template. The scaling of the number of valid row vectors with the length of the template is shown in Table 4.1.

I combine row vectors to construct the unique conformations, S. Valid conformations cannot feature multiple monomeric units bound to each template site. Row vectors are said to overlap if at any site $0 \leq i < n$ more than one row, j, contains a 1. Iterating through all combinations of row vectors is exponentially more computationally expensive with growing template length. Filtering valid rows would require the recomputation of the overlap of repeated pairs of row vectors many times. Therefore, I construct the set of template-bound conformations Σ hierarchically, first generating the empty template conformation, then all conformations that contain one contiguous polymer or monomer, then all conformations of two contiguous, non-overlapping polymers, and so on, to avoid the recomputation of overlaps. Intuitively, as more contiguous polymers (each corresponding to a valid row vector) are added to the template, there are fewer valid locations in which a polymer can reside as there are more opportunities for overlap.

I begin by computing and storing whether each pair of row vectors overlap. I construct a symmetric row adjacency matrix, rA, that contains a 0 at site (i, j) and (j, i) if the i'th and j'th row vectors overlap and 1 if they do not. For each row of rA, a set L_{I} of indices j for which $rA(i, j) = 1$ is constructed. The set L_{I} contains the indices of the row vectors that are non-overlapping with the row vector with index i.

The set of template-bound conformations, Σ, is then constructed iteratively. I start by constructing the set of all conformations in which a single polymer occupies the template, Σ_1 , e.g. one valid row vector and $n-1$ rows of 0's . For each conformation $S_{1,i} \in \Sigma_1$, I store a list of all the row vectors that overlap with it $L_{1,i}$ e.g. a list of the rows which can be added to the conformation $S_{1,i}$ without overlap.

Next, I construct the conformation set containing two template-bound contiguous polymers Σ_2. To construct the set of 2 polymer conformations $S_{2,i} \in \Sigma_2$, I combine every conformation of templates bound by single

polymers $S_{1,i}$ with every row in the list of non-overlapping row vectors $L_{1,i}$. Then I construct a list of all row vectors that are non-overlapping with $S_{2,i}$, $L_{2,i}$, by calculating the intersection of $L_{1,i}$ with the overlap set of each row vector j added, L_j.

Next, I construct the conformations containing three contiguous polymers Σ_3. To construct the set of 3 polymer conformations $S_{3,i} \in \Sigma_3$, I combine every 2-bound template conformation $S_{2,i}$ with every row vector in the list of non-overlapping row vectors $L_{2,i}$. Then I construct a list of all row vectors that are non-overlapping with $S_{3,i}$, $L_{3,i}$, by calculating the intersection of $L_{2,i}$ with the overlap of each row vector j added, L_j.

This process is repeated until the set of n-bound conformations (same number of polymers as the template length) Σ_n has been constructed. The complete set of template-bound configurations, Σ is given by the union of the empty template conformation, Σ_0, the set of 1-bound ,Σ_1, 2-bound Σ_2, 3-bound Σ_3, ... and n-bound Σ_n conformations. The size of the set Σ defines the number of chemical "template-bound" chemical species in the model, each corresponding to a unique conformation of polymers on the template. The scaling of the number of "in-solution" and "template-bound" species with the template length n is shown in Table 4.1.

4.2.4 Transitions between conformations

Having encoded and enumerated all "in-solution" and "template-bound" chemical species in the model, I then construct the network of reactions that inter-convert these species.

All transitions (1 through 16) in Fig. 4.2 and Fig. 4.3 between conformations that do not involve polymerisation or depolymerisation (reactions 13, 14, 15 and 16), involve a modification to just one row of the matrix encoding the template-bound conformation, S. Polymerisation and depolymerisation involve a modification to two rows of the conformation matrix, S. These two connections either join together two separate polymer conformations or separate a single polymer into two separate conformations.

I build the network of internal transitions between the conformations by first identifying how the transitions (1 through 16) in Fig. 4.2 and Fig. 4.3 change the row vectors that comprise the conformation matrices S. For each type of reaction in 4.2 and 4.3, I first identify all valid row vectors that this type of reaction could apply to in the "forward" direction. I call these vectors the "from" row vectors. Applying the reaction modifies the row vector, resulting in a "to" row vector. For each reaction type (aside from reactions that involve polymerisation and depolymerisation 13 and 14), I run through all row vectors and identify which of them are "from" vectors for this particular reaction. For each "from" vector, I calculate all possible "to" row vectors that can be reached from this "from" vector for this particular type of reaction. For depolymerisation, I identify all row vectors ("to" vectors) which contain contiguous template-bound monomeric units that comprise part of a polymer (e.g. stretches of

1's). The "from" vectors are all possible pairs of vectors (buffered with 0s) that can be generated by splitting the "to" vector at at pairs of 1's (e.g to row [1,1,1] from pairs of rows ([1,0,0],[0,1,1]) and ([1,1,0],[0,0,1])).

Once I have constructed these mappings for each reaction type, I iterate over the rows of each conformation matrix $S \in \Sigma$ and check whether each reaction type (1 through 16) is applicable. If the conformation matrix S contains a row vector that is a "from" row vector for a particular reaction, then this reaction is possible from this conformation. Given the reaction is possible, I check whether replacing the "from" row vector in S with each the corresponding "to" row vector $S \rightarrow S_{\text{new}}$ generates a new in a valid conformation S_{new} (e.g. a conformation already in the set of valid conformation $S_{\text{new}} \in \Sigma$). If the new conformation is valid, then the reaction with reaction id, r, can map the original conformation with index i to the destination conformation with index j. In this case, I set the conformation adjacency matrix entries $A(i,j) = A(j,i) = 1$, otherwise they are 0. Fig. 4.4a and Fig. 4.4b show the conformation adjacency matrix for templates of length 2 and 3. This conformation adjacency matrix A encodes which template-bound chemical species can be interconverted by a direct transition.

The stoichiometry matrix ∇ is constructed from the entries in the conformation adjacency matrix A. The stoichiometry matrix has as many rows as there are chemical species (both "in solution" and "template bound") and as many columns as there are unique reactions between conformations (given by the number of non-zero entries in the upper right triangular form of the conformation adjacency matrix A). The i'th row of column j of the stoichiometry matrix will be -1 or 1 if one unit of species i is respectively consumed or generated in reaction j, and 0 if the species i is not involved in reaction j. Fig. 4.4 shows the stoichiometry matrix for template lengths 2 and 3.

4.2.5 Model parameterisation: Reducing the number of free rate constants

Having constructed the set of chemical species and the stoichiometry matrix that describes the chemical reaction network, the rate constants $k_1 \ldots k_{16}$ needed to be determined. Here I employ physical arguments about the processes involved in the reactions depicted in Fig. 4.3 and the principle of detailed balance to ensure the thermodynamic self-consistency of the model and to reduce the number of free parameters in the model from 16 to four rate constants and 8 standard free energies [129].

The principle of local detailed balance states that the ratio of the rate constants of a pair of reversible reactions with rate constants k_f and k_r in the forward and reverse direction is constrained by the standard free energy change, ΔG^o, of the reaction

$$\frac{k_f}{k_r} = e^{-\Delta G^o/(k_B T)}, \tag{4.5}$$

where T is the temperature in Kelvin, k_B is the Boltzmann constant and the standard free energy is measured at standard concentrations of 1M [129]. I begin by assuming that all bimolecular binding reactions proceed with mass action kinetics and the same diffusion-limited rate constant and therefore, $k_1 = k_2$. Applying Eqn. 4.5, we can constrain the unbinding rate constants

$$k_7 \equiv k_1 e^{\Delta G^o_{1,7}/(k_B T)}[1M] \tag{4.6}$$

$$k_8 \equiv k_1 e^{\Delta G^o_{2,8}/(k_B T)}[1M]. \tag{4.7}$$

I assume that the rate constants for the unzipping of polymers are equal to the rate constants associated with the unbinding to solution of polymers,

$$k_{10} \equiv k_7 = k_1 e^{\Delta G^o_{1,7}/(k_B T)}[1M] \tag{4.8}$$

$$k_{11} \equiv k_8 = k_1 e^{\Delta G^o_{2,8}/(k_B T)}[1M]. \tag{4.9}$$

By extension, I constrain the rate constants of zipping, k_4 and k_5 by again applying Eqn. 4.5 such that

$$k_4 \equiv k_{10}\, e^{-\Delta G^o_{4,10}/(k_B T)} = k_1\, e^{(\Delta G^o_{1,7} - \Delta G^o_{4,10})/(k_B T)}[1M] \tag{4.10}$$

$$k_5 \equiv k_{11}\, e^{-\Delta G^o_{5,11}/(k_B T)} = k_1\, e^{(\Delta G^o_{2,8} - \Delta G^o_{5,11})/(k_B T)}[1M]. \tag{4.11}$$

As a result of the free energy differences $\Delta G^o_{1,7} - \Delta G^o_{4,10} = \Delta G^o_{2,8} - \Delta G^o_{5,11}$ being equal, the rate constants are also equal, $k_4 = k_5$.

Reactions 3 and 9 are toehold-mediated strand displacement reactions with a toehold equal to the information domain length in the forward direction and equal to what I have referred to as the toehold domain length in the reverse direction. The reaction heals a number of C/C mismatches in the forward direction and must overcome the same number in reverse. If reaction 3 proceeds with a rate constant k_3, then we can constrain k_9 as

$$k_9 \equiv k_3 \, e^{\Delta G^o_{3,9}/(k_B T)}. \tag{4.12}$$

For the sake of consistency, we assume that the rate constant of unzipping at the leading edge k_{12} is the same as the rate constant of unbinding from the leading edge by lock invasion k_9

$$k_{12} \equiv k_9 = k_3 \, e^{\Delta G^o_{3,9}/(k_B T)}. \tag{4.13}$$

Applying the principle of detailed balance, we can also set the leading edge unzipping rate

$$k_6 \equiv k_{12} \, e^{-\Delta G^o_{6,12}/(k_B T)}[1M] = k_3 \, e^{(\Delta G^o_{3,9} - \Delta G^o_{6,12})/(k_B T)}[1M]. \tag{4.14}$$

I constrain the rate of depolymerisation $k_{14} \equiv k_{13} \, e^{\Delta G^o_{13,14}/(k_B T)}$ by the rate constant for polymerisation k_{13}. Finally, the reverse leak reaction is constrained by $k_{16} = k_{15} \, e^{\Delta G^o_{15,16}/(k_B T)}[1M]$.

Following this method the number of free rate constants in the model is reduced from 16 to 4, k_1, k_3, k_{13} , k_{15} with 8 standard free energy changes. In the following section, I describe how modifications to standard free energy change predictions calculated using Nupack were needed [153, 154, 155].

4.2.6 Model parameterisation: Calculating standard free energies

I used Nupack to calculate the standard free-energy changes of the various transitions described in the model [153, 154, 155]. I set the temperature to 25°C and the concentration of salt as 1M [NaCl], in accordance with the conditions described in Cabello-Garcia's thesis and characterisation of HMSD [21, 22]. In order to estimate the free-energy of a structure in Nupack, a "dot-parens-plus" (DPP) description of the base-pairing pattern in the structure, and the sequences of the individual strands in the structure must be supplied. The DPP notation used to describe each conformation before and after the transition and the corresponding sequences used with illustrative domain lengths (information domain length = 5, backbone domain length 7, toehold length 3, end domain length 4, with mismatches at positions 2 and 5 in the lock backbone domain and only at position 2 in the template backbone domain) is provided in Table 4.2. The sequence of the information, toehold and end domains were the same as in Cabello-Garcia's thesis [21], and the backbone domains were set to polyC and

Name	Dot-Parens-Plus Before	Dot-Parens-Plus After	Sequences
$\Delta G^o_{1,7}$	..... + ...	.....(((((+))))).....	CCGGGGGGGGCGGAT ATCCGGGGGGGGATCCG
$\Delta G^o_{2,8}$	..... + ...	.....(((((((((+))))))))).....	CCGGGGGGGGCGGATGTGA TCACATCCGGGGGGGGATCCG
$\Delta G^o_{3,9}$	.....((.((.((((+)))).)).)) + ...	(((((((.(((... + ... + ...)))).)))))))	CGGATCCCCCCCCGG CCGGCGGCGG ATCCGGGGGCGGATCCG
$\Delta G^o_{4,10}$	((((((((((((((+)))))))))))..... + ...))))))	((((((((((((((+))))))))))((((+)))))).....))))))	CGGATCCCCCCCCGG CCGGGGGGGGCGGAT ATCCGGGGGGGGATCCG
$\Delta G^o_{5,11}$	((((((((((((((+)))))))))))..... + ...))))))	((((((((((((((+))))))))))(((((((((+)))))))).....))))))	CGGATCCCCCCCCGG CCGGGGGGGGCGGATGTGA TCACATCCGGGGGGGGATCCG
$\Delta G^o_{6,12}$	((((((((((((((+))))))))))).....((.((.((((+)))).)).)) + ...))))))	((((((((((((((+))))))))))(((((((.((((... + ... +))))).))))))).....)))))	CGGATCCCCCCCCGG CCGGGGGGGGCGGATCCCCCCCCCGG CCGGCGGCGG GGGGCGGATCCGGGGGCGGATCCG
$\Delta G^o_{13,14}$	(((((((.((((... + ...(((((+)))))))).))))))))	((((((((((((((+))))))))))(((((+)))))).....))))))	CGGATCCCCCCCCGG CCGGGGGGGGCGGAT ATCCGGGGGCGGATCCG
$\Delta G^o_{15,16}$	(((((((.((((... + ... + ...)))).)))))))	((((((((((((((+)))))))))))..... + ...))))))	CGGATCCCCCCCCGG CCGGGGGGGGCGGAT ATCCGGGGGCGGATCCG

Table 4.2: The standard free-energy change for each reaction shown in Fig. 4.2 (and repeated in Fig. 4.3) was calculated using Nupack [153, 154, 155]. A dot-parens-plus structure defining the conformation of the strands before and after the transition and a sequence for each strand was required. In DPP notation the symbol . represents an unpaired base, (and) represent paired bases, and + represents a break between strands. The lengths shown here are illustrative. The information domain length is 6, the end domain is 3, the toehold is 2, and the backbone domain is 10. All mismatches were of the type C/C. As the backbone-domains are never assumed to break spontaneously and only engage in strand displacement reactions, the specific sequence of the backbone domain was neglected. The free energies were estimated at a temperature of 25 °Cin 1M NaCl, in accordance with the conditions reported in Ref.[21, 22]. Nupack calculates standard free energies at a standard concentration of 1 mole fraction, therefore to calculate the standard free energy at a standard concentration of 1M an offset $(m-1)k_BT\ln(\rho_{H2O}/(1\,\text{mol/liter}))$ was applied where $\rho_{H2O} \approx 55.14$ mol/liter and m is the number of bound strands in the complex. The standard free energy change was used to constrain the ratio of rate constants. An additional correction function was applied to all free energy changes that involved the formation of a loop ($\Delta G^o_{4,10}$, $\Delta G^o_{5,11}$, $\Delta G^o_{6,12}$, and $\Delta G^o_{13,14}$), as discussed further in Fig.4.5 and Table 4.3

Name	Dot-Parens-Plus Before	Dot-Parens-Plus After	Sequences
ΔG^o Bulge loop	(((((((..... +.....)))))))	(((((((.....((((((+)))))))))))))	CGGATGTCCCCCCCCCCGAAGTG CACTTCACATCCG
ΔG^o dsDNA spanning loop	((((((((((+))))))..... +.....))))))	((((((((((+)))))((((((+))))))....))))))	CGGATGTGGGGG CCCCCCGGATG CATCCGCCCCCCCCCCACATCCG

Table 4.3: Forming a loop structure in a complex of DNA strands incurs an entropic penalty to the free energy that grows with the length of the loop [147]. The DPP and sequences given in this table were used to compare the scaling of the free energy with loop length in the case where the strand opposite the loop contained secondary structure in the region spanning the loop $\Delta G^o_{\text{dsDNA spanning loop}}$ to the scaling without secondary structure $\Delta G^o_{\text{Bulge loop}}$. The scaling is shown in Fig. 4.5.

polyG respectively. Any mismatches included in the model were of the type C/C.

Nupack defines concentrations in units of mole fractions, not molarities, as described in footnote 13 of Ref. [153]. To estimate the standard free energy at a standard concentration of 1M, an offset of $+(m-1)\,k_BT\ln(\rho_{H_2O}/(1\text{mol/liter}))$ (with $\rho_{H_2O} = 55.14$mol/liter) was applied to the standard free energy predicted by Nupack for a complex of m strands.

A further correction to the standard free energy estimate needed to be made in the case of transitions that formed loop structures $\Delta G^o_{4,10}, \Delta G^o_{5,11}, \Delta G^o_{6,12}$ and $\Delta G^o_{13,14}$, as can be seen in Fig. 4.2d,e,f,g. A standard assumption in simple polymer models is that the entropic penalty associated with forming a loop structure scales logarithmically with the length of the loop [147], as shown by the red dots in Fig. 4.5 in the case that there is no secondary structure in the ssDNA region of the complementary strand. However, if secondary structures are present in the strand/s spanned by the bulge loop, Nupack predicts a linear scaling with the loop length (see the blue dots in Fig. 4.5). We suspect that this is an error, and therefore applied a correction function. The DPP and sequences used to generate the curves shown in Fig. 4.5 are shown in Table.4.3. The correction function was insensitive to domain lengths other than the backbone domain, and therefore the correction function was solely a function of the loop length.

To correct the length such that the standard free energy scales logarithmically with the loop length L, I fitted a curve $y_{noSS}(L) = c_1 + m_1\ln(L)$ to the standard free energy changes with no secondary structure in the region spanned by the loop (red dots in Fig. 4.5), and fitted a linear curve $y_{SS}(L) = c_2 + m_2 L$. I applied a correction function $C(L) = m_1\ln(L/20) - m_2(L-20) = 0.765 + 1.028\ln(L) - 0.192\,L$, to the standard free energy, calculated for each transition that formed a loop structure $\Delta G^o_{4,10}, \Delta G^o_{5,11}, \Delta G^o_{6,12}$ and $\Delta G^o_{13,14}$. The correction function is 0 for a loop length of 20, $C(L=20) = 0$ by choice, it reduces the sensitivity of the standard free energy to changes to the backbone length, and is presented as the black dot-dashed line in Fig. 4.5.

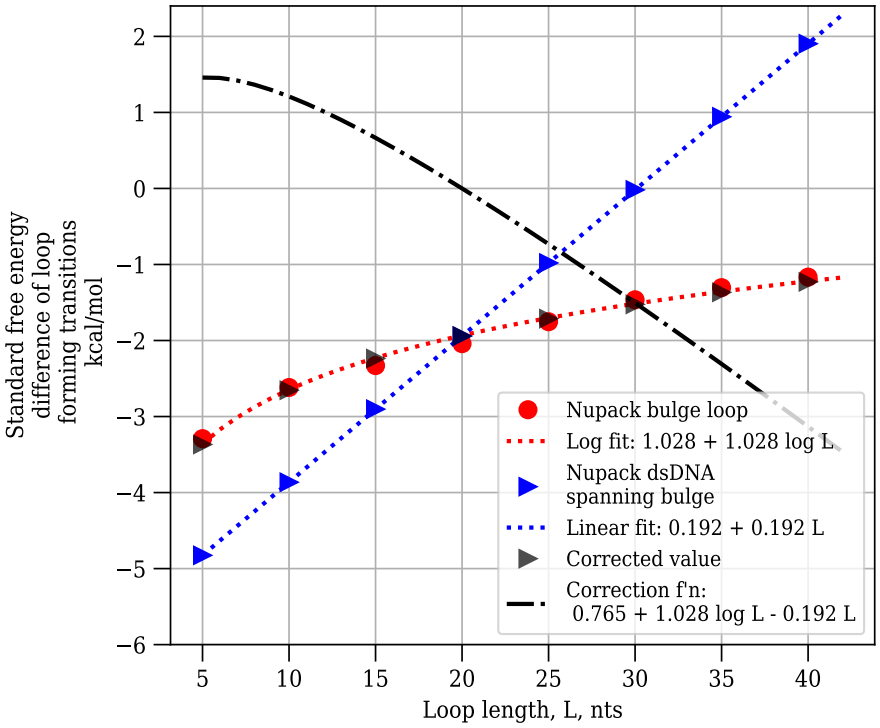
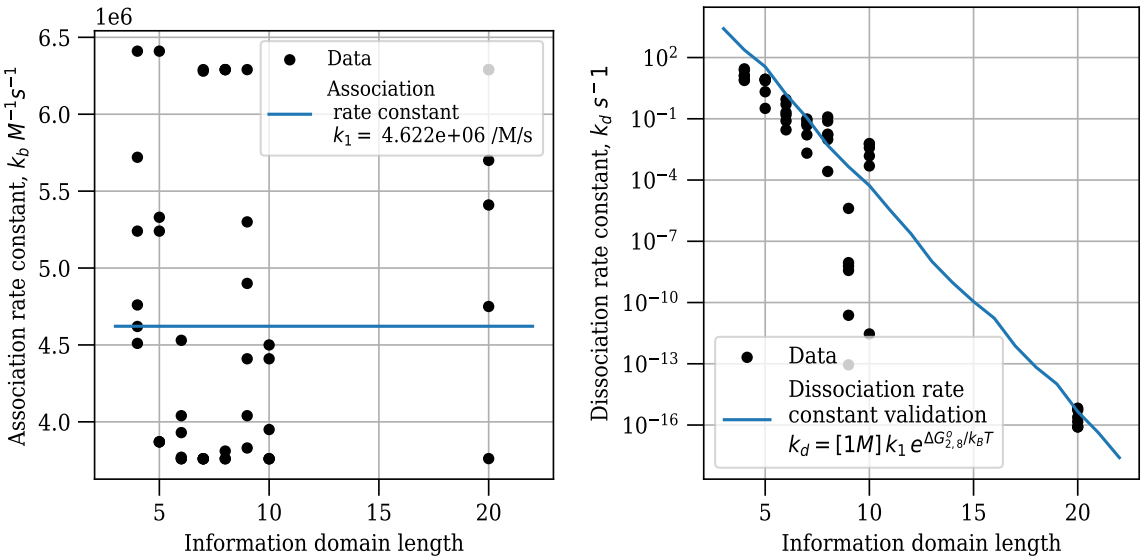


Figure 4.5: A correction function was used to ensure that the standard free energies of the transitions that form a loop structure $\Delta G^o_{4,10}$, $\Delta G^o_{5,11}$, $\Delta G^o_{6,12}$, and $\Delta G^o_{13,14}$ (see Fig. 4.2d,e,f,g) scale logarithmically with the length of the loop. Nupack's model predicts standard free energy changes scale logarithmically with loop length when there is no secondary structure in the single-stranded region opposite the loop (red dots and lines), but gives linear dependence when there is secondary structure in the strand opposite the loop (blue dots and lines). A correction function (black dot-dashed line) was calculated to correct this discrepancy by taking the difference between the log and linear fits and applying an offset such that the correction function was 0 at a backbone length of 20. The correction function is shown as the black dot-dashed line.



(a) The diffusion-limited binding rate, $k_1 = k_2$ was set as the mean (on a log scale) of the diffusion-limited binding rates measured in Refs. [21, 22] .

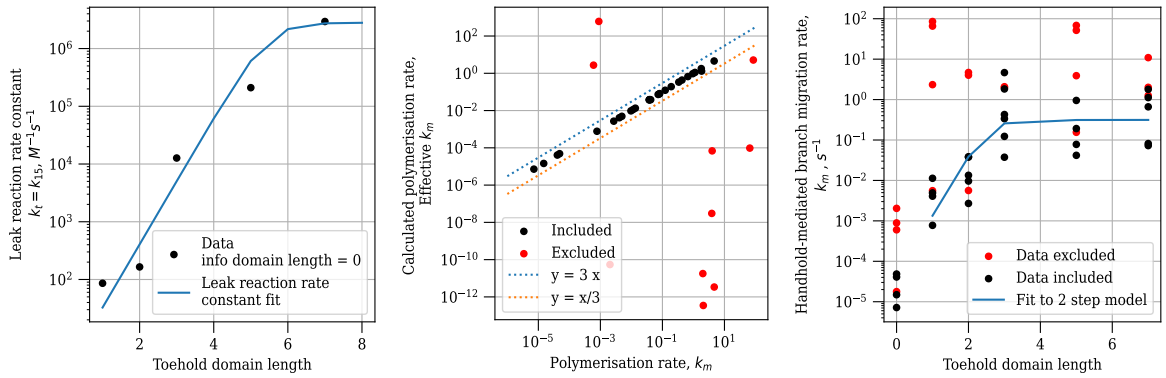
(b) The diffusion-limited binding rate k_2 was validated by comparing the dissociation rate measured experimentally to the rate predicted by the model $k_d \approx k_2 \exp -\Delta G^o_{2,8}/k_B T$. The model shows good agreement with the data without tuning $\Delta G^o_{2,8}$.

Figure 4.6: a)Determining and b) validating the second-order binding rate constant which was set to $k_1 = k_2 = 4.62 \times 10^6 M^-1 s^-1$.

4.2.7 Model parameterisation: Rate constant estimation

Here I describe the method by which I inferred the rate constants k_1 of bimolecular diffusion-limited binding, k_3 and k_{15} of second-order toehold-mediated strand displacement, and k_{13} of first order polymerisation or strand displacement mediated by the template from the data characterising the kinetics of handhold-mediated strand displacement presented by Cabello-Garcia in Refs [22, 21]. In these works, Cabello Garcia measured the kinetics of a handhold-mediated strand displacement process that closely resembles the transitions that take place on the final two sites of the template in the model presented here, with a backbone domain of length 24. In characterising the system at 25 °Cin 1M NaCl, Cabello-Garcia fitted the kinetic rate constants associated with monomer binding to the template, k_b, the rate constant of the leak pathway, k_t, the rate constant associated with monomer dissociation k_d, and the first order rate of polymerisation k_m. This data is presented in Table B.7 of the appendix in Cabello-Garcia's thesis [21].

The diffusion-limited rate constant associated with bimolecular binding k_1 was determined to be $k_1 = 4.62 \times 10^6 \mathrm{M}^{-1} s^{-1}$ by exponentiating the average of the logarithm of the binding constants $\ln(k_b)$ presented in Table B.7 of Ref [21]. k_1 is compared to the data points k_b in Fig. 4.6a. As validation, in Fig. 4.6b I compared Cabello-Garcia's dissociation rate constants k_d to end unbinding rate constants (with comparible end domain lengths) $k_8 = k_1 e^{\Delta G^o_{2,8}/k_B T}$ of Eqn. 4.7 and saw fair agreement.



(a) The toehold-mediated strand displacement model was calibrated by fitting an internal parameter to the rate of the leak reaction measured in Refs. [21, 22].

(b) Excluding inconsistent data points to fit the polymerisation rate k_{13}.

(c) Fitting the polymerisation rate k_{13} to the handhold-mediated branch migration rate measured in Refs. [21, 22]

Figure 4.7: (a) The leak reaction rate measured in Refs. [21, 22] is a strand displacement process, initiated by the toehold. The leak rate was used to set the branch migration rate $r_{bm,Irmisch}$, an internal parameter used in the strand displacement model developed by Irmisch [156]. The calibrated model was used to predict the rate constants k_3 and k_{15} with the appropriate toehold lengths. **(b)** The handhold-mediated strand displacement rates were each quoted twice in Refs. [21, 22]. I checked for consistency between the values using equation Eqn. 4.15. Values within a factor of 3 with each other (shown in black) were included for fitting the parameters of my model. Values different by more than a factor of 3 were excluded (shown in red). **(c)** Values of the branch migration rate that were within tolerance (same points in black) were used to fit the parameters of the polymerisation rate model k_{13} of Eqn. 4.16. Data for toeholds of length 0 were not included in the fit.

Rate constants k_3 and k_{15} are both associated with a toehold-mediated strand displacement reaction that may heal mismatches in the incumbent strand. I used the strand displacement rate model R_{Irmisch} developed by Irmisch et. al to estimate the rate constants k_3 and k_{15} [156]. The model treats the strand displacement process as discrete-state, continuous-time first-passage time problem over a discrete-state energy landscape shaped by mismatches. The branch migration rate, $k_{\text{bm, Irmisch}}$, affects the overall rate of stepping backwards and forwards through this landscape. At low concentrations, these strand displacement reactions are second-order in concentration. I assumed that the second-order rate constant in the direction that heals mismatches would be roughly the same if there were one or multiple healed mismatches. I also assumed that mismatches not healed by the process have no effect on the second-order rate constant of the process.

Irmisch's model takes the relevant toehold and branch-migration domain lengths and the branch migration rate, $r_{\text{bm, Irmisch}}$, as input (I left other parameters as their default values). The branch migration rate, $r_{\text{bm, Irmisch}}$, was determined by fitting the leak reaction rate $k_{15}(r_{\text{bm, Irmisch}}) = R_{\text{Irmisch}}(r_{\text{bm, Irmisch}})$ to the experimentally determined leak rate constants k_t (rate of the reaction in the absence of a handhold domain/information bearing domain in Ref. [21]), as shown in Fig. 4.7a. The fitting process gave $r_{\text{bm, Irmisch}} = 2.27 \times 10^4 s^{-1}$.

Having determined the rate constants k_1, k_3 and k_{15}, I fitted the polymerisation rate k_{13} to the branch

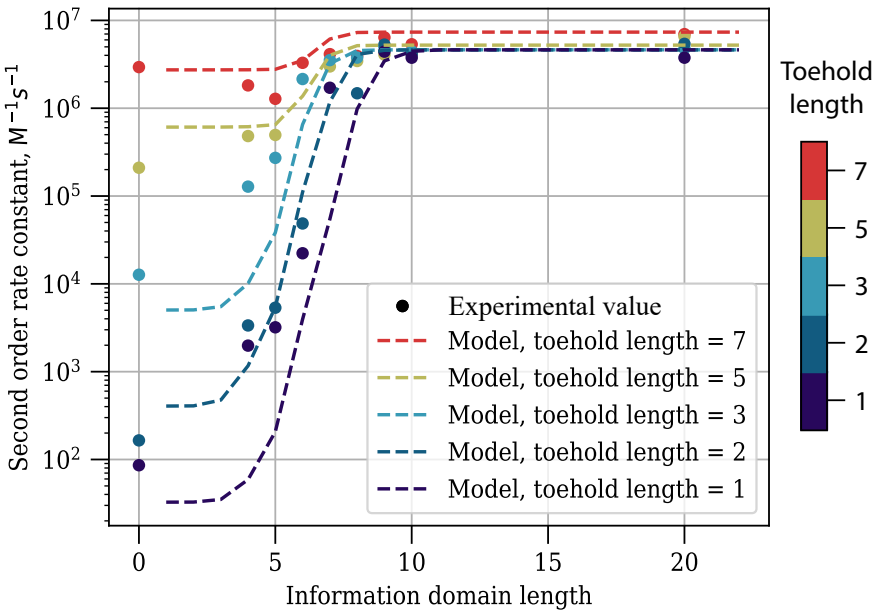


Figure 4.8: Having determined all of the rate constant parameters for the model using the data collected by Cabello-Garcia, I validated the model against the second-order rate constant using Eqn. 4.15 for a range of toehold and information domain lengths. The model shows a reasonable agreement with the experimental data, most often underestimating the second-order rate at lower information domain lengths.

migration polymerisation rate k_m presented in Table B.7 of the appendix in Cabello-Garcia's thesis [21] in two steps. The overall process of handhold-mediated strand displacement that Cabello-Garcia characterised was often well-modelled as a second-order process with a second-order rate constant given by

$$k_{2nd} = k_t + \frac{k_m k_b}{k_d + k_m}. \tag{4.15}$$

First, I checked the self-consistency of the values quoted in Table B.7 of the appendix in Cabello-Garcia's thesis [21], by plotting the quoted rate constant k_m against the rate constant k_m, as shown in Fig. 4.7b, calculated from the other quoted rate constants by inverting Eqn. 4.15. Some of the quoted second-order rate constant values shown in red lay far away from the calculated rate constant values and therefore they were excluded from the following step.

Next, I plotted the experimentally determined rate of polymerisation k_m against the toehold length, as shown in Fig. 4.7c, and noticed saturating behaviour at longer toehold lengths. This suggested that the polymerisation rate could be modelled as a two step process

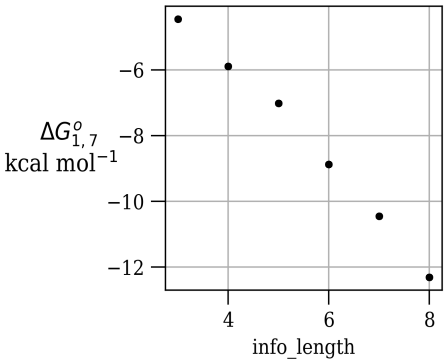
$$k_{13} = k_m = \left(k_{m,2}^{-1} + (k_{m,1} \exp(-\Delta G^o_{\text{th}}/k_B T))^{-1}\right)^{-1}, \tag{4.16}$$

where $k_{m,1}$ is a first-order rate constant associated with forming the toehold connection, ΔG^o_{th} is the standard free energy change associated with base-pairing with the toehold domain, $k_{m,2}$ is a first-order rate constant associated with the branch migration process that is rate limiting at long toehold lengths. I estimated the free energy ΔG^o_{th} using Nupack [153, 154, 155] and fitted the rate constants $k_{m,1}$, $k_{m,2}$ to the empirical values of k_m, as shown in Fig. 4.7c. The fit resulted in $k_{m,1} = 1.35 \times 10^{-5}\, s^{-1}$, $k_{m,2} = 0.314\, s^{-1}$.

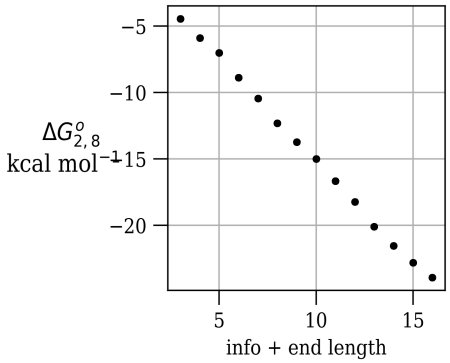
Having constrained all the rate constants in the model, I compared my model's prediction of the second-order reaction rate to the second-order rate constant quoted in Appendix Table B.7 in Cabello-Garcia's thesis [21], as shown in Fig. 4.8, and saw reasonable agreement, given the somewhat inconsistent patterns within the data itself.

