

Modelling The Folding Pathway of DNA Nanostructures

Behnam Najafi

Wadham College
University of Oxford

*A thesis submitted for the degree of
Doctor of Philosophy*

Michaelmas 2022

Abstract

DNA origami is a robust technique for bottom-up nano-fabrication. It encodes a target shape into uniquely addressable interactions between a set of short ‘staple’ strands and a long ‘scaffold’ strand. The mechanisms of self-assembly, particularly regarding kinetics, need to be better understood. Origami design usually relies on optimising the thermodynamic stability of the target structure, and thermal annealing remains the most fool-proof assembly protocol. This work focuses on studying the folding pathway of three types of origami through simulations: a reconfigurable T-junction origami, several traditional origami, and origami with coated scaffolds.

The T-junction origami is intended as an economically feasible method of changing the uniqueness of interactions. My contribution to this work is characterising the basic structural motif through oxDNA, a nucleotide-resolution model of DNA. The thesis then focuses on extending a domain-level model of DNA origami to study several experimental origami designs. We reveal design-dependent free energy barriers using biased simulations and relate this to the observed hysteresis in experiments. We also highlight the role of specific design elements in determining the folding pathway. A novel method of lowering the temperature of error-free assembly using coated scaffolds is then presented, with simulations indicating the existence of an activation barrier. By exposing particular regions of the scaffold, we can lower assembly time and temperature.

Modelling The Folding Pathway of DNA Nanostructures



Behnam Najafi
Wadham College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2022

This thesis is dedicated to
Farshad Najafi

Acknowledgements

The completion of this thesis would not have been feasible without the unwavering support of numerous individuals.

First and foremost, I express my deepest gratitude to my supervisors, Prof. Andrew Turberfield and Prof. Ard Louis. Their persistent encouragement, guidance, and mentorship were instrumental in helping me navigate the complexities of DNA nanotechnology. I also appreciate Prof. Jonathan Doye from the Theoretical Chemistry department, who acted as an unofficial supervisor and shared valuable insights and assistance throughout my research. Special thanks to Dr Jonathan Bath for his collaboration in experimental work and assistance with publications, as well as Dr Katherine Young, whose experimental endeavours complemented my theoretical results. I extend my appreciation to past and present colleagues who made the work environment enjoyable and provided moral and intellectual support in various ways. Your camaraderie and warmth have truly enriched my research experience, and I will undoubtedly miss the delightful cakes and entertaining costumes.

My heartfelt appreciation goes to my friends at Oxford—Emil, Zhao, Vira, Marija, Louis, Kaveh, Rafa, and Michaela. Your friendship and support have been invaluable throughout my time at Oxford. I treasure our memories together, from bonfires at Port Meadow to the countless hours spent in the Wadham MCR—a special acknowledgement to Michaela, who endured my PhD-related meltdowns and offered unwavering love and support.

I extend my deepest appreciation to my parents for their steadfast love, encouragement, and understanding throughout this journey. To my siblings—Mohsen, Hesam, and Kimiya—your confidence in my abilities and your unwavering support have served as sources of strength and motivation.

Lastly, I thank Wadham College of Oxford University for the enriching academic and social environment it provided as a student and for the necessary resources, facilities, and conducive atmosphere for conducting my research.

Abstract

DNA origami is a robust technique for bottom-up nano-fabrication. It encodes a target shape into uniquely addressable interactions between a set of short ‘staple’ strands and a long ‘scaffold’ strand. The mechanisms of self-assembly, particularly regarding kinetics, need to be better understood. Origami design usually relies on optimising the thermodynamic stability of the target structure, and thermal annealing remains the most fool-proof assembly protocol. This work focuses on studying the folding pathway of three types of origami through simulations: a reconfigurable T-junction origami, several traditional origami, and origami with coated scaffolds.

The T-junction origami is intended as an economically feasible method of changing the uniqueness of interactions. My contribution to this work is characterising the basic structural motif through oxDNA, a nucleotide-resolution model of DNA. The thesis then focuses on extending a domain-level model of DNA origami to study several experimental origami designs. We reveal design-dependent free energy barriers using biased simulations and relate this to the observed hysteresis in experiments. We also highlight the role of specific design elements in determining the folding pathway. A novel method of lowering the temperature of error-free assembly using coated scaffolds is then presented, with simulations indicating the existence of an activation barrier. By exposing particular regions of the scaffold, we can lower assembly time and temperature.

Contents

List of Figures	xiii
List of Tables	xvii
List of Abbreviations	xix
1 Introduction	1
1.1 Self-Assembly	2
1.2 DNA Nanotechnology	4
1.2.1 DNA Molecule	5
1.2.2 Design Elements and Early Motifs	7
1.2.3 DNA Origami	9
1.2.4 DNA Bricks	14
1.3 Modelling Tools in DNA Nanotechnology	16
1.3.1 The oxDNA Model	18
1.4 Simulation Techniques	20
1.5 Thesis Outline	24
2 Reconfigurable T-junction Origami	27
2.1 Introduction	28
2.2 Characterising T-motifs	30
2.2.1 Conformations	33
2.2.2 Free Energy	37
2.3 Characterising Linkers	40
2.4 T-junction Origami	44
2.5 Summary and Outlook	48
3 Domain Level Model of DNA Origami	51
3.1 Introduction	51
3.2 The Chemical System	52
3.2.1 Theoretical Framework	53
3.2.2 Simulation Algorithm	57
3.2.3 Equilibrium and Standard Gibbs Free Energy	58

3.2.4	Nearest-Neighbour Model of DNA Thermodynamics	60
3.3	State Space	61
3.4	Free Energy Contributions	67
3.5	Rate Model	73
3.6	Sampling Techniques	76
3.7	Comparison to Experiment	81
4	Thermodynamics and Kinetics of DNA Origami Assembly	89
4.1	Introduction	89
4.2	Toy Systems	90
4.2.1	Equilibrium Properties	90
4.2.2	Simulations under a Temperature Ramp	98
4.3	Real Origami	101
4.3.1	System $\mathcal{R}2$	103
4.3.2	System $\mathcal{R}3$	109
4.3.3	System \mathcal{T}	115
4.4	Summary and Outlook	118
5	Nucleated Assembly at Low Temperature	121
5.1	Introduction	122
5.2	Preliminary Experiments	122
5.3	Simulations at Low Temperature	128
5.3.1	Non-Coated Origami	128
5.3.2	Coated Origami	133
5.4	Nucleation	138
5.5	Propagation	143
5.6	Controlling The Folding Pathway	145
5.7	Summary and Outlook	150
6	Conclusions	153
Appendices		
A	DNA Sequences	159
A.1	Scaffold Strands	159
A.1.1	pKD1	160
A.1.2	pJB101	161
B	Fitting Staple Data	163

C	Appendix to Chapter 4	165
C.1	Temperature Dependence of Two-Row Systems	165
C.2	Open Scaffold Systems	168
C.3	Four-Helix Systems	172
C.4	Seam Variants of $\mathcal{R}2$	175
C.5	Sequence-dependent Staple Data	178
C.6	Domain Distribution of $\mathcal{R}3$	181
D	Appendix to Chapter 5	183
D.1	Domain Energies	183
	References	185

List of Figures

1.1	Illustration of Free Energy Landscape in self-assembly.	2
1.2	The DNA Molecule.	6
1.3	Design elements in DNA nanotechnology.	8
1.4	Illustration of DNA Origami	10
1.5	The oxDNA model.	18
2.1	The T-motif and proposed Tile.	29
2.2	T-motif characterisation scheme.	31
2.3	T-motif umbrella sampling windows.	33
2.4	Coaxial stacking probability as a function of the number of base pairs.	35
2.5	Coaxial stacking probability as a function of bulge loop size.	35
2.6	Angle between the three arms of T-motifs.	36
2.7	Geometric distributions of T-motif stacking states.	37
2.8	Free energy profile of T-motifs.	38
2.9	Comparison of T-motif stabilities to experiment.	40
2.10	Effect of linker length and sticky end polarity on orientation.	41
2.11	Cooperative binding of U-shaped motif.	43
2.12	Tolerance of the U-shaped motif to linker length.	44
2.13	Design of T-junctions origami.	47
3.1	Sample origami state space.	63
3.2	Ambiguity in loop assignments.	64
3.3	Strand exchange model.	66
3.4	Position of seam staples and edge staples.	73
3.5	Allowed transitions in the model.	75
3.6	Testing the umbrella sampling method.	81
3.7	Comparison of annealing and melting curves between simulation and experiment for different origami designs.	84
3.8	Fitting procedure for annealing and melting transitions.	85
3.9	Comparison of the domain-level model to experiment.	86
3.10	Comparison of the domain-level model to experiment: effect of staple concentration and temperature ramp.	87

4.1	Free energy of two-row toy systems.	91
4.2	Effect of varying model parameters on free energy profiles.	95
4.3	Probability that each staple is fully bound in two-row systems.	97
4.4	Two-row systems under temperature ramps.	100
4.5	Position of shorter domains.	103
4.6	Incorporation temperature and Hill coefficients for individual staples during annealing and melting in $\mathcal{R}2$ origami.	104
4.7	Incorporation probability at different temperatures during annealing and melting in $\mathcal{R}2$ origami.	105
4.8	Free energy profiles as a function of the number of bound domains and the number of bound seam domains for system $\mathcal{R}2$ at $T_M^{\text{eq}}=68.1^\circ\text{C}$	106
4.9	Free energy profiles of seam variants of $\mathcal{R}2$	108
4.10	2D and sequence-dependent free energy profiles of $\mathcal{R}2$	110
4.11	Incorporation temperature and Hill coefficients for individual staples during annealing and melting of $\mathcal{R}3$ origami.	111
4.12	Incorporation probability at different temperatures during annealing and melting of $\mathcal{R}3$	111
4.13	Free energy profile of $\mathcal{R}3$ origami.	113
4.14	Free energy profile of \mathcal{T} origami.	116
4.15	Incorporation temperature and Hill coefficients for individual staples during annealing and melting of \mathcal{T}	117
4.16	Incorporation probability at different temperatures during annealing and melting of \mathcal{T}	117
5.1	The coated origami system.	123
5.2	Position of fluorophores and FRET pairs.	125
5.3	Initial investigation of the folding process.	127
5.4	Averaging over simulation trajectories.	129
5.5	Isothermal simulations of non-coated scaffold.	132
5.6	Effect of sequence-specificity on non-coated origami.	133
5.7	Simulation trajectories in coated and non-coated origami.	134
5.8	Effect of temperature on coated and non-coated origami.	135
5.9	Single trajectories of the folding pathway of coated origami.	136
5.10	Averaged trajectories of the folding pathway of coated origami.	137
5.11	Comparison of simulation to experiment.	140
5.12	Nucleation barrier in coated origami.	141
5.13	2D free energy profile of coated origami.	142
5.14	Nucleation pathway in coated origami.	142
5.15	Comparison of nucleation and propagation time-scales.	145
5.16	Comparison to experiment, full coating set.	148

5.17	Comparison to experiment, coating staples at the seam omitted. . .	148
5.18	Comparison to experiment, coating staples at the 5' side of seam staples omitted.	149
5.19	Comparison to experiment, coating staples at the 3' side of seam staples omitted.	149
5.20	Comparison to experiment, Non-seam coating staples omitted. . . .	150
B.1	Staple fitting procedure	163
C.1	Two-row staples at various temperatures.	166
C.2	Effect of temperature on free energy profiles of two-row systems. . .	167
C.3	Open scaffold systems.	169
C.4	Open scaffold systems.	170
C.5	Open scaffold systems.	171
C.6	Four-row systems.	173
C.7	Four-row systems.	174
C.8	Staple data for seam variations of $\mathcal{R}2$	176
C.9	Staple data for seam variations of $\mathcal{R}2$	177
C.10	Sequence-specific staple data for $\mathcal{R}2$	179
C.11	Sequence-specific staple data for $\mathcal{R}3$	179
C.12	Sequence-specific staple data for \mathcal{T}	180
C.13	Individual domain data for $\mathcal{R}3$	181
D.1	Variation in domain melting temperature in coated system.	184
D.2	Nucleation sites in the coated system.	184

List of Tables

3.1	Custom rate model used in this work.	76
-----	--	----

List of Abbreviations

nt	Nucleotide.
bp	Base pair.
ssDNA	Single-stranded DNA.
dsDNA	Double-stranded DNA.
AFM	Atomic force microscopy.
FRET	Förster resonance energy transfer.
TMSD	Toehold-mediated strand displacement.
MC	Monte Carlo.
VMMC	Virtual move Monte Carlo.
US	Umbrella sampling.
WHAM	Weighted histogram analysis method.
FJC	Freely-jointed chain
DLM	Domain-level model

1

Introduction

Contents

1.1 Self-Assembly	2
1.2 DNA Nanotechnology	4
1.2.1 DNA Molecule	5
1.2.2 Design Elements and Early Motifs	7
1.2.3 DNA Origami	9
1.2.4 DNA Bricks	14
1.3 Modelling Tools in DNA Nanotechnology	16
1.3.1 The oxDNA Model	18
1.4 Simulation Techniques	20
1.5 Thesis Outline	24

This thesis is focused on modelling the folding pathways of self-assembled DNA nanostructures. I begin the chapter by providing a brief overview of natural and man-made molecular self-assembly. I then delve into the specific topic of DNA nanotechnology, including a discussion of the relevant aspects of the DNA molecule and experimental advances in the field. Lastly, I review the models of DNA in the literature and end the chapter with brief discussions about the simulation techniques used in this work. A short thesis outline is sketched at the end of this chapter.

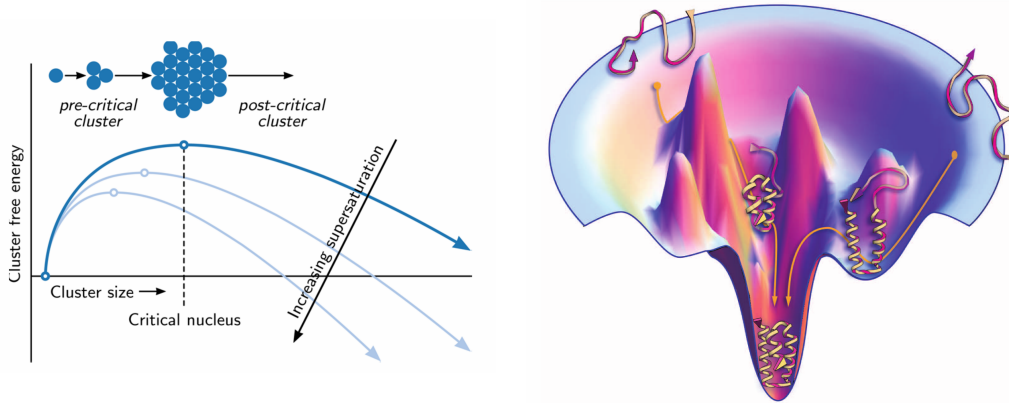


Figure 1.1: Illustration of Free Energy Landscape in self-assembly. Left: Free-energy profile of a simple crystal at various values of degree of supersaturation (from Ref. [2]). Right: Funnel-shaped energy landscape of proteins. The rims represent high-energy unfolded states; the global minimum is the native state. The local minima (folding intermediates) provide a funnel towards the native state, and the height of local free energy barriers determine the overall rate of folding (from Ref. [3]).

1.1 Self-Assembly

Self-assembly is the naturally occurring process by which a disordered system of components spontaneously organises into an ordered state. One example of this is the crystallisation of salt, where the local interactions between sodium and chloride ions result in periodic structures. Molecular self-assembly can also result in the formation of finite structures, such as micelles[1], which form when surfactant lipid molecules in a solution are driven by the hydrophobic effect to minimise the contact of their tail groups with water molecules. However, if the hydrophilic head groups of the surfactant molecules are too closely packed, they will repel each other due to electrostatic forces to minimise their free energy. This leads to the formation of aggregates with different shapes, such as spherical micelles that sequester the hydrophobic tails in their centre. The optimal number of monomers in an individual micelle is determined by the balance between these two forces.

Self-assembly pathways can be surprisingly complex. They are typically best described using collective variables and are best understood in the language of phase transformations, even if, technically, for many finite-sized assemblies, the system is not quite in the infinite limit. To understand the transition from a disordered to an

ordered phase, a free energy landscape based on an appropriate order parameter can be helpful. In both examples discussed above, nucleation is the initial step in forming the self-assembled phase. According to classical nucleation theory [4], [5], the nucleation barrier arises from a negative volume free energy contribution ($\propto r^3$), counteracted by a positive interfacial free energy contribution ($\propto r^2$). Clusters with a radius larger than the critical radius r^* will grow, while clusters with a radius smaller than r^* will disappear. The rate at which nuclei form is determined by the height of the nucleation barrier ΔG^* , which is controlled by supersaturation. As illustrated in Figure 1.1, high supersaturation leads to the formation of multiple nuclei and rapid growth, which often leads to defects.

Self-assembly can also happen within a single large molecule. Intra-molecular self-assembly is usually referred to as folding, with protein folding being the most common example. The sequence of amino acids along the peptide chain encodes the protein's three-dimensional shape. The process of determining how the one-dimensional sequence of amino acids maps to the three-dimensional structure is known as the protein folding problem. One puzzling aspect is the timescale on which it occurs, which is typically in the range of milliseconds to seconds. This is much too short a time for a random search through the vast number of possible conformations of a protein. Known as Levinthal's paradox [6], this puzzle is best explained through principles of statistical mechanics. Rather than being a random search, folding can be viewed as taking an ensemble of pathways guided on a hyper-dimensional free energy landscape (Figure 1.1). Because conformations closer to the native state have lower free energy, folding is funnelled through these metastable states along the way to its native state, while the folding rate is determined by the local energy barriers between metastable states. Viewed through this lens, information about the shape of the energy landscape is crucial in predicting both equilibrium and dynamic properties of the protein [7].

This description of the folding problem can also be extended to other biomolecules, such as RNA or DNA, which is of particular interest to this thesis. Researchers have made significant progress towards characterising free energy landscapes of

biomolecules by studying intermediate, partially-folded states. Single-molecule techniques[8] such as AFM (atomic force microscopy) and laser optical tweezers have allowed scientists to take snapshots of the folding process and confirm the existence of these intermediate states. However, the underlying principles that govern protein folding and the detailed mechanisms of the folding process still need to be fully understood and are the subject of ongoing research. Structure prediction through molecular mechanics simulations is limited by a lack of accurate force fields, especially with regard to interactions with the solvent. In addition, larger molecules present a computational challenge, where the time scales of folding cannot be reached by detailed atomistic simulations, rendering the study of kinetics infeasible without making approximations. Recently, advancements in machine learning have led to breakthroughs in the field of protein structure prediction [9], but these have not yet been used to work out the kinetics of protein folding.

Compared to biological materials such as protein-RNA complexes, most man-made self-assembling structures have been relatively simple, using only a few types of distinct building blocks. In this regard, DNA nanotechnology is a relatively recent field that presents remarkable opportunities. The individual components in DNA nanostructures can be made uniquely addressable, allowing the construction of objects that rival the complexity of natural biological structures. While being a promising nanofabrication technique in its own right, the reliability of DNA nanotechnology also provides us with a playground to study the self-assembly and microscopic physics of biological processes as more general phenomena. Since the thermodynamics of DNA molecules are relatively well understood, arbitrary objects can be designed and used in experimental setups to study other biomolecules [10], [11].

1.2 DNA Nanotechnology

DNA nanotechnology began with the seminal works of Nadrian Seeman [12], [13] in his attempt to design periodic crystals using DNA. In addition to creating 2D and 3D crystals, the field has grown to routinely produce finite structures with

complex geometries and curvatures and implement dynamic devices that utilise the DNA molecule. The interested reader can find detailed reviews of recent developments in Refs. [14], [15]. Here, I will summarise the most important principles that are relevant to this work.

1.2.1 DNA Molecule

The discovery of deoxyribonucleic acid (DNA) as a carrier of genetic information is one of the turning points in biological sciences. While DNA was first isolated in 1891, it was not until 1953 that Watson and Crick proposed their famous model of double-helical DNA [16]. As shown in Figure 1.2, the double-helix is composed of two strands that run in opposite directions. Each strand is a chain of nucleotides, and each nucleotide is composed of a sugar, a phosphate and a nitrogenous base. A set of covalent bonds across alternating sugar and phosphate groups stabilise the single strands and are said to form the *backbone*. The alternating nature of sugars and phosphates gives the strands directionality, and the double helix can only form when the two strands run in opposite directions. By convention, strands are drawn as arrows pointing from the 5' to the 3' direction, ending in a sugar unit.

The bases are planar and come in four variants: adenine (A), cytosine (C), guanine (G), and thymine (T). Although other pairings are possible [17], the most common base-pairing rules are the Watson-Crick A-T bonds and C-G bonds. These rules arise from the alignment of base planes necessary for hydrogen bonding to occur through the spacial proximity of donor and acceptor pairs within each base. For a pair of complementary bases, it is feasible to align several hydrogen bond donors and acceptors simultaneously (two for an A-T pair and three for a G-C pair). In contrast, only one HB donor-acceptor pair can be aligned at a time when dealing with non-complementary bases, resulting in a less stable base pair. Thus C-G bonds are stronger than A-T bonds, which are in turn stronger than non-complementary bonds. Two single strands (ssDNA) with complementary regions are said to *hybridise* to form double-stranded DNA (dsDNA).

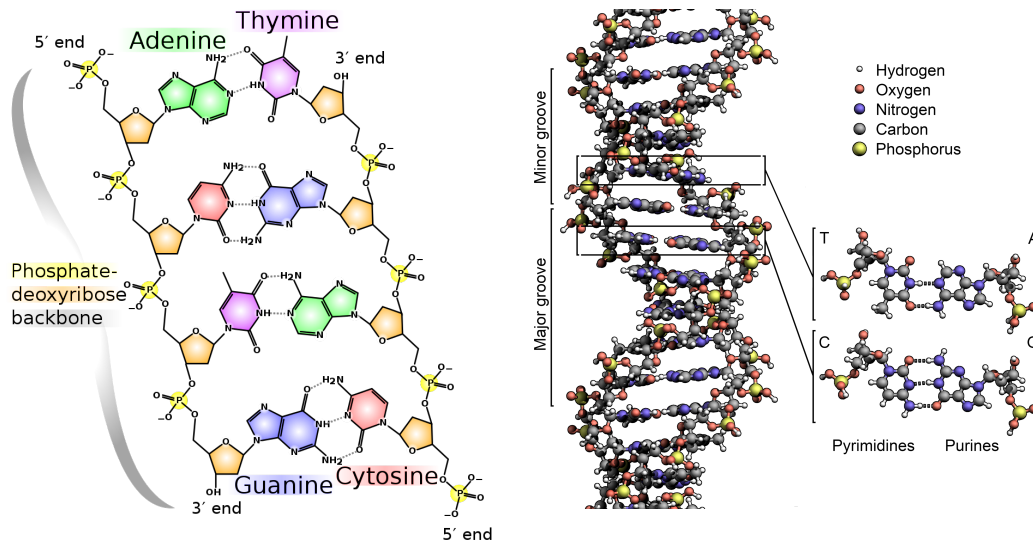


Figure 1.2: The DNA Molecule. Left: chemical composition of DNA is shown with covalent bonds drawn as solid lines and hydrogen bonds as dashed lines. Right: double-helical structure of B-DNA, showing the major and minor grooves. Adapted from Ref. [18].

The most common type of DNA geometry is that of B-DNA (shown in Figure 1.2), which occurs naturally in cells. B-DNA is a right-handed double-helix with a 2 nm diameter, a helical pitch of 3.4 nm (approximately 10.5 base pairs), and major and minor grooves of 1.2 nm and 2.2 nm, respectively. The stability of double-helix is primarily determined by the stacking of the aromatic rings of adjacent base pairs [19]. In an aqueous solution, B-DNA geometry arises as a result of minimising the interaction of hydrophobic bases and water molecules, shielding the stacked rings in the core of the double-helix. On the other hand, the electrostatic repulsion between phosphate groups in the backbone has a destabilising effect on the geometry of B-DNA. This can be mitigated by the presence of cations in the solution that effectively screen the electrostatic field of the phosphate groups. Under non-physiological conditions, DNA can adopt other geometries (e.g. Z-DNA, A-DNA) as dictated by the balance of the aforementioned forces.

The specificity of base-pairing makes DNA the ideal molecule for storing and transmitting genetic information. Since two copies of the sequence of bases are contained in each double-helix, each strand can be used as a blueprint to synthesise

a copy in a process called replication that occurs in cells aided by the DNA *polymerase* enzymes. This forms the basis of transmitting genetic information from parent to offspring. The information encoded in DNA is also used to direct functions within each cell as described by the central dogma of molecular biology: DNA is first transcribed into RNA, which is in turn used to synthesise proteins that perform complex functions.

Outside the cell, oligonucleotides, or oligos for short, are DNA strands consisting of up to 200 nucleotides. They can be produced through chemical synthesis and can be purchased from commercial vendors. However, the production process is prone to errors, and as a result, longer sequences are more difficult and expensive to create. For strands of DNA longer than 200 nucleotides, DNA can be obtained directly from plasmids of live organisms. In this thesis, plasmid-derived strands are used in DNA origami structures and folded into the desired shape by interacting with chemically synthesised oligos.

1.2.2 Design Elements and Early Motifs

Inspired by the biological function of DNA, the structural design of nanostructures proceeds via pairing regions of sequence-complementary on separate strands. The regions of sequence-complementary are often called *domains* of interaction (see Figure 1.3). Each strand is usually designed to have multiple domains and branch between duplexes of the intended structure, forming *crossovers* from one domain to another. Each domain is designed to have a unique sequence, and the target structure is encoded as the global free energy minimum because it maximises the number of correct base pairs [20]. However, partial sequence complementarity between domains is inevitable, leading to the possibility of undesired interactions. While the effect on thermodynamics can be mitigated using longer domains, these off-target interactions can lead to kinetic traps that may not be overcome within experimental time scales.

Early motifs in DNA nanotechnology were based on naturally occurring Holliday junctions [23], [24]. In biological processes such as genetic recombination, the junctions are mobile due to their symmetrical sequence, allowing relative movement

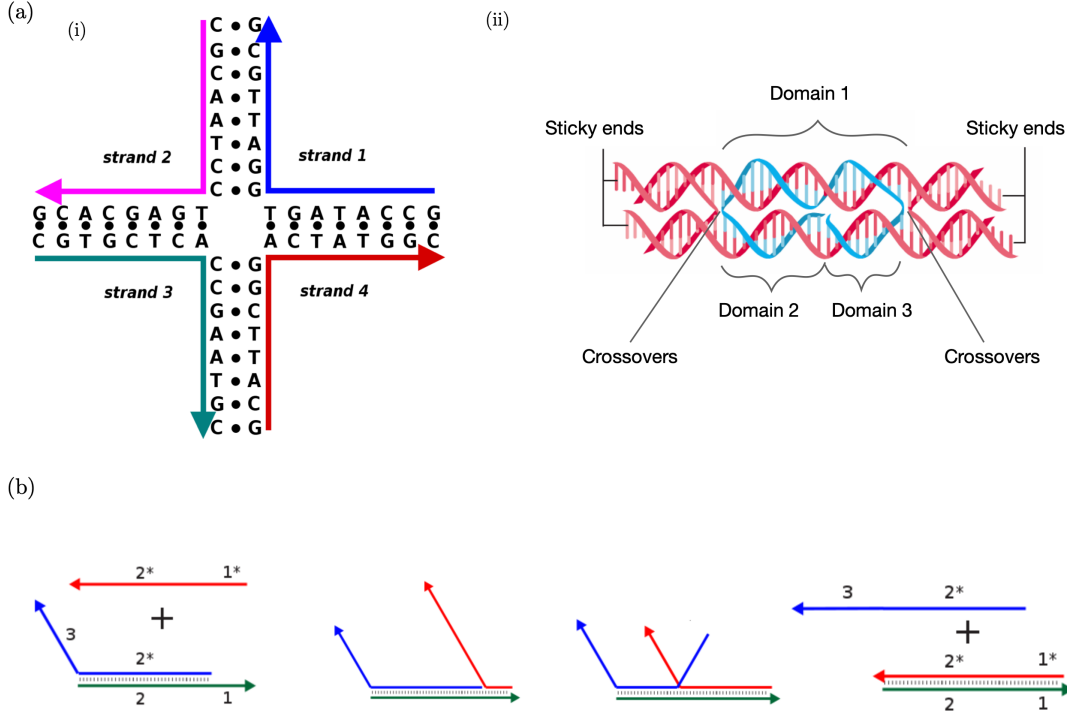


Figure 1.3: Design elements in DNA nanotechnology. (a) (i) A four-way junction formed by the interaction of 4 strands, each consisting of two domains. Taken from [21] (ii) A double-crossover tile, formed by 4 crossovers between 4 strands to create two parallel helices with backbone ‘nicks’ at the position of the crossovers. The blue strand has 3 domains of interactions, leading to an additional nick in the backbone of the bottom duplex. The sticky ends can be used to join the tile to other motifs within larger assemblies. Adapted from [22]. (b) Illustration of toehold-mediated strand displacement (TMSD) with complementary domains labelled with a star (*). The incumbent strand (green) is initially bound to the target strand (blue) via domain 2. To displace the target strand, the invader strand (red) binds via domain 1 (the toehold) to the incumbent, followed by branch migration at domain 2. Adapted from Ref. [21].

of the strands while preserving base pairing. Artificially-designed counterparts aim to immobilise the junctions using asymmetric sequences. The geometry of the arms is determined by the coaxial stacking of duplexes at the junction. An example of a 4-way junction is shown in Figure 1.3(a)(i). More generally, a class of N -way branched junctions, in which N double-helices emanate from a single point, have been successfully assembled [12], [13], [25], [26]. While N -way junctions have been incorporated within larger structures with varying degrees of success [27], they lack the rigidity necessary to be used in larger well-defined geometries. The double-crossover (DX) tile [28] is a more rigid structural motif consisting of two parallel

double-helices (Figure 1.3(a)(ii)). Reciprocal strand exchange between the duplexes leads to discontinuities in the backbone of the resulting double-helices. These are often called backbone *nicks* because of their similarity to the end-product of nicking enzymes. At these positions, bases are still able to coaxially stack to provide stability to the motif. Other variants using more complex crossovers between parallel duplexes have also been realised [29], [30]. These structural motifs can be linked together using single-stranded *sticky ends* to create a variety of higher-order structures in hierarchical assembly protocols, although more simple crystals have been assembled in one-pot reactions. Examples of structures that have been realised in the literature include periodic lattices [31]–[33], polyhedra [34]–[37] and nanotubes [38].

Another useful feature of DNA for technological purposes is its ability to undergo strand displacement reactions. As shown in Figure 1.3(b), toehold-mediated strand displacement (TMSD) can be viewed as a multi-step process, resulting in the replacement of one strand with another on a third ‘incumbent’ strand. If the outgoing target strand can bind back to the incumbent via a second toehold (not shown), the reaction becomes reversible and is referred to as strand exchange. TMSD has been exploited to create dynamic DNA devices [39] such as tweezers that switch between open and closed states [40], motors that are able to ‘walk’ on a track [41], [42], and to perform computations [43], [44]. This thesis does not study dynamic devices but strand-exchange is used in chapter 5 as a mechanism to induce competition between two assembly targets.

1.2.3 DNA Origami

DNA origami was first introduced in 2006 by Paul Rothemund [45] and has become one of the most widely used assembly techniques in DNA nanotechnology. In this scheme, many short oligos, known as staples, are designed to individually interact with two or more regions on a long scaffold strand (Figure 1.4). Through hybridisation with the staples, the distant interaction domains are brought into close proximity, pinning the scaffold into the desired shape. Rothemund created a variety of 2D shapes including a smiley face and a world map using different

sets of staples to fold a circular strand derived from a bacteriophage genome. In a typical assembly protocol, the staple strands are mixed in an excess (100-fold in Rothemund’s case) with the scaffold in solution and heated to 95°C before being cooled down to room temperature, a process referred to as annealing. The initial ‘denaturing’ step is required to relieve any secondary structure in the scaffold strand. While Rothemund was able to achieve good yields (> 90%) in under two hours, the optimal cooling rate will generally depend on the complexity of the target structure.

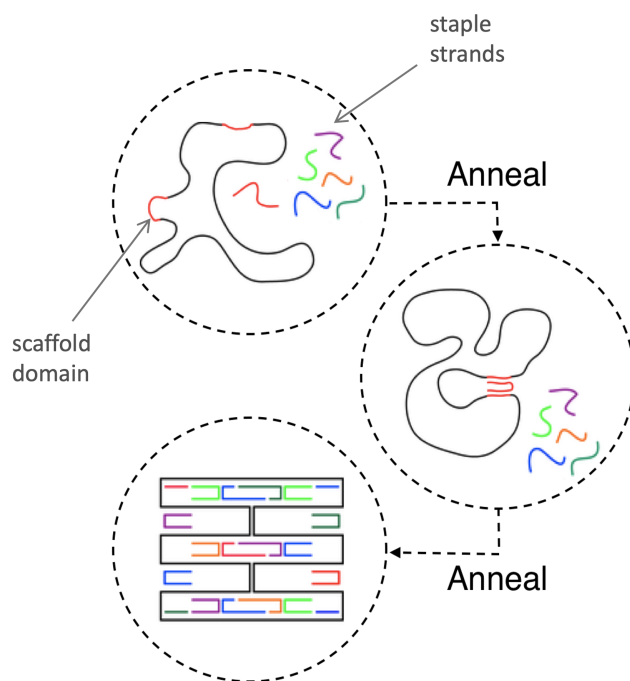


Figure 1.4: Illustration of DNA Origami. Interaction domains on ‘staple’ strands are designed to be complementary to distant regions on the scaffold strand. Hybridisation of domains, usually under an annealing protocol, pins the scaffold into the target shape.

Given a geometry between parallel duplexes, the helical pitch of DNA dictates where crossovers can be positioned. The spacing of crossovers can be adjusted to extend origami to three dimensions [46] and to induce twist and curvature [47]. A plethora of 2D and 3D shapes have been assembled, including a cage formed by linking planar sheets of DNA using flexible single-stranded hinges [48], structures with complex curvatures such as ellipsoids and nano-flasks [49], polyhedra made of wireframe meshes [50], and multi-layer origami on square, honey-comb

and hexagonal lattices [46], [51], [52]. Crucially, these complex structures can be assembled in one-pot reactions. This is in contrast to tile-based approaches, where good yields often require multi-stage assembly with careful stoichiometric control through purification steps.

There have been some experimental studies on the folding behaviour of DNA origami. These studies use several techniques to characterise both folding intermediaries and end-products of assembly. I give a brief summary of these methods below.

- Gel electrophoresis is used to assess the extent of successful assembly as a whole. In this ensemble approach, the samples are placed in a porous gel and an electric field is applied across the two ends. Different products in the sample migrate at different rates and are thus separated based on their size, charge, and shape. These can be visualised by staining the gel with a fluorescent molecule that binds non-specifically to DNA. Alternatively, specific strands can be labelled with fluorescent dyes prior to assembly. These methods are usually used to assess the different species present in the sample, with the assumption that the gel environment or fluorescent labels do not alter the behaviour of those species.
- Fluorescence spectroscopy probes monitor assembly in real-time. Intercalating dyes that bind differently to dsDNA and ssDNA can be used to track the extent of hybridisation in the whole structure. Specific information about parts of the structure can also be obtained using the Förster resonance energy transfer (FRET) mechanism. In this scheme, strands modified with donor/acceptor pairs of dyes undergo a distance-based change in fluorescence. However, information can only be obtained on the population level and folding behaviour may be altered by the presence of such dyes.
- Atomic Force Microscopy (AFM) is a single-molecule technique used to image the samples at atomic resolution by measuring intermolecular forces between the tip of the probe and the sample. The samples need to be prepared on an atomically flat surface, limiting the scope of this method to 2D objects.

- Transmission electron microscopy (TEM) and cryo-EM can be used to reconstruct 3D images of single structures and provide information about structural defects.

In a significant study, the Dietz group showed that rapid isothermal folding of origami is possible in a narrow temperature range [53]. They investigated a series of 3D objects and a 2D sheet, using fluorescence microscopy to characterise transitions in real-time, and gel electrophoresis and AFM to confirm the final structures. By assessing the extent of folding/unfolding under step-wise temperature ramps, they found a high degree of cooperativity between staples: the temperature ranges at which the transition occurred was narrower ($\Delta T \sim 4\text{C}$) than the range of melting temperatures of individual binding domains ($\Delta T > 30\text{C}$). They also observed a high degree of hysteresis between annealing and melting transitions, pointing to non-equilibrium processes. The hysteresis widened under faster temperature ramps, mainly due to a shift in the folding transition. This points to the existence of free energy barriers that play a more prominent role in the folding process. They confirmed this in isothermal assembly protocols in which the temperature was chosen based on peaks of annealing and melting fluorometric curves. They were able to fold structures at a few degrees below the peak of the annealing transition on object-dependent timescales. The 2D sheet was assembled in under 5 minutes, whereas the 3D objects were assembled in ~ 1 hour. They were also able to unfold all structures in under 10 minutes by incubation at temperatures set to the peak of the melting transition. These results suggest the existence of a nucleation barrier, particularly in the case of 3D origami.

Song *et al* [54] and Wah *et al* [55] developed methods to monitor the folding pathway of 2D DNA origami by AFM imaging at different stages of the transition. They confirm the existence of intermediate stages of folding, where staples cooperate locally to form parts of the structure. In the case of Ref. [54], they heat a pre-formed origami in liquid directly on mica surface and transiently image the structure. They report an unfolding pathway that begins at the edges of the origami, ending with seam staples that connect the two sides. However, this pathway is highly correlated

to the hybridisation energies of the domains of their particular structure. The fact that the transition takes place while immobilised on mica makes it difficult to draw conclusions about origami in bulk solution. In the case of Ref. [55], they report that folding and unfolding happen through reverse pathways.

Wei *et al* [56] did a more systematic study of origami using FRET pairs. They observed a higher degree of hysteresis in the transition during melting and annealing in their 3D structures compared to their 2D structures. This was attributed to the more complex nature of connections and the higher degree of cooperativity required in the 3D case. In the 2D case, they also gauge the effect of omitting certain staples from the design. They found that the effect of omitting distant staples is minimal compared to those near the FRET pair. However, drawing conclusions about the folding pathway is difficult using this method. Leaving out staples near the FRET pair leaves the scaffold single-stranded, which inevitably increases the flexibility of that particular region and causes reduced resonance between the FRET pair.

More recently, Schneider *et al* [57] completely mapped the incorporation time of staples of a 3D helix bundle under an isothermal folding protocol. They used all possible combinations of FRET-labelled strands between terminal domains on individual strands and in pairwise combinations of staples. They report a specific sequence of events in which staples are cooperatively incorporated in groups. Staples that form longer connections generally form later but their presence mediates the binding of other staples that connect similar regions of the scaffold. Incorporation times of staples were also found to be only weakly dependent on their hybridisation free energy pointing to a high degree of cooperativity. They also investigated the effect of removing crossovers (by cutting staples into pieces so that they no longer form connections) on folding intermediaries that were present in gel electrophoresis measurements. Truncating the fastest staples (those that form early when using the full set) led to the disappearance of the intermediate band and dramatically increased total folding time. Truncating the slowest staples had no effect on the time taken to form the intermediate structure but caused delays in folding the

full structure. These results confirm the existence of local cooperativity and the importance of binding intermediaries in origami folding.

Studies mentioned above utilise designs in which every interaction domain is unique. In such studies, the target structure is designed as the global free energy minimum, and assembly protocols are optimised to reduce the extent of misbound structures that often negatively affect yields through kinetic traps. Introducing domain redundancy to reduce the absolute addressability of interactions can lead to interesting insights. Dunn *et al* [58] investigated the folding pathway of a polymorphic tile, where two copies of the same circular scaffold were ligated, allowing staples to form crosslinks between the duplicated region of the ‘dimer’ scaffold. Depending on the binding configuration of the staples, a multitude of distinguishable shapes with high yield are made possible by design, with each shape representing a local free energy minimum in a vast free energy landscape. The distribution of shapes observed through AFM imaging was well-predicted by the domain-level model of Dannenberg *et al* [59]. The change in this distribution as certain crosslinks were modified was used to probe the kinetics of assembly. They found that folding could be biased towards certain shapes by adjusting the strength of crosslinks. Staples that formed long-range connections were shown to be particularly important in determining the dominant pathway and the ultimate distribution. Interestingly, certain modifications did not alter the relative thermodynamic stability of one particular fold over another, but strongly shifted the distribution towards a particular shape. This work demonstrates the importance of the kinetics of self-assembly, whereby deliberately altering intermediate states by allowing certain connections to form earlier can significantly alter the outcome.

1.2.4 DNA Bricks

The success of DNA origami compared to previous tile-based approaches was attributed to the existence of efficient folding pathways, made possible by the cooperativity of forming connections on the scaffold strands. While the exact mechanisms of folding are still not well-understood, intricate structures could

be reliably produced without any need for perfect stoichiometric control. The high excess of staples used in origami was thought to contribute to its success by allowing the reversal of misfolded structures through strand exchange. This dogma was challenged by the DNA ‘brick’ experiments of Ke *et al* [60], [61], in which a wide array of 2D and 3D structures were successfully assembled using only short single strands. In this scheme, each strand can bind via four uniquely addressable binding domains to four other strands to form lattices with different geometries [62]. Importantly, the strands are annealed in one-pot assembly protocols without the need for purification steps.

The success of DNA bricks in producing structures that rivalled DNA origami in size and complexity [63] with as many as 10,000 unique components [64] was somewhat surprising and motivated several theoretical investigations. The Frenkel group used theoretical arguments [65], [66] as well as lattice-based [67], [68] and off-lattice [69] patchy-particle models to show the importance of nucleation in brick assembly. They predict the existence of nucleation barriers that depend on the temperature and the coordination number of the lattice. The barrier needs to be small enough to be surmountable in experimental time scales, and yet large enough for nucleation to be a rare process to avoid monomer depletion. This leads to a narrow temperature range under which successful assembly can occur. They also make arguments about the role of misbinding that applies to addressable assembly more broadly. The use of uniquely-addressable components results in extra entropic penalties compared to equivalent single-component systems, since there is only one way to arrange the distinguishable components into the target shape [65]. This extra entropy pushes the assembly temperature to lower values at which misbinding is difficult to reverse. Annealing protocols aid assembly by providing conditions under which the thermodynamic drive for designed reactions is high enough and yet errors can still be corrected. By passing through this temperature window, sufficient correct binding can take place in the initial stages before the process can be completed at lower temperatures. Fonseca *et al* [70] reach similar conclusions using a kinetic model that is parametrised with more realistic free energies.

1.3 Modelling Tools in DNA Nanotechnology

Many theoretical models of DNA that include various levels of detail have been developed. The selection of the model is dictated by the length and time scales of the problem at hand. At the finest level, quantum mechanical approaches can provide detailed information about the nature of interactions, for example on base-stacking energies [71]. These first-principles models are the gold standard in modelling DNA but they come at a high computational cost and are often used to parametrise force fields in atomistic models [72]. In atomistic frameworks such as AMBER [73] or CHARMM [74], positions of individual atoms are evolved according to force fields using computational techniques such as molecular dynamics or the Monte Carlo method. The force fields are either derived experimentally or from more granular first-principles methods. Recently, these models have been used to study systems as large as DNA origami [75], [76], albeit for \sim microseconds. Generally, all-atom molecular dynamics simulations can provide useful information about the mechanical behaviour of small DNA systems in thermal equilibrium and interactions with the solvent [76], [77]. However, sampling rare events at the length scales relevant to DNA nanotechnology is still far beyond the reach of such detailed frameworks.

At the other extreme, there are continuum models that describe DNA as a semi-flexible polymer. These include the freely-jointed chain [78] or the worm-like chain [79] and are mainly suitable for studying large-scale DNA behaviour on length-scales much greater than the persistence length of DNA. These ideas have been incorporated into modelling frameworks such as CanDo [80], which is a finite-element model regularly used in the field for structural prediction of DNA origami structures [81].

Statistical models of DNA also play an important role in the field. Particularly, the model of Santalucia *et al* [82], [83] is often used to describe the thermodynamics of DNA denaturation. The model considers the transition between hybridised and single-stranded states as a perfect two-state transition. It assumes that the stability of a given base pair is dependent only on its adjacent bases, allowing for 10 combinations of such nearest-neighbours (NN). This NN model is fitted to experimental data

on a large number of DNA sequences and provides a thermodynamic database for calculating the standard state changes in free energy during the melting transition of any given sequence. While simple, the model is widely considered to be a reliable predictor of melting temperatures of duplexes, since the model is fitted to UV-spectrometry of the melting transition of short DNA strands. However, the existence of certain outlying sequences and recent advances in experimental methods have motivated studies that go beyond the simple nearest-neighbour approach [84], [85].

Between the two extremes are coarse-grained models of DNA. These models aim to compromise between the detailed description of atomistic models and the efficiency of continuum models. The level of abstraction is often dictated by the property of interest and the system being investigated. The common theme is that degrees of freedom of a finer model are integrated out into some larger basic unit to allow faster computation with reduced coordinates while keeping the unit small enough to capture relevant effects. Many such models have been introduced with the basic unit ranging from a whole domain [59], [86], to 7 base-pair segments [87], to a single base-pair [88], to single nucleotides [89]–[92]. There are several considerations to keep in mind when deciding on a level of coarse-graining. Some models describe both ssDNA and dsDNA and are able to capture hybridisation, while others are designed for structural predictions and can only represent thermal fluctuations in the assembled state. Another factor to consider is whether the solvent is modelled explicitly or approximated using a mean-field approach. The models can also be classified by their approach to parametrisation: bottom-up models derive their effective interactions from models at a finer level of granularity, while top-down models fit their parameters to experimental data such as melting temperatures and persistence lengths.

There are two models that are of particular importance to this work: oxDNA, which is a nucleotide-level model developed by Thomas Ouldridge [91] and the domain-level model (DLM) developed by Frits Dannenberg [59]. oxDNA is used in chapter 2 to model a small DNA motif and its incorporation within a larger scaffolded DNA structure. The DLM is extended in chapter 3, and used in chapters

4 and 5 to study the folding of DNA origami. I will discuss oxDNA briefly here and leave the discussion of the DLM to chapter 3.

1.3.1 The oxDNA Model

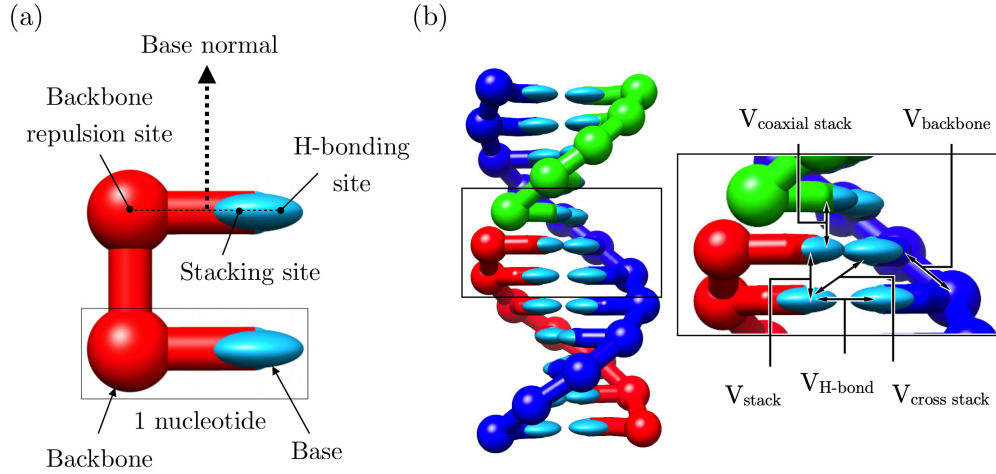


Figure 1.5: The oxDNA model. Representation of (a) two oxDNA nucleotides and (b) an 11 base-pair double helix that illustrates the various interactions in the oxDNA model.

As shown in Figure 1.5, each nucleotide in the oxDNA model is represented as a rigid body that can interact with other nucleotides in a pair-wise fashion through three interaction sites: a backbone connectivity site representing the sugar-phosphate backbone, and two sites responsible for stacking and hydrogen bonding representing the base. The full interaction potential in the most recent version (oxDNA2) is given by:

$$V_{\text{oxDNA2}} = \sum_{\text{NN}} (V_{\text{backbone}} + V_{\text{stack}} + V_{\text{exc}}) + \sum_{\text{other}} (V_{\text{HB}} + V_{\text{cross stack}} + V_{\text{exc}} + V_{\text{coaxial stack}} + V_{\text{DH}}),$$

where the first sum runs over adjacent nucleotides on a strand and the second sum runs over all other pairs of nucleotides. The functional forms of these interaction potentials can be found in Ref. [93].

The potentials are designed to provide a physically realistic description of the effects of the underlying chemical interactions. The excluded volume potential

applies to all nucleotides. The backbone term is isotropic, and the stacking and hydrogen bonding terms are anisotropic. The orientation of each base is specified by a normal vector, giving a notional plane for the base. The relative angles of these planes are used to modulate interactions. For example, hydrogen bonding is modulated by angular terms that favour linear alignment of all four backbone and hydrogen-bonding sites and anti-alignment of the normal vectors, while stacking is modulated by angular terms that favour the alignment of base normals with the vector between stacking sites, encouraging co-planar stacks. The Debye-Hückel term was introduced in oxDNA2 to describe screened electrostatic interactions that take into account different salt conditions. Additionally, different widths for the major and minor double helical grooves were introduced resulting in a more realistic description of duplex DNA geometry. This was achieved by adjusting the position of the backbone site without changing the duplex radius, causing the three interaction sites to lie in a plane rather than a straight line. Interaction energies have been fitted to experimental data in order to reproduce much of the structural, mechanical and thermodynamic properties of single- and double-stranded DNA [91], and to incorporate the sequence-dependence of Watson-Crick base pairing interaction strengths [94].

The model is widely used in the field and has been applied to study a variety of basic biophysical mechanisms and structural and dynamic systems. It has been applied to probe the mechanisms of DNA hybridisation [95], hair-pin formation [96], and toehold-mediated strand displacement [97], making experimentally accurate predictions about the dependence of displacement rate on toehold length. On the structural front, oxDNA was used to accurately predict conformations of star-shaped tiles and their dependence on bulge loop size [98], and ‘hinged’ motion in joint structures [99]. It has also proved an excellent tool to probe the operation of dynamic systems such as DNA tweezers [100] and DNA walkers [101], [102]. DNA origami has also been modelled using oxDNA. The structural and mechanical properties of several large assemblies were well predicted [103], and the assembly pathway of a small DNA origami was also directly simulated [104]. Although the study was done with

extremely high strands concentrations to increase efficiency, it provides interesting insights into kinetic phenomena and the types of misbinding that can occur.

1.4 Simulation Techniques

The oxDNA package is publicly available for use with both CPU and GPU implementations. My work was carried out using the CPU implementation and I will give a brief overview of the simulation methods relevant to chapter 2. The simulation techniques such as the Gillespie algorithm [105] used in other chapters are described later in chapter 3.

Simulations in oxDNA can be carried out using molecular dynamics or Monte Carlo implementations. Molecular dynamics simulations proceed by numerically integrating Newton’s equations of motion for each particle, with forces determined by the derivative of the oxDNA potential. To enforce a set temperature in the implicit solvent, a ‘thermostat’ algorithm is employed that resets a subset of the particles’ velocities to that of the Maxwell-Boltzmann distribution at the set temperature. This is intended to mimic collisions between the particles and the solvent; the frequency at which the velocities are reset is chosen based on considerations of computational efficiency and realistic diffusive dynamics.

However, if equilibrium properties of the system, rather than the dynamics are of interest, then Monte Carlo simulations [106] are often the method of choice and can provide increased efficiency. The purpose of Monte Carlo simulations is to generate a set of microstates that are representative of the equilibrium probability distribution for a given ensemble. The effect of the solvent as the heat bath is therefore inherently included if the generated set of microstates reflects the equilibrium distribution at the set temperature. In the language of statistical mechanics, this corresponds to sampling the microstates of the canonical ensemble. For an N-body system with Hamiltonian $H(\vec{p}^N, \vec{r}^N) = T(\vec{p}^N) + U(\vec{r}^N)$, elementary statistical mechanics dictates that the equilibrium probability of the system in configuration \vec{r}^N is given by $P^{\text{eq}}(\vec{r}^N) = e^{-\beta U(\vec{r}^N)} / Z$, where $U(\vec{r}^N)$ is the internal energy of the microstate, Z is the partition function of the ensemble, and $\beta = 1/k_B T$

is the inverse of thermal energy. The question remains on how to generate a new microstate from an existing one in a way that correctly samples the equilibrium distribution. This generation of a sequence of microstates can be considered a Markov process with the probability of transition from configuration μ to ν given by $\pi_{\mu\nu}$. To sample the equilibrium distribution, the Markov process must reproduce $P^{\text{eq}}(\vec{r}^N)$ as a steady-state distribution [107]. One way this can be ensured is through the condition of detailed balance:

$$P_{\mu}^{\text{eq}}\pi_{\mu\nu} = P_{\nu}^{\text{eq}}\pi_{\nu\mu} \quad (1.1)$$

Metropolis Monte Carlo Algorithm

An elegant computational method that ensures detailed balance was proposed in 1953 by Metropolis *et al* [108] and remains popular to this day. Given an initial configuration μ , a new configuration ν is generated with probability $P_{\mu\nu}^{\text{trial}}$ and accepted with probability $P_{\mu\nu}^{\text{acc}}$. The transition probability $\pi_{\mu\nu}$ can be considered the product of the two terms $\pi_{\mu\nu} = P_{\mu\nu}^{\text{trial}}P_{\mu\nu}^{\text{acc}}$. In the Metropolis scheme, these are chosen to satisfy:

$$P_{\mu\nu}^{\text{trial}} = P_{\nu\mu}^{\text{trial}}, \quad P_{\mu\nu}^{\text{acc}} = \min\{1, e^{-\beta(E_{\nu}-E_{\mu})}\}, \quad (1.2)$$

where E_{μ} and E_{ν} are the energies $U(\vec{r}^N)$ of each configuration. Since the choice of trial moves is arbitrary, large and even unphysical moves can be used to equilibrate the system quickly. This makes the Metropolis algorithm extremely popular.

Virtual Move Monte Carlo

The Metropolis algorithm is not suitable for studying systems of strongly bound particles [109] such as strands in oxDNA. Most large moves of individual monomers are rejected because they cause a significant increase in energy. To increase the efficiency of sampling the collective diffusion of nucleotides that are strongly bound as strands, oxDNA implements a cluster move algorithm. In virtual move Monte Carlo (VMMC) developed by Whitlam and Giessler [109], [110], trial moves involve a collection of particles. To implement collective moves, a ‘seed’ particle is randomly

chosen and a cluster of particles is iteratively grown based on the strength of their pairwise interactions. A collective move is then attempted for the cluster as a whole, preserving their pairwise interactions. VMMC is particularly useful because both the interactions in the initial state and the energy change due to the move are considered when building a cluster. This makes the moves more sensitive to the underlying potential energy landscape, resulting in a better approximation of diffusive motion.

Umbrella Sampling

While VMMC increases the sampling efficiency in oxDNA simulations, the system can still get trapped in metastable states. This can present a problem if we are interested in exploring regions of configurational space that are separated by high energy barriers. The umbrella sampling method [111] can be used to bias simulations and in effect flatten energy barriers. Given we are interested in states defined by an order parameter $A(\vec{r}^N)$, the ensemble average in the canonical ensemble is given by:

$$\langle A \rangle = \int d\vec{r}^N P^{\text{eq}}(\vec{r}^N) A(\vec{r}^N) = \frac{1}{Z} \int d\vec{r}^N e^{-\beta U(\vec{r}^N)} A(\vec{r}^N). \quad (1.3)$$

The umbrella sampling method speeds up transitions between metastable states by sampling a biased probability distribution

$$P^{\text{biased}}(\vec{r}^N) = w[A] P^{\text{eq}}(\vec{r}^N) = e^{-\beta[U(\vec{r}^N) + V^{\text{bias}}(\vec{r}^N)]}, \quad (1.4)$$

where $w[A]$ is a weighting factor, and $V^{\text{bias}}(\vec{r}^N)$ is determined by the value of A in configuration \vec{r}^N . In practice, sampling the biased distribution involved changing the acceptance criterion of the MC moves in accordance with the biasing weights of the initial and final configurations. The equilibrium distribution can be retrieved from the biased distribution after the MC run as:

$$P^{\text{eq}}(A) = \frac{P^{\text{biased}}(A)}{c_w w[A]}, \quad (1.5)$$

where c_w are normalisation factors. The idea is to choose a set of weights that leads to a flat distribution ($P^{\text{biased}}(A) = \text{const}$), but the correct choice requires *a priori* knowledge of the equilibrium distribution. Usually, a starting guess is

used for the set of weights and iteratively improved until a flatness criterion is reached. Given that the flatness criterion is met, we can recover the unbiased ensemble average of A through:

$$\langle A \rangle = \frac{\langle A/w \rangle_w}{\langle 1/w \rangle_w}. \quad (1.6)$$

Weighted Histogram Analysis Method

A single umbrella sampling simulation may be insufficient for exploring the whole range of values of the order parameter A . In such cases, we can break up the simulation into individual windows, and explore a subset of the range of A in each window. Since the full partition function is not known, only relative probabilities can be obtained in each window. Therefore, it is essential that the windows explore overlapping ranges of A . The results can then be combined using the WHAM equations of Kumar *et al* [112]. The full derivation is tedious [113] but it relies on writing the optimal equilibrium distribution $P^{(0)}(A)$ over the whole range A as a weighted sum of the unbiased distributions obtained in each window. The weighting factors are subject to normalisation conditions and are chosen to minimise the variance of $P^{(0)}(A)$. For K umbrella sampling windows, this results in the following WHAM equations:

$$P^{(0)}(A) = \frac{\sum_k N_k(A)}{\sum_k n_k w_k(A) e^{f_k}}, \quad (1.7)$$

where $N_k(A)$ is the count of order parameter value A in the k^{th} window, n_k is the number of samples in the k^{th} window, and f_k is the free energy of the k^{th} window, given by:

$$f_k = -\ln \sum_{\{A\}} w_k(A) P^{(0)}(A), \quad (1.8)$$

where the sum is taken over all values of A , and $w_k(A)$ is the biasing weight in the k^{th} window. These two equations must be solved self-consistently. Beginning with a trial set f_k , the process is repeated until a convergence criterion is met.

Single Histogram Reweighting

The single-histogram re-weighting method, developed by Ferrenberg and Swendsen [114], can be used to obtain ensemble averages of order parameters at different temperatures. During the simulation, states are grouped according to their energies E and values of the order parameter of interest A . If we assume that the density of states is independent of temperature, the probability distribution $P(E, A; T_0)$, where T_0 is the simulation temperature, can then be used to obtain the probability distribution $P(E, A; T) \propto P(E, A; T_0)e^{(\beta - \beta_0)E}$ at another temperature T . The extrapolation process is more accurate if the target temperature T is close to T_0 due to the increased number of shared states with energy E . In this thesis, I used this method once in chapter 2 to estimate the free energy of T-motif formation at 4°C from oxDNA simulations at $T_0=25^\circ\text{C}$. To do this, I extrapolated the results to nearby temperatures $T \in \{19, 21, 23, 27, 29, 31\}^\circ\text{C}$ and applied a linear fit to the resulting free energies.

1.5 Thesis Outline

The main purpose of this work is to use theoretical methods to shed light on assembly pathways in DNA origami, an important technique in the construction of DNA nanostructures. The technique is routinely used to precisely control matter at the nanoscale with design principles that mainly rely on empirical rules. To achieve the ultimate aim of producing precise molecular machinery, reconfigurable structures that respond to external stimuli need to be reliably constructed, and the dynamics of their operations reliably predicted.

Chapter 2 is geared towards the design of a reconfigurable origami based on a relatively new structural motif: the DNA T-junction. In this work, I apply the oxDNA model to study and characterise the structural motif to inform its use within larger constructions. The work forms the basis of two publications, which were done in collaboration with Katherine Young who did the majority of the experiments. The first publication [115] (under review) concerns characterising the conformations

and stability of a variety of T-motifs. The second publication [116] uses the findings to construct a higher-order origami-like object that incorporates domain redundancy in its design. The aim is to shed some light on the extent of cooperativity as the addressability of interactions is reduced. Interactions are generally weaker in these systems, which also allows them to be assembled at room temperature.

Chapter 3 will lay down the modelling framework used to study DNA origami systems in chapters 4 and 5. This domain-level model is an extension of that of Ref. [59], which was previously used to study the folding pathway of a polymorphic tile [58]. I extend the model to staples with three domains and apply the model to several experimental systems in chapter 4. I focus on the thermodynamic properties of DNA origami, particularly with regard to the free energy landscape and its implication for optimising assembly protocols. In chapter 5, I incorporate a description of strand exchange into the model, effectively allowing two sets of staples to compete during the folding process. This work is again done in collaboration with Katherine, who did all of the experiments. We show that coating the scaffold strand with single strands prior to mixing with the staples allows error-free folding to occur at lower temperatures. The model predicts the existence of a nucleation barrier and the overall predictions are largely consistent with experimental assembly times.

2

Reconfigurable T-junction Origami

Contents

2.1	Introduction	28
2.2	Characterising T-motifs	30
2.2.1	Conformations	33
2.2.2	Free Energy	37
2.3	Characterising Linkers	40
2.4	T-junction Origami	44
2.5	Summary and Outlook	48

The work described in this section was jointly carried out by me and my colleagues Katherine G. Young and Jonathan Bath, who did the majority of the experimental work. The results are published in Refs. [115] and [116]. In the case of Ref. [115], atomistic simulations are currently being carried out in the Doye group at Oxford to supplement the oxDNA simulations. We are awaiting the final results to confirm the structures are stable. In the case of Ref. [116], the final T-junction origami structures were also studied under the effect of tensile forces [117] in molecular dynamics simulations in oxDNA. These simulations were carried out by William Sant in the Doye group, and I took part in data analysis and interpretation of results.

2.1 Introduction

The initial aim of this work was to probe the mechanisms of self-assembly of DNA origami by creating a system in which the interaction domains could be easily adjusted. This is difficult to do in origami systems since the scaffold strand is usually derived from a natural source. A novel T-shaped junction that was introduced by Hamada *et al* [118] in 2009 motivated us to design such a system and avoid the prohibitively high cost of manufacturing (multiple) custom scaffold strands. The work in this section is centred around characterising and using this ‘T-junction’ as a structural motif. Unlike previous motifs that rely on reciprocal strand exchange between parallel helices, the T-motif consists of a bulge loop hybridised to a single-stranded sticky end, resulting in a right-angled geometry (see Figure 2.1a). We exploit the fact that both these elements can be designed with arbitrary base sequences and incorporate them within larger assemblies to adjust interactions.

The T-motif and the initially proposed origami tile are shown in Figure 2.1a. The bulge loop and sticky end connect two separate duplexes at a T-junction. The motif is stabilised by coaxial stacking of bases at the junction into a well-defined geometry. The proposed origami tile is shown in Figure 2.1b. This design consists of a pre-assembled, circular, scaffolded structure consisting of alternating duplex and bulge loop segments. To create custom base sequences, the scaffold strand is hybridised to ‘adaptor’ strands that contain the sequences of the bulge loops. To fold the structure into a rectangular shape, a set of pre-assembled ‘linkers’ are used to form connections across the bulge loops. Each linker is formed by two strands and consists of a duplex region flanked by two sticky ends on either side. These sticky ends can be of different polarity depending on the strand on which they occur. The polarity of the sticky end (5′ or 3′) determines the type of interaction with the bulge loop. The type of interaction along with the helical pitch of B-DNA dictate the length of duplex regions that can be used in the tile. To understand the possible design space, we did a systematic study of T-junctions of both polarities and the effect of linker length using a combination of simulation and experiment. The results are presented in the next sections.

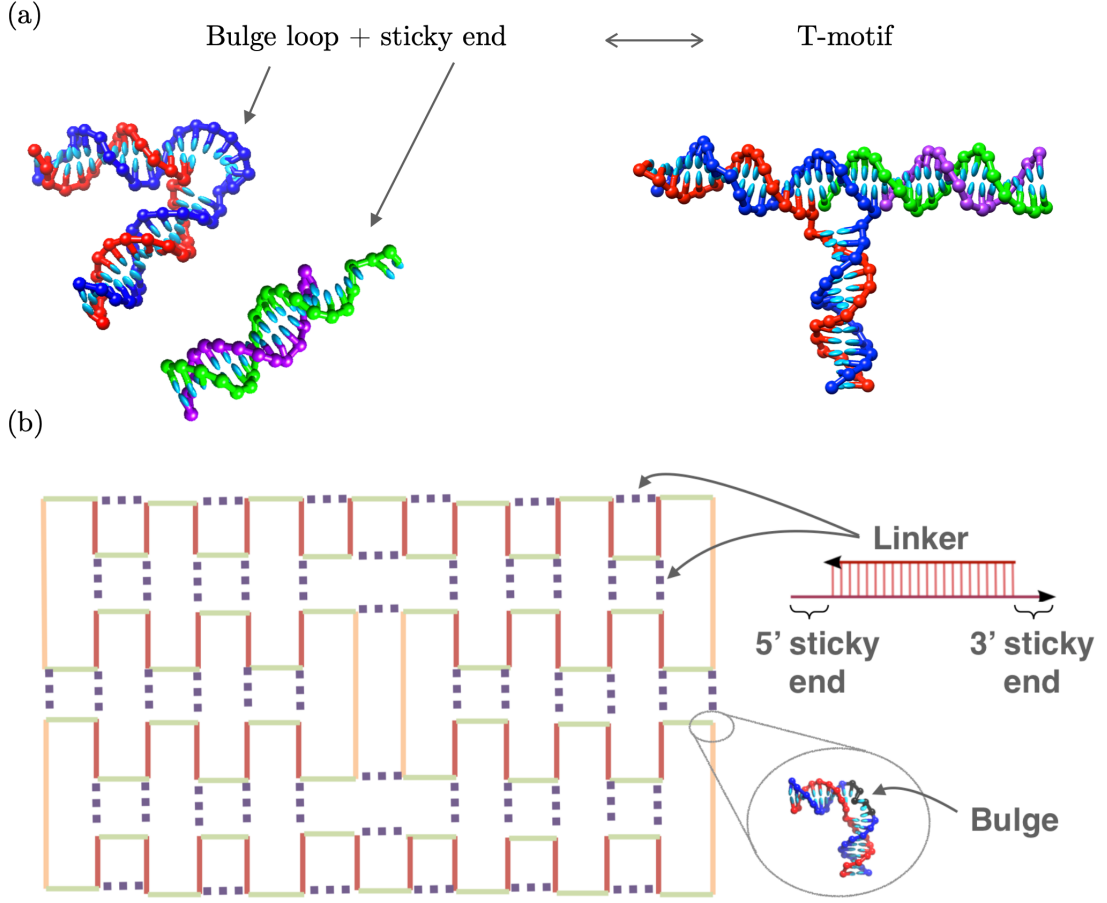


Figure 2.1: The T-motif and proposed Tile. (a) The T-motif is formed by the interaction of a bulge loop on one duplex with the sticky end on another duplex. (b) The proposed T-junction origami. A long ‘scaffold’ strand of DNA runs through all red, green and yellow regions, fully hybridised to shorter ‘adaptor’ strands in duplexes separated by bulge loops. The single-stranded bulge regions are encoded into the shorter adaptor strands. To fold the structure, a set of ‘linkers’ (shown as dotted lines) interact with the bulge loops through sticky ends on either side to form T-junctions.

There have been other attempts at utilising T-junctions in larger assemblies. In the original work, Hamada *et al* [118] created a variety of periodic and finite structures. They used basic units formed with two strands that include bulge loops and sticky ends to create periodic lattices. By adjusting the length of the non-interacting duplex regions, they were able to effectively control the ‘coordinates’ of their lattice to form both ladder-like and wheel-like shapes. Similar 1D chain arrays were created in Ref. [119], with the basic unit of the lattice formed by a self-interacting single strand that includes a bulge loop and a sticky end. The

Mao group [120], [121] were able to extend this to three dimensions and create a variety of polyhedra with T-junctions at the vertices. Recently, T-junctions were used in an algorithmic self-assembly protocol on a 2D DNA lattice to calculate the value of π [122]. However, all of these studies use T-junctions with 5' polarity. The reasoning for this is not clear since the use of both polarities would increase the versatility of shapes that could be designed.

2.2 Characterising T-motifs

There are several factors that determine the geometry and stability of T-junctions. In Figure 2.2, I present a simple scheme that can be used to clarify the discussion and characterise the T-motifs. I use the notation $\mathbf{T}_{\mathcal{P}}^L[n_{bp}, s]$ to specify a T-motif. This refers to a T-motif in which the size of the bulge loop and sticky end are L bases, and n_{bp} base pairs have formed between the bulge and sticky end at the junction. The polarity $\mathcal{P} \in \{3', 5'\}$ of the sticky end is consequential because it determines the groove that the junction spans. Because of the difference in groove width, the steric constraints and optimal size are different for the two cases. If the junction domain is too large to ‘fit’ in the groove, the stacking interactions s_A and s_B can no longer simultaneously form to stabilise the junction due to steric constraints. On the other hand, there is a competing stacking interaction s_C that can distort the geometry of smaller junctions. Therefore, the T-motifs are also assigned a stacking state $s \in \{s_0, s_A, s_B, s_C, s_{AB}, s_{BC}\}$. A well-formed T-motif is in state s_{AB} (both of the motifs in Figure 2.2b are in this state), s_0 denotes no coaxial stacking, s_A and s_B are partially stacked states, s_C is the misfolded state (Figure 2.2d), and s_{BC} is a possible combination that will be discussed later. Snapshots of all states are shown in Figure 2.5.

I use oxDNA to study the thermodynamics stability and geometry of T-motifs with bulge sizes between 3 and 9 nucleotides. All-atom simulations are too computationally expensive to probe the hybridisation of the bulge loop to the sticky end. On the other hand, models that coarse-grain to units larger than a nucleotide don’t provide a good description of duplex fraying and coaxial stacking,

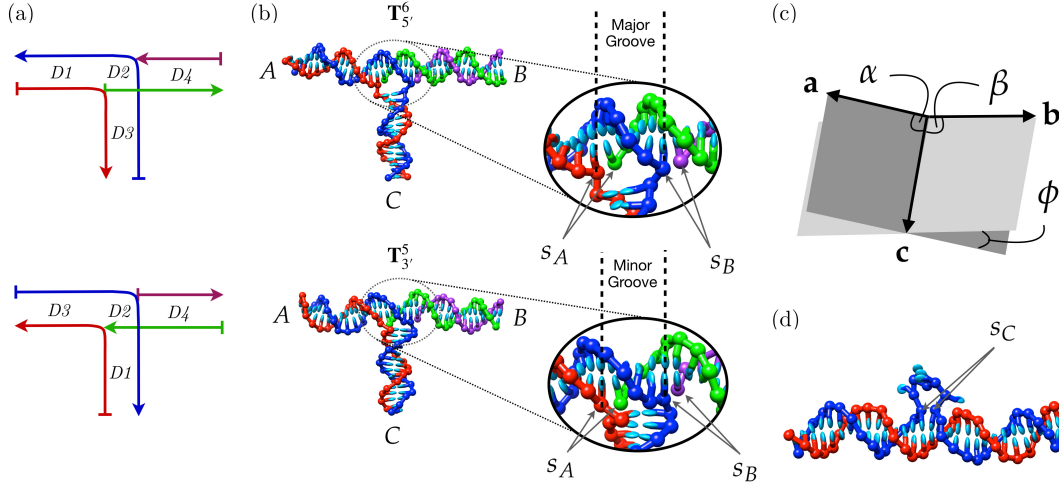


Figure 2.2: T-motif characterisation scheme. (a) A fully formed T-motif consists of four domains of interaction ($D1D4$) between four strands. To form the motif, the complex ($D1D2D3$) containing the bulge interacts at the junction domain $D2$ with the complex ($D2D4$) containing the sticky end. The coloured arrows point from 5' to 3' directions. The polarity of the sticky end determines the geometry of the junction. If the sticky end terminates in the 5' direction, the side branch of the junction is duplex $D3$ and the main branch is the multi-domain 'nicked' duplex $D1D2D4$. For a 3' sticky end, the side branch is $D1$ and the main branch is $D3D2D4$. (b) The motif can be seen as a main branch AB and a side branch C . The geometry of B-DNA means that for a 5' sticky end, C will straddle AB at the major groove, whereas for a 3' sticky end, the junction will occur at the minor groove. The main branch is stabilised by coaxial stacking sites s_A and s_B on either side of $D2$. The notation $T_{\mathcal{P}}^L$ refers to a T-motif with $D2$ domain of size L base pairs and sticky end polarity $\mathcal{P} \in \{3', 5'\}$. (c) Vectors \mathbf{a} , \mathbf{b} and \mathbf{c} correspond to the orientations of the three arms of the junction. They are calculated by summing over base-normal vectors of the nucleotides in the 11 middle base pairs of the corresponding domain. The angles α and β are defined as positive for $(\mathbf{a} \times \mathbf{c}) \cdot (\mathbf{b} \times \mathbf{c}) < 0$. The dihedral angle ϕ is positive for $[(\mathbf{a} \times \mathbf{c}) \times (\mathbf{b} \times \mathbf{c})] \cdot \mathbf{c} < 0$. All three angles have positive values in the configuration shown here. (d) The coaxial stacking site s_C can occur between the base pairs of $D1$ and $D3$, which competes with the formation of the T-motif, especially for shorter bulge loops. Taken from Ref [115].

both of which are important for T-motif formation. This makes oxDNA an easy choice. It has an explicit description of stacking and captures the conformations of single-stranded DNA [91], which is crucial for modelling the sticky end. It accurately predicts the tendency of bound duplexes to fray near their melting temperature and captures the sequence-dependent thermodynamics of hybridisation [94], which we use for direct comparison to experiment. It has been previously applied to characterise bending angles of bulge loops as a function of their size [123] and can accurately predict the geometries of DNA polyhedra that feature

flexible loops. Importantly, the most recent version of the model (oxDNA2) includes an adapted geometric representation that captures the difference in widths of the major and minor grooves [93]. This will be crucial to this study when comparing the behaviour of the 5' and 3' variants of T-junctions.

There are some effects that the oxDNA model does not capture. In oxDNA2, new experimental data was used to re-parametrise the oxDNA potential to account for the different stacking strengths of consecutive A/T bases along single strands. But this was not applied to the coaxial stacking interactions, which are still independent of base identity. Therefore, we will not be able to gauge the effect of sequence on the stacking states described above. Hoogsteen pairing interactions and sugar-edge bonds are also not included in the model. These have been shown to play a role in determining the flexibility of RNA bulge loops [124], which may also be true in DNA.

All simulations in this section are run using virtual move Monte Carlo (VMMC) to speed up the diffusion of strands (see section 1.4). The sequences of the strands are the same as those used in our experiments. For each T-motif, I separate the simulations into three umbrella sampling windows shown in Figure 2.3 to increase the efficiency of sampling rare states (see section 1.4). In the first window, the bulge loop complex and the duplex with the sticky end are prepared with all their respective native bonds fully hybridised. The end-to-end distance between the sticky end and bulge is used as an order parameter and the distribution is biased towards states with smaller separation, resulting in a flat distribution over the chosen distance ranges. This allows efficient sampling of diffusive states and the state in which the first base pair has formed between the bulge loop and the sticky end. In the second window, the number of base pairs between the bulge loop and the sticky end is used as the order parameter with the weights set to produce a flat distribution. States in which the separation of the bulge and the sticky end is too large are assigned zero weight but unbound states with small separation are still sampled to allow shared configurations between the two windows. Each window is repeated five times, with each run consisting of $\sim 10^9$ VMMC steps. The temperature in all simulations was set to 25°C and the monovalent salt concentration

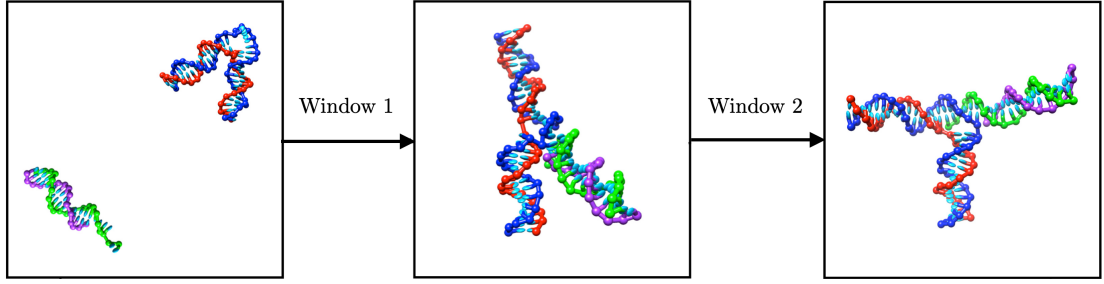


Figure 2.3: Umbrella sampling windows. The first window samples diffusive states and the second window captures the hybridisation of the sticky end to the bulge loop region.

to $[\text{Na}^+] = 0.5\text{M}$. The size of the simulation box with periodic boundary conditions is set to correspond to a relatively high strand concentration of $42\ \mu\text{M}$. The size of the box was chosen to increase computational efficiency. The results can easily be extrapolated to other concentrations by scaling the unbiased partition function of the unbound states relative to the bound states [125].

2.2.1 Conformations

To understand the role of coaxial stacking, it is instructive to look at how the stacking probabilities change as more hydrogen bonds are formed between the bulge and the sticky end. For a given number of base pairs n_{bp} , I take a subset of the configurations and ensure that they are separated by $\sim 10^4$ VMMC steps to minimise any correlations. The sites of interest $\{s_A, s_B, s_C\}$ are characterised as stacked if the coaxial stacking interaction energy between the two nucleotides is above a threshold value in order to avoid counting fleeting interactions. This is usually chosen to be the same as the threshold for hydrogen bonding and corresponds to 0.093 times the well-depth of the potential ($0.6\ \text{kcal mol}^{-1}$ at 25°C).

Figure 2.4 shows the probability of stacking for 5' and 3' T-motifs with bulge loops ranging from 3 and 9 nucleotides. In the unbound case ($n_{bp}=0$), the most probable configurations are unstacked (s_0), while the degree to which the ‘misbound’ state s_c is prevalent depends on the size of the bulge. The dependence of s_c on bulge size is consistent with the results of Ref. [123], with the probability decreasing as loop size is increased from 3 to 6 nt, and increasing with loop sizes greater

than 6 nt. The equivalent data for loop sizes not shown here can be found in the supplementary material of Ref. [115].

The T-motif $\mathbf{T}_{5'}^7$ has an optimal loop size. Binding the first few base pairs increases the probability of coaxial stacking at site s_A , followed by coaxial stacking site s_B for high n_{bp} . When all bases are hybridised ($n_{bp} = L = 7$), $\mathbf{T}_{5'}^7$ is correctly formed with high probability in stacking state s_{AB} . If a longer than optimal loop size is used (e.g. $\mathbf{T}_{3'}^9$), the T-motifs behave in the same way in the initial stages (small n_{bp}). However, stacking at site s_A is no longer possible when all bonds are formed because 9 bps cannot fit in the correctly-stacked junction. This results in a non-planar structure shown in the top right of Figure 2.5. The preference for such T-motifs to relieve stress by breaking s_A rather than s_B is because it is more feasible without breaking base pairs in the side branch duplex. Conversely, if the loop size is too short (eg. $\mathbf{T}_{5'}^3$), the probability of s_C is increased in the unbound state and binding at site s_B is favoured over s_A . This can be explained by a 4th stacking site s_{across} , which was characterised for bulge loops in Ref. [123]. They showed that duplexes with longer bulge loops have a higher probability of breaking the stacking site *across* the bulge (opposite s_C on the red strand). This stacking interaction competes with s_A and is unlikely to be broken in shorter bulge loops, which explains the prevalence of s_B and s_{BC} for T-motifs with shorter loops.

Figure 2.5 shows the probability of stacking states for fully bound T-motifs ($n_{bp} = L$). The ideal T-motif will be in stacking state s_{AB} with high probability when fully bound. This is true of 5' T-motifs with $4 \leq L \leq 8$ and 3' T-motifs with $4 \leq L \leq 6$. This is due to the difference in groove width at which the junction occurs. The major groove (22 Å) of B-DNA is wider than the minor groove (12 Å) [126], allowing more base pairs to fit without steric limitations on coaxial stacking on either side. The 3' T-motif has a shorter optimal range because it spans the minor groove and allows better correct stacking for shorter bulges.

I characterise the geometry of T-motifs based on the scheme shown in Figure 2.2c. Vectors pointing in the direction of the three arms are calculated using the base-normal vectors of individual bases of the corresponding duplex. These are

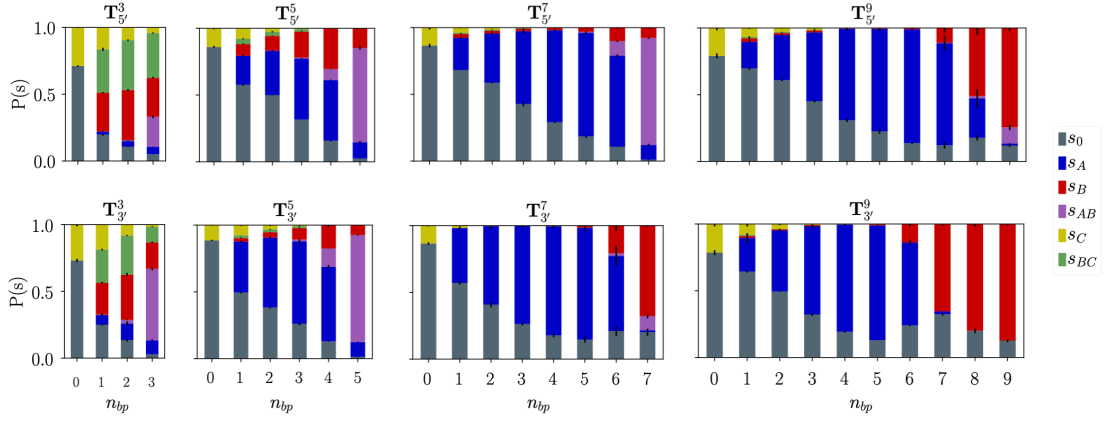


Figure 2.4: Coaxial stacking probability as a function of the number of base pairs. The top and bottom panel show results for the 5' and 3' variants, respectively. For each variant, four bulge loop sizes $L \in [3, 5, 7, 9]$ are shown. The x-axes refer to the number of hydrogen bonds formed between the sticky end and the bulge loop. Error bars represent the standard error of the mean computed from configurations in five independent VMMC trajectories. Adapted from Ref. [115].

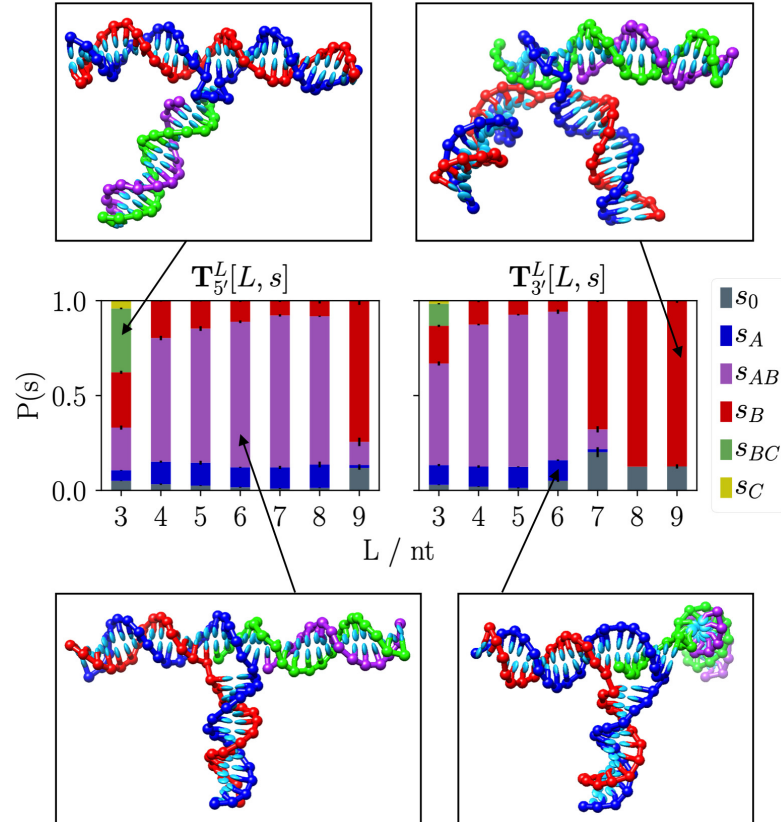


Figure 2.5: Coaxial stacking probability as a function of bulge loop size. For 5' T-motifs (left), the desired stacking state (s_{AB}) is most probable for bulge sizes between 4 and 8 nt. For 3' T-motifs (right), the range is reduced to between 4 and 6 nt. Error bars represent the standard error of the mean computed from configurations in five VMMC trajectories. Taken from Ref. [115].

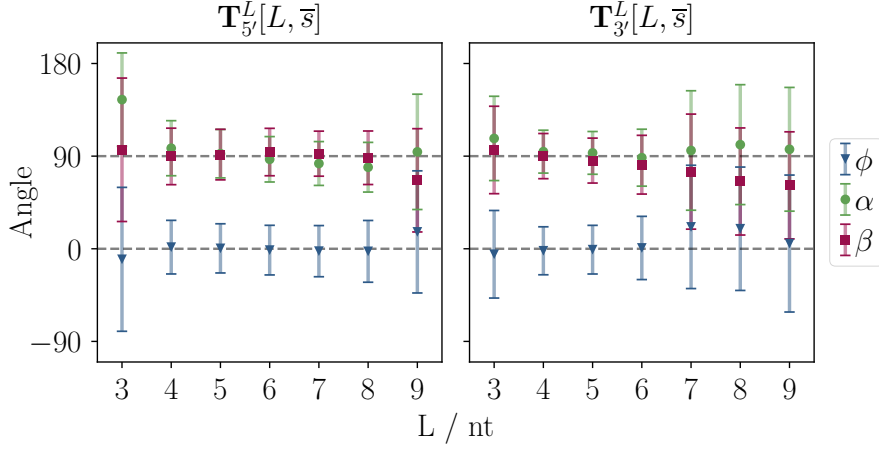


Figure 2.6: Angle between the three arms of T-motifs. Mean values and standard deviations are calculated assuming a wrapped normal distribution for three angles between arms (defined in Figure 2.2c). All stacking states with fully bound bulge loops $n_{bp} = L$ are included in the data set. A planar, right-angled geometry corresponds to $\alpha \simeq \beta \simeq 90^\circ, \phi \simeq 0$. Taken from Ref. [115].

averaged over both strands, using only 11 of the central bases to avoid distortions caused by the helical pitch. The resulting distributions are used to calculate means and variances shown in Figure 2.6. The range of loop sizes for which the T-motifs are planar ($\alpha \simeq \beta \simeq 90^\circ, \phi \simeq 0$) coincides with the range in which they are correctly stacked (s_{AB}). Correct stacking leads to a stable geometry as seen by the reduced standard deviation of the angles around their mean value.

Figure 2.7 looks at the relationship between stacking and T-motif geometry more closely. The end-to-end distance R_{AB} of the main duplex AB (see Figure 2.2b) is calculated using the positions of the four terminal nucleotides to avoid any distortion caused by fraying. A well-formed T-motif is expected to have a planar geometry ($\phi = 0$), with the main arm in an approximately straight line, maximising values of R_{AB} . Since the junction is part of the main arm, the maximum value of R_{AB} depends on the size of the bulge loop. This criterion is satisfied by the majority of the sampled configurations for the range of loop sizes mentioned above. The distributions for R_{AB} and ϕ are also separately shown for individual stacking classes in Figure 2.7. The distribution of partially or incorrectly stacked states is much wider than that of s_{AB} . It is clear that the large deviations from the mean in T-motifs with extreme loop sizes are caused by configurations that are partially or incorrectly

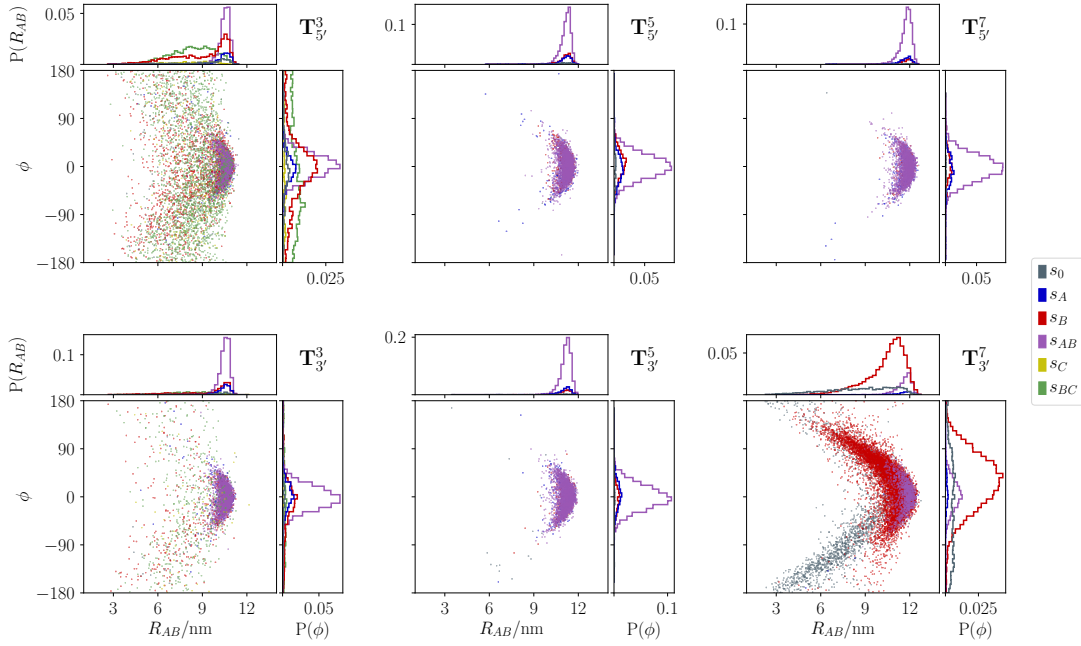


Figure 2.7: Geometric distributions of T-motif stacking states. The Scatter plots show the relationship between the dihedral angle, ϕ , and the end-to-end distance R_{AB} along the main duplex for a selection of T-motifs (see Figure 2.2 for definitions). The projections of the individual probability distributions of R_{AB} and ϕ are also shown in each panel. Each coloured distribution includes configurations in a single stacking state. Only states with fully-bound bulge loops ($n_{bp} = L$) are included in the dataset. A well-formed T-motif corresponds to stacking state s_{AB} (purple), dihedral angle $\phi = 0^\circ$ and maximal R_{AB} . Taken from Ref. [115].

stacked. The full set of distributions for all angles at each value of n_{bp} , separated by the stacking states can be found in the supplementary material of Ref. [115].

2.2.2 Free Energy

To construct the free energy profile during the formation of T-motifs, I use the weighted histogram analysis method described in 1.4 to connect the two sampling windows. Simulations were carried out with all strands at 42 μM and scaled to molar concentration for comparison to experiment. The free energy profiles for simulated T-motifs are shown in Figure 2.8. The profiles $\Delta G(n_{bp}) = -k_B T \ln[Z(n_{bp})/Z(n_{bp} = 0)]$ were constructed over the order parameter n_{bp} , where $Z(n_{bp})$ is the unbiased partition function of states with n_{bp} base-pairs. The scaling to adjust the concentration was applied to $Z(n_{bp} = 0)$. For all T-motifs, there is an initial concentration-dependent

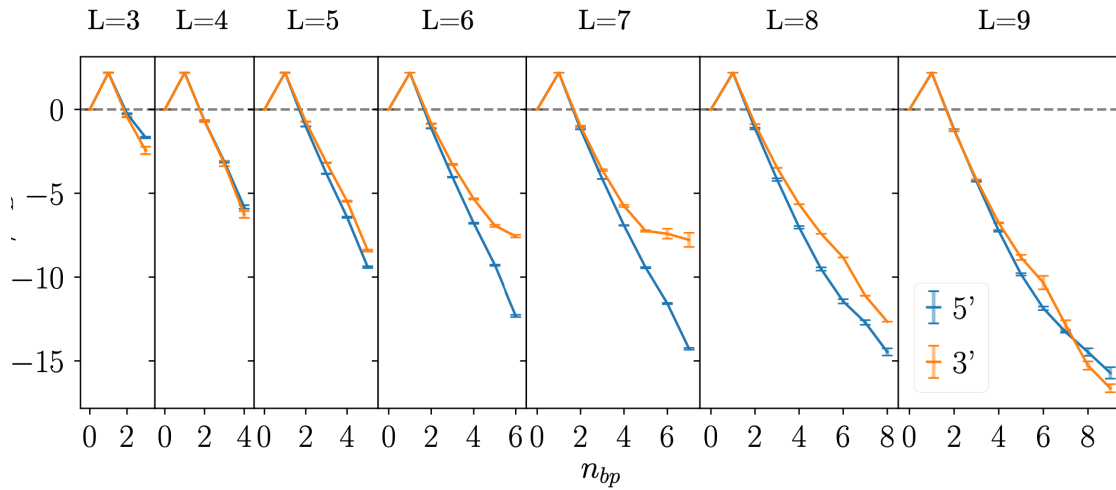


Figure 2.8: Free energy profile of T-motifs. Free-energy profiles for the formation of T-motifs with bulge loop sizes between 3 and 9 nucleotides at 25°C and molar concentrations. Results are averages obtained from 5 independent simulations and error bars represent uncertainties of the WHAM method.

entropic cost associated with bringing the two interacting regions (the bulge loop and sticky end) close together to form the first base pair. Once this barrier is crossed, hybridisation of more base pairs proceeds through enthalpic gains that outweigh entropic penalties leading to a progressive decrease in free energy. For an optimal loop size, the duplex in the junction ‘clicks’ into shape with the formation of the final base pairs. In this optimal case, the additional stability provided by correct coaxial stacking leads to the largest drop in free energy for the last base pair. If the loop size is larger or smaller than the groove width for each T-motif variant, steric constraints counteract the free energy loss of hybridisation. In the case of $L = 3$ or $L = 4$, the 5′ T-motif is less stable than the 3′ variant because stacking is more probable in the smaller minor groove. In the case of $L = 7$, the loop size is ideal for the 5′ variant but the 3′ variant benefits little from full hybridisation. In extremely long loop sizes, the junction loses its well-defined geometry. These cases (eg. \mathbf{T}_3^9) lack any coaxial stacking when fully hybridised. However, a ‘bump’ can be seen in the intermediate stages of binding, which is presumably due to the countering forces of coaxial stacking and further hybridisation.

The experimental work, carried out by Katherine Young, was aimed at estimating the thermodynamic stability of the T-motifs. For each polarity, three loop sizes

$L \in \{5, 6, 7\}$ were assessed using gel electrophoresis. One of the strands in the duplex that contains the bulge was labelled with a fluorophore. The four strands that make up the T-motif were annealed in a standard buffer solution from 95°C to 25°C. Polyacrylamide gel electrophoresis was then used to observe the complexes. We observed that the mobility of the gel band changed when the concentration of the strands containing the bulge loop was kept fixed, but the concentration of the sticky end duplex was increased. This is due to the binding and unbinding of the components in the gel. Since they have similar mobilities, a band shift is observed as they co-migrate through the gel. Through the use of reference markers, the shift in mobility was accurately measured as a function of sticky end duplex concentration. Based on the assumption that the shift in mobility is proportional to the proportion of complexes that are in the bound state, we extracted dissociation constants for each T-motif. The full experimental details can be found in Ref. [115] or Katherine’s thesis [127].

The temperature of the gels in which the dissociation constants were estimated was 4°C. To compare the extracted values to simulations, I used the single histogram reweighing method (see section 1.4) to extrapolate the simulation partition functions to nearby temperatures. The simulations were done at 25°C and the extrapolation process is more accurate if the target temperature is close to the simulation temperature because there are more microstates that share the same energy. A linear fit was then applied to the free energy data to extrapolate down to 4°C, assuming that the change in entropy is constant over this temperature range. Two examples of the fitting process are shown in Figure 2.9a. Standard changes in Gibbs free energy estimated experimentally and via simulation are compared in Figure 2.9b. While the trend is fully captured by oxDNA, there are several reasons for the slight disparity. To calculate the dissociation constants, an estimation of the concentrations of the strands in the gel lane was required, which introduces a systematic uncertainty into the experimental results. The lack of sequence-dependence of coaxial stacking interactions in oxDNA can also be a contributing factor.

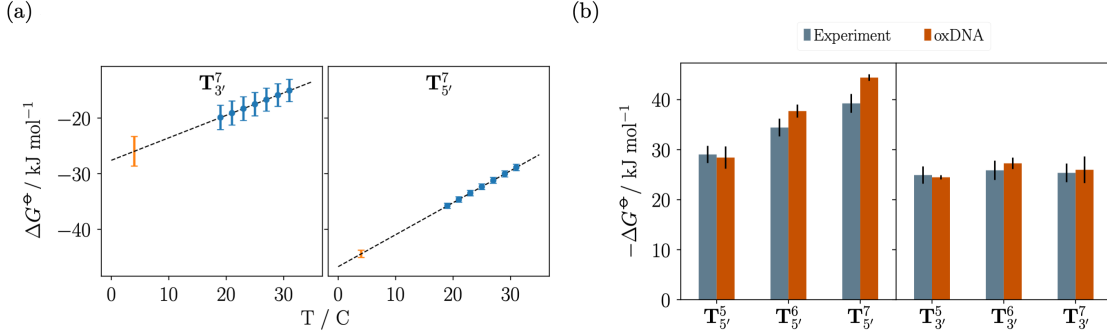


Figure 2.9: Comparison of T-motif stabilities to experiment. (a) Examples of extrapolation to lower temperatures. The simulations were performed at 25°C . Blue marks are results obtained from simulations at 25°C , extrapolated to six temperatures using single histogram reweighing. Error bars represent standard errors in averaging over 5 independent simulations. The orange point is the final value used in panel (b) and is obtained by applying a linear fit to the simulation data. (b) Gibbs free energies for the formation of T-motifs with a range of bulge sizes at 4°C and molar concentration. Experimental error bars reflect uncertainties in estimating band shifts and the systematic error associated with estimating local concentrations in gel electrophoresis.

2.3 Characterising Linkers

The work described in this section was carried out to provide a better understanding of how to incorporate T-motifs into larger DNA nanostructures. The most basic structure using two T-motifs that can be constructed is shown in Figure 2.10a. It consists of two bulge loops connected by a short single strand through hybridisation of both terminal ends, resulting in two T-junctions of opposite polarity. Space can be added between the two T-junctions by using a duplex with sticky ends to link the bulge loops (Figure 2.10b,c). We will refer to these duplexes with terminal sticky ends as ‘linkers’. The two sticky ends on the linker can have either the same or opposite polarity depending on which component strand is extended. Following the convention from the last section, I will use the notation ${}_{\mathcal{P}1}^{B1}\mathbf{L}{}_{\mathcal{P}2}^{B2}$ to denote a linker with a duplex region of L bps, bulge loop sizes of $B1$ and $B2$ nucleotides, and sticky end polarities $\mathcal{P}1$ and $\mathcal{P}2$.

I ran VMMC simulations for various linker lengths, bulge loop sizes, and the three possible combinations of sticky-end polarities. Figure 2.10d is a simple demonstration of the dependence of the orientation of the two arms on these parameters. As the number of base pairs in the long duplex containing the two

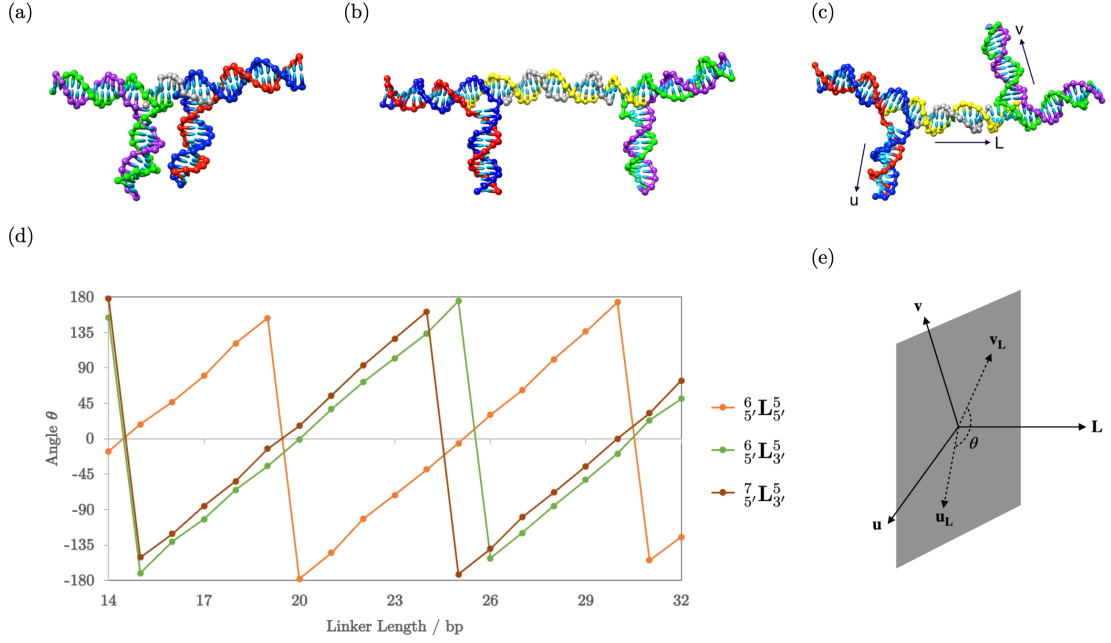


Figure 2.10: Effect of linker length and sticky end polarity on orientation. (a) Two duplexes with 6 and 5 nt bulge loops connected using a zero-length ‘linker’, which is simply an 11 nt single strand. (b) Same as (a) but the linker includes a 20 bp duplex region. (c) Same as (b) but the linker length is reduced to 15 bp rotating the arms of the two T-motifs (\mathbf{u} and \mathbf{v}). (e) Ensemble averages during VMMC trajectories of the angle θ between the side branches of the connected T-motifs as a function of linker length L . The dataset only contains fully-hybridised states, separated by 10^4 VMMC steps to reduce correlation. (d) Schematic representation of the connected T-motifs used to calculate the angle between the side branches of the T-motifs. Vectors in the direction of the linker, \mathbf{L} , and the directions of the side branches of the two T-motifs \mathbf{u} and \mathbf{v} were calculated by summing over base-normal vectors of 11 central nucleotides of one of the strands of the duplex. To measure the relative orientation of the two arms, \mathbf{u} and \mathbf{v} are projected onto the plane with normal $\hat{\mathbf{L}}$.

junctions increases, the two arms rotate relative to each other following the helical pitch of B-DNA ($\sim 10.5\text{bps}$). When opposite polarities are used to connect 6 nt and 5 nt bulge loops using a linker length of 20 bps ($6_{5'}\mathbf{L}_{3'}^5$), the two arms point in the same direction ($\theta \approx 0$). This linker length corresponds to two full turns of the DNA between the junctions and a full turn for the combination of bps inside the junctions. Given the same linker length, if a longer bulge loop is used on one side ($7_{5'}\mathbf{L}_{3'}^5$), the arms are slightly rotated. The results are similar for half-turn linkers when using the same polarity ($6_{5'}\mathbf{L}_{5'}^5$).

The next step in evaluating appropriate design features was to connect the two

rotating arms, forming a U-shaped motif, as shown in Figure 2.11a. The linker still binds in two stages because the hybridisation stage happens on a much faster timescale than the diffusive stage. In experiments, we assembled a complex with alternating duplex and bulge regions and four sets of linkers. The linkers are all 20 bps in length and have both 5' and 3' sticky ends but linkers differ in the sequence of the sticky ends. In this case, we used symmetric bulge loop sizes of 6 nt for both interactions. For each sticky end, I separated the simulations into two umbrella sampling windows in the same way as described above for single T-motifs: one for unfavourable association and one for favourable hybridisation.

The resulting free energy profiles are shown in Figure 2.11b. Two cases are shown depending on which sticky end binds first. In both cases, there is a concentration-dependent entropic penalty to bring the first sticky end near the targeted bulge loop. This is followed by the energetically downhill hybridisation of the first T-junction. The second ‘intra-molecular’ entropic cost for the second T-junction is lower than the first because the effective concentration of the sticky end at the target bulge loop is increased. As expected, the free energy of the final state (total base-pairs = 12) is the same in both cases but the landscape is more favourable if binding is initiated on the 5' side. This shows that while the 3' T-motif is not stable at 25°C and 12.5 μ M, cooperativity with the 5' end can result in successful binding. This picture is corroborated by the experimental results shown in Figure 2.11c. The mobility of the complex only changes if the 5' end has a complementary sequence to its target bulge loop. Full experimental details and strand sequences can be found in Refs. [116], [127].

Because the helical pitch of DNA is not an integer number, crossover positions in normal origami designs often need to be optimised by shortening some of the interaction domains. This is also true in designing T-junction origami. There, it can be beneficial to utilise a variety of bulge sizes and linker lengths to adjust scaffold turns and reduce overall stress in a large structure. We proceeded to study the tolerance to changes in loop size and linker length in two different cases. These ‘braced’ and unbraced cases are shown in Figure 2.12. The corresponding simulation

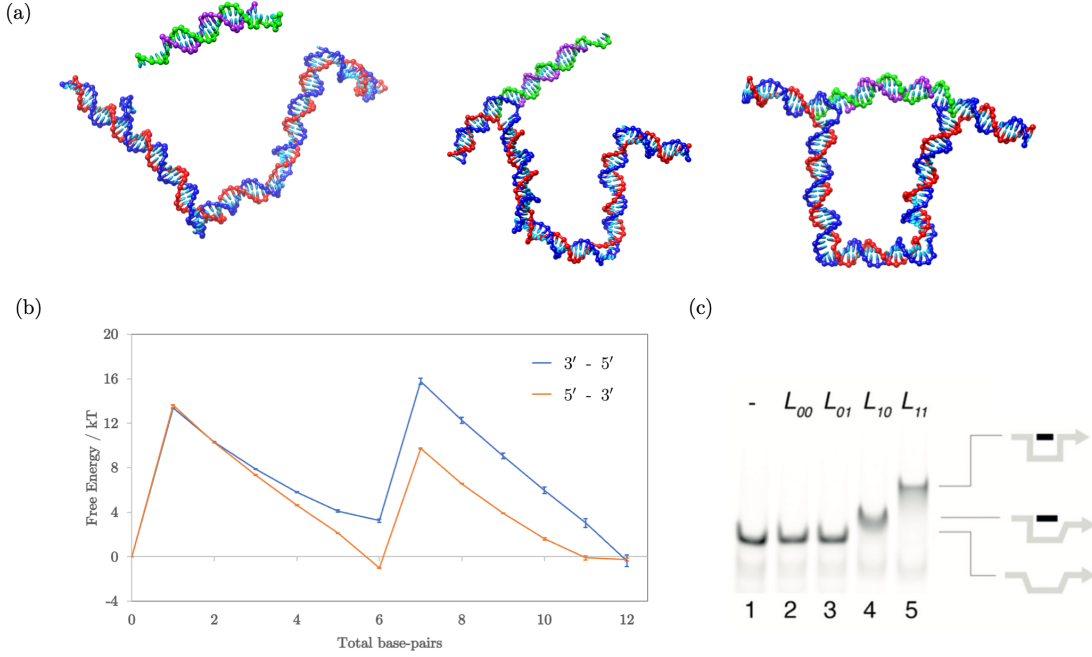


Figure 2.11: Cooperative binding of U-shaped motif. (a) The U-shaped motif consists of a linker and a pre-assembled complex with alternating duplex and 6 nt bulge loop regions. The hybridisation of the two sticky ends of the linker with two bulge loops leads to a planar geometry. (b) Free energy profile of the U-motif over the number of base pairs between the sticky ends of the linker and the two target bulge loops. Two profiles are shown: binding of the 3' sticky end followed by the 5' sticky end, and the reverse case. Simulations were performed at 25°C in periodic boxes corresponding to 12.5 μ M strand concentration. Results are averaged over 5 VMMC trajectories in 4 umbrella sampling windows and error bars correspond to errors of the WHAM method. (c) The long template strand (red in (a)) was labelled with a fluorophore and mixed with one of four linkers in each lane. Lane 1 has no linkers present. Other lanes are labelled by the sequence complementarity between the sticky ends and the bulge loops. For example, L_{01} has a complementary 3' sticky end and a non-interacting 5' sticky end. While the 5' sticky end can bind independently (lane 4), the 3' sticky end can only bind cooperatively (lanes 3 vs. 5).

and experimental results show that there is a good degree of tolerance towards linker length. The structures are able to form in a wide range ($18 \leq L \leq 24$) under experimental conditions. The probability of correct stacking (both T-junctions in state s_{AB}) is also highest in this range. For the braced case, the overall stacking probability decreases, presumably due to the lower flexibility of the structure.