4.2.8 Parameter sets span the design space

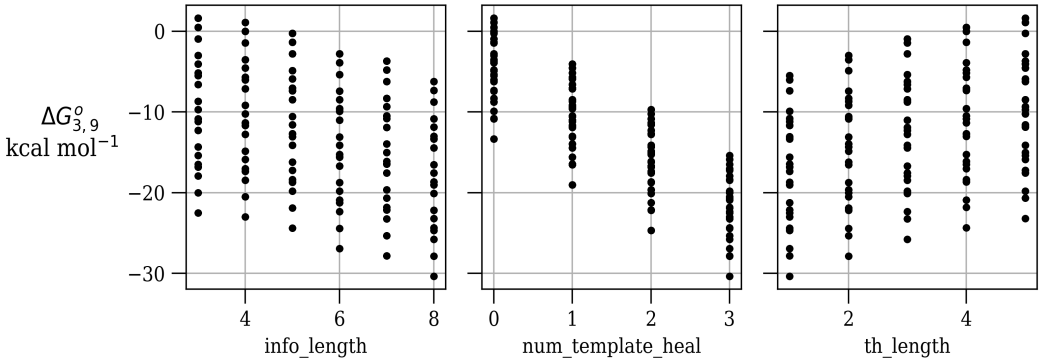
I generated a set of design parameters (domain lengths and number of mismatches) for the model that spanned the design-space. The effect these parameters have on the rate constants and free energies are shown in Figures 4.9 through 4.14. The information domain length took values in the range 4 through 9. The backbone domain length took values 20, 30 and 40. The toehold length took values between 1 and 5, and the end domain length took values in the range (0,1,2,4,6,10). The total number of mismatches in the system was varied in the range (0,1,3,4) and the number of mismatches healed when locked monomers bind to the template was in the range (0,1,2,3,4). The sum of the number of mismatches healed in polymerisation and locked monomer binding is equal to the total number of mismatches in the system. Hence if no mismatches were healed during lock displacement binding, then all the mismatches must have been healed during polymerisation.



(a) The standard free energy change associated with the binding of a polymer to the template from solution by an information-bearing domain, $\Delta G^o_{1,7}$, decreases approximately linearly with the length of the domain.

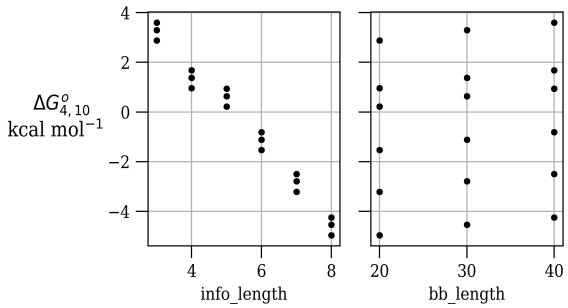


(b) The standard free energy change associated with the binding of a polymer to the template from solution by an end domain (with length info length + end length), $\Delta G^o_{2,8}$, decreases approximately linearly with the total length of the domain.

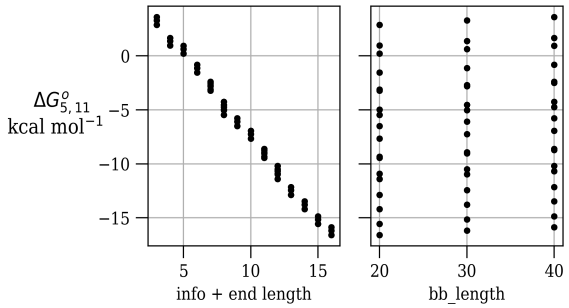


(c) The standard free energy change associated with the binding of a polymer to the template from solution with the displacement of the lock strand, $\Delta G^o_{3,9}$, decreases approximately linearly with the length of the information bearing domain and the toehold length, and most substantially with the number of C/C mismatches that are repaired in the strand displacement process (num_template_heal).

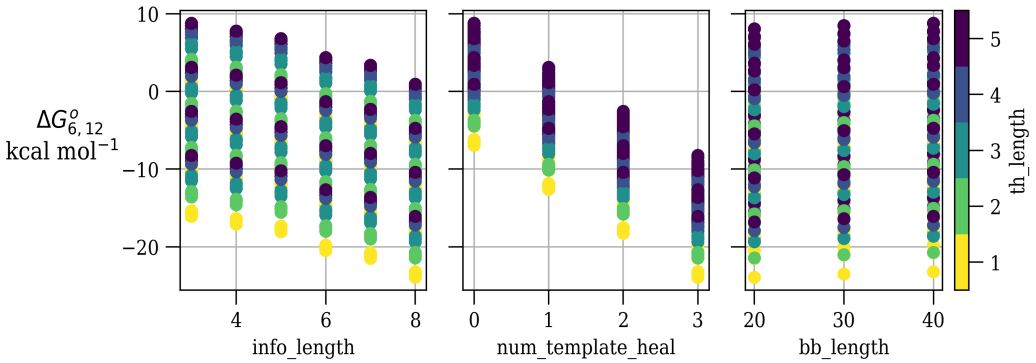
Figure 4.9: Plots of the calculated standard free energies against input parameters. Input parameters not shown have no affect on the dependent variables.



(a) The standard free energy change associated with the zipping (tail binding to the template) of a polymer tail to the template, $\Delta G^o_{4,10}$, decreases approximately linearly with the length of the information-bearing domain and increases with the backbone length due to the increasing entropic penalty associated with bulges [153, 154, 155, 147].

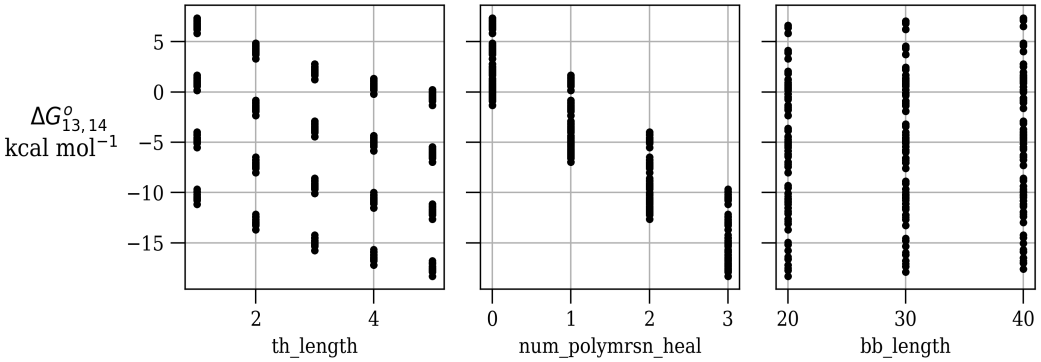


(b) The standard free energy change associated with the zipping (tail binding to the template) of a polymer tail at the end of the template, $\Delta G^o_{5,11}$, decreases with the length of the end domain (info length + end length) and increases slightly with the backbone length due to the associated entropic penalty of forming a loop.

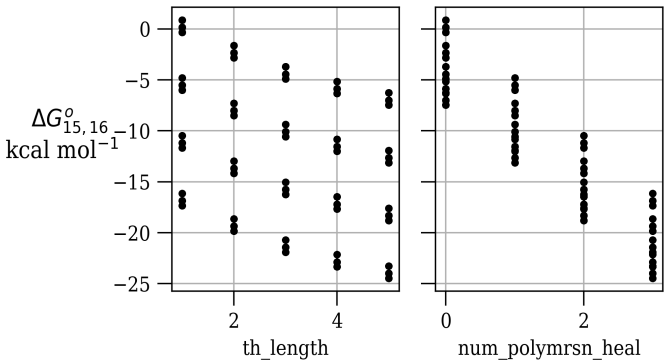


(c) The standard free energy change associated with the zipping (tail binding to the template) of the leading edge of a polymer to the template with lock displacement, $\Delta G^o_{6,12}$, decreases with the length of the information-bearing domain and increases slightly with the backbone length. The number of C/C mismatches healed in the strand displacement process imparts the largest component to the free energy change.

Figure 4.10: Plots of the calculated standard free energies against independent input variables. Input parameters not shown have no effect on the dependent variables.

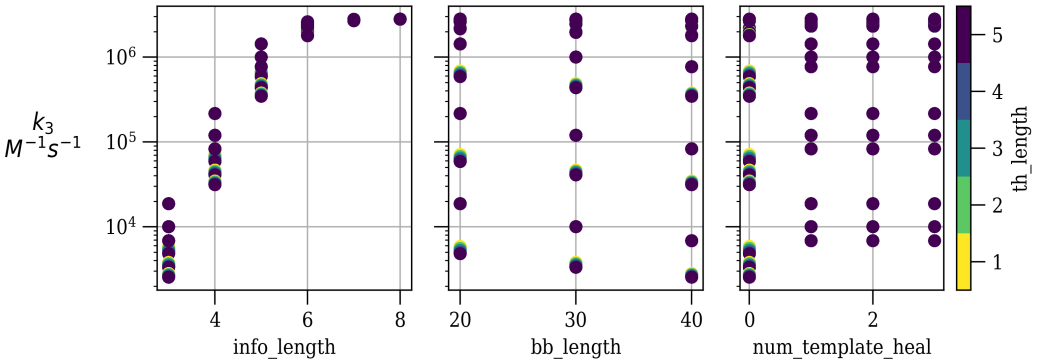


(a) The standard free energy change associated with polymerisation, $\Delta G^o_{13,14}$, decreases with the length of the toehold domain, increases with the backbone length, and decreases with the number of C/C mismatches repaired in the process.

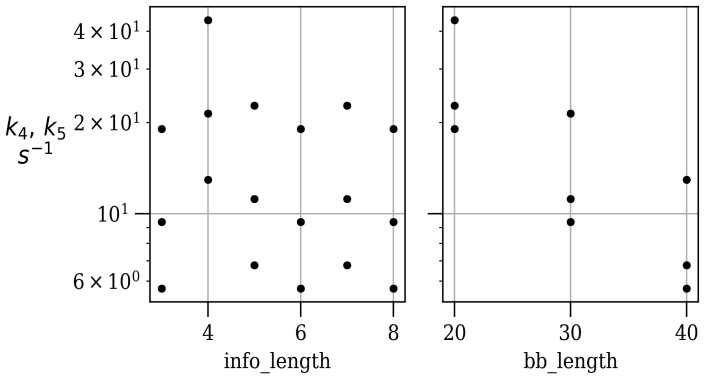


(b) The standard free energy change associated with the leak reaction, $\Delta G^o_{15,16}$, in which a polymer in solution undergoes toehold-mediated strand exchange (non-handhold/template mediated strand displacement polymerisation), decreases with the number of mismatches healed in the process (num_polymrsn_heal) and the toehold length .

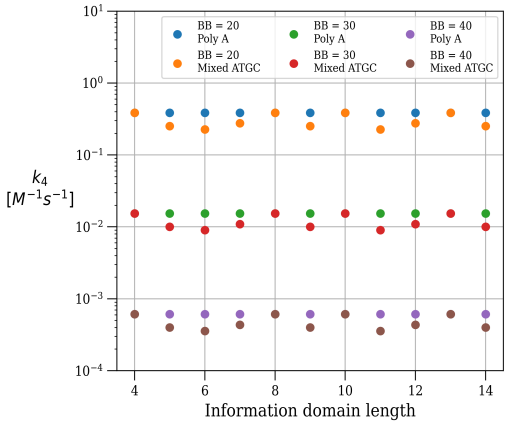
Figure 4.11: Plots of the calculated standard free energies against independent input variables. Input parameters not shown have no effect on the dependent variables.



(a) The second-order rate constant associated with polymers binding to the template from solution at the leading edge (with lock displacement), k_3, calculated using Irmisch's method [156]. k_3 is always less than the diffusion-limited binding constant, $k_1 \approx 4.6 \times 10^6 M^{-1}s^{-1}$. k_3 increases with the information domain length and whether a mismatch is present, and decreases slightly with the backbone domain length. The presence of additional mismatches was assumed not to change the rate substantially.

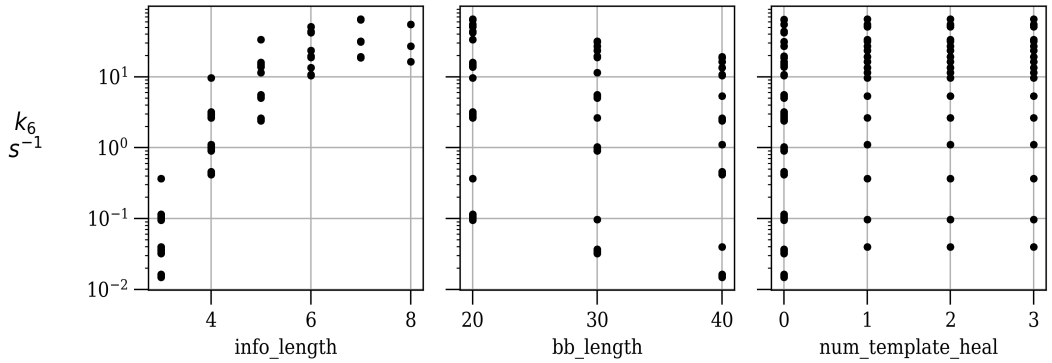


(b) The first order rate constant of polymers zipping to the template (without lock displacement), k_4, decreases with the backbone length. The toehold and information-bearing domain lengths make minor changes to the rate.

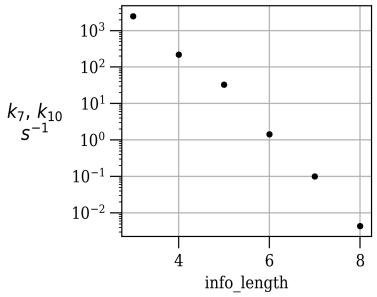


(c) The variation in the first order rate constant of polymers zipping to the template (without lock displacement), k_4 and k_5, can be attributed to the sequence of the information-bearing domain. A poly-A information domain shows no variation with the length of the domain, while a mixed sequence gives rise to slight variations in the rate constant.

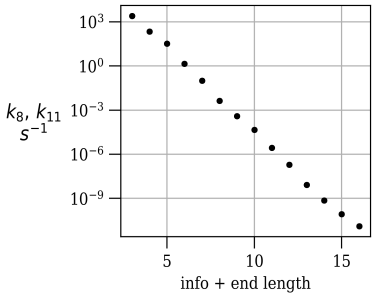
Figure 4.12: Plots of the calculated rate constants against independent input variables. Input parameters not shown have no effect on the dependent variables.



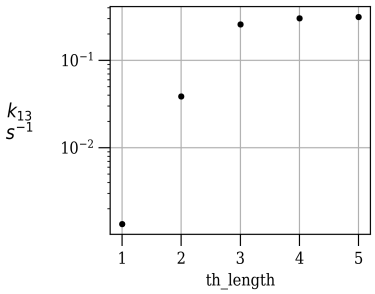
(a) The first order rate constant of polymers zipping to the leading edge of a polymer to the template with lock displacement, k_6, decreases with the backbone length, increases with the information bearing domain length and whether there is a C/C mismatch healed in the process.



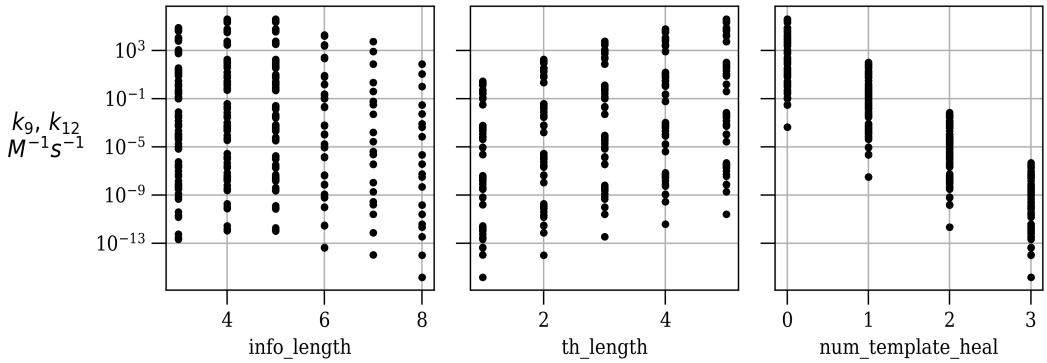
(b) The first order rate constant of a polymer unbinding the template from an information-bearing domain, k_7 and k_{10}, decreases only with the length of the domain itself.



(c) The first order rate constant of a polymer unbinding the template from the end of the template, k_8 and k_{11}, decreases with the total length of the domain.

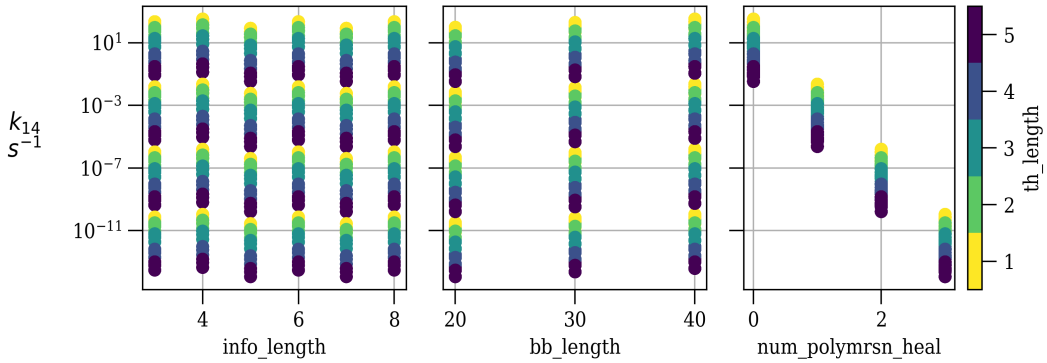


(d) The first order rate constant of polymerisation, k_{13}, increases with the toehold length and saturates above a length 3.

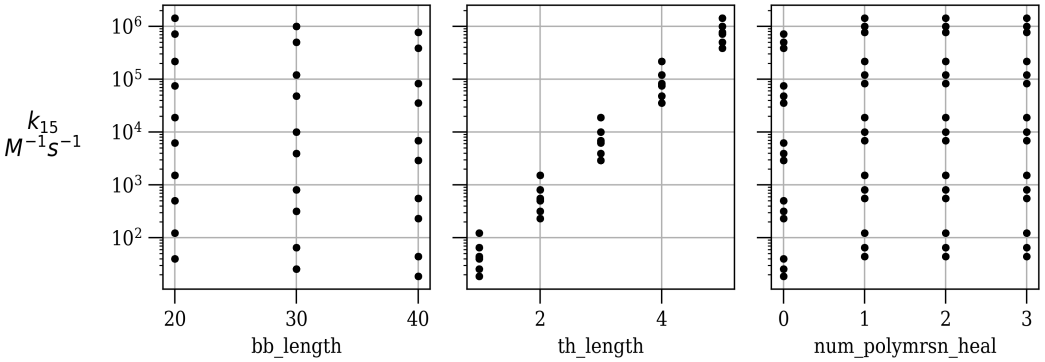


(e) The second-order rate constant of the leading-edge unbinding of a polymer from the template to solution via lock strand displacement, k_9 and k_{12}, decreases with the information bearing domain length, the number of C/C mismatches repaired in the process, and increases with the toehold domain length.

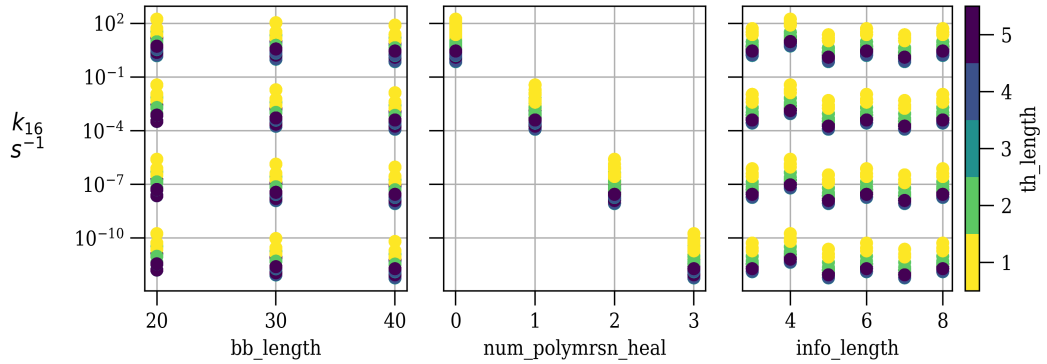
Figure 4.13: Plots of the calculated rate constants against independent input variables. Input parameters not shown have no effect on the dependent variables.



(a) The first order rate constant of depolymerisation, k_{14} increases with the backbone domain length and decreases with the number of C/C mismatches created in the process. The rate constant changes slightly with the toehold domain length and the information-bearing domain length.



(b) The second-order rate of the leak reaction in the forward direction k_{15} increases with the toehold length, whether a mismatch is healed in the process, and decreases with the backbone domain length.



(c) The second-order rate of the leak reaction in the reverse direction k_{16} increases with the backbone length and the number of C/C mismatches healed in the polymerisation process. The toehold length and the information-bearing domain length have a minor effect on the rate constant.

Figure 4.14: Plots of the calculated rate constants against independent input variables. Input parameters not shown have no effect on the dependent variables.

4.2.9 Solving the system of ODEs

Having constructed the model and parameterised it based upon previous experimental characterisation [22, 21], the set of differential equations in Eqn. 4.4 could be solved. The model was solved using the Radau method of the solve_ivp function from the python package Scipy.integrate [133]. The Radau method is designed to solve stiff differential equations. The system of differential equations were integrated from time $t = 10^{-3}s$ to $t = 10^{10}s = 317$ years on a variable step-size, logarithmic timescale to ensure that the short and very long-time dynamics were captured, and to ensure that the equilibrium state was reached. The unit of concentration was set to 1nM, and the second-order reaction rate constants were scaled appropriately to be in units per nM per second. A relative tolerance of 10^{-4} and an absolute tolerance of 10^{-18} were used, such that the simulations were numerically accurate to about four decimal places at 10^{-18}nM, which was required to prevent numerical errors over long timescales.

As mentioned in Section 4.2.8, I generated a data set that spanned the model's parameter space.

For each combination of these input parameters, I ran four slightly modified simulation. In the first simulation, all reaction pathways were permitted. In the second simulation, the leak reaction was suppressed by setting $k_{15} = k_{16} = 0$. I refer to simulations in which the leak reaction was suppressed as simulations "without leak". In the third simulation, the leak reaction was permitted, but the binding of dimers in solution to the template via any reaction pathway (including the leak reaction) was prohibited. Unbinding of dimers from the template was always permitted. I refer to simulations in which the dimers were prevented from rebinding to the template as "inert dimer" simulations. In the fourth simulation, both the leak reaction and the rebinding of dimers was prohibited. I refer to such simulations as "without leak and with inert dimers".

For each simulation, unless stated otherwise, the concentration of the template was 1nM and the concentration of monomers in solution at the start of the simulation was [M1] = [M2] = [M3] = 10 nM. Initially, there were no dimers or trimers in solution. The temperature of the system was set to 25 °C.

The parameter sweep involved generating and solving 19,440 individual simulations. In the following subsection, I describe the ideal properties of a trimeric template copying system, I devise some measures by which these properties can be quantified, and, based upon the measures, I develop metrics from which I generate a score. By changing the weighting of these metrics, I explore how changing the input parameters can lead to differing behaviour.

4.2.10 Performance

An ideal template copying system generates a high yield of copolymers that have exactly the same length and sequence as the template that encoded them and a low yield of copolymers of any other length or sequence. Furthermore, the ideal system produces accurate copies quickly, with minimal product inhibition (rebinding of product to the template or slowing of the reaction due to failure to release product), and leaves the template unchanged and reusable at the end of the reaction cycle.

In the model presented in this chapter, the sequence of the copolymers produced by the template is neglected, as described in section 4.2.1, as the binding of non-complementary monomers to the template is expected to be minimal, or at least much slower than the binding of complementary monomers, due to the two-step process involved in binding as seen in Fig.4.1b. and Fig.4.2c. Therefore, we expect inaccurate sequences generated by the system to arise from other processes such as non-template directed polymerisation (leak reaction) shown in Fig. 4.1h, or by the premature release and subsequent rebinding and growth of incomplete copolymers (e.g. a dimer released into solution rebinding at a different site). By comparing the results of simulations in which the

leak reaction is allowed or prohibited and in which dimer rebinding is allowed or prohibited, I can assess the prevalence of these undesirable reaction pathways.

The best-performing systems will have the following qualities:

1. High yield: The reaction will convert the maximum concentration of monomers to trimers.

2. High rate: The reaction will proceed quickly.

3. Minimal leak: The leak reaction will have minimal impact on the kinetics of the reaction. The reaction will proceed with the same yield, rate, and reaction order (first, second...) regardless of whether the leak reaction is permitted or not.

4. Tight control over the product length distribution: The system will generate polymers that match the length of the template. In our case we expect to observe a high ratio of trimers to dimers.

5. Minimal rebinding of incomplete polymers: To minimise another source of sequence inaccuracies, the system will show little early release of polymers and their subsequent rebinding and growth.

4.2.11 Measureables

I compared a number of measures for each of the qualities described above (most not shown). Some of the measures were based upon fitting n'th order saturating reaction kinetics to the trimer concentration timeseries [Y(t)]. The n'th order kinetics are the solution to the following differential equation, where k_n is the n'th order rate constant, $[T_0]$ = 1 nM is the template concentration, and [M] is the concentration of reactant, and [C] is the reaction yield,

$$\frac{d[\mathrm{Y}]}{dt} = k_n[\mathrm{T}_0][\mathrm{M}]^n = k_n[\mathrm{T}_0]([\mathrm{C}]-[\mathrm{Y}])^n \tag{4.17}$$

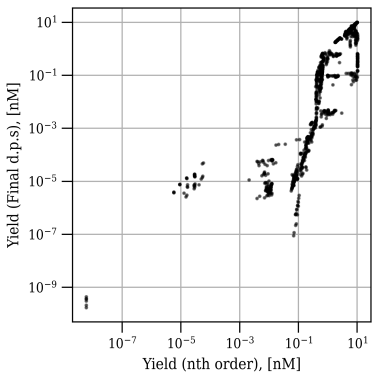
that yields a solution

$$[\mathrm{Y(t)}] = [\mathrm{C}] - \left([\mathrm{C}]^{1-n} + (n-1)\,k_n[\mathrm{T}_0]t\right)^{1/(1-n)}. \tag{4.18}$$

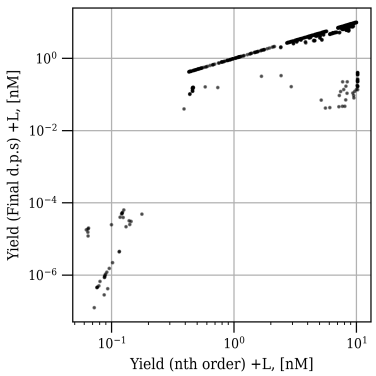
The time to half maximum concentration is given by

$$t_{1/2,n} = \frac{2^{n-1}-1}{(n-1)\,k_n[\mathrm{T}_0][\mathrm{C}]^{n-1}} \tag{4.19}$$

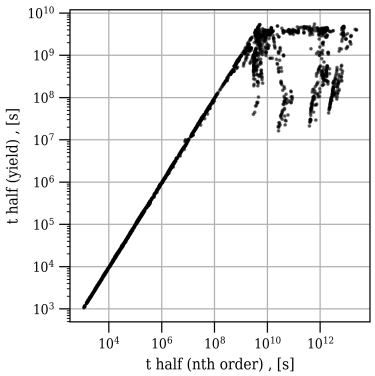
The measures I investigated were as follows:



(a) Cross comparison of two measures of trimer yield: Yield measured as the limiting concentration of a fit to n'th order reaction kinetics, and yield measured as the mean of the trimer concentration at the final 6 points. Both were measured in the model without the leak reaction. Higher is better.



(b) Cross comparison of two measures of trimer yield: Yield measured as the limiting concentration of a fit to n'th order reaction kinetics, and yield measured as the mean of the trimer concentration at the final 6 points. Both were measured in the model with the leak reaction. Higher is better.



(c) A cross comparison of $t_{1/2}$ measures, measured without the leak reaction pathway. Measured as the $t_{1/2}$ of a fit to n'th order kinetics and by interpolating the time at which the yield reached half of the maximum yield (measured at the last 6 data points). Lower is better.

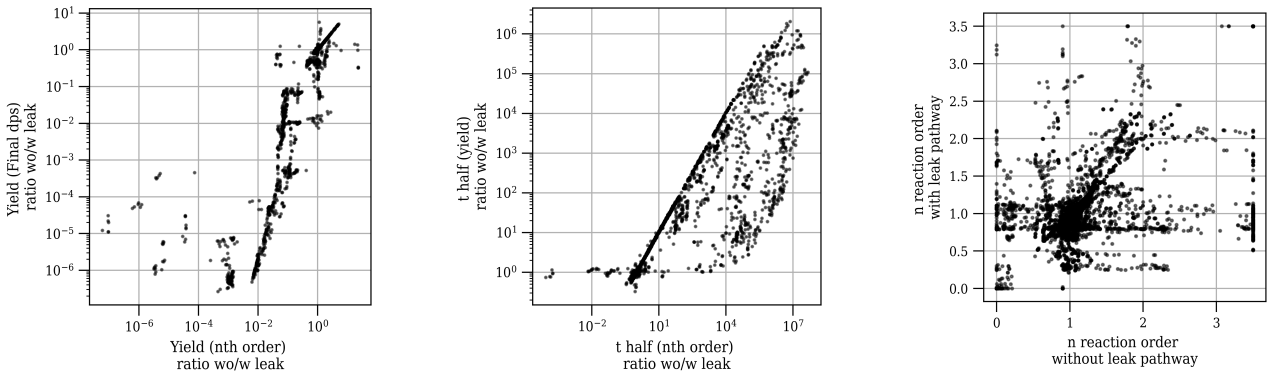
Figure 4.15: A comparison of measures of the reaction trimer yield and time to half trimer yield $t_{1/2}$.

Yield measures

The yield was measured with the leak reaction prohibited as the saturating concentration reached under a fit to n'th order reaction kinetics (Eqn. 4.18) or as the mean value of a spline fitted through the final 6 time points. In Fig. 4.15a these two measures are compared. For 186 simulations a fit to the data using Eqn. 4.18 could not be determined, because the reactions did not reach completion within the time window chosen. In the majority of cases, a good fit was attained, and the agreement between measures was good when the yield was high. The fit to the last 6 data points was chosen as the measure of yield, for the sake of simplicity and so the 186 simulations that failed to fit could be included in the analysis.

Reaction speed measures

The speed of the reaction was also measured under the model in which the leak reaction was prohibited. The speed of the reaction was measured based on 1) the fit to n'th order kinetics as the time to half the final concentration, which can be obtained from the fitted concentration, rate constant and reaction order using Eqn. 4.19 and 2) by interpolation of the time at which the trimer concentration reached half of the yield measured as the mean of the final 6 data points. For consistency and because of the failed fit to 186 reactions that did not reach saturation, the interpolated time to half yield was chosen as the measure of the reaction speed. The two measures are compared in Fig. 4.15c.



(a) A comparison of the ratio of yield without and with the leak reaction pathway, measured by the final n data points or by a fit to n'th order kinetics. Closer to 1 is better.

(b) A comparison of the ratio of $t_{1/2,-L}/t_{1/2,+L}$ measured by interpolation or by a fit to n'th order kinetics. Closer to 1 is better.

(c) A comparison of reaction order n resulting from a fit to n'th order kinetics without and with the leak. Closer to the line $y = x$ is better, indicating that the leak reaction doesn't change the reaction order.

Figure 4.16: A comparison of the ratios of kinetic measures measured without and with the leak reaction pathway.

Effect of the leak reaction on yield

The ratio of the trimer yield measured *without* the leak reaction pathway to the yield measured *with* the leak reaction pathway was chosen as the measure of the effect of leak on yield. Ideally, this measure would lie close to 1. This ratio, measured using the fit to the n'th order reaction kinetics and as the fit to the final data points are compared in Fig. 4.16a. The ratio measured by the final 6 data points was chosen for consistency.

Effect of the leak reaction on reaction timescale

The ratio of the reaction timescale measured *without* the leak reaction pathway to the yield measured *with* the leak reaction pathway was chosen as the measure of the effect of leak on speed. Ideally, this measure would also lie close to 1. This ratio, measured using the fit to the n'th order reaction kinetics and the method of interpolation are compared in Fig. 4.16b. The ratio using the method of interpolation was chosen as the measure of the effect of leak on the reaction timescale.

Effect of the leak reaction on the reaction kinetics order

The ratio of the reaction order measured *without* the leak reaction pathway to the yield measured *with* the leak reaction pathway was chosen as the measure of the effect of leak on speed. Ideally, this measure would also lie close to 1. This ratio could only be measured using the fit to the n'th order reaction kinetics and is shown in Fig. 4.15. Points lying close to the line passing through the origin with gradient 1 were ideal. Where the fit could not be determined (in 186 cases), the score attributed to this measure was set to 0 as described in more

detail in Section 4.2.12.

Trimer to dimer ratio measures

The ratio of trimers to dimers in solution toward the end of the reaction in the model with leak reactions permitted was chosen as the measure of length control. This was measured at the following points in time: 1) at a time $t = 10^8$ seconds, 2) at 20 times the time to half yield measured by the interpolation method, and 3) by the mean of the concentrations of trimers and dimers at the final 6 points in time. The ratio measured by the final 6 data points was chosen as the measure of length control.

Dimer release and rebinding

The ratio of trimers to dimers in the model in which leak reactions were permitted but dimer rebinding was prohibited (inert dimers) was chosen as the measure of dimer release and rebinding. If dimers are released from the template, they will not rebind to the template in this model and will stay in solution. Therefore, we can determine the cumulative release of dimers over the simulation. As dimers cannot rebind to the template, after a long period of time all of the mass of the system will turn to dimers, and this ratio will tend to 0, rendering the measure less informative. Therefore I chose to measure this quantity at the point in time at which the trimer concentration was greatest. Ideally, this ratio is high, indicating that few dimers were released over the timescale of the reaction.

4.2.12 Scoring metrics based upon the measures

For each measure described in section 4.2.11, I devised a simple scoring function. For each metric i, I calculated a raw score, $S_{raw,i} = f_i(m_i)$, using a function f that was chosen to flatten the distribution of measure samples $m_i \in M_i$. The raw score was centered by subtracting the mean of the measure distribution $\overline{f_i(m_i)}$ and standardised by dividing through by the standard deviation of the centered measure distribution $\sigma f_i(m_i) - \overline{f_i(m_i)}$, to arrive at a unitless score function S_i:

$$S_i \equiv \frac{f_i(M_i) - \overline{f_i(m_i)}}{\sigma\left(f(m_i) - \overline{f(m_i)}\right)} \tag{4.20}$$

Yield metric

For yields y, the yield metric was calculated as

$$S_{yield}(y) = \frac{y - \bar{y}}{\sigma\,(y - \bar{y})}, \tag{4.21}$$

and thus high yields y score highly.