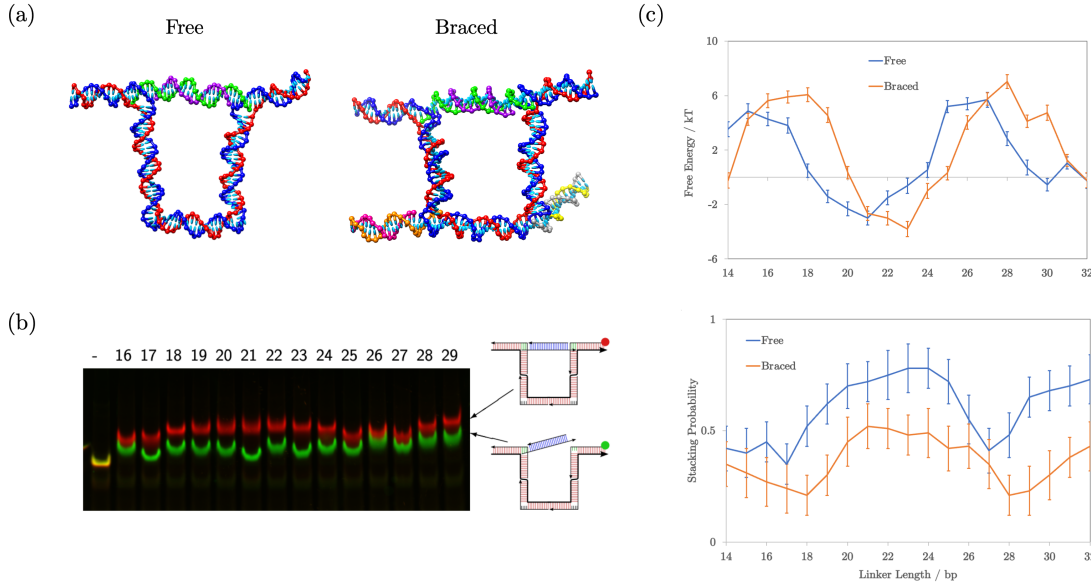


Figure 2.12: Tolerance of the U-shaped motif to linker length. (a) The U-shaped motif is ‘braced’ when the bottom bulge loops are filled with dummy duplexes with sticky ends. In all results presented, the 5′ sticky end is 6 bases long, and the 3′ sticky end is 5 bases long. (b) Gel electrophoresis results show the dependence on mobility on linker length. The red bands contain complementary sticky ends for both bulge loops. The green bands contain linkers with complementary sequences on the 5′ sticky end only. A similar pattern was seen for the ‘braced’ case with all bands shifted to a higher mobility. From Ref. [127]. (c) The top panel shows the free energy of the state with both bulge loops fully bound. For the braced case, the reference state is bound by both dummy linkers. The bottom panel shows the probability of correct stacking (all 4 stacking sites) in the fully bound state as a function of linker length. Simulations were done at 20°C in boxes corresponding to 12.5 μM strand concentrations. Error bars represent standard deviations around averages of 5 VMMC trajectories.

2.4 T-junction Origami

We approached the design of the origami-like structures based on the results of the previous sections. The aim was to demonstrate that origami can form with minimal encoding of interactions. To keep things simple, we used 20 bps linkers, 6 nt bulge loops for all 5′ insertions, and 5 nt bulge loops for all 3′ insertions. The design is illustrated in Figure 2.13. A single-stranded scaffold strand of 537 nt was obtained from a natural source and annealed with a set of 28 shorter adaptor strands to pre-assemble a complex with alternating duplex and bulge loop regions. The distances between loops were optimised by adjusting the length of the intermediate scaffolded duplexes. These were chosen to correspond to 1.5 and 2 helical turns

of DNA to align with the linker length. This complex was then mixed with a set of linkers to fold the target structure.

The bulge loops encoding the interactions with linkers occur on the adaptor strands, allowing us to adjust the specificity of interactions with ease in the future. In this work, only four loop sequences and three types of linkers were used in order to encode multiple possible misfolded structures. There are 7 loops of each type and $7!$ available configurations for two-domain binding of each linker. If two types of linkers are used, there are a total of $\sim 10^7$ possible configurations, not including any states where more than 7 copies of a linker are bound, or any states that are misbound due to partial sequence complementarity. This is in contrast to traditional origami where all interactions are designed to be unique. Some misfolds are impossible or less likely than others. For example, the half-bound linker shown in Figure 2.13b cannot bind to the pink bulge loop immediately to its left without extreme bends in duplex regions of the scaffold.

Two distinct structures were designed using two sets of linkers. For the extended structure (Figure 2.13d), 7 copies of $L_{a,b}$ and 6 copies of $L_{c,d}$ were used to interact with 26 bulge loops and leave 2 bulge loops unbound. For the compact structure (Figure 2.13c), two of the loops were omitted from the design to create a more rigid continuous helix on one side of the structure. The compact design uses 6 copies of $L_{a,b}$ and 6 copies of $L_{d,c}$ and was intended to be distinguishable from the extended motif during analysis. The set of linkers and the pre-assembled scaffold complexes were mixed and assembled under annealing and isothermal folding protocols. The resulting structures were analysed using gel electrophoresis and AFM imaging. The full experimental details can be found in Ref. [116]. I will summarise some key findings here.

The yield of both structures as assessed by AFM imaging was low compared to normal origami. This is due to the deliberate increase in domain redundancy, leading to a large number of possible misfolds that involve designed interactions. However, the results were the same for assembly under annealing protocol and isothermal folding at room temperature (20°C or 30°C for 1 hour). In normal origami, assembly

at room temperature leads to a drastic increase in the number of misfolded structures. The designed interactions in T-junction origami (5-6 bps) are much weaker than those in normal origami (> 16 bps). At room temperature, a misbound domain in normal origami is very unlikely to reverse, whereas a misbound T-junction is able to unbind, and the linker can reconfigure into a more thermodynamically favourable state. The inclusion of strong (5') and weak (3') interactions could also be aiding the process. Some of the misfolded extended structures showed a characteristic zig-zag shape, which could correspond to linkers bound at the 5' end alone.

The extended structure produced a better yield than the compact structure. In the extended design, the binding of one linker does not alter the distance between the target bulge loops for other linkers, whereas in the compact design, the scaffold 'loops around' as central linkers bind, reducing the distance between target bulge loops for subsequent linkers. This can be summarised by calculating the contact order [128] in a similar way to protein folding. The compact structure has a contact order of 0.24 compared to 0.1 for the extended motif. This is a measure of the average distance between 'native' contacts on the structure. A higher contact order corresponds to a higher degree of cooperativity between the interactions, which would result in slower folding.

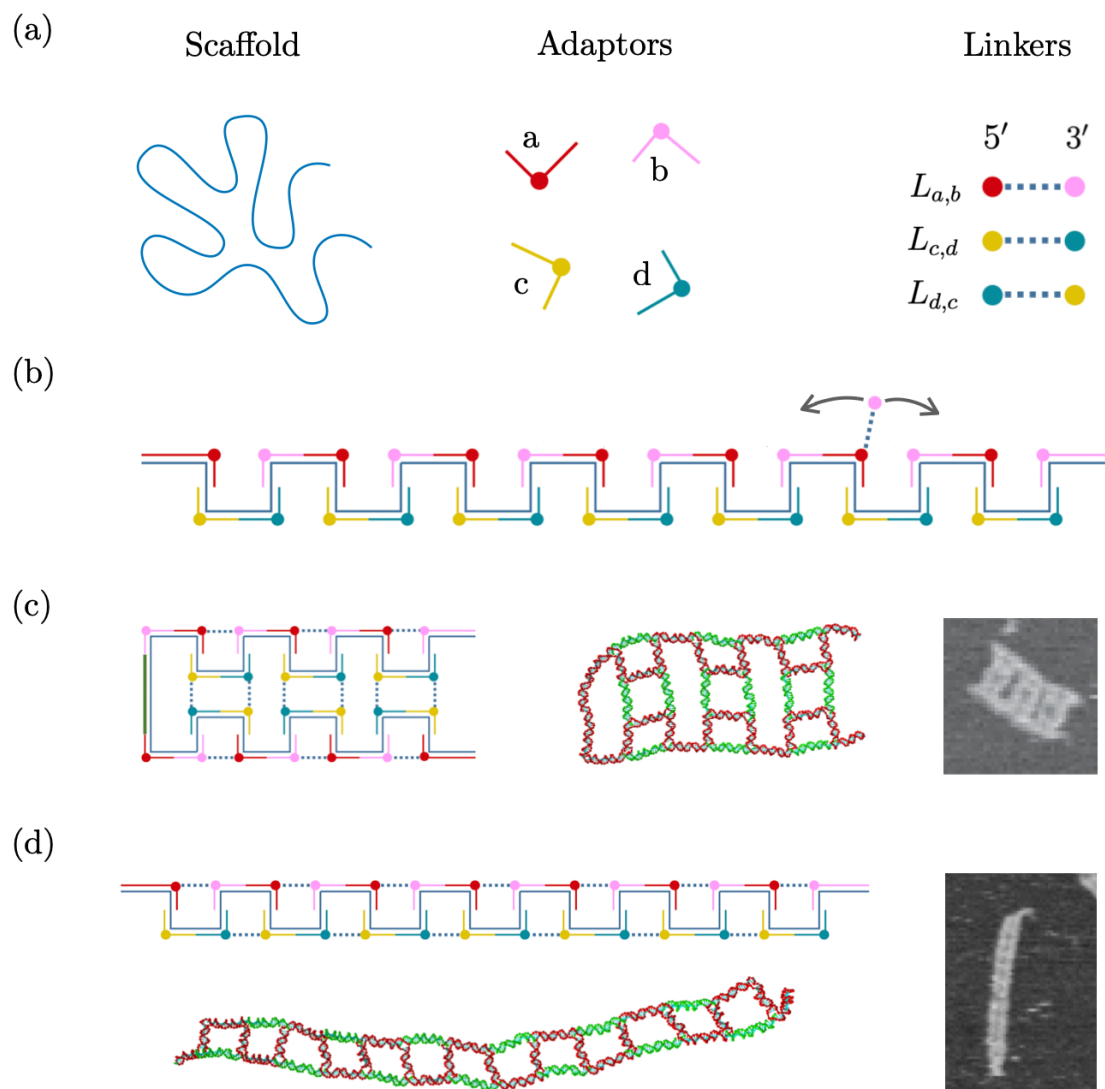


Figure 2.13: Design of T-junctions origami. (a) A long single scaffold is annealed with a set of adaptor strands and forms a complex of alternating duplex and bulge loop regions. The bulge loops (shown as coloured circles) are part of the shorter adapter strands and encode the interactions with the sticky end of linkers. Only four bulge sequences were used in the design. Linkers with 5' and 3' sticky ends have complementary sequences to a pair of bulge loops. (b) A linker bound at one end can form connections with one of 7 possible bulge loops. (c) Using linkers $L_{a,b}$ and $L_{d,c}$ leads to a compact structure. The pre-assembled complex is shown in red and the linkers are shown in green in oxDNA snapshot. An AFM image is also shown. (d) Same as (c) but using linkers $L_{a,b}$ and $L_{c,d}$, leading to an extended ladder-like structure. Adapted from [116], [127].

2.5 Summary and Outlook

In traditional origami structures, each domain of scaffold-staple interaction is typically designed to have a unique sequence and the target structure is encoded as the global free energy minimum. Origami design relies on enhancing the free energy difference between the target structure and misbound states, with empirically derived principles used to optimise assembly protocols. The folding pathway is not well-understood and assembly protocols are limited to thermal annealing or isothermal conditions in a narrow temperature range. Moreover, the specificity of interactions is constrained by the base sequence of the scaffold strand, which is usually extracted from naturally occurring genomes. Studying the effect of sequence specificity on origami formation would require custom-built scaffold strands, which would be economically unfeasible if the domain sequence is to be systematically varied.

We designed a system that utilises a relatively new structural motif, the T-junction, as a way to study the folding pathway in an economically feasible way. The stability and geometry of T-motifs were systematically characterised using oxDNA simulations. We showed that contrary to previous studies, both polarities of T-motifs can be used, with the range of sizes dictated by the groove width that the junction spans. Further, we showed that in larger assemblies, the relative weakness of 3' junctions can be compensated through cooperative binding. We successfully assembled two medium-sized, origami-like structures using both the stronger 5' and the weaker 3' interactions.

The structures were able to form at room temperature with yields that were higher than would be expected given the high degree of domain redundancy. The system can be used to study the effect of interaction specificity by exchanging a few short adaptor strands. The relatively weak interactions (compared to normal origami) that encode the target shape also allow a degree of reconfigurability. The ability to cause a global change in the structure using some stimuli is one of the focuses of the field [129]. Therefore, if structures can be assembled and reconfigured at room temperature, they could prove useful in realising numerous biomedical applications [130]–[132]. The design presented here could be a suitable

system to explore reconfigurability by sequentially adding certain adaptor strands. Proof of principle work on this has already been successfully carried out (see section 4.2.2.4 in Ref. [127]).

Future work could be focussed on adjusting the specificity of interactions to better understand the limiting cases. For example, the number of unique interactions can be systematically varied to see its effect on the assembly yield. At the moment, we believe that reconfigurability may be enhanced by the use of both weak and strong interactions. All 5' and all 3' linkers could be tested to check the effect on both reconfigurability and yield. Further, competitive interactions can be explored by using linkers that compete for loops. A similar system to the one shown here can easily be designed to achieve this.

Domain-level models can be developed to accompany experimental work. Similar models to the one I will describe in the next section can be created to estimate the entropic costs of linker binding. Such a domain-level model could easily be parametrised using oxDNA simulations. This approach was taken in modelling DNA bricks [70] to accurately predict nucleation and assembly pathways.

3

Domain Level Model of DNA Origami

Contents

3.1	Introduction	51
3.2	The Chemical System	52
3.2.1	Theoretical Framework	53
3.2.2	Simulation Algorithm	57
3.2.3	Equilibrium and Standard Gibbs Free Energy	58
3.2.4	Nearest-Neighbour Model of DNA Thermodynamics	60
3.3	State Space	61
3.4	Free Energy Contributions	67
3.5	Rate Model	73
3.6	Sampling Techniques	76
3.7	Comparison to Experiment	81

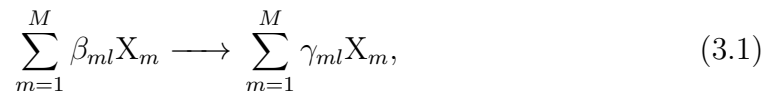
3.1 Introduction

In this chapter, I will outline the model used to study DNA origami systems presented in subsequent chapters. I begin with a general discussion about modelling chemical systems, distinguishing between the theoretical foundations of deterministic and stochastic approaches in section 3.2. The focus then shifts to the specific model of DNA origami, which follows the same principles as the model of Dannenberg *et al*[59]. In section 3.3, I discuss how the state space of the model was expanded

to allow the simulation of origami systems with three-domain staples and origami systems in which two sets of staples compete to fold the scaffold strand into their target shape. A full description of the state space, types of transition considered, thermodynamics and sampling techniques forms the remainder of the chapter. I end the chapter with a set of tests to verify the model performs as expected. The code is publicly available at Ref. [133].

3.2 The Chemical System

Our aim is to model a chemical system in a closed volume V with M unique chemical species (DNA strands or complexes) diluted in an aqueous solution. The solvent (water) acts as a thermal reservoir but is otherwise considered non-reactive and in excess of the DNA strands. The chemical species interact through different reversible reaction channels in forward and backward directions. Decomposing the reaction network into L unidirectional reaction channels, a general reaction between DNA strands in the l^{th} channel can be written as:



where X_m are the chemical species, and $\beta_{ml}, \gamma_{ml} \in \mathbb{Z}_0^+$ are stoichiometric coefficients of reactants and products, respectively. There are two ways to formulate the time evolution of chemical reactions in spatially homogenous systems: stochastic and deterministic. In general, the stochastic approach contains more information about the system, which can lead to behaviours not predicted by the deterministic approach [134]. Our aim is to simulate the system using a kinetic Monte Carlo (KMC) algorithm [135], which is a numerical method used to solve the chemical master equation (CME) in the stochastic framework. However, we parametrise reaction rates using experimental data that are fitted based on the deterministic framework. Here, I include a discussion on the differences between these approaches and how the reaction rates relate to one another.

3.2.1 Theoretical Framework

The time-evolution of a random variable X with a range of possible values x and probability density $P(x, t)$ in a stochastic process that satisfies the Markov property (memoryless-ness) is described by the Chapman-Kolmogorov equation [136]. If the range of possible values of X is discrete, the stochastic process is said to form a Markov chain in the state space x . The dynamics of the system are characterised by random jumps from state i to another state j , in which X takes values x_i and x_j . The master equation (the differential form of the Chapman-Kolmogorov equation) for the system can be written as:

$$\frac{dP(x_i, t)}{dt} = \sum_{j \in x} W_{ji}P(x_j, t) - W_{ij}P(x_i, t), \quad (3.2)$$

where W_{ij} are propensities (probabilities per unit time) of the transition from state i to state j . The master equation is a statement of the conservation of probability. One way to obtain the stationary solution ($dP_i/dt = 0$) is through the condition of detailed balance:

$$\frac{W_{ij}}{W_{ji}} = \frac{P(x_j, t)}{P(x_i, t)}. \quad (3.3)$$

If this stationary state corresponds to the thermodynamic equilibrium of the physical system at temperature T , then $P(x_i) \propto e^{E_i/k_B T}$ where E_i is the internal energy of the microstate x_i and k_B is the Boltzmann constant, then we have:

$$\frac{W_{ij}}{W_{ji}} = e^{-(E_j - E_i)/k_B T} \quad (3.4)$$

Any kinetic process satisfying Eq(3.4) will lead to equilibrium at temperature T - a useful fact that is used to predict equilibrium states through Monte Carlo simulations.

Stochastic models of chemical reactions treat reactions as discrete, random events, leading to a continuous-time Markov Chain. The state of the system is typically described in terms of the number of molecules of reactive species $\vec{n}(t) = (n_1(t), \dots, n_M(t))^1$ with probability density $P(\vec{n}_i, t)$ of each state $i \in S$. For

¹Formally, n_m are the ranges of underlying stochastic variables N_m , which take the value $n_{m,i}$ when the system is in state i .

the reaction network in (3.1), we can define a stoichiometric matrix with entries $a_{ml} = \gamma_{ml} - \beta_{ml}$ to describe all possible transitions between states. Each column of this matrix $\vec{a}_l = (a_{1l}, \dots, a_{Ml})$ specifies the change in the number of all chemical species due to the l^{th} reaction. Assuming that within infinitesimal intervals, only single reactions lead to transitions², we can associate a propensity $w_l(\vec{n}_i)$ to the l^{th} reaction channel, given the system is in state i . This propensity can be written as the product of physical and statistical factors:

$$w_l(\vec{n}_i) = \kappa_l h_l(\vec{n}_i) = \kappa_l \prod_{m=1}^M \binom{n_m}{\beta_{ml}}, \quad (3.5)$$

where n_m are the components of \vec{n}_i , κ_l is the (stochastic) rate constant for the l^{th} reaction, and $h_l(\vec{n}_i)$ takes into account the combinatorial probability of collisions between the reactants. The CME can then be written as:

$$\frac{dP(\vec{n}_i, t)}{dt} = \sum_{l=1}^L w_l(\vec{n}_i - \vec{a}_l) P(\vec{n}_i - \vec{a}_l, t) - w_l(\vec{n}_i) P(\vec{n}_i, t). \quad (3.6)$$

This gain-loss equation is the central equation of the stochastic formulation of chemical kinetics. In going from the propensity of reactions from one state to another W_{ij} to the propensity of a reaction channel to fire $w_l(\vec{n}_i)$, we have made the following assumption: within a small time interval dt , the probability of a single firing of l^{th} channel is given by $w_l dt$ to the first order in dt , and the probability that within the same interval, more than one firing occurs is given by $o(dt)$, where $o(dt)/dt \rightarrow 0$ as $dt \rightarrow 0$.

If we assume that the number of molecules n_m is large enough to produce at least one firing of each reaction channel, we can approximate the statistical contribution to the propensity as [137]:

$$h_l = \prod_{m=1}^M \frac{n_m^{\beta_{ml}}}{\beta_{ml}!}. \quad (3.7)$$

²This is Gillespie's main assumption. It also leads to the requirement that for KMC to be applicable, the rate of the elementary reaction in the system must be separated in time-scale from other reactions.

Substituting (3.7) into (3.5) and considering the expectation value, we can express the propensity in a way that can be compared to the deterministic description of chemical reactions:

$$\langle w_l \rangle = \frac{\kappa_l}{\prod_{m=1}^M \beta_{ml}!} \left\langle \prod_{m=1}^M n_m^{\beta_{ml}} \right\rangle. \quad (3.8)$$

Deterministic models of chemical reactions replace the probabilities with molar concentration of each species, $c_m = \langle n_m \rangle / N_A V$, which vary smoothly over time. This is justified in the thermodynamic limit ($V \rightarrow \infty, n_m \rightarrow \infty$ at constant c_m), where the molecular densities remain relatively constant in the timescale of statistical fluctuations. In the deterministic framework, the state of the systems is described by the molar densities of all reactive species, $\vec{c} = (c_1, \dots, c_M)$, which is evolved in time according to a set of M ordinary differential equations of the type:

$$\frac{dc_m(t)}{dt} = \sum_{l=1}^L a_{ml} R_l = \sum_{l=1}^L \left[a_{ml} k_l \prod_{m=1}^M c_m^{\beta_{ml}} \right], \quad (3.9)$$

where k_l are the (deterministic) rate constant in the l^{th} channel, and R_l are the rate laws for the l^{th} channel. Note that rate laws generally require empirical determination and the second equality in (3.9) is only valid for elementary reactions³. The above equation is a statement of material balance, which the deterministic formulation of chemical reaction kinetics in any closed volume is assumed to obey. Changing from concentrations to particle numbers, we get:

$$\frac{d\langle n_m(t) \rangle}{dt} = \sum_{l=1}^L \left[a_{ml} k'_l \prod_{m=1}^M \langle n_m \rangle^{\beta_{ml}} \right], \quad k'_l = \frac{k_l}{(N_A V)^{B_l-1}}, \quad (3.10)$$

where $B_l = \sum_{m=1}^M \beta_{ml}$ is the molecularity of the l^{th} channel. The factor k'_l can be seen as the equivalent of the deterministic rate constant if molecular numbers are used instead of concentrations.

Equation (3.10) implies that the average number of reactions in the l^{th} channel in the interval $(t, t + dt)$ is given by $\langle \lambda_l \rangle = k'_l \prod_{m=1}^M \langle n_m \rangle^{\beta_{ml}} dt$. To arrive at a deterministic equivalent to propensity, consider λ_l as a discrete random variable (with

³A reaction is elementary if the reaction order (sum of the exponents in the rate law) is the same as the molecularity, which is the case when the reactants come in direct contact to react.

expectation value $\langle \lambda_l \rangle$) that can take any value $r_l \in \mathbb{Z}_0^+$ and is distributed according to probability distribution $P(\lambda_l = r_l)$. We can insert the central assumption of the stochastic approach into this probability distribution by setting:

$$P(\lambda_l = r_l) = \begin{cases} w_l dt + o(dt) & \text{if } r_l = 1 \\ 1 - w_l dt + o(dt) & \text{if } r_l = 0, \\ 0 & \text{if } r_l > 1 \end{cases} \quad (3.11)$$

where $o(dt)$ is the probability that more than one firing of the l^{th} reaction will occur. Ignoring the $o(dt)$ term in the expectation value value of $\langle \lambda_l \rangle = \sum_{r_l} r_l P(\lambda_l = r_l) = \langle w_l dt \rangle + o(dt)$, we arrive at the deterministic equivalent to the propensity:

$$\langle w_l \rangle = k_l' \prod_{m=1}^M \langle n_m \rangle^{\beta_{ml}}. \quad (3.12)$$

We can now equate (3.12) to (3.8) and make the assumption inherent in the deterministic approach that $\langle n_a n_b \rangle = \langle n_a \rangle \langle n_b \rangle$, which assumes that there are no correlations between species ($a \neq b$) or any self-correlations leading to random fluctuations ($a = b$). We can therefore relate the deterministic and stochastic rate constants by:

$$\kappa_l = k_l' \prod_{m=1}^M \beta_{ml}! = \frac{k_l}{(N_A V)^{B_l-1}} \prod_{m=1}^M \beta_{ml}! \quad (3.13)$$

Note that the assumption of zero covariance is an assumption of the deterministic model. In fact, $\langle f(Y) \rangle \neq f(\langle Y \rangle)$ for any non-linear function f , where Y is a random variable. Substituting (3.13) into (3.8) and comparing to (3.12), we see that the concentrations c_m in the deterministic framework are an exact description of the expectation values $\langle C_m \rangle$ of the random variables $C_m = N_m/N_A V$ in the stochastic framework if $\langle \prod_{m=1}^M n_m^{\beta_{ml}} \rangle = \prod_{m=1}^M \langle n_m^{\beta_{ml}} \rangle$ holds, which requires linearity of $\prod_{m=1}^M n_m^{\beta_{ml}}$. This is not true for reactions of high order in which case the deterministic and stochastic approaches may lead to entirely different behaviours [138], [139]. As we shall see in the next sections, all of the reactions we consider have stoichiometric coefficients of at most one and molecularity of at most two. Therefore, for the purpose of this work, the deterministic and stochastic rate constants are either equivalent or related by a simple scaling factor $1/N_A V$.

3.2.2 Simulation Algorithm

In principle, the CME can predict the state of the chemical system $P(\vec{n}, t)$ at any time t , given an initial state $P(\vec{n}_i, t_0)$. However, the CME cannot usually be solved directly because the size of the state space grows rapidly as the number of reactive species and the complexity of the reaction network increase. A second option would be to numerically solve the set of ordinary differential equations (3.9) in the deterministic framework. Apart from the inherent assumptions in the deterministic approach (eg. not valid for low molecular populations), this also runs into scaling issues as system size grows. A third option is to use a Stochastic Simulation Algorithm (SSA) [140] to generate traces that follow the same probability distribution as the CME. Originally conceived by Gillespie, these algorithms have been further developed over the years to become a powerful tool in modelling chemical and biological processes.

The most widely SSA is the Gillespie algorithm, also known as the kinetic Monte Carlo (kMC) algorithm. Its utility lies in the fact that it reformulates the question “given $P(\vec{n}_i, t_0)$, what is $P(\vec{n}, t)$?” as “given current state \vec{n}_i , what is the time taken τ for the next reaction to occur, and which reaction l will it be?”. To answer the second question probabilistically while conforming to the Master equation, Gillespie introduces the *reaction* probability density function $P(l, \tau | \vec{n})$ and postulates that the next reaction will occur with probability $P(l, \tau | \vec{n})d\tau$ within the time interval $[t + \tau, t + \tau + d\tau]$, and will be of type l . By sampling this distribution iteratively, adjusting \vec{n} and repeating the process, an entire realisation of the process can be generated.

To do this, we can break the joint distribution into two parts $P(l, \tau | \vec{n})d\tau = P_0(\tau | \vec{n})w_l(\vec{n})d\tau$, where $P_0(\tau | \vec{n})$ is the probability that no reaction of any type will occur in $[t, t + \tau]$, and $w_l(\vec{n})d\tau$ is the probability that a single reaction of type l will occur in the interval $[\tau, \tau + d\tau]$. To obtain the functional form of $P_0(\tau | \vec{n})$, we can break up the interval τ into N subintervals of length ϵ . The probability that no reaction of any type will occur during ϵ is $P_0(\epsilon | \vec{n}) = \prod_{l=1}^L (1 - w_l(\vec{n})\epsilon + o(\epsilon)) = 1 - \sum_{l=1}^L w_l(\vec{n})\epsilon + o(\epsilon)$, where $o(\epsilon)$ is the probability that more than one reaction occurs.

Because the system remains in state \vec{n} after each subinterval, we can take the limit of arbitrarily small ϵ to get $P_0(\tau|\vec{n}) = \lim_{N \rightarrow \infty} [P_0(\epsilon|\vec{n})]^N = \exp[-\sum_{l=1}^L w_l(\vec{n})\tau]^4$. Finally, the reaction probability density can be written as:

$$P(l, \tau|\vec{n}) = e^{w_0(\vec{n})\tau} w_l(\vec{n}) \quad (3.14)$$

where we have defined $w_0(\vec{n}) = \sum_{l=1}^L w_l(\vec{n})$.

To sample this distribution, we must devise a method to select a random pair (l, τ) from the distribution. Since $P(l, \tau|\vec{n})$ is the product of two independent functions, τ can be drawn independently from $e^{w_0(\vec{n})\tau}$, and l can be drawn independently from $w_l(\vec{n})$. To achieve this algorithmically, we can generate two random numbers $r_1, r_2 \in [0, 1]$ and select a random pair (l, τ) by the following prescription:

$$\sum_{j=1}^{l-1} w_j < r_1 w_0 \leq \sum_{j=1}^l w_j \quad , \quad \tau = \frac{1}{w_0} \ln \frac{1}{r_2} \quad (3.15)$$

3.2.3 Equilibrium and Standard Gibbs Free Energy

In chemical systems, it is common to determine the thermodynamic driving force of reactions at some reference state, which can then be scaled to apply to systems under different conditions. The choice of the standard state is arbitrary but by convention, thermodynamic quantities, such as enthalpy, entropy or Gibbs free energy are denoted by superscript \ominus when measured at pressure $P^\ominus = 1\text{bar}$, temperature $T^\ominus = 37^\circ\text{C}$ and concentration of $c^\ominus = 1\text{M} = 1\text{molL}^{-1}$. Because chemical reactions occur in solution, the choice of the standard state is made to correspond to conditions that would be observed in the limit of infinite dilution and ideal behaviour. Here, we discuss how equilibrium concentrations under one set of conditions are related to the measured thermodynamic databases at the standard state.

In the deterministic framework, if we take all reactions to be reversible, and require that the concentration current between each reversible pair vanishes by setting $R_l^+ = R_l^-$ in (3.9), we arrive at the deterministic equivalent to the condition

⁴We have used the property that $o(\epsilon)/\epsilon \rightarrow 0$ as $\epsilon \rightarrow 0$

of detailed balance:

$$\frac{k_l^+}{k_l^-} = \prod_{m=1}^M \frac{\{c_m\}^{\beta_{ml}^-}}{\{c_m\}^{\beta_{ml}^+}}, \quad (3.16)$$

where curly brackets denote equilibrium concentrations. Detailed balance can be shown to be necessary and sufficient for the system to reach equilibrium. The condition for thermodynamic equilibrium in such a system can also be stated in terms of a lack of driving force in each reaction channel. Under constant pressure and temperature, equilibrium corresponds to no change in the Gibbs free energy due to the reaction. This change can be written in terms of the chemical potential of each reactive species:

$$\Delta G_l = \sum_{m=1}^M (\beta_{ml}^+ - \beta_{ml}^-) \mu_m, \quad (3.17)$$

where μ_m is the chemical potential of species m . This can be related to conditions of the standard state through:

$$\mu_m = \mu_m^\ominus + RT \ln(a_m), \quad (3.18)$$

where $a_m = \eta_m c_i / c^\ominus$ is the activity (effective concentration) of species m under the new conditions with respect to the reference state, and η_m is the corresponding activity coefficient. If the solution is sufficiently dilute⁵, we can write the chemical potential as:

$$\mu_m = \mu_m^\ominus + RT \ln\left(\frac{c_m}{c^\ominus}\right). \quad (3.19)$$

Substituting into (3.17) equations, we have:

$$\Delta G_l = \Delta G_l^\ominus + RT \sum_{m=1}^M (\beta_{ml}^+ - \beta_{ml}^-) \ln\left(\frac{c_m}{c^\ominus}\right), \quad (3.20)$$

where we have defined $\Delta G_l^\ominus = \sum_{m=1}^M (\beta_{ml}^+ - \beta_{ml}^-) \mu_m^\ominus$ as the standard state change in Gibbs free energy. By requiring that $\Delta G_l = 0$, we can relate ΔG_l^\ominus to equilibrium concentrations:

$$\Delta G_l^\ominus = -RT \sum_{m=1}^M (\beta_{ml}^+ - \beta_{ml}^-) \ln\left(\frac{\{c_m\}}{c^\ominus}\right). \quad (3.21)$$

⁵Such that the quotient of the activity coefficients is constant with reference to the standard state.

Assuming the detailed balance condition corresponds to thermodynamic equilibrium, we can combine (3.16) and (3.21) to arrive at:

$$\frac{k_l^+}{k_l^-} = e^{-\Delta G_l^\ominus / RT} (c^\ominus)^{-\sum_{m=1}^M (\beta_{ml}^+ - \beta_{ml}^-)} \quad (3.22)$$

In what follows, we make use of measured ΔG^\ominus values for simple bimolecular reactions (where $\sum_{m=1}^M (\beta_m^+ - \beta_m^-) = 1$), and make use of (3.20) and (3.22) to make adjustments to our model.

3.2.4 Nearest-Neighbour Model of DNA Thermodynamics

The gold standard for DNA thermodynamics in the literature is currently the Nearest-Neighbour model of Santalucia [141], [142]. The model assumes that the stability of a given Watson-Crick base-pair depends only on the identity and direction of immediately adjacent bases and provides a database to calculate the standard state free energy difference $\Delta G_{\text{NN}}^\ominus(T) = \Delta H_{\text{NN}}^\ominus - T\Delta S_{\text{NN}}^\ominus$ between bound and unbound states. The database is derived from applying a two-state model to UV spectrometry measurements of melting temperature of DNA from various sources[143], where it is assumed that intermediate states are too short-lived to contribute to the absorbance. It provides a temperature-independent parameter for enthalpy and entropy of each of the 10 possible combinations of nearest-neighbours, plus additional parameters for sequence-dependent initiation of duplex formation and salt dependence. Given the sequence of two interacting oligos A and B, the NN model describes the transitions as a perfect two-state transition, with equilibrium concentrations given by:

$$\frac{\{AB\}}{\{A\}\{B\}} = e^{-\Delta G_{\text{NN}}^\ominus(T)/RT} M^{-1}, \quad (3.23)$$

where M is the standard concentration, and R is the gas constant. The NN model is an accurate predictor of melting temperatures, defined as the temperature at which half the strands are in duplex form:

$$T_{\text{NN}}^{\text{Melt}} = \frac{\Delta H_{\text{NN}}^\ominus}{\Delta S_{\text{NN}}^\ominus + R \ln ([A]_0 - [B]_0/2)}, \quad (3.24)$$

where $[A]_0$ and $[B]_0$ are the total concentrations of each strand, with the more abundant strand taken to be A if the two are not of equal concentrations. Since both

the enthalpy and entropy are taken to be independent of temperature, the accuracy of the model in predicting $T_{\text{NN}}^{\text{Melt}}$ points to the specific heat of DNA being constant across a wide range of temperatures or to the balance of its contribution to entropy and enthalpy. Therefore, we do not make any corrections to the standard state measurements when applied at the range of temperatures in our model. Another consideration when using the database is the salt dependence of the transition. This dependence is characterised empirically for solutions with monovalent salt through a simple correction.

3.3 State Space

The model we will use for DNA origami follows the same principles as that of Ref. [59] with some additions to the state space to allow staples with three domains and staples from multiple staple sets to bind to the scaffold. Origami design usually proceeds by designing sections of shorter staple strands to pair with sections of a long scaffold strand such that when all pairs are hybridised, the scaffold strand folds into a target shape. I refer to these sections of designed sequence complementarity as “domains” of interaction, which are typically 5 to 20 base pairs in length. Domains are usually adjacent on a staple strand but their counterparts on the scaffold strand are not. This allows staple strands to form crossovers between helices to bring non-adjacent domains on the scaffold together, thus pinning part of the scaffold into a loop. I will refer collectively to multiple domains on the scaffold that are designed to bind to a single staple as a *staple-complement*. By adjusting the position (and therefore the sequence) of staple-complements, the same scaffold strand can have a multitude of target shapes. I will refer to an entire set of staples that are designed to fold the scaffold into a single target shape as a *staple set*.

The philosophy of the model is to explicitly assign states that we think are particularly important to origami formation. We integrate out intermediate states such as the diffusive motion of staples or scaffold, or the transient “zippering” states of domain hybridisation. Instead of explicitly assigning states to these configurations, we consider them intermediate states during transitions in the model

and account for them thermodynamically. In what follows, we ignore mis-bonds due to partial sequence complementarity between domains or between any regions of the scaffold and staples that are not intended to bind to each other. We also ignore any interactions between distinct scaffolds and self-interactions of different regions of any particular scaffold. However, we do allow blocked states in which staple-complements bind to multiple copies of the same staple type, leading to kinetic traps. Therefore, the state of a partially-formed origami can be described by the binding configuration of its staple-complements.

Consider an origami such as the one in Figure 3.1. The scaffold is designed to be folded by a staple set consisting of M_1 one-, M_2 two-, and M_3 three-domain staples. The state of any such partially-folded scaffold can be specified by a vector $\vec{s} = (s_1, s_2, \dots, s_M)$ with $M = M_1 + M_2 + M_3$ components, where s_m is the binding configuration of the m^{th} staple-complement. Let $S = \{\vec{s}_i\}$ be the set of all possible states for the origami. One-domain staple-complements can be in one of 2 states: unbound (s_0) or bound (s_1). Two-domain staple-complements can be in one of 5 states, four of which do not form a crossover ($s_{00}, s_{10}, s_{01}, s_{12}$), and one in which both domains are bound by the same staple, forming a crossover (s_{11}). Three-domain staple-complements can be in one of 15 states, in which eight do not form any crossovers ($s_{000}, s_{100}, s_{010}, s_{001}, s_{120}, s_{012}, s_{102}, s_{123}$), six that form a single crossover ($s_{110}, s_{101}, s_{011}, s_{112}, s_{121}, s_{122}$), and one that forms two crossovers (s_{111}). Therefore, the size of the state space for an origami with such a staple set is $|S| = 2^{M_1} \times 5^{M_2} \times 15^{M_3}$.

As shown in Figure 3.1, the state of any origami can be represented by a graph, with edges representing either domains or crossovers. Each staple crossover places constraints on the conformation of the scaffold strand by creating an additional loop. Our aim is to represent the state of origami by identifying a *unique* set of loops in any partially-folded state so that we may assign it an unambiguous free energy to represent its relative entropy. As discussed in Ref. [59] where it is called the ‘global’ approach, this task can be reduced to systematically identifying faces of the planar embedding of the graph. When all edges that represent domains are added,

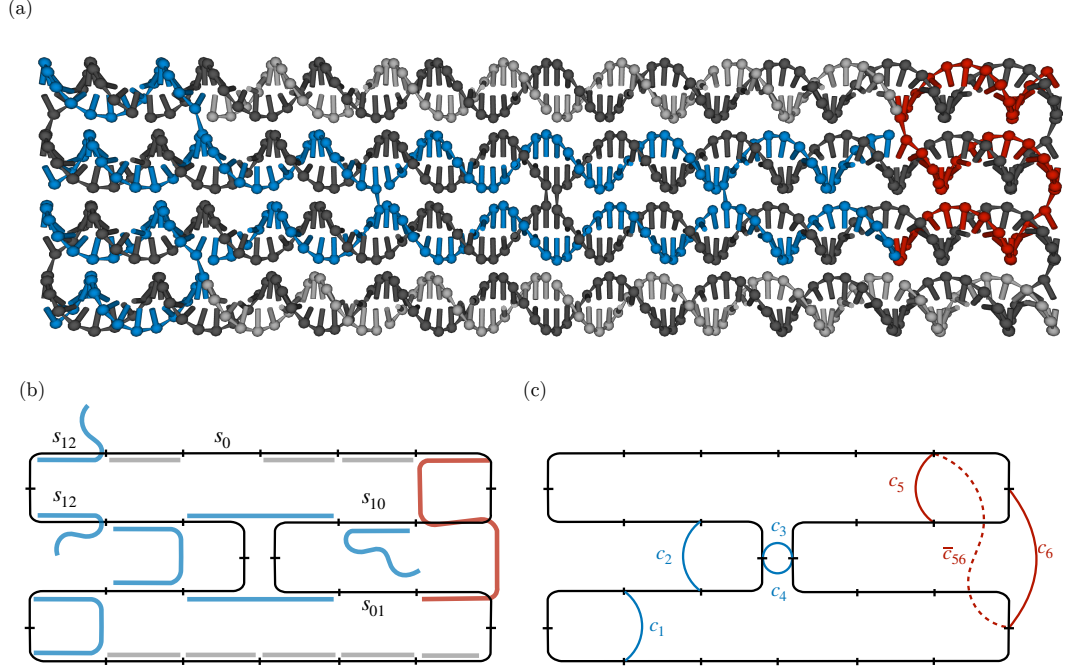


Figure 3.1: Sample origami state space. (a) A sample origami design consisting of a scaffold (dark grey) and staples with one, two, and three domains, shown in grey, blue and red respectively. (b) Schematic representation on an intermediate state of the origami shown in (a), highlighting the state s_{12} , in which the staple-complement is bound by two staples of the same type. (c) Graph representation of the partially-folded state shown in (b), with an additional dashed crossover (\bar{c}_{56}); this crossover only forms if the three-domain staple is in state s_{101} or s_{121} , in which case the graph is not planar.

the planar embedding of the graph has two faces: an inner loop and an outer loop, which can be represented as a traversal of the nodes in clockwise/anti-clockwise directions. Each additional crossover adds an extra face but there is ambiguity in the way the face can be drawn. As shown in Figure 3.2, in order to algorithmically select a unique planar embedding for all partially-folded states, it is necessary to specify as an input whether each crossover is internal or external to the embedding.

Not all graphs have a planar embedding, for example, those representing the state of 3D origamis or some states of 2D origamis with three-domain staples (Figure 3.2c). Although the graphs associated with the states of such origami are still unique, it is not possible to unambiguously assign a unique set of loops (and hence conformational entropy) to every state in our model. In such cases, it is not possible to construct a thermodynamically well-defined model using our algorithmic

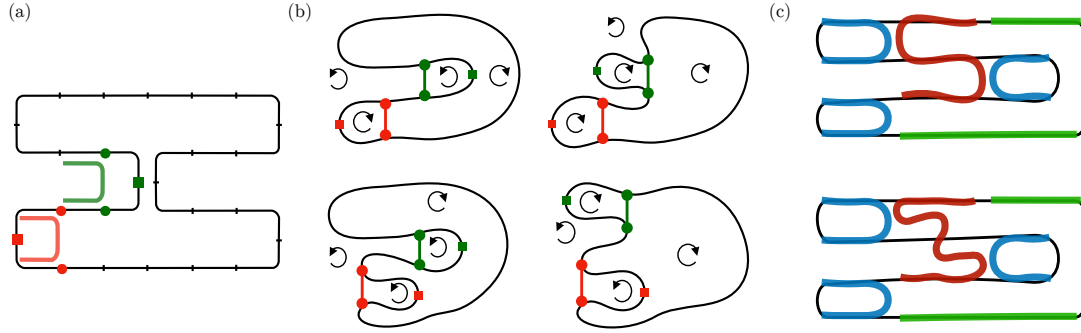


Figure 3.2: Ambiguity in loop assignments. (a) A partially-folded origami consisting of two crossovers. (b) Four possible planar embeddings of the graph representing (a). In each embedding, clockwise / anti-clockwise arrows indicate a face that is inside/outside the scaffold. The correct embedding is the one in which the large ‘outside’ face traverses the red square node but not the green square node (top left); the selection of this embedding requires a priori labelling of each crossover as being either inside or outside the scaffold. (c) States in which a 3-domain staple is fully-bound (s_{111}) or bound by its end domains (s_{101}). The graph for the s_{101} case is non-planar.

approach but we can develop path-dependent dynamics by considering free energy changes during transitions that create or remove crossovers. This approach is called ‘local’ in Ref. [59] and we will return to this in the next sections.

Consider two staple sets, each designed to fold the same scaffold into a different shape as shown in Figure 3.3a. For simplicity, let us take all domains to be of equal length d in both staple sets. It is always possible to rotate the scaffold sequence relative to each staple set such that each domain in the first staple set shares its sequence with parts of two domains of the second staple set. Starting from fully aligned domains, shifting the scaffold by n nucleotides relative to the first staple set leads to ‘toehold’ domains of lengths n (with $m = n$ in Figure 3.3 if all domains are of equal length) and ‘displacement’ domains of $x = d - n$. Note that the role of the displacement and toehold domains can be reversed depending on the staples that are involved. We introduce an approximation where we do not allow part-bound domains in the state space, so toeholds become available only when a domain in the other staple set (fully) unbinds. This allows us to simplify the state space such that the state of a single scaffold in the presence of multiple staple sets can be represented by the concatenation of the individual state vectors of each staple set. Thus, the state of the scaffold can be described by a vector $\vec{s} = (s_1^1, s_2^1, \dots, s_M^1, s_1^2, s_2^2, \dots, s_M^2)$,

where s_m^k specifies the configuration of the m^{th} staple-complement in the k^{th} staple set; the size of the state space is difficult to count but is bounded by $|S_1| \times |S_2|/2$.

To represent competition between staples from different sets without allowing part-bound domains in the state space, we aim to coarse-grain out the intermediate steps of strand exchange while capturing the dependence of the kinetics on the thermodynamics of hybridisation. Strand exchange reaction kinetics have been previously shown to be well-approximated by bimolecular models with second-order rate constants [144], [145]. Zhang and Winfree [146] propose a three-step model, schematically shown in Figure 3.3b, and derive a relation between the bimolecular rate constant k_{mn} and the rate constants of the individual reactions of the three-step model:

$$k_{mn} = \frac{k_f k_m k_b}{k_n k_m + k_m k_b + k_n k_b} \quad (3.25)$$

This derivation amounts to solving the set of ODEs in Equation 3.9 for the reversible transitions $I \leftrightarrow Y_1 \leftrightarrow Y_2 \leftrightarrow F_3$ in Figure 3.3, subject to the quasi-steady state conditions $\dot{c}_{Y_1} = \dot{c}_{Y_2} = 0$. This quasi-steady state assumption is valid due to the timescale separation between the equilibration of the branch migration step (domain x in Figure 3.3b) and the overall reaction. This condition naturally imposes an upper limit on the concentrations of the reactants for which the approximation is valid. If this critical concentration is breached, the bimolecular reactions cannot be assumed to occur on a slower timescale than the intermediate unimolecular reactions, leading to inaccuracies in k_{nm} .

In the first of the three steps, an incoming staple binds at the toehold (m for the green staple) with a diffusion-limited ⁶ rate constant k_f . The second step is a random walk process, with a rate that depends on the length of the branch migration domain. The rate of the final step, k_n is dependent on the stability of the toehold of the outgoing staple:

$$k_m = \frac{2}{x} k_f e^{\Delta G_{NN}^{\ominus}(m)/RT} s^{-1} \quad (3.26)$$

⁶This is not strictly true. Kinetically, hybridisation can be viewed as a two-step process: a slow rate-limiting nucleation step to bind a few bps, followed by a fast ‘zipping’ of other bases [147], [148].

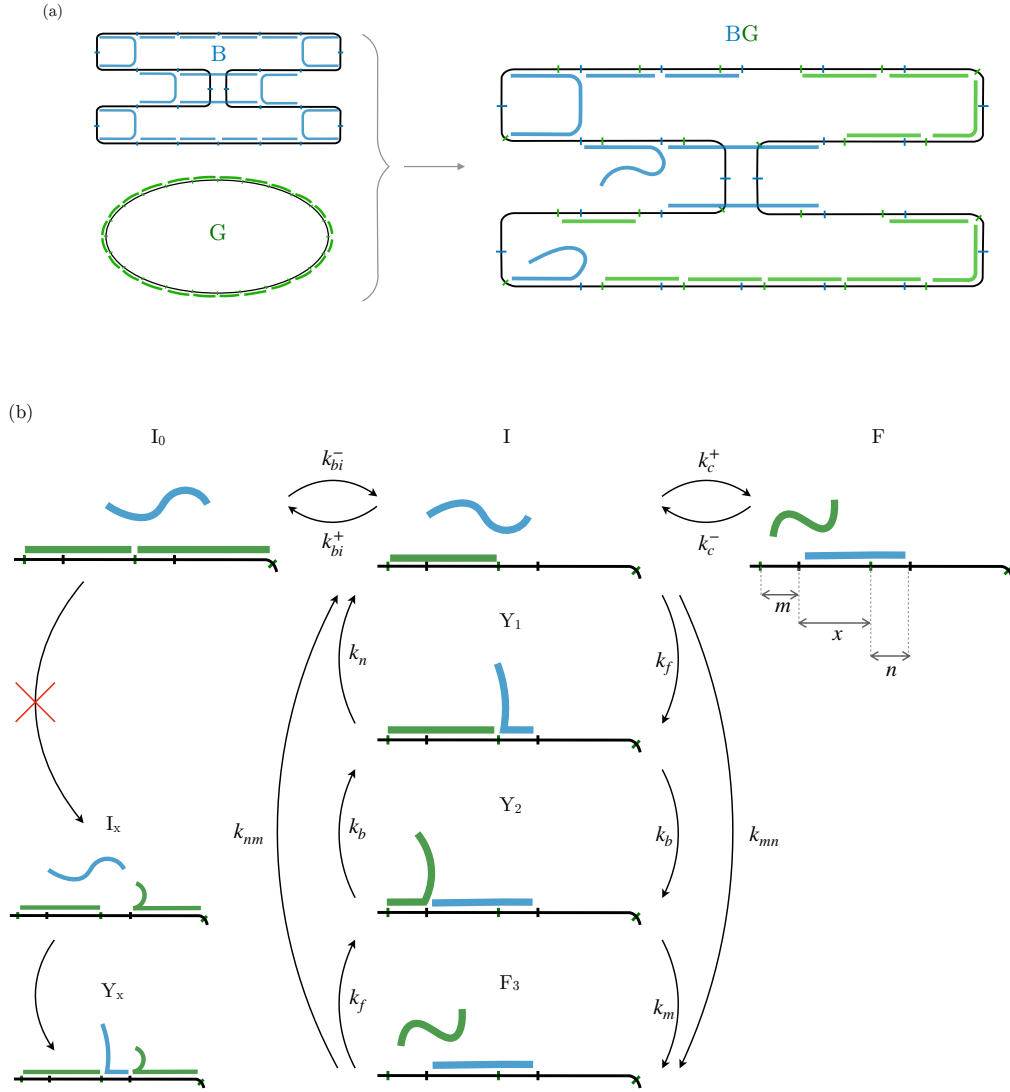


Figure 3.3: Strand exchange model. States considered in origami with two competing staple sets. (a) A staple set that is designed to fold the scaffold into a rectangle (B) and a staple set that only has single-domain strands (G) have sequence complementarity with the same scaffold. When drawn on a fixed positioning of the scaffold (BG), the domains in G do not align with domains in B, allowing competition between the two staple sets through toehold-mediated strand exchange. (b) A toehold exchange event on a subsection of the scaffold strand, in which a blue staple displaces a green staple. State I_0 refers to a fully-coated initial state, which cannot undergo an exchange reaction since partially-bound domains such as that in I_x are not considered in the model. The three-step process $I \rightarrow Y_1 \rightarrow Y_2 \rightarrow F_3$ is coarse-grained to a one-step process $I \rightarrow F_3 \equiv F$. In the model, the transition proceeds through a simple unbinding event $I_0 \rightarrow I$, followed by a displacement event $I \rightarrow F$. The lengths of toehold and displacement domains of a particular reaction depend on the length of the original hybridisation domains of the staples in the competing staple sets. All resulting sub-domains (n , m and x) act as toeholds or displacement domains depending on the staples involved in the reaction.

where $\Delta G_{NN}^{\ominus}(m)$ is the nearest-neighbour free energy change due to binding at toehold m at standard conditions and the factor $2/x$ is to ensure correct relative concentrations of the states at equilibrium (see below). By experimentally determining the values of k_{nm} for a large number of sequences, Zhang and Winfree [146] were able to fit the values of the strand association rate constant $k_f = 3.5 \times 10^6 \text{M}^{-1}\text{s}^{-1}$ and branch migration rate constant $k_b = 1.0 \times (20/x)^2 \text{s}^{-1}$, where x is the length of the branch migration domain. The factor of 20 is because the experimental best-fit value of 1s^{-1} was obtained for branch migration domain length of $x = 20$ nt; this is scaled here by a factor of x to reflect the change in the length of the random walk relative to the fitted case.⁷ To add the relevant contributions arising from the origami environment, we first note that $k_{nm}/k_{mn} = k_n/k_m = e^{[\Delta G^{\ominus}(m) - \Delta G^{\ominus}(n)]/RT} = e^{-\Delta G_{NN}^{\ominus}(n,m)/RT}$. Therefore, we can relate the forward and backward rate constants for the reaction $I \leftrightarrow F$ within the origami environment by:

$$\frac{k_c^+}{k_c^-} = \frac{k_{mn}}{k_{nm}} e^{-\Delta G_{\text{origami}}^{\ominus}(I,F)} \quad (3.27)$$

where $\Delta G_{\text{origami}}^{\ominus}(I,F)$ represents any additions to the nearest-neighbour free energy arising from the origami environment, which will be discussed in the next section.

3.4 Free Energy Contributions

In line with the approach taken in Ref. [59], we assume there are three independent contributions to the free energy of each state \vec{s}_i , relative to the fully unbound state under standard conditions:

$$\Delta G^{\ominus}(\vec{s}_i) = \Delta G_{NN}^{\ominus}(\vec{s}_i) + \Delta G_{CX}^{\ominus}(\vec{s}_i) + \Delta G_{SH}^{\ominus}(\vec{s}_i), \quad (3.28)$$

where $G_{NN}^{\ominus}(\vec{s}_i)$ is the contribution from DNA base-pairing, $G_{CX}^{\ominus}(\vec{s}_i)$ takes into account coaxial stacking energies between adjacent domains that are bound, and

⁷An iso-energetic branch migration with N steps can be modelled as a random walk with an average time-scale of $N^2\tau$, where τ is the time-step of a single step.

$G_{\text{SH}}^{\ominus}(\vec{s}_i)$ takes into account the change in entropy due to reduced freedom of the scaffold strand due to the formation of staple crossovers.

The duplex term $G_{\text{NN}}^{\ominus}(\vec{s}_i)$ only takes into account the free energy change of hybridisation for isolated duplex formation in bulk solution and is estimated using the well-established nearest-neighbour model of Santalucia [141], [142].

Coaxial stacking refers to the stacking of bases in a duplex in cases where the backbone is ‘nicked’, which has a stabilising effect on the duplex[149]. The term $G_{\text{CX}}^{\ominus}(\vec{s}_i)$ takes into account the stabilisation provided by adjacent domains on the scaffold that are bound to staples. Given two such scaffold domains, the situation is equivalent to a nick across the backbone of a complementary strand composed of domains of staples. A stacking contribution $\Delta G_{\text{CX}}^{\ominus}(T) = n\langle\Delta G_{\text{NN}}^{\ominus}(T)\rangle$, using the sequence-averaged free-energy gain per base-pair $\langle\Delta G_{\text{NN}}^{\ominus}(T)\rangle$ of the NN model, is added whenever two adjacent domains of the scaffold are both hybridised. The free parameter n is a fitting parameter tuned by comparison to experimental data.