The speed metric was calculated as

$$S_{t_{1/2}}(t_{1/2}) \equiv \frac{-\log_{10}(t_{1/2}/(1\text{s})) + \overline{\log_{10}(t_{1/2}/(1\text{s}))}}{\sigma\left(-\log_{10}(t_{1/2}/(1\text{s})) + \overline{\log_{10}(t_{1/2}/(1\text{s}))}\right)}, \tag{4.22}$$

where $t_{1/2}$ is the measure of time to half yield measured in seconds. Lower values of $t_{1/2}$ score higher.

The leak yield ratio metric was calculated as

$$S_{r_y}(M_{r_y}) \equiv \frac{M_{r_y} + \overline{M_{r_y}}}{\sigma\left(M_{r_y} + \overline{M_{r_y}}\right)}, \tag{4.23}$$

where $M_{r_y}(r_y) = 10^{-\left|\log_{10}(r_y)\right|}$ and $r_y = \frac{\text{yield without leak}}{\text{yield with leak}}$. Values of r_y closer to 1 score higher.

The leak timescale ratio metric was calculated as

$$S_{r_t}(M_{r_t}) \equiv \frac{M_{r_t} + \overline{M_{r_t}}}{\sigma\left(M_{r_t} + \overline{M_{r_t}}\right)}, \tag{4.24}$$

where $M_{r_t}(r_t) = 10^{-|\log_{10}(r_t)|}$ and $r_t = \frac{t_{1/2} \text{ without leak}}{t_{1/2} \text{ with leak}}$.Values of r_t closer to 1 score higher.

The leak reaction order ratio metric was calculated as

$$S_{r_o}(M_{r_o}) \equiv \frac{M_{r_o} + \overline{M_{r_o}}}{\sigma\left(M_{r_o} + \overline{M_{r_o}}\right)}, \tag{4.25}$$

where $M_{r_o}(r_o) = 10^{-|\log_{10}(r_o)|}$ and $r_o = \frac{n \text{ without leak}}{n \text{ with leak}}$ and n is the fitted reaction order. Where the reaction order could not be determined by fitting, a score of 0 (the mean score) was given. Values of r_o closer to 1 score higher.

The trimer-to-dimer ratio metric was calculated as

$$S_{r_{T:D}}(r_{T:D}) \equiv \frac{\log_{10}(r_{T:D}) - \overline{\log_{10}(r_{T:D})}}{\sigma\left(\log_{10}(r_{T:D}) - \overline{\log_{10}(r_{T:D})}\right)}, \tag{4.26}$$

where $r_{T:D}$ is the measure of the ratio of trimers to dimers at the endpoint of the reaction.

The trimer-to-dimer ratio with inert dimers metric was calculated as

$$S_{r_{T:D,ID}}(r_{T:D,ID}) \equiv \frac{\log_{10}(r_{T:D,ID}) - \overline{\log_{10}(r_{T:D,ID})}}{\sigma\left(\log_{10}(r_{T:D,ID}) - \overline{\log_{10}(r_{T:D,ID})}\right)}, \tag{4.27}$$

where $r_{T:D,ID}$ is the measure of the ratio of trimers to dimers at the time of maximum trimer concentration under the model in which dimers are prevented from rebinding (inert dimers).

The total score S is a weighted sum, with weights ω_i of the individual metrics i described above

$$S = \sum_i \omega_i S_i \,. \tag{4.28}$$

The raw scores, the standardised metrics and the total score distributions are shown in Fig. 4.17. In the following Section 4.3, I determine the parameters that contribute to optimising each metric independently.

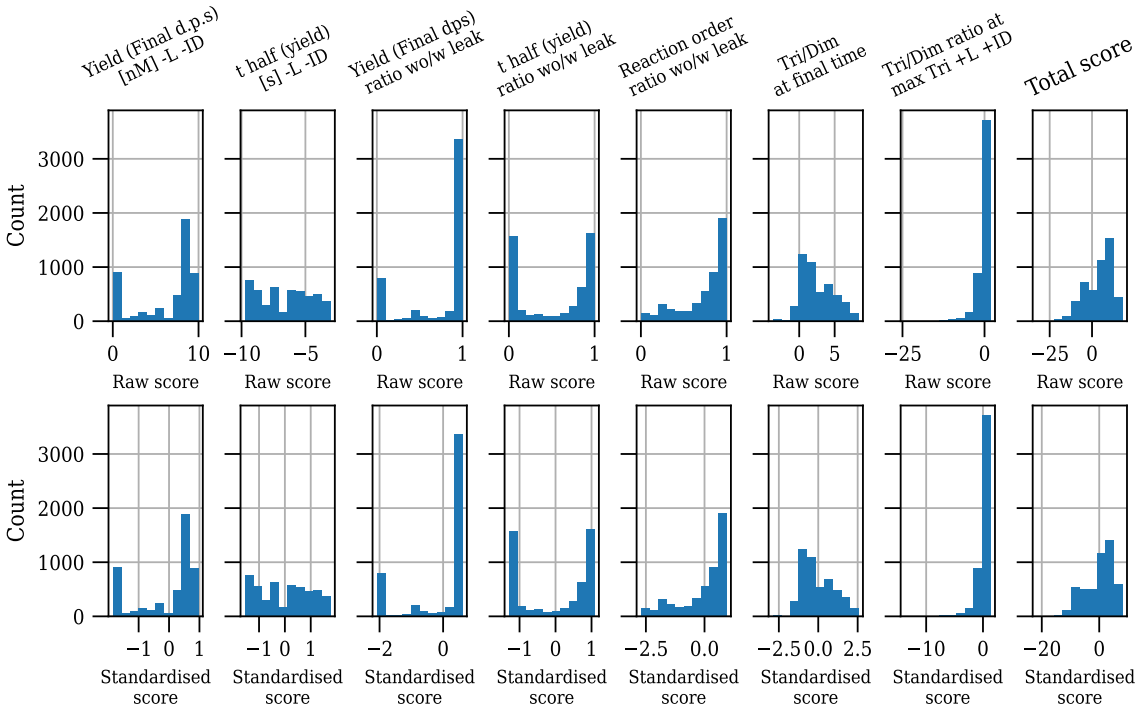


Figure 4.17: A score S_i was derived for each sample of each measurable M_i. The raw score $S_{raw,i}$ was a direct function of the measurable $f_i(M_i)$. The standardised score was calculated by subtracting the population mean of the distribution and dividing through by the population centered standard deviation as the $S_i \equiv \frac{f_i(M_i) - \overline{f_i(m_i)}}{\sigma\left(f(m_i) - \overline{f(m_i)}\right)}$.

4.3 Results: Optimising for independent metrics

In this section, I analyse how the input parameters contribute to optimising each of the metrics defined in section 4.2.12. In each case, I identify the top m performing parameter values for the metric in question and calculate their distribution over variables such as the domain lengths, the number of mismatches, the standard free energies, and the rate constants. If a variable has a large effect on the metric in question, the distributions of the variable in the cohort of top m performers and the whole ensemble should be quite different. I quantify this difference by calculating the relative entropy or Kullbach-Leibler Divergence (KLD) between the distribution over the top cohort P and the ensemble Q. For samples of the metric m in the space M, the KLD is given by

$$D_{KL}(P||Q) = \sum_{m \in M} P(m) \log \frac{P(m)}{Q(m)} \tag{4.29}$$

A high $D_{KL}(P||Q)$ indicates that the distributions are different, and hence the metric in question is sensitive to the variable over which the distribution is measured.

4.3.1 Optimising Yield

I begin by analysing the parameters that maximise the yield metric in the absence of the leak reaction by applying a weighting function $\omega = (1,0,0,0,0,0,0)$. The distributions of scores across the input parameters are shown in Fig. 4.18, across the standard free energies in Fig. 4.19. Plots of the concentration of chemical species in solution as a function of time for two parameter sets which had high trimer yields are shown in Fig. 4.21.

We can see in Fig. 4.18, that the yield is maximised (close to 10 nM) for systems with short information and end domain lengths and a large number of mismatches healed in the polymerisation reaction. As shown in Fig. 4.19, in this region of the parameter space, the affinity between copies and the template is very low as $\Delta G^o_{1,7} = \Delta G^o_{2,8}$ and $\Delta G^o_{4,10} = \Delta G^o_{5,11}$ are all very high, and the polymerisation reaction is effectively irreversible, $\Delta G^o_{13,14} \ll 0$. The unbinding rate constants that do not involve lock displacement k_7, k_8, k_{10}, k_{11} are therefore all high and the depolymerisation reaction is very slow k_{14}. Therefore, the system irreversibly polymerises monomers that quickly detach from the template. Note that we can see evidence of the rebinding of dimers and their subsequent conversion into trimers in the two simulations that generated the highest yield in Fig. 4.21. Eventually, all monomers are polymerised into trimers, and the affinity of the trimers for the template is so low that most of the trimers are not template bound, but are free in solution.

The yield of trimers can be constrained by the thermodynamics of the system at equilibrium. At equilibrium, the standard free energy of polymerisation ΔG^o_{pol}, the concentration of trimers in solution (yield) [Y] (assuming

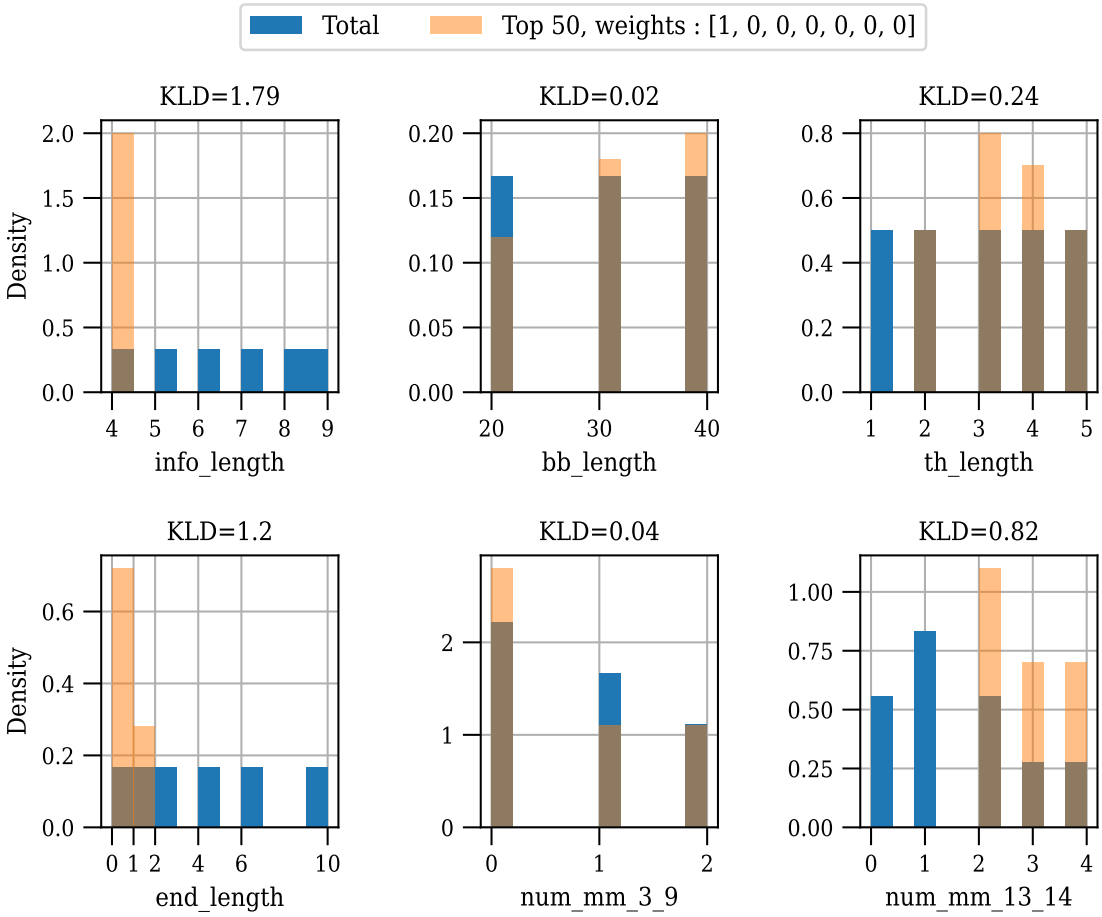


Figure 4.18: The input parameter distributions that optimised the trimer yield under a weighting function, $w = (1,0,0,0,0,0,0)$ (yield weight $= 1$ and all other weights set to 0) shown in orange compared with the whole ensemble distributions in blue. The top 50 parameter sets featured the shortest information domain lengths, short-end domain lengths, two or more mismatches healed in the polymerisation reaction, and a toehold greater than 1. These parameters indicate that the yield is maximised by systems in which the polymerisation process is irreversibly driven and the affinity between the formed trimers and the template is minimal.

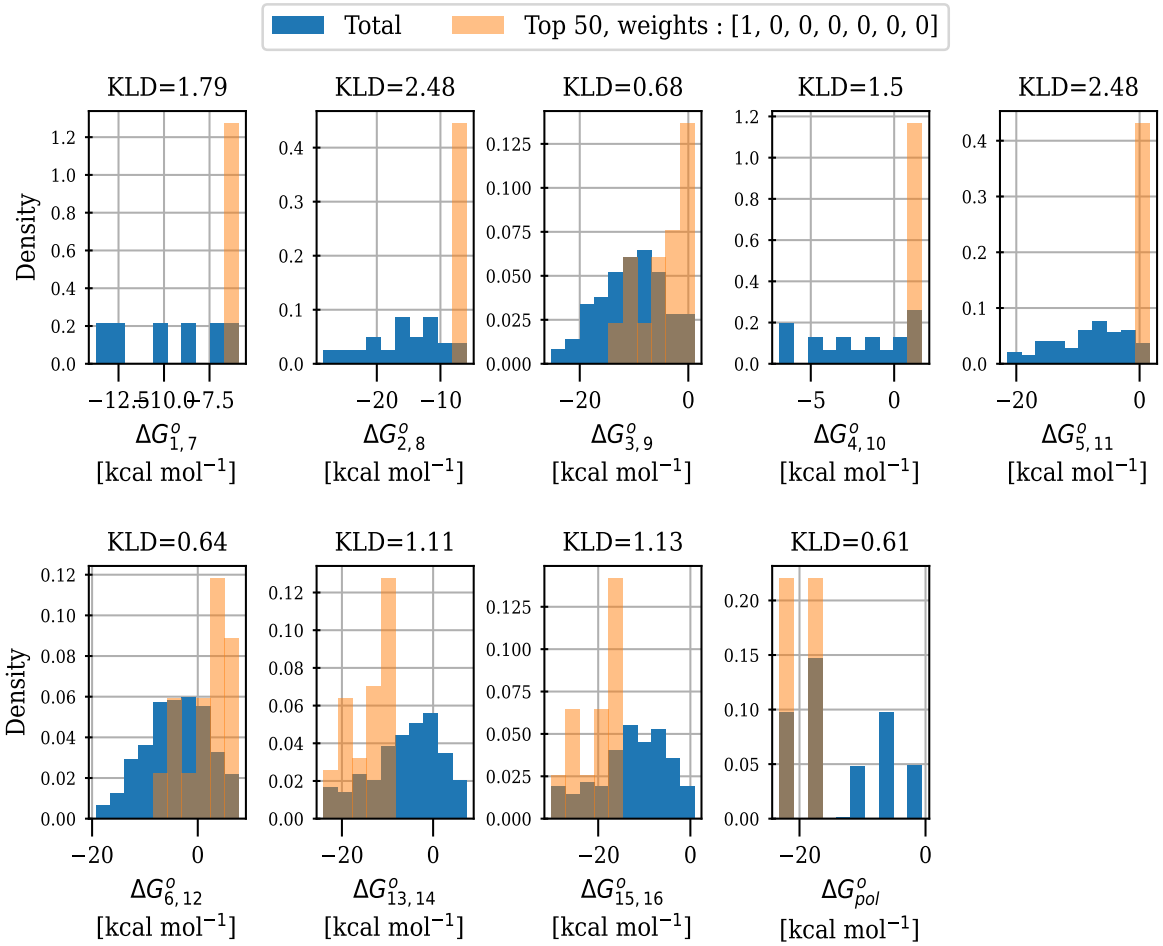
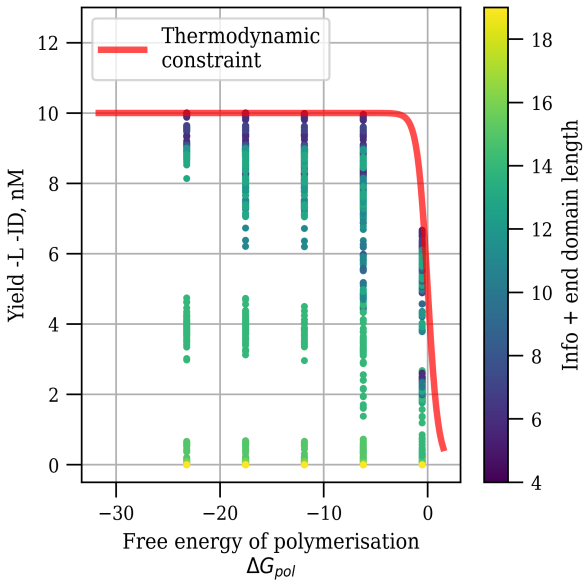
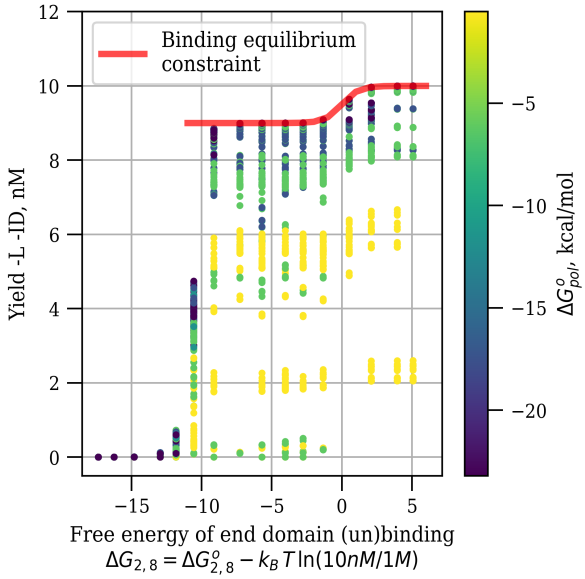


Figure 4.19: The same data that is presented in 4.18, except here I plot the standard free energy distributions ΔG^o, not the rate constants, that optimised the trimer yield under a weighting function, $w = (1,0,0,0,0,0,0)$ (yield weight $= 1$ and all other weights set to 0), shown in orange compared with the whole ensemble distributions in blue. The top 50 parameter sets had very weak binding-related standard free energies $\Delta G^o_{1,7}, \Delta G^o_{2,8}, \Delta G^o_{4,10}, \Delta G^o_{5,11}$ and very strong polymerisation-related standard free energies $\Delta G^o_{13,14}, \Delta G^o_{15,16}, \Delta G^o_{\text{pol}}$, indicating that trimers are thermodynamically favoured and the equilibrium occupancy of the template is low.



(a) The free energy of forming a trimer, given in Eqn. 4.32, places an upper bound on the trimer yield. It is necessary that one or more C/C mismatches were healed in the polymerisation process to reach the highest yield.



(b) The equilibrium concentration of trimers in solution, given by Eqn.4.36, places an upper bound on the trimer yield. The retention of trimers on the template increases with the length of the information-bearing domains. If the free energy change associated with the binding of a trimer to the template is positive, the maximal yield of 10nM can be reached. Otherwise, if $\Delta G^o_{2,8} \ll 0$, at least 1nM worth of trimer is bound to the template in equilibrium.

Figure 4.20: The maximum trimer yield is limited by **(a)** a thermodynamic constraint on the free energy change of polymerisation and **(b)** the binding equilibrium between trimers and the template.

no trimers are left bound to the template - which I will challenge shortly), the concentration of monomers [M1]=[M2]=[M3]=[M] and the concentrations of lock strands [L1]=[L2]=[L]=10nM, satisfy

$$0 = 2\Delta G^o_{\text{pol}} + k_B T \ln([\text{Y}]/1\text{M}) + 2 k_B T \ln([\text{L}]/1\text{M}) - 3 k_B T \ln([\text{M}]/1\text{M}) \,, \tag{4.30}$$

assuming no dimers are formed in equilibrium. The conservation of the mass of species implies

$$[\text{Y}] + [\text{M}] = [\text{M}_0] \,, \tag{4.31}$$

where $[\text{M}_0]$ is the initial concentration of each monomer. I solve for the trimer yield [Y] using Eqn. 4.30 and Eqn. 4.31 to get

$$[\text{Y}] = [\text{M}_0] + \left(\frac{[\text{M}_0]\lambda}{2}\right)^{1/3} \left([\text{L}] e^{\Delta G^o_{\text{pol}}/k_B T}\right)^{2/3} - \left(\frac{2}{[\text{M}_0]\lambda}\right)^{1/3} \left(l e^{\Delta G^o_{\text{pol}}/k_B T}\right)^{4/3} /3 \,, \tag{4.32}$$

where

$$\lambda = \sqrt{\frac{4}{3^3}\left(\frac{[\text{L}] e^{\Delta G^o_{\text{pol}}/k_B T}}{[\text{M}_0]}\right)^2 + 1} - 1 \,. \tag{4.33}$$

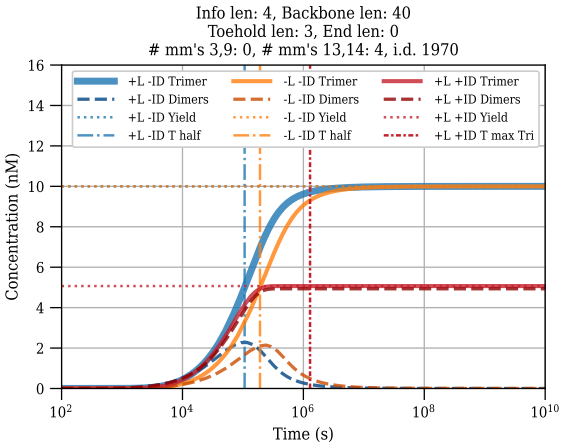
Equation 4.32 can be approximated as

$$[\text{Y}] \approx \frac{[\text{M}_0]}{1 + \left(\frac{[\text{L}]}{[\text{M}_0]} e^{\Delta G^o_{\text{pol}}/k_B T}\right)^{2/3}} \,, \tag{4.34}$$

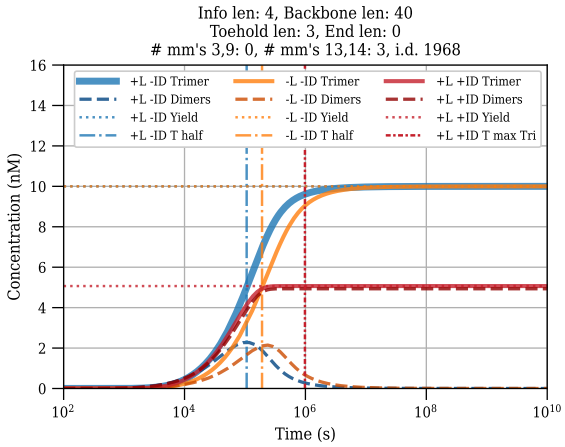
from which we can easily see that increasing the free energy of the polymerisation reaction, will drive the conversion of monomers to trimers. Equation 4.32 is shown as a red line in Fig. 4.20a that clearly constrains the maximum yield achievable. This constraint indicates that just a single C/C mismatch is thermodynamically necessary to achieve a yield close to 10nM. The key benefit of additional mismatches may be to help to prevent unwanted reaction pathways from dominating the dynamics.

If the system is strongly driven such that a high yield of trimers is thermodynamically accessible, we can place an additional constraint on the yield of free trimers in solution $[\text{Y}_{\text{free}}]$ by analysing the equilibrium between trimers and the final site on the template. I assume that no chemical species other than trimers exist in solution and that trimers bind to the template predominantly by the end domain. The equilibrium between template-bound trimers $[\text{Y}_{\text{bound}}]$ and unbound templates [T] and trimers $[\text{Y}_{\text{free}}]$ satisfies

$$0 = \Delta G^o_{2,8} + k_B T \ln[\text{Y}_{\text{bound}}] - k_B T \ln[\text{Y}_{\text{free}}] - k_B T \ln[\text{T}] \,. \tag{4.35}$$



(a) Timeseries with highest trimer yield.



(b) Timeseries with second highest trimer yield.

Figure 4.21: Timeseries of the concentrations of chemical species in solution (trimers - solid lines, total concentration of dimers - dashed lines) for simulations with the leak pathway (tag: +L, -ID, blue lines), without the leak pathway (tag: -L, -ID, yellow lines), and with the leak pathway and "inert dimers" (tag: +L, +ID, red lines). The trimer yield is shown by a dotted horizontal line and the time to half maximum yield $t_{1/2}$ is shown as a vertical dot-dashed line. The input parameters (domain lengths etc.) are given in the title of the plot. The timeseries that reached the highest yields both had very short information and end domain lengths and a large number of mismatches that were healed in the polymerisation step, driving the conversion of monomers to dimers and minimising the retention of trimers on the template at equilibrium. The process by which the trimers were formed involved the release, rebinding and subsequent polymerisation of dimers, as can be seen by the peak in the dashed lines around $t = 10^5 s$, and the lower trimer equilibrium reached in the model where dimer rebinding was prohibited (red lines).

Given that the total concentration of the template is conserved $[\mathrm{T}_0] = [\mathrm{Y}_{\mathrm{bound}}] + [\mathrm{T}]$ and that the total amount of free trimer is only slightly less than the total amount of trimer in the system $[\mathrm{Y}_{\mathrm{free}}] = [\mathrm{Y}] - [\mathrm{Y}_{\mathrm{bound}}] \approx [\mathrm{Y}]$ due to the low template concentration (which is only approximately true in our case), I can solve for the total amount of free trimer in solution,

$$[\mathrm{Y}_{\mathrm{free}}] = [\mathrm{Y}] - \frac{[\mathrm{T}_0]}{1 + \exp(\Delta G^o_{2,8}/k_B T)/[\mathrm{Y}]}\,. \tag{4.36}$$

Equation 4.36 is shown as a red line in Fig. 4.20b, which neatly constrains the maximum yield of trimers in solution. The effects of product retention can be mitigated by increasing the monomer concentration to increase the trimer concentration Y_{T}, reducing the length of the end domain to reduce $\Delta G^o_{2,8}$, or by reducing the amount of template in the system [T]. As the affinity between trimers and the template increases, a greater fraction of the trimers bind to the template, resulting in yields between 7nM (three trimers stuck to the template) and 9nM (a single trimer stuck to the template).

4.3.2 Optimising reaction speed

Here I analyse the factors that contribute to increasing the overall speed of the trimerisation reaction in the absence of the leak reaction pathway. The fastest reactions were identified by setting the weight of the timescale metric to 1 and all other weights to 0. As we can see in Fig. 4.22, the top 50 parameter sets feature long toeholds that speed up the rate of polymerisation (see k_{13} in Fig. 4.23), a short end domain and an intermediate information domain length (that keeps monomers bound to the template long enough to polymerise together, but not so strongly that the rate of detachment becomes slow, as we shall see in the following model), and an even distribution of mismatches healed in the monomer binding reaction 3,9 and the polymerisation reaction 13,14, making both reactions effectively irreversible. From the timeseries in Fig. 4.24 (yellow), we can see that the system produces polymers rapidly with a $t_{1/2} \approx 10^3 s$, the trimers reach a high yield, but this is achieved through the release, rebinding and subsequent polymerisation of dimers into trimers.

In Fig. 4.25a, I show that a simple model, derived below, upper bounds the observed reaction rate and explains the intermediate information domain length. I begin by asserting that the rate of trimer formation is limited either by $R_{\text{pol end}}$, the rate at which the final monomer, M3, binds and polymerises at the end of the template; or by $R_{\text{displaced unzipping}}$, the rate at which a monomer or dimer binds under the tail of an unzipping polymer; or by k_8 the rate at which a trimer unbinds from the end domain.

$R_{\text{pol end}}$, the overall rate (per molar template) at which the final monomer, M3, binds and polymerises at the end of the template is determined by the product of the rate of binding $k_2\,[\text{M}]$ and the probability that the bound monomer polymerises rather than unbinds $p_{\text{pol}} = \frac{k_{13}}{k_8+k_{13}}$, and the probability that once polymerised, the polymer does not depolymerise, $p_{\text{no depol}} = \frac{k_{13}}{k_{14}+k_{13}}$,

$$R_{\text{pol end}} = k_2\,[\text{M}]\frac{k_{13}}{k_8+k_{13}}\frac{k_{13}}{k_{14}+k_{13}}\,. \tag{4.37}$$

The rate at which a unit binds underneath an unzipping polymer $R_{\text{displaced unzipping}}$ is approximated similarly, by taking the product of the rate of monomer binding to a site $k_3\,[\text{M}]$, and the probability that the site is not occupied by the tail of a polymer, $\frac{k_{10}}{k_4+k_{10}}$, to get

$$R_{\text{displaced unzipping}} = \frac{k_3\,[\text{M}]\,k_{10}}{k_4+k_{10}} \tag{4.38}$$

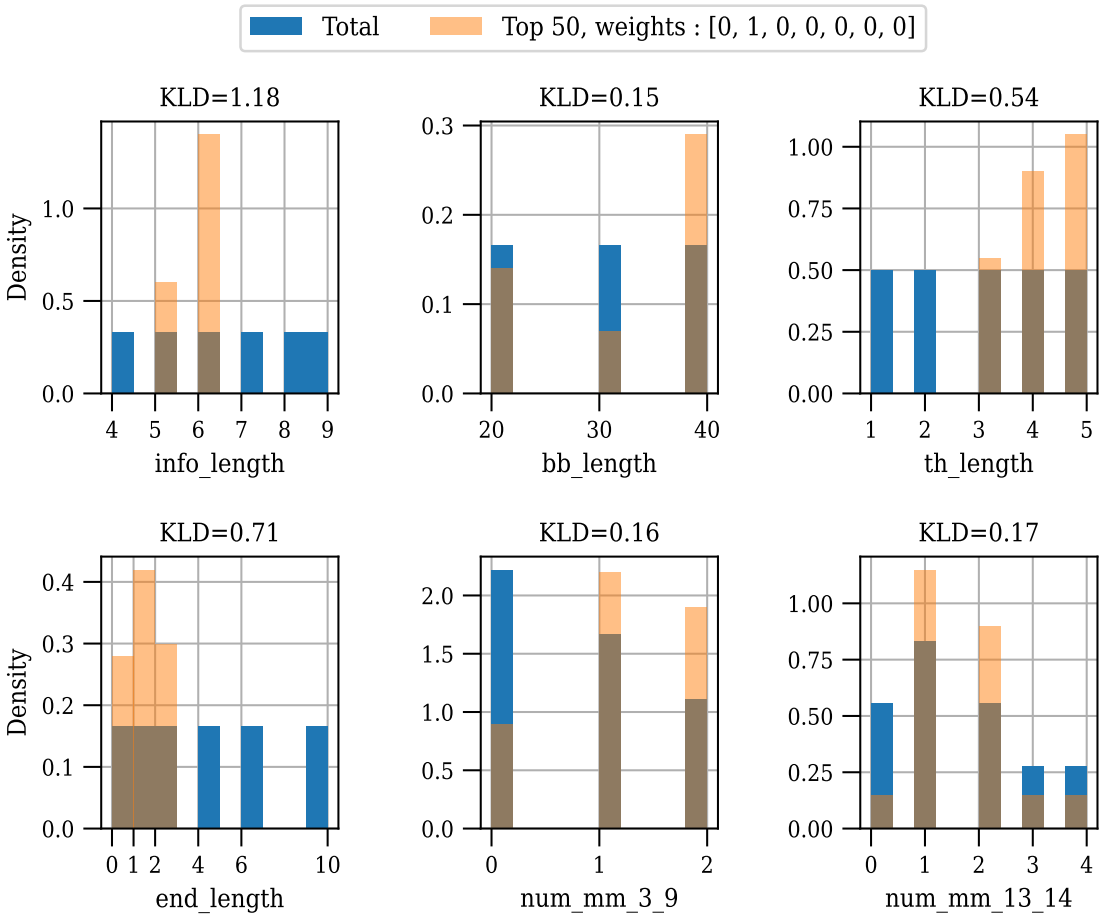


Figure 4.22: The input parameter distributions that minimised the time to half maximum yield $t_{1/2}$ (maximised the reaction speed) shown in orange compared with the whole ensemble distributions in blue. Here the weighting function used was $w = (0,1,0,0,0,0,0)$. The top 50 parameter sets had information domains between 5 and 6, end domains of 2 or less, and toeholds of 3 or greater. The information domain and toeholds are long enough that the toehold-mediated strand displacement reaction k_3 and the polymerisation reaction k_{13} are both saturated as shown in Fig.4.23. The end domain and information domain together are not so long that the process is slowed down by the unbinding of trimers from the end of the template. A single mismatch in both the monomer binding and polymerisation processes (3,9 and 13,14) is sufficient to ensure irreversibility. Slightly longer backbones reduce the tail zipping rate, increasing the speed at which polymers unbind the template.