The shape term $G_{\text{SH}}^{\ominus}(\vec{s}_i)$ takes into account the entropic penalty of loop formation as staples pin parts of the scaffold strand together, reducing its conformational freedom. It can be decomposed into individual loop penalties:

$$G_{\text{SH}}^{\ominus}(\vec{s}_i) = \sum_{\text{loop} \in L(\vec{s}_i)} G_{\text{loop}}^{\ominus}, \quad (3.29)$$

where $G_{\text{loop}}^{\ominus}$ is the length-dependent free energy of forming a single loop under standard conditions and $L(\vec{s}_i)$ is the set of loops in state \vec{s}_i chosen according to the convention outlined in section 3.3. To estimate $G_{\text{loop}}^{\ominus}$, consider a heterogenous freely-jointed chain in a volume $v^{\ominus} = 1/(N_A \times c^{\ominus})$, which is the volume occupied by a single scaffold strand at standard concentration ($N_A = 6.022 \times 10^{23} \text{mol}^{-1}$ is the Avogadro constant). Let $P_{\text{loop}}^{r_c}$ be the probability that the two ends of the scaffold region that contain the loop come within an arbitrarily small distance r_c . Let $P_{v^{\ominus}}^{r_c}$ be the probability that two unconnected molecules in v^{\ominus} come within a distance r_c . Then the difference between the free energy of association of isolated strands under standard conditions and the additional free energy of forming a loop is given by:

$$G_{\text{loop}}^{\ominus, \text{FJC}} = -RT \ln \left(\frac{P_{\text{loop}}^{r_c}}{P_{v^{\ominus}}^{r_c}} \right) = -RT \ln \left(\frac{v^{\ominus} P_{\text{loop}}^{r_c}}{v_{r_c}} \right), \quad (3.30)$$

where $v_{r_c} = \pi r_c^3$ is the volume with radius r_c . To estimate $P_{\text{loop}}^{r_c}$, we model the section of the scaffold that is pinned into a loop as a heterogenous freely-jointed chain, consisting of single- and double-stranded segments. The probability distribution for the end-to-end distance of the chain can be written as:

$$P(r) = 4\pi r^2 \left(\frac{3}{2\pi E[r^2]} \right)^{3/2} e^{\left(\frac{-3r^2}{2E[r^2]} \right)}, \quad (3.31)$$

where $E[r^2] = \sum_{i=1}^K N_i b_i^2$ is the mean-square distance between the two ends consisting of K segments and b_i is Kuhn length of the i^{th} segment. The probability that the ends of the chain are within a distance r_c is then given by:

$$P_{\text{loop}}^{r_c} = \int_0^{r_c} P(r) dr \approx \left(\frac{3}{2\pi E[r^2]} \right)^{3/2} \int_0^{r_c} 4\pi r^2 dr = \left(\frac{3}{2\pi E[r^2]} \right)^{3/2} v_{r_c} \quad (3.32)$$

where we have assumed $r_c \ll E[r^2]$. Substituting back into 3.33, we have:

$$G_{\text{loop}}^{\ominus, \text{FJC}} = -RT \ln \left[v^{\ominus} \left(\frac{3}{2\pi E[r^2]} \right)^{3/2} \right] = -RT \gamma \ln \left(\frac{C_{\text{FJC}}}{E[r^2]} \right), \quad (3.33)$$

where $\gamma = 3/2$ is the loop exponent for a freely-jointed chain, and we have defined $C_{\text{FJC}}(\gamma) = \frac{3}{2\pi} v^{\ominus 1/\gamma} = 0.67 \text{nm}^2$. The loop exponent gives the scaling of the volume available to a chain with its contour length. Excluded-volume effects increase the value of loop exponent in real chains. The model of a freely-jointed chain here to describe the DNA segments is an approximation and we expect the real behaviour of the chain to show large deviations from this description. Therefore, we consider γ and C as fitting parameters. The net effect of increasing C is to systematically increase the stability of all loops, whereas increasing γ decreases the stability of all loops but enhances the entropic penalty of forming long loops relative to short loops.

To account for more realistic polymer chain behaviour, Dannenberg *et al*[59] consider γ as a fitting parameter and ignore the γ -dependence of C ; instead they choose to adjust C together with γ to maintain the loop cost for an 18 nucleotide bulge loop to be the same as in the freely-jointed case. For a given γ , the value of C_γ is given by:

$$C_\gamma = C_{3/2}^{3/2\gamma} (E_{18}[r^2])^{1-3/(2\gamma)}, \quad (3.34)$$

where $E_{18}[r^2]$ is the loop cost for an 18 nt bulge loop. The loop cost in the model can then be written as:

$$G_{\text{loop}}^{\ominus} = -RT\gamma \ln \left(\frac{C_{\gamma}}{E[r^2]} \right). \quad (3.35)$$

In evaluating $E[r^2]$, I follow the same procedure as Ref.[59]. Double-stranded domains of n_{bp} base-pairs are considered a rigid single segment of Kuhn length $n_{\text{bp}}l_{\text{ds}}$, where $l_{\text{ds}} = 0.34\text{nm/bp}$ is the contour length of dsDNA [150]. Single-stranded domains of n_{nt} nucleotides are considered $n_{\text{nt}}l_{\text{ss}}/\lambda_{\text{ss}}$ segments of Kuhn length $\lambda_{\text{ss}} = 1.8\text{nm}$, where $l_{\text{ss}} = 0.6\text{nm/nt}$ is the contour length per nucleotide of ssDNA. Crossovers are represented as a single segment of ‘Kuhn’ length λ_{ss} . The mechanical behaviour of single-stranded DNA is not fully understood and these values were chosen to be roughly consistent with experimental data [151].

To evaluate $G_{\text{SH}}^{\ominus}(\vec{s}_i)$, we consider the origami as a graph $H(\vec{s}_i) = (V, E(\vec{s}_i))$, with vertices $v \in V$ at the junctions between each scaffold domain, and edges $e \in E(\vec{s}_i)$ representing staple crossovers or single- or double-stranded scaffold domains. Each edge is assigned a weight $W(e)$ as follows:

$$W(e) = \begin{cases} (n_{\text{bp}}l_{\text{ds}})^2 & \text{if } e \text{ is a double-stranded domain,} \\ n_{\text{nt}}l_{\text{ss}}\lambda_{\text{ss}} & \text{if } e \text{ is a single-stranded domain,} \\ \lambda_{\text{ss}}^2 & \text{if } e \text{ is a crossover.} \end{cases} \quad (3.36)$$

Assuming that $H(\vec{s}_i)$ is planar, such that we can choose a unique planar embedding and identify a set of faces $F(\vec{s}_i)$, where each face $F_k \in F(\vec{s}_i)$ is a subgraph of $H(\vec{s}_i)$, then we can assign a weight $W(F_k)$ to each face and calculate $G_{\text{SH}}^{\ominus}(\vec{s}_i)$ as:

$$G_{\text{SH}}^{\ominus}(\vec{s}_i) = -RT\gamma \sum_{F_k \in F(\vec{s}_i)} \ln \left(\frac{C_{\gamma}}{W(F_k)} \right), \quad (3.37)$$

where $W(F_k) = \sum_{e \in F_k} W(e)$ is the total weight of the face.

For graphs that are not planar, it is not possible to identify loops unambiguously. Therefore, it is not possible to use this procedure to construct a thermodynamically consistent model. In this case, Dannenberg *et al*[59] suggest a “local” model in which only the most immediate loop is considered during any transition that forms or breaks a crossover. Consider such a transition from state \vec{s}_i to state \vec{s}_j in

which a crossover forms between vertices v_1 and v_2 . Then the change in Gibbs free energy during the transition is given by:

$$\Delta G_{\text{SH}}^{\ominus, \min}(\vec{s}_i, \vec{s}_j) = -RT\gamma \ln \left(\frac{C_\gamma}{\lambda_{\text{ss}}^2 + D(v_1, v_2)} \right), \quad (3.38)$$

where $D(v_1, v_2)$ is the weight of the shortest path between the connected vertices, calculated using Dijkstra's weighted algorithm on the graph $H(\vec{s}_i)$. Note that this method of calculating the shape contribution is path-dependent and cannot be used to construct a global energy model with respect to a single reference state.

There are cases in which only a few states that involve partly bound three-domain staples do not have a planar embedding, but otherwise, the state space of the origami is planar (see Figure 3.2c). In these cases, it is still possible to use the global model with special consideration for any state \vec{s}_{NP} involving staple-complements in the s_{101} configuration. To do this, I apply a mixture of local and global approaches. First, I remove the crossovers that violate planarity from the graph $H(\vec{s}_{\text{NP}})$ and calculate $G_{\text{SH}}^{\ominus}(\vec{s}_{\text{P}})$ using (3.37) based on the new planar graph $H(\vec{s}_{\text{P}})$. Then, for each staple crossover that was removed, I calculate $\Delta G_{\text{SH}}^{\ominus, \min}(\vec{s}_{\text{P}}, \vec{s}_k)$ using (3.38), where \vec{s}_k is the non-planar state in which only the crossover in question is formed. The total $G_{\text{SH}}^{\ominus}(\vec{s}_{\text{NP}})$ is then calculated as the sum of $G_{\text{SH}}^{\ominus}(\vec{s}_{\text{P}})$ and the $\Delta G_{\text{SH}}^{\ominus, \min}(\vec{s}_{\text{P}}, \vec{s}_k)$ for all such crossovers. This results in a thermodynamically well-defined model and is used in chapter 4 to simulate a real origami design.

Dannenber *et al* [59] extensively studied the parameter space by comparing their simulation results to fluorescence measurements of a simple 2D DNA origami during annealing and melting. They found that the best fit to experimental data on hysteresis, the width of the transition, and melting and annealing temperatures were obtained with $n = 2$ and $\gamma = 2.5$ (corresponding to $C_{2.5} = 2.8\text{nm}^2$). In their work, they allowed coaxial stacking between any two domains that are in adjacent positions on the scaffold. As we shall see in chapter 4, coaxial stacking plays an important part in determining the free energy landscape and the dynamics of the systems. I altered the coaxial stacking rules to better reflect the scaffold turns in the origami. This change affects only staples that bind at positions of scaffold crossovers (where

the scaffold strand is routed from one duplex to the next in the target structure). These staples are shown in Figure 3.4 for the origami that was previously studied. I slightly altered the rules to only allow coaxial stacking between domains that occur on the same duplex in the target structure. This affects seam staples and edge staples because they are placed at positions of scaffold turns. For edge staples, coaxial stacking of the two domains of the same fully bound staple is not allowed because the two domains are on different duplexes in the bound state. For seam staples, coaxial stacking is allowed between the two domains if the seam staple is correctly bound (the staple-complement is in state s_{11}), but not allowed if two copies of the seam staple are bound to the scaffold (the staple-complement is in state s_{12}). These rules provide a more realistic picture than the simple stacking of all adjacent domains. I repeated the parametrisation procedure for the origami they considered and various other origami designs, with the results presented in the next sections.

Some possible complications to the stacking rules were left to further work. For example, pairs of seam staples may deviate from this picture if one is correctly bound (s_{11}), but the other is bound by two copies (s_{12}). In this case, it may be more realistic to allow stacking in the s_{12} state because the domains will align due to the crossover made by the other seam staple that is correctly bound.

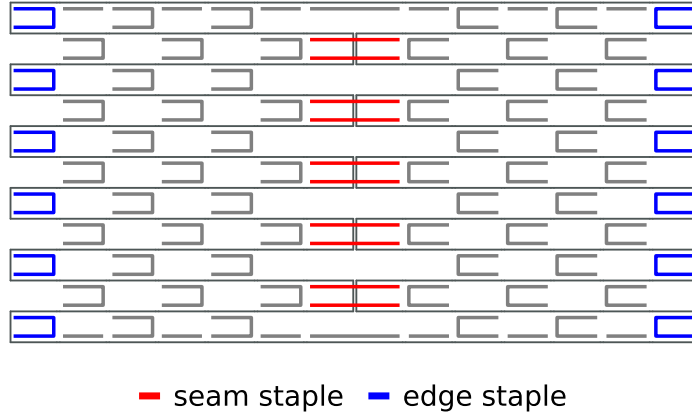


Figure 3.4: Position of seam staples and edge staples. Domains of edge staples cannot coaxially stack when the staple is fully bound (s_{11}). Domains of seam staples can coaxially stack when the staple is fully bound (s_{11}) but cannot stack when two copies of the staple are bound (s_{12}). All other adjacent domains can coaxially stack if they are on the same helix.

3.5 Rate Model

To simulate our system, we consider the various conformations of the scaffold strand as distinct chemical species and keep the chemical potential of the staple strands constant. This corresponds to the assumption that staple strands are in large excess compared to the scaffold. Transitions ($\vec{s}_i \rightarrow \vec{s}_j$) considered in the model are shown in Figure 3.5 and take the general form:



where \mathbf{X}_i and \mathbf{X}_j are partially-folded scaffold strands with or without bound staples, \mathbf{I} and \mathbf{J} are two staple strands, $\beta_i, \beta_j \in \{0, 1\}$ are stoichiometric coefficients that identify the numbers of free \mathbf{I} and \mathbf{J} strands in each state, and k_{ij} and k_{ji} are deterministic reaction rate constants. The reaction propensities given by (3.8) can be written for our system as:

$$w_{ij} = k_{ij} (N_A V)^{-\beta_i} n_{\mathbf{X}_i} n_{\mathbf{I}}^{\beta_i} = k_{ij} c_{\mathbf{I}}^{\beta_i} \quad (3.40)$$

$$w_{ji} = k_{ji} (N_A V)^{-\beta_j} n_{\mathbf{X}_j} n_{\mathbf{J}}^{\beta_j} = k_{ji} c_{\mathbf{J}}^{\beta_j} \quad (3.41)$$

where n_k and c_k are the number of molecules and molar concentrations of species k , respectively. In the second equality, we have assumed a ‘scaffold-eye’ view such that the number of molecules of the scaffold strand is one ($n_{X_i} = n_{X_j} = 1$), and used $n_I = c_I N_A V$ and $n_J = c_J N_A V$, which is valid for smoothly-varying staple concentrations. We can use the condition of detailed balance (3.22) to relate the forward and backward rates at equilibrium:

$$\frac{k_{ij}}{k_{ji}} = e^{-\Delta G^\ominus(i,j)/RT} (c^\ominus)^{(\beta_j - \beta_i)} \quad (3.42)$$

where $\Delta G^\ominus(i, j)$ refers to the change in Gibbs free energy for the reaction under standard conditions. Since only the relative rate constants are fixed by this condition, we have the freedom to set either the forward or backward rate. The specific rate model is usually chosen based on considerations of computational performance and based on available experimental data. Since the main computational cost in the model is the evaluation of ΔG_{SH} , I use the model described in Table 3.1 for transitions of type (a)-(e) depicted in Figure 3.5. This custom model is faster and does not significantly differ from the more traditional Metropolis and Kawasaki algorithms. However, it allows us to minimise the number of evaluations of ΔG_{SH} .

For type (a), we assume that the forward rate constant is independent of temperature and sequence and use the experimentally determined [146] rate constant $k_f = 3.5 \times 10^6 \text{M}^{-1} \text{s}^{-1}$ for forward reactions of this type. This is intuitive since the binding rate of an isolated duplex is less strongly dependent on duplex stability than its dissociation. For type (b), we fix the unbinding rate to that of an isolated duplex, and account for the entropic changes of the scaffold ($\Delta G_{\text{SH}}^\ominus$) in the rate constant for the forward reaction. For type (c), we follow the two-step phenomenological model of Ref. [146] as described in section 3.3. Since the reaction is symmetric, we split the Gibbs free energy equally in both directions as in the Kawasaki method. For type (d), the reaction is no longer symmetric since a loop forms in the $i \rightarrow j$ direction but breaks in the $j \rightarrow i$ direction. To minimise computational costs of calculating loop costs, we follow a similar approach to that in (b) and only

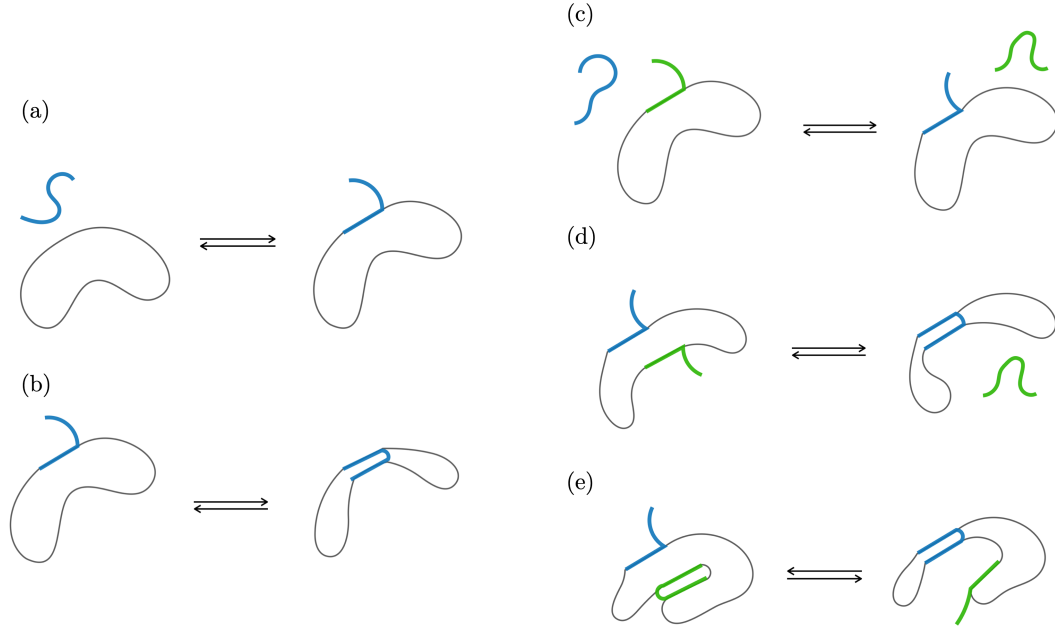


Figure 3.5: Allowed transitions in the model. (a) Bi-molecular binding/unbinding of a first domain of a staple to the scaffold. (b) Uni-molecular binding/unbinding of a further domain of a partly-bound staple to the scaffold. (c)-(e) Toehold-mediated strand exchange reactions involving staples designed to bind to overlapping regions of the scaffold.

consider $\Delta G_{\text{SH}}^{\ominus}$ in the forward direction. For the symmetric reaction type (e), we follow the same approach as (c).

There are transitions involving three-domain staple-complements that require special consideration. As noted in section 3.3 (see Figure 3.2c), the graph involving staple-complements in the s_{101} or s_{121} configurations is not planar and requires special considerations in calculating its ΔG_{SH} in the global approach. There are also topological considerations in the transition $s_{101} \leftrightarrow s_{111}$ since the staple and scaffold domains cannot twist around each other to form a double-helix without breaking the flanking duplexes. Therefore, this transition is forbidden from occurring directly, but there are still indirect paths between the two states via intermediate unbinding of other domains that allow it to occur.

Type	β_i	β_j	w_{ij}	w_{ji}
(a)	1	0	$k_f c_I$	$k_f e^{\Delta G^{\ominus}(i,j)/RT} c^{\ominus}$
(b)	0	0	$k_f e^{-\Delta G_{SH}^{\ominus}(i,j)/RT} c^{\ominus}$	$k_f e^{[\Delta G_{NN}^{\ominus}(i,j) + \Delta G_{CX}^{\ominus}(i,j)]/RT} c^{\ominus}$
(c)	1	1	$k_{nm} e^{-[\Delta G_{CX}^{\ominus}(i,j) + \Delta G_{SH}^{\ominus}(i,j)]/2RT} c_I$	$k_{mn} e^{-[\Delta G_{CX}^{\ominus}(j,i) + \Delta G_{SH}^{\ominus}(j,i)]/2RT} c_J$
(d)	0	1	$k_{nm} e^{-\Delta G_{SH}^{\ominus}(i,j)/RT} c^{\ominus}$	$k_{mn} e^{-\Delta G_{CX}^{\ominus}(j,i)/RT} c_J$
(e)	0	0	$k_{nm} e^{-[\Delta G_{CX}^{\ominus}(i,j) + \Delta G_{SH}^{\ominus}(i,j)]/2RT} c^{\ominus}$	$k_{mn} e^{-[\Delta G_{CX}^{\ominus}(j,i) + \Delta G_{SH}^{\ominus}(j,i)]/2RT} c^{\ominus}$

Table 3.1: Custom rate model used in this work. The propensities w_{ij} and w_{ji} for transitions of type (a) to (e) shown in Figure 3.5.

3.6 Sampling Techniques

We aim to sample the grand canonical ensemble⁸ (μVT) in a hypothetical volume containing one scaffold strand in a pool of staples at excess concentration. Under the law of mass action, the reaction rates in the forward and backward directions are given by:

$$R_{ij} = k_{ij} c_{X_i} c_I^{\beta_i} \quad (3.43)$$

where c_{X_i} is the molar concentration of the scaffold strand at state \vec{s}_i , and c_I and c_J are molar concentrations of the staple strands. In a fixed volume, we could obtain the steady-state concentration of the different conformations of the scaffold strand by setting $R_{ij} = R_{ji}$ and keeping the concentration of the staple strands constant ($c_I = \{c_I\}$). Combining this with (3.42) the equilibrium concentrations are then given by:

$$\frac{\{c_{X_j}\}}{\{c_{X_i}\}} = e^{-\Delta G^{\ominus}(i,j)/RT} \left(\frac{c_I}{c^{\ominus}} \right)^{\beta_i} \left(\frac{c^{\ominus}}{c_J} \right)^{\beta_j}. \quad (3.44)$$

⁸The grand canonical ensemble is a statistical ensemble of particles in a fixed volume V at thermodynamic equilibrium with a reservoir at temperature T while allowing the exchange of particles with the reservoir to maintain a constant chemical potential μ .

Consider a reaction pathway $\psi_{pq} \equiv \vec{s}_p \rightarrow \vec{s}_{p+1} \rightarrow \dots \rightarrow \vec{s}_{q-1} \rightarrow \vec{s}_q$ between non-adjacent states \vec{s}_p and \vec{s}_q . If detailed balance is satisfied, the change in free energy is path-independent and we have:

$$\Delta G^\ominus(\vec{s}_p, \vec{s}_q) = \sum_{\vec{s}_i, \vec{s}_j \in \psi_{pq}} \Delta G^\ominus(\vec{s}_i, \vec{s}_j), \quad (3.45)$$

where the sum can be taken over any path linking the states. Let \vec{s}_0 be the state of the scaffold with no staples and let $\Delta G^\ominus(\vec{s}_0) = 0$. It is easy to see that under the rate model described above, the equilibrium concentrations can be written relative to the empty state as:

$$\frac{\{c_{X_j}\}}{\{c_{X_0}\}} = e^{-\Delta G^\ominus(j)/RT} \prod_{st \in \vec{s}_j} \frac{c_{st}}{c^\ominus}, \quad (3.46)$$

where c_{st} are the concentrations of staple strands that are bound to the scaffold at state \vec{s}_i , $\{c_{X_0}\}$ is the equilibrium concentration of the empty scaffold, and $\Delta G(\vec{s}_i) = \Delta G(\vec{s}_0, \vec{s}_i)$. Instead of simulating a fixed volume, we take the scaffold-eye view and follow a single scaffold in a hypothetical volume. The probability of finding a the scaffold in state \vec{s} at equilibrium is given by:

$$p(\vec{s}) = \frac{1}{Z} e^{-\Delta G^\ominus(\vec{s})/RT} \prod_{st \in \vec{s}} \frac{c_{st}}{c^\ominus}, \quad (3.47)$$

where Z is a normalisation factor. Sampling this distribution allows us to compute ensemble averages of observables of interest $A(\vec{s})$ such as the number of domains occupied on the scaffold. If the $A(\vec{s})$ takes a set of possible values $\{a\}$, the probability distribution over A is given by:

$$p(a) = \sum_{\vec{s}} p(\vec{s}) \delta_{a, A(\vec{s})}, \quad (3.48)$$

where $\delta_{a, A(\vec{s})} = 1$ if $A(\vec{s}) = a$ and zero otherwise. We can then construct the free energy profile $F(a) = -k_B T \ln(p(a))$ over the values of the observable.

When there are large free energy barriers between the macrostates defined by the observable, the simulation becomes trapped and may not explore all macrostates in reasonable timescales. We can use a biasing scheme to artificially flatten the energy barriers and speed up transitions between the local minima. The biasing scheme

described here is similar to the Umbrella Sampling technique and was recently applied to DNA brick assembly [70]. It involves specifying an order parameter $Q(\vec{s})$ in terms of one or more observables of interest, for example, $Q(\vec{s}) = Q(a)\delta_{a,A(\vec{s})}$ for a single observable A . Each value of the observable is assigned a weight $\phi[Q(\vec{s})]$ such that the biased probability distribution is given by $p_\phi(\vec{s}) = \phi^2[Q(\vec{s})]p(\vec{s})$. To sample from the biased distribution, the reaction propensities are changed such that detailed balance condition in the biased simulation is still satisfied:

$$p_\phi(i)w_\phi(i, j) = p_\phi(j)w_\phi(j, i), \quad (3.49)$$

where I have used the simplified notation $\vec{s}_i = i$ for clarity. Substituting in $p_\phi(\vec{s})$ into the above equation, we can relate the biased propensities to the unbiased ones through:

$$w_\phi(i, j) = \frac{\phi[Q(j)]}{\phi[Q(i)]}w(i, j). \quad (3.50)$$

The weights are chosen such that the biased probability distribution remains flat over the range of the order parameter. Then the distribution over the observable is given simply by:

$$p(a) = \sum_i \frac{p_\phi(i)}{\phi^2[Q(i)]}\delta_{a,A(i)} = p_\phi(a)/\phi^2[Q(a)], \quad (3.51)$$

where $p_\phi(a)$ is the probability distribution of A in the biased simulation.

Here, I present a simple system to gauge the convergence of the ensemble averages using this biasing scheme. The origami $\mathcal{R}2^{4 \times 4}$ shown in Figure 3.6(a) is designed to fold into a rectangle of four helices, each consisting of four 16bp domains, using 6 two-domain and 4 single-domain staples. The number of possible states of the scaffold is given by $5^6 \times 2^4 = 250,000$. It is therefore possible to calculate the exact probability distribution by manually computing the free energy of each state according to (3.47). I also run several biased SSA simulations to compute the free energy landscape $p(n_d)$ according to (3.51), where n_d is the number of bound domains on the scaffold. The initial state in all simulations was chosen at random. There are three factors to consider when comparing the simulation results

to the exact distribution. First, I can vary N_{step} , which is the number of transitions during a single simulation with a specific random number generator seed. Second, I vary N_{seed} , which is the number of simulations that are averaged over, allowing a measure of the uncertainty. Third, I change the input weights $\phi(n_d)$ to gauge the effect of the flatness criterion on the accuracy of the results.

The results are shown in Figure 3.6(c)-(f) for the local and global model, for two sets of weights $\phi_1(n_d)$ and $\phi_2(n_d)$, averaged over $N_{\text{seed}} \in \{1, 5, 10, 50, 99\}$ simulations, each with $N_{\text{step}} \in \{10^4, 2 \times 10^4, 4 \times 10^4, 10^5, 10^6\}$ transitions. In each case, I take the average of $p_\phi(n_d)$ across the considered seeds and compute $\Delta F(n_d) = \ln(p(0)) - \ln(p(n_d))$, propagating the standard deviation as a measure of the uncertainty. The set of weights $\phi_1(n_d)$ produces an almost flat $p_{\phi_1}(n_d)$ distribution and converges quickly with increasing total number of transitions $N_{\text{seed}} \times N_{\text{step}}$ to the exact result. However, the number of seeds is the more important factor with $N_{\text{seed}} = 10$ case converging much more quickly than $N_{\text{seed}} = 5$, even with a small number of steps per simulation. As a general rule of thumb, I generate at least 100 paths for systems discussed in this thesis. Interestingly, the local model produces nearly identical results. The largest contributor to the free energy is always the loop immediately associated with the crossover during a transition. Since the local model takes this loop cost into account, the deviations from detailed balance remain small during the simulation. I will sometimes use the local model for some of the systems; this will be indicated in each figure. I performed similar simulations and compared them to the exact results for other origami designs of this size $|S| < 10^6$. In general, I found that a flatness criterion of less than 10% produced nearly identical results to the exact case. For larger origami sizes, where the state space grows exponentially, I break up the state space into smaller sampling windows, with at least three points of the order overlapping between each adjacent window. For each window, the weights are iteratively improved until the flatness criterion is satisfied.

There are interesting features in the 1D free energy profile $\Delta F(n_d)$. Most staples have two domains where the binding of the second domain closes a loop. This

explains the drop in free energy for even n_d . Most states contributing to the region $n_d > 12$ are ones in which all two-domain staples are already bound, with less stable single-domain staples coming in, explaining the monotonous increase in free energy in this region. We also see a large decrease in free energy for $n_d = 4$ and this is due to the stability provided by the pair of seam staples in the middle of the origami. In the absence of other staples, each seam staple closes a long loop but the second seam staple closes the shortest possible loop: a double scaffold crossover. This structure is further stabilised by the stacking interaction between the two domains of each seam staple, leading to cooperative behaviour between the four domains that are involved in the structure. To verify this hypothesis, we can look at 2D free energy profiles where the number of seam domain n_{seam} is used alongside n_d as an order parameter. In Figure 3.6(b), I compare the exact 2D profile with SSA simulations in which n_{seam} was used as a biased order parameter and a flat distribution $p_\phi(n_{\text{seam}})$ was obtained. It is clear that the minima occur for $n_{\text{seam}} = 4$ with relatively high barriers at odd n_{seam} . Note that no bias was applied to n_d in the SSA case (although it is possible) and some states such as $n_{\text{seam}} = 0, n_d = 12$ were left unsampled. The main features of the free energy profile are still reproduced using the SSA approach.

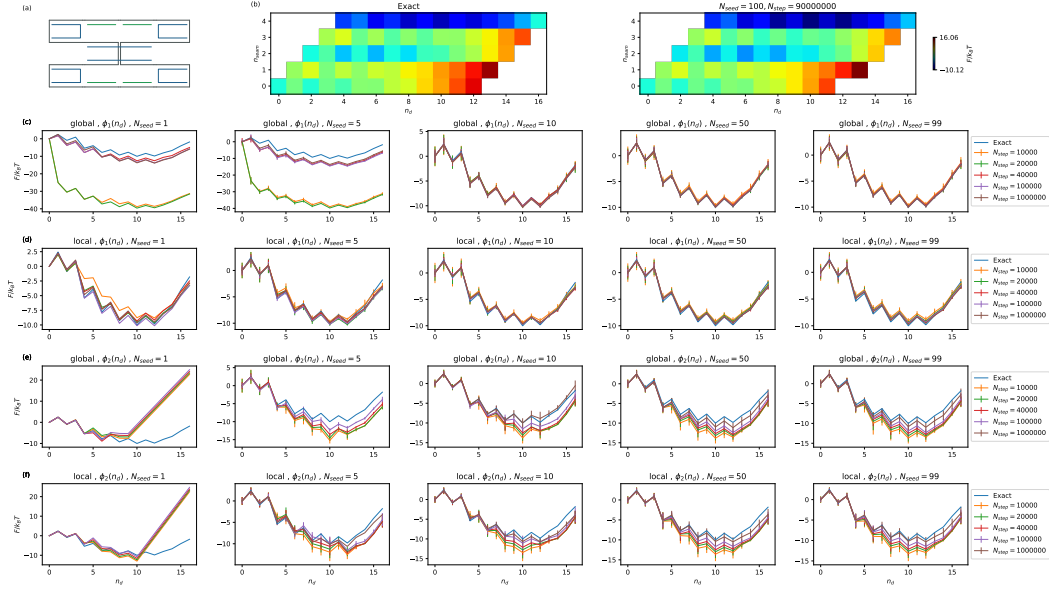


Figure 3.6: Testing the umbrella sampling method. All simulations are performed with $n = 2$ and $\gamma = 2.5$ at 66.8°C . (a) The origami $\mathcal{R}2^{4 \times 4}$ consists of 6 two-domain staples and 4 single-domain staples. (b) Exact 2D free energy profile is compared to SSA simulations averaged over 100 trajectories, each with 9×10^7 steps. In each trajectory, the number of seam domains n_{seam} was used as the biased order parameter to obtain a flat distribution but the second order parameter, the number of bound domains, n_d was not weighted. (c) Free energy profile plotted as a function of n_d , with a set of weights $\phi_1(n_d)$ that produced a flat distribution $p_{\phi_1}(n_d)$ is plotted along with the exact free energy profile. Results are shown for different values of N_{step} , which is the maximum number of steps a trajectory is allowed to run. In each column, results are averaged over a set number of trajectories N_{seed} . (d) Same as (c) but the local model is used instead of the thermodynamically consistent global model. (e) Same as (c) but the weights used ϕ_2 no longer produce a flat distribution. (f) Same as (e) but the local model is used.

3.7 Comparison to Experiment

Initially, the experiments described here were designed and carried out by Jonathan Bath, who kindly ordered the required oligos based on his previous work. The scaffold strand has been used extensively in the Turberfield group; it was derived by Katherine Dunn who synthesised a custom plasmid (pKD1) from the pUC19 vector. A second scaffold strand (pJB101) was synthesised in a similar manner by Jonathan Bath for this work. The two strands are both 2646 nt in length, but a 572 nt section is replaced with a different sequence in the case of pJB101 as shown in Appendix A. The set of fluorescence experiments presented here was done by me while varying certain parameters to produce the figures in this section.

In each experiment, the scaffold and staple strands were mixed in a standard buffer solution with 125mM MgCl_2 and an intercalating dye (EvaGreen from IDT) that changes fluorescence upon binding to double-stranded DNA. The scaffold strand concentration is set at 50 nM in all experiments, but the following parameters were varied:

- The heating/cooling rates: $\pm 0.5, 1, 2, 4$ Kelvin per minute.
- The staple concentrations: 100, 200, 400 nM.
- Scaffold strands: pKD1 and pJB101 (sequences shown in Appendix A)
- Topologies: \mathcal{E} , $\mathcal{R}2$, $\mathcal{R}3$, \mathcal{T} , $\mathcal{R}2_H$, $\mathcal{R}2_N$ (see Figure 3.7)

I ran a set of simulations with the same parameters for comparison to the experiments. In addition, I varied the rate constant k_f , which sets the timescale of the model and is equivalent to changing the cooling/heating rate. Each simulation was run under a stepwise temperature ramp $50^\circ\text{C} \leq T(t) \leq 80^\circ\text{C}$ in 0.5°C increments. The amount of time in each temperature increment was adjusted to achieve the desired cooling/heating rate. The fluorescence was taken as a proxy for the total number of bps of the scaffold in duplex form n_{bp} when comparing results simulation and experimental results. For each simulation n_{bp} was recorded as a rolling average in 0.5°C increments and averaged over 60 heating and cooling cycles.

In Figure 3.8, I outline the procedure used to compare fluorescence measurements to the simulation data. It is not immediately obvious how the simulation and experimental results can be compared. The experimental data is difficult to interpret since staple-staple interactions cause background noise in the data, especially at lower temperatures. Given the general coarseness of the experimental data and the simplifications of the model, it is better to aim for qualitative agreement on a general set of trends that change as parameters are varied rather than aiming for quantitative numerical agreement. Therefore, I use T_{50} , the temperature that the origami is half-bound (during annealing or melting) as a measure of stability, and use the fitted Hill coefficient n_H as a measure of cooperativity between domains.

Other methods such as taking the derivative of the fluorescence gave similar results and were used to visually confirm trends. However, it was not as consistent because some designs include domains of different lengths leading to several peaks in the temperature profile of the derivative, making comparison difficult.

The results under a temperature ramp of 1°C/min at 100 nM staple concentration are shown in Figure 3.9 for the origamis in Figures 3.7. The simulations capture the subtle effects of changing the origami design on both the stability, characterised by $\bar{T}_{50} = (T_{50}^{\text{anneal}} + T_{50}^{\text{melt}})/2$ and average cooperativity \bar{n}_H . Both \bar{T}_{50} and \bar{n}_H follow the same trend in simulation and experiment. Exact values do not agree and this is expected given the simplicity of the model and the coarseness of the experimental data. However, the effect of small changes such as breaking a seam staple leads to decreased stability and cooperativity in both sets of data. Similar agreement was obtained for other temperature ramps and concentrations.

In Figure 3.10, I explore the effect of cooling rate and staple concentration on the observed hysteresis, which is characterised by the difference in T_{50} or n_H during melting and annealing. There are some anomalies in experimental results at very slow cooling rates. Nonetheless, hysteresis tends to increase in both simulation and experiment at faster cooling rates and with decreasing staple concentration. I found $k_f = 5 \times 10^5 \text{M}^{-1}\text{s}^{-1}$ gave the best agreement in the scaling of the hysteresis with experimental results and fixed this value for further simulations.

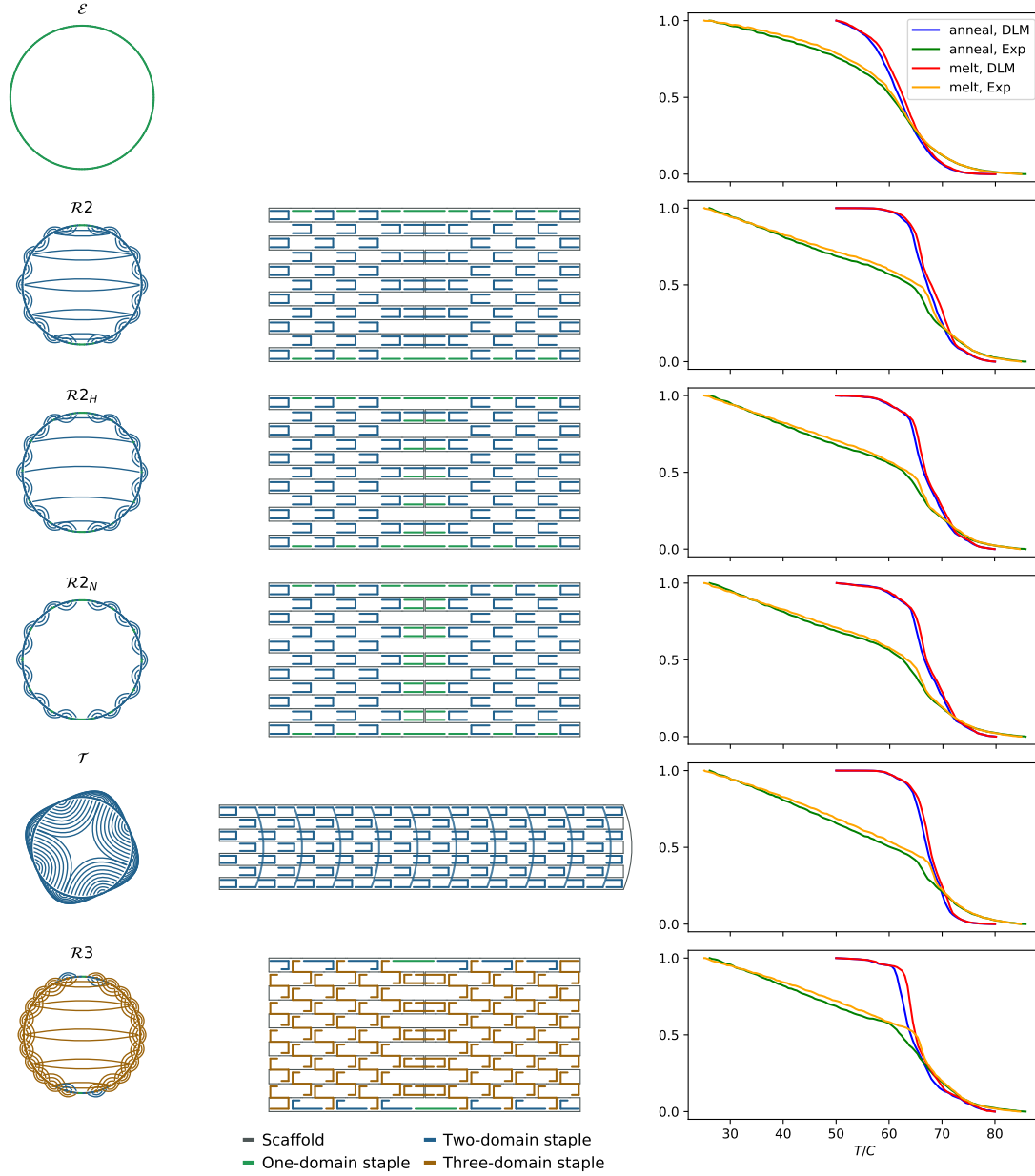


Figure 3.7: Comparison of annealing and melting curves between simulation and experiment for different origami designs. For each design, staple connectivity on rectangular and circular layouts of the scaffold are shown schematically. Normalised melting and annealing curves from domain-level model (DLM) simulations and fluorometric experiments are also shown on the right. \mathcal{E} is an empty design similar to a plasmid. $\mathcal{R}2$ and $\mathcal{R}3$ are designed to fold into flat rectangular shapes. $\mathcal{R}2_H$ and $\mathcal{R}2_N$ are half-seam and no-seam variations of $\mathcal{R}2$. \mathcal{T} is an 8-helix tube. Note that the transitions occur in the same temperature range in simulation and experiment.

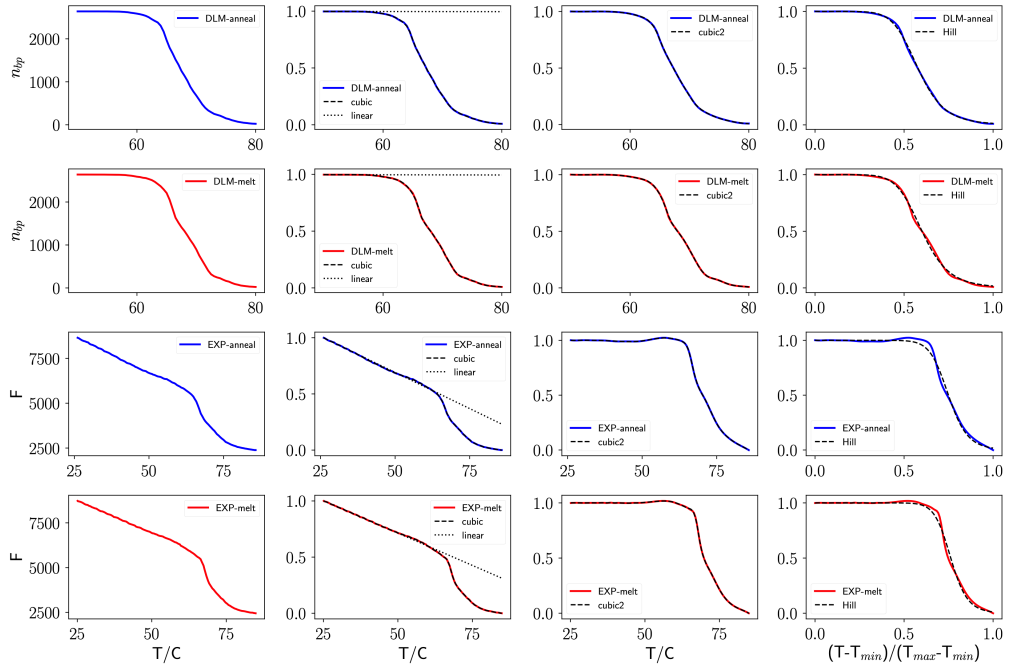


Figure 3.8: Fitting procedure for annealing and melting transitions. The first column shows raw annealing/melting data for the design $\mathcal{R}2$, with staples at a concentration of 100 nM. The top two rows show the number of base pairs as a function of temperature during annealing/melting simulations. The bottom two rows show fluorescence data from annealing/melting experiments. The second column shows equivalent normalised data; a linear fit is applied to correct for background fluorescence arising from staple-staple interactions. The third column shows the corrected data. In the fourth column, a transformation is applied to the temperature axis, and the data is fitted to the function $f(x) = 1/(1 + (x/x_{50})^{n_H})$, extracting n_H ; the reverse transformation yields T_{50} .

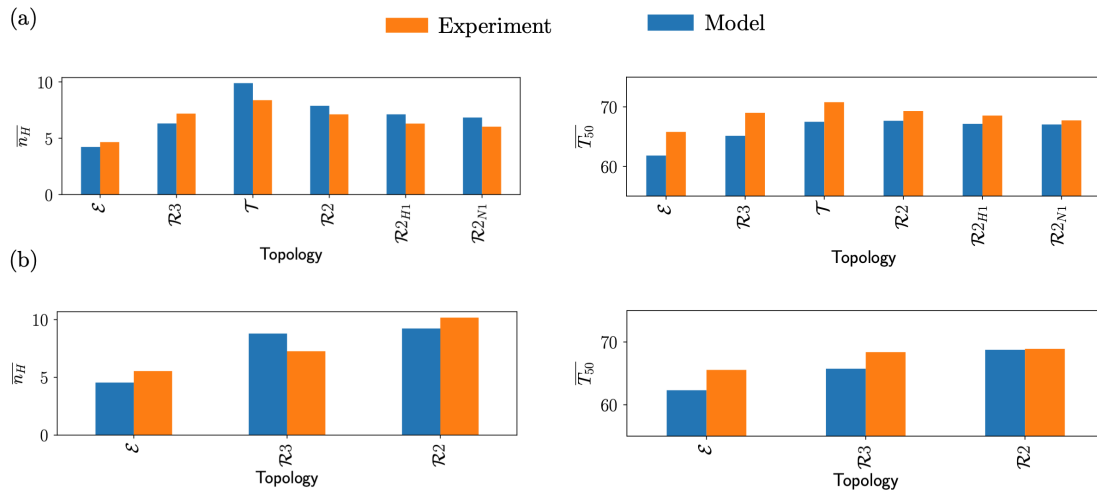


Figure 3.9: Comparison of the domain-level model to experiment. Extracted values of T_{50} and n_H averaged over melting and annealing simulations/experiments for various systems, with staples at 100nM and a temperature ramp of 1°C min^{-1} . (a) Using pJB scaffold. (b) Using pKD scaffold (only some designs).

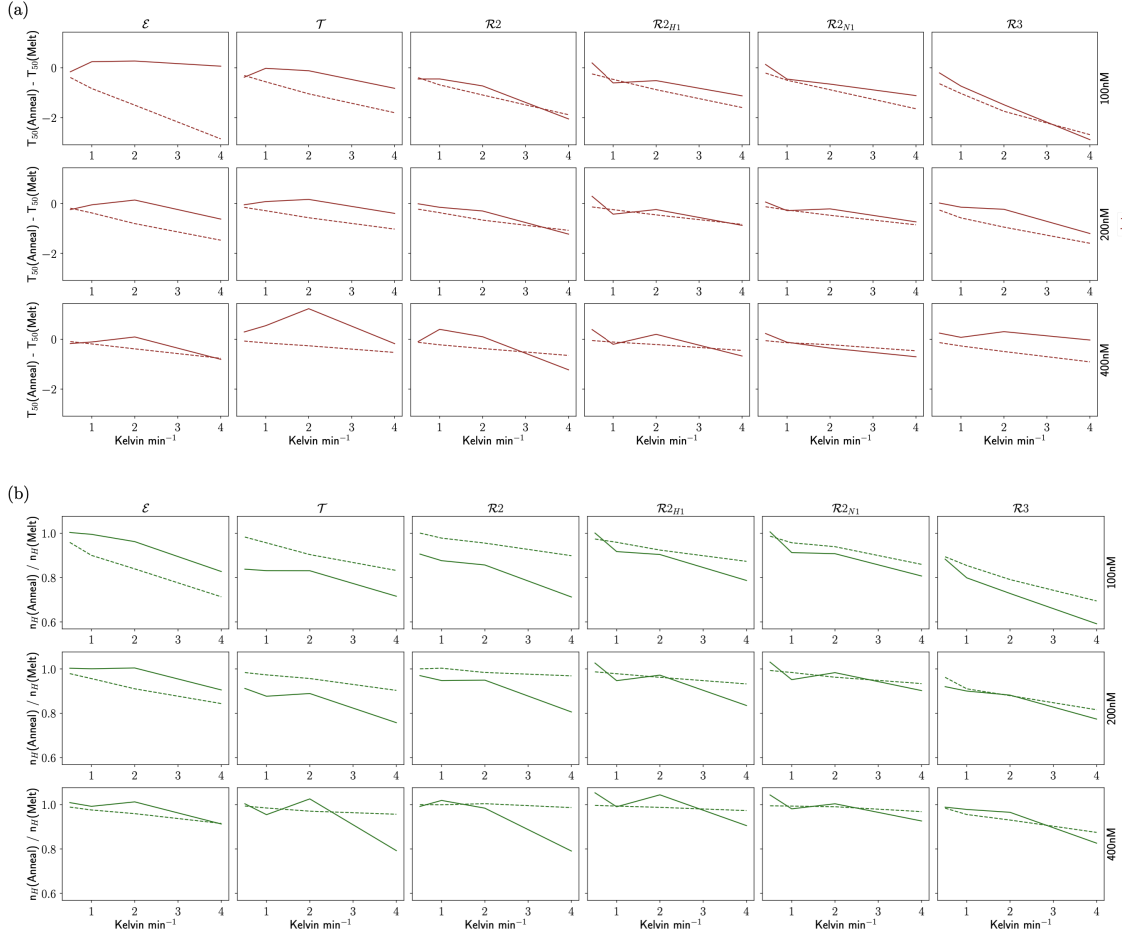


Figure 3.10: Comparison of the domain-level model to experiment: effect of staple concentration and temperature ramp. Observed hysteresis in simulation (dashed lines) and experiment (solid lines) as a function of cooling rate at three staple concentrations (rows) for six topologies (columns). (a) Hysteresis in T_{50} . (b) Hysteresis in n_H .

4

Thermodynamics and Kinetics of DNA Origami Assembly

Contents

4.1	Introduction	89
4.2	Toy Systems	90
4.2.1	Equilibrium Properties	90
4.2.2	Simulations under a Temperature Ramp	98
4.3	Real Origami	101
4.3.1	System $\mathcal{R}2$	103
4.3.2	System $\mathcal{R}3$	109
4.3.3	System \mathcal{T}	115
4.4	Summary and Outlook	118

4.1 Introduction

In this section, I apply the domain-level model (DLM) described in the last section to explore the folding pathway of DNA origami in the presence of a single set of staples. We begin with toy systems to understand the effect of various design features and model parameters. We will then move to real origami systems and compare kinetic data from annealing and melting simulations, resolved at the level of individual domains, to explain features of experimental melting curves for different topologies.

These will be accompanied by free energy profiles to explain the thermodynamic causes of the different folding pathways and the role of nucleation.

4.2 Toy Systems

For very small systems, it is possible to solve the probability distribution of states of the origami exactly. Before moving on to more realistic systems, it is useful to see how the free energy profile changes as more complex design features are added to these toy systems. Some of the systems presented here are not physically viable and involve crossovers between duplexes that do not respect DNA helicity. I designed a set of small systems to gauge the effect of various commonly-used design features.

4.2.1 Equilibrium Properties

For each system, I calculate the free energy of every possible state as follows:

$$\Delta G^\ominus(\vec{s}) = \sum_{d \in \vec{s}} \Delta G_{\text{NN},d}^\ominus(T) + n \sum_{cx \in \vec{s}} \langle \Delta G_{\text{NN}}^\ominus(T) \rangle_{\text{bp}} - RT\gamma \sum_{F_k \in F(\vec{s})} \ln \left(\frac{C_\gamma}{W(F_k)} \right), \quad (4.1)$$

where the first sum is over the scaffold domains that are bound in \vec{s} , the second sum is over design-dependent coaxial stacking neighbours and the third sum is over the faces of the planar embedding of the graph. Once the total loop costs $W(F_k)$ are calculated for each state of a given design, it is easy to change model parameters or the system temperature and see the effect on the free energy profile. The free energy profile is constructed using the equilibrium probabilities (3.47) and (3.48) across various reaction coordinates. For computational simplicity, I define the melting temperature T_M^{eq} as the temperature at which the target structure has the same free energy as the single-stranded scaffold. Partly-formed structures may have lower free energy than either empty or fully-formed states.

Figure 4.1 shows free energy profiles and staple binding probabilities for four systems using the default parameter set ($n=2$, $\gamma=2.5$, $C=C_{2.5}$) and staple concentrations at 100nM. The sequence of the scaffold strand is taken to be the first 128 nucleotides of the pKD1 strand in Appendix A. The sequence-averaged results in the figure take the hybridisation energies of all domains to be the same as the

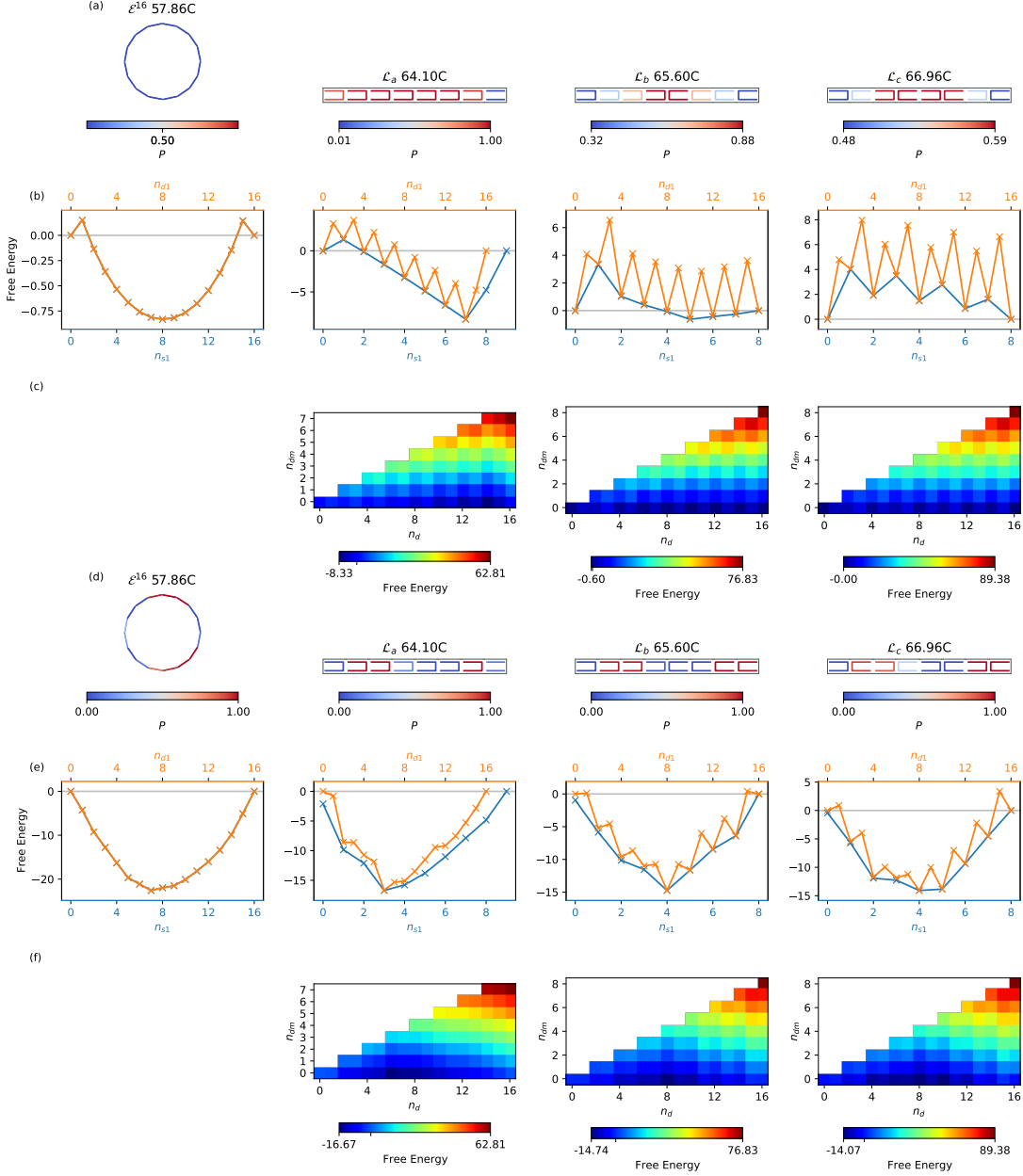


Figure 4.1: Thermodynamics of small empty and two-row systems at their melting temperature. (a) The probability of correct binding for each staple present in the system at the melting temperature (shown at the top) using sequence-averaged duplex energies. (b) Free energy profiles of systems shown in (a) as a function of the number of correctly bound staples n_{s1} and number of correctly bound domains n_{d1} . (c) Free energy profile as a function of two order parameters: n_d is the total number of bound or blocked domains, and n_{dm} is the number of blocked domains. (d),(e),(f) are sequence-dependent versions of (a),(b),(c).

average of the sequence-specific case. In this case, the final state has the same free energy regardless of sequence-specificity, which results in the same melting temperature T_M^{eq} given our definition.