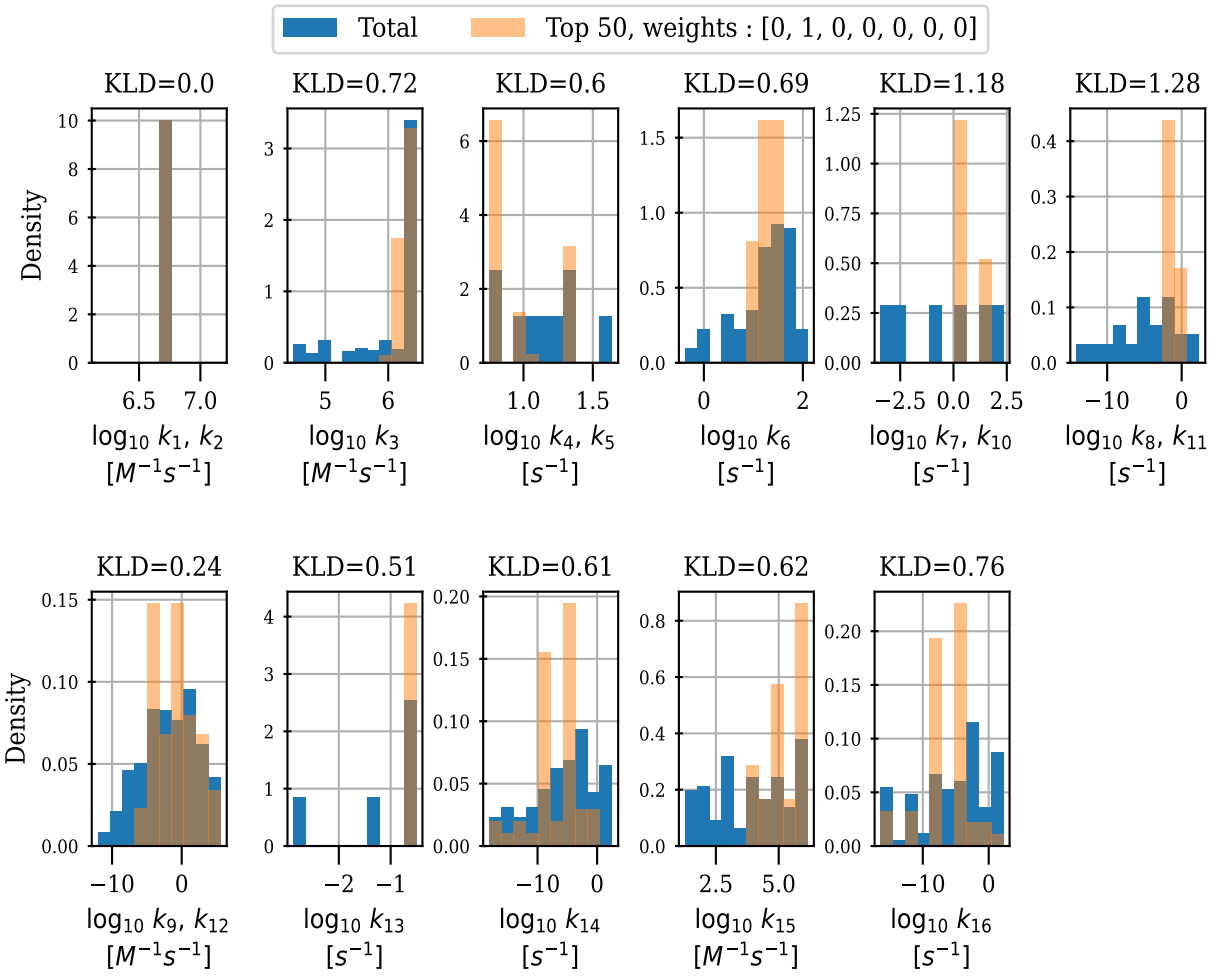
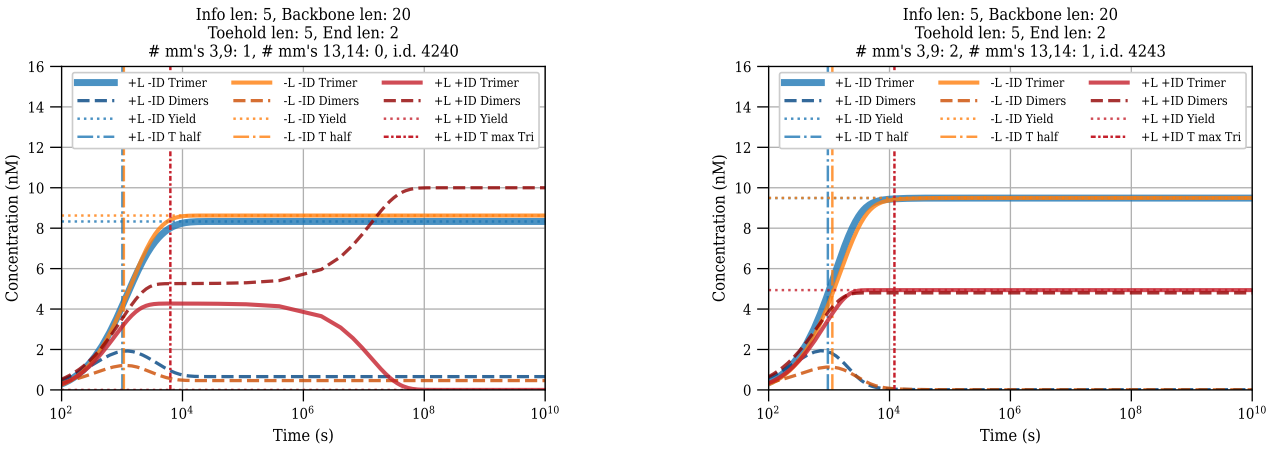


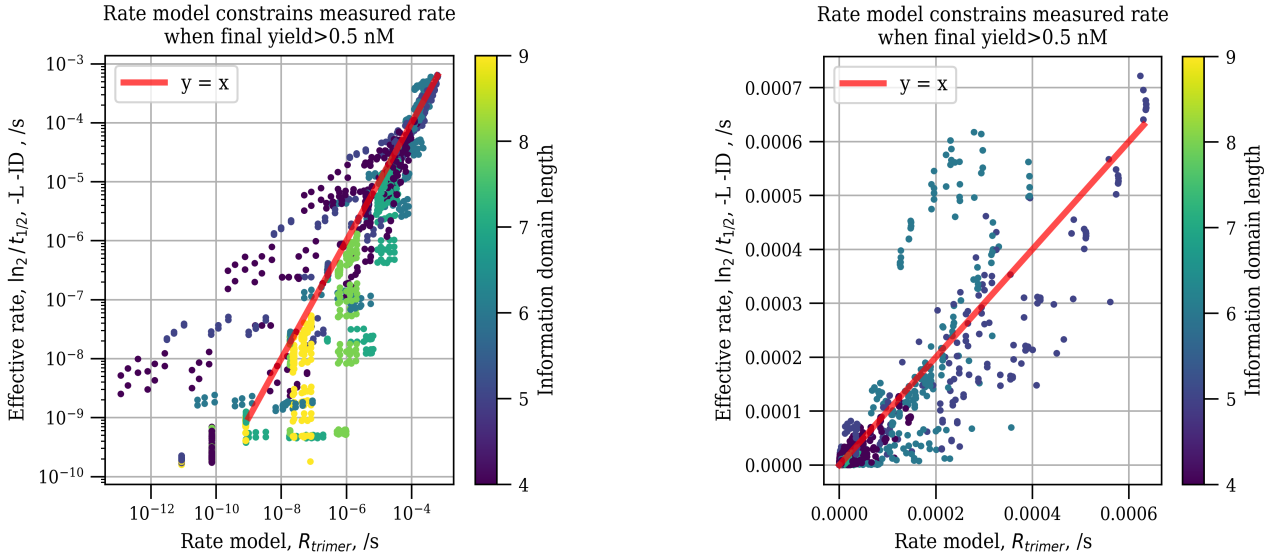
Figure 4.23: The same data as in Fig. 4.22, but instead showing the rate constant distributions that minimised the time to half maximum yield $t_{1/2}$ (maximised the reaction speed) in orange compared with the whole ensemble distributions in blue. Here the weighting function used was $w = (0,1,0,0,0,0,0)$. The top 50 parameter sets had information domains between 5 and 6, end domains of 2 or less, and toeholds of 3 or greater. The toehold-mediated strand displacement binding reaction rate constant k_3 and the polymerisation reaction rate constant k_{13} are both saturated. The tail and end unbinding and unzipping rate constants k_7,k_8,k_{10},k_{11} are both fairly high and the depolymerisation process is quite slow k_{14}. Together, this points to a reaction kinetics in which the binding of monomers and the polymerisation process are both fast and balanced by the unbinding reaction rates.



(a) Parameter set with shortest $t_{1/2}$.

(b) Parameter set with second shortest $t_{1/2}$.

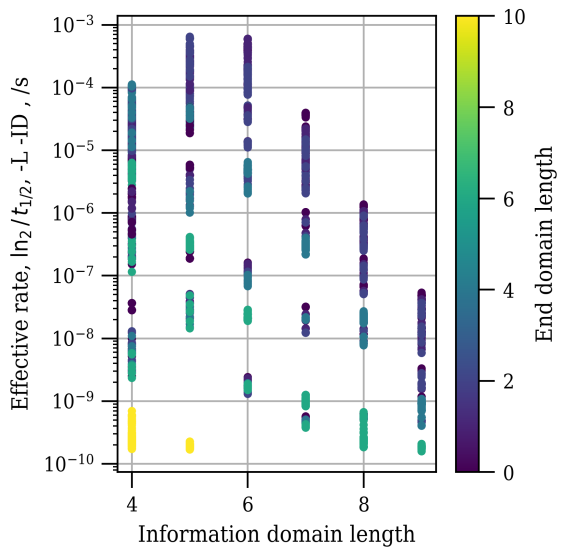
Figure 4.24: The two parameter sets with the shortest $t_{1/2}$ reached equilibrium in roughly 10^4s. These parameter sets had very long toeholds, that sped up the rate of polymerisation, and short information domains, that increased the rate of trimer detachment.



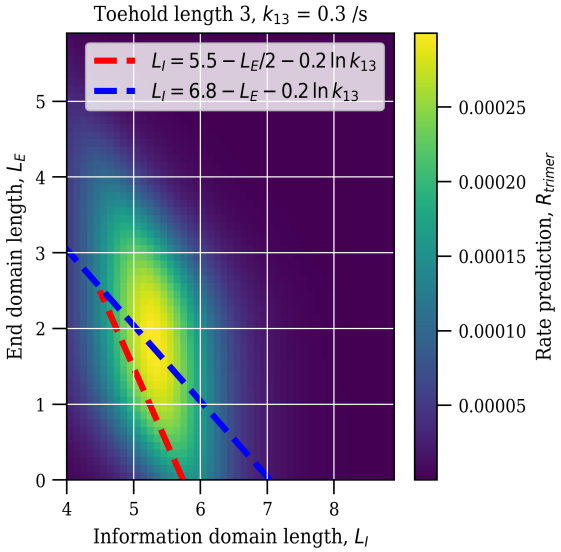
(a) Log-Log scale

(b) Linear-linear scale

Figure 4.25: The model R_{trimer} of Eqn.4.40 approximates the effective reaction rate $\ln 2/t_{1/2}$ of the system when the leak reaction is suppressed $k_{15} = k_{16} = 0$. data points where the final yield was less than 0.5nM were excluded, as many of these did not reach equilibrium. The model underestimates the reaction rate for very short information domain lengths, likely because these reactions are sped up by the release and rebinding of dimers which was not considered in the model.



(a) The effective rate $\ln 2/t_{1/2}$ is maximum for an information domain length between 5 and 6 and end domain lengths of 2 or less. For longer information domains, shorter end domains increase the rate, while for shorter information domains longer end domains enhance the rate.



(b) A heatmap of the rate model R_{trimer} given in Eqn.4.40 for a toehold of length 3 or greater (such that $k_{13} \approx 0.3/s$) shows that the optimal information domain length is between 5 and 6 and the optimal end domain length is between 1 and 2, in good agreement with the optimal parameters presented in Fig.4.22.

Figure 4.26: The optimal information domain length lies between 5 and 6 as shown by the simulation results in **(a)** and as predicted by the model in **(b)**.

With a correction factor of $1/10$, the overall rate of trimer formation per molar template under this simple model is

$$R_{\text{trimer}} = 0.1 \left(R^{-1}_{\text{displaced unzipping}} + R^{-1}_{\text{pol end}} + (k_8[\text{T}])^{-1} \right)^{-1} \tag{4.39}$$

$$= 0.1 \left(\frac{k_4 + k_{10}}{k_3\,[\text{M}]\,k_{10}} + \frac{k_8 + k_{13}}{k_{13}\,k_2\,[\text{M}]} \frac{k_{14} + k_{13}}{k_{13}} + k_8^{-1} \right)^{-1}. \tag{4.40}$$

Equation 4.40 approximates the effective rate constant for the reaction when the monomer concentration is set to [M] = 10nM as shown in Fig. 4.25a. For shorter information domains, the model underestimates the rate at low rates.

The polymerisation reaction rate k_{13} saturates with a toehold of length greater than 2, hence the top 50 fastest parameters had toeholds longer than 2, as seen in Fig.4.22. For a toehold of 3 or greater, a single mismatch healed in the polymerisation reaction was generally sufficient to make the reaction irreversible, hence the fastest parameters also had one or more mismatches, as shown in Fig.4.22.

Assuming that the polymerisation process is irreversible, $k_{13} \gg k_{14}$, the model in Eqn. 4.40 explains why intermediate to short information domain lengths and short end domain lengths result in a higher rate, as shown in Fig. 4.26a. When the information domain and end domains are very short, the process is rate limited by the

polymerisation step at the final site on the template, due to a low template occupancy. This is avoided when the rate of polymersiation $k_{13} > k_8$ is faster than the rate of detachment. For toeholds of length 1, 2 or 3 and greater, the polymerisation rate is approximately $k_{13} \approx 10^{-4.3}\exp(10L_{\mathrm{TH}}/3) = (10^{-3}, 0.04, 0.3)s^{-1}$, and the rate of detachment at the end of the template is $k_8 \approx k_2\exp(-1.5(L_{\mathrm{I}}+L_{\mathrm{E}})/0.6)$. Substituting these approximations into $k_{13} > k_8$, we arrive at an inequality on the total length of the end domain (end + information domain length) and the toehold length

$$L_{\mathrm{E}} + L_{\mathrm{I}} > 10.1 - 1.3L_{\mathrm{TH}}\,, \tag{4.41}$$

for toeholds of length $L_{\mathrm{TH}} = (1, 2, \geq 3)$ (e.g. toeholds of length $L_{\mathrm{TH}} \geq 3$ should be treated as though they are length 3). Eqn. 4.41 gives $L_{\mathrm{E}} + L_{\mathrm{I}} > (8.8, 7.5, 6.2)$ for toeholds of length 1, 2 and 3 or greater respectively.

When the information domain is very short, the rate of displaced unzipping $R_{\text{displaced unzipping}} = \frac{k_3[\mathrm{M}]k_{10}}{k_4+k_{10}} \approx k_3[\mathrm{M}]$ drops as the rate of binding decreases k_3. The rate of displaced unzipping is greatest for $L_{\mathrm{I}} \approx 5$, where a balance between the rates of tail unbinding and monomers binding from solution balance. When the information domain length is greater than or equal to $L_{\mathrm{I}} \geq 5$, the rate of monomer binding saturates $k_3 \approx k_2 \approx 10^{6.7}M^{-1}s^{-1}$, beyond which increasing the length of the information domain exponentially decreases the rate of unzipping. Hence, when the information domain is long the rate is limited by the unzipping step involved in $R_{\text{displaced unzipping}}$.

Furthermore, if the end domain length is also long, the process may be slowed further by unbinding at the end of the template k_8. For information domain lengths greater than $L_{\mathrm{I}} > 5$, the rate of displaced unzipping and end unbinding balance when the end domain length is approximately $L_{\mathrm{E}} \approx 2.6$. Hence, for information domain lengths greater than $L_{\mathrm{I}} > 5$, if the end domain length is $L_{\mathrm{E}} \leq 2$, the displaced unzipping process is rate limiting, and if the end domain is length $L_{\mathrm{E}} \geq 3$ or longer, unbinding at the end of the template will be rate limiting.

For a short end domain $L_{\mathrm{E}} \leq 2$, such that the unbinding is rate limited by the displaced unzipping process, for an information domain of length $L_{\mathrm{I}} > 5$, such that the binding rate k_3 is saturated, and given sufficient mismatches that the polymerisation process is irreversible $k_{13} \gg k_{14}$, we can reduce R_{trimer} to:

$$R_{\text{trimer}} \approx \left(\frac{k_4+k_{10}}{k_3[\mathrm{M}]k_{10}} + \frac{k_8+k_{13}}{k_{13}k_2[\mathrm{M}]}\right)^{-1} \tag{4.42}$$

Substituting in the approximate scaling of the rates with the domain lengths, ($k_8 \approx k_2\exp(-1.5(L_{\mathrm{I}}+L_{\mathrm{E}})/0.6)$ and $k_{10} \approx k_4\exp(-1.5(L_{\mathrm{I}}-5)/0.6)$, differentiating with respect to the information domain length L_{I} and solving for its zero, we find that the information domain length that optimises speed lies approximately along the line,

$$L_{\mathrm{I}} \approx 5.5 - 0.5L_{\mathrm{E}} - 0.2\ln k_{13} \tag{4.43}$$

which yields $L_{\mathrm{I}} \approx (6.8 - 0.5L_{\mathrm{E}}, 6.1 - 0.5L_{\mathrm{E}}, 5.7 - 0.5L_{\mathrm{E}})$ for toeholds of length $L_{\mathrm{TH}} = (1, 2, \geq 3)$.

The constraints in Eqn.4.41 and Eqn. 4.43 predict the optimal domain lengths. For a toehold of length $L_{\mathrm{TH}} = 3$, combining the constraint in Eqn. 4.43 with the constraint set by polymerisation and the occupancy of the end domain in Eqn.4.41, $(L_{\mathrm{E}} + L_{\mathrm{I}} > 6.2)$, and given an information domain long enough that the monomer binding rate is saturating $L_{\mathrm{I}} = 5$, we get that the end domain should be about $L_{\mathrm{E}} \approx 1.5$, which is in good agreement with the optimal domain lengths seen in Fig.4.22 and . For end lengths between 0 and 2, the optimal information domain lengths for toeholds $L_{\mathrm{TH}} \geq 3$, is between 5 and 6, exactly in agreement with the optimal parameters identified in Fig.4.22. Thus, these simple models predict the optimal behaviour of the full simulations well.

Summary

Here I have demonstrated that the reaction proceeds fastest when the toehold is long and there are sufficient mismatches healed in the polymerisation step to ensure that it is effectively irreversible, $k_{13} \gg k_{14}$. With a simple model, I have shown that the information domain should be longer than $L_{\mathrm{I}} \geq 5$ to ensure hybridisation-limited monomer binding and that the combined end domain length should be long enough that the polymerisation step is not limited by the occupancy of the final site on the template: $L_{\mathrm{E}} + L_{\mathrm{I}} > 6.5$, for toeholds of length 3 or greater. If the information domain is $L_{\mathrm{I}} > 5$, then the optimal information domain length lies approximately along the line $L_{\mathrm{I}} \approx 5.5 - 0.5L_{\mathrm{E}} - 0.2\ln k_{13}$. In the absence of the leak reaction pathway, the information domain lengths that optimised the reaction speed were between 5 and 6, and the optimal end domain length was between 0 and 2.

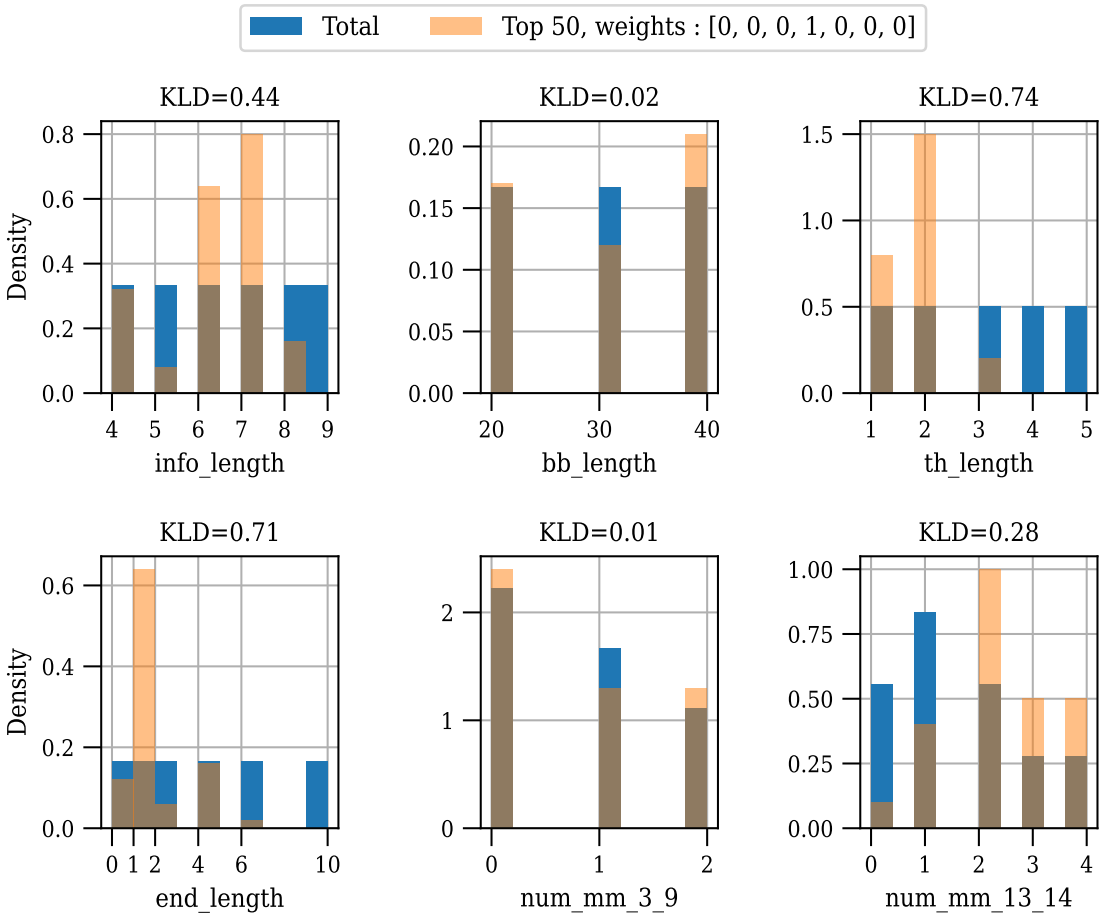


Figure 4.27: The input parameter distributions that showed the minimal variation in $t_{1/2}$ when measured without and with the leak reaction pathway are shown in orange with the whole ensemble in blue. Toeholds of length 2 or less, information domains of 6 or 7 and an end length of about 1 resulted in the minimum variation in $t_{1/2}$. A greater number of mismatches healed in the polymerisation step 13,14 also decreased the effect of the leak on the kinetics.

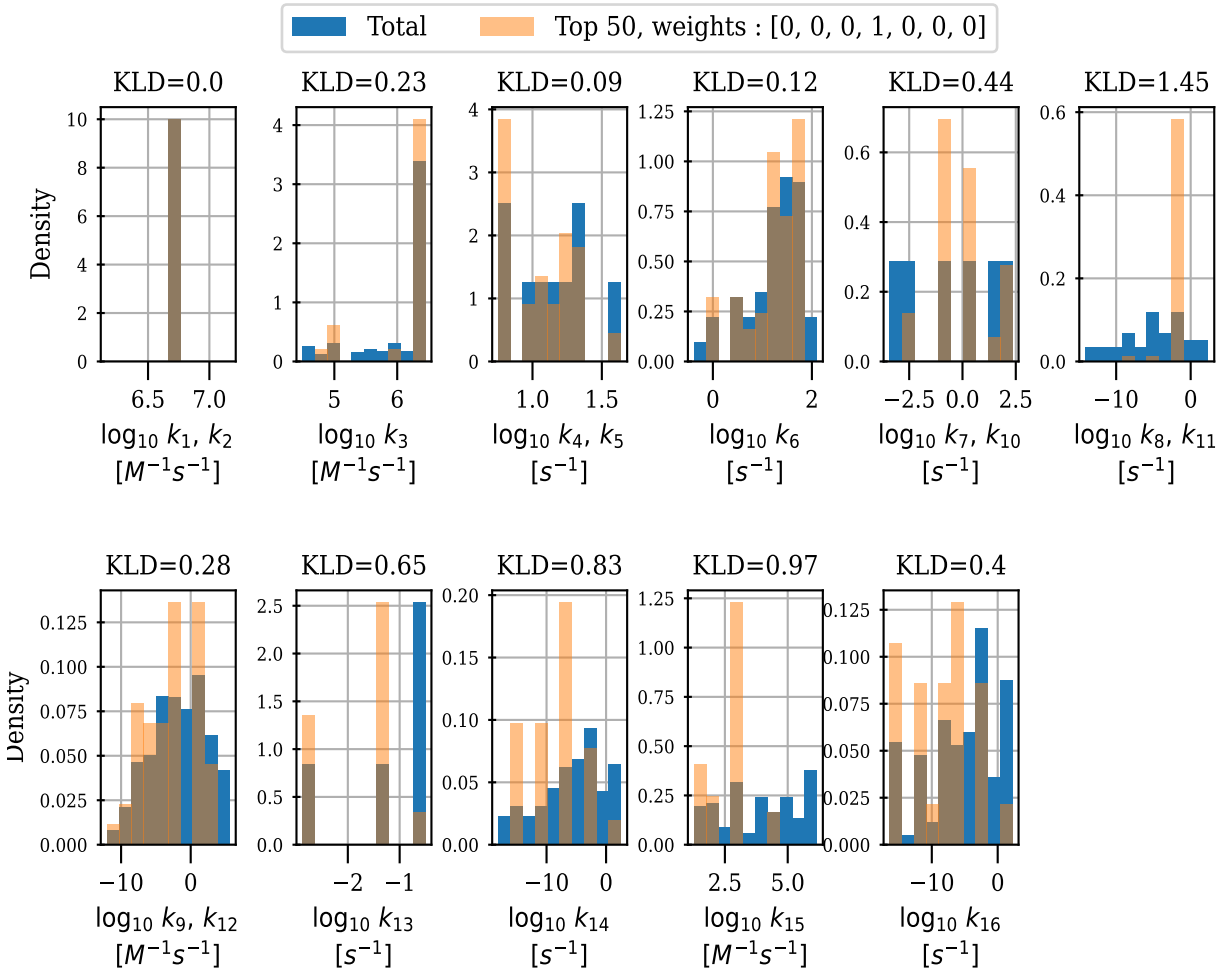


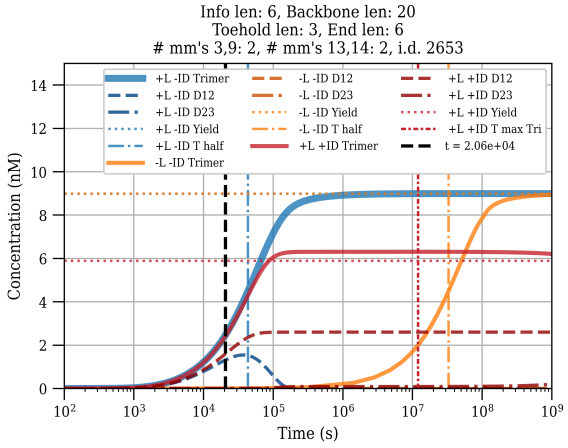
Figure 4.28: The parameters with time scales that were minimally affected by the leak pathway featured relatively slow polymerisation and leak rate constants, k_{13}, k_{15}, that tend to be irreversible $k_{13} \gg k_{14}, k_{15} \gg k_{16}$, and rapid binding of monomers from solution, k_3, and a tightly controlled rate of unbinding from the end of the template k_8.

4.3.3 Optimising for minimal leak reaction

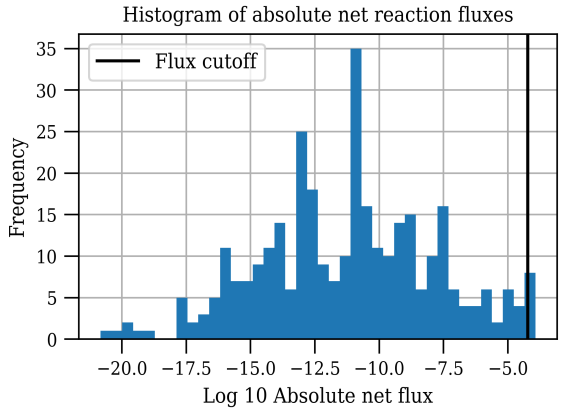
I analyse the factors that can be controlled to minimise the effect of the leak reaction (non-template directed polymerisation) on the kinetics of the reaction, by focusing on the ratio of the reaction $t_{1/2}$ without and with the leak reaction pathway permitted, $r_{t_{1/2}} = \frac{t_{1/2,-L}}{t_{1/2,+L}}$. Here I set the weight for this metric to 1 and to 0 for all other metrics. The top 50 performing parameters feature short toeholds, short end domains, a large number of mismatches healed in the polymerisation reaction, and moderately long information domain lengths, as can be seen in Fig. 4.27. Thus the polymerisation and leak reaction rates are very slow and irreversible, (k_{13}, k_{15} skew high and k_{14}, k_{16} skew low), the binding of monomers from solution is fast, (k_3 is high), and the rate of unbinding from the end of the template k_8 is tightly controlled, as seen in Fig. 4.28. This points to a dynamics in which the timescale of trimer formation and release at the end of the template is shorter than the timescale of trimerisation via the leak pathway.

I looked at a point in the parameter space in which the kinetics was much faster with the leak than without it in Fig. 4.29. Here the system has a longer end domain that slows the rate of release of trimers from the final site on the template, resulting in a much longer $t_{1/2}$ without the leak pathway than with it, as shown in Fig. 4.29a. In Fig.4.29b, I show a histogram of the absolute reaction fluxes between states at a time $t = 2 \times 10^4$ seconds in the system with the leak pathway permitted. In Fig.4.29c, I present a network constructed by discarding all reactions with an absolute flux less than a threshold value. We see a cycle of reactions wherein a trimer remains tightly bound to the end template site and trimers are formed via the leak pathway on the rest of the template. This evidence supports the notion that the leak reaction will dominate the kinetics if the timescale of forming and releasing a trimer via the standard reaction pathway is slow in comparison. In the flux cycle illustrated in Fig. 4.29, the speed of the standard pathway is slow due to the relatively long end domain. Instead, if the end domain were very short, the monomer occupancy of the last template site may be very low, and thus the speed of the standard pathway could be rate-limited by polymerisation.

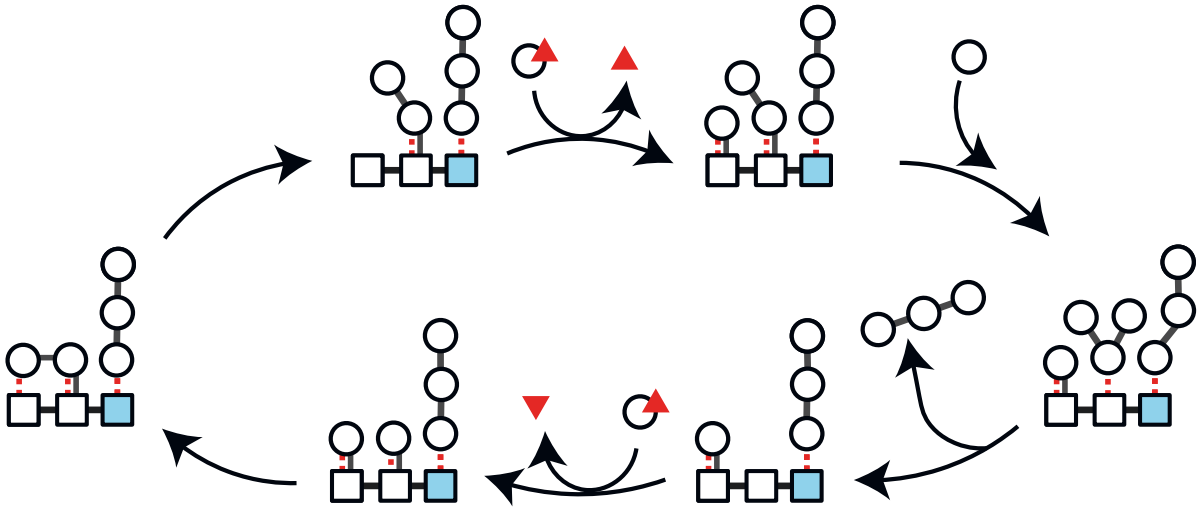
I devised another simple model that constrains the rate of the reaction when the leak is present. I posit that the rate of trimer formation in the presence of the leak pathway $R_{\text{trimer, leak}}$ is approximately the sum of the rate of forming and releasing a trimer via the non-leak pathway, R_{trimer}, and the rate of forming a trimer via the leak pathway $R_{\text{leak only}}$. This gives $R_{\text{trimer, leak}} \approx R_{\text{trimer}} + R_{\text{leak only}}$. The rate of trimer formation via the standard pathway is given by Eqn.4.40. I assume that the rate of trimer formation via just the leak reaction pathway $R_{\text{leak only}}$ is limited by either the rate of leak polymerisation k_{15}[M], or the rate of product release, which itself is either rate limited by the rate at which a tail unzips and is replaced by a unit with concentration [M] $k_{10}k_3$[M]$/(k_{10}+k_4)$ (process depicted in Fig. 4.30) or the rate of detachment from the information domain k_7.



(a) A timeseries where the reaction timescale $t_{1/2}$ was three orders of magnitude larger when the leak was allowed versus when it was prohibited. The reaction fluxes at $t = 2 \times 10^4 s$ were analysed to determine the reaction pathway by which the leak sped up the reaction.



(b) A histogram of the logarithm of the absolute reaction fluxes of the system in Fig.4.29a. The fluxes below a certain threshold, shown by the black vertical line, were excluded.



(c) The fluxes that were above the threshold shown in Fig.4.29b form a cycle that net consumes three monomers and generates a trimer. The trimerisation process is not template-directed, rather the final monomer is incorporated via the leak reaction pathway.

Figure 4.29: The kinetics of a system in which the toehold length is moderately long and the end domain is long are greatly sped up by the leak reaction pathway. Analysis of the main reaction fluxes through the model at short times indicates that the rapid production of trimers is a result of the leak reaction pathway. Here the end domain is so tightly bound and the toehold is long enough that the leak reaction exceeds the unbinding rate $k_{15}[\mathrm{M}] \gg k_8$.

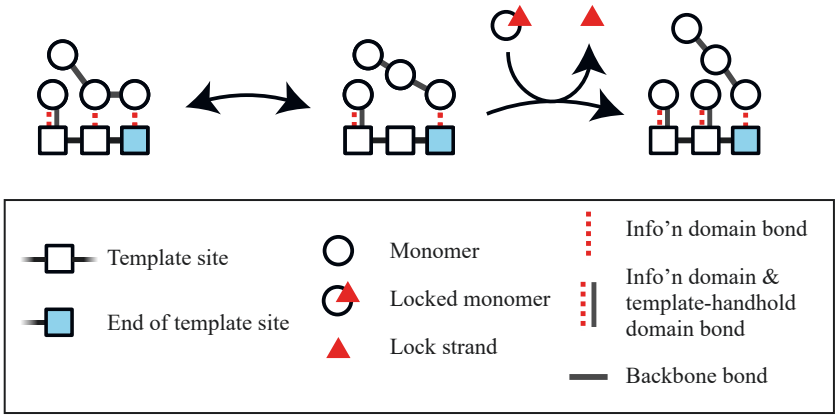


Figure 4.30: An illustration of the process in which the tail of a polymer unzips from the template and is prevented from re-zipping by the binding of a monomer.