The empty system \mathcal{E}^{16} is plasmid-like and consists of 16 domains/staples, each taken to be 16 bp in length. Relative to other systems, the combinatorial entropy needed for formation is high since each of the 16 scaffold domains requires a bimolecular reaction to hybridise. On the other hand, the conformational entropy of the scaffold is also high in the target structure since the addition of staples does not restrict the scaffold through loops. All adjacent domains are taken to be stacking neighbours and the only source of cooperativity is coaxial stacking between single-domain staples. In the sequence-averaged case (Figure 4.1b) at the melting temperature (57°C), binding of any individual staple is unfavourable since the melting temperature of an isolated 16 bp duplex at the same concentration is approximately $\sim 52^\circ\text{C}$. This leads to positive free energy for the macrostate $n_{s1}=1$. Subsequent staples can be energetically favourable due to stacking; the well-shaped free energy profile is a result of changing multiplicity of the macrostates, with $n_{s1}=8$ having the highest multiplicity. Binding of the final staple results in twice the coaxial stacking energy gain, leading to a drop in free energy.

The two-helix system \mathcal{L}_a consists of 9 staples, 7 of which form loops. Given the lower number of staples, \mathcal{L}_a is more stable than \mathcal{E}^{16} and melts at a higher T_M^{eq} . The number of stacking interactions is reduced in \mathcal{L}_a relative to \mathcal{E}^{16} since domains at the edges of the scaffold helices are not considered stacking neighbours. Once the first domain of a two-domain staple is bound, the local concentration of the staple is increased at the binding site of the second domain leading to more favourable binding in steps of $n_{d1}=2$. The addition of each staple further lowers the loop cost for subsequent staples, leading to cooperative behaviour between staples. Coaxial stacking still plays a dominant role, as seen by the higher probability of central domains being bound than edge domains; the balance between the two effects modulated by n and γ will be discussed further below.

Figures 4.1c,f show the free energy of blocked (the staple is in state s_{12})¹ states on a 2D landscape. The probability of blocked states is extremely low in all cases including the sequence-dependent case.

The system \mathcal{L}_b consists of 8 two-domain staples and is, therefore, more stable than \mathcal{L}_a . A new feature is visible in the free energy profile in steps of 4 domains. This is because the shortest possible loop is closed with the addition of a second crossover when the two central staples are both correctly bound. This design feature is based on the double-crossover tile, which was one of the earliest rigid motifs used in DNA nanotechnology and is used extensively in a majority of origami designs.

The system \mathcal{L}_c also consists of 8 two-domain staples. However, the double-crossover design element is used four times in \mathcal{L}_c , leading to a higher melting temperature than \mathcal{L}_b . In this case, at the melting temperature, the most stable states are either the fully bound or the empty state.

In the sequence-dependent case (Figure 4.1d-f), the heterogeneity of binding energies changes the free energy landscape by causing CG-rich staples to bind first. In the case of the empty system \mathcal{E}^{16} , this introduces an asymmetry into the landscape in line with the distribution of CG-rich base pairs across domains. The free energy profiles of the other systems show a high degree of similarity to \mathcal{E}^{16} because domains are placed at the same position on the scaffold in all the designs. Compared to the average energy case, the probability of partial staple binding increases. This can be seen in Figure 4.1e by the non-zero free energy of states with $n_{s1} = 0$.

In Figure 4.2, I explore the effect of varying the staple concentration, stacking strength n , loop exponent γ and associated parameter C on the free energy profiles at T_M^{eq} . An equivalent figure in which the parameters are varied at constant temperature is shown in Appendix C.1 for reference. Increasing the staple concentration raises T_M^{eq} and lowers the cost association of the first arm of every staple as expected.

Increased stacking strength leads to an increase in T_M^{eq} and universally raises the free energy barriers by increasing local cooperativity between adjacent staples. Note that the free energy barrier does not rise with increasing n if the temperature

¹blocked does not refer to the binding of domains with partial sequence complementarity. It refers to states where two copies of the same staple are bound to the scaffold.

is kept constant (Appendix C.1). This behaviour is in line with a recent study by Cumberworth *et al* [152]. Their model is a less coarse lattice model of DNA that treats coaxial stacking across continuous helices (along multiple bound domains) by a predefined set of rules designed to obey steric constraints. This more complex treatment allows for adjacent domains to break coaxial stacking, even if both domains are bound, leading to possible non-planar geometries. They studied similarly small systems and found that increasing the stacking strength leads to increased nucleation barriers, while systems with larger coordination numbers (higher number of domains per staple) naturally have higher barriers than their lower coordination number counterparts.

Increasing γ at constant C destabilises all loops while exaggerating the difference in stability between loops. This applies to both differences in stability due to both the length of the loop and the type of segments in the loop. Increasing C at fixed γ stabilises all loops equally, leading to an increase in T_M^{eq} . The effect can be seen more clearly in Appendix C.1, where the temperature is kept constant as parameters are varied. We couple the changes in γ and C so that the values of ΔG_{loop} remain physically reasonable and give the best agreement with experimental data. The effect of increasing γ in this way on T_M^{eq} depends on the lengths of loops present in the system. As discussed in the previous chapter, we maintain the loop cost for an 18 nt bulge when we couple the changes in γ and C . Therefore, as γ is increased, any loops with a higher cost than an 18 nt bulge become more costly, but loops with a lower cost than an 18 nt bulge become less costly. The only loops in these systems that become more stable with increasing γ are those formed by double crossovers since all other loops are longer than 18 nt. Since there are many such loops in \mathcal{L}_c , this results in slightly increased T_M^{eq} as γ is increased. Importantly, for $\gamma < 4$, that binding of the second staple that forms such a double crossover is still less favourable than a continuous 32 bp duplex, even when coaxial stacking strength is set to $n = 2$. If γ is raised further however, the picture becomes physically unreasonable because four bound domains (two staples) in a double crossover configuration become more stable than if the staples were laid in a line adjacent to each other.

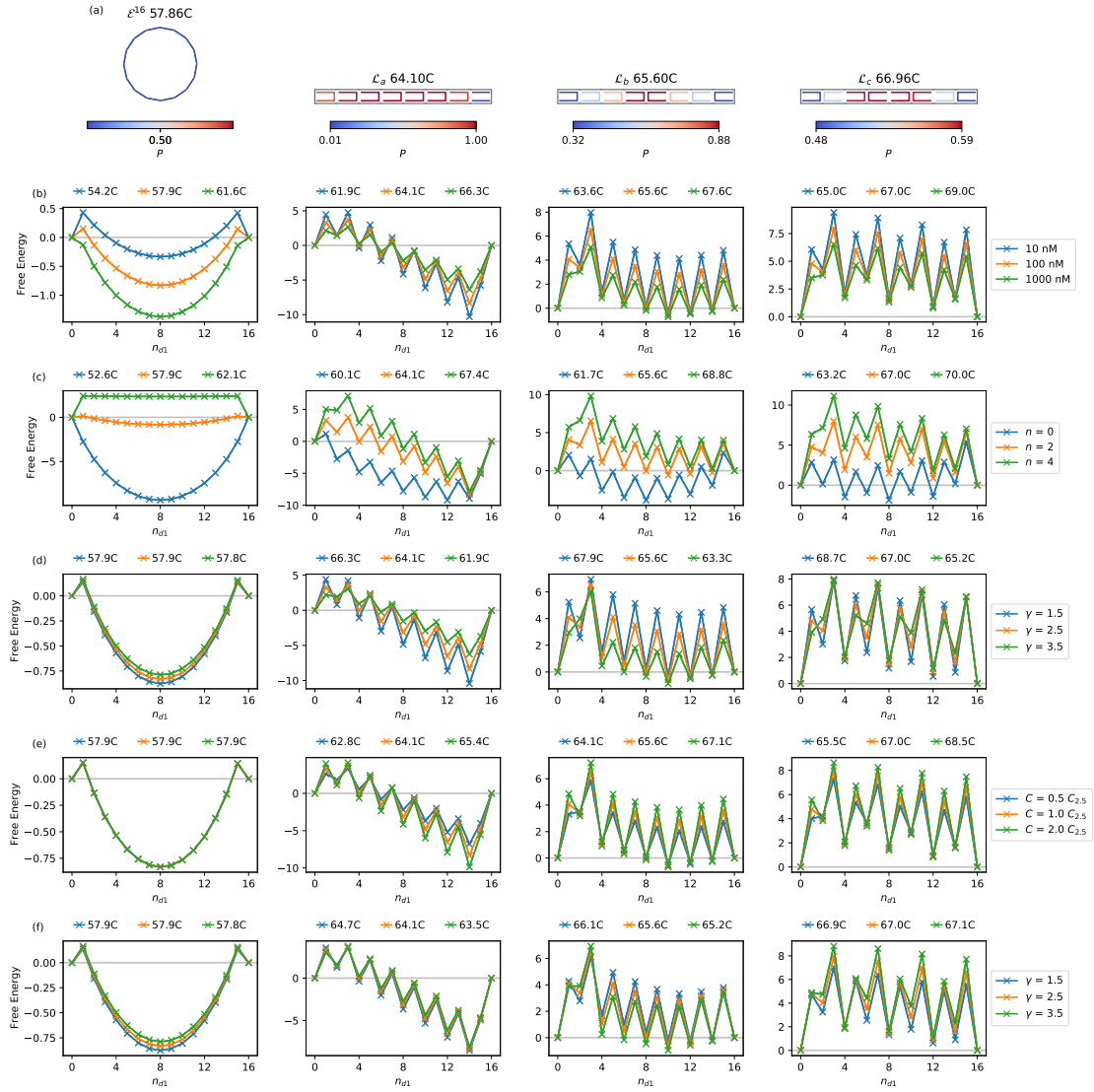


Figure 4.2: Effect of varying model parameters on free energy profiles of the sequence-averaged model at the melting temperature. (a) Default parameter set is used: staple concentration = 100nM, $n=2$, $\gamma=2.5$, $C = C_{2.5}$. (b) Staple concentration is varied. (c) Stacking strength n is varied. (d) γ is varied at fixed $C = C_{2.5}$. (e) C is varied at fixed $\gamma = 2.5$. (f) γ and C are varied, keeping the free energy of an 18 nt single-stranded loop constant.

Figure 4.3 shows slices of the most likely bound staples as the number of staples in the system is increased. Consider the \mathcal{L}_a system, with stacking turned off, and $n_{st} = 1$. The most likely bound staple is the rightmost two-domain staple, which is adjacent to the two single-domain staples at the edge. The reason this staple is more likely to be bound than the left-most staple is that it closes a loop of the same length, but the loop consists entirely of the single-stranded scaffold. For the edge staple on the left, the shortest loop consists of a duplex segment due to the binding of its first domain. This behaviour can also be seen in system \mathcal{L}_c , $n = 0$, $n_{st} = 1$, where the two near-edge staples are more likely to be bound than the edge staples. It is also in line with experimental data presented in Ref. [153]. In their study, they consider a small ‘origami’ consisting of a scaffold with four 16 bp domains with an inner and outer staple in a double-crossover configuration, similar to the two rightmost or leftmost staples in \mathcal{L}_c . By measuring the incorporation temperature of each staple in the presence or absence of the other staple, they showed that the inner staple is less likely to form in both cases, while binding both staples becomes more likely in the presence of the other. This can be explained by considering the half-bound states of each staple in the absence of the other staple. For the edge staple, the unbound arm of the staple and its complementary scaffold domain are on opposite sides of its bound arm. For the near-edge staple, they are on the same side, making binding of the near-edge staple easier in the absence of the edge staple than the binding of the edge staple in the absence of the near-edge staple. Once one of the staples is bound, the scaffold region is already prepared in a favourable condition, making the second staple much more stable. Both these effects are captured qualitatively by our model through the difference in loop costs, although the exact differences in binding energies depend on the modulation of γ and n .

In Figure 4.3, another interesting feature can be seen in system \mathcal{L}_b . When $\gamma \neq 0$, the most stable single staples ($n_{st} = 1$) are edge staples since they close the shortest loop. However, when pairs of staples are considered ($n_{st} > 2$), the central pair of staples close the shortest possible loop and are more likely to be bound. This cooperativity by double crossover structures was also observed in a

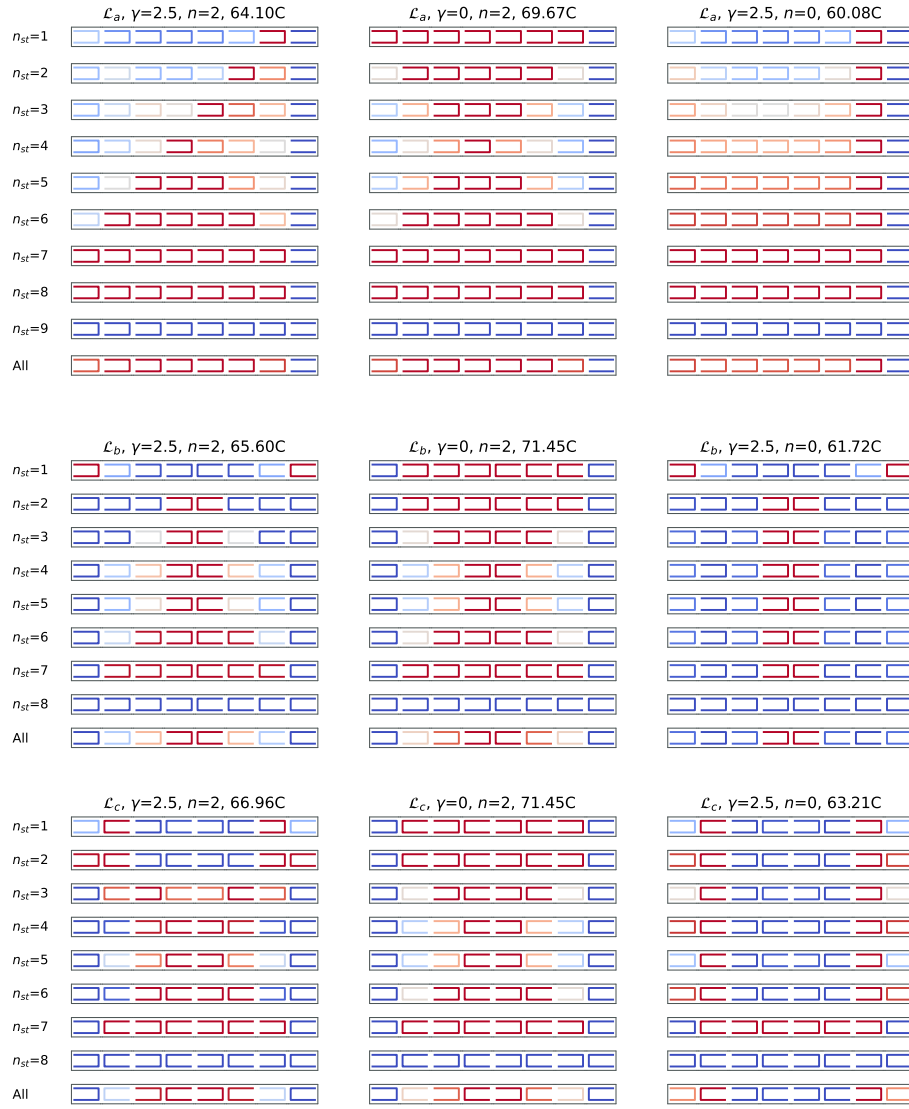


Figure 4.3: Cross-sections of staple binding probabilities. Probability that each staple is fully bound (blue is low and red is high) at cross-sections of the number of fully-bound staples n_{st} . At each cross-section, colours are normalised over microstates corresponding to the value of n_{st} ; colours in rows labelled by “All” are normalised over the whole range of n_{st} . In the first column, default n and γ values are used (the ‘All’ row for each system is the same as the one in Figure 4.2a). In the second column, the looping penalty is turned off. In the third column, coaxial stacking is tuned off.

study by Snodin *et al*[104], in which a small origami was simulated at nucleotide level of detail using oxDNA. This cooperativity is also seen in \mathcal{L}_b , but across all staples, since every staple can form a double crossover.

Next consider the $\gamma = 0$, $n = 2$ case. The result is the same for all systems (with a minor difference for \mathcal{L}_a due to single-domain staples) as loop-dependent interactions are turned off. Since edge staples lack one stacking neighbour compared to other staples, they are less likely to be bound. This in turn lowers the probability of coaxial stacking for near-edge staples, leading to a distribution that peaks at the centre of the origami. With $\gamma = 2.5$, the effect of introducing coaxial stacking is to increase the binding probability of central staples for systems \mathcal{L}_a and \mathcal{L}_c . For \mathcal{L}_b , the effect is more subtle since the central staples are already the most stable due to the double crossover in the centre.

4.2.2 Simulations under a Temperature Ramp

In the previous section, we looked at the consequences of certain design features and various model parameters on equilibrium properties. Here, I will explore the response of the systems as they are driven out of equilibrium under temperature ramps. The temperature is changed in steps and kept constant over a period of time corresponding to the average rate of cooling/heating. The total number of base pairs on the scaffold n_{bp} is used as a measure of the degree of folding during annealing and melting. To look at folding at the level of staples, I fit the binding probability of each staple as a function of temperature with a Hill function and extract T_{50} and n_H parameters as outlined in Appendix B. T_{50} is the first temperature at which the staple is bound/unbound with 50% probability. The Hill coefficient n_H characterises the sharpness of the transition.

For each system, I ran SSA simulations under various conditions, averaging each set of results over 150 trajectories. Figure 4.4 shows the simulation results. Using the default parameter set of the model, we can see the least hysteresis in \mathcal{L}_a and the most hysteresis in \mathcal{L}_c . This is consistent with the higher free energy barriers seen in the profile of \mathcal{L}_c . For the cooling rate considered, the time of incubation

at each temperature may not be sufficient to allow barrier crossing. This would introduce an asymmetry in the annealing/melting pathways that increases with barrier size. Staple T_{50} values are shown in Figure 4.4a. Staples in the centre of the origami tend to contribute the most to hysteresis due to stacking. In Appendix C.1, the relative equilibrium probabilities of binding for staples at various temperatures are shown for the systems considered here. The annealing distributions here tend to resemble the higher temperature equilibrium cases and the melting distributions resemble the lower temperature cases.

Figure 4.4b shows the effect of lowering the concentration from 100nM to 10nM (staples are still assumed to be in excess of the dilute scaffold strand). Annealing and melting temperatures are both reduced and hysteresis is increased at lower concentrations. This is expected from the observed increase in barrier heights (Figure 4.2a) since all bimolecular reactions become less favourable leading to a more rugged energy landscape. The temperature shift is drastically larger during annealing because the binding rate of the first domain of each staple is concentration-dependent. The larger shift in the annealing curve than the melting curve was also observed in experiments by Wei *et al* [56], with a more pronounced effect in 3D origami than in the 2D case.

The effect of faster cooling/heating is shown in Figure 4.4c. It results in a rise in melting temperature and a fall in annealing temperature, causing larger hysteresis. In general, one would expect this to happen when a system is driven out of equilibrium by fast temperature ramps since the system does not have enough time to respond to changes in temperature. In our rate model, binding rates are not temperature-dependent. However, the unbinding rate for all domains is extremely temperature-dependent. This generally results in a faster temperature response during melting since the system starts in a fully bound state.

To understand the features of annealing and melting curves better, it is instructive to compare T_{50}^A and T_{50}^M to T_M^{eq} from Figure 4.2. As an example, consider \mathcal{L}_b using the default parameter set. From Figure 4.2, we expect 10 domains to be bound at $T_M^{\text{eq}} = 65.6\text{C}$ (the definition of T_M^{eq} is not 50% binding probability).

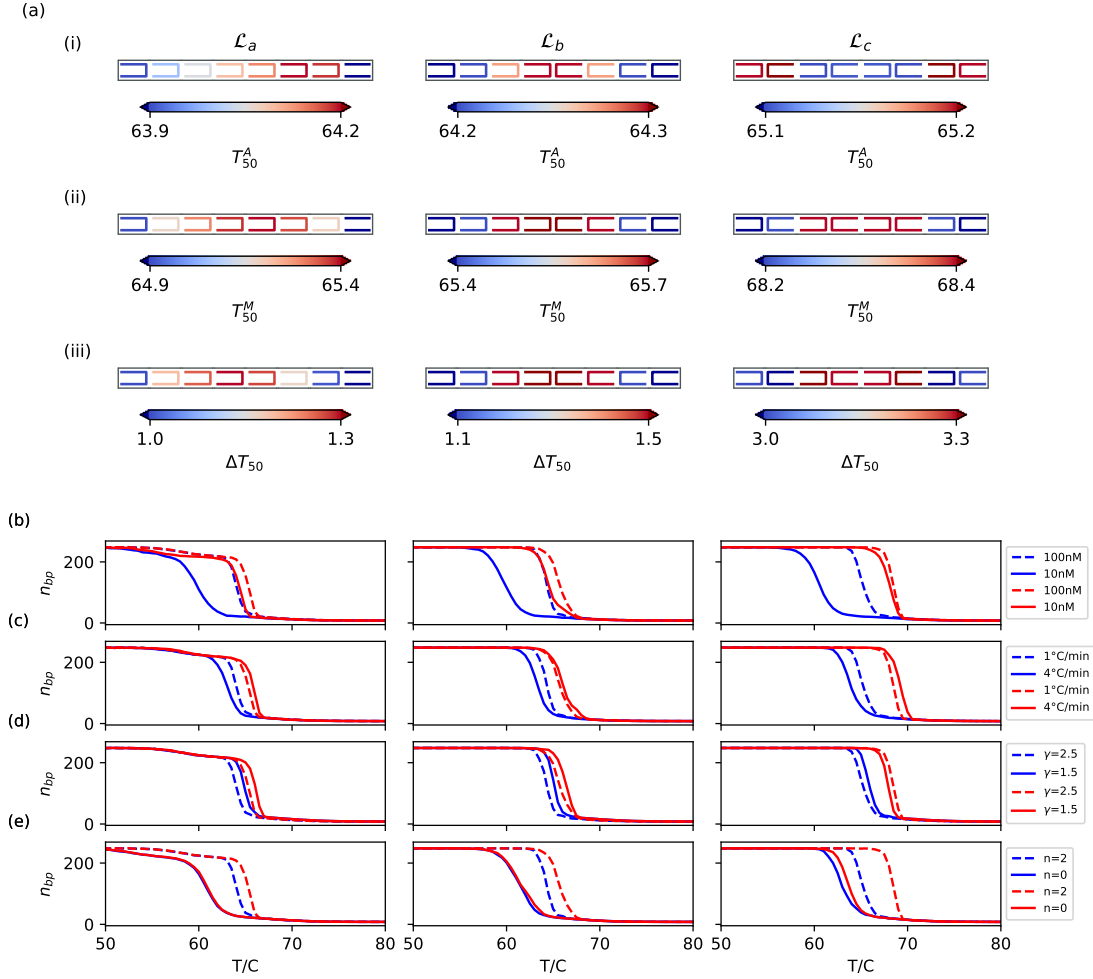


Figure 4.4: Two-row systems under temperature ramps. (a) Incorporation temperature of staples during annealing (i) and melting (ii). Hysteresis is shown in (iii). The default parameter set (100nM staple concentration, $\gamma=2.5$, $n=2$, cooling rate of $1^\circ\text{C}/\text{minute}$) is used in (a) and dashed curves in (b)-(e). One of the parameters is changed during annealing (blue) and melting (red) to plot solid lines.

In Figure 4.4 (dashed lines), we see that at this temperature, approximately 10 out of 16 domains are bound during melting, but during annealing, only two domains are bound. While a feature of our chosen rate model, there is experimental evidence to suggest this behaviour is valid. Sobczack et al [53] observed a similar process experimentally in 3D origami. They performed isothermal folding and melting experiments at T_{50}^A and T_{50}^M found during cooling and melting protocols for their systems. They found that unfolding occurred on the timescales close

to the incubation periods of their temperature ramps, but unfolding occurred at higher temperatures if the incubation time was increased, leading them to conclude that folding is the out-of-equilibrium process.

The overall effect of changing γ (Figure 4.4d) depends on the nature of the loops involved. For \mathcal{L}_a and \mathcal{L}_b , reducing γ pushes both T_{50}^A and T_{50}^M higher as the majority of loops become more stable. In the case of \mathcal{L}_c , we would expect both T_{50}^A and T_{50}^M to be lowered by reducing γ based on the equilibrium melting temperatures. However, T_{50}^A increases because the free energy barriers are generally lowered with decreasing γ , pushing the annealing curve towards equilibrium. This is not the case for \mathcal{L}_a and \mathcal{L}_b (see Figure 4.2f), where barriers are somewhat increased.

Reducing the stacking strength n (Figure 4.4e) to zero makes the transition much broader. In \mathcal{L}_a and \mathcal{L}_b , lack of coaxial stacking effectively eliminates hysteresis. This is expected from the low free energy barriers (Figure 4.2c). However, barriers associated with the formation of double crossovers are still significant in \mathcal{L}_c , leading to a small degree of hysteresis.

4.3 Real Origami

In the previous section, we looked at a set of small systems to examine how the model parameters and design features affect the thermodynamics and kinetics of DNA origami formation. We also looked at how the kinetics of origami relates to its equilibrium properties. Here, we will fix the model parameters to the default set ($n=2$, $\gamma=2.5$, $C_{2.5}=2.8 \text{ nm}^2$, $k_f=0.5 \times 10^6 \text{ M}^{-1}\text{s}^{-1}$), which gave the best agreement with experimental systems. The experimental systems we consider are the same as those in section 3.7.

Because these are experimentally viable designs, some of the domains are shorter by 1 bp than others in order to optimise crossover locations and relieve stress in the helices of the target structure. The standard domain sizes are 8, 16, 24, and 32 bps long. Shorter domains (coloured in red in Figure 4.5) are either 15 or 23 bps long. This introduces some asymmetry in the folding pathways presented in this section.

The results in this section are obtained using three different parametrisations of domain hybridisation energies:

- Sequence-specific ($\langle \text{pKD} \rangle$): each domain has a specific hybridisation energy depending on its position of the pKD scaffold (see Appendix A for the sequence).
- Sequence-averaged ($\langle \text{pKD} \rangle$): each domain has hybridisation energy corresponding to $\Delta G_{\text{NN}} = d \langle \Delta G_{\text{BP}} \rangle_{\text{pKD}}$, where d is the domain length and $\langle \Delta G_{\text{BP}} \rangle_{\text{pKD}}$ is the average hybridisation energy of each base pair in the sequence-specific case.
- Average-base ($\langle \text{ATCG} \rangle$): each domain has hybridisation energy corresponding to $\Delta G_{\text{NN}} = d \langle \Delta G_{\text{BP}} \rangle_{\text{NN}}$, where d is the domain length and $\langle \Delta G_{\text{BP}} \rangle_{\text{NN}}$ is the average hybridisation energy of all 16 combinations of bases in the nearest-neighbour model of Santalucia [142].

The two average parameterisations ($\langle \text{pKD} \rangle$ and $\langle \text{ATCG} \rangle$) essentially unmask some of the Physics that are washed out by sequence heterogeneity. The results are similar in both cases except for the slight change in melting temperatures. When comparing to the sequence-specific case, we expect the free energy of the target structure for $\langle \text{pKD} \rangle$ at any specific temperature to be the same as that of the target structure in the sequence-specific. This is not true for $\langle \text{ATCG} \rangle$. Therefore, $\langle \text{pKD} \rangle$ is a more suitable parametrisation. For the system $\mathcal{R}2$ and its seam variants, all results are available and I will present the $\langle \text{pKD} \rangle$ version. However, due to time constraints, I was only able to obtain $\langle \text{ATCG} \rangle$ results for the systems \mathcal{T} and $\mathcal{R}3$; simulations with the $\langle \text{pKD} \rangle$ parametrisation are ongoing for these systems.

Furthermore, it can take a lot of time to find optimum weights for umbrella sampling simulations since the procedure needs to be repeated for every system, temperature, and order parameter used. Therefore, I will present only some thermodynamic results in this section. To add to our analysis, I have calculated the exact thermodynamics of various small systems that have similar features to

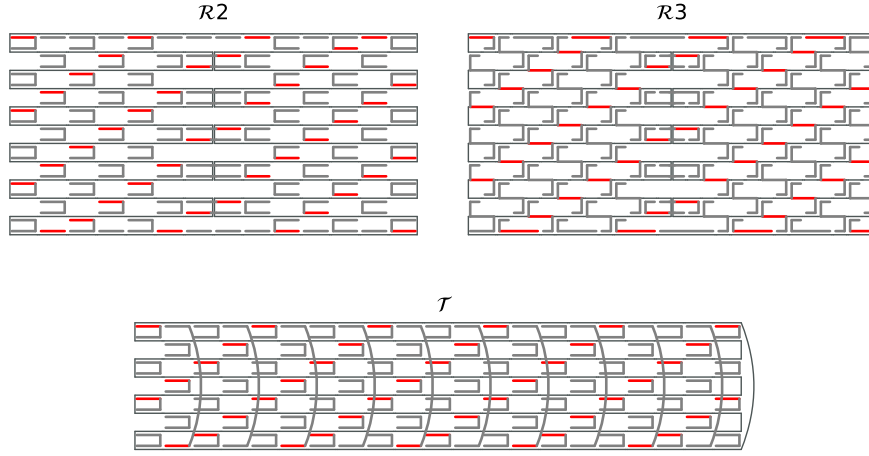


Figure 4.5: Position of shorter domains. Domains that are shorter (15 or 23 base pairs) than the default lengths (8, 16, 24, 32 bps) are coloured in red for the three experimental designs.

the real systems in this section. The results are presented in Appendix C.3 and are equivalent to those in the previous section.

4.3.1 System $\mathcal{R}2$

The $\mathcal{R}2$ origami consists of one- and two-domain staples. It was studied previously in Ref. [59], and used to demonstrate the possibility of guiding the folding pathway of origami by making small adjustments to the design. In particular, ‘seam’ staples that connect the two sides of the origami were shown to play a prominent role in determining the folding pathway. Here, I look at the thermodynamic and kinetic behaviour of the system to understand the role played by seam staples in this design.

Figure 4.6 shows the Hill function parameters (see Appendix B for fitting method) during annealing and melting at the level of individual staples. Single domain staples are the least stable with the lowest T_{50}^A and T_{50}^M . Their transition between bound and unbound states is also broader than two-domain staples, as evidenced by low Hill coefficients n_H^A and n_H^M . Two-domain staples show sharper transitions in general due to the cooperativity between their constituent domains.

The cooperativity between seam and seam-adjacent staples at each central turn of the scaffold gives rise to large hysteresis as seen in $\Delta T_{50} = T_{50}^M - T_{50}^A$ in Figure

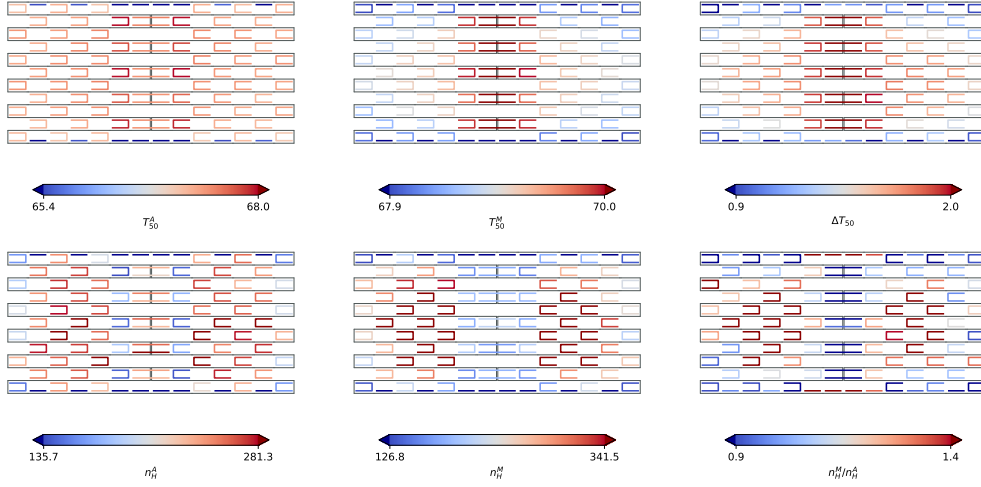


Figure 4.6: Incorporation temperature T_{50} and Hill coefficients n_H for individual staples during annealing and melting are shown for the system $\mathcal{R}2$ using sequence-averaged hybridisation energies $\langle \text{pKD} \rangle$. The observed hysteresis in each parameter is shown in the right column. For clarity, extreme values below or above one standard deviation of the distribution are coloured in dark red/blue.

4.6. Interestingly, the three pairs of central seam staples have high values of n_H^A but all seam staples have similar values of n_H^M . There is also some asymmetry between n_H^A values of seam-adjacent staples: the four staples that have 15 bp domains have higher n_H^A values. The Hill coefficient characterises the sharpness of the transition with respect to temperature. Therefore, staples that benefit most from the presence of other staples have higher n_H values. It is clear that the central pairs of seam staples benefit from the presence of the top and bottom pairs during annealing. During melting however, the pathway is not fully reversed but the steepness of their transition depends more on the length of the constituent domains: seam staples with 15 bp domains have higher n_H^M values.

Compared to other staples, seam-adjacent staples have the highest T_{50}^A but low to average values of n_H^A . When the system is annealed, these staples are the first ones to come in since they close the shortest scaffold loops. Staples at edges have a slightly higher loop cost than seam-adjacent staples due to the presence of a double-stranded segment containing their own first domain. It is therefore surprising

that T_{50}^A values of edge staples are similar to body staples. To understand this, I show the incorporation probability of individual staples at temperatures above and below the melting temperature in Figure 4.7. Edge staples and seam-adjacent staples do come in at higher temperatures, but they are not stable enough to reach 50% binding probability. As the temperature is lowered, the next structures to form with some probability are pairs of seam staples at the top and bottom of the origami. These pairs of staples form double crossovers, which are the shortest possible loops in the origami. However, the first staple of each pair closes a rather large loop so the coaxial stacking stability due to the presence of the seam-adjacent staples is required to aid its closure at these temperatures. The energy barrier associated with double crossover formation at the seams makes the transition less reversible than those of simple loop closure. Once pairs of seam staples are stable, folding proceeds from the centre of the origami outwards due to cooperativity by coaxial stacking. The incorporation temperature of staples is determined by their distance from the centre of the origami leading to the rhombus-shaped pattern in some of the figures (for example P_{67}^A or P_{69}^M).

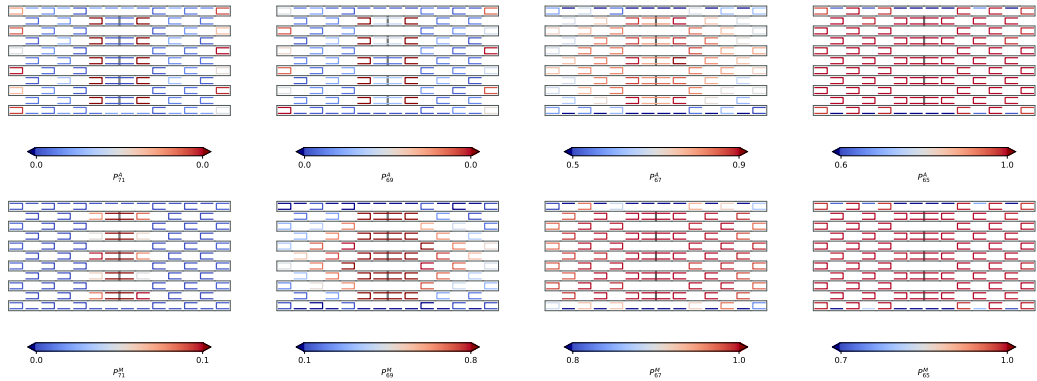


Figure 4.7: Incorporation probability P_T at different temperatures T during annealing (top row) and melting (bottom row) for the $\mathcal{R}2$ origami using the $\langle \text{pKD} \rangle$ parametrisation. T is decreased at a rate of 1°C per minute in 30s intervals from 80°C to 40°C during annealing and increased at the same rate during melting. Snapshots are shown at 71°C (right), 69°C , 67°C and 65°C (left).

To confirm this picture, I ran umbrella sampling simulations (see section 3.6) using the number of bound domains n_d as the biased order parameter. To sample

the whole space efficiently, I broke up the simulation in 18 sampling windows W_i , with $n_d \in [\min(W_i), \max(W_i)]$. The limits were designed such that there were 3 overlapping points between consecutive sampling windows: $\max(W_i) = \min(W_{i+1}) - 2$. Each window results in a partial free energy profile $F_{W_i}(n_d) = -k_B T \ln(p_{W_i}^{\text{eq}}) + c_{W_i}$ after unbiasing the distribution. The windows were then ‘patched up’ by fixing c_{W_i} such that $F_{W_1}(n_d = 0) = 0$ and the sum of square distances between overlapping regions of $F_{W_i}(n_d)$ and $F_{W_{i+1}}(n_d)$ was minimised, allowing sequential reconstruction of $F(n_d)$ over the whole region.

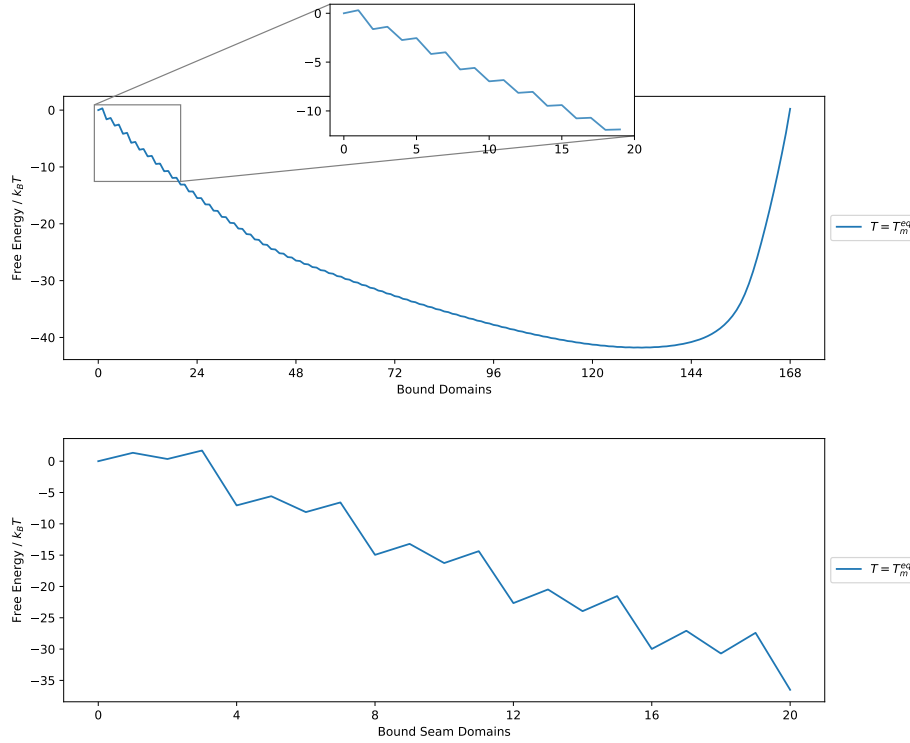


Figure 4.8: Free energy profiles as a function of the number of bound domains and the number of bound seam domains for system $\mathcal{R}2$ at its $T_M^{\text{eq}} = 68.1^\circ\text{C}$. Results are shown for the $\langle \text{pKD} \rangle$ parametrisation.

The free energy profile at 68.1°C is shown in Figure 4.8, using either the total number of bound domains or the number of bound seam domains as an order parameter. At this temperature, there is a sawtooth-like decrease in free energy at low n_d , followed by a change of slope for $40 < n_d < 140$, and an increase in

free energy in the region $n_d > 140$. The initial drop in free energy is due to the binding of seam-adjacent and seam staples, which together add up to 40 domains. Other staples in the origami are less stable at this temperature, leading to the change in the slope of the free energy profile past 40 domains. The 16 single-domain staples are especially unstable, which causes the curve to change steepness towards the end. The free energy profile as a function of seam domains confirms this hypothesis and also explains why the seam staples exhibit high hysteresis. While there is no global barrier, the formation of double crossover structures does require crossing local barriers.

We can also understand the role of seam staples by cutting them in half so that they cannot make long-range connections or form double crossovers. In Figure 4.9, we look at the consequences of these changes on the free energy profile and on hysteresis. In the half-seam variant $\mathcal{R}2_H$, one of each pair of seam staples is cut in half (replaced by two single domain staples). In the no-seam variant $\mathcal{R}2_N$, all seam staples are cut in half. Cutting one of every pair of seams ($\mathcal{R}2 \rightarrow \mathcal{R}2_H$) drastically changes the hysteresis profile of the remaining seam staples. More detailed staple data for each variant (Appendix C.4) also shows that staples close to the seams are affected much more than others. However, the change $\mathcal{R}2_H \rightarrow \mathcal{R}2_N$ does not alter the behaviour of seam-adjacent staples significantly.

The free energy profiles also confirm the strong local cooperativity between seam-adjacent and seam staples and a lack of long-range cooperativity. The change in the free energy of the target structures can be attributed to both changes in loop lengths and to the reduced amount of coaxial stacking between scaffold duplexes at the seams. Both changes to seam structure result in the same reduction in coaxial stacking (one for each staple cut), but the change $\mathcal{R}2 \rightarrow \mathcal{R}2_H$ causes a more dramatic change because ultra-short double crossover loops can no longer form. For $\mathcal{R}2$, the most stable state is when all seam and seam-adjacent domains are bound. The profile for $\mathcal{R}2_H$ deviates from $\mathcal{R}2$ after ~ 4 domains, which are likely the top and bottom seam-adjacent staples since they form short loops and do not have 15 bp domains. The profiles behave similarly in the middle region

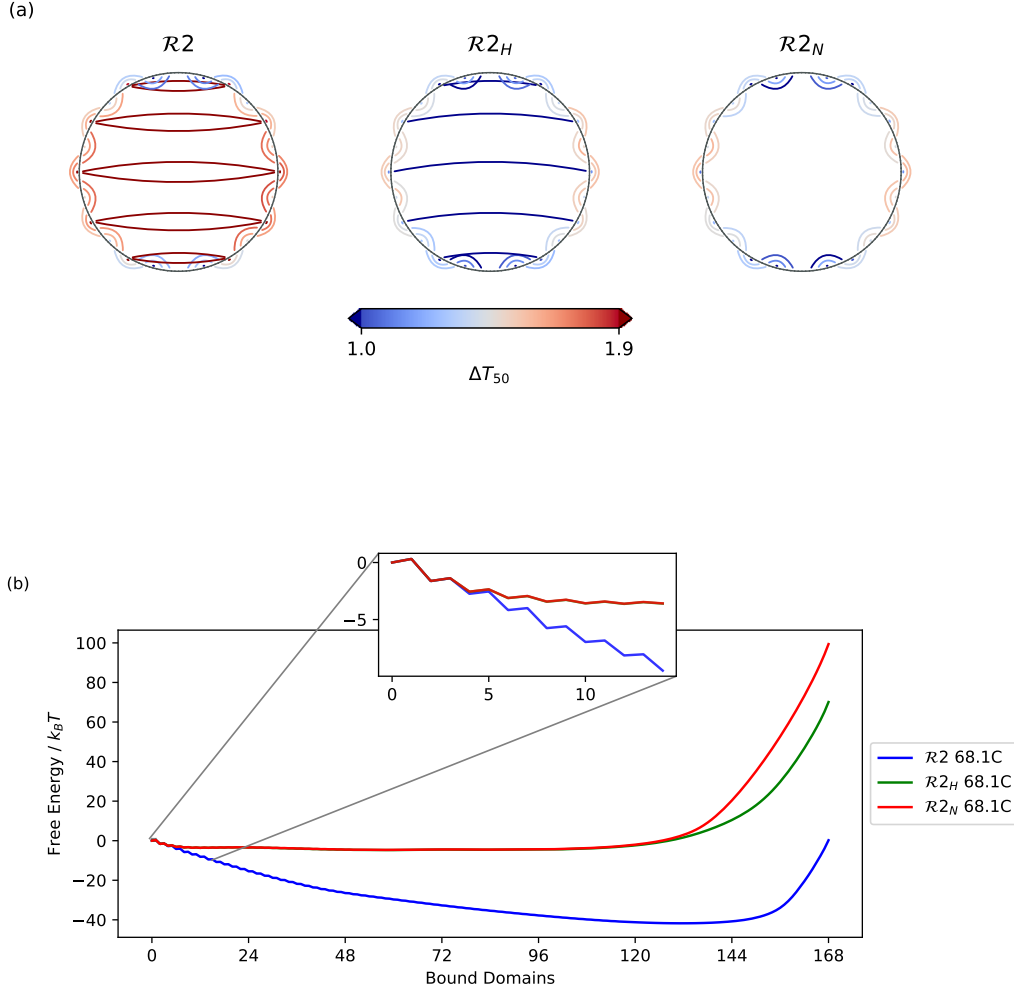


Figure 4.9: Results are shown for the $\langle \text{pKD} \rangle$ parametrisation. (a) Hysteresis $\Delta T_{50} = T_{50}^M - T_{50}^A$ seen in crossovers for seam-variant of $\mathcal{R}2$ plotted on a circular layout of the scaffold strand. (b) Free Energy profiles of seam variants of $\mathcal{R}2$ at 68.1°C.

because there is little change to loop closure for body staples. For body staples, the presence of seam staples has little effect on the shortest loop that they close, which contributes the most to the change in shape energy.

While the use of the sequence-averaged parametrisation of the model helps highlight the physics of the assembly, it turns out that many of the features are washed out, or masked, by the heterogeneity of the hybridisation energies when sequence dependence is taken into account. Figure 4.10 shows the results of umbrella

sampling simulations at various temperatures around the melting temperature of the origami. Taking into account the sequence dependence increases the depth of the free energy profile by introducing heterogeneity into domain hybridisation energies while shifting the minimum towards states with CG-rich domains present.

Figure 4.10b,c show 2D free energy profiles in the average and sequence-specific parametrisation. In each case, the number of bound seam domains n_s was used as the biased order parameter. This results in a partial sampling of regions of state space over the second order parameter n_d . This imperfect sampling reveals important features of the free energy profile since states with the lowest free energy are still likely to be sampled². In the sequence-average case, we can see sharp drops in free energy at steps of 4 seam domains. In the sequence-specific case, the barriers with respect to seam domains are preserved to a degree. Looking at ΔT_{50} values in melting and annealing simulations (Appendix C.5), seam staples tend to have the most hysteresis even in the sequence-specific case. This can be attributed to the roughness of the landscape as seam staples bind/unbind. Since this feature is somewhat preserved in the sequence-specific case, it may explain the lower hysteresis seen in experiments for half-seam and no-seam variants.

4.3.2 System $\mathcal{R}3$

The $\mathcal{R}3$ system folds into a planar rectangle in the same aspect ratio as $\mathcal{R}2$ but with different scaffold connectivity. The design consists of one-, two- and three-domain staples, with double crossovers spread out across the whole structure. Experimental data suggest that $\mathcal{R}3$ folds at a much lower temperature and shows greater hysteresis than $\mathcal{R}2$ using either of the two scaffold strands (see section 3.7).

Figure 4.11 shows T_{50} and n_H values for staples during annealing and melting. The top and bottom duplexes of the origami contain staples with long domains. There is some asymmetry in the placement of the long 32 bp single-domain staples at the top and bottom duplexes (these are not shown in Figure 4.11 because their extreme values distort the distribution). These staples anneal and melt at higher

²It is possible to sample the whole space by applying a bias to both order parameters but I have not done so due to time-constraints.

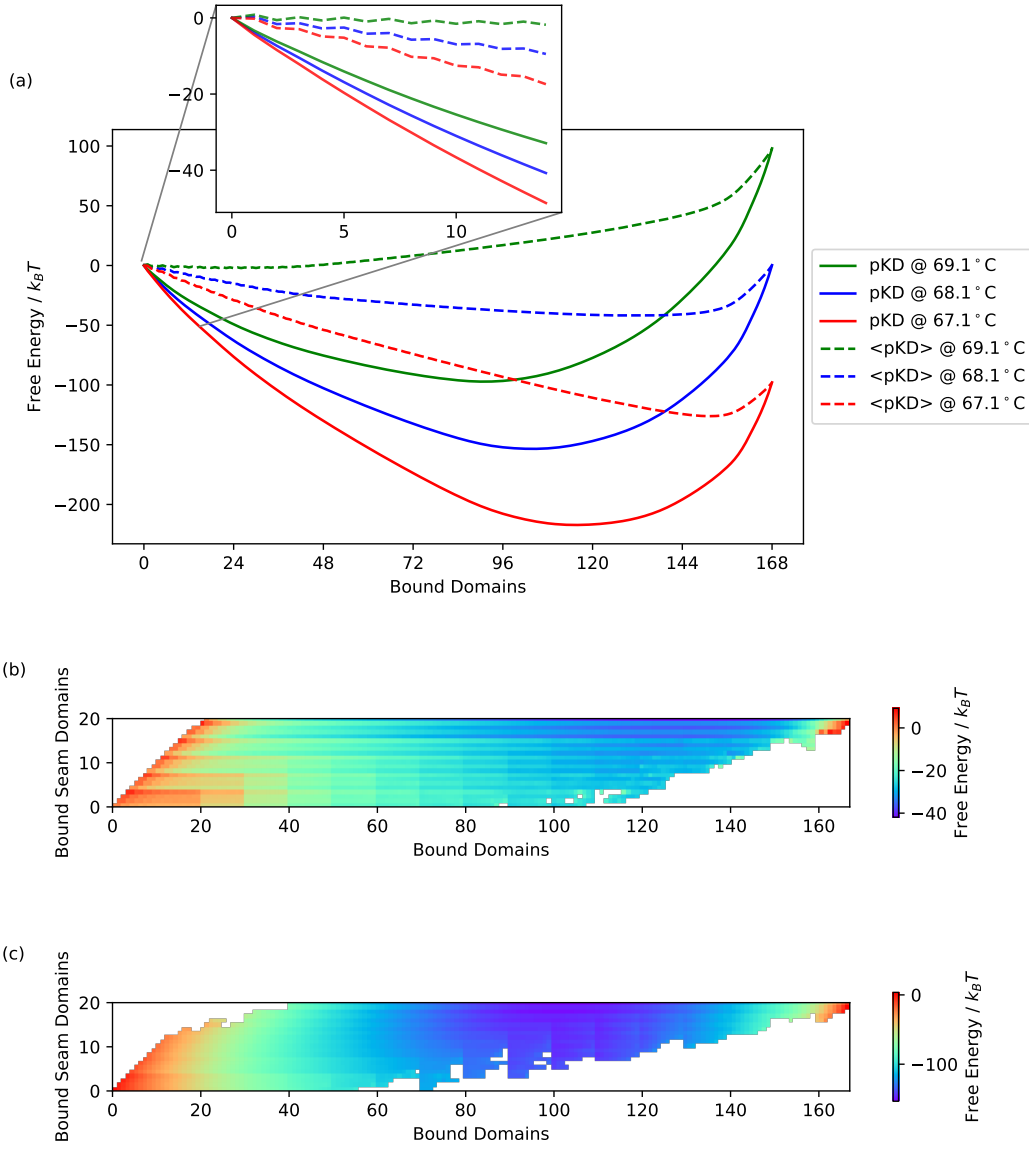


Figure 4.10: (a) Free energy profiles as a function of the number of bound domains for system $\mathcal{R}2$ at various temperatures using sequence-averaged ($\langle \text{pKD} \rangle$) and sequence-specific (pkD) parametrisation of the model. (b) 2D profile using the $\langle \text{pKD} \rangle$ parametrisation at 68.1°C. (c) 2D profile using sequence-specific parametrisation at 68.1°C.

temperatures due to the stability provided by their binding energies. There is also some asymmetry at the corners of the origami: there are two-domain staples at the top left and bottom right corners and three-domain staples at the other two corners. The three-domain staple at the bottom left and top right corners close one short loop by binding two 16 bp domains, and a longer loop by binding one 16 bp and one

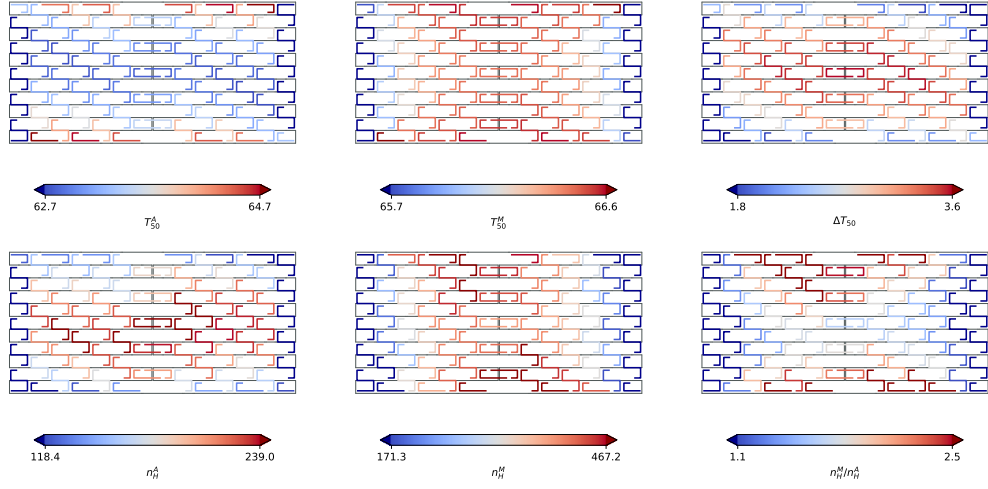


Figure 4.11: Incorporation temperature T_{50} and Hill coefficients n_H for individual staples during annealing and melting are shown for the system $\mathcal{R}3$ using average-base $\langle \text{ATCG} \rangle$ hybridisation energies. The observed hysteresis in each parameter is shown in the right column. For clarity, the long 32 bp domains at the top and bottom were not included in the colouring set; extreme values below or above one standard deviation of the remaining distribution are coloured in dark red/blue.

8 bp domain. These asymmetries give rise to the pattern seen in the folding pathway during annealing, where folding begins at the top and bottom helices but shifts towards the more stable corners. This pattern can also be seen clearly in Figure 4.12.

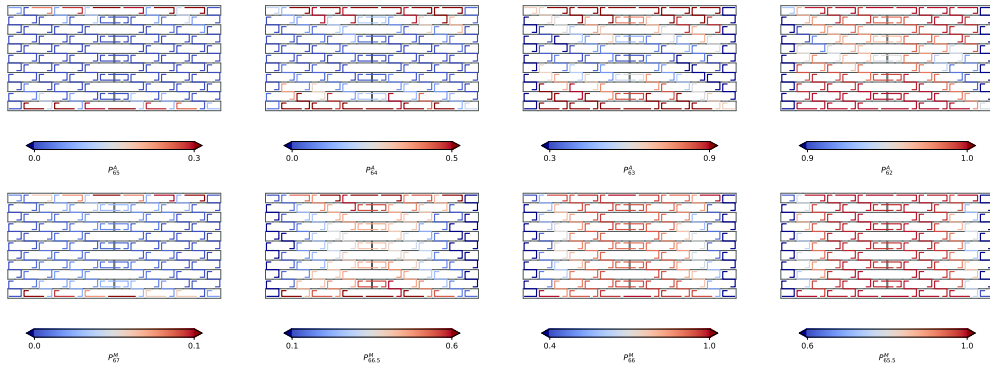


Figure 4.12: Incorporation probability P_T at different temperatures T during annealing (top row) and melting (bottom row) of $\mathcal{R}3$ using the $\langle \text{ATCG} \rangle$ parametrisation. T is decreased at a rate of 1°C per minute in 30s intervals from 80°C to 40°C during annealing and increased at the same rate during melting. Snapshots are shown at 65°C , 64°C , 63°C and 62°C for annealing and at 67°C , 66.5°C , 66°C and 65.5°C for melting.