Using the same correction factor of 0.1 as in Eqn. 4.40, this assumption gives us

$$R_{\text{leak only}} \approx 0.1 \left(\frac{1}{k_{15}[M]} + \frac{k_{10}+k_4}{k_{10}k_3[M]} + \frac{1}{k_7} \right)^{-1}, \tag{4.44}$$

and

$$R_{\text{trimer, leak}} = R_{\text{trimer}} + R_{\text{leak only}} \tag{4.45}$$

$$= 0.1 \left(\left(\frac{k_{10}k_3\,[\text{M}]}{k_4+k_{10}} \right)^{-1} + \left(\frac{k_{13}k_2\,[\text{M}]}{k_8+k_{13}} \right)^{-1} + k_8^{-1} \right)^{-1} \tag{4.46}$$

$$+ 0.1 \left(\frac{1}{k_{15}[M]} + \frac{(k_{10}+k_4)}{k_{10}k_3[M]} + \frac{1}{k_7} \right)^{-1}. \tag{4.47}$$

In Fig. 4.31a, I show that $R_{\text{trimer, leak}}$ approximates the observed $t_{1/2}$, when [M] = 10nM.

The ratio of $t_{1/2}$ measured without leak pathway to $t_{1/2}$ with the leak reaction pathway, $r_{t_{1/2}} = \frac{t_{1/2,-L}}{t_{1/2,+L}}$ is then approximately given by $r_{t_{1/2}} \approx \frac{R_{\text{trimer, leak}}}{R_{\text{trimer}}}$ which simplifies to

$$r_{t_{1/2}} \approx \left(1 + \frac{R_{\text{leak only}}}{R_{\text{trimer}}} \right)^{-1}. \tag{4.48}$$

Eqn. 4.48 provides an approximate upper bound on the ratio of timescales as shown in Fig. 4.31b.

The effect of the leak pathway on the kinetics is minimised when ratio $r_{t_{1/2}} \approx 1$ when $R_{\text{leak only}} \ll R_{\text{trimer}}$. This condition can be reduced to

$$\frac{1}{k_{15}[M]} + \frac{1}{k_7} \gg \frac{k_8+k_{13}}{k_{13}k_2\,[\text{M}]} + \frac{1}{k_8}, \tag{4.49}$$

which can be further simplified to

$$\frac{1}{k_{15}[M]} \gg \frac{k_8}{k_{13}\,k_2\,[\mathrm{M}]} + \frac{1}{k_8}, \tag{4.50}$$

where I have used the fact that $k_7 \gg k_8$ if the end length is greater than 1, and that the process is not generally limited by the binding rate of monomers ($k_2[\mathrm{M}] \approx 10^{1.7}/s$ is generally much greater than all other terms). We can simplify further by approximating the rate constants involved in Eqn.4.50. The leak polymerisation rate is approximately $k_{15}[\mathrm{M}] \approx 10^{-7.5+1.1L_T}$, where L_T is the toehold length. The polymerisation rate on the template is approximately $k_{13} = \left(10^{0.5} + 10^{5-1.9L_T}\right)^{-1}$. The rate of unbinding is approximately $k_8 \approx 10^{6.7-1.1(L_\mathrm{I}+L_\mathrm{E})}$, where L_I and L_E are the lengths of the information and end domains.

In the limit that the occupancy of the final template site doesn't limit the polymerisation rate, $\frac{k_8}{k_{13}\,k_2\,[\mathrm{M}]} \ll \frac{1}{k_8}$, we find that the constraint in Eqn.4.50 is satisfied by shorter toeholds. If the leak timescale is at least a factor 10^X longer than the end detachment timescale, the leak will have minimal effect, and therefore we can reduce

$$\frac{1}{k_{15}[\mathrm{M}]} \approx 10^X \frac{1}{k_8}, \tag{4.51}$$

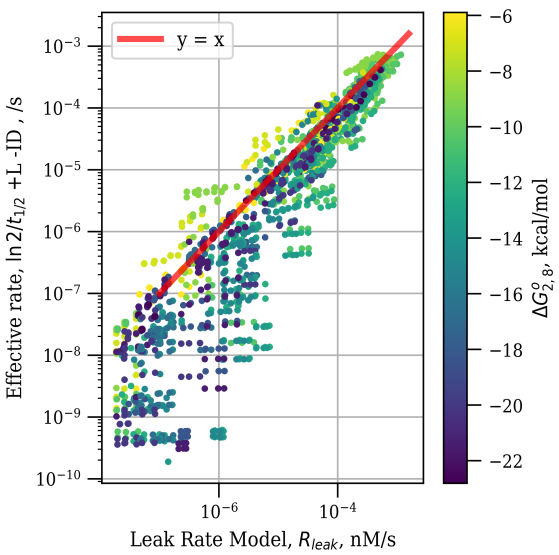
to

$$L_\mathrm{I} + L_\mathrm{E} < 12.9 - 0.9X - L_T. \tag{4.52}$$

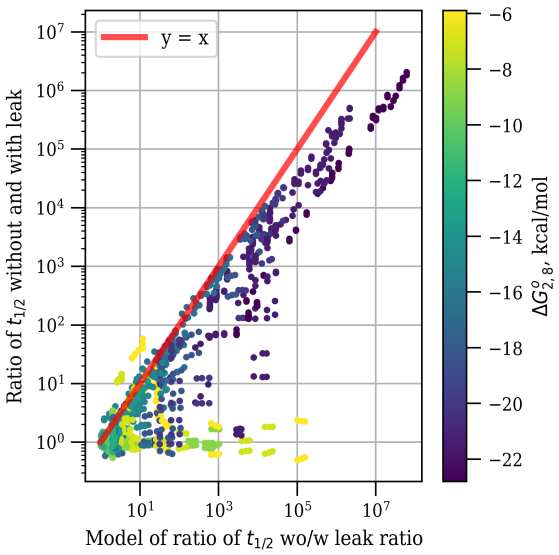
Choosing $X = 2$, which would give a 1 in 100 leak to polymerisation event ratio (effectively an error-rate in this model), and a toehold of length 2, we get $L_\mathrm{I} + L_\mathrm{E} < 9$. The constraint given by Eqn. 4.41 on balancing the occupancy of the template and the polymerisation rate gives $L_\mathrm{E} + L_\mathrm{I} > 7.5$ for a toehold length of 2. Hence, a combined length $L_\mathrm{I} + L_\mathrm{E} \approx 8$ satisfies both conditions, minimising the leak reaction while keeping the occupancy of the end of the template high enough for fast polymerisation via the standard pathway. Given that the rate of displaced unbinding and the rate of end detachment balance for an end length $L_\mathrm{E} \approx 2$, the information domain should be about $L_\mathrm{I} \approx 6$, in good agreement with the optimal parameters in Fig.4.27. A shorter toehold slightly reduces the free energy change associated with the polymerisation reaction (transitions 13 and 14); this explains the observation that more mismatches are required to ensure the irreversibility of the reaction for shorter toeholds.

Summary

In summary, the leak has a minimal effect on the timescale of the reaction if the standard reaction pathway is fast. Crucially, the rate of end detachment should be significantly faster than the rate of polymerisation by the leak pathway, given by Eqn.4.52, and this can be achieved with shorter toeholds. However, if the toehold is



(a) The leak rate model given by Eqn.4.47 approximates the effective rate $\ln 2/t_{1/2}$ measured when the leak pathway is permitted.



(b) The ratio of rate models given by Eqn.4.48 provides an upper bound on the ratio of timescales $t_{1/2}$ measured without and with the leak reaction pathway. When the information and end domains are very short $\Delta G^o_{2,8} \approx -6$ the bound is weak.

Figure 4.31: The rate of trimer production was estimated with a simple model in Eqn.4.47

shortened, the information domains and end domains must be lengthened slightly to not slow down the rate of polymerisation at the end of the template (Eqn.4.41). Furthermore, the reduced toehold length requires that more mismatches be healed in the polymerisation reaction to ensure that the polymerisation process is irreversible, $k_{13} \gg k_{14}$. An end domain of approximately $L_{\mathrm{E}} \approx 2$ ensures that the displaced unzipping and end unbinding rates are balanced, to deliver a high overall speed.

4.3.4 Minimisation of dimer release

Here I analyse the balance of parameters that prevent the early release of dimers. A weight of 1 was given to the ratio of trimers to dimers in the model in which dimer binding reactions (including polymerisation via the leak reaction pathway) are prohibited. The trimer-to-dimer ratio is measured at the time of maximal trimer concentration. A weight of 0 was given to all other metrics.

In Fig. 4.32, we see that the parameter values that minimise dimer release also minimise the effects of the leak reaction, as the kinetic ratios without and with the leak reaction are all close to 1. In Fig. 4.33, we see that the input parameters of the top 50 performing parameters had information domain lengths of 7 or 8, end domain lengths less than 3, short toeholds of length less than 3 and more than two mismatches healed in the polymerisation reaction.

How can dimer release be eliminated? From looking at the measures in Fig.4.32, eliminating the propensity for the system to be subject to leaks reduces the number of dimers that are released. For instance, dimers could be produced by the leak pathway by polymerising from solution via reaction k_{15} and quickly detaching in systems with short information domains and long toeholds. Therefore, minimising the leak reaction pathway speed, by satisfying the constraints derived in the previous section would suppress most dimer releases.

However, in the absence of the leak reaction pathway, dimer D12 could be generated under the tail of a template-bound trimer if the trimer remains bound long enough for the dimer to fall off. In this regime, the rate of trimer production would be k_8 and the rate of dimer production would be rate limited by the rate of reinvasion of the lock strand k_9[L]. Hence, the expected ratio of trimers to dimer D12 in solution should be $\frac{k_8}{k_9[\mathrm{L}]}$, and the yield of trimers would then be given by

$$Y = [\mathrm{M}]/(1 + k_9[\mathrm{L}]/k_8)\,, \tag{4.53}$$

where [M]=10nM is the initial concentration of each monomer. Here I have assumed that all monomers are converted into dimer D12 or trimers, which is possible if the polymerisation free energy is sufficiently strong. A necessary condition for the yield to be high is $k_9[\mathrm{L}] \ll k_8$. Approximating the rate of lock invasion as $k_9 \leq 0.1\exp(3L_{\mathrm{TH}})$ (k_9 is approximately equal to this if there are no mismatches healed in template binding reactions 3 and 9, and less otherwise) and approximating the rate of unbinding at the end of the template as $k_8 \approx 10^{6.7}\exp(-2.5\,(L_{\mathrm{I}} + L_{\mathrm{E}}))$, I find that dimer release should be minimial if

$$L_{\mathrm{E}} + L_{\mathrm{I}} < 14.5 - 1.2L_{\mathrm{TH}} - 0.9X\,, \tag{4.54}$$

which, for $X = 2$ and $L_{TH} = 2$ reduces to $L_E + L_I < 10.3$. We can see that this constraint on the lock reinvasion rate (Eqn.4.54) is satisfied if the constraint for leak minimisation in Eqn.4.52 is satisfied. The reason for this is simple. Both the lock invasion and the leak reaction pathways are toehold-mediated strand displacement processes, with equal-length toeholds. However, the lock reinvasion rate is never greater than the leak invasion rate, as the reaction is slowed by the potential presence of mismatches and by the energetic cost that must be overcome to break the information domain bonds with the template ($\Delta G_{6,12}$). Therefore, if the standard reaction pathway is faster than the leak pathway such that the leak's effect on the kinetics is minimal, so too will be the lock's effect on early dimer unbinding. If the concentration of the lock were much higher, this would necessitate more mismatches healed in the 3,9 reaction or, less conveniently, a further shortening of the toehold.

In Fig.4.34a, I show that the simple model of Eqn.4.53 predicts the yield of trimers reached in the model without the leak pathway and with dimer rebinding prohibited. When the leak pathway is allowed, as shown in Fig.4.34b the constraint of Eqn. 4.54 still applies to the parameters that reached maximum yields. In Fig.4.34b, the bright yellow region corresponds to parameters that reached a high yield without the leak pathway and high to medium yield with it, indicating the leak reaction increased dimer production. The dark purple region had a very low trimer yield without the leak pathway, but a higher yield when the leak pathway was permitted, indicating that the leak reaction increased trimer production (likely by the non-template directed polymerisation of monomer M3 to a template-bound dimer D12).

Summary

The model presented in this section suggests that early dimer release can be prevented by slowing the leak reaction pathways relative to the standard reaction pathways (as explored in the previous section), by having a sufficiently short toehold, set by the constraint in Eqn.4.52. By keeping the ratio k_8/k_9[L] high, trimers can be made to fall off from the end domain before dimers are displaced by lock re-invasion. If the toehold is short enough to prevent the leak reaction, then it will also be sufficient to prevent the lock re-invasion pathway, especially if some mismatches are healed in reaction 3,9, and if the lock concentration is kept below k_8/k_9.

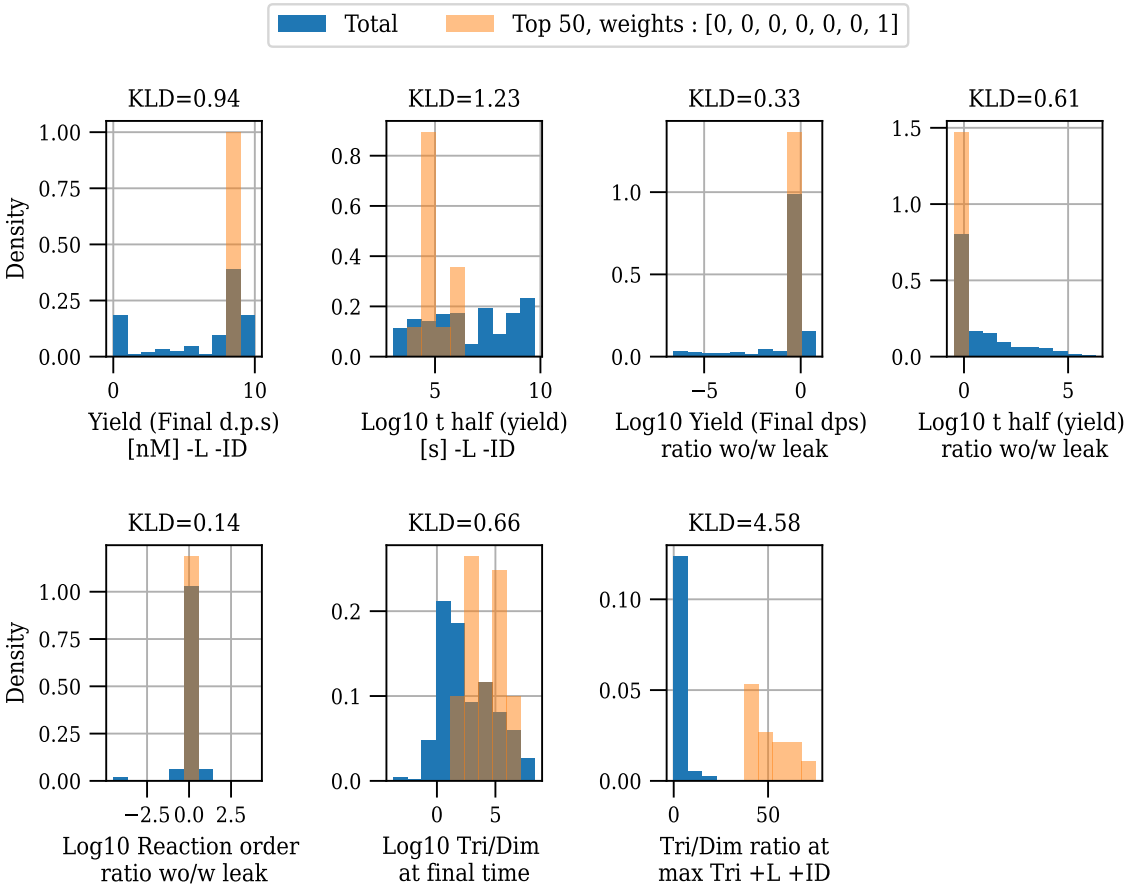


Figure 4.32: The parameter values that minimised the propensity for the system to release dimers early also minimise the propensity for the system to be affected by the leak reaction, as the ratios of the kinetic parameters measured without and with leak are all close to 1.

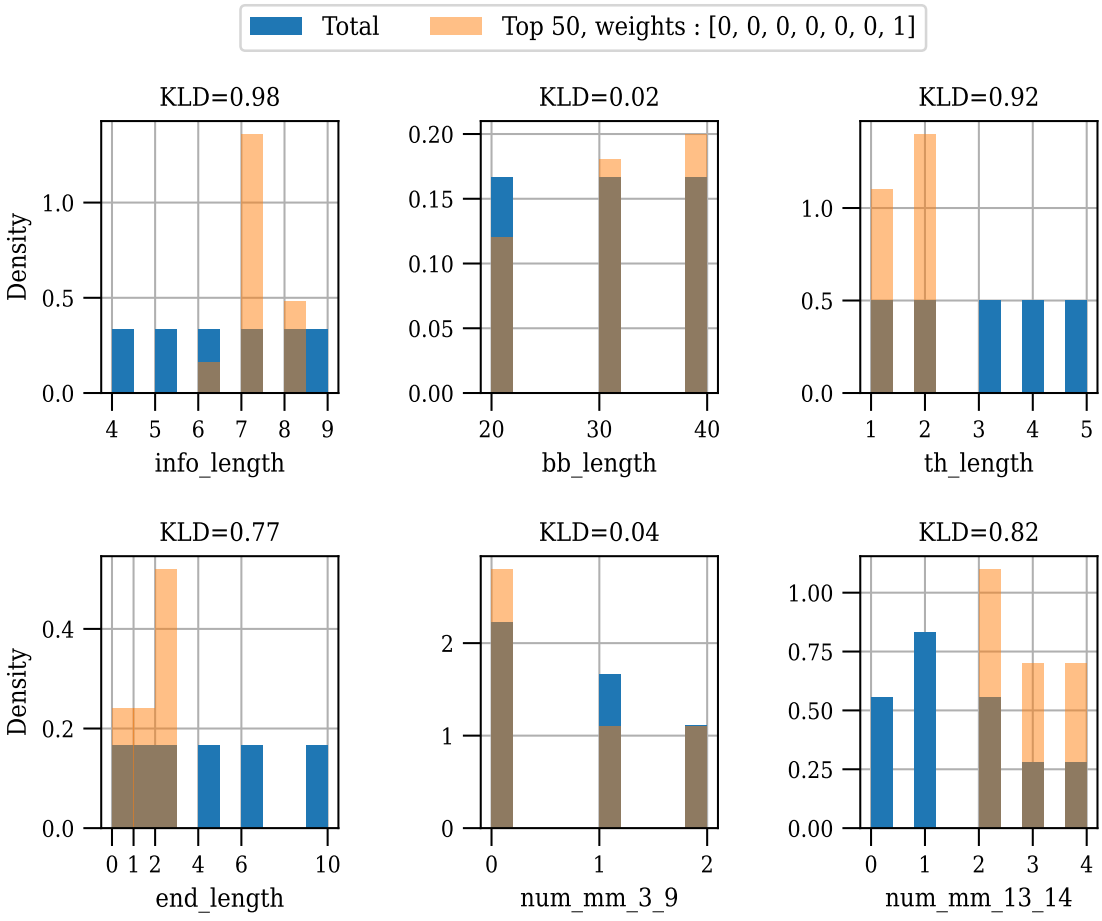
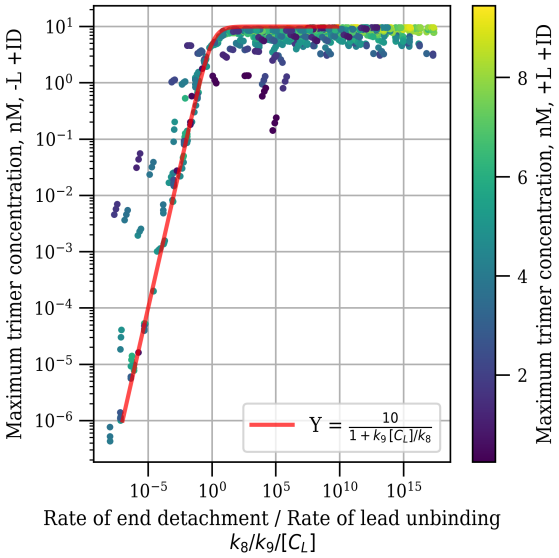
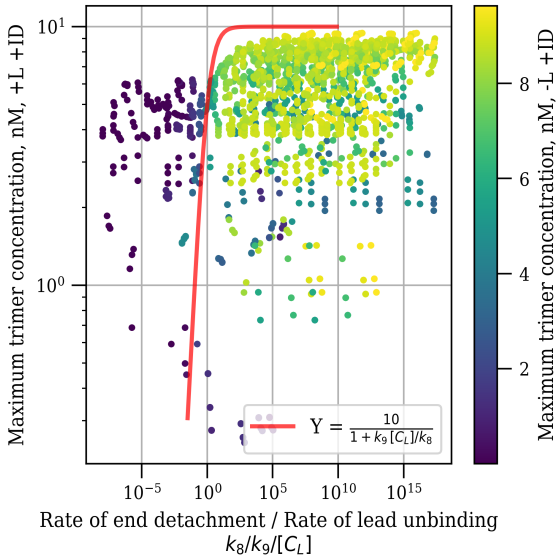


Figure 4.33: The release of dimers is minimal in systems with short toeholds, long information domains and end domains of length 2 or less, and greater than two mismatches healed in the polymerisation reaction 13,14.



(a) A simple model (Eqn. 4.53) that posits that the trimer concentration in solution is set by the relative rates of dimer (k_9) and trimer release (k_8), predicts the trimer yield when the leak pathway is prohibited.



(b) The model in Eqn. 4.53 places a necessary condition on whether the maximum trimer concentration can be reached (10nM) when the leak reaction is allowed. The leak reaction pathway enables many more paths for dimer release, and, as such, minimising the rate of lock reinvasion does not prevent the leak from producing dimers instead.

Figure 4.34: The trimer yield was compared to a simple model (Eqn.4.53). Parameters sets that had zero polymerisation mismatches (low thermodynamic drive) and that had timescales $t_{1/2} \geq 10^7 s \approx 115$ days were excluded from these plots because most did not reach equilibrium, and hence the yield was not a good measure of the equilibrium concentration.

4.3.5 Optimising across metrics

Having explored the factors that optimise the trimer yield and the overall reaction speed, that minimise the effect of the leak reaction and minimise early dimer release, in this section, I analyse the parameter sets that performed best across most metrics. I used a weighting of (1,1,0,1,0,0,10), to optimise for yield and speed without the leak pathway, to minimise the difference in speeds without and with the leak pathway, and to strongly suppress the early release of dimers. I chose not to score against the ratio of trimer yields and reaction order without and with the leak, as these were generally optimised by systems which had similar speeds without and with the leak, and which had high yield and speed without the leak pathway. I chose not to optimise against the ratio of trimers to dimers measured in the model where dimer rebinding was allowed as this was lower bounded by the ratio measured at the time of maximum trimer concentration in the model with inert dimers. Therefore, optimising the trimer-to-dimer yield in the model with dimer rebinding prohibited also effectively optimised the ratio in the model with dimer rebinding. I weighted the ratio of trimers to dimers in the model with inert dimers highest as this deficiency is a source of loss of both length and sequence control.

The distributions of the optimal input parameters over the top 25 parameters are shown in Fig. 4.36. The best-performing systems feature an information domain length $L_{\mathrm{I}} \approx 7$, a toehold length of $L_{\mathrm{TH}} \leq 2$, three or four mismatches healed in the polymerisation reaction 13,14, and an end length of $L_{\mathrm{E}} \leq 2$. The information domain length is long enough that the rate of monomer binding k_3 saturates $L_{\mathrm{I}} > 5$. According to the constraint in Eqn. 4.52, the toehold length L_{TH} and the combined end and information domain length $L_{\mathrm{E}} + L_{\mathrm{I}}$ are short enough that the rate of unbinding at the end of the template is faster than the leak reaction pathway, and hence the system is minimally affected by leaks. Furthermore, minimising the leak reaction pathway also minimises the rate of lock reinvasion and the rate of dimer release (Eqn.4.54), and as such, the system produces a high ratio of trimers to dimers (50 as seen in Fig.4.35). Finally, the combined end and information domain length are of a length that the polymerisation process at the end of the template is effectively "one-shot" (as constrained by Eqn. 4.41), contributing to a high overall speed (most timescales are between $t_{1/2} \approx 10^4$ to 10^5 in Fig.4.35).

The large number of mismatches healed over the course of the reaction contributes to a strong thermodynamic drive toward trimer formation, as predicted by Eqn.4.32. The yield stays close to 9nM, and given the long information domain length and small end domain length, in equilibrium, one or more trimers will be bound to the template as predicted by Eqn. 4.36. The top 25 show a mix of 3 and 4 mismatches in the polymerisation reaction. One of the four could be spared for the monomer binding reaction 3,9 to further suppress lock reinvasion (though this is already minimal due to the long information domain and short toehold). Opting for an information domain of 6nts, as shown in Fig.4.37c, increases the speed over an information domain of 7nts, as shown in Fig.4.37a,

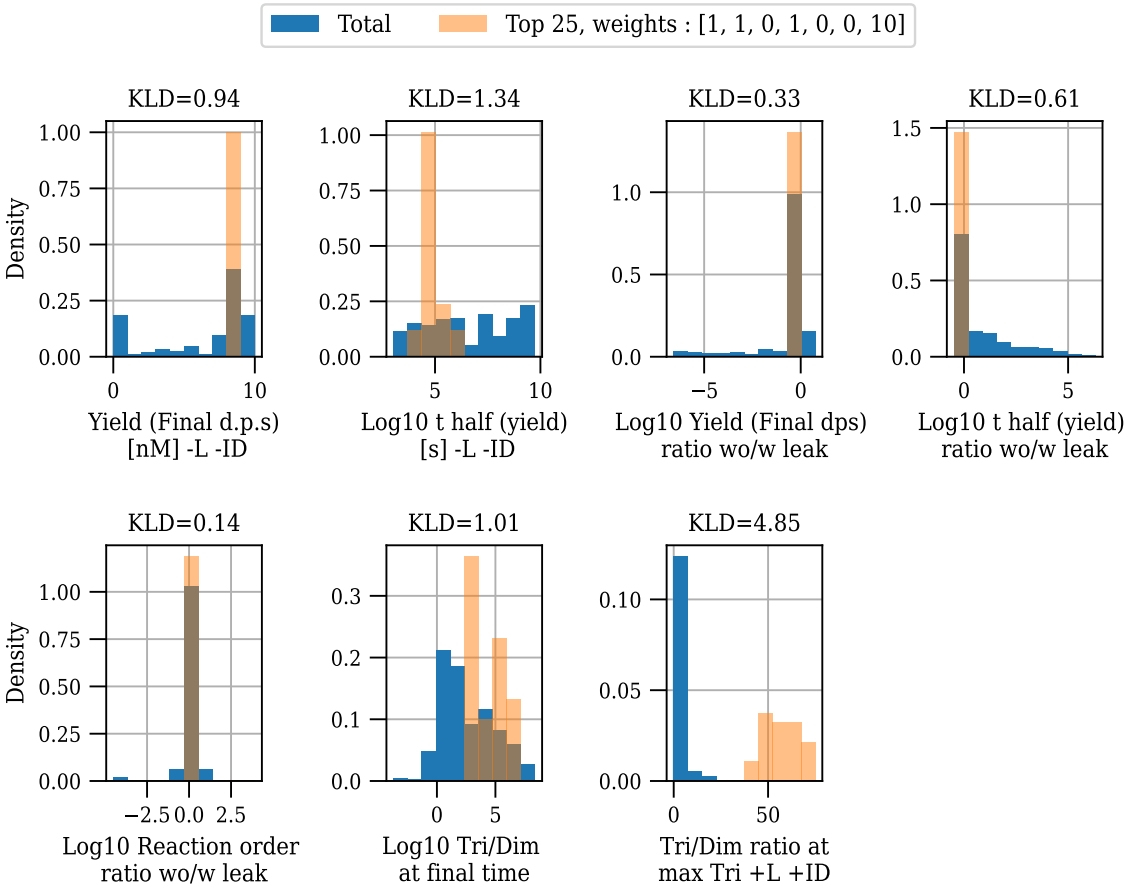


Figure 4.35: The weightings were set to $w = (1,1,0,1,0,0,10)$ to optimise for trimer yield, fast reaction speeds, the minimal effect of the leak reaction pathway on the reaction speed, and minimal release of dimers into solution. The top 25 scoring parameter sets all reached a yield of 9nM, had a $t_{1/2} \approx 10^5 s \approx 1$ day, had kinetics that was minimally affected by the leak pathway, and rarely produced dimers.

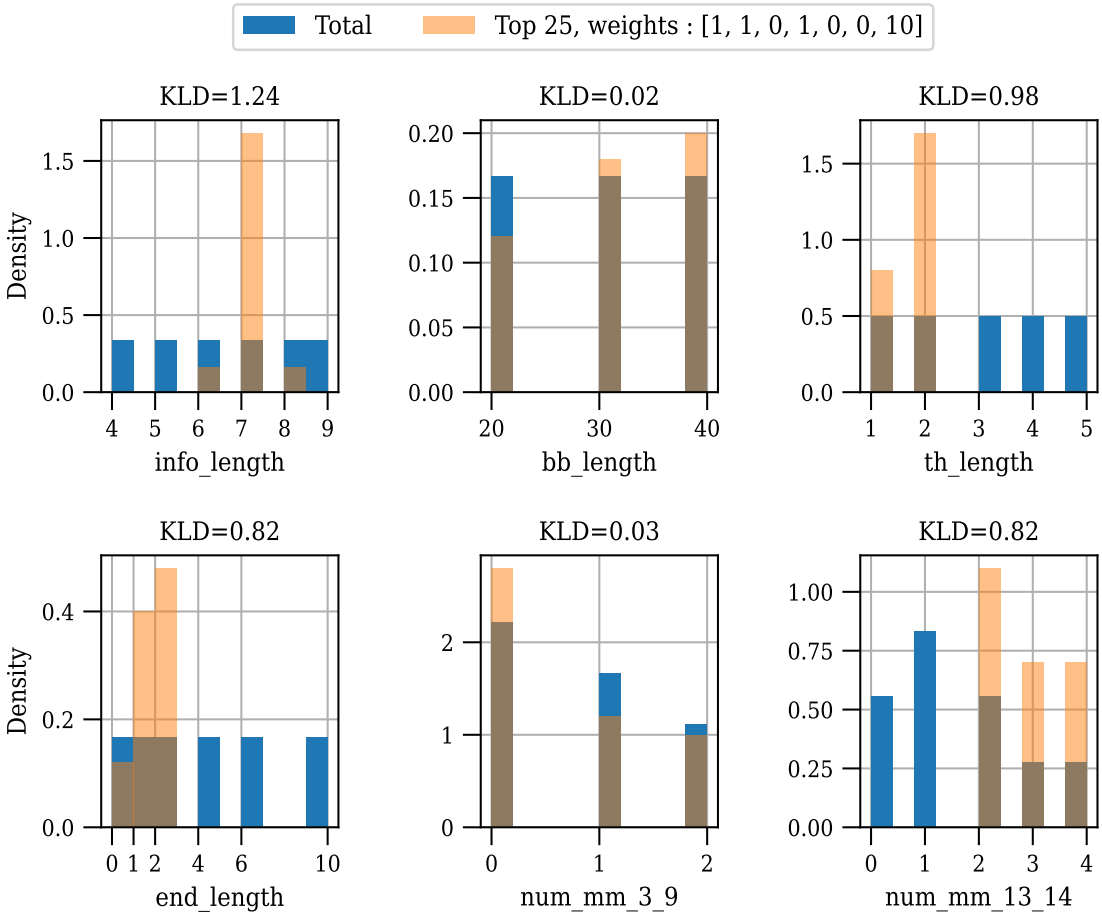


Figure 4.36: The weightings were set to $w = (1,1,0,1,0,0,10)$ to optimise for trimer yield, fast reaction speeds, the minimal effect of the leak reaction pathway on the reaction speed, and minimal release of dimers into solution. The top 25 scoring parameters had information domain lengths of 7, end domain lengths of 1 or 2, a toehold of 2 or less, a large number of mismatches healed in the polymerisation reaction 13,14, and somewhat tended to have slightly longer backbones.

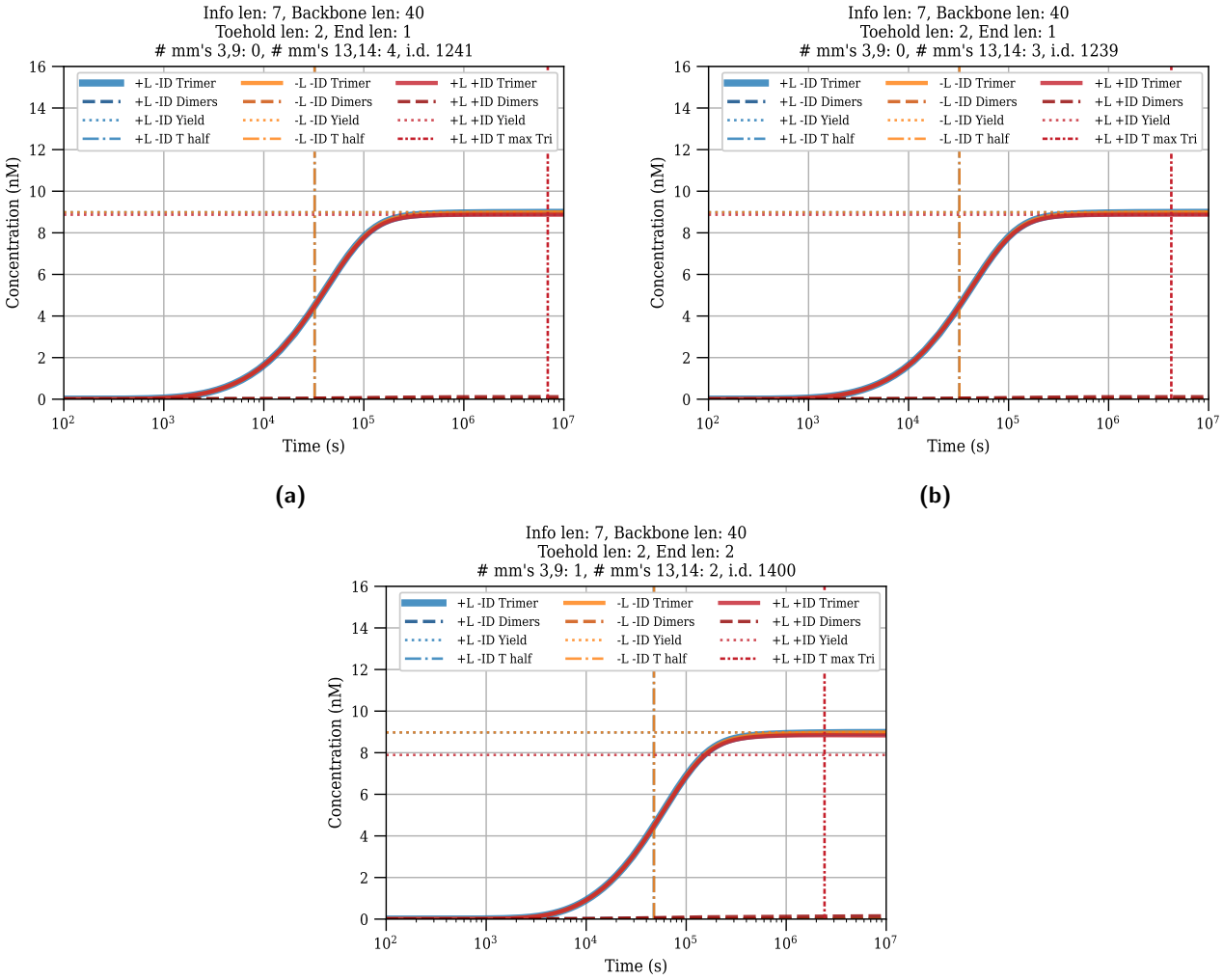
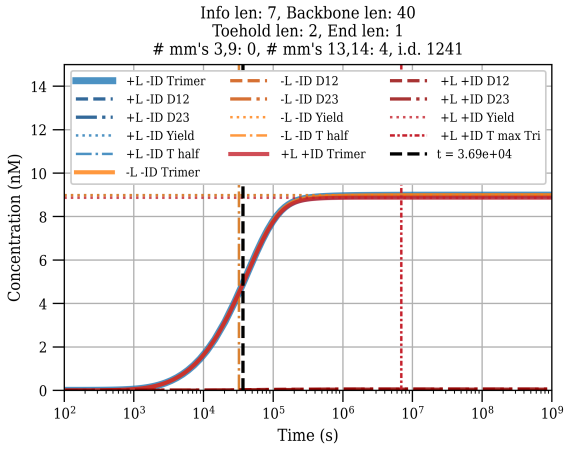


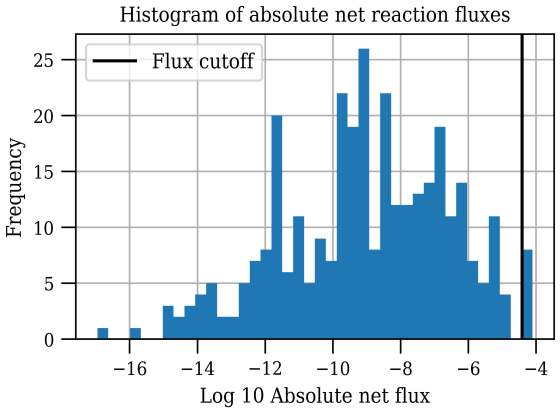
Figure 4.37: Timeseries of the top performing parameter sets. The simulations show high reaction speeds, minimal dimer release and minimal effect of the leak reaction pathway.

though it comes at the cost of a slightly higher production of dimers. The optimal parameters show a slight tendency toward longer backbone domain lengths. This decreases the rate of zipping k_4, which increases the rate of displaced unzipping, by increasing the probability that a tail is not bound to the template.

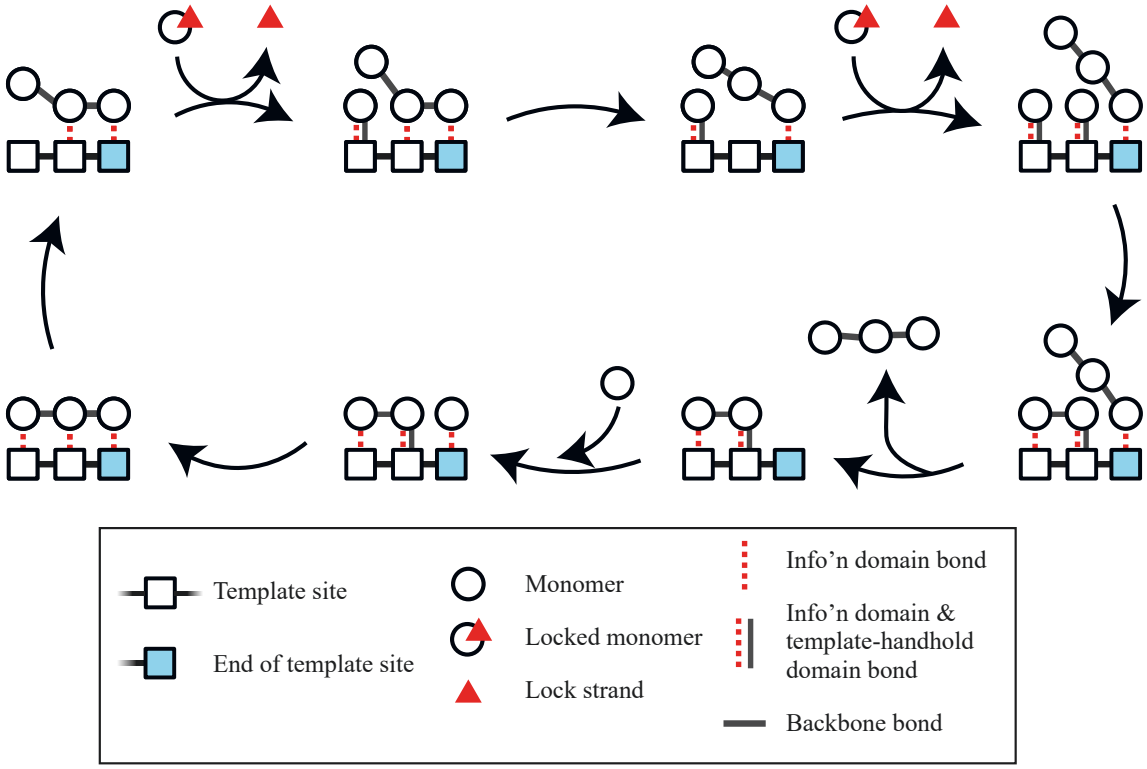
In Fig. 4.38, I show a system working optimally. Here, the maximum reaction fluxes drive a reaction cycle in which there is an orderly displacement of the tails of polymers during their growth. Such a cycle was predicted to achieve high sequence and length control in chapter 3.



(a) A parameter set that scored highly was investigated at $t = 3.7 \times 10^4 s$.



(b) The reactions with absolute fluxes greater than a threshold were chosen for plotting.



(c) The production of trimers at $t = 3.7 \times 10^4 s$ proceeds by the reaction cycle shown above. The system adopts a dense conformation on the template, predicted in chapter 3. Monomers bind underneath the tail of a trimer, preventing it from zipping back to the template. The trimer falls off and is replaced by a monomer, which polymerises into the growing chain.

Figure 4.38: Flux analysis of an optimal parameter set reveals that trimer production is driven by the simultaneous binding of monomers and growth under the tail of the trimer.

4.4 Conclusions

In this chapter, I have developed a model of a templated copying system that explicitly accounts for the rebinding of products to the template and a non-template-directed leak reaction pathway. The model was parameterised based on estimates of the kinetic and energetic properties of a DNA strand displacement reaction designed to perform the copying reaction. The standard free energies of the transitions involved were determined using Nupack [153, 154, 155], and the kinetic rate constants were estimated by comparison against kinetic parameters fitted to experimental data [21, 22]. The rates of toehold-mediated strand displacement reactions were estimated using the calculator developed by Irmisch [156]. I constructed the chemical network reaction model computationally using a bottom-up approach from a minimal set of fundamental reaction motifs that I applied depending on the conformational state of the system. The explicit sequence dependence of the model was neglected on the grounds that the main sources of loss of length control and sequence control result from the early release of fragments of polymers, the rebinding of fragments of polymers from solution, or the polymerisation of incorrect units in a non-template directed pathway.

I used the model to explore the design space for the DNA strand reaction network on a template of length three. I used four key metrics to evaluate the performance of the system: 1) the overall yield of trimers measured at long times where the leak reaction pathway was prohibited, 2) the speed, measured via the time to half yield when the leak reaction was prohibited, 3) the effect of the leak reaction on the kinetics which was measured via the ratio of the time to half yield in the model without the leak pathway to the model with the leak pathway, and 4) the propensity of the system to release fragmented polymers measured as the ratio of trimers to dimers at the time at which the trimer concentration was maximal in a model in which dimers were prohibited from rebinding to the template. I analysed the properties of the system that would optimise each of these qualities independently.

I found that the yield of trimers is thermodynamically constrained by the overall free energy of polymerising monomers. One C/C mismatch was generally sufficient to reach a maximal yield, as shown in Fig.4.32, though more mismatches helped to increase irreversibility. Further, I found that for appreciably high template concentrations the equilibrium of trimers bound to the template and in solution was constrained by the free energy of binding between trimers and the template, as expected.

In the absence of the leak reaction pathway, long toeholds increased the rate of polymerisation and a balance of the end and information domain lengths resulted in the highest overall reaction speed. The monomer binding was a toehold-mediated strand displacement process with a rate saturated for an information domain length of 5 or greater. Longer information domains decreased the rate at which the tails detached from the template and the rate at which monomers were released from the end of the template, resulting in a slower overall

speed. If the combined length at the end of the template was long enough such that the rate of end monomer unbinding matched the polymerisation rate, the process of monomer binding and polymerising at the end of the template proceeded in one shot, and hence the reaction speed was roughly optimal. For a shorter toehold, the polymerisation rate drops; therefore, a longer information and end domain was required to balance the polymerisation rate and the rate of monomer unbinding. Increasing the length of information and end domains beyond this level decreased the overall rate, which was limited by the rate of unbinding at the end of the template or the rate of unzipping the tail. An additional mismatch was required for shorter toeholds to ensure that polymerisation remained irreversible.

The presence of the leak reaction pathway placed constraints on the toehold lengths. The leak reaction pathway dominated the kinetics of the system when the rate of the standard pathway was low (e.g. when polymerisation at the end of the template or the unbinding of a polymer became slow). The leak trimerisation pathway circumvented binding at the end of the template. Hence, when the rate of leak polymerisation was slower than the rate of detachment at the end of the template, the leak had little effect on the kinetics. This placed a constraint on the toehold length and the end domain length, which had to be sufficiently short that the leak was slower than end unbinding, but not so short that polymerisation at the end of the template in the standard reaction pathway was not slowed by the occupancy of the final template site. This was achieved with shorter toeholds and longer information domains than were optimal for the system in the absence of the leak pathway.

The model presents that the early release of dimers is prevented if the leak reaction pathway is also slow relative to the standard pathway. The leak reaction rate and the pathway for dimer release involve a toehold-mediated strand displacement reaction. However, the leak pathway generally runs with, rather than against, the free-energetic landscape. Hence slowing the leak reaction relative to the standard pathway also slowed the dimer release pathway.

The backbone length made relatively little difference to the system's overall performance. The change in the entropic contribution to the free energy was small over the loop lengths sampled (20 to 40) and because the fastest systems were mostly rate limited by unbinding steps, the change in the branch migration rates over longer backbone domains that factor into the binding processes had little effect. A longer backbone would be preferable as, while it slightly reduces the stability of template-bound polymers, it provides space for extra C/C mismatches. Putting the mismatches too close together could cause bulges or unwanted side reactions.

The constraints on the input parameters described in this chapter should, at least in algebraic form, apply to an experimental implementation of the DNA strand displacement network. In parameterising the model, firstly, I assumed that the GC content of all domains, aside from the backbone domain, was equal and that the backbone domain was polyC or polyG, with infrequent C/C mismatches. Secondly, I assumed that the position

of the mismatches and the presence of mismatches that are unchanged by a strand displacement process had little effect on the kinetics, in a simplification of the true properties of these systems [156] (Indeed, members of the Ouldridge lab have recently started a project to better understand the kinetics of strand displacement with neutral, unchanged mismatches). Each of these phenomena will affect the kinetics of each reaction in the model. Therefore, I recommend that the rate constants and standard free energies be calculated based upon a detailed characterisation of the system in question, or at least by detailed calculation of the rate constants based upon the actual sequences being used. For instance, if the information domains and end domains have varying GC contents, the relationship between the bond length and the lifetime of the bound state will vary, necessitating shorter or longer domain lengths to satisfy the various constraints.

Careful placement of mismatches could enhance their impact on the kinetics [157]. Toehold-mediated strand displacement reactions in which the invader (incumbent) strand contains a mismatch proceed more slowly (quickly) if the mismatch is early in the branch migration domain than if the mismatch is later in the branch migration [157, 156]. Therefore placing the mismatches closer to the toehold could result in greater suppression of the rate of lock reinvasion, but could also speed up the rate of polymerisation and the leak reaction. Placing a mismatch healed in the polymerisation reaction close to the toehold increases the likelihood of fraying of the dsDNA ends, effectively increasing the accessible toehold length. Given that Cabello-Garcia's characterisation of HMSD indicates that the polymerisation process has a rate that saturates for toeholds of length 3nts or longer (see Fig. 4.7) and given typical toehold-mediated strand displacement reactions saturate with a toehold of length approximately 5nts, placing a mismatch close to a short toehold (of length about 2nts) could greatly speed up the rate of the leak reaction relative to polymerisation. Therefore, I suggest that mismatches healed in the polymerisation process should be placed far from the toehold, and mismatches healed during monomer binding should be placed far from the information domain and closer to the toehold. The latter would have the effect of reducing the likelihood of non-complementary monomer binding and would also further suppress the reinvasion of the lock strand. Care should be taken that the mismatch is not positioned so close to either backbone domain that it risks fraying the last few nucleotides, as such an event could lead to undesirable side reactions between strands. In practice, a G/C "clamp" sequence should be included at the end of the branch migration domains to prevent these sorts of unwanted fraying events.

As discussed in this chapter, the leak reaction pathway may be the main route to the loss of sequence and length control for this template copying reaction system. The leak reaction pathway involves non-template-directed polymerisation and was also shown to promote the release and rebinding of dimers. Ameliorating the leak reaction pathway in any experimental design is highly desirable. The effects of the leak reaction pathway could be measured experimentally through the use of fluorescently-labelled strands. Monomers 1, 2, and a

variation of monomer 3 that has a non-complementary information domain should be used. Dimer 23 and trimer 123 will only be able to form via the leak reaction pathway (non-template directed polymerisation). The formation of dimers 23 and trimers 123 should be quantified, with the use of a probe strand to avoid affecting the kinetics and thermodynamics of the reaction [22, 158, 159].

A natural question is how well the system will copy longer templates and whether there is a limit that can be set. In chapter 3, I demonstrated that if the rate of unbinding the leading edge of partially grown polymers was similar to the rate of unbinding at the end of the template, then simply an energetic difference that scaled with the logarithm of the template length was sufficient to promote the release of complete polymers over partially grown ones. However, when the occupancy of the end of the template became low enough that the polymerisation process at the end of the template was slower than the rate of release of partially grown polymers, length control was lost.

Thus, the work in chapter 3 presented a rather large window of the parameter space over which tight length control could be achieved. However, in the model considered in this chapter, the leak reaction pathway presents a new challenge. The overall rate of the leak reaction will grow with the length of the template n, and the number of distinct monomer types in solution m (e.g. m types of monomer 1, m types of monomer 2 and so on). In section 4.3.3, I argued that the leak reaction has a minimal effect on the kinetics when the speed of the standard pathway exceeds the rate of the leak pathway, which could be reduced to the constraint of Eqn.4.52 $k_{15}[\mathrm{M}] \ll k_8$, which says that the rate of detachment at the end of the template must be much faster than the rate of non-template direction polymerisation. If there are n template sites at which the leak can occur and m times as many monomers in solution, with a total concentration $m[M]$, then the constraint becomes $mnk_{15}[M] \ll k_8$. This suggests that to prevent the leak reaction, the rate of release at the end domain must be increased. However, increasing the rate of monomer detachment at the end of the template reduces the occupancy, and hence slows the rate of polymerisation at the end of the template, slowing the reaction as a whole. The constraint in Eqn. 4.41 $k_8 < k_{13}$ states that template-directed polymerisation must occur faster than the final monomer detaches from the template for the process to be efficient. Taken together this implies that the rate of unbinding at the end of the template must satisfy

$$mnk_{15}[\mathrm{M}] \ll k_8 < k_{13}\,. \tag{4.55}$$

If we are to allow 1 in 10^X, leak events for each trimer detachment event, the maximum template length n and number of monomer types m is approximately upper-bounded by

$$mn < \frac{k_{13}\,10^{-X}}{k_{15}[\mathrm{M}]}\,. \tag{4.56}$$

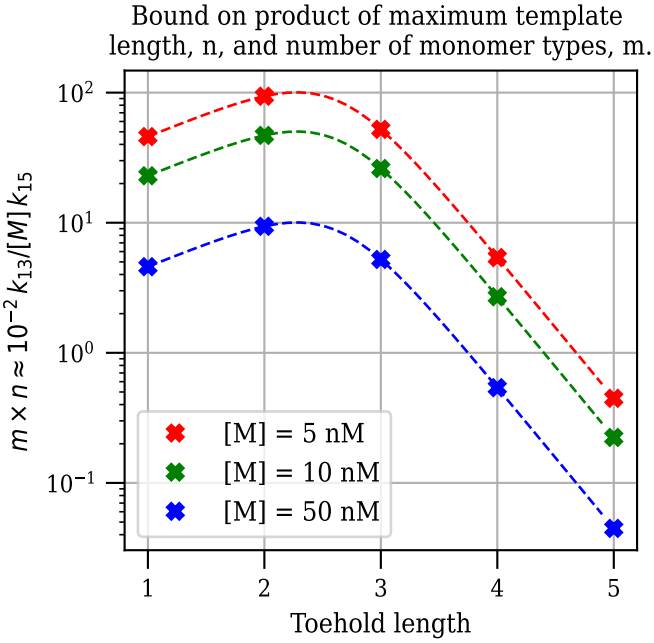


Figure 4.39: The product of the maximal template length n and the number of types of monomer m is proportional to the maximum information, I, that can be reliably copied by the system at an error rate of 1 leak event 1 for every 100 polymerisation events (X=2). I is constrained by the ratio of the leak reaction rate to the polymerisation rate. The optimal toehold length is approximately 2. Increasing the monomer concentration increases the total rate of the leak reaction pathway, and hence, decreases the maximum information that can be reliably copied. For a toehold of length 2, the maximum amount of information that can be reliably transferred is approximately 25 bits on a template of length 17 with 3 distinct symbols.

The equality of the condition in Eqn. 4.56 ($mn = \frac{k_{13}\, 10^{-X}}{k_{15}\,[\mathrm{M}]}$) is plotted in Fig. 4.39 for $X \approx 2$. This constraint is a measure of the maximal information content that can be stored and effectively copied by the reaction. The information in nats that can be encoded in a template of length n with m distinct symbols is $\ln(m^n)$. If the length of the template and the number of distinct symbols are constrained by $m \times n = c$, where the constant c is set by Eqn.4.56, the maximum informational content I that can be reliably copied at an 'error rate' of 1 leak event in 100 successful trimer releases can be estimated by solving

$$\frac{d}{dn} \ln\left((c/n)^n\right) = 0 \tag{4.57}$$

for n, which yields $I = \frac{c}{e}$, measured in nats. Hence, the constraint in Eqn.4.56 relates the polymerisation and leak reaction rates to the maximum information transfer that can be achieved in the copying operation. The toehold length that maximises the bound in Eqn.4.56 is 2, as seen in Fig. 4.39. For a monomer concentration $[M] \approx 10nM$, $m \times n \approx 50$, which suggests that the system would be capable of reliably copying 25 bits (binary) worth of information with $m = e \approx 3$ different types of monomers on a template of length $n \approx 50/3 \approx 17$, with one leak event per 100 trimers. This bound could be increased by decreasing the error tolerance X, decreasing the

monomer concentration $[M]$, or, perhaps, by incorporating error-checking steps into the polymer growth process that counteract erroneous events. A candidate motif, a version of HMSD incorporating a kinetic proofreading cycle that can discriminate between monomers with higher specificity, is being developed in the Ouldridge lab (unpublished).

This argument about the performance of the system on longer templates neglects that for longer the templates, the larger the product that the reaction forms, the greater the steric hindrance may become. For instance, if long DNA strands engulf the template, the rate of monomers binding to the template may decrease to such a point that the polymer growth process slows. Furthermore, the longer the product becomes, the greater the chance that a single link in a polymer breaks (potentially by spurious lock reinvasion). Therefore, longer templates will need more mismatches to exponentially suppress spontaneous polymer breaking. I note that in this chapter, the non-template-mediated interactions between copies have not been considered.

Simulating a system with a template length greater than 4 or 5 is infeasible with this analysis method. The construction of the model involves the enumeration of all possible conformations of n or fewer contiguous polymers bound to a template of length n, each of which is treated as an individual chemical species, the number of which conformations grows as a quadratic exponential of n. The number of transitions between these conformations grows even faster. The enumeration of all possible template configurations was desirable for this model as I sought to enable all possible pathways toward desirable or undesirable behaviour on a template of length 3, and, by varying parameters, I sought to maximise just the desirable performance. On longer templates, an approach that either reduces the number of transitions and chemical species in the chemical reaction network, by excluding low-likelihood states and transitions that are much slower than the timescale of interest, may be needed if the system is going to be solved as a series of coupled deterministic ODEs. Rewriting the computational CRN construction and using ODE solvers in C++ may reduce the memory overhead and the run-time by an order of magnitude or two, which would be sufficient to solve the ODEs for template lengths 4, 5, or 6. However, a trajectory-based approach, similar to that described in chapter 3, that has the advantage of low memory usage would be more appropriate for much longer templates.

Chapter 5

Conclusions

Processes, such as transcription, translation and DNA replication, in which a template copolymer directs the synthesis of a "copy" copolymer are fundamental to the molecular complexity of living systems that exist today [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. These biological systems can autonomously create long, complex copolymers, that eventually separate from their template, in stable environments driven by chemical fuel molecules such as ATP. This thesis aims to understand the properties of minimal systems of this kind from a computational and theoretical perspective and apply these insights towards creating a synthetic template copying system.

Earlier research in this field has focused on the accuracy of copy production, with older studies largely overlooking the thermodynamic consequences of product separation [109, 110, 111, 112, 113, 114, 115, 116]. More recently, the Ouldridge research group has investigated the impact of separation, demonstrating that, in constrast to systems without separation, accurate copying with product separation is only possible far from equilibrium [26, 23, 24, 27, 117]. However, these works assumed that the processes involved generate copies with a precisely defined length, and questions related to how the length distribution of copies can be controlled have gone unanswered. Product inhibition (the tendency for longer copies to have a higher affinity for the template than shorter copies) and substrate depletion pose challenges in controlling the length distribution for such systems, as both processes slow down the overall reaction. In Chapters 2, 3, and 4, I investigated the consequences of these effects and searched for mechanisms to overcome them.

5.1 Summaries

5.1.1 Chapter 2

To build a templated copying system, we must be able to quantify success. Therefore, in Chapter 2, I argued for the use of the time-averaged turnover frequency (TATF) as a suitable metric for comparing templated reactions that are affected by product inhibition and substrate depletion. I calculated the TATF for the following dimeric reactions, an ordered bi-uni reaction with competitive product inhibition, an ordered bi-uni reaction with mixed product inhibition, and a random order bi-uni reaction with competitive inhibition under the assumption that the processes were irreversible and that the dynamics of the external species (monomers and dimers) were slow compared to the dynamics of the template-bound species, such that quasi-steady-state rate laws held. I analyzed the TATF after 99% of the initial substrate concentration had been converted to product in low, intermediate, and high concentration regimes, each of which yielded approximations that were simple ratios of kinetic constants such as k_{cat} and various Michaelis constants. At low substrate concentrations, the near-complete conversion of substrate to product may not be possible, and therefore the approximations may break down.

I only calculated the TATF for dimeric systems and it could be approximated for longer-length templates with various product inhibition patterns. However, as the length of the system increases, the number of states that must be modelled and the complexity of the inhibition patterns necessarily increases too. Thus, extending the model may be challenging for trimeric and longer tempaltes. A related computational approach to exploring these state spaces was used in Chapter 4.

The TATF would serve as a useful metric for the evaluation of the performance of templated copying systems both in experimental and *in silico* settings, and could provide a simple route to the experimental measurement of kinetic constants. The utility of the approach presented in Chapter 2 is that it offers a method to reduce the complexity of the steady-state rate laws (many parameters) and cast different mechanisms into a comparable form that directly measures the desired performance of the system with just a few key parameters. The reduction in the degrees of freedom comes from assuming irreversibility, equimolar initial substrate concentrations, and the integral involved in calculating the TATF.

In future work, the TATF should be measured for various templated reactions. This would involve consolidating results from existing literature in this area and using the TATF to compare their performance. The approximations in the high, intermediate, and low concentration regimes should be tested. Secondly, an ideal copying system generates copies with a high average turnover and does so accurately. Therefore, a metric that gauges the system's specificity under substrate depletion and product inhibition effects would be valuable. As the substrates comprising the correct copy copolymer are depleted, a system with low discrimination between

monomer types may tend to produce more inaccurate copies as the reaction progresses. The TATF of individual products, and ratios thereof, might provide useful figures of merit. The TATF could be better approximated in the low substrate and non-irreversible limits, where incomplete conversion of substrate to product is expected.

5.1.2 Chapter 3

On longer templates with more binding sites, more complex mechanisms of product inhibition arise, making steady-state rate and, therefore, TATF calculations challenging, thus limiting the extent to which theory and analytical expressions could be used to explore the dynamics of these systems. Therefore, in chapter 3 (which is based upon the results presented in Ref. [1]), I took a different approach, and developed a bottom-up, stochastic model of a templated polymerisation system on a long template. I designed the model to search for mechanisms and regions of the parameter space in which length control could be achieved on very long templates. I stipulated a local set of transitions between local conformations and simulated the system with the Gillespie algorithm to evolve the system globally [128]. I assumed that the concentration of polymers in solution was infinitely dilute, such that no rebinding occurred, as would be a suitable initial condition in most experimental settings.

First, I focused on finding reaction mechanisms that could create copies with a well-defined length (identical length to the template). Secondly, I asked whether these mechanisms could be compatible with mechanisms to discriminate between monomers. I found that naive models of templated copolymerisation are scuppered by product inhibition, and no length control could be achieved. If the affinity of monomers for the template is high, either long stable polymers form on the template and shorter polymers form under their fraying edges, or, if the free energy of polymerisation is lower, polymers are unstable and only fragmented copies can be produced.

I introduced a bond-weakening mechanism, by which the energy released in polymerisation between copy monomers is directed into the disruption of the complementary interaction between one of the copy monomers (the left-most monomer in the polymerised pair was chosen arbitrarily) and the template. If this effect was strong enough, in certain regions of the parameter space, it induced a directionality to the system, wherein polymers adopted dense, brush-like conformations and were attached to the template only at their leading edge (right-most monomer in the chain). Additionally, weakening the copy-template bonds at the terminal site on the template, where full-length copies were bound, biased the release of full-length copies above the shorter fragments. These mechanisms, which rely on the formation of densely-packed conformations on the template, were a surprising discovery, as they are far removed from the idealised, sparsely-occupied mechanisms assumed in earlier models of autonomous templated copying with separation [23, 24, 27, 117]. Indeed, the set of challenges for sparse copying that were initially expected, such as the spurious nucleation of growth at unwanted sites on the template,

and control over the start and end points of polymer growth, have no bearing on the mechanisms identified in Chapter 3.

Having identified a mechanism that would generate only full-length copies, I introduced sequence-specific interactions to the model and asked how accurate copies could be generated. I simulated the model again with discrimination that modulated either the unbinding rates of monomers (off-rate discrimination) or the binding rates (on-rate discrimination). The length control mechanism requires that polymers be attached strongly by their leading edge, and thus, under off-rate discrimination, incorporating incorrect monomers into a chain sped up the rate of detachment of the chain, which tended to make the system produce fragmented polymers. On-rate discrimination, however, did not suffer this effect and was compatible with the length control mechanisms, which gave accuracy and precise length control.

5.1.3 Chapter 4

Finally, in Chapter 4, I asked how the mechanisms identified in Chapter 3 could be implemented in an explicit model of a DNA strand-displacement system. Previous experimental work has shown that handhold-mediated DNA strand displacement (HMSD) can autonomously copy dimeric templates with minimal product inhibition, driven only by a high free energy source of chemical fuel, the concentration of the monomers themselves [21]. However, whether such a system could copy much longer templates remains to be seen. To guide this effort, I developed a deterministic ODE model, parameterised it on the available experimental data [21] and used NUPACK for free energy calculations [153, 154, 155]. I modelled a system with a similar set of transitions to those featured in Chapter 3 but with the addition of the exchange of polymers of all lengths between the templates and the environment and certain "leak" reaction pathways that HMSD can be afflicted by. In this system, the main source of sequence inaccuracies is expected to be the leak pathway, not the direct binding of incorrect monomers to the template. Therefore, explicit sequence effects were neglected and the effect of the leak pathway on the kinetics was taken as a proxy for accuracy. Part of the challenge involved in this work was developing an efficient iterative computational method to construct the full state space of the model, which grew rapidly with the template length. To reduce the computational complexity of the model, I focused on modelling a trimeric template. Furthermore, a trimeric template is the shortest template for which length-control of the products (producing dimers as opposed to trimers) could present an issue. I explored the design space by varying domain lengths in the DNA-based system's model, changing the copier's kinetic and thermodynamic properties.

I assessed the performance using four metrics: trimer yield, trimer production rate, the impact of the leak reaction pathway on kinetics (a proxy for accuracy), and the propensity for the system to release dimers (a

measure of length control). I analysed which design parameters (domain lengths, mismatch positions, etc.) independently optimised each metric. Subsequently, I determined which parameters optimised performance across all metrics, grounding these results in constraints on the design space for ideal copiers, thereby I have proposed explicit designs which are currently being tested experimentally. Eliminating the leak reaction through careful domain length selection also removed the system's propensity to release dimers. The cumulative rate of the leak reaction pathways increases proportionally with the template length and the number of distinct monomer types in solution and hence poses a challenge for copying long, information-dense templates. By extrapolating the constraints on performance, I estimated that this system should be able to copy approximately 25 bits of information (achievable with a ternary alphabet of monomers and a 17-unit long template) at an error rate of 1 leak event per 100 trimers produced.

The number and complexity of the ODEs in the model grow rapidly with the template's length; hence, simulating systems with template lengths greater than 4 is infeasible with this current version. A variety of techniques could be used to increase the effectiveness of the analysis approach. While rewriting the simulation in a more efficient language, such as C++, could provide 100x speed and memory overhead improvements, this change may only result in being able to simulate templates of length 5 or 6. Therefore, reducing the model's state space by excluding certain highly improbable conformations and transitions or employing a trajectory-based approach, as was used in Chapter 3, may prove more effective strategies. I note here, that had I not adopted an approach to modelling these systems that didn't enforce a preconceived global dynamics (the models were built up from simple, local rules), then I would not have been able to identify the key mechanisms in Chapter 3 and 4. Assuming that polymers could gradually unzip from the template, rather than just directly detach by breaking multiple copy-template bonds, was key. Therefore, care should be taken in reducing the set of permissible transitions.

5.2 Outlook

5.2.1 Trimers and self-replicators

An exciting chapter in this nascent field of autonomous chemical copying is just beginning; these systems are being built. The findings of Chapter 4 have served as a starting point for design of a trimeric copying system which is now being tested in the lab. Experimental validation will be a valuable test of the model presented in the chapter. Better access to experimental data could be used to update the model, refit the parameters, provide mechanistic insights into observed phenomena, and iterate the system design again. Further still, self-replicating dimeric systems, in which each copy can also serve as a template, that operate truly autonomously are being

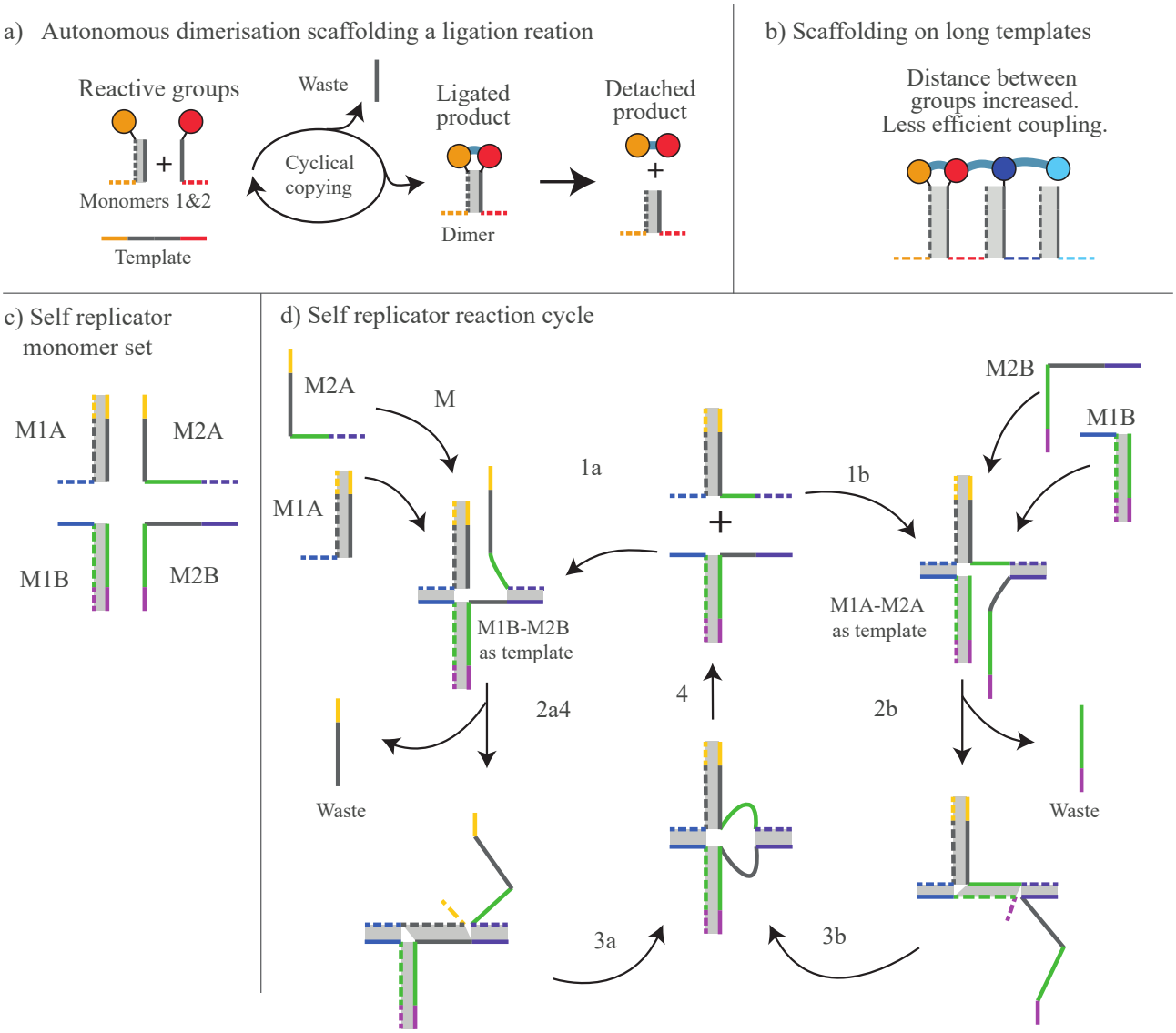


Figure 5.1: a) A templated dimerisation reaction could scaffold the ligation of two reactive groups with minimal product inhibition. The template could be a biomarker for a disease state. b) The scaffolding of longer molecules may be challenging if the reactive groups are not brought into close enough proximity. c) Four monomers with an additional "handhold" domain are required to create a self-replicator. The monomers correspond to two families A and B, that template one another. The mismatches required to drive the reaction are not shown for clarity. d) Adapted from Fehily's thesis [160]. Step 1(a,b), the monomers of each family template the dimerisation of the other family. Step 2(a,b), the monomers bind to the templates and the lock strand is displaced as waste. Step 3(a,b), the monomers polymerise to form a dimer. Step 4, the monomers detach from the parent strand and can template further reactions.

tested. A design and an illustrative replication cycle is shown in Fig.5.1 c and d. Early characterisations by Luke Fehily, a Masters student in the Ouldridge group, indicate that the system is capable of generating a sigmoidal growth curve [160], a key indicator of autocatalysis [161]. This work could serve as a basis for understanding the dynamics of evolution in a simple, model system. Again, as was the case with copiers, it is unclear whether such a system could scale to self-replicate on long templates, and there may be a need for modelling the system using the methods of this thesis. I co-supervised another Masters student, Nick Simpson, in generalising the model presented in Chapter 3 to simulate self-replicating systems such as this, the source code for which is available here [162].