Apart from the above exceptions, all other staples have one 15-16 bp middle domain and two 8 bp domains; their crossovers close one long and one short loop in an otherwise empty origami. Body staples cooperate to form two double crossovers, edge staples can only form one, and seam staples can form one double crossover with an adjacent staple, and one scaffold double crossover with another seam staple. It is clear from Figure 4.12 that folding propagates diagonally through the origami by forming double crossovers until an edge is reached. It begins from the top and bottom duplexes with the binding of two-domain staples with long arms. These are able to form double crossovers with three domain staples in diagonal top-left to bottom-right patterns. There are 11 diagonal propagation channels, most clearly seen in P_{63}^A , with 5 starting at the bottom, 5 starting at the top, and a symmetric long diagonal from top-left to bottom-right corners. It is also clear that seam staples do not hinder this propagation. We can also see this pattern in the hysteresis (ΔT_{50} in Figure 4.11) values. Staples that are furthest from their 'seed' at the top/bottom duplexes are the most hysteretic.

Unfolding occurs through a different pathway to folding. The least stable structures in the system are crossovers at the edges of the origami. Once edge staples are unbound, their neighbouring staples become destabilised by the lack of a double crossover on one side. Unfolding propagates diagonally from all edge staples towards the centre, meeting resistance from the highly stable domains at the top and bottom rows. This leads to the X-shaped pattern seen in the pathway.

Because staples have arms of different lengths, one might expect that individual staple domains may become stable at different temperatures. However, as shown in Appendix C.6, the same folding/unfolding pattern is seen in binding probabilities of individual domains, and staples are rarely part-bound during annealing and melting. It should also be noted that staple-blocking events are very rare. In the simulations shown here, blocking events occurred at a rate of approximately 1 per 10^5 steps and were corrected within < 10 steps. An average number of steps per simulation was 10^7 for the $\mathcal{R}3$ system, with the majority of transitions occurring near the melting temperature.

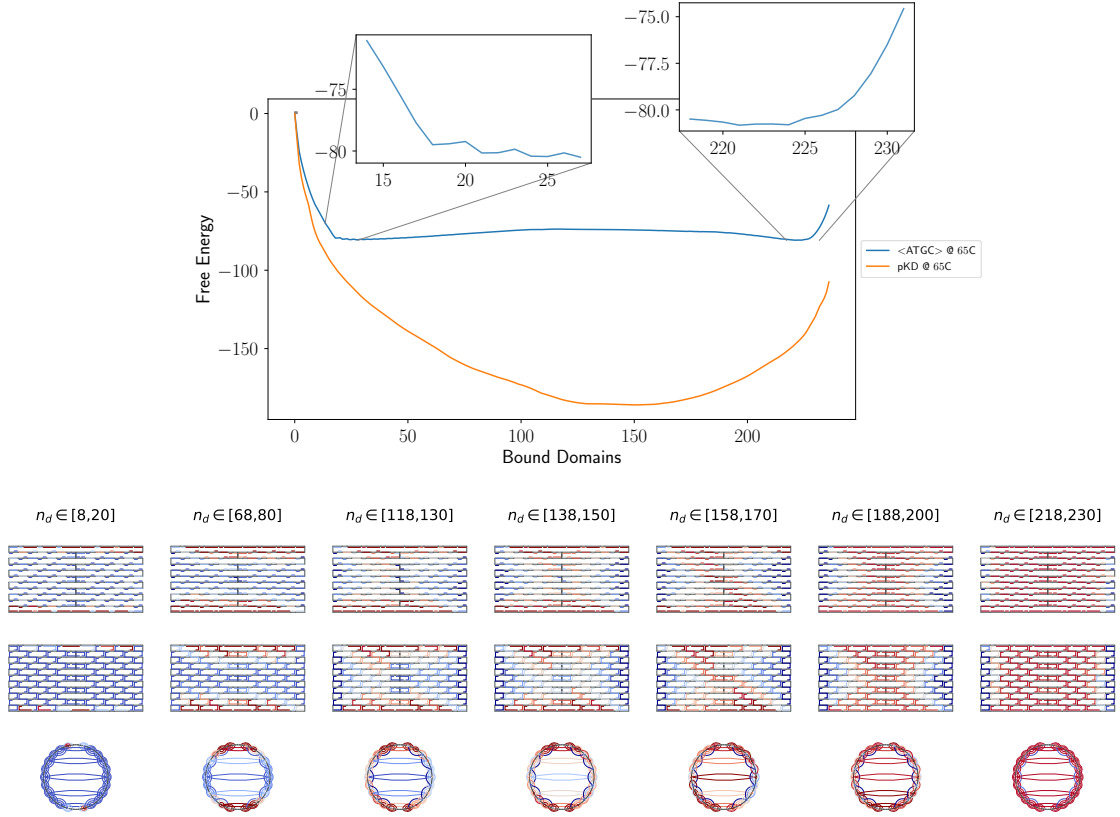


Figure 4.13: Free energy profile as a function of the number of bound domains for system $\mathcal{R}3$ at 65°C in the sequence-specific (pKD) and average-base ($\langle\text{ATGC}\rangle$) parametrisations. The free energy of the target structure is not the same in the two cases because of the different amounts of CG-content in the two parametrisations. Normalised probability distributions of domains/staples/crossovers being bound at selected umbrella sampling windows are also shown for the $\langle\text{ATGC}\rangle$ case. The top panel shows which domains are most likely to be bound. The middle panel shows which staples are most likely to be *fully* bound. The bottom panel shows which crossovers are most likely to be bound. Blue and red correspond to low and high probability respectively.

Figure 4.13 shows the free energy profile of $\mathcal{R}3$ at 65°C as a function of the number of domains. The average-base parametrisation profile shows an initial drop in free energy followed by a long and gentle energy barrier and a steep rise in free energy at the end of the profile. Each of the lower panels shows the relative likelihood of individual objects being fully bound in selected umbrella sampling windows. In each window, the simulation is forced to visit states within a certain range of bound domains n_d with equal probability. The initial sharp drop in free energy is clearly due to the high-stability domains at the top and bottom of the origami. The rise at the

end of the free energy profile ($n_d \in [226, 236]$) is caused by the unstable edge staples. There are a total of 10 such staples, each forming two crossovers, one of which is less stable than the other, causing one of the staple domains to be particularly unstable. This explains the sharp rise in free energy with the 10 final domains.

It is more difficult to attribute a single factor to the energy barrier in the intermediate region $n_d \in [26, 220]$. Generally, there is a higher likelihood of finding part-bound staples in these simulations compared to simulations under a temperature ramp. This may be a contributing factor to the energy barrier since many of the shorter staple domains can only bind cooperatively through the formation of double crossovers with surrounding staples. Seam staples also appear to contribute to this barrier. Particularly, the long diagonal crossing the central seams is likely to be bound at $n_d \sim 150$, which coincides with the change in the slope of the free energy.

The existence of this barrier was also confirmed by unbiased isothermal simulations started at different initial states. Simulations that were started with an empty scaffold quickly (within 10s) reached $n_d \sim 20$ but did not fold to completion within an hour. Simulations started from the target structure oscillated around $n_d \sim 220$ within the same timescales. Simulations started with n_d along various states along the barrier either ended in one or the other of these local minima. The existence of the two local minima is consistent with the high hysteresis seen in melting and annealing simulations. As shown in Figure 4.12, at 65°C, neither folding nor annealing have begun under the temperature ramp of 1°C per minute.

The sequence-specific energy profile does not show a similar barrier. However, larger hysteresis is seen in $\mathcal{R}3$ than $\mathcal{R}2$ in both simulation and experimental results. The individual staple data (Appendix C.5) shows no identifiable pattern so it is difficult to attribute the higher hysteresis in $\mathcal{R}3$ to the discussion of the folding pathway above. The main source of hysteresis in both systems is the formation of double crossovers. Since a majority of the staples in $\mathcal{R}3$ cooperate by forming double crossovers, it is not surprising to see higher hysteresis.

4.3.3 System \mathcal{T}

The \mathcal{T} design folds into a 3D 8-helix tube and consists entirely of two-domain staples. Experimental data suggest that \mathcal{T} folds at a similar temperature (1°C lower) and shows greater hysteresis than $\mathcal{R}2$ but less hysteresis than $\mathcal{R}3$. Since \mathcal{T} is a 3D system, the graph representation of its states does not have a planar embedding and I use the local model for simulations in this section.

Figure 4.14 characterises the thermodynamic behaviour of the system at 67°C. In an otherwise empty origami, the most stable staples are on the right side, forming 32 nt bulge loops. The next most stable staples are those on the left edge, which close loops containing 16 bp double-stranded segments and 16 nt single-stranded segments. Since they are at the edge, the four staples on the left lack one coaxial stacking interaction compared to the four on the right. At equilibrium, the binding of these eight staples is favoured when the total number of domains present is low, but binding of the right side is favoured once more staples are present. We can identify an energy barrier with the formation of a nucleus that propagates from right to left. There are also some diagonal features in the stability distributions of the staples associated with shorter 15 bp domains.

Figures 4.15 and 4.16 show the folding pathway of the system under temperature ramps. Melting follows a similar pathway to the equilibrium distributions, beginning from the less stable left side. During annealing, however, both sides form at similar times, although propagation from the left side is faster. Hysteresis ranges from 1.9°C to 3.1°C (lower values fall below one std and are coloured dark blue) and is concentrated in the centre of the origami. Sequence-specific data (Appendix C.5) shows that a similar degree of hysteresis remains in the system to that of $\mathcal{R}2$ but less than $\mathcal{R}3$. In the case of $\mathcal{R}3$, the folding pathway is dominated by the formation of double crossovers and hysteresis arises because of the cooperativity due to this mechanism. There are no double crossovers in \mathcal{T} and cooperativity mainly arises from coaxial stacking interactions, which leads to the central staples being most hysteretic. The mechanism of cooperativity in $\mathcal{R}2$ is a mixture of these two effects. While there are fewer coaxial stacking interactions in $\mathcal{R}2$ than

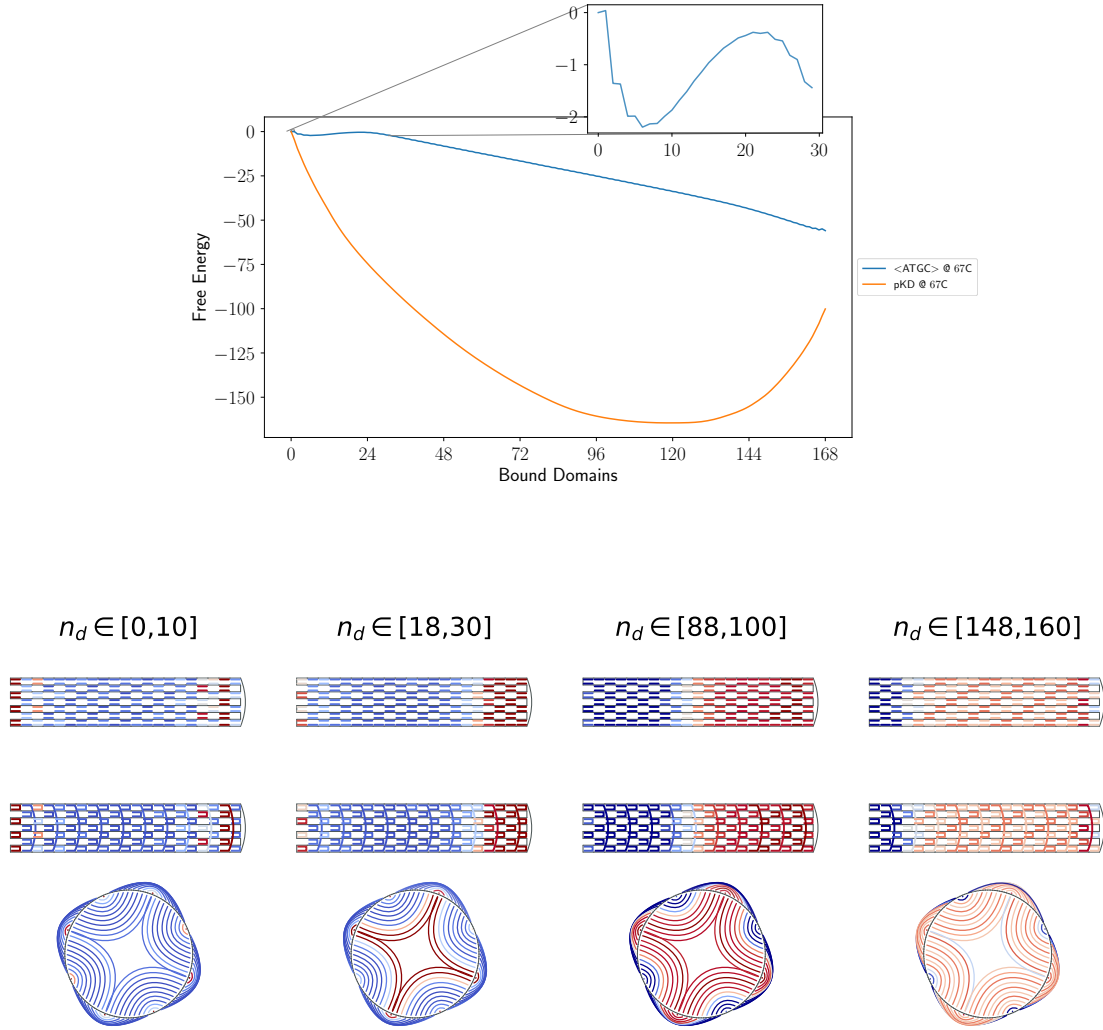


Figure 4.14: Free energy profile as a function of the number of bound domains for system \mathcal{T} at 67°C in the sequence-specific (pKD) and average-base ($\langle\text{ATCG}\rangle$) parametrisations. The free energy of the target structure is not the same in the two cases because of the different amounts of CG-content in the two parametrisations. Normalised probability distributions of domains/staples/crossovers being bound at selected umbrella sampling windows are also shown for the $\langle\text{ATCG}\rangle$ case. The top panel shows which domains are most likely to be bound. The middle panel shows which staples are most likely to be *fully* bound. The bottom panel shows which crossovers are most likely to be bound. Blue and red correspond to low and high probability respectively.

the \mathcal{T} , this is offset by the existence of double crossovers at the seams, leading to a similar degree of overall hysteresis.

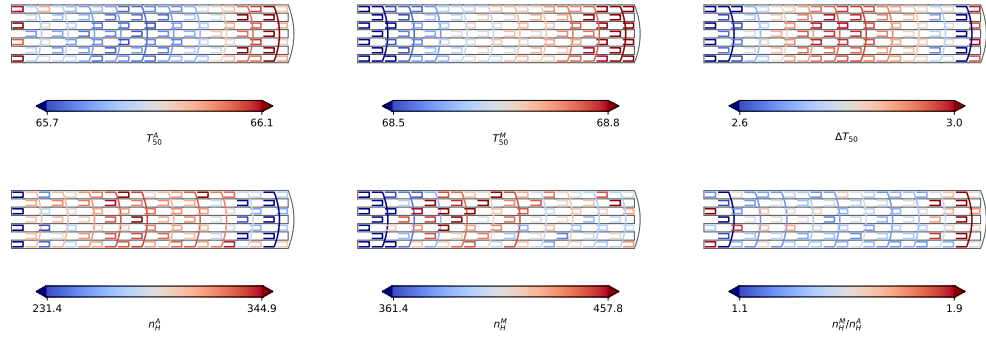


Figure 4.15: Incorporation temperature T_{50} and Hill coefficients n_H for individual staples during annealing and melting are shown for the system \mathcal{T} using average-base hybridisation energies $\langle \text{ATCG} \rangle$. The observed hysteresis in each parameter is shown in the right column. Extreme values below or above one standard deviation of the remaining distribution are coloured in dark red/blue.

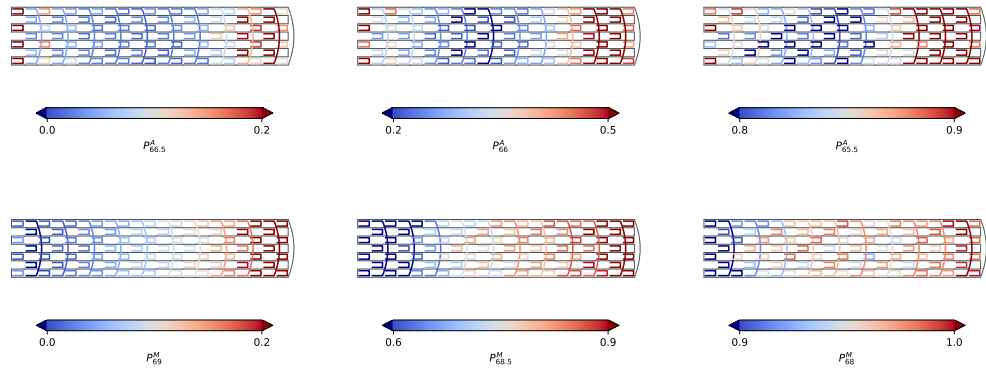


Figure 4.16: Incorporation probability P_T at different temperatures T during annealing (top row) and melting (bottom row) in the average-base ($\langle \text{ATCG} \rangle$) parametrisation. T is decreased at a rate of 1°C per minute in 30s intervals from 80°C to 40°C during annealing and increased at the same rate during melting. Snapshots are shown at 66.5°C , 66°C , and 65.5°C for annealing and at 69°C , 68.5°C , 68°C for melting.

4.4 Summary and Outlook

We studied the behaviour of the domain-level model using several toy systems, highlighting the role of loop closure in determining the folding pathway. Particularly, we found double crossovers to be particularly important in determining the thermodynamics of the systems. They introduce free energy barriers that lead to a more rugged landscape, which slows down folding dynamics. The barrier arises because cooperativity between four domains is needed to form these structures. Once bound, the four cooperating domains close extremely short loops, making them unlikely to unbind. This is consistent with detailed oxDNA simulations [104] of the assembly pathway of small origami. Simulations of these systems under temperature ramps confirm the picture. The more double crossovers are introduced, the more rugged the free energy landscape, the slower the dynamics, which ultimately leads to a higher degree of hysteresis between the folding and unfolding transitions.

We also investigated several real origami systems for which we have experimental data available for comparison. We characterised the free energy profiles of these systems, which show design-specific features. Compared to the smaller toy systems, the higher degeneracy of the macrostates defined by the number of domains makes it more difficult to associate design elements with the features of the free energy landscape. Other order parameters that focus on specific design features (such as bound seam domains in $\mathcal{R}2$) were found to be more appropriate, corroborating the principles revealed by the toy systems. Further, sequence-specificity often masks important physics of the process, making it necessary to use the average-sequence parametrisation of the model to reveal the underlying principles of folding.

We simulated the effect of temperature ramps on these systems and connected the asymmetry in the folding and unfolding transitions to our thermodynamic data. Our simulation results are consistent with the experimental data on the change in hysteresis with each design. Particularly, the $\mathcal{R}3$ design that utilises different domain lengths and many double crossovers shows a greater degree of hysteresis. By resolving the transition at the level of individual staples, we also mapped the folding pathway for these systems. Consistent with experimental

studies [56], we found that unfolding nearly follows the reverse pathway as that of folding. The asymmetry in the two pathways is increased in the $\mathcal{R}3$ design but the sequence-specific case remains largely symmetric.

The model provides an appropriate framework to study origami folding. Its simplicity has allowed us to study relatively large origami systems compared to other studies that include a more detailed geometric representation of DNA [104], [152]. The simple treatment of stacking could be improved in the future. The current value of $n = 2$ required for agreement with experimental data is probably compensating for neglected effects. For example, when a staple is partly-bound to an otherwise well-folded scaffold, the single-stranded scaffold region is held in place by other staples, making the rebinding of the unbound domain more likely. This is not fully captured by our treatment of scaffold loops and the high coaxial stacking may be making up for such effects. The simple treatment of loop costs could also be improved in the future, although it is difficult to see how this could be done analytically. One approach could be to use a more detailed model such as oxDNA to parametrise the cost for specific types of loops; such an approach was taken by Fonseca *et al* [70] to simulate DNA brick assembly.

Further improvements to the model could focus on allowing linear (not circular) scaffolds and generalising the model to multi-domain staples. This would allow comparison to a wider range of experimental data such as that of Ref. [57]. The next step could be incorporating a description of misbinding due to partial sequence complementarity. The inclusion of this type of misbinding would allow us to study kinetically trapped structures that are prevalent in the isothermal assembly of origami at low temperatures. This is discussed further in the next chapter.

5

Nucleated Assembly at Low Temperature

Contents

5.1	Introduction	122
5.2	Preliminary Experiments	122
5.3	Simulations at Low Temperature	128
5.3.1	Non-Coated Origami	128
5.3.2	Coated Origami	133
5.4	Nucleation	138
5.5	Propagation	143
5.6	Controlling The Folding Pathway	145
5.7	Summary and Outlook	150

In this section, I apply the model to an origami system in which the scaffold strand is coated with short oligonucleotides. A set of staple strands is used to isothermally displace the coating staples resulting in a well-folded origami. The experiments in this section were done by Katherine Young in the Turberfield group and more detailed experimental results can be found in her thesis [127]. I will present selected results from Katherine's work to accompany simulation results.

5.1 Introduction

The overall aim of this chapter is to design a system where two origamis with distinct shapes can successfully fold the same scaffold strand in a single mixture. We took inspiration from the competitive annealing experiments of Majikes *et al*[154], where two staple sets were used to produce a distribution of chimeric structures that inherit parts of both designs. In order to assemble a single well-formed (rather than chimeric) product in a competitive scenario, we envisaged engineering a free energy barrier that could be manipulated to favour one target over another. One way to achieve such an energy barrier is by covering the scaffold strand with a set of short oligos. Folding on this ‘coated’ scaffold can then be initiated in a particular region by strategically omitting some of the coating strands. If the nucleation sites of the two competing staple sets are distinct, the coated scaffold can be predisposed towards one shape by omitting a few coating strands at that site. It is necessary that such competitive folding occurs under isothermal conditions at relatively low temperatures in order to avoid dissociation of the coating oligos. Therefore, we began studying the isothermal folding of coated scaffolds as a first step towards achieving this aim. In the context of this chapter, one set of folding staples competes with one set of coating staples. We leave including a third set to further work.

5.2 Preliminary Experiments

In Figure 5.1, I present the system of competing staple sets that is studied in this chapter. In experiments, the scaffold strand (pkD sequence in Appendix A) was first thermally annealed with a set of 16 nt single-domain staples, resulting in the \mathcal{E} plasmid-like structure from the previous chapter. I will refer to these staples as “coating” staples (they are called “helper” staples in Katherine’s thesis). The folding staples ($\mathcal{R}2$ from the previous chapter) were then added to the mixture and incubated at constant temperature for a set period of time. The domains of the coating set are shifted along the scaffold by 6 nt with respect to those of the folding set. This results in the creation of 6 nt and 10 nt toeholds when one staple is missing

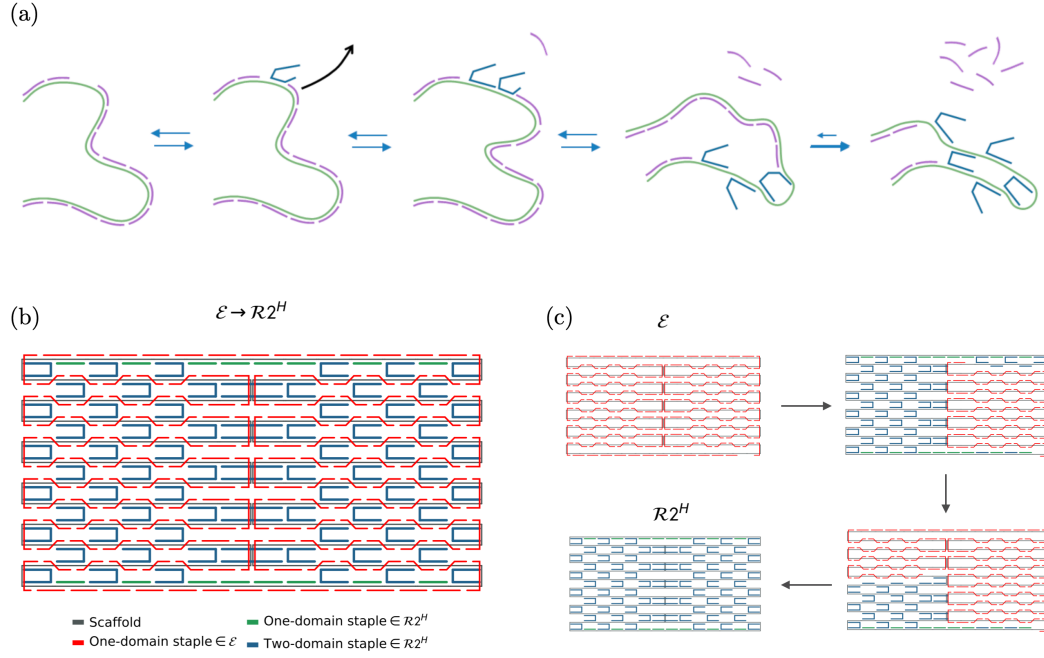


Figure 5.1: The coated origami system. (a) A scaffold strand (green) is coated with short single-domain staples (pink). Through toe-hold mediated strand displacement, the blue staples displace the pink staples. (b) Both sets of staples, one folding into archetype \mathcal{E} , and the other into a variation of $\mathcal{R}2$ are drawn adjacent to their binding sites on the scaffold. A vacant 16 nt domain creates a 6 nt or 10 nt toehold for invasion of an adjacent domain by a staple of the other type via toehold-mediated strand displacement. (c) Sample set of intermediate states is shown for the process of $\mathcal{R}2$ staples replacing \mathcal{E} staples.

(see section 3.3) allowing the staples from the two staple sets to compete for binding sites on the scaffold. In all experiments, the concentration of the scaffold strand was 100nM and the concentration of both coating and folding staples was 200 nM. For computational simplicity, a slight change was made to the $\mathcal{R}2$ design in simulations: the two 32 bp domains of the system (at the top and bottom of the structure) were broken into two 16 bp domains each. This specific variant is sometimes referred to as $\mathcal{R}2^H$ in the figures but I will use the two names interchangeably in this section.

For the folding staples to displace the coating staples, the target structure of the folding staple set must be thermodynamically more stable than that of the coating staples. The coating and folding sets have the same number of base-pairing and stacking interactions but differ in the number of inter-helical cross-links. The \mathcal{E} set consists of 168 single-domain staples. The $\mathcal{R}2$ set consists of 16 one-domain

staples and 76 two-domain staples that form crossovers between the helices of the target structure. Due to the cross-links, the conformational entropy of $\mathcal{R}2$ is lower than \mathcal{E} . The close-packing of helices in $\mathcal{R}2$ also has a destabilising effect caused by electrostatic repulsion between DNA backbones. However, we can expect these effects to be more than compensated for by the configurational entropy gain due to the lower number of staples used in $\mathcal{R}2$ than \mathcal{E} . Both experimental and simulation results in annealing protocols with single-stranded templates (see previous chapter) confirm that $\mathcal{R}2$ has a higher melting temperature than \mathcal{E} . We also observed that the cooperativity between cross-linking staples, and seam staples in particular, results in a sharper transition in $\mathcal{R}2$ than \mathcal{E} .

After the one-hour incubation period at specific temperatures, the samples are analysed using agarose gel electrophoresis; 20 samples can be analysed at a time. Gel electrophoresis can distinguish between different structures present in the sample based on their different electrophoretic mobilities. A voltage is applied across the gel which causes the charged structures to move through the gel at rates determined by their size, shape and charge. Typically, the structures are visualised by either staining the DNA bands in the gel with a fluorescent dye or by using a staple strand that is labelled with a fluorophore. Since a well-folded $\mathcal{R}2$ origami is more compact than the coated-template \mathcal{E} , we expect it to move faster through the gel. In the experiments presented in this chapter, up to three staple strands are labelled with fluorophores. The positions of these fluorescently-labelled DNA strands are shown in Figure 5.2a. Two strands from the folding $\mathcal{R}2$ set are labelled with blue and red fluorophores and one strand from the coating \mathcal{E} set is labelled with a green fluorophore. The labelled coating strand (shown in green) shares some part of its binding site on the scaffold with the labelled ‘body’ staple (shown in red). All results presented in this chapter use 0.7% agarose gel with a 60V voltage applied for two hours at room temperature. A subset of the samples were analysed using AFM. Because AFM is much more time-consuming and expensive, it is mainly used here as a way to confirm our interpretation of

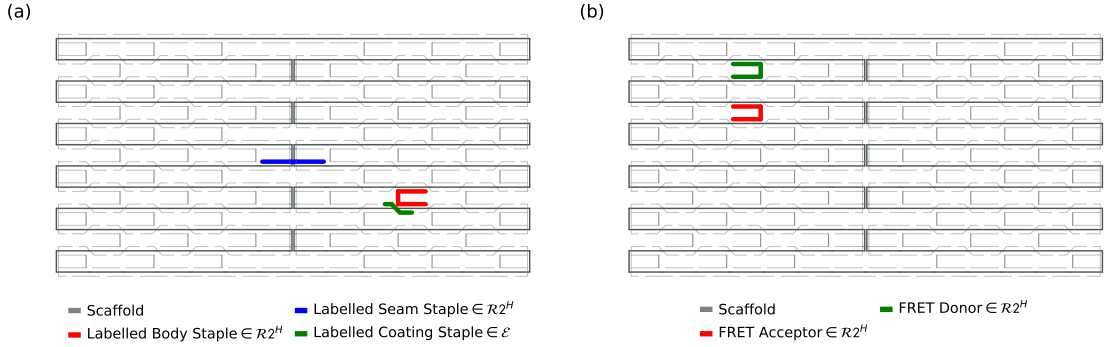


Figure 5.2: Position of the fluorescently-labelled strands used in experiments in: (a) gel electrophoresis experiments and (b) real-time temperature-dependent FRET experiments.

gel electrophoresis experiments. Gel electrophoresis and AFM provide single time-point information about the final structure. In some cases, real-time temperature-controlled fluorescence measurements using a FRET-labelled donor/acceptor pair of staple strands were used to probe the binding state of a particular region of the structures. The positions of the labelled strands are shown in Figure 5.2b.

Figure 5.3 shows the experimental and simulation results from our initial investigations. Simulations were prepared with an initial state corresponding to every domain of the scaffold bound to staples from the \mathcal{E} set. Isothermal trajectories were generated with the default parameter set ($n = 2$, $\gamma = 2.5$ and $k_f = 0.5 \times 10^5 \text{M}^{-1} \text{s}^{-1}$) using the sequence-specific parametrisation of the model. The number of base pairs of the scaffold occupied by $\mathcal{R}2$ was used as a proxy for the degree of folding. The sample trajectories suggested that the transition is highly sensitive to temperature. A few trajectories were used to estimate the temperatures at which the transition occurs on reasonable time scales. The results suggest that folding should occur on 1-hour time scales between 50°C and 55°C .

For the initial experiments, coated scaffold strands were incubated with folding $\mathcal{R}2$ strands for different periods of time and analysed using gel electrophoresis (Figure 5.3c). The seam staple (blue in Figure 5.2a) was not labelled in this set of experiments. The competing ‘coating’ and ‘folding’ staples were labelled with red and green fluorophores, respectively. Structures with different mobilities can be

observed in each band, with a well-defined red band appearing at 59°C, 57°C and 55°C after incubation periods of 0.5, 1 and 4 hours, respectively. These temperatures are well below the melting temperature of the $\mathcal{R}2$ origami, which is 68°C. The fainter lower mobility red bands appearing in some lanes seem to correspond to higher-order structures and will be discussed further in the next sections. Two of the samples were also analysed with AFM to confirm that the strong red bands correspond to well-folded $\mathcal{R}2$ tiles. We can see that well-folded structures do appear at 55°C after 1 hour and the yield improves if incubation time is increased to 4 hours.

AFM images also show that in the absence of coating staples, folding of $\mathcal{R}2$ at temperatures as low as 55°C produces defective structures. AFM images of isothermal assembly of $\mathcal{R}2$ with non-coated scaffold strands are shown in Figure 5.3b. The equivalent results of a standard annealing protocol are also shown for reference. In standard non-coated origami, the yield of isothermal assembly is generally understood to suffer at temperatures well below the melting temperature¹ because misbinding causes kinetic traps that are not easily reversed. Annealing protocols are thought to improve yields by increasing the likelihood of designed interactions occurring and allowing the reversal of off-target interactions if they occur in earlier stages of assembly. Still, isothermal assembly was still shown to be viable by Sobczak *et al*[53]. In their experiments, single-stranded scaffolds are first heated to 95°C to get rid of any unwanted secondary structure. This is followed by a period of incubation with staple strands at temperatures of $\sim 7^\circ\text{C}$ below the melting temperature of the origami as determined by peak fluorescence. They were able to achieve near 100% yields from this protocol in very rapid assemblies compared to annealing protocols.

In the case of our origami, the melting temperature of the $\mathcal{R}2$ tile is 69°C. Therefore, at 55°C, isothermal assembly of the single-stranded scaffold leads to extremely poor yields, and increasing the incubation time does not improve yields as unwanted interactions are difficult to reverse at this temperature. This was confirmed by Katherine in experiments where the incubation time was increased

¹Melting temperature is usually determined by the temperature at which half of the scaffold is bound to staples and determined by scanning a range of temperatures in annealing protocols.

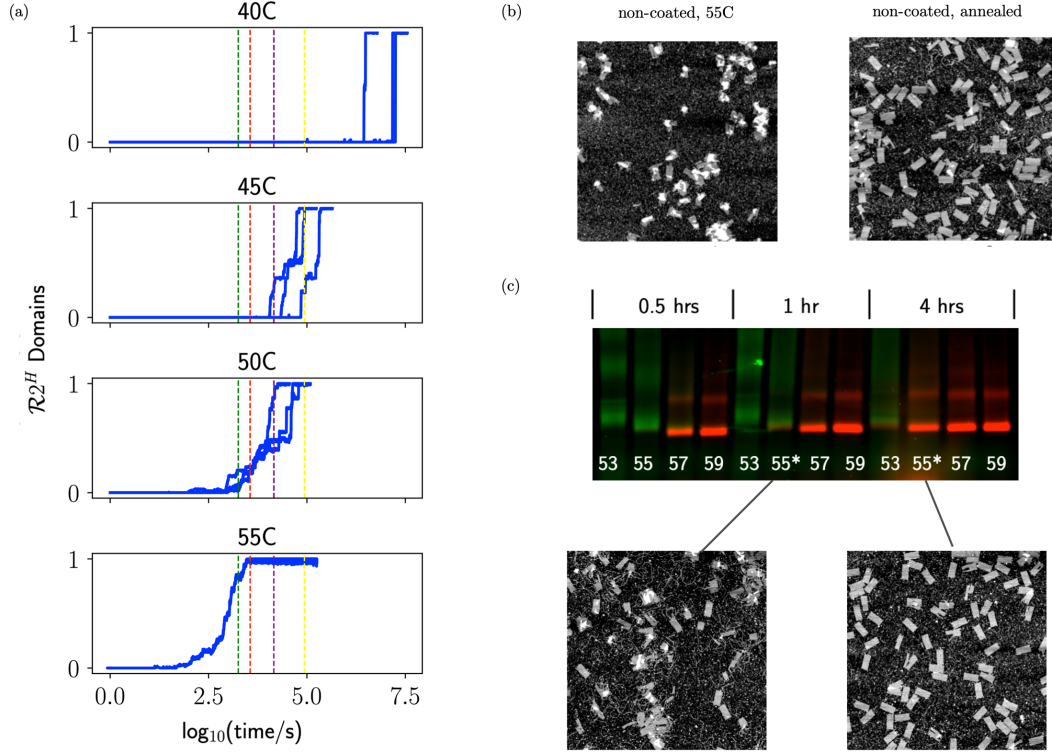


Figure 5.3: Initial investigation of the folding process. (a) Three sample simulation trajectories at four temperatures are shown. The vertical green, red, purple and yellow lines indicate 0.5, 1, 4 and 24 hours respectively. The simulation results suggest complete folding happens in 1 hour at $\sim 55^\circ\text{C}$. (b) AFM images of assembly of $\mathcal{R}2^H$ with single-stranded scaffold after one hour incubation at 55°C (left) and after annealing (right). (c) Folding of $\mathcal{R}2^H$ with a coated scaffold. The scaffold is annealed with \mathcal{E} strands, with one \mathcal{E} staple labelled with a green fluorescence label. The coated origami is then incubated with $\mathcal{R}2^H$ strands, with one $\mathcal{R}2^H$ staple labelled with a red fluorescence label. Gel electrophoresis results after incubation of the coated template with $\mathcal{R}2^H$ staples for 0.5, 1 and 4 hours are shown. The transition is sharp and strongly dependent on temperature and incubation time. Two samples at 55°C are analysed using AFM, showing enhanced assembly yields after 4 hours.

to 4 hours and no change was observed in the gel band compared to the 1-hour case. Coating the scaffold strand with \mathcal{E} strands improves the yield at 55°C at 1-hour incubation and results in near-perfect yields at 4 hours of incubation. The coating staples, therefore, help the process by blocking regions of the scaffold to reduce kinetically trapped, defective structures. Encouraged by these results, we proceeded to study the folding process further.

5.3 Simulations at Low Temperature

The main obstacle to the successful isothermal assembly at room temperature is the increased probability of kinetically trapped interactions, which fall into two types:

- Misbonding: non-native scaffold-scaffold or staple-scaffold hybridisation of regions on strands that have partial sequence complementarity but are not designed to interact.
- Staple blocking: binding of multiple copies of a multi-domain staple to the scaffold without forming the designed crossovers.

As shown in simulations by Snodin *et al*[104], at temperatures just below the melting temperature of the origami staple blocking is the dominant cause of incomplete assembly at high excess of staple to scaffold strands. They also showed that misbonding is the dominant factor at temperatures well below the melting temperature of the origami at stoichiometric staple concentrations. It is not clear which type of undesired interaction is more important in the high-concentration and low-temperature regime. The domain-level model does not include a description of misbinding but states with blocked staples can have significant lifetimes at low temperatures.

5.3.1 Non-Coated Origami

To understand the low-temperature behaviour of the model, let us first examine the isothermal assembly of the $\mathcal{R}2$ origami in the absence of any staples from the \mathcal{E} set (similar to the last chapter). I ran simulations at various temperatures, with the initial state corresponding to the fully unbound scaffold, and without the presence of any competing \mathcal{E} strands. The simulations were stopped when one of the following criteria was fulfilled: the target state (defined by all possible crossovers bound) is reached, or the number of steps exceeded 10^7 . An upper limit on the number of steps is necessary for isothermal simulations because some of the trajectories never reach the target state. At high temperatures, this is because the target state is not the

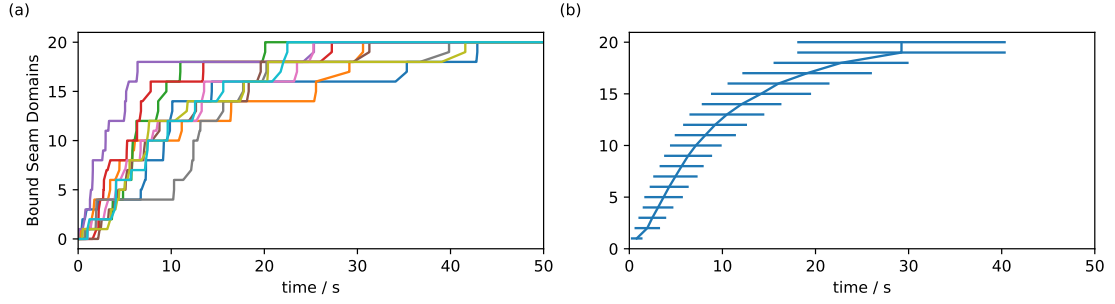


Figure 5.4: Averaging over simulation trajectories. The number of seam domains bound to the scaffold (out of a possible 20 seam domains) during simulation of $\mathcal{R}2$ origami at 30°C. (a) 10 sample trajectories. Each colour is an independent trajectory. (b) Time taken to reach the values of the order parameter is averaged over 50 trajectories. Error bars represent one standard deviation from the mean.

equilibrium state. At low temperatures, some trajectories get kinetically trapped due to staple blocking and the simulation never reaches equilibrium. In the most extreme case, all scaffold domains are bound to staples, but the number of crossovers does not reach its maximum value because some of the staples are in the s_{12} state.

Figure 5.4 shows sample trajectories generated at 30°C using the average-sequence parametrisation of the model. Due to the variable time-step of the model, the value of the observables as a function of time can vary greatly across individual trajectories, which presents a difficulty in obtaining trajectory-averaged results. Kinetically trapped structures that do not reach equilibrium within the allotted number of steps also complicate the process. To track folding, we can define order parameters $A(\vec{s})$, with a set of values $\{a\}$, that correspond to some observable of interest, such as the number of bound domains. We are interested kinetic behaviour of the system: the time taken for the system to explore values of the order parameter. Depending on the behaviour of the system, I will present the results by averaging one of the following:

- The first time $\{\tau_0(a)\}$ for each trajectory to reach each point in $\{a\}$.
- The total time $\{\tau(a)\}$ each trajectory spends at each point in $\{a\}$.

To highlight the role of staple blocking at low temperatures, Figure 5.5 shows the isothermal assembly of $\mathcal{R}2$ origami, with τ_0 values averaged over trajectories.

Histograms of the total time spent by each staple in each of its possible states were also recorded for each trajectory with averaged results shown in the figure. At 69°C and above, simulations never reach the target state. The results are consistent with the free energy profile of $\mathcal{R}2$ (Figure 4.8). The average occupancy at 69°C is ~ 40 domains, with seam and seam-adjacent staples the most likely staples to be bound. Seam-adjacent domains close the shortest loops and their presence mediates the binding of seam staples through coaxial stacking. Given enough time to overcome the barriers associated with double-crossover formation, these staples are able to bind. At this temperature, blocked states are short-lived and occur mostly at domains that are close to seam-adjacent staples due to the additional coaxial stacking stability.

At 68°C, the majority of two-domain staples are stable in the presence of seam staples. This is due to the coaxial stacking cooperativity with the stable structures formed by seam and seam-adjacent staples. The folding pathway follows the same steps as 69°C, followed by the binding of other staples from the middle outwards. Single-domain staples at the top and bottom duplexes are still unstable. Folding at the seams happens an order of magnitude faster than at 69°C due to the lowered free energy barrier. From 67°C to 62°C, the results are similar to 68°C but the time taken to reach the target state (defined by all crossovers bound) reduces. Single-domain staples are also able to bind in this temperature range.

As the temperature is lowered to 61°C, 16 bp domains become stable without the presence of other staples. The folding pathway is no longer dictated by loop costs or coaxial stacking and hybridisation energies become dominant in determining the total free energy change of reactions. At temperatures below 61°C, staple blocking becomes significant, with most blocking occurring at the seams. This causes the system to spend a long time escaping the kinetic trap, increasing the time taken to reach the target state (note the change of the x-axis range for 45°C and 25°C).

It is difficult to gather enough statistics to distinguish individual blocked staples since they are rare events. Nonetheless, a pattern does emerge pointing to more prominent blocking at the seams. This could be due to the longer loops that seam staples close. In the model, binding rates are not temperature dependent, whether

the transition involves loop closure or is a bimolecular reaction involving a free staple. However, unbinding rates are strongly dependent on temperature. At low temperatures, any domains that bind are unlikely to unbind. The presence of these bound domains increases the loop cost for seam staples compared to the cost for an otherwise empty scaffold. This reduces the binding rate for the second domain of the seam staple and increases the relative probability of a blocking reaction occurring at the seam.

These results suggest the existence of an energy barrier associated with binding the last remaining crossovers at low temperatures. No such features associated with staple blocking were observed in the free energy profiles from umbrella sampling simulations shown in the previous chapter for two reasons. Firstly, the profiles were calculated at temperatures close to the melting temperature (above 65°C). At these temperatures, any blocked states are short-lived as seen in Figure 5.5 (crossovers reach their maximum value before domains). Secondly, the features would not be present on profiles that use domains as the order parameters since there are no kinetic traps associated with domains. In reality, other sources of misbinding would introduce such energy barriers as domains can also become misbound. Due to time constraints, I have not run umbrella sampling simulations at lower temperatures using other order parameters.

While the model does not include a full description of misbonding, increased blocking at low temperatures does shed some light on why annealing protocols, and isothermal folding close to the melting temperature, result in much better yields than at low temperatures. If annealed at an appropriate rate, by the time the system gets to temperatures with significant blocking, staples that are most likely to be blocked under isothermal conditions are already fully bound to the scaffold. In the particular case of $\mathcal{R}2$, the stability provided by double-crossover formation allows the seams to bind at high temperatures. By the time the annealing process gets to low temperatures, the loop closure cost at the seam is extremely low. If a domain unbinds at the seam, it is the least likely to bind a second copy, given the low loop cost. If annealing is done too fast, the seams no longer

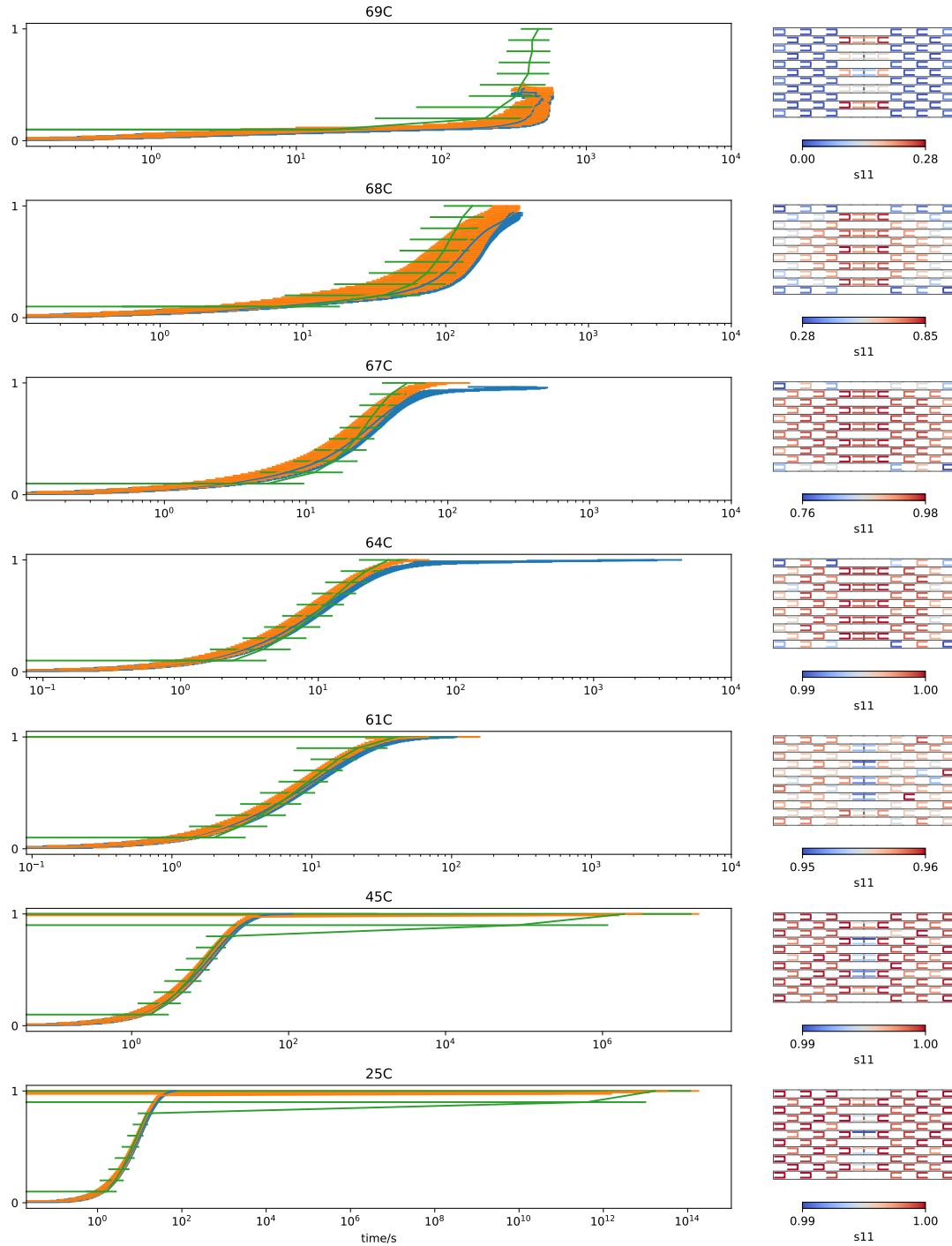


Figure 5.5: Isothermal simulations of non-coated scaffold with $\mathcal{R}2$ staples in the absence of any other staples. At each temperature, the left panel shows the time taken to reach values of each order parameter, with the range normalised between 0 and 1 for a clearer comparison. The horizontal error bars are standard deviations when averaging over trajectories. The right panel shows the proportion of time spent by each staple in the fully bound (s_{11}) state. Results are averages over 1000 trajectories using the sequence-average ($\langle \text{pKD} \rangle$) parametrisation of the model.

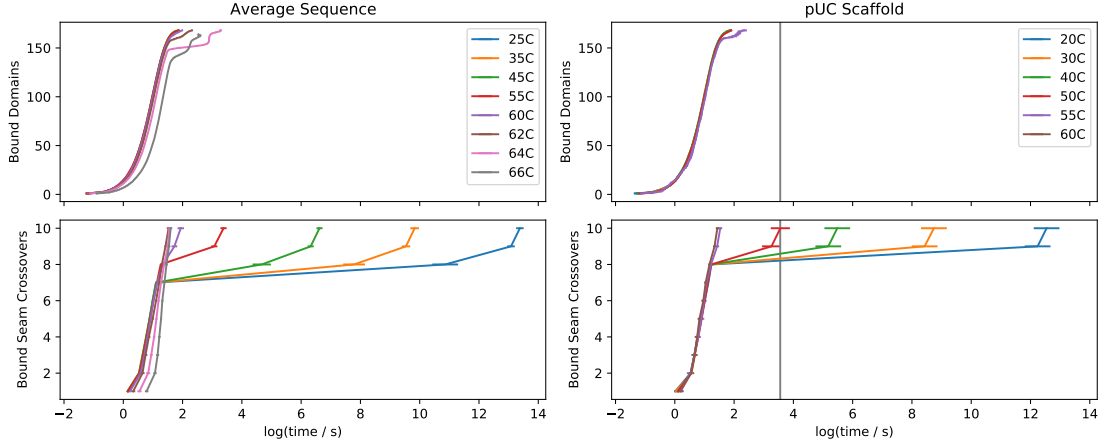


Figure 5.6: Effect of sequence-specificity on non-coated origami. Comparing the times for bound domains and bound seam crossovers for the average-sequence and sequence-specific isothermal folding of $\mathcal{R}2$ origami. The one-hour time is marked with a grey line in the sequence-specific case. For temperatures above 50°C, both models show that staple blocking can anneal out within an hour.

have enough time to overcome the barrier associated with double crossovers at high temperatures. Combined with staple blocking, this contributes to the higher hysteresis seen under faster temperature ramps.

Our experimental data suggest that isothermal assembly of the non-coated origami after 1 hour at 62°C results in similar yields to annealing but at 55°C, yields are poor. Figure 5.6 shows that the model predicts all staple blocking can be corrected after an hour at temperatures above 50°C. This discrepancy is likely due to other types of misbinding, which are not taken into account by the model.

5.3.2 Coated Origami

I ran simulations at various temperatures to compare the isothermal assembly of coated and non-coated origami. In both cases, \mathcal{E} staple and $\mathcal{R}2$ staples were included in the system, but the initial state of the scaffold was chosen as fully single-stranded in the non-coated case, or fully bound by \mathcal{E} strands in the coated case. To gain some familiarity with the simulation output, single trajectories produced by the model at 40°C and 50°C are shown in Figure 5.7. In the non-coated case, the $\mathcal{R}2$ staples rapidly bind to the scaffold, with little binding by \mathcal{E} staples. The folding time did not change significantly if \mathcal{E} staples were excluded from the simulation

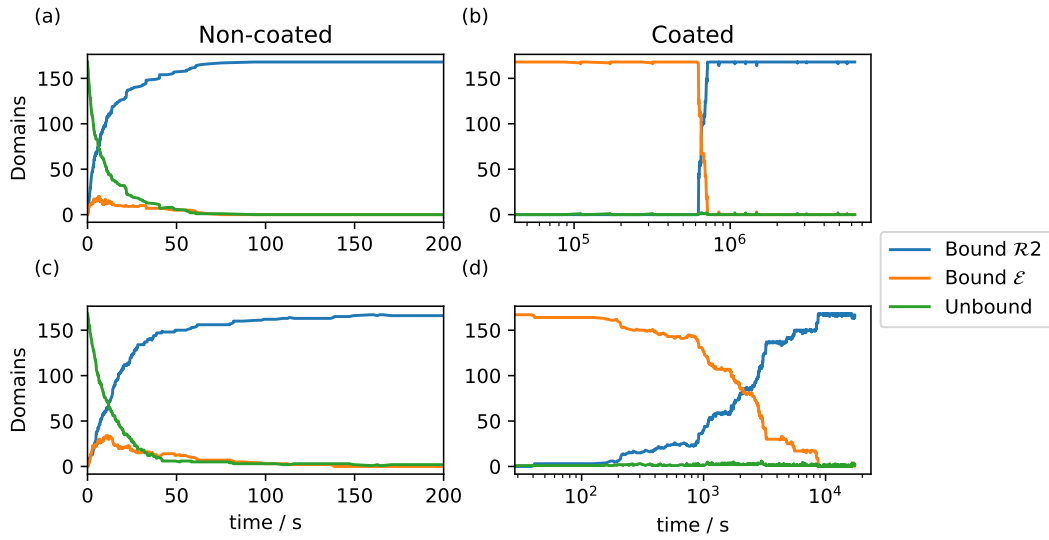


Figure 5.7: Simulation trajectory for isothermal assembly of non-coated (a),(c) and coated (b),(d) origami. The temperature is set to 40°C in (a),(b) and 50°C in (c),(d). In all trajectories, sequence-specific parametrisation of the model is used.

and the trajectories of $\mathcal{R}2$ staples remained nearly identical. In the coated case, there is an initial waiting time in every trajectory determined by the stability of the fully-coated state, which is strongly temperature-dependent. Folding does not begin until a toehold is made available for $\mathcal{R}2$ staples. Once an \mathcal{E} staple unbinds, the probability of re-binding is high due to the stability provided by coaxial stacking at both ends of the domain because all other \mathcal{E} domains are already bound. The process repeats a number of times before the first $\mathcal{R}2$ staple binds. The relative probability of rebinding of the \mathcal{E} staple and displacement of a further \mathcal{E} staple by an $\mathcal{R}2$ staple is determined by the specific sequence of the binding domains. Therefore, folding usually begins at positions where the offset between overlapping \mathcal{E} and $\mathcal{R}2$ domains causes the $\mathcal{R}2$ domain to have higher stability than the competing \mathcal{E} domain. This is not immediately obvious and will be discussed further below.

Figure 5.8 shows the proportion of the scaffold bound by $\mathcal{R}2$ staples as a function of time, averaged over 1000 independent trajectories. In the non-coated case, the results are similar for both sequence-averaged and sequence-specific simulations. In the case of coated origami (Figure 5.8b,d), the transition is highly sensitive to temperature and the introduction of sequence-specificity leads to faster activation of

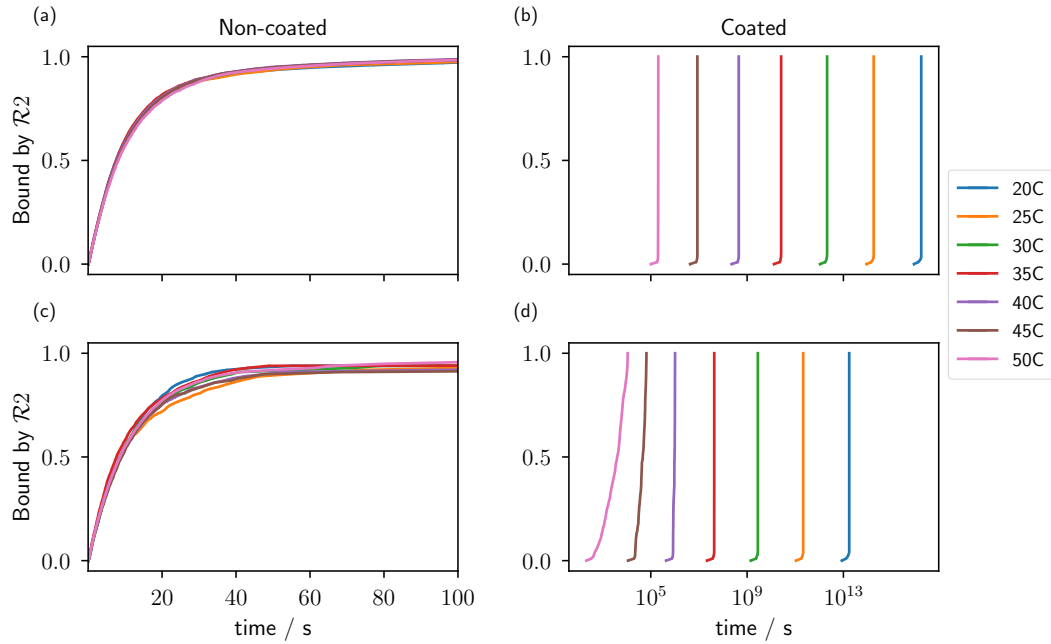


Figure 5.8: Simulation results for isothermal assembly of non-coated (a),(c) and coated (b),(d) origami, using average base-pair (a),(b) and sequence-specific (c),(d) parametrisation of the model. At each temperature, the proportion of the scaffold bound by the folding staples is averaged over 1000 trajectories.

the process at all temperatures. This is because the difference between the stabilities of competing domains is heterogenous in the sequence-specific case, which introduces favoured binding sites on the scaffold for the process to begin. As can be seen in Figure 5.7, the number of toeholds available on the scaffold (unbound regions) does not change significantly at temperatures below 50°C. This points to the growth of a single nucleus along the path of the scaffold through strand exchange.

To track the folding pathway, histograms of the time spent by individual staples, domains and crossovers in each possible state were recorded during simulations. Figure 5.9 shows two sample trajectories for the sequence-specific and average-sequence simulations. The nucleation site is random for the sequence-averaged case as can be seen by the nucleus forming in two different regions in Figures 5.9(c) and (d). In the sequence-specific case, I found that most of the trajectories nucleate in the same region at the bottom left corner of the origami as shown in (a) and (b). In both cases, the nucleus grows along the path of the scaffold in both

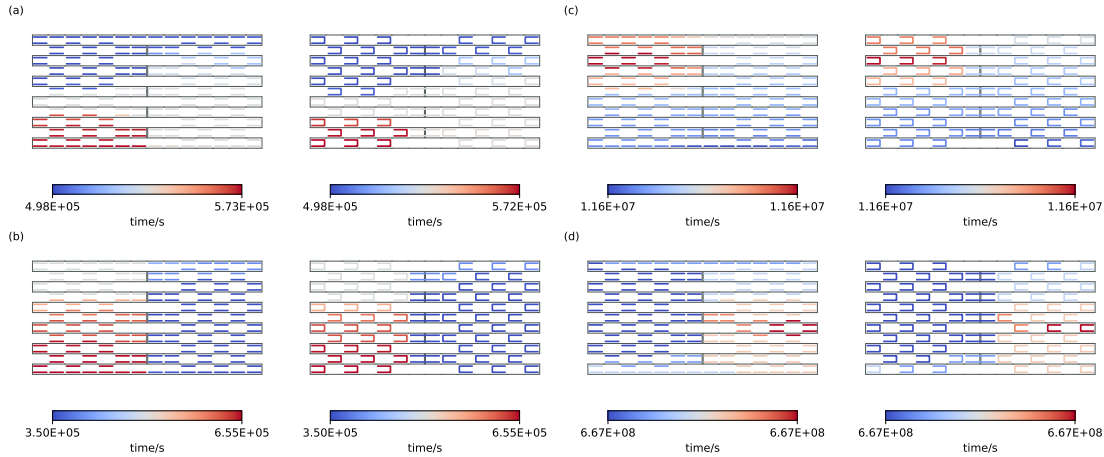


Figure 5.9: Simulation trajectories showing the folding pathway of $\mathcal{R}2$ staples on coated origami at 40°C. The colours correspond to the amount of time the staple/domain spends in the bound state: red regions bind first, and blue regions bind last. (a) and (b) are two independent trajectories using the sequence-specific model. (c) and (d) are two independent trajectories using the average-sequence model. In each panel, two sets of data are shown: time spent by domains in the bound state (left) and time spent by staples in the s_{11} state (right).

directions. In each panel, it is also instructive to compare the folding pathway of domains to those of staples, particularly at the seams. We can see that one domain of each seam staple binds far faster than the second domain. Therefore, it is evident that folding does not propagate through the seams to the opposite side. To get from one side to the other, the nucleus propagates through the top and bottom rows. Each seam staple is only able to fully bind and form a crossover when the nucleus has reached both of its domains.

In each trajectory, the simulation is stopped when all $\mathcal{R}2$ staples have been bound to the scaffold in the correct configuration for 50% of time, resulting in a wide range of total trajectory times. To average individual staple or domain data over multiple trajectories, it is useful to first normalise individual trajectory data sets to the corresponding end-point in time before averaging to show the sequence of events more clearly. Figure 5.10 shows the average proportion of time spent by domains and crossovers in the bound state. In the sequence-specific case, a single favoured nucleation site is clearly visible at 40°C and 45°C. Simulations at other temperatures showed that at 47°C and below, this is the only nucleation site, with

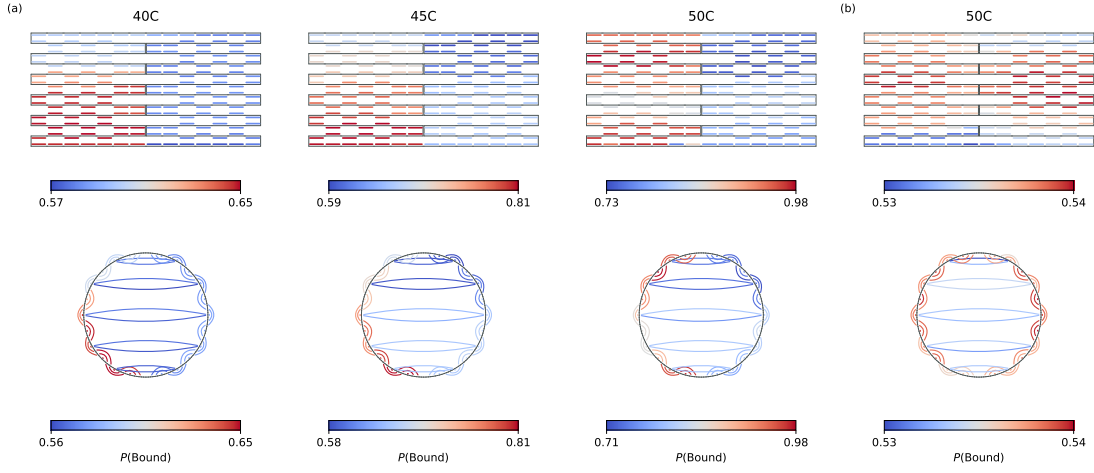


Figure 5.10: Simulation results averaged over 1000 trajectories, showing the folding pathway of $\mathcal{R}2$ staples on coated origami. The top and bottom rows show the average fraction of time each domain and crossover spends in the bound state. (a) Sequence-specific case at 40°C, 45°C, and 50°C. (b) Sequence-averaged case at 50°C.

a second nucleus appearing at 48°C and a third at 52°C.

It is not immediately clear whether the preferred nucleation site is a result of a weakly-bound \mathcal{E} strand or the difference between the stabilities of the competing staples. In Appendix D.1, I compare the melting temperature of individual $\mathcal{R}2$ domains with those of competing \mathcal{E} domains along the scaffold. The least stable \mathcal{E} domains are positioned at the top left corner. The competing $\mathcal{R}2$ domains at the same position also lack stability because they share some of the same region on the scaffold. However, the highest difference in melting temperatures of competing domains is at the bottom corner, which is where nucleation usually occurs in the simulations.

In the sequence-averaged case, the only factor that affects the nucleation site is the length of the loop that the initial incoming $\mathcal{R}2$ staple closes. Propagation is mediated by closing further loops as the nucleus reaches an adjacent row. The top and bottom rows are unique because they contain 8 single-domain staples that do not form crossovers. In all other rows, the only domains that do not participate in loop closure are seam domains (1 on each side). This results in slower propagation at the top and bottom duplexes due to the local crossover environment, leading to the pattern in Figure 5.10b.

5.4 Nucleation

Figure 5.11a shows detailed experimental results for isothermal folding of the coated scaffold. In this set of experiments, all three labelled strands (see Figure 5.2) are used. At temperatures below 50°C, the only band that can be identified is the one associated with the coated template (lane A). At 51°C, a faint band with slightly higher mobility can be observed, with more intermediate structures appearing at 52°C. Folding is nearly complete within 1 hour at 53°C. The presence of well-defined intermediate structures can be attributed to the formation of seam cross-links, since the binding of each pair of seam staples strongly affects the size of the structure, leading to a well-defined mobility. This was confirmed by further experiments at various temperatures between 52°C and 53°C (see Katherine’s thesis[127]). Five distinct bands were observed in the green channel, corresponding to 5 pairs of seam crosslinks in $\mathcal{R}2$. However, when the experiments were repeated under identical conditions but with two pairs of seam staples omitted from the $\mathcal{R}2$ set, only three distinct bands were observed.

Simulation results are shown in Figure 5.11 for temperatures between 50°C and 55°C. Given that the bands in experiments correspond to structures with different seam cross-links, I have shown the number of bound $\mathcal{R}2$ domains as well as the number of $\mathcal{R}2$ seam crossovers as a function of time, which acts as a proxy for the defined bands. There is a great degree of agreement between simulation and experimental results, which are summarised below.

- At 50°C, the simulations predict that 40% of the scaffold is bound by $\mathcal{R}2$ staples, but no seam crossovers have formed after 1 hour. This is in line with the thickening of the single band at 50°C compared to band A because binding of non-seam staples would result in structures of similar (but not identical) mobility to that of the coated template.
- At 51°C, the simulations suggest that the amount of $\mathcal{R}2$ -bound scaffold increased to 60% and 2 seam crossovers have formed. In the gel, the main band is moved slightly lower, and a faint second band appears.

- At 52°C, $\mathcal{R}2$ -bound scaffold increased to 80% and seam crossovers are increased to 6. This is reflected in the gel by the well-defined band of intermediate mobility since the presence of 6 seam crossovers (which can be 3 pairs of adjacent seam staples) would increase the mobility of the structure.
- At 53°C, the simulations predict that folding is nearly complete and all seam staples are bound, in line with experimental data.
- At 54°C, the transition is completed well within an hour.

While the agreement is encouraging, it should not be over-interpreted because there are some discrepancies between our sets of experimental data. Particularly, the gels in Figure 5.11 contradict those of Figure 5.3 at 55°C for 1-hour incubation. We note that the two sets of experimental data utilise different dyes: two dyes are used in Figure 5.3, while three dyes are used in Figure 5.11. This could potentially explain the discrepancy as dyes are known to affect the melting temperature of origami. Furthermore, we do not have any AFM images of the origami with all three dyes, making it difficult to draw conclusions. Another set of experimental data would be helpful for the final publication of our results.

The long lag times seen in coated origami (Figure 5.8) suggest that coating the scaffold with \mathcal{E} strands introduces a nucleation barrier with the time required for crossing sharply increasing with decreasing temperature. I ran umbrella sampling simulations to calculate the free energy profile using the number of $\mathcal{R}2$ bound domains as the biased order parameter. The results are shown in Figure 5.12. There is a large free energy barrier associated with binding the first $\mathcal{R}2$ domain. I have also shown 2D profiles, using the number of scaffold domains bound by \mathcal{E} strands as a second reaction coordinate. For the first $\mathcal{R}2$ domain to bind, a \mathcal{E} strand must dissociate, and another must be displaced by an $\mathcal{R}2$ staple. At 50°C, the most favourable path is along the minimum number of toeholds (unbound scaffold domains). In the sequence-specific case, the regions with weaker \mathcal{E} strands are more likely to unbind, resulting in a lower nucleation barrier. At 60°C, a change in gradient can be seen in both cases toward the end of the profile. In

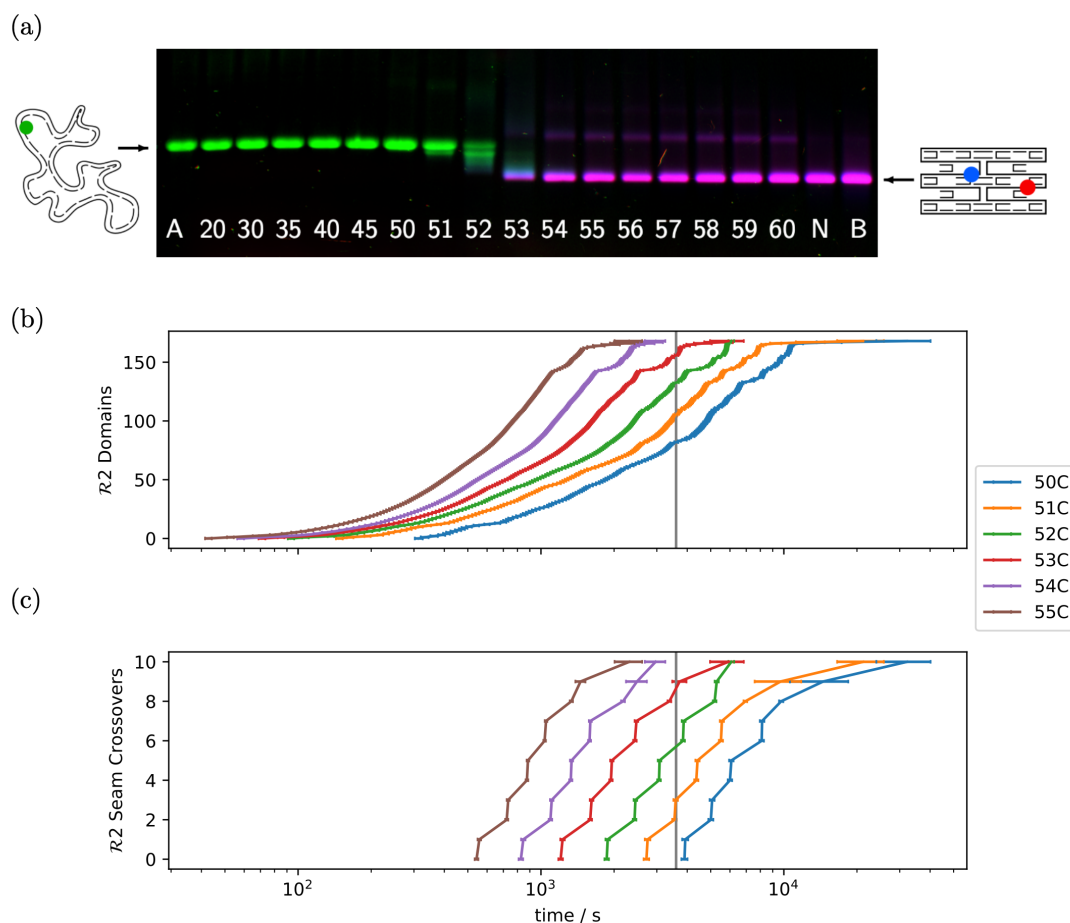


Figure 5.11: Comparison of simulation to experiment. (a) Gel electrophoresis results for the folding of the coated template. Lane A is the template annealed with \mathcal{E} strands. Lane N is the template annealed with $\mathcal{R}2$ staples. Middle lanes show the structures after incubation of the coated template with $\mathcal{R}2$ staples for 1 hour at the temperature indicated. (b) Simulation results showing the number of domains on the scaffold bound by $\mathcal{R}2$ staples. The grey vertical line marks 1 hour. Results for each temperature are averaged over 500 trajectories, with horizontal error bars representing standard errors. (c) Same as (b) but the number of crossovers formed by seam staples is shown.

the sequence-average case, this is due to the single-domain staples at the top and bottom duplexes of the $\mathcal{R}2$ design. In the sequence-specific case, it becomes unfavourable to bind some of the weaker domains and the free energy minimum no longer corresponds to the fully bound structure.

There is another interesting feature in the profile: sharper drops in free energy at 2, 6 and 8 domains. This is likely due to loop closure. In Figure 5.14, I show two scenarios for nucleation and the pathway for propagation. If nucleation occurs at

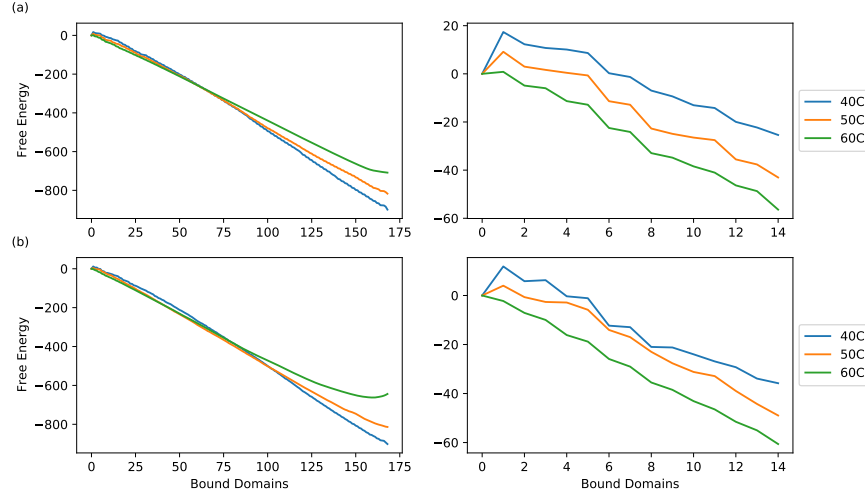


Figure 5.12: Nucleation barrier in coated origami. Free energy profile as a function of number $\mathcal{R}2$ domains bound to the scaffold. The figures on the right side are higher-resolution figures for up to 14 bound domains. (a) is the average-sequence case. (b) is sequence-specific.

the edge of the origami (Figure 5.14a), loop formation can happen by binding two domains and further loops can be closed at 6 and 10 domains. To proceed further with loop closure, 4 more domains are required because two domains at the seams cannot form loops. If nucleation occurs at the seams, 4 domains are required to bind to close the first loop and 8 domains to close the second loop. Thus, nucleation at the edge is favoured because binding the first domain of the staple creates a toehold for binding its second domain; this is not the case at the seam. Finally, I did observe some nuclei that form and propagate along a single helix in simulations. In the average-sequence parametrisation, this is a random walk along the helix that gets reversed by \mathcal{E} strands binding back. If no loops are closed by $\mathcal{R}2$ staples, it is favourable to reverse the process due to coaxial stacking of \mathcal{E} strands.

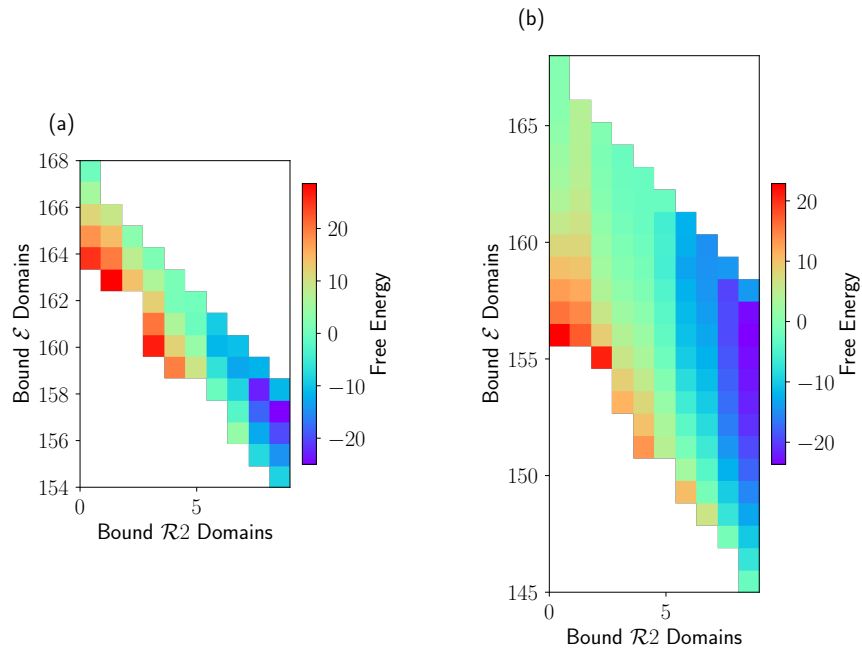


Figure 5.13: 2D free energy profile. Free energy profile as a function of number $\mathcal{R}2$ domains and \mathcal{E} domains at 50°C for (a) the average-sequence case, and (b) the sequence-specific scaffold. States above the diagonal are physically impossible states, whereas states below the diagonal are not sampled due to their low probability. Zero free energy corresponds to the state of the fully coated scaffold.

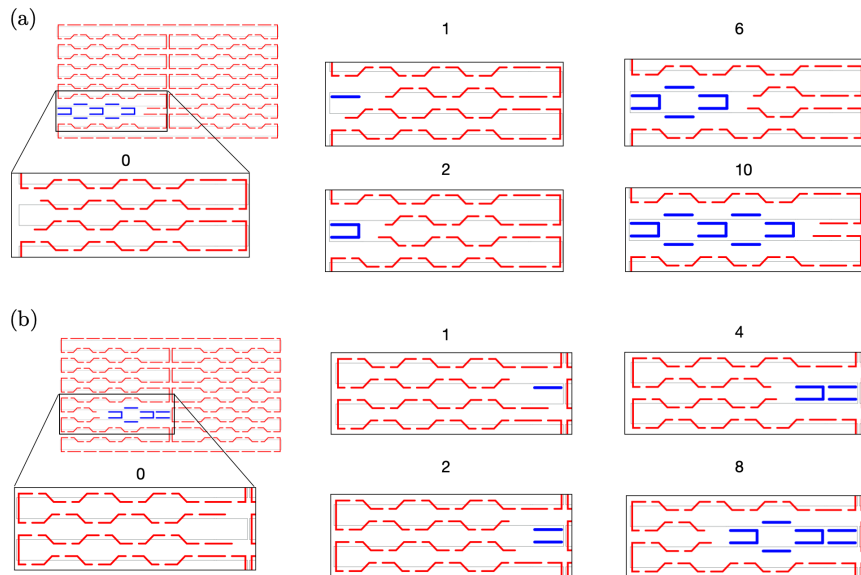


Figure 5.14: Nucleation pathway. Sample snapshots of simulations showing two possible nucleation pathways: (a) nucleation at the edge, and (b) nucleation at the seam. Bound \mathcal{E} domains are coloured red and bound $\mathcal{R}2$ domains (or staples) are coloured blue. The number of bound $\mathcal{R}2$ domains is shown above each state.

5.5 Propagation

If an $\mathcal{R}2$ staple is bound at the edge, then it can act as a nucleus that can propagate in both directions using available toeholds. In one direction, strand exchange reactions proceed via toeholds of 6 nt and branch migration domains of 10 nt.² This corresponds to $n = m = 6$ and $x = 10$ in Figure 3.3. In the other direction, the toeholds are 10 nt and branch migration domains are 6 nt. Given that the coating strands compete to use the same toehold/migration domain on each side, the effect of changing toehold lengths is minimal. In the average-sequence case, folding proceeds at a slightly different rate. In the sequence-specific case, the rate will be changed at different positions on the scaffold, depending on the sequence of the toeholds.

An interesting question is whether the folding pathway of the system can be controlled, either via manipulating nucleation or propagation. Figure 5.15 shows the average time $\tau(a \in [a_l, a_h])$ spent by trajectories at different values of a , where a is the number of bound crossovers in $\mathcal{R}2$. Each trajectory is stopped when the maximum number of crossovers ($a=76$) is reached. In the fully-coated system (left panel), nucleation time $\tau(a = 0)$ decreases exponentially with temperature in both the average-sequence and sequence-specific cases. This is due to the exponential temperature dependence of the unbinding rate of \mathcal{E} domains.

The time taken for completion $\tau(a \in [71, 76])$ also shows logarithmic dependence on temperature. This is due to staple blocking. Correction of blocking can proceed via strand exchange with \mathcal{E} strands, only if folding hasn't proceeded too far past the blocked domain. Otherwise, an unbinding event of a whole 16 bp domain is required to overcome the kinetic trap, causing long wait times with a similar temperature dependence to nucleation. It should be noted that the results are averages over all trajectories. While a minority of trajectories visit blocked states, their inclusion strongly skews the average results. In experiments, trapped trajectories would negatively affect the yield and increase the prevalence of higher-order structures. For example, the second domain of a staple in the s_{12} state can form a link to a second scaffold, leading to a dimeric structure.

²Toeholds of 5 nt and 9 nt also exist because some domains are 15 bps long.

Propagation is more weakly dependent on temperature and takes longer in the sequence-specific case. In the average-sequence case, the rate constant (Equation 3.25) reduces to $k_{nn} = k_f k_b / (k_n + 2k_b)$, where $k_b = 4s^{-1}$ for branch migration domains of 10 nt, and k_n varies exponentially with temperature. In the absence of coaxial stacking and scaffold shape contributions, this would result in an unbiased random walk along the scaffold by competing $\mathcal{R}2$ and \mathcal{E} domains, with an exponential dependence of the rate on temperature through the k_n term. The addition of shape and stacking terms to the expression for propensities w_{ij} (see Table 3.1) biases the walk towards states with maximum coaxial stacking (favouring \mathcal{E}), and a maximum number of loops (favouring $\mathcal{R}2$). In the sequence-specific case, the walk is also biased by the variability in free energies of the toeholds, which may lead to slower propagation in particular regions of the scaffold. At higher temperatures, the likelihood of multiple nuclei increases and the rate of propagation increases faster with temperature.

In the right panel of Figure 5.15, I repeated the simulation but omitted one of the \mathcal{E} domains. This has a dramatic impact on nucleation time but a minimal impact on propagation and completion times. This is because propagation times are mainly determined by strand exchange reactions. Since a toehold is available at all times when an \mathcal{E} strand is omitted, the nucleation barrier associated with binding the first $\mathcal{R}2$ domain is lowered, but the propagation time is not significantly altered. The impact on nucleation is not as dramatic in the sequence-specific case because the nucleation barrier is smaller due to the weaker domains. It would be interesting to see if the impact increases if the omitted coating strand is located at the preferred nucleation site of the scaffold.

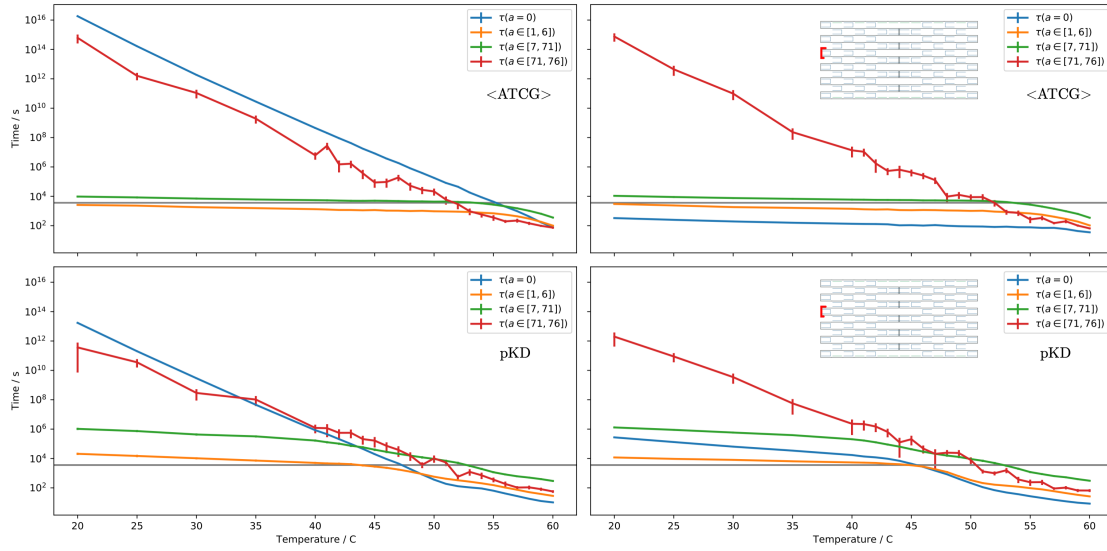


Figure 5.15: Comparison of nucleation and propagation time-scales. Time spent within various ranges of values of the number of bound crossovers in simulations. $\mathcal{R}2$ has a total of 76 crossovers in the target state. Average-sequence and sequence-specific cases are shown on the top and bottom panels, respectively. In the left panel, the full coating set \mathcal{E} is used. In the right panel, one coating strand (position shown in red) is omitted. The one-hour mark is shown as horizontal grey lines. Results at each temperature are averaged over 500 trajectories, and error bars indicate standard errors.

5.6 Controlling The Folding Pathway

Simulation results in the previous section suggest that the nucleation barrier can be lowered significantly by omitting a single \mathcal{E} strand, but propagation time remains as high as 10 days. Here, I will present some of our attempts to reduce propagation time by leaving out a subset of the coating strands. In simulations presented in this section, trajectories were generated with an initial state corresponding to the scaffold coated with all available \mathcal{E} strands. The minimum time taken to reach each value of the order parameters is recorded and averaged over 500 trajectories. Each trajectory is allowed to run until the target structure is reached, allowing sufficient simulation steps to correct any kinetically trapped structures. Each set of simulations is accompanied by gel electrophoresis results after an incubation period of 1 hour.

Figures 5.16 to 5.20 show 5 sets of experiments. In each figure, the gel assay is shown in (a), the positions of the omitted coating strands are shown in (b), and simulation results are shown in (c). The number of seam crossovers bound (out of a

possible 10) after 1 hour can be taken as a proxy for the electrophoretic mobility of the structure. The percentage of trajectories that undergo a single staple blocking event (two copies of the same staple bound) is also shown at each temperature. A higher blocking percentage may correlate to more prominent lower-mobility bands associated with higher-order structures. These bands were not seen in annealed samples but were observed in isothermal experiments with coated scaffolds to a varying degree. We believe they are dimeric or trimeric structures formed by cross-linking staples across two or more copies of the scaffold.

Figure 5.16 shows the results for a fully coated scaffold. At 30°C, the lag time before folding is of the order of 100 years. The one-hour mark is indicated by the vertical grey line. As discussed in the previous section, while no seam crossovers have formed at 50°C, we can expect to see fully formed structures at 55°C. Figure 5.17 shows the equivalent results if the 10 coating strands that compete with seam staples are omitted. Experimental results suggest that complete folding occurs at a lower temperature (50°C) than the fully-coated scaffold but the frequency of higher-order structures is also increased. Simulations suggest that at 20°C, we can expect 2-3 seam crossovers to have formed, corresponding to a well-defined band in the gel. The expected number of bound seam crossovers after one-hour increases gradually with increasing temperature and kinetic traps are more prominent compared to the fully-coated case.

Figures 5.18 and 5.19 show the equivalent results when 10 \mathcal{E} strands on either side of the seam staples are omitted, resulting in available toeholds at the seams of length 10 nt or 6 nt for 5' or 3' cases respectively. Both simulation and experimental results suggest that the overall effect is similar to omitting \mathcal{E} strands at the seams. The presence of different intermediate products in each case suggests that the folding pathway changes if different sets of omitted strands are used. Figure 5.20 shows that omitting \mathcal{E} strands at non-seam sites does not impact the folding time as dramatically as omission at the seams. Simulations indicate that seam crossovers form on longer timescales than when seam or seam-adjacent \mathcal{E} strands

are omitted. However, simulations predict that a higher degree of folding should occur at 50°C than shown by the gel.

AFM images of the structures[127] were consistent with the prediction that omitting \mathcal{E} strands at the seams and those adjacent to the seams result in better yields than the fully-coated template. As seen in the previous chapter, seam staples play a crucial role in the $\mathcal{R}2$ tile, since they form a nearly irreversible double-crossover structure. In the fully coated case, the nucleus propagates from one side to the other via either the top or bottom duplexes. The double crossovers can only form once the nucleus propagates to the other side of the tile. This can be a slow process, given the presence of the $\mathcal{R}2$ single-domain staples at the top and bottom duplexes.³ If nucleation occurs at the seam, the double crossovers can form early, leading to a nucleus that can propagate on both sides.

There is some downside to propagation at the seams. It might not be the best strategy because it may increase the likelihood of staple blocking at the seams. Both simulation and gel results suggest blocking (or misbinding) prevents the full formation of the structures. A more detailed analysis of the extent of this effect will be helpful in the future.

³In the experimental system, there are two 32 bp single-domain staples the top and bottom duplexes, which may speed up the propagation.

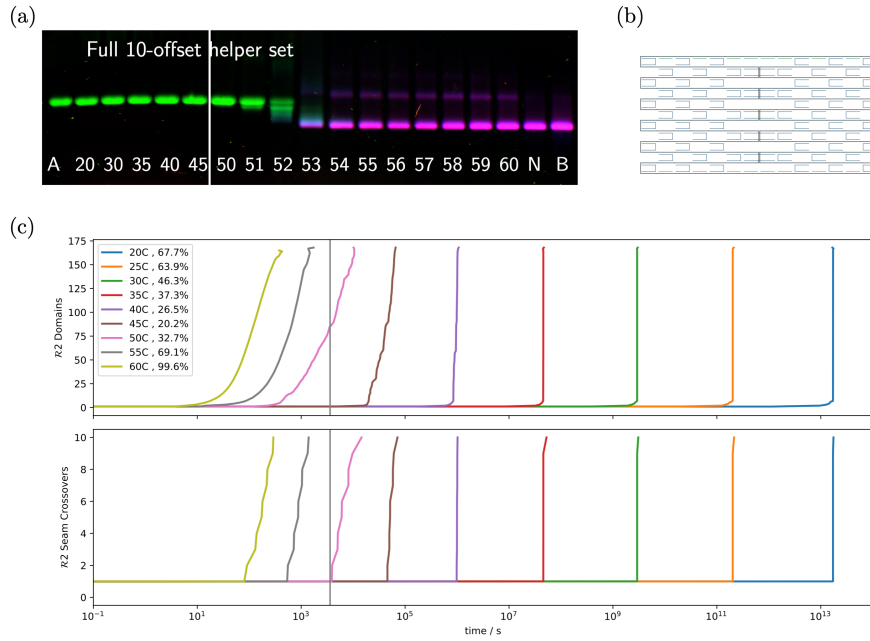


Figure 5.16: Comparison to experiment. Full \mathcal{E} set used. (a) Gel results. (b) No \mathcal{E} strands are omitted from the coating set in this case. (c) Minimum time $\tau_0(a)$ taken to explore values of order parameters.

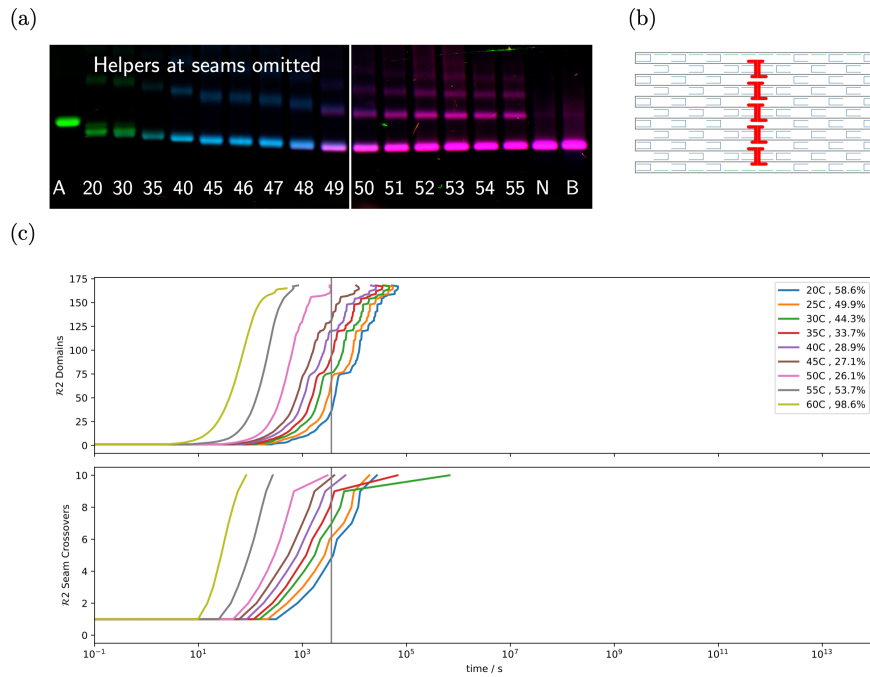


Figure 5.17: Comparison to experiment. Coating staples at the seam were omitted. (a) Gel results. (b) Position of omitted \mathcal{E} strands shown in red. (c) Minimum time $\tau_0(a)$ taken to explore values of order parameters.

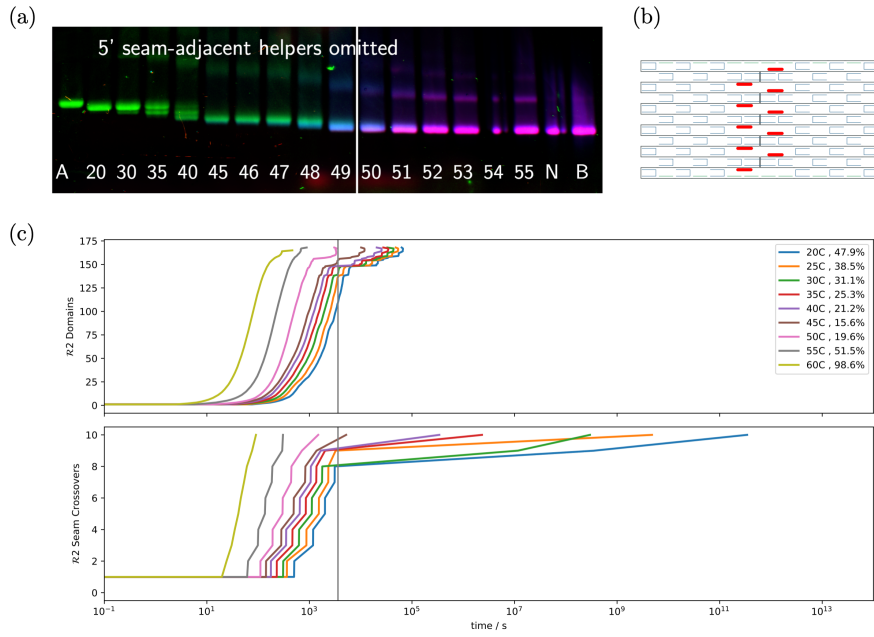


Figure 5.18: Comparison to experiment. Coating staples at the 5' side of seam staples omitted. (a) Gel results. (b) Position of omitted \mathcal{E} strands shown in red. (c) Minimum time $\tau_0(a)$ taken to explore values of order parameters.

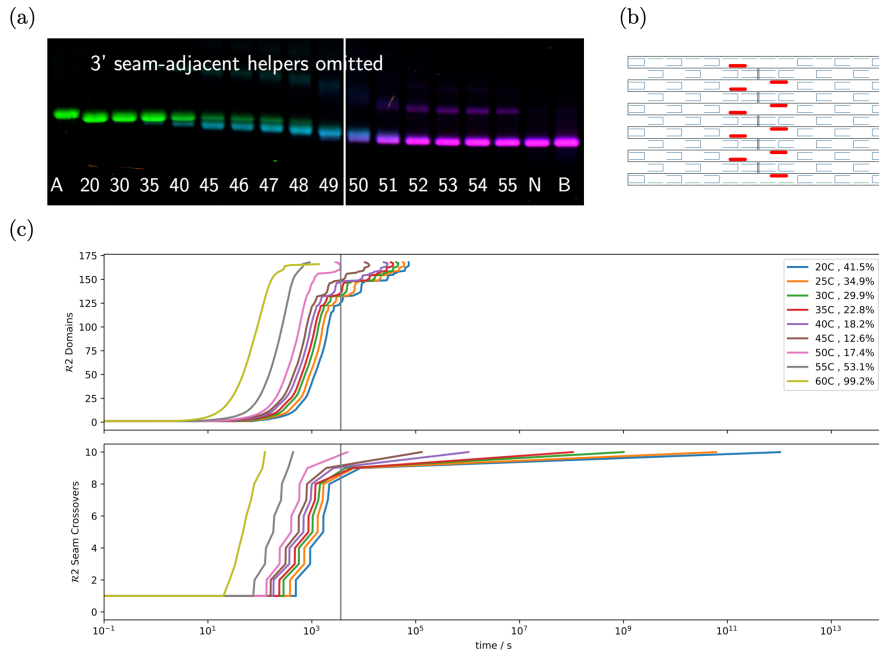


Figure 5.19: Comparison to experiment. Coating staples at the 3' side of seam staples omitted. (a) Gel results. (b) Position of omitted \mathcal{E} strands shown in red. (c) Minimum time $\tau_0(a)$ taken to explore values of order parameters.

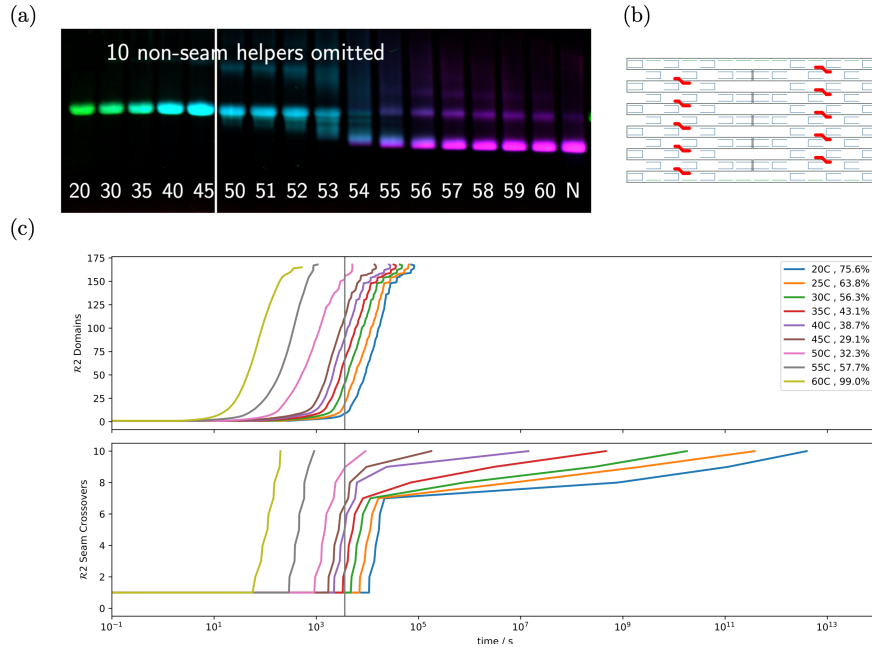


Figure 5.20: Comparison to experiment. Non-seam coating staples omitted. (a) Gel results. (b) Position of omitted \mathcal{E} strands shown in red. (c) Minimum time $\tau_0(a)$ taken to explore values of order parameters.

5.7 Summary and Outlook

In applications that involve in vivo assembly, rapid isothermal folding through efficient pathways is required at low temperatures. This remains one of the main challenges in DNA nanotechnology because of the irreversible nature of misbinding on experimental timescales. In this chapter, I presented a simple approach to lower the temperature at which isothermal assembly is viable. By coating the scaffold, an activation barrier can be introduced into the system allowing for controlled assembly. The timescales predicted by the domain-level model for nucleation and propagation largely agree with the experimental data without any adjustments to fitting parameters. This agreement is somewhat surprising because misbinding due to partial sequence complementarity is not described in the model. However, while the model would not be suitable to studying low-temperature assembly in normal origami, the reduced amount of misbinding by shielding the scaffold with coating strands may explain the agreement.

We attempted to control the folding pathway by omitting sets of coating strands.

This was successful to a degree when coating strands at the seams were omitted, resulting in a reduction of the folding temperature by $\sim 5^\circ\text{C}$. At lower temperatures, staple blocking becomes significant due to the lack of the shielding effect of the omitted coating strands. It would be interesting to test other strategies, for example by using longer domains in the folding set that span multiple coating strands. I leave this to future work.

The $\mathcal{R}2$ origami may not be the ideal system for fast isothermal assembly using a coated scaffold. We have seen the dominant role played by double crossovers in determining the folding pathway of origami. In the $\mathcal{R}2$ origami, such stable structures can only form by connecting distant parts of the scaffold. This can slow the folding process on coated scaffolds because the nuclei propagate along the path of the scaffold. A design similar to $\mathcal{R}3$ in the last chapter may be more useful because double-crossovers can form at all intermediate stages, creating nearly irreversible structures along the propagation path.

Currently, the model only allows competition between two sets of staples (eg. \mathcal{E} and $\mathcal{R}2$). This could be expanded in the future to study coated scaffolds in a pool of two other staple sets (eg. \mathcal{E} , $\mathcal{R}2$, and $\mathcal{R}3$). This would allow us to optimise the design of future one-pot competitive experiments, where the ‘winner’ origami can be selected by omitting coating strands at its preferred nucleation site.

6

Conclusions

In this thesis, I investigated the folding pathway of several types of DNA origami using modelling tools in close collaboration with experimental work. In chapter 2, I used oxDNA [91], [93], a coarse-grained model with nucleotide resolution, to study a novel origami-like structure based on the DNA T-junction. In chapters 3-5, I extended the domain-level model of Dannenberg et al. [59] to study the thermodynamics and kinetics of origami with single-stranded and coated scaffolds.

A relatively new structural motif based on the DNA T-junction has recently been used to assemble several DNA nanostructures. I characterised the stability and geometry of the T-motif using oxDNA to inform its use in larger assemblies. I found that the range of conditions under which the T-motif is stable is determined by the polarity of its sticky end insertion. Contrary to previous works, we were able to use both the 5' and 3' junctions to successfully assemble a large origami-like structure. The relatively weak interactions used in this design compared to normal origami allowed us to lower the assembly temperature to 20°C. The yields were higher than expected, especially given we only used four unique interactions to encode the target structure. We propose that the use of weaker 3' interactions aid assembly at low temperature by allowing reversal of kinetically trapped structures. The design can serve as a template to systematically study the changes in the folding pathway as interactions are altered. This can be done in an economically feasible way by

changing a few short adaptor strands. Future work could focus on this aspect, and a domain-level model can be specifically parametrised to accompany experiments.

I extended the domain-level model to simulate systems with three-domain staples and performed a series of experiments to check the model captured the effect of changes in origami design. I obtained fluorometric melting and annealing curves for four origamis (\mathcal{E} , $\mathcal{R}2$, $\mathcal{R}3$, \mathcal{T}) and several of their variants. The model does respond correctly to changes in design without the need for design-specific parametrisation. I tuned the timescale of the model to reflect changes in hysteresis at different rates of heating / cooling. To investigate the thermodynamic properties of the systems, I developed an umbrella sampling scheme to bias simulations and calculate the free energy profiles of the origamis. I found design-specific free energy barriers that lead to slow dynamics and used this to explain the variations in hysteresis across the origamis. Kinetically trapped states due to binding of two or more copies of the same staple were found to be transient in these simulations. In the future, we could focus on including other types of misbinding due to partial sequence complementarity into the model. More accurate 2D free energy profiles could also provide some insight into the role of kinetically trapped structures.

We then introduced a novel origami system that relies on the folding set of staples ($\mathcal{R}2$) displacing oligos on a pre-coated scaffold strand (\mathcal{E}). In traditional origami, thermal annealing protocols aid assembly by slowly modulating the strength of interactions over time. This allows the designed interactions to occur at temperatures under which the misbound states are still transient. The structure is nearly fully formed by the time kinetic trapping becomes persistent. Using our coating method, we were able to substantially lower the temperature of error-free isothermal assembly. To simulate the system, I extended the model to account for strand exchange transitions between the two sets of staples. The predictions of the model are surprisingly accurate, especially given the lack of misbinding due to partial-sequence complementarity. Biased simulations revealed the existence of a nucleation barrier and predicted the folding time could be substantially reduced by omitting coating strands that compete with seam staples. This prediction was confirmed in

experiment. Future work could focus on including more than two sets of staples in the model. This would allow us to simulate competition between two distinct shapes, one of which could be selected through strategic omission of coating strands.

Appendices



DNA Sequences

Strand sequences for the T-motifs and T-junction origami can be found in Ref. [115] and Ref. [116], respectively. Here, I show the scaffold strands used in chapters 3-5.

A.1 Scaffold Strands

The two scaffold strand sequences shown here were used in simulations using the domain-level model. More details on strand sequences used in chapter 5 can be found in Ref. [127].

A.1.1 pKD1

GACGAAAGGGCCTCGTGATACGCCTATTTTTATAGGTTAATGTCATGATAATAATG
 GTTTCTTAGACGTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTG
 TTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGAT
 AAATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTC
 GCCCTTATTCCCTTTTTTTGCGGCATTTTGCCCTTCCTGTTTTTTGCTCACCCAGAAAC
 GCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTACATC
 GAACTGGATCTCAACAGCGGTAAAGATCCTTGAGAGTTTTTCGCCCCGAAGAACGTT
 TTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATT
 GACGCCGGGCAAGAGCAACTCGGTGCGCGCATACACTATTCTCAGAATGACTTGG
 TTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGA
 ATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGA
 CAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCA
 TGTAACCTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGAC
 GAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAA
 CTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGC
 GGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATT
 GCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGG
 GGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGC
 AACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAG
 CATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTTAAAACT
 TCATTTTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTTGATAATCTCATGACCA
 AAATCCCTTAACGTGAGTTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATC
 AAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAAACAAA
 AAAACCAACGCTACACGCGGTGGTTTTGTTTGGCGGATCAAGAGCTACCAACTCTTT
 TTCCGAAGGTAACCTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTG
 TAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGC
 TCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCG
 GGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTGCGGCTGAACGGG
 GGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATAC
 CTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACA
 GGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGG
 GGGAAACGCCTGGTATCTTTATAGTCCTGTGCGGTTTTCGCCACCTCTGACTTGAGC
 GTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAA
 CGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTT
 CTGCGTTATCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGAT
 ACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGG
 AAGAGCGCCCAATACGCAAACCGCCTCTCCCCGCGGTTGGCCGATTCAATTAATGC
 AGCTGGCACGACAGGTTTCCCGACTGGAAGCGGGCAGTGAGCGCAACGCAATTA
 ATGTGAGTTAGCTCACTCATTAGGCACCCAGGCTTTACACTTTATGCTTCCGGCT
 CGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTACACAGGAAACAGCTATG
 ACCATGATTACGCCAAGCTCCTCAGCAATTCAGTGGCCGTCGTTTTACAACGTCGT
 GACTGGGAAAACCCCTGGCGTTACCCAACCTTAATCGCCTTGACGACATCCCCCTTT
 CGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTG
 CGCAGCCTGAATGGCGAATGGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTG
 CGGTATTTACACCGCATATGGTGCACCTCTCAGTACAATCTGCTCTGATGCCGCAT
 AGTTAAGCCAGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGT
 CTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTG
 TCAGAGGTTTTACCGTCATCACCGAAACGCGCGA

A.1.2 pJB101

Lower case letters represent portions that differ from pKD1.

```
gctgaggccttccaacagttgcgagcctgaatggcgaatggcgctgatgcggtatcttccttacgcatctgtgcggtatttc
acaccgcatacgtcaaagcaaccatagtagcgccctgtagcggcgcattaagcgcggcggtgtggtggttacgcgcagcgtgaccg
ctacacttgccagcgcctagcggcgctcctttcgcttttctcccttcttctcgccacgttcgcccgtttccccgtaagctctaaatc
gggggctcccttagggttccgatttagtgcttacggcacctcgaccccaaaaaacttgatttgggtgatggttcacgtagtgggccatc
gccctgatagacggttttgcctttgacgttgagtcacgttctttaatagtggaactctgttccaaactggaacaacactcaaccct
atctcgggctattcttttgattataagggatttgcgatttcggcctattggttaaaaaatgagctgatttaacaaaaatttaacgga
attttaacaaaatattaacgtttacaattttATGGTGCACCTCTCAGTACAATCTGCTCTGATGCCGCA
TAGTTAAGCCAGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTG
TCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGT
GTCAGAGGTTTTTCACCGTCATCACCGAAACGCGCGAGACGAAAGGGCCTCGTGAT
ACGCCTATTTTTATAGGTAAATGTCATGATAATAATGGTTTTCTTAGACGTCAGGTG
GCACTTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTCTAAATACAT
TCAAATATGTATCCGCTCATGAGACAATAACCCCTGATAAAATGCTTCAATAATATTG
AAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTGCGCCCTTATTCCTTTTTTG
CGGCATTTTGCCTTCCTGTTTTTGTCTACCCAGAAACGCTGGTGAAAGTAAAAGAT
GCTGAAGATCAGTTGGGTGCACGAGTGGGTACATCGAACTGGATCTCAACAGCG
GTAAGATCCTTGAGAGTTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTT
AAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAAC
TCGGTCCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCCAGTCACA
GAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAA
CCATGAGTGATAAACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAG
GAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACCTCGCCTTGATCGTTG
GGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCCAGATGCCT
GTAGCAATGGCAACAACGTTGCGCAAACCTATTAACCTGGCGAACTACTTACTCTAGC
TTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTT
CTGCGCTCGGCCCTTCCGGCTGGCTGGTTTTATTGCTGATAAATCTGGAGCCGGTG
AGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCG
TATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGA
CAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACCTGTCAGACCAAG
TTTACTCATATATACTTTAGATTGATTTAAACTTTCATTTTTTAATTTAAAGGATC
TAGGTGAAGATCCTTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTT
GTTCCACTGAGCGTCAGACCCCCGTAGAAAAGATCAAAGGATCCTTCTTGAGATCCTT
TTTTTCTGCGCGTAATCTGCTGCTTGCAAAACAAAAAACACCCGCTACCAGCGGTG
GTTTGTGTGCGGATCAAGAGCTACCAACTCTTTTTTCCGAAGGTAACCTGGCTTCAG
CAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACT
TCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTG
GCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGT
TACCGGATAAGGCGCAGCGGTGCGGGCTGAACGGGGGGTTTCGTGCACACAGCCAG
CTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAA
AGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGG
TCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAACGCCTGGTATCTTTA
TAGTCCTGTGCGGTTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGT
CAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTTACGGTTCCT
GGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCT
```


B

Fitting Staple Data

In Figure B.1, I outline the procedure used to extract T_{50} and n_H values for individual staples.