5.2.2 Scaffolding the production of non-nucleic products

Modifying the copying architecture to chaperone the assembly of non-nucleic acid molecules, echoing the operation of the ribosome, would offer a route to making these systems more functional. The dynamics of these systems are measured by tagging the DNA strands with fluorophores and quencher molecules [22, 21]. Instead, tagging the strands with reactive groups could enable templated ligation of non-nucleic molecules [71], as shown in Figure 5.1 a. An autonomous dimerisation system could trigger the formation of a therapeutic molecule in the presence of a template that is a biomarker of a disease state in a manner that is less impeded by the effects of product inhibition than systems have been in the past [69, 70, 64, 73, 64]. Indeed, an autonomous system that incorporates kinetic proofreading to enhance the sequence sensitivity of a templated reaction has recently been developed in the Ouldridge lab (unpublished), which could be a suitable basis for such a system. I could also foresee the self-replicating system deployed, perhaps with sensitivity-enhancing kinetic proofreading, to template the production of a molecule in response to a biomarker, as the autocatalytic growth of the system may provide a superior signal to noise relative to standard copying systems. As shown in Figure 5.1 b, chaperoning the assembly of molecules from even longer templates could open routes toward the reliable, programmatic production and exploration of the large chemical space of sequence defined-copolymers, a “holy-grail” of polymer science [55]. Ensuring the spatial proximity of the reactive groups on templates of length three or greater may be one challenge to overcome.

5.2.3 Autonomous templated computation

Copying is a computational operation that simply correlates two physically separated objects. A more comprehensive class of computations could be achieved by modifying the architecture to transfer information along the “backbone” of the growing copolymer, as demonstrated in Fig. 5.2. This abstract representation of a

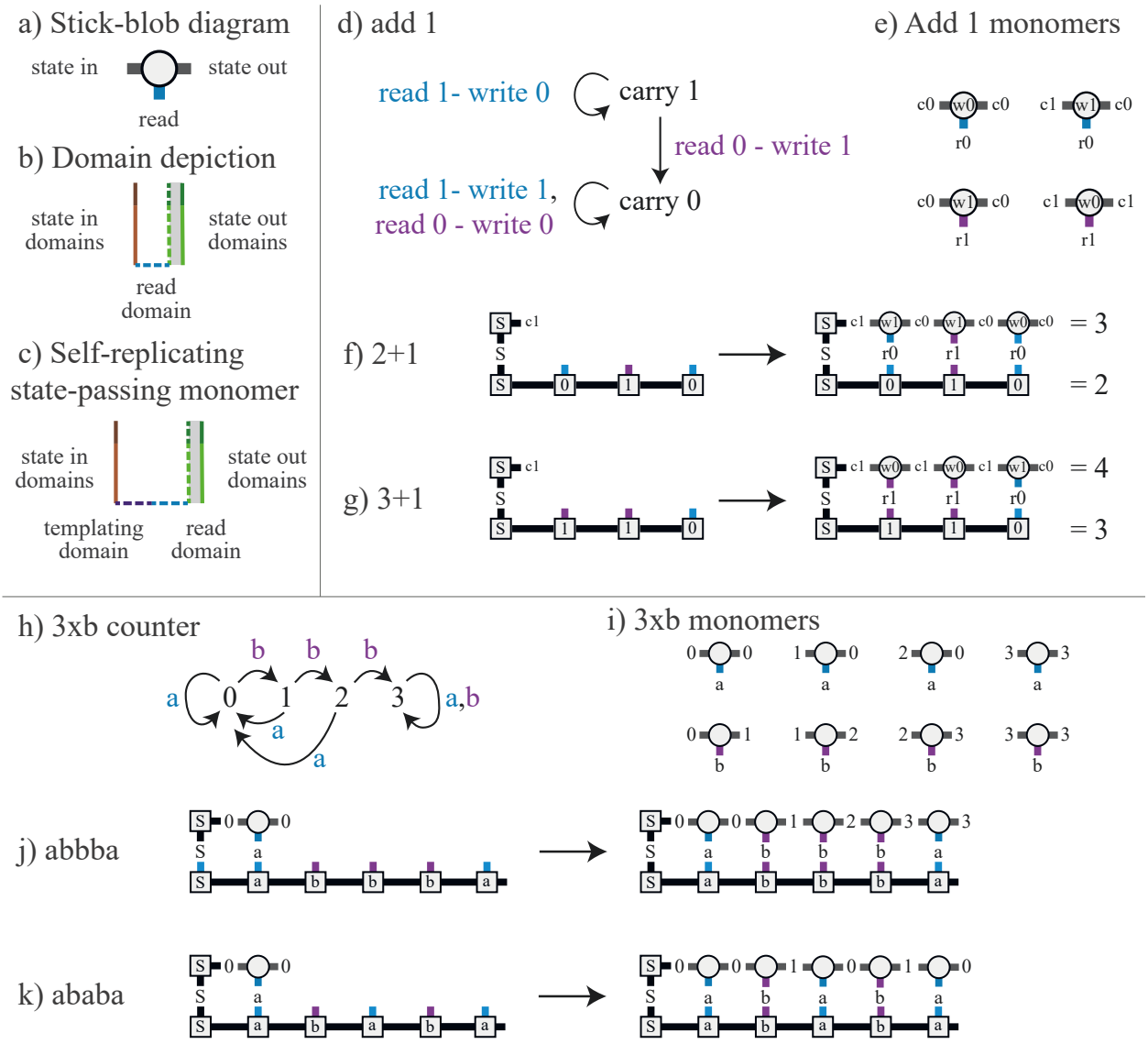


Figure 5.2: Copying, the focus of this thesis, is a simple type of computation. A more complex function on the template input could be achieved by introducing information/state passing domains in the backbone domains. **a)** An abstract representation of a state-passing monomer. **b)** A domain-level depiction of a state-passing monomer capable of HMSD. **c)** Given an extra templating domain, the state-passing monomer may also be able to act as a template, thereby creating a system that computes across self-replicating generations. A similar system without the ability to spontaneously separate products is described in Ref. [45]. **d)** A finite state diagram of a process that can read a template and write to another "output" copolymer with a single bit of internal memory, to add 1 to an input string in binary. **e)** Monomers required to synthesise a copolymer with a sequence that is 1 greater than the template sequence in binary. **f)** Adding 1 to a template that contains 2 in binary (reading with smaller indices going from left to right). The S indicates a start monomer, which may or may not be required to nucleate the computation in practice. For clarity, I have shown the systems remaining deposited on the template. If the mechanisms investigated in this thesis were to be used, I would expect that the layers of computation would eventually detach the template. **g)** Adding 1 to 3 encoded in binary. **h)** A finite state diagram of a process that counts whether there are three contiguous b's in the template string. If an a is read, the count resets to 0, else increments by 1, until three b's have been read. **i)** The monomer set required to implement the process in h. **j)** The template with sequence abbba contains three contiguous b's, and hence, the terminal monomer reaches the state 3. **k)** The sequence ababa does not contain three contiguous b's and, hence, the terminal monomer does not reach state 3.

computing system could constitute a single step in an iterative boolean circuit, a form of computation previously implemented in DNA tile-based systems [45]. A key distinction is that if this system were implemented with the mechanisms explored in this thesis, the products of the computation would autonomously separate from the input template, allowing the template to reseed the computation. Multiple layers of computation, akin to an autocatalytic computer, could be possible if a further templating domain were introduced to the monomers, as illustrated in Fig. 5.2c. The computation requires that a monomer be polymerised only if it is complementary to the template and the previous monomer in the chain. The question of how the output of such a computation would be measured is difficult. Ensuring that the desired endpoint of a computation can only self-replicate, rather than compute further, may enhance the concentration of correct outputs against incorrect ones through competition. This minimal system would serve as yet another testbed for exploring our theoretical and practical understanding of molecular computation.

5.2.4 Stochastic thermodynamics of length-control

Finally, interesting theoretical questions about the far-from-equilibrium nature of controlling the length distribution of copies remain open. In previous work on finite templates, Poulton demonstrated that the overall thermodynamics of polymer copying with separation was controlled by the exchange of matter with baths, but did not consider the production of fragmented copies [24, 27]. When considering that products of all lengths could be produced, it is yet more entropically challenging to produce a molecule with the exact sequence of the template. A detailed treatment of this process using the tools of stochastic thermodynamics could ask questions of the entropy production required and, consequently, the efficiency with which a flux of length and sequence controlled products is maintained against a mixed bath of products [163, 164, 26, 24]. Further, the stochastic thermodynamic properties of systems containing multiple competing templates could also be explored.

5.3 Reflection

Minimal models of physical systems are powerful in their simplicity. When constructing such a model, we choose to represent only the essential features, in the hope that a nimble and focused model will lead us to a clear perspective into the principles and mechanisms driving whichever system has caught our interest. In this thesis, that system was the foundation of biomolecular complexity – the information-copying machinery of life.

These systems have ancient origins and are terribly intricate. Ribosomes, for instance, are incredibly complex, show great variety across the taxa of life, are implicated in many cellular processes, and have been described as "an ancient molecular fossil that provides a telescope to the origins of life" [Bowman et al. [17]] [15, 16, 54, 19]. Minimal models omit the specifics of these particular systems for the purpose of asking deeper questions of the underlying physical principles that govern their operation. The convention in the field has been to focus just on the operation of templated copying, the process of growing a copolymer under the guidance of a template [109, 110, 111, 112, 113, 114, 115, 116, 26, 23, 24, 27, 1, 117].

Previous work in this area [109, 110, 111, 112, 113, 114, 115, 116] had overlooked the thermodynamic consequences of separation of the copy and the template [26]. The extant biological copying machines operate in constant environments, powered by chemical fuel molecules, and can produce accurate copies of templates that eventually separate from their template [26]. More recent models have accounted for separation, but focused only on sequence control of the process, not the mechanisms for length control of products [26, 23, 24, 27, 117]. In this thesis, I focused on elucidating molecular mechanisms that could deliver ideal copies, with minimal production of inaccurate or fragmented copies in a non-enzymatic setting. The principles that I have uncovered, with the contribution and support of many talented colleagues and mentors, are helping to guide the design of a novel class of synthetic molecular information copiers, that may, soon, copy as life does – autonomously.

Bibliography

[1] J. Juritz, J. M. Poulton, and T. E. Ouldridge. “Minimal mechanism for cyclic templating of length-controlled copolymers under isothermal conditions”. In: *The Journal of Chemical Physics* 156.7 (2022). Link.

[2] B. Alberts, A. Johnson, J. Lewis, D. Morgan, and M. Raff. *Molecular Biology of the Cell*. Taylor & Francis Group, 2014.

[3] L. Brocchieri and S. Karlin. “Protein length in eukaryotic and prokaryotic proteomes”. In: *Nucleic Acids Research* 33.10 (2005). Link.

[4] M. P. Stumpf, T. Thorne, E. De Silva, R. Stewart, J. A. Hyeong, M. Lappe, and C. Wiuf. “Estimating the size of the human interactome”. In: *PNAS* 105.19 (2008). Link.

[5] UniProt. Size of the UniProt database of protein sequences. Link [Accessed: 14/04/2023]. 2023.

[6] S. D. Bell. “Prime-time progress”. In: *Nature 2006 439:7076* 439.7076 (2006). Link.

[7] C. O’Connor and J. U. Adams. *Essentials of Cell Biology*. Link. 2004.

[8] S. P. Bell and A. Dutta. “DNA Replication in Eukaryotic Cells”. In: *Annual Review of Biochemistry* 71.1 (2003). Link, pp. 333–374.

[9] L. J. Reha-Krantz. “DNA polymerase proofreading: Multiple roles maintain genome stability”. In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1804.5 (2010). Link.

[10] C. M. Joyce and S. J. Benkovic. “DNA Polymerase Fidelity: Kinetics, Structure, and Checkpoints”. In: *Biochemistry* 43.45 (2004). Link.

[11] A. Vannini. “A structural perspective on RNA polymerase I and RNA polymerase III transcription machineries”. In: *Biochimica et biophysica acta* 1829.3-4 (2013). Link.

[12] R. H. Ebright. “RNA polymerase: Structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II”. In: *Journal of Molecular Biology* 304.5 (2000).

[13] Y. Hirose and J. L. Manley. “RNA polymerase II and the integration of nuclear events”. In: *Genes and Development* 14.12 (2000). Link.

[14] M. V. Rodnina. “Translation in Prokaryotes”. In: *Cold Spring Harbor Perspectives in Biology* 10.9 (2018). Link.

[15] G. E. Fox. “Origin and Evolution of the Ribosome”. In: *Cold Spring Harbor Perspectives in Biology* 2.9 (2010). Link.

[16] A. S. Petrov et al. “History of the ribosome and the origin of translation”. In: *PNAS* 112.50 (2015). Link.

[17] J. C. Bowman, A. S. Petrov, M. Frenkel-Pinter, P. I. Penev, and L. D. Williams. “Root of the Tree: The Significance, Evolution, and Origins of the Ribosome”. In: *Chemical Reviews* 120.11 (2020). Link.

[18] R. Nikolay, D. van den Bruck, J. Achenbach, and K. H. Nierhaus. “Ribosomal Proteins: Role in Ribosomal Functions”. In: *eLS*. Link. John Wiley & Sons, Ltd, 2015.

[19] G. Caetano-Anollés and D. Caetano-Anollés. “Computing the origin and evolution of the ribosome from its structure - Uncovering processes of macromolecular accretion benefiting synthetic biology.” In: *Computational and structural biotechnology journal* 13 (2015). Link.

[20] F. Crick. “Central dogma of molecular biology”. In: *Nature* 227.5258 (1970). Link.

[21] J. Cabello-Garcia. “Handhold-mediated strand displacement: a mechanism for non-equilibrium and catalytic templating”. Link. PhD thesis. Imperial College London, 2022.

[22] J. Cabello-Garcia, W. Bae, G. B. V. Stan, and T. E. Ouldridge. “Handhold-Mediated Strand Displacement: A Nucleic Acid Based Mechanism for Generating Far-from-Equilibrium Assemblies through Templated Reactions”. In: *ACS Nano* (2021). Link.

[23] J. M. Poulton, P. R. ten Wolde, and T. E. Ouldridge. “Nonequilibrium correlations in minimal dynamical models of polymer copying.” In: *PNAS* 116.6 (2019). Link.

[24] J. M. Poulton and T. E. Ouldridge. “Edge-effects dominate copying thermodynamics for finite-length molecular oligomers”. In: *New Journal of Physics* 23.6 (2021). Link.

[25] W. Kühlbrandt. “Structure and Mechanisms of F-Type ATP Synthases”. In: *Annual Review of Biochemistry* 88.1 (2019). Link.

[26] T. E. Ouldridge and P. R. ten Wolde. “Fundamental Costs in the Production and Destruction of Persistent Polymer Copies”. In: *Physical Review Letters* 118.15 (2017). Link.

[27] J. Poulton. “The long and short of template copying”. Link. PhD thesis. Imperial College London, 2020.

[28] A. Murugan, Z. Zeravcic, M. P. Brenner, and S. Leibler. “Multifarious assembly mixtures: Systems allowing retrieval of diverse stored structures”. In: *PNAS* 112.1 (2015). Link.

[29] T. Yamaguchi, F. Makino, T. Miyata, T. Minamino, T. Kato, and K. Namba. “Structure of the molecular bushing of the bacterial flagellar motor”. In: *Nature Communications 2021 12:1* 12.1 (2021). Link.

[30] T. P. Wörner, A. Bennett, S. Habka, J. Snijder, O. Friese, T. Powers, M. Agbandje-McKenna, and A. J. R. Heck. “Adeno-associated virus capsid assembly is divergent and stochastic”. In: *Nature Communications* 12.1 (2021). Link.

[31] L. Thomas. Molecule of the Month: ATP Synthase. Link [Accessed: 14/04/2023]. 2005.

[32] J. Tan et al. “Structural basis of assembly and torque transmission of the bacterial flagellar motor”. In: *Cell* 184.10 (2021). Link.

[33] D. Goodsell. Molecule of the Month: Bacteriophage phiX174. Link [Accessed: 14/04/2023]. 2000.

[34] T. Dokland, R. Bernal, A. Burch, S. Pletnev, B. Fane, and M. Rossmann. Procapsid of Bacteriophage PhiX174. Link [Accessed: 14/04/2023]. 1999.

[35] P. Sartori and S. Leibler. “Lessons from equilibrium statistical physics regarding the assembly of protein complexes”. In: *PNAS* 117.1 (2020). Link.

[36] N. C. Seeman and H. F. Sleiman. “DNA nanotechnology”. In: *Nature Reviews Materials* 3.1 (2017). Link.

[37] A. Murugan, D. A. Huse, and S. Leibler. “Speed, dissipation, and error in kinetic proofreading”. In: *PNAS* 109.30 (2012). Link.

[38] N. C. Seeman. “Nucleic acid junctions and lattices”. In: *Journal of Theoretical Biology* 99.2 (1982). Link.

[39] P. W. Rothemund. *Folding DNA to create nanoscale shapes and patterns*. Link. 2006.

[40] S. Dey et al. “DNA origami”. In: *Nature Reviews Methods Primers* 1.1 (2021). Link.

[41] P. Zhan et al. “Recent Advances in DNA Origami-Engineered Nanomaterials and Applications”. In: *Chemical Reviews* (2023). Link.

[42] C. Lin, Y. Liu, S. Rinker, and H. Yan. “DNA Tile Based Self-Assembly: Building Complex Nanoarchitectures”. In: *ChemPhysChem* 7.8 (2006). Link.

[43] Y. Ke, L. L. Ong, W. M. Shih, and P. Yin. “Three-dimensional structures self-assembled from DNA bricks”. In: *Science* 338.6111 (2012). Link.

[44] L. L. Ong et al. “Programmable self-assembly of three-dimensional nanostructures from 10,000 unique components”. In: *Nature 2017 552:7683* 552.7683 (2017). Link.

[45] D. Woods, D. Doty, C. Myhrvold, J. Hui, F. Zhou, P. Yin, and E. Winfree. “Diverse and robust molecular algorithms using reprogrammable DNA self-assembly”. In: *Nature 2019 567:7748* 567.7748 (2019). Link.

[46] J. K. Leman et al. “Macromolecular modeling and design in Rosetta: recent methods and frameworks”. In: *Nature Methods* 17.7 (2020). Link.

[47] I. Anishchenko et al. “De novo protein design by deep network hallucination”. In: *Nature 2021 600:7889* 600.7889 (2021). Link.

[48] R. Divine et al. “Designed proteins assemble antibodies into modular nanocages”. In: *Science* 372.6537 (2021). Link.

[49] H. Shen et al. “De novo design of self-assembling helical protein filaments”. In: *Science* 362.6415 (2018). Link.

[50] N. H. Kim, H. Choi, Z. M. Shahzad, H. Ki, J. Lee, H. Chae, and Y. H. Kim. “Supramolecular assembly of protein building blocks: from folding to function”. In: *Nano Convergence* 9.1 (2022). Link.

[51] J. Jin, E. G. Baker, C. W. Wood, J. Bath, D. N. Woolfson, and A. J. Turberfield. “Peptide Assembly Directed and Quantified Using Megadalton DNA Nanostructures”. In: *ACS Nano* 13.9 (2019). Link.

[52] L. Li and G. Chen. “Precise Assembly of Proteins and Carbohydrates for Next-Generation Biomaterials”. In: *Journal of the American Chemical Society* 144.36 (2022). Link.

[53] R. Feynman. Image of Feynman’s blackboard. Link [Accessed: 14/04/2023]. 2022.

[54] I. Nanikashvili, Y. Zarai, A. Ovseevich, T. Tuller, and M. Margaliot. “Networks of ribosome flow models for modeling and analyzing intracellular traffic”. In: *Scientific Reports* 9.1 (2019). Link.

[55] J.-F. Lutz. “Sequence-controlled polymerizations: the next Holy Grail in polymer science?” In: *Polymer Chemistry* 1.1 (2010). Link.

[56] D. L. Usanov, A. I. Chan, J. P. Maianti, and D. R. Liu. “Second-generation DNA-templated macrocycle libraries for the discovery of bioactive small molecules”. In: *Nature Chemistry* 10.7 (2018). Link.

[57] R. K. O’Reilly, A. J. Turberfield, and T. R. Wilks. “The Evolution of DNA-Templated Synthesis as a Tool for Materials Discovery”. In: *Accounts of Chemical Research* 50.10 (2017). Link.

[58] J.-F. Lutz, M. Ouchi, D. R. Liu, and M. Sawamoto. “Sequence-Controlled Polymers”. In: *Science* 341.6146 (2013). Link.

[59] J. D. Neve, J. J. Haven, L. Maes, and T. Junkers. “Sequence-definition from controlled polymerization: the next generation of materials”. In: *Polymer Chemistry* 9.38 (2018). Link.

[60] E. Laurent, R. Szweda, and J.-F. Lutz. “Synthetic Polymers with Finely Regulated Monomer Sequences: Properties and Emerging Applications”. In: *Macromolecular Engineering* (2022). Link.

[61] G. Gody, P. B. Zetterlund, S. Perrier, and S. Harrisson. “The limits of precision monomer placement in chain growth polymerization”. In: *Nature Communications 2016 7:1* 7.1 (2016). Link.

[62] S. Samokhvalova and J.-F. Lutz. “Macromolecular Information Transfer”. In: *Angewandte Chemie* 135.16 (2023). Link.

[63] Y. He and D. R. Liu. “Autonomous multistep organic synthesis in a single isothermal solution mediated by a DNA walker”. In: *Nature Nanotechnology* 5.11 (2010). Link.

[64] J. Michaelis, A. Roloff, and O. Seitz. “Amplification by nucleic acid-templated reactions”. In: *Org. Biomol. Chem.* 12.18 (2014). Link.

[65] G. von Kiedrowski. *Minimal Replicator Theory I: Parabolic Versus Exponential Growth*. Link. Springer, Berlin, Heidelberg, 1993.

[66] A. Luther, R. Brandsch, and G. V. Kiedrowski. “Surface-promoted replication and exponential amplification of DNA analogues”. In: *Nature 1998 396:6708* 396.6708 (1998). Link.

[67] R. Issac and J. Chmielewski. “Approaching exponential growth with a self-replicating peptide”. In: *Journal of the American Chemical Society* 124.24 (2002). Link.

[68] P. G. Moerman, M. Gavrilov, T. Ha, and R. Schulman. “Catalytic DNA Polymerization Can Be Expedited by Active Product Release”. In: *Angewandte Chemie International Edition* 61.24 (2022). Link.

[69] A. Shibata, H. Abe, and Y. Ito. “Oligonucleotide-Templated Reactions for Sensing Nucleic Acids”. In: *Molecules* 17.3 (2012). Link.

[70] M. D. Pisa and O. Seitz. “Nucleic Acid Templated Reactions for Chemical Biology”. In: *ChemMedChem* 12.12 (2017). Link.

[71] O. Seitz. “Templated chemistry for bioorganic synthesis and chemical biology”. In: *Journal of Peptide Science* 25.7 (2019). Link.

[72] L. B. Pellejero, M. A. D. Nijenhuis, F. Ricci, and K. V. Gothelf. “Protein-Templated Reactions Using DNA-Antibody Conjugates”. In: *Small* 19.13 (2023). Link.

[73] H. Abe and E. T. Kool. “Destabilizing universal linkers for signal amplification in self-ligating probes for RNA”. In: *Journal of the American Chemical Society* 126.43 (2004). Link.

[74] A. Roloff and O. Seitz. “Reducing product inhibition in nucleic acid-templated ligation reactions: DNA-templated cycligation”. In: *ChemBioChem* 14.17 (2013). Link.

[75] A. Shibata, T. Uzawa, Y. Nakashima, M. Ito, Y. Nakano, S. Shuto, Y. Ito, and H. Abe. “Very Rapid DNA-Templated Reaction for Efficient Signal Amplification and Its Steady-State Kinetic Analysis of the Turnover Cycle”. In: *Journal of the American Chemical Society* 135.38 (2013). Link.

[76] D. Chang, K. T. Kim, E. Lindberg, and N. Winssinger. “Accelerating Turnover Frequency in Nucleic Acid Templated Reactions”. In: *Bioconjugate Chemistry* 29.1 (2018). Link.

[77] A. Osuna Gálvez and J. W. Bode. “Traceless Templated Amide-Forming Ligations”. In: *Journal of the American Chemical Society* 141.22 (2019). Link.

[78] D. Braun. “PCR by thermal convection”. In: *Modern Physics Letters B* 18.16 (2004). Link.

[79] X. He, R. Sha, R. Zhuo, Y. Mi, P. M. Chaikin, and N. C. Seeman. “Exponential growth and selection in self-replicating materials from DNA origami rafts”. In: *Nature Materials* 16.10 (2017). Link.

[80] R. Zhuo, F. Zhou, X. He, R. Sha, N. C. Seeman, and P. M. Chaikin. “Litters of self-replicating origami cross-tiles”. In: *PNAS* 116.6 (2019). Link.

[81] D. Núñez-Villanueva, M. Ciaccia, G. Iadevaia, E. Sanna, and C. A. Hunter. “Sequence information transfer using covalent template-directed synthesis”. In: *Chemical Science* 10.20 (2019). Link.

[82] D. Núñez-Villanueva and C. A. Hunter. “Molecular replication using covalent base-pairs with traceless linkers”. In: *Organic & Biomolecular Chemistry* 17.44 (2019). Link.

[83] D. Núñez-Villanueva and C. A. Hunter. “Replication of Sequence Information in Synthetic Oligomers”. In: *Accounts of Chemical Research* 54.5 (2021). Link.

[84] D. Núñez-Villanueva and C. A. Hunter. “Controlled mutation in the replication of synthetic oligomers”. In: *Chemical Science* 12.11 (2021). Link.

[85] A. Kühnlein, S. A. Lanzmich, and D. Braun. “tRNA sequences can assemble into a replicator”. In: *eLife* 10 (2021). Link.

[86] T. Li and K. C. Nicolaou. “Chemical self-replication of palindromic duplex DNA”. In: *Nature* 369.6477 (1994). Link.

[87] R. Schulman, B. Yurke, and E. Winfree. “Robust self-replication of combinatorial information via crystal growth and scission”. In: *PNAS* 109.17 (2012). Link.

[88] C. B. Mast and D. Braun. “Thermal trap for DNA replication”. In: *Physical Review Letters* 104.18 (2010). Link.

[89] C. B. Mast, S. Schink, U. Gerland, and D. Braun. “Escalation of polymerization in a thermal gradient”. In: *PNAS* 110.20 (2013). Link.

[90] M. Kreysing, L. Keil, S. Lanzmich, and D. Braun. “Heat flux across an open pore enables the continuous replication and selection of oligonucleotides towards increasing length”. In: *Nature Chemistry* 7.3 (2015). Link.

[91] M. Colomb-Delsuc, E. Mattia, J. W. Sadownik, and S. Otto. “Exponential self-replication enabled through a fibre elongation/breakage mechanism”. In: *Nature Communications* 6.1 (2015). Link.

[92] J. M. A. Carnall, C. A. Waudby, A. M. Belenguer, M. C. A. Stuart, J. J.-P. Peyralans, and S. Otto. “Mechanosensitive Self-Replication Driven by Self-Organization”. In: *Science* 327.5972 (2010). Link.

[93] J. A. Baross and S. E. Hoffman. “Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life”. In: *Origins of life and evolution of the biosphere 1985 15:4* 15.4 (1985). Link.

[94] A. Salditt, L. M. R. Keil, D. P. Horning, C. B. Mast, G. F. Joyce, and D. Braun. “Thermal Habitat for RNA Amplification and Accumulation”. In: *Physical Review Letters* 125.4 (2020). Link.

[95] L. E. Orgel. “Evolution of the genetic apparatus”. In: *Journal of Molecular Biology* 38.3 (1968). Link.

[96] F. H. Crick. “The origin of the genetic code”. In: *Journal of Molecular Biology* 38.3 (1968). Link.

[97] W. Gilbert. “Origin of life: The RNA world”. In: *Nature* 319.6055 (1986). Link.

[98] G. F. Joyce. “RNA evolution and the origins of life”. In: *Nature* 338.6212 (1989). Link.

[99] H. Mutschler, A. Wochner, and P. Holliger. “Freeze-thaw cycles as drivers of complex ribozyme assembly”. In: *Nature Chemistry* 7.6 (2015). Link.

[100] L. E. Orgel. “Prebiotic Chemistry and the Origin of the RNA World”. In: *Critical Reviews in Biochemistry and Molecular Biology* 39.2 (2004). Link.

[101] J. Attwater, A. Wochner, and P. Holliger. “In-ice evolution of RNA polymerase ribozyme activity”. In: *Nature Chemistry* 5.12 (2013). Link.

[102] R. Ball and J. Brindley. “Hydrogen peroxide thermochemical oscillator as driver for primordial RNA replication”. In: *Journal of The Royal Society Interface* 11.95 (2014). Link.

[103] J. W. Szostak. “The eightfold path to non-enzymatic RNA replication”. In: *Journal of Systems Chemistry 2012 3:1* 3.1 (2012). Link.

[104] P. W. Kudella, A. V. Tkachenko, A. Salditt, S. Maslov, and D. Braun. “Structured sequences emerge from random pool when replicated by templated ligation”. In: *PNAS* 118.8 (2021). Link.

[105] J. H. Rosenberger, T. Göppel, P. W. Kudella, D. Braun, U. Gerland, and B. Altaner. “Self-Assembly of Informational Polymers by Templated Ligation”. In: *Physical Review X* 11.3 (2021). Link.

[106] T. Göppel, J. H. Rosenberger, B. Altaner, and U. Gerland. “Thermodynamic and Kinetic Sequence Selection in Enzyme-Free Polymer Self-Assembly inside a Non-equilibrium RNA Reactor”. In: *Life* 12.4 (2022). Link.

[107] L. E. Orgel. “Unnatural selection in chemical systems”. In: *Accounts of chemical research* 28.3 (1995). Link.

[108] T. A. Lincoln and G. F. Joyce. “Self-sustained replication of an RNA enzyme”. In: *Science* 323.5918 (2009). Link.

[109] C. H. Bennett. “Dissipation-error tradeoff in proofreading”. In: *BioSystems* 11.2-3 (1979). Link.

[110] M. Ehrenberg and C. Blomberg. “Thermodynamic constraints on kinetic proofreading in biosynthetic pathways”. In: *Biophysical Journal* 31.3 (1980). Link.

[111] F. Cady and H. Qian. “Open-system thermodynamic analysis of DNA polymerase fidelity”. In: *Physical Biology* 6.3 (2009). Link.

[112] D. Andrieux and P. Gaspard. “Nonequilibrium generation of information in copolymerization processes”. In: *PNAS* 105.28 (2008). Link.

[113] P. Gaspard and D. Andrieux. “Kinetics and thermodynamics of first-order Markov chain copolymerization”. In: *Journal of Chemical Physics* 141.4 (2014). Link.

[114] P. Sartori and S. Pigolotti. “Kinetic versus energetic discrimination in biological copying”. In: *Physical Review Letters* 110.18 (2013). Link.

[115] P. Sartori and S. Pigolotti. “Thermodynamics of Error Correction”. In: *Physical Review X* 5.4 (2015). Link.

[116] M. Esposito, K. Lindenberg, and C. Van Den Broeck. “Extracting chemical energy by growing disorder: Efficiency at maximum power”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2010.1 (2010). Link.

[117] B. Qureshi, J. Juritz, J. M. Poulton, A. Beersing-Vasquez, and T. E. Ouldridge. “A universal method for analyzing copolymer growth”. In: *The Journal of Chemical Physics* 158.10 (2023). Link.

[118] J. J. Hopfield. “Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity”. In: *PNAS* 71.10 (1974). Link.

[119] J. Ninio. “Kinetic amplification of enzyme discrimination”. In: *Biochimie* 57.5 (1975). Link.

[120] R. Plasson, A. Brandenburg, L. Jullien, and H. Bersini. “Autocatalysis: At the Root of Self-Replication”. In: *Artificial Life* 17.3 (2011). Link.

[121] M. P. Robertson and G. F. Joyce. “The Origins of the RNA World”. In: *Cold Spring Harbor Perspectives in Biology* 4.5 (2012).

[122] A. V. Tkachenko and S. Maslov. “Spontaneous emergence of autocatalytic information-coding polymers”. In: *The Journal of Chemical Physics* 143.4 (2015). Link.

[123] A. V. Tkachenko and S. Maslov. “Onset of natural selection in populations of autocatalytic heteropolymers”. In: *The Journal of Chemical Physics* 149.13 (2018). Link.

[124] A. S. Tupper and P. G. Higgs. “Rolling-circle and strand-displacement mechanisms for non-enzymatic RNA replication at the time of the origin of life”. In: *Journal of Theoretical Biology* 527 (2021). Link.

[125] P. C. Id and P. G. H. Id. “Computer simulations of Template-Directed RNA Synthesis driven by temperature cycling in diverse sequence mixtures”. In: *PLOS Computational Biology* 18.8 (2022). Ed. by S.-J. Chen. Link.