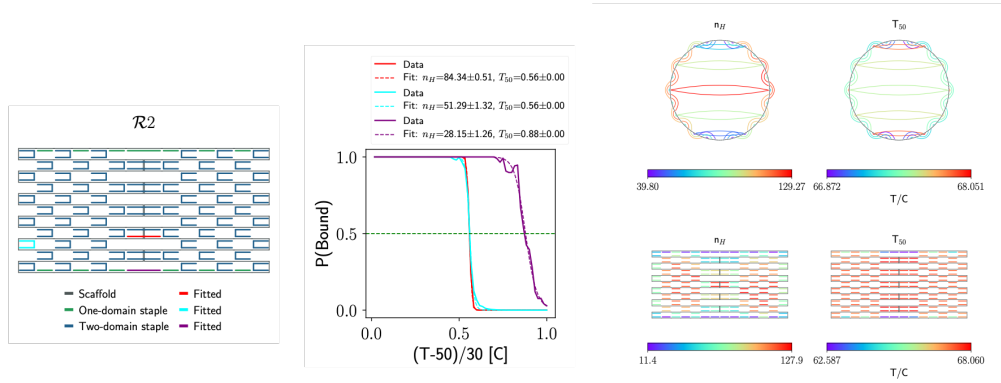


Figure B.1: Procedure used to construct staple heat maps. Occupation probability of each domain/crossover as a function of temperature is averaged over ~ 100 annealing cycles and fitted with a Hill function to extract T_{50} and n_H values.

The parameter T_{50} gives the temperature at which the domain or crossover is expected to be bound with 50% probability and the parameter n_H is a measure cooperativity of the staple with other staples in the system.

C

Appendix to Chapter 4

C.1 Temperature Dependence of Two-Row Systems

Figure C.1 shows the relative probabilities of staples being bound at each temperature for the three systems discussed in the main text.

Figure C.2 shows the change in free energy landscapes as model parameters are varied but the temperature is kept constant. In Figure 4.2 of the main text, the temperature was adjusted to the melting temperature of the origami.

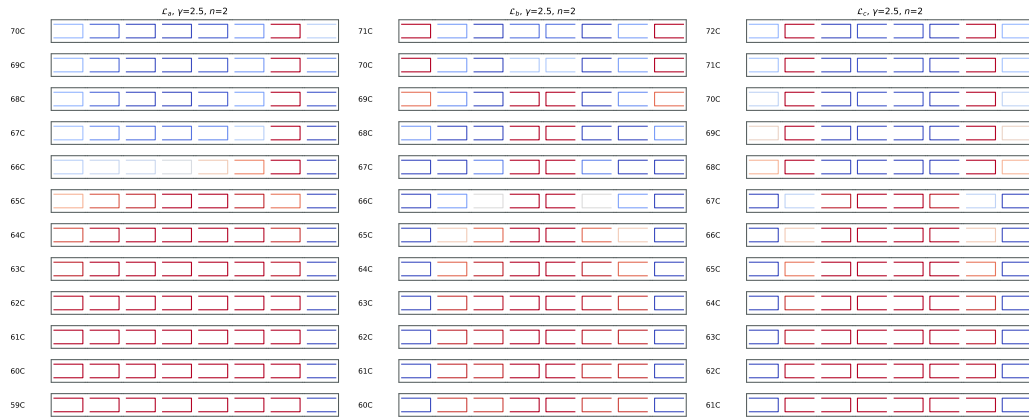


Figure C.1: Relative probability of staples being fully bound at temperatures around the melting temperature of the three origamis \mathcal{L}_a , \mathcal{L}_b and \mathcal{L}_c . Staples with highest probability of binding are coloured red and with least probability coloured blue at each temperature. Note that colours at different temperatures correspond to different probabilities.

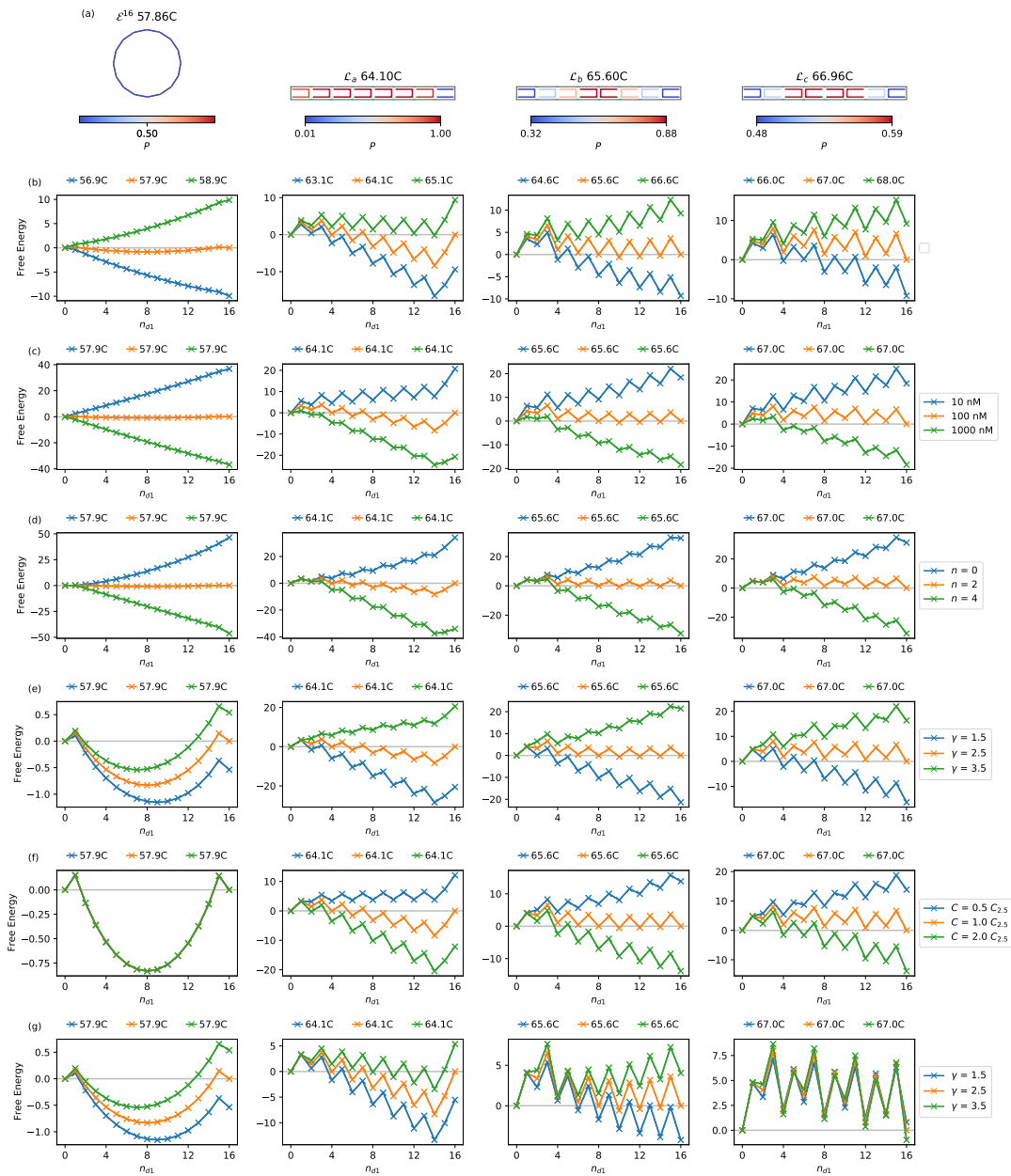


Figure C.2: Temperature is varied in Row (a) using default parameter set. Other rows are equivalent of Figure 4.2 with temperature kept constant.

C.2 Open Scaffold Systems

For direct comparison to Ref. [152], I recently modified the code to calculate the exact probability distributions for open scaffold systems (other scaffolds considered here are all closed single stranded loops). In Appendix C.2, I present three such systems: the “Two-Row” system is the same as that in Ref. [152], and the “Three-RowS” system is a smaller version of the “Three-Row” system of Ref. [152]. The Two-Row system behaves similarly to \mathcal{L}_c , with adjacent staples with double-crossovers defining the energy profile. States with multiple partly bound staples are rare, as evidenced by the reversion of the $\Delta F(n_{d1})$ to $\Delta F(n_{st1})$ in steps of two domains. However, these states are much more likely in the Three-row system, leading to a much larger barrier as a function of domains than staples.

I have not implemented the C++ code to simulate open systems in the SSA framework. Therefore, I have not tested these systems under temperature ramps. Currently, there is a simple python implementation of the code to evaluate the probability distribution of small open systems. In the future, linear scaffolds could be added to the main SSA code.

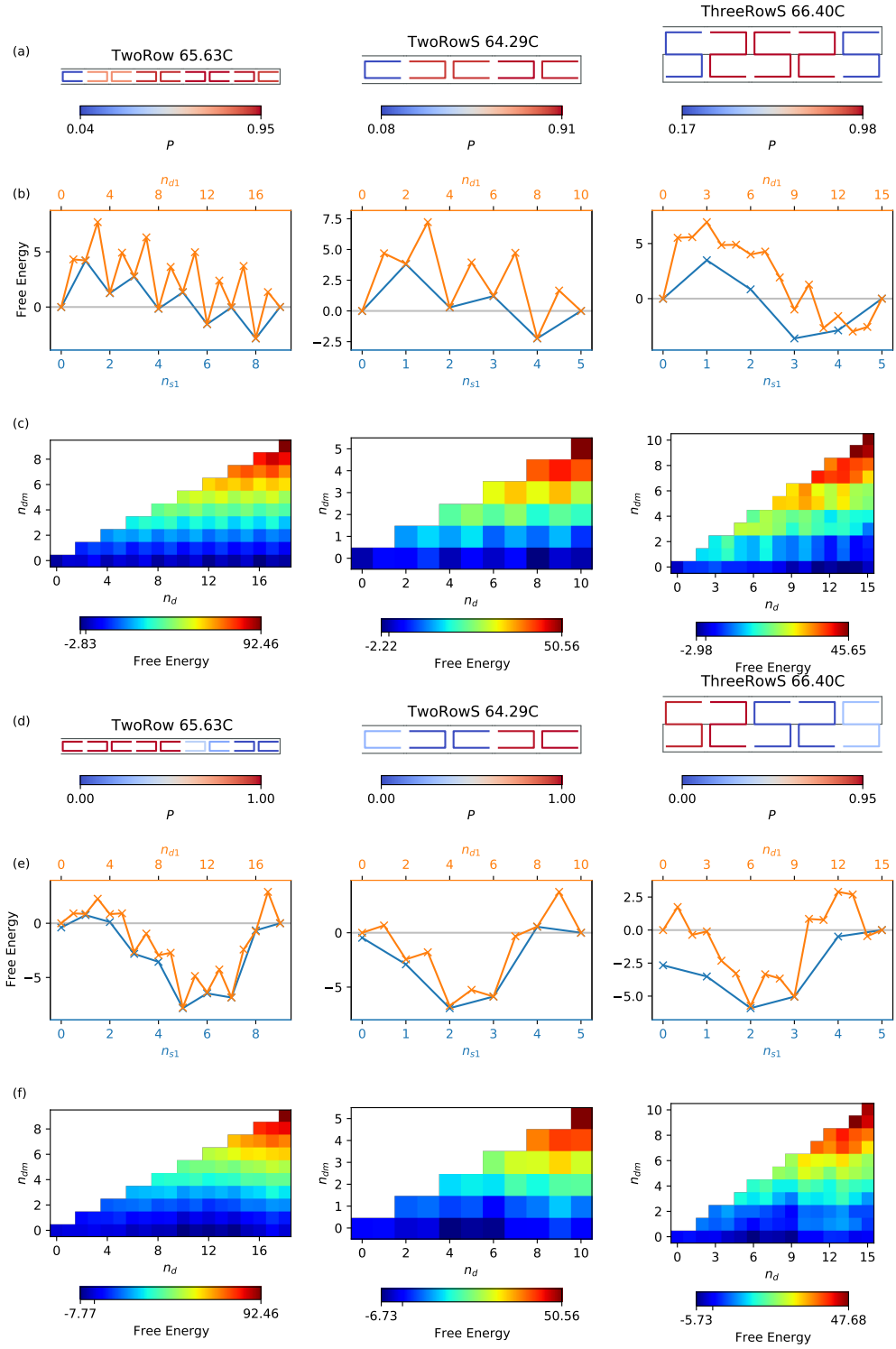


Figure C.3: Equivalent of Figure 4.1 for open-scaffold systems.

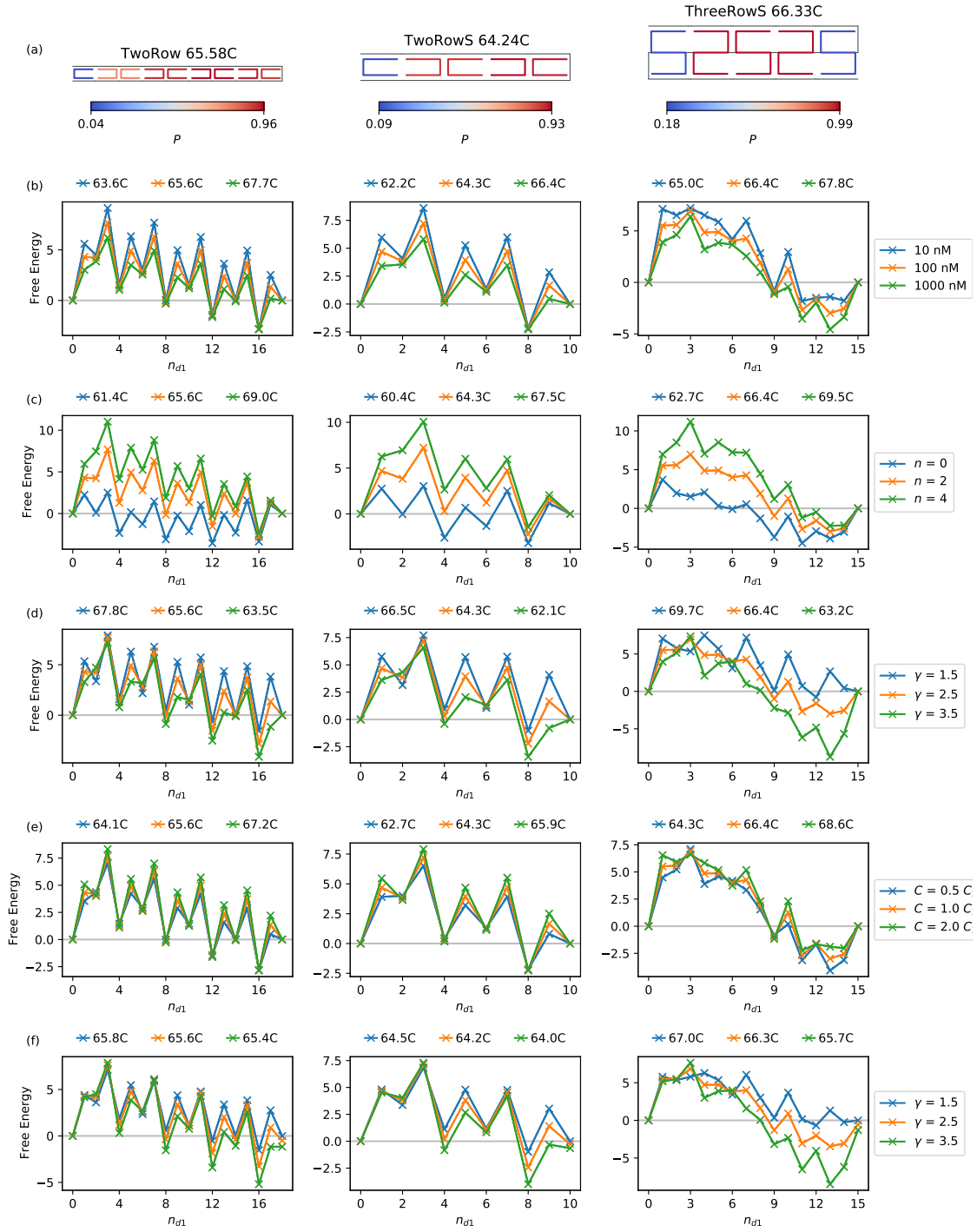


Figure C.4: Equivalent of Figure 4.2 for open-scaffold systems.

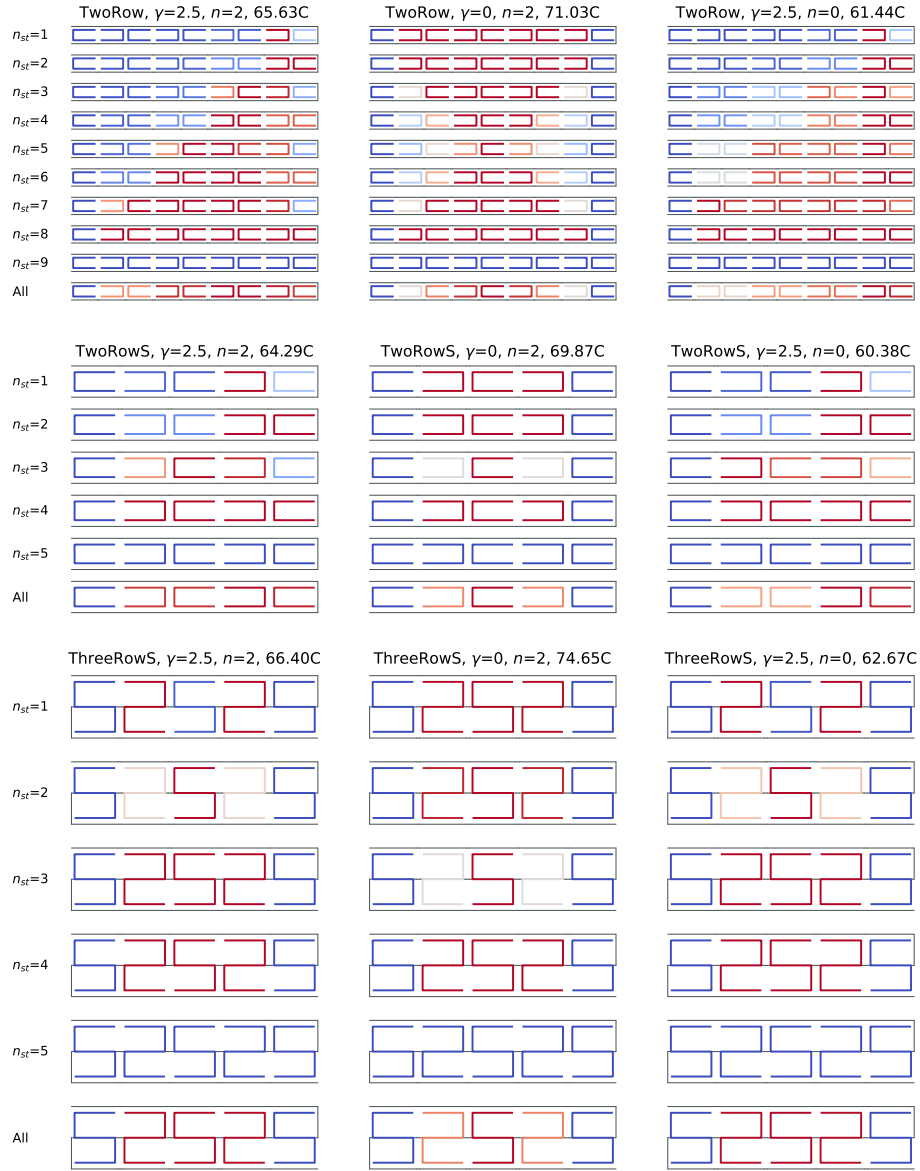


Figure C.5: Equivalent of Figure 4.3 for open-scaffold systems.

C.3 Four-Helix Systems

Reference figures for discussions in the main text. These 4-row systems have some features of the larger systems that were simulated.

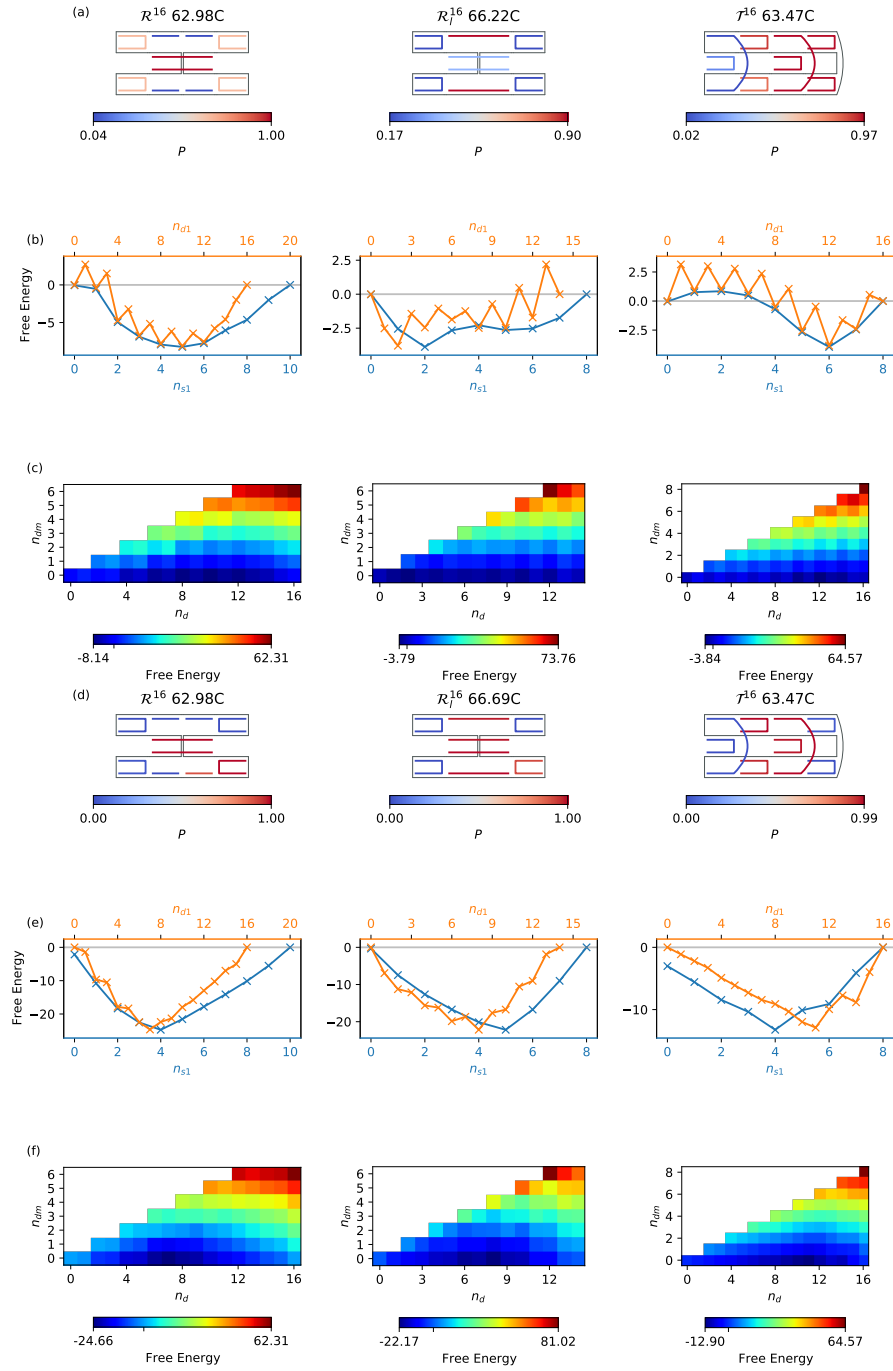


Figure C.6: Equivalent of Figure 4.1 for 4-helix systems.

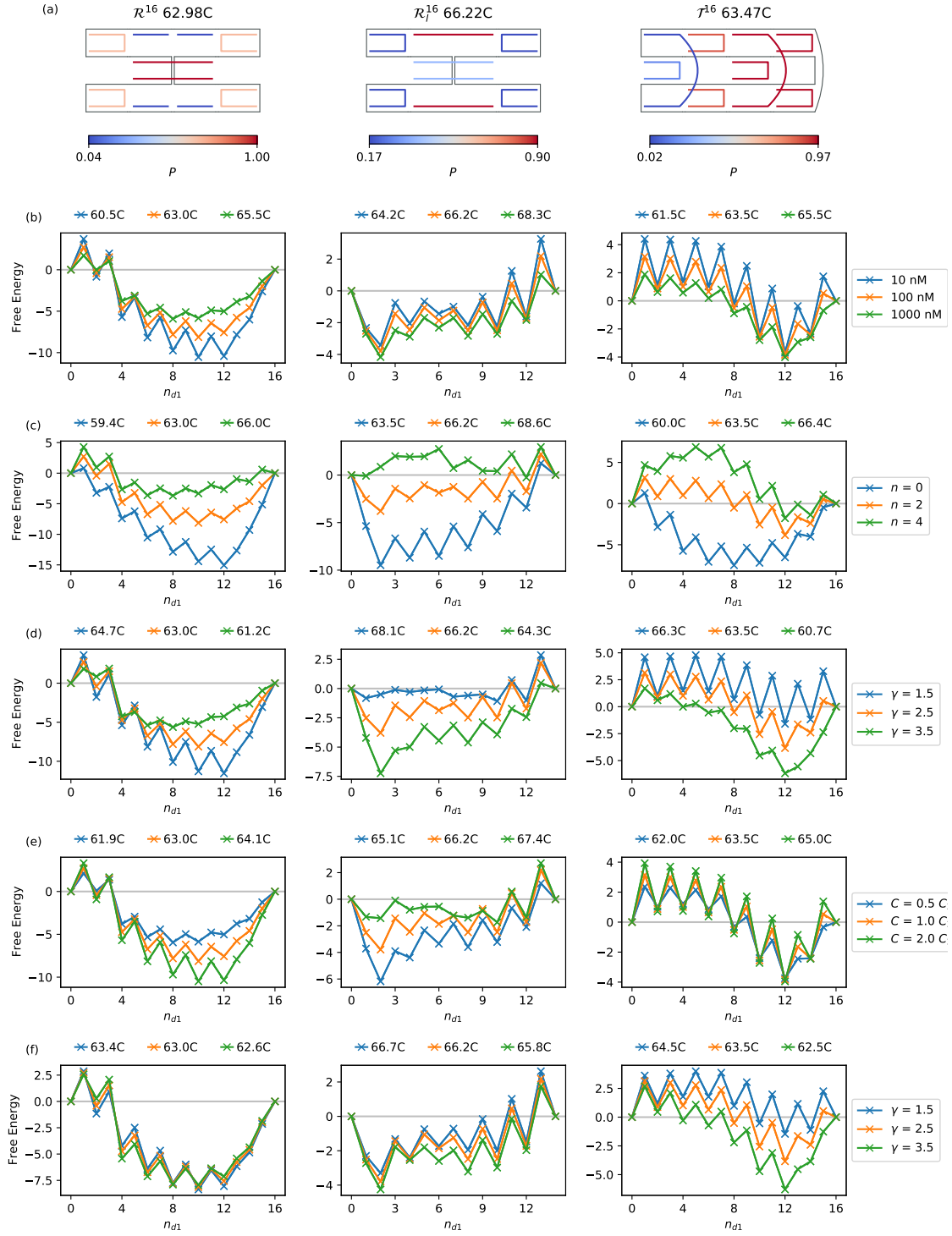


Figure C.7: Equivalent of Figure 4.2 for 4-helix systems.

C.4 Seam Variants of $\mathcal{R}2$

These figures show more detail to accompany the discussion of Figure 4.9 in the main text.

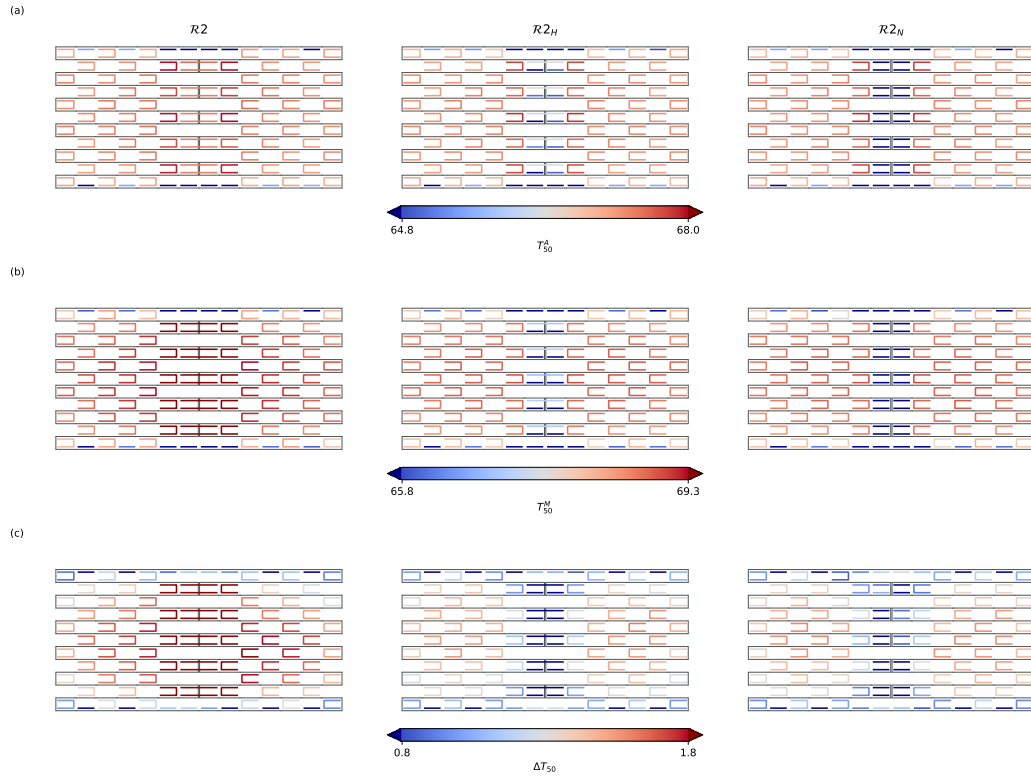


Figure C.8: T_{50}^A , T_{50}^M and $\Delta T_{50} = T_{50}^A - T_{50}^M$ are shown in top, middle, and bottom rows for seam-variants of $\mathcal{R}2$ (left), $\mathcal{R}2_H$ (middle) and $\mathcal{R}2_N$ (right).

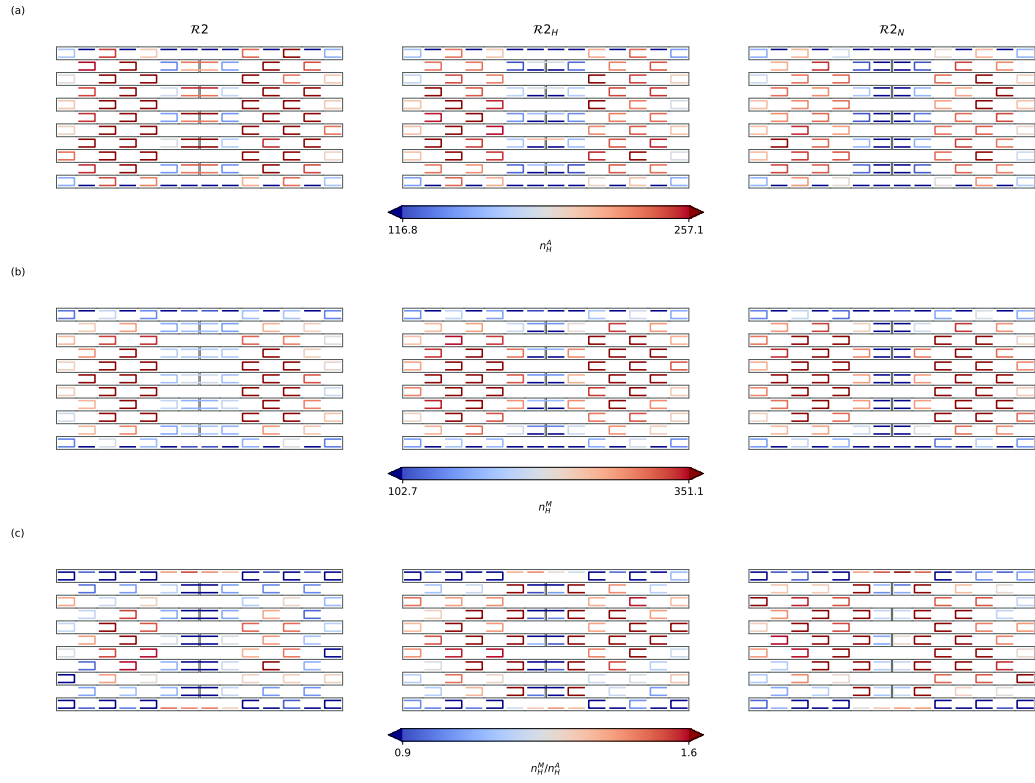


Figure C.9: Hill coefficients n_H^A , n_H^M and n_H^M/n_H^A are shown in top, middle, and bottom rows for seam-variants of $\mathcal{R}2$ (left), $\mathcal{R}2_H$ (middle) and $\mathcal{R}2_N$ (right).

C.5 Sequence-dependent Staple Data

These figures show the sequence-specific equivalents of the average-sequence results presented in the main text. Sequence-specificity often masks other important effects; using sequence-averaged parametrisation of the model allows us to discuss the important features and underlying physics more easily.

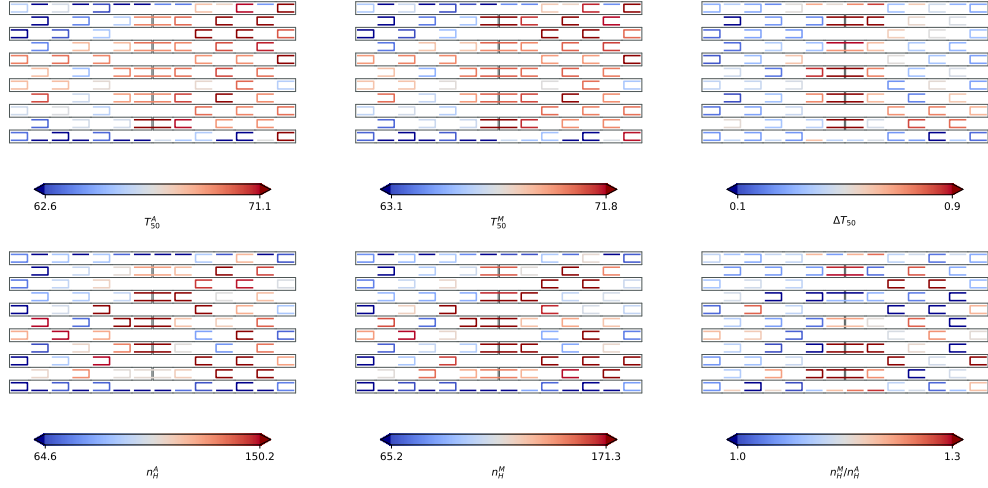


Figure C.10: Equivalent of Figure 4.6 but using sequence of pKD scaffold.

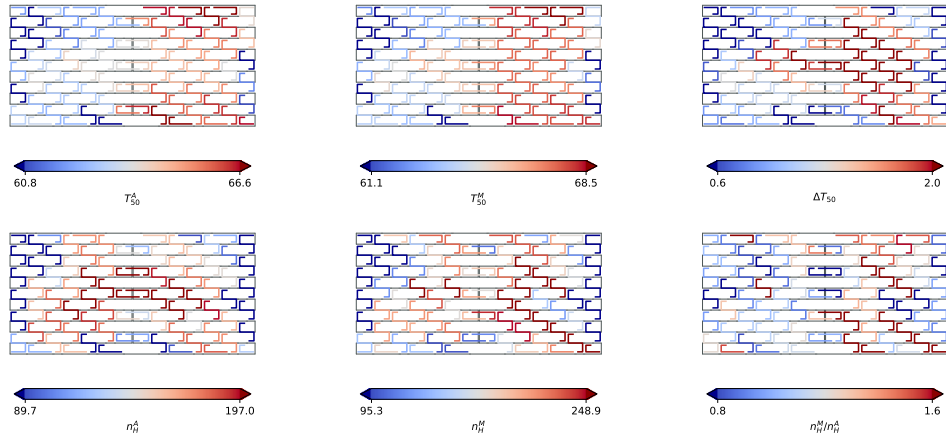


Figure C.11: Equivalent of Figure 4.11 but using sequence of pKD scaffold.

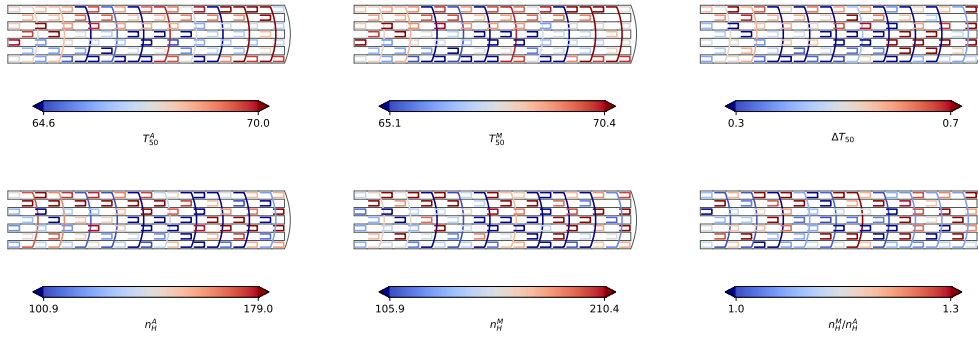


Figure C.12: Equivalent of Figure 4.15 but using sequence of pKD scaffold.

C.6 Domain Distribution of $\mathcal{R}3$

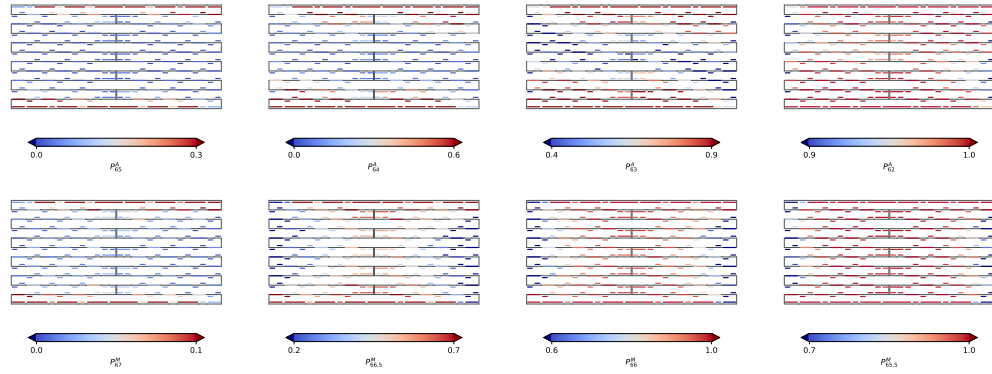


Figure C.13: Probability that each domain is bound at various temperatures during annealing (top) and melting (bottom).

This figure is shown for comparison to Figure 4.12 in the main text. It shows that while staples have domains of different sizes, folding and unfolding mainly proceeds at the level of staples. ie. all domains of a single staple bind / unbind at roughly the same time.

D

Appendix to Chapter 5

D.1 Domain Energies

Melting temperature of individual domains in bulk solution calculated using the two state model of SantaLucia:

$$T_{\text{NN}}^{\text{Melt}} = \frac{\Delta H_{\text{NN}}^{\ominus}}{\Delta S_{\text{NN}}^{\ominus} + R \ln ([A]_0 - [B]_0/2)}, \quad (\text{D.1})$$

where $\Delta H_{\text{NN}}^{\ominus}$ and $\Delta S_{\text{NN}}^{\ominus}$ are enthalpic and entropic nearest neighbour contributions, and $[A]_0$ and $[B]_0$ are strand concentrations. I used equal strand concentration of 100nM for simplicity. The aim is to identify regions of the scaffold with weak and strong interactions when different staple sets are used. Four nucleation sites can be identified using the method in Figure D.1. These sites appear in alphabetic order with increasing temperature in isothermal simulations (Figure D.2). Interestingly, site B is occupied later in the process at 60°C because both the $\mathcal{R}2$ and \mathcal{E} domain interactions are weak at this site.

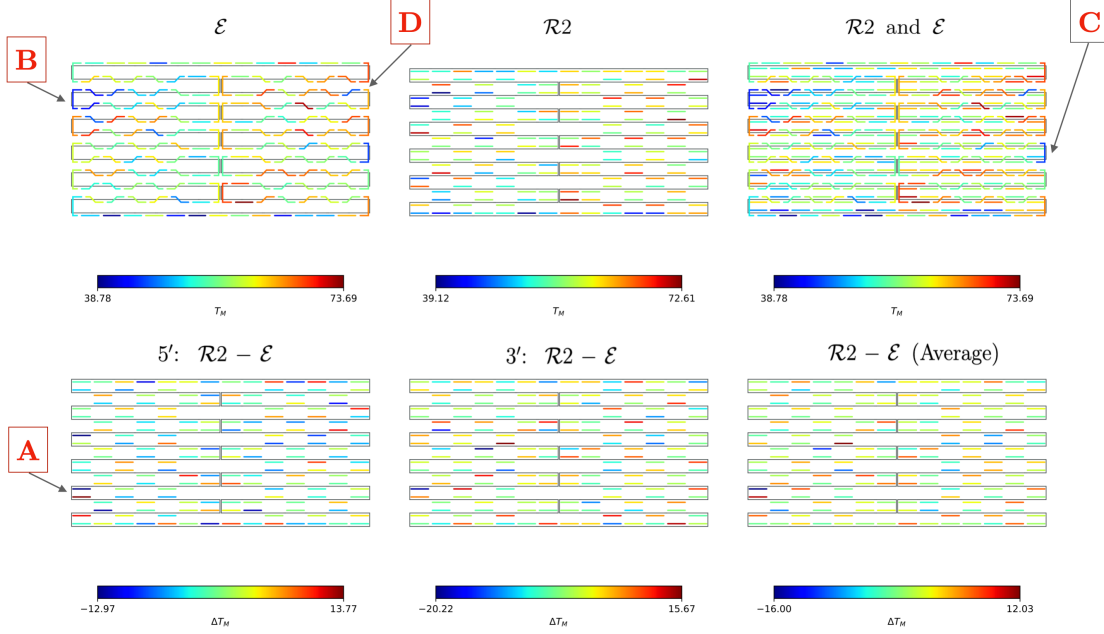


Figure D.1: Top left: melting temperatures of \mathcal{E} domains. Top middle: melting temperatures of $\mathcal{R}2$ domains. Top right: melting temperatures of $\mathcal{R}2$ and \mathcal{E} domains shown together. Bottom left: difference between T_M of $\mathcal{R}2$ domain and its 5'-adjacent competing \mathcal{E} domain. Bottom middle: same as bottom left but 3' case. Bottom right: Average of differences calculated last two cases. Four nucleation sites that appear in the order A, B, C, D as the temperature is raised are marked (see Figure D.2).

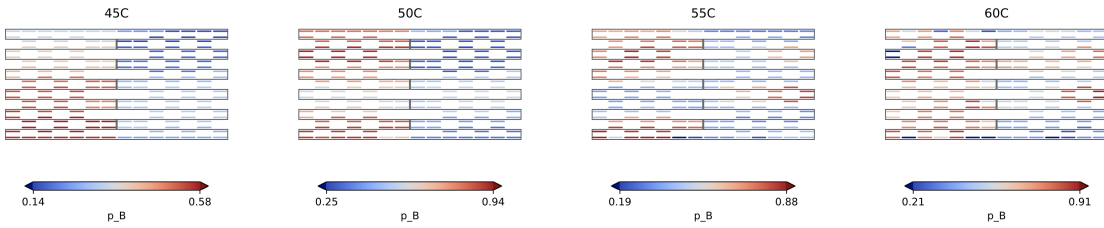


Figure D.2: Proportion of simulation time spend in the bound state by each domain is shown at 45°C, 50°C, 55°C and 60°C. The red regions bind first. The four nucleation sites identified in the Figure D.1 correspond to the red regions in this figure. Results for all temperatures below 45°C shows the same distribution as that of 45°C, nucleating at site A alone. As the temperature is raised, sites B, C, and D sequentially become viable nucleation sites.

References

- [1] R. Hadgiivanova and H. Diamant, “Kinetics of surfactant micellization: A free energy approach,” *The Journal of Physical Chemistry*, 2010.
- [2] W. Jacobs and D. Frenkel, “Self-assembly of structures with addressable complexity,” *Journal of the American Chemical Society*, 2016.
- [3] K. A. Dill and M. J. L., “The protein-folding problem, 50 years on.,” *Science*, vol. 338, no. 6110, pp. 1042–6, 2012.
- [4] H. Vehkamäki, *Classical Nucleation Theory in Multicomponent Systems*. Springer-Verlag Berlin Heidelberg, 2006.
- [5] R. Sear, “Nucleation: Theory and applications to protein solutions and colloidal suspensions,” eng, *Journal of Physics - Condensed Matter*, vol. 19, no. 3, 2007.
- [6] C. Levinthal, “How to fold graciously,” *Mossbauer spectroscopy in biological systems*, vol. 67, pp. 22–24, 1969.
- [7] S. S. Plotkin and J. N. Onuchic, “Understanding protein folding with energy landscape theory part i: Basic concepts,” *Quarterly Reviews of Biophysics*, vol. 35, no. 2, pp. 111–167, 2002.
- [8] H. Miller, Z. Zhou, J. Shepherd, A. J. M. Wollman, and M. C. Leake, “Single-molecule techniques in biophysics: A review of the progress in methods and applications,” *Reports on Progress in Physics*, vol. 81, no. 2, p. 024 601, Dec. 2017. [Online]. Available: <https://dx.doi.org/10.1088/1361-6633/aa8a02>.
- [9] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, *et al.*, “Highly accurate protein structure prediction with alphafold,” *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.
- [10] E. Pfitzner, C. Wachauf, F. Kilchherr, B. Pelz, W. M. Shih, M. Rief, and H. Dietz, “Rigid dna beams for high-resolution single-molecule mechanics,” *Angewandte Chemie International Edition*, vol. 52, no. 30, pp. 7766–7771, 2013. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201302727>.
- [11] P. C. Nickels, B. Wünsch, P. Holzmeister, W. Bae, L. M. Kneer, D. Grohmann, P. Tinnefeld, and T. Liedl, “Molecular force spectroscopy with a DNA origami-based nanoscopic force clamp,” *Science*, vol. 354, no. 6310, pp. 305–307, 2016.
- [12] N. C. Seeman, “Nucleic acid junctions and lattices,” *J. Theor. Biol.*, vol. 99, no. 2, pp. 237–247, 1982.
- [13] N. C. Seeman and N. R. Kallenbach, “Design of immobile nucleic acid junctions,” *Biophys. J.*, vol. 44, no. 2, pp. 201–209, 1983.

- [14] N. C. Seeman and H. F. Sleiman, "Dna nanotechnology," *Nature Reviews Materials*, vol. 3, no. 1, pp. 1–23, 2017.
- [15] M. Madsen and K. V. Gothelf, "Chemistries for dna nanotechnology," *Chemical reviews*, vol. 119, no. 10, pp. 6384–6458, 2019.
- [16] J. D. Watson and F. H. C. Crick, "Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids," *Nature*, vol. 171, no. 4356, pp. 738–740, 1953.
- [17] K. Hoogsteen, "The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine," *Acta Crystallographica*, vol. 16, no. 9, pp. 907–916, 1963.
- [18] [Online]. Available: en.wikipedia.org/wiki/DNA.
- [19] M. D. Frank-Kamenetskii, "Biophysics of the dna molecule," *Physics Reports*, vol. 288, no. 1-6, pp. 13–60, 1997.
- [20] A. V. Pinheiro, D. Han, W. M. Shih, and H. Yan, "Challenges and opportunities for structural dna nanotechnology," *Nature nanotechnology*, vol. 6, no. 12, pp. 763–772, 2011.
- [21] [Online]. Available: https://en.wikipedia.org/wiki/DNA_nanotechnology.
- [22] N. C. Seeman, "Nanotechnology and the double helix," *Scientific American*, vol. 290, no. 6, pp. 64–75, 2004.
- [23] R. Holliday, "A mechanism for gene conversion in fungi," *Genetics Research*, vol. 5, no. 2, pp. 282–304, 1964.
- [24] H. Potter and D. Dressler, "On the mechanism of genetic recombination: Electron microscopic observation of recombination intermediates.," *Proceedings of the National Academy of Sciences*, vol. 73, no. 9, pp. 3000–3004, 1976.
- [25] Y. Wang, J. E. Mueller, B. Kemper, and N. C. Seeman, "Assembly and characterization of five-arm and six-arm dna branched junctions," *Biochemistry*, vol. 30, no. 23, pp. 5667–5674, 1991.
- [26] X. Wang and N. C. Seeman, "Assembly and characterization of 8-arm and 12-arm dna branched junctions," *Journal of the American Chemical Society*, vol. 129, no. 26, pp. 8169–8176, 2007.
- [27] J. H. Chen, N. R. Kallenbach, and N. C. Seeman, "A specific quadrilateral synthesized from dna branched junctions," *Journal of the American Chemical Society*, vol. 111, no. 16, pp. 6402–6407, 1989.
- [28] T. J. Fu and N. C. Seeman, "DNA double-crossover molecules," *Biochemistry*, vol. 32, no. 13, pp. 3211–3220, 1993.
- [29] T. H. LaBean, H. Yan, J. Kopatsch, F. Liu, E. Winfree, J. H. Reif, and N. C. Seeman, "Construction, analysis, ligation, and self-assembly of dna triple crossover complexes," *Journal of the American Chemical Society*, vol. 122, no. 9, pp. 1848–1860, 2000.
- [30] Z. Shen, H. Yan, T. Wang, and N. C. Seeman, "Paranemic crossover dna: A generalized holliday structure with applications in nanotechnology," *Journal of the American Chemical Society*, vol. 126, no. 6, pp. 1666–1674, 2004.

- [31] E. Winfree, F. Liu, L. A. Wenzler, and N. C. Seeman, "Design and self-assembly of two-dimensional DNA crystals," *Nature*, vol. 394, no. 6693, p. 539, 1998.
- [32] Y. He, Y. Tian, A. E. Ribbe, and C. Mao, "Highly connected Two-Dimensional crystals of DNA Six-Point-Stars," *J. Am. Chem. Soc.*, vol. 128, no. 50, pp. 15 978–9, 2006.
- [33] J. Zheng, J. J. Birktoft, Y. Chen, T. Wang, R. Sha, P. E. Constantinou, S. L. Ginell, C. Mao, and N. Seeman, "From molecular to macroscopic via the rational design of a self-assembled 3D DNA crystal," *Nature*, vol. 461, p. 74, 2009.
- [34] W. M. Shih, J. D. Quispe, and G. F. Joyce, "A 1.7-kilobase single-stranded dna that folds into a nanoscale octahedron," *Nature*, vol. 427, no. 6975, pp. 618–621, 2004. [Online]. Available: <https://doi.org/10.1038/nature02307>.
- [35] R. P. Goodman, I. A. T. Schaap, C. F. Tardin, C. M. Erben, R. M. Berry, C. F. Schmidt, and A. J. Turberfield, "Rapid chiral assembly of rigid DNA building blocks for molecular nanofabrication," *Science*, vol. 310, no. 5754, pp. 1661–1665, 2005.
- [36] C. Zhang, M. Su, Y. He, X. Zhao, P. Fang, A. E. Ribbe, W. Jiang, and C. Mao, "Conformational flexibility facilitates self-assembly of complex DNA nanostructures," *Proc. Natl. Acad. Sci. USA*, vol. 105, no. 31, pp. 10 665–10 669, 2008.
- [37] Y. He, M. Su, P. Fang, C. Zhang, A. E. Ribbe, W. Jiang, and C. Mao, "On the chirality of self-assembled DNA octahedra," *Angew. Chem. Int. Ed.*, vol. 49, pp. 748–751, 2010.
- [38] P. W. Rothmund, A. Ekani-Nkodo, N. Papadakis, A. Kumar, D. K. Fygenson, and E. Winfree, "Design and characterization of programmable dna nanotubes," *Journal of the American Chemical Society*, vol. 126, no. 50, pp. 16 344–16 352, 2004.
- [39] J. Bath and A. J. Turberfield, "Dna nanomachines," *Nature nanotechnology*, vol. 2, no. 5, pp. 275–284, 2007.
- [40] B. Yurke, A. J. Turberfield, A. P. Mills, F. C. Simmel, and J. L. Neumann, "A dna-fuelled molecular machine made of dna," *Nature*, vol. 406, no. 6796, pp. 605–608, 2000.
- [41] J.-S. Shin and N. A. Pierce, "A synthetic dna walker for molecular transport," *Journal of the American Chemical Society*, vol. 126, no. 35, pp. 10 834–10 835, 2004.
- [42] P. Yin, H. Yan, X. G. Daniell, A. J. Turberfield, and J. H. Reif, "A unidirectional dna walker that moves autonomously along a track," *Angewandte Chemie*, vol. 116, no. 37, pp. 5014–5019, 2004.
- [43] C. Mao, T. H. LaBean, J. H. Reif, and N. C. Seeman, "Logical computation using algorithmic self-assembly of dna triple-crossover molecules," *Nature*, vol. 407, no. 6803, pp. 493–496, 2000.
- [44] P. W. K. Rothmund, N. Papadakis, and E. Winfree, "Algorithmic self-assembly of dna sierpinski triangles," *PLoS biology*, vol. 2, no. 12, e424, 2004.

- [45] P. W. K. Rothemund, "Folding DNA to create nanoscale shapes and patterns," *Nature*, vol. 440, no. 7082, pp. 297–302, 2006. [Online]. Available: <http://dx.doi.org/10.1038/nature04586>.
- [46] S. M. Douglas, H. Dietz, T. Liedl, B. Högberg, F. Graf, and W. M. Shih, "Self-assembly of DNA into nanoscale three-dimensional shapes," *Nature*, vol. 459, no. 7245, pp. 414–418, 2009. [Online]. Available: <http://dx.doi.org/10.1038/nature08016>.
- [47] H. Dietz, S. M. Douglas, and W. M. Shih, "Folding DNA into twisted and curved nanoscale shapes," *Science*, vol. 325, pp. 725–730, 2009.
- [48] E. S. Andersen, M. Dong, M. M. Nielsen, K. Jahn, R. Subramani, W. Mamdough, M. M. Golas, B. Sander, H. Stark, C. L. Oliveira, *et al.*, "Self-assembly of a nanoscale dna box with a controllable lid," *Nature*, vol. 459, no. 7243, pp. 73–76, 2009.
- [49] D. Han, S. Pal, J. Nangreave, Z. Deng, Y. Liu, and H. Yan, "Dna origami with complex curvatures in three-dimensional space," *Science*, vol. 332, no. 6027, pp. 342–346, 2011.
- [50] E. Benson, A. Mohammed, J. Gardell, S. Masich, E. Czeizler, P. Orponen, and B. Högberg, "Dna rendering of polyhedral meshes at the nanoscale," *Nature*, vol. 523, no. 