[126] H. Fellermann, S. Tanaka, and S. Rasmussen. “Sequence selection by dynamical symmetry breaking in an autocatalytic binary polymer model”. In: *Physical Review E* 96.6 (2017). Link.

[127] A. V. Tkachenko and S. Maslov. “Emergence of catalytic function in prebiotic information-coding polymers”. Link. 2023.

[128] D. T. Gillespie. “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions”. In: *Journal of Computational Physics* 22.4 (1976). Link.

[129] T. E. Ouldridge. “The importance of thermodynamics for molecular systems, and the importance of molecular systems for thermodynamics”. In: *Natural Computing* 17.1 (2018). 1702.00360.

[130] S. Bo and A. Celani. “Multiple-scale stochastic processes: Decimation, averaging and beyond”. In: *Physics Reports* 670 (2017). Link.

[131] R. Rao and M. Esposito. “Nonequilibrium Thermodynamics of Chemical Reaction Networks: Wisdom from Stochastic Thermodynamics”. In: *Physical Review X* 6.4 (2016). Link.

[132] M. Pekař. “Thermodynamics and foundations of mass-action kinetics”. In: *Progress in Reaction Kinetics and Mechanism* 30.1-2 (2005). Link.

[133] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II*. Vol. 14. Link. Springer Berlin Heidelberg, 1991.

[134] W. Cleland. “The kinetics of enzyme-catalyzed reactions with two or more substrates or products. II. Nomenclature and rate equations”. In: *BBA - Biochimica et Biophysica Acta* 67.2 (1963). Link.

[135] W. W. Cleland. “The kinetics of enzyme-catalyzed reactions with two or more substrates or products. II. Inhibition: Nomenclature and theory”. In: *BBA - Biochimica et Biophysica Acta* 67.C (1963). Link.

[136] W. W. Cleland. “The kinetics of enzyme-catalyzed reactions with two or more substrates or products. III. Prediction of initial velocity and inhibition patterns by inspection”. In: *BBA - Biochimica et Biophysica Acta* 67.C (1963). Link.

[137] T. N. Grossmann, A. Strohbach, and O. Seitz. “Achieving Turnover in DNA-Templated Reactions”. In: *ChemBioChem* 9.14 (2008). Link.

[138] C. Walter. “The Use of Product Inhibition and Other Kinetic Methods in the Determination of Mechanisms of Enzyme Action”. In: *Advance in Chemical Physics* 7.1 (1964). Link.

[139] S. Ainsworth. “The mechanistic interpretation of initial velocity and product inhibition data arising from enzyme catalysed reactions whose catalytic cycle is unbranched”. In: *Journal of Theoretical Biology* 50.1 (1975). Link.

[140] M. Yu, M. L. B. Magalhães, P. F. Cook, and J. S. Blanchard. “Bisubstrate Inhibition: Theory and Application to N-Acetyltransferases”. In: *Biochemistry* 45.49 (2006). Link.

[141] S. Banta and I. Wheeldon. “Theory-Based Development of Performance Metrics for Comparing Multireactant Enzymes”. In: *ACS Catalysis* 10.2 (2020). Link.

[142] V. Leskovac, S. Trivić, D. Peričin, and J. Kandrač. “Deriving the rate equations for product inhibition patterns in bisubstrate enzyme reactions”. In: *Journal of Enzyme Inhibition and Medicinal Chemistry* 21.6 (2006). Link.

[143] R. Ohs, K. Fischer, M. Schöpping, and A. C. Spiess. “Derivation and identification of a mechanistic model for a branched enzyme-catalyzed carboligation”. In: *Biotechnology Progress* 35.6 (2019). Link.

[144] S. Kozuch and J. M. L. Martin. ““Turning Over” Definitions in Catalytic Cycles”. In: *ACS Catalysis* 2.12 (2012). Link.

[145] R. J. Fox and M. D. Clay. “Catalytic effectiveness, a measure of enzyme proficiency for industrial applications”. In: *Trends in Biotechnology* 27.3 (2009). Link.

[146] S. Banta, B. A. Swanson, S. Wu, A. Jarnagin, and S. Anderson. “Optimizing an Artificial Metabolic Pathway: Engineering the Cofactor Specificity of Corynebacterium 2,5-Diketo-d-gluconic Acid Reductase for Use in Vitamin C Biosynthesis†”. In: *Biochemistry* 41.20 (2002). Link.

[147] J. SantaLucia and D. Hicks. “The thermodynamics of DNA structural motifs”. In: *Annual Review of Biophysics and Biomolecular Structure* 33 (2004). Link.

[148] A. Deshpande and T. E. Ouldridge. “Optimizing enzymatic catalysts for rapid turnover of substrates with low enzyme sequestration”. In: *Biological Cybernetics* 114.6 (2020). Link.

[149] D. Y. Zhang and E. Winfree. “Control of DNA strand displacement kinetics using toehold exchange”. In: *Journal of the American Chemical Society* 131.47 (2009). Link.

[150] N. Srinivas, T. E. Ouldridge, P. Šulc, J. M. Schaeffer, B. Yurke, A. A. Louis, J. P. Doye, and E. Winfree. “On the biophysics and kinetics of toehold-mediated DNA strand displacement”. In: *Nucleic Acids Research* 41.22 (2013). Link.

[151] H. Zur and T. Tuller. “Predictive biophysical modeling and understanding of the dynamics of mRNA translation and its evolution”. In: *Nucleic Acids Research* 44.19 (2016). Link.

[152] L. Zhou, S. C. Kim, K. H. Ho, D. K. O’Flaherty, C. Giurgiu, T. H. Wright, and J. W. Szostak. “Non-enzymatic primer extension with strand displacement.” In: *eLife* 8 (2019). Link.

[153] R. M. Dirks, J. S. Bois, J. M. Schaeffer, E. Winfree, and N. A. Pierce. “Thermodynamic Analysis of Interacting Nucleic Acid Strands”. In: *SIAM Review* 49.1 (2007). Link.

[154] M. E. Fornace, N. J. Porubsky, and N. A. Pierce. “A Unified Dynamic Programming Framework for the Analysis of Interacting Nucleic Acid Strands: Enhanced Models, Scalability, and Speed”. In: *ACS Synthetic Biology* 9.10 (2020). Link.

[155] M. E. Fornace, J. Huang, C. T. Newman, N. J. Porubsky, M. B. Pierce, and N. A. Pierce. “NUPACK: Analysis and Design of Nucleic Acid Structures, Devices, and Systems”. Link. 2022.

[156] P. Irmisch, T. E. Ouldridge, and R. Seidel. “Modeling DNA-Strand Displacement Reactions in the Presence of Base-Pair Mismatches”. In: *J. Am. Chem. Soc* 142 (2020). Link.

[157] R. R. F. Machinek, T. E. Ouldridge, N. E. C. Haley, J. Bath, and A. J. Turberfield. “Programmable energy landscapes for kinetic control of DNA strand displacement”. In: *Nature Communications* 5.1 (2014). Link.

[158] C. Wang, J. H. Bae, and D. Y. Zhang. “Native characterization of nucleic acid motif thermodynamics via non-covalent catalysis”. In: *Nature Communications 2016 7:1* 7.1 (2016). Link.

[159] B. G. Moreira, Y. You, and R. Owczarzy. “Cy3 and Cy5 dyes attached to oligonucleotide terminus stabilize DNA duplexes: Predictive thermodynamic model”. In: *Biophysical Chemistry* 198 (2015). Link.

[160] L. E. Fehily. “DNA Self-Replication by Handhold Mediated Strand Displacement”. MA thesis. Imperial College London, 2022.

[161] A. I. Hanopolskyi, V. A. Smaliak, A. I. Novichkov, and S. N. Semenov. “Autocatalysis: Kinetics, Mechanisms and Design”. In: *ChemSystemsChem* 3.1 (2021). Link.

[162] N. Simpson. “Framework for Length Control in Minimal Autonomous Copying Systems”. MA thesis. Imperial College London, 2022.

[163] U. Seifert. “Stochastic thermodynamics: principles and perspectives”. In: *The European Physical Journal B* 64.3-4 (2008). Link.

[164] U. Seifert. “Stochastic thermodynamics, fluctuation theorems and molecular machines”. In: *Reports on Progress in Physics* 75.12 (2012). Link.

Appendix A

Additional material supporting chapter 2

The contents of this appendix are adapted from work I published in Ref. [1] with supervision from Thomas Ouldridge and Jenny Poulton.

A.1 Extreme parameter values create long embedded processes

In Fig. A.1, we show that extreme parameter values result in long embedded Markov processes which are time consuming to simulate with the Gillespie algorithm [128], resulting in simulations that produce few polymers. We plot the averaged mean polymer length and averaged polymer count produced by parameter values which generated fewer than 20 polymers in Fig. A.1. When $\Delta G_{\text{spec}} = 6$, the copy-template interaction is very weak, and the dynamics are dominated by repeated binding and unbinding of monomers to the template. Hence when the polymerisation rate is also slow, $k = 10^{-4}$ and $k = 10^{-2}$, many repeated binding and unbinding events occur before the formation of a dimer, and very few polymers were produced by the simulation within the run time. When $\Delta G_{\text{spec}} = -10$, the copy-template interaction is very strong, unbinding is slow. Hence, for fast polymerisation rates k and low values $-\Delta G_{\text{pol}}$, the dynamics are dominated by repeated polymerisation and depolymerisation of neighbouring units, and it takes a long time to simulate the formation and release of dimers. In cases where few outputs of polymers were given, it doesn't necessarily give long polymers, but we have no evidence that there is a tendency to produce long polymers reliably.

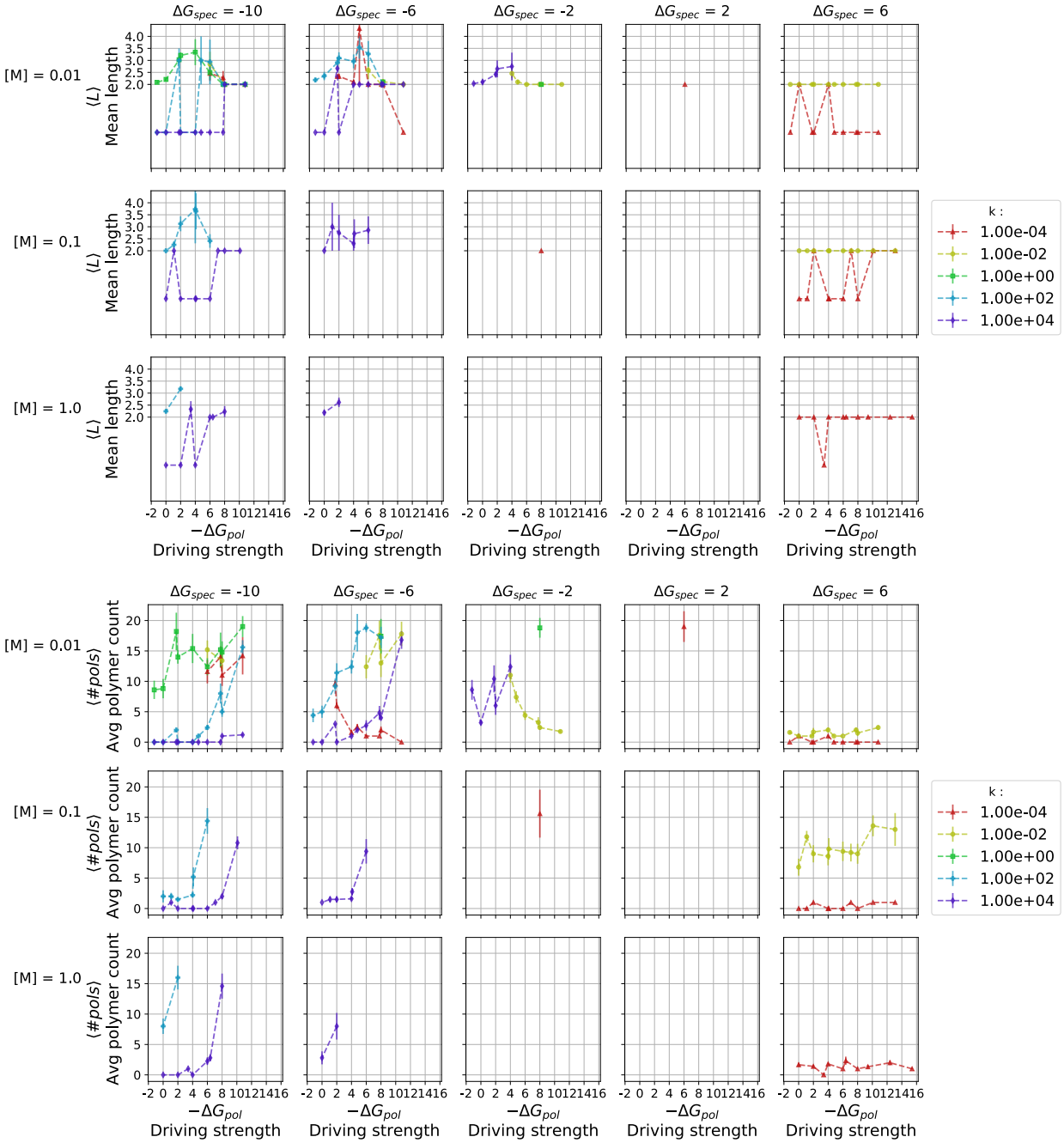


Figure A.1: Extreme parameter values result in long embedded processes which are time consuming to simulate with the Gillespie algorithm [128]. These plots are the average length and average polymer count produced by simulations with average polymer count below 20. When ΔG_{spec} is high, slow polymerisation rates k lead to undersampling, and when ΔG_{spec} is low, fast polymerisation rates and low monomer concentrations $[\text{M}]$ lead to undersampling. Undersampling is due to long embedded processes. Figure adapted with permission from Ref. [1] under a CC BY license.

A.2 Full product length distributions accompanying main text Fig. 3.12

In Section 3.3.3 we demonstrated that weakening the copy-template interaction strength at the final site of the template with an energetic factor ΔG_{end} was sufficient to selectively release complete polymers. In Fig. 3.12, we presented the only the complete-length component, p_{complete}, of the product length distribution as ΔG_{end} was varied. In Figs. A.2, A.3 and A.4 we show the entire product length probability distribution, $p(\text{length})$ against ΔG_{end}, for all values of $[\text{M}] = \{1,3,10\}$, $\Delta G_{\text{gen}} = \{-12,-10,-8\}$, and $L_T = \{10,30,100\}$. We fix $k = 1$, $\Delta G_{\text{spec}} = -4$, and $\Delta G_{\text{BB}} = -20$, which puts the system into a dense brush regime in which long copies with $\langle L \rangle \sim L_T/2$ are produced in Section 3.3.2 when $\Delta G_{\text{end}} = 0$. p_{complete} corresponds to the slice of the length distribution for polymers which have the same length as the template $p(\text{length} = L_T) \equiv p_{\text{complete}}$.

As discussed in Section 3.3.3, when $\Delta G_{\text{end}} = 0$, in the surface plots below we see a relatively flat, uniform length distribution. When is increased ΔG_{end} to moderate values complete polymers dominate the length distribution. From the length distributions shown below we can clearly see that the dip in p_{complete} that features in Fig. 3.12 is due to a peak in the production of short polymers. At the highest values of ΔG_{end}, no complete polymers are produced ($p(\text{length} = L_T) \equiv p_{\text{complete}} \approx 0$) and the rest of the length distribution is flat.

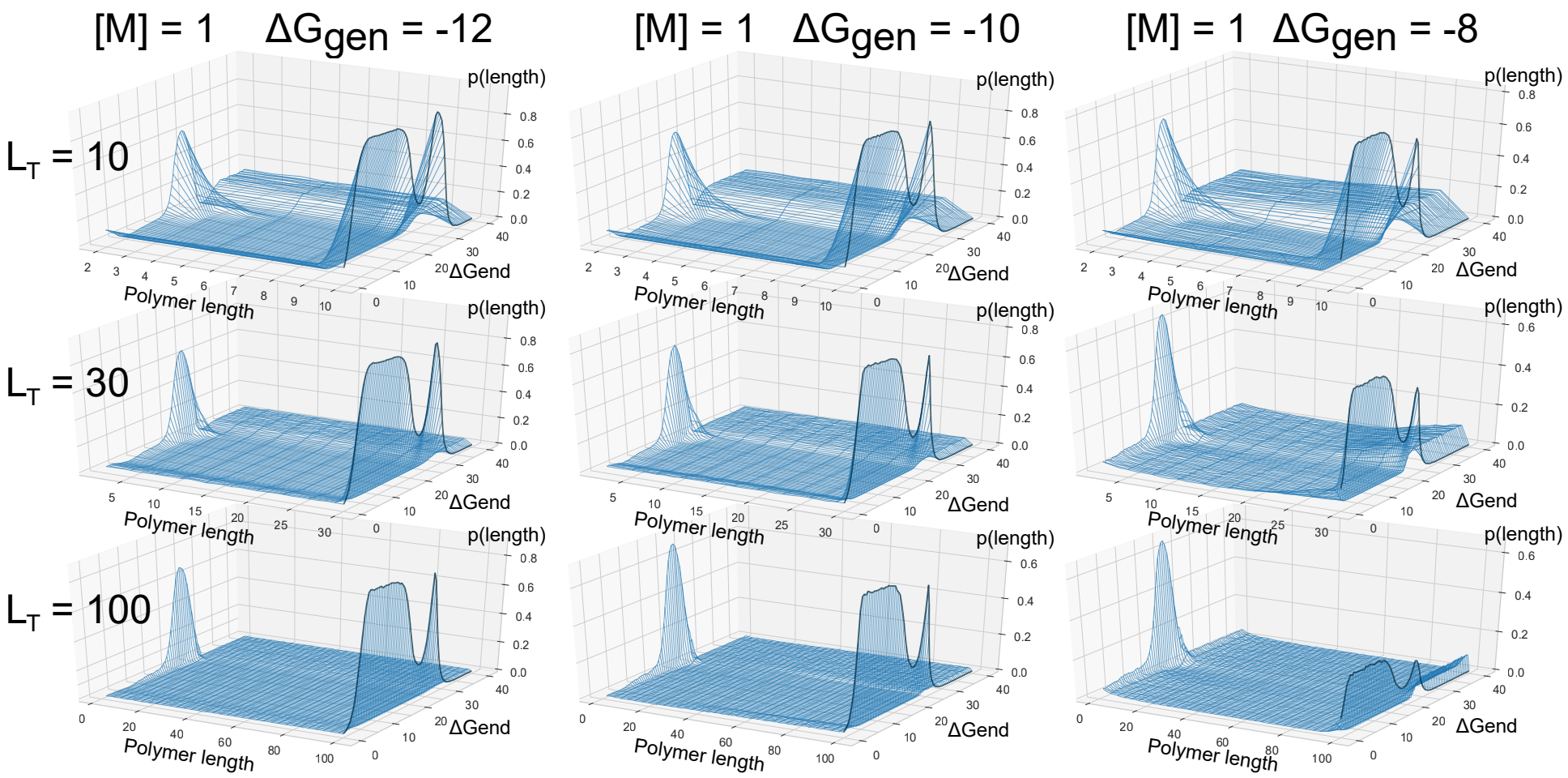


Figure A.2: The product probability length distributions $p(\text{length})$ averaged over 5 independent repeats of ensembles of up to 1000 polymers at each value of $\Delta G_{\text{end}} = 0, 0.5, ..., 40$, for $[\text{M}] = 1$, $\Delta G_{\text{gen}} = \{-12, -10, -8\}$, and $L_T = \{10, 30, 100\}$. The line corresponding to p_{complete} is accentuated in black. Figure adapted with permission from Ref. [1] under a CC BY license.

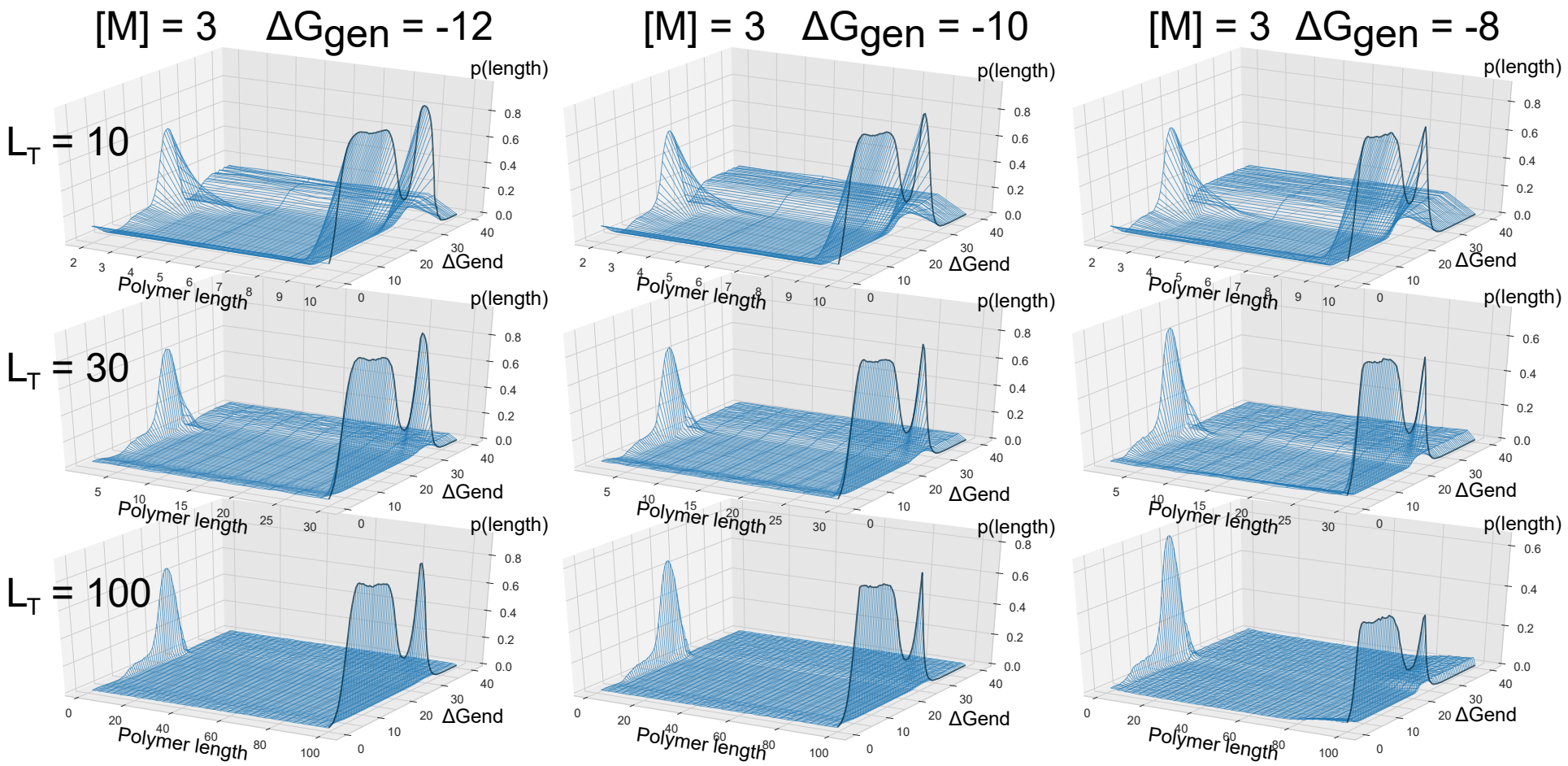


Figure A.3: The product probability length distributions $p(\text{length})$ averaged over 5 independent repeats of ensembles of up to 1000 polymers at each value of $\Delta G_{\text{end}} = 0, 0.5, ..., 40$, for $[\text{M}] = 3$, $\Delta G_{\text{gen}} = \{-12, -10, -8\}$, and $L_T = \{10, 30, 100\}$. The line corresponding to p_{complete} is accentuated in black. Figure adapted with permission from Ref. [1] under a CC BY license.

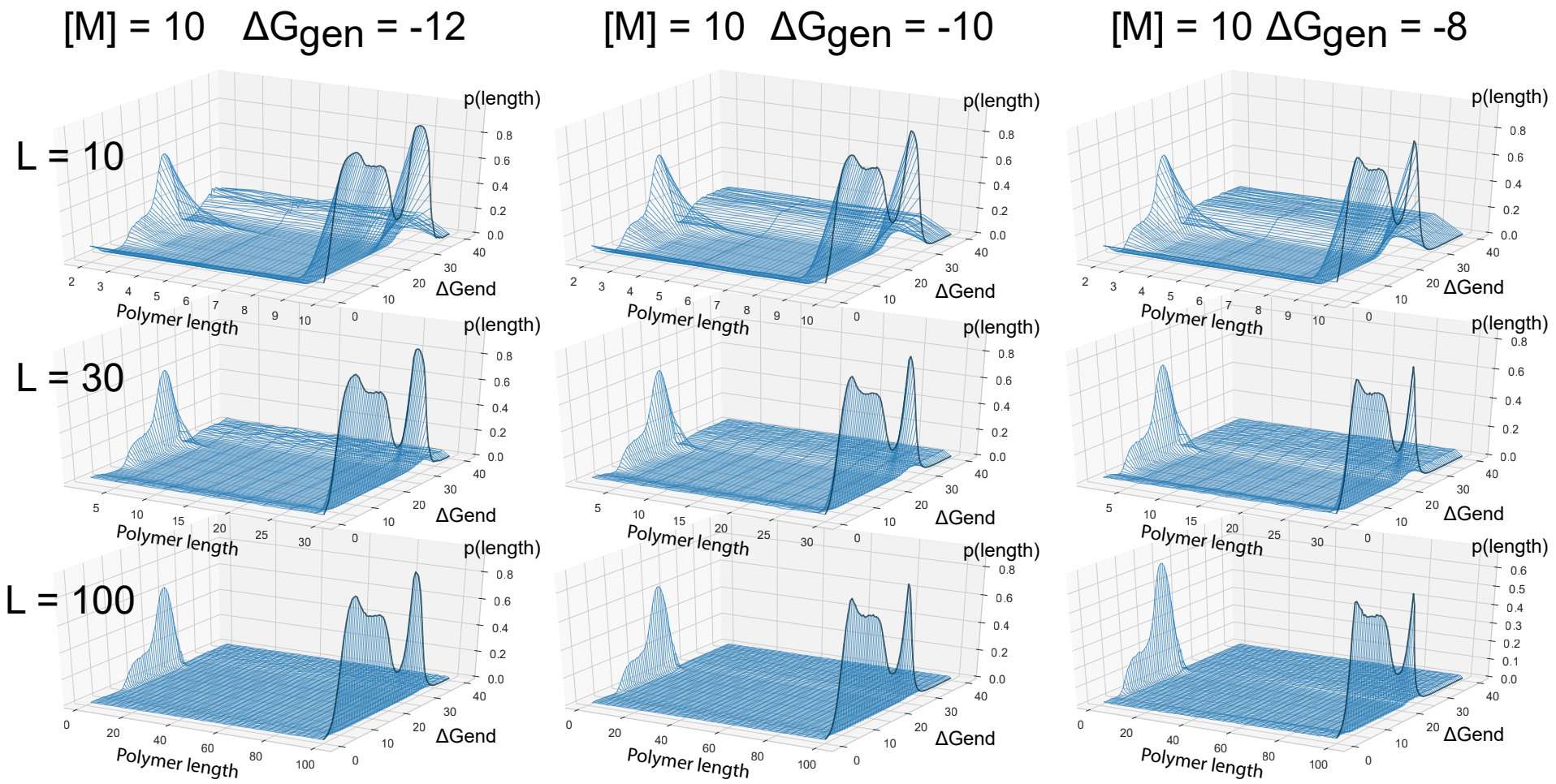


Figure A.4: The product probability length distributions $p(\text{length})$ averaged over 5 independent repeats of ensembles of up to 1000 polymers at each value of $\Delta G_{\text{end}} = 0, 0.5, ..., 40$, for $[\text{M}] = 10$, $\Delta G_{\text{gen}} = \{-12, -10, -8\}$, and $L_T = \{10, 30, 100\}$. The line corresponding to p_{complete} is accentuated in black. Figure adapted with permission from Ref. [1] under a CC BY license.

A.3 The asymmetric destabilisation and weak end mechanism is more reliable for polymer tails with higher template affinity

In addition to the results presented in Section 3.3.3, in Fig. A.5 we demonstrate that the combination of the asymmetric destabilisation of polymer tails induced by ΔG_{gen} and the weakening of the end of the template with ΔG_{end} can result in the reliable production of complete polymers p_{complete} even when the tails of polymers tend to detach from the template faster than they rebind with $k_0 exp(\Delta G_{\text{spec}}) \gg k_0[\text{M}_{\text{eff}}]$, or when $\ln \Delta G_{\text{spec}} > \ln[\text{M}_{\text{eff}}] \approx 4.6$. Here we set $k_0 = k = 1$, $[\text{M}] = 1, [[\text{M}_{\text{eff}}]] = 100, \Delta G_{\text{BB}} = -20, \Delta G_{\text{gen}} = \{-16, -12\}, \Delta G_{\text{spec}} = \{-4, 2, 8\}$ and $\Delta G_{\text{end}} = \{0, 0.5, ...\}$ and $L_T = \{10, 30, 100\}$, and calculate the average of p_{complete} from up to 2000 polymers produced by 5 independent simulations at each point in the parameter space.

In Fig. A.5, as ΔG_{spec} is increased, p_{complete} begins to fall at lower values of ΔG_{end}. This cusp in p_{complete} is well reproduced by the simple model presented in Section 3.3.3. As ΔG_{end} increases, the occupancy of the final template site reduces, which increases the average timescale on which polymers are formed (completed) $\langle \tau_{\text{form}} \rangle$. When the timescale to form a complete polymer $\langle \tau_{\text{form}} \rangle$ is of the same order as the timescale on which the L_T incomplete polymers detach from the template $\langle \tau_{\text{uniform}} \rangle$, the flux of uniform length polymers exceeds the flux of completed polymers, and hence p_{complete} drops. This drop, which defines the useful range of ΔG_{end}, occurs at lower values of ΔG_{end} when the specific bond strength ΔG_{spec} is weaker (more positive).

Additionally in Fig. A.5, we note that for $\Delta G_{\text{spec}} = 8$, p_{complete} fails to increase substantially upon weakening the end of the template, especially for long templates. We note that the for $\Delta G_{\text{gen}} = -16$, at $\Delta G_{\text{end}} = 0$, the system lies in region **III'** of Fig. 3.8, where a dense brush of displaced polymer tails is expected to form on the template, which results in a uniform product length distribution. However, for $\Delta G_{\text{gen}} = -12$, at $\Delta G_{\text{end}} = 0$, the system lies in region **V** where the spontaneous detachment of polymers enables fragments to grow, biasing the production of shorter polymers, as the system doesn't reach a maximally-displaced conformation. Hence, for $\Delta G_{\text{gen}} = -16$, $\Delta G_{\text{spec}} = 8$, p_{complete} fails to increase substantially upon weakening the end of the template because the workable range of ΔG_{end} is slim. However, for $\Delta G_{\text{gen}} = -12$, $\Delta G_{\text{spec}} = 8$, p_{complete} is further reduced because of the tendency for the system to release incomplete polymers before reaching a maximally-displaced conformation in which weakening the end of the template can selectively release completed polymers.

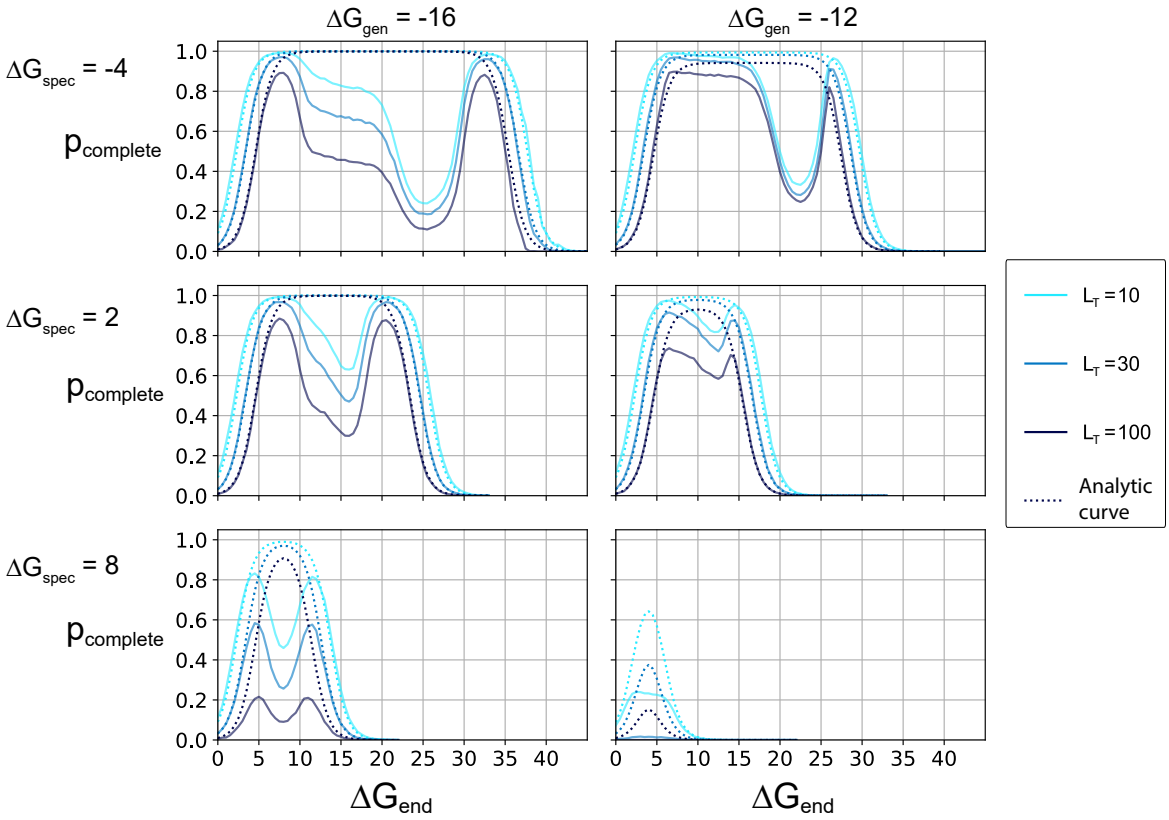


Figure A.5: Strengthened the specific bond strength, ΔG_{spec}, (more negative), increases the range of ΔG_{end} over which p_{complete} is high. The simple analytical model (dotted lines) presented in Section 3.3.3 still reproduces the boundaries over which p_{complete} is high with no free fitted parameters. The simple model excludes the effects of dimer formation that occur at intermediate values of ΔG_{end}, which makes p_{complete} lower than the predicted value. Figure adapted with permission from Ref. [1] under a CC BY license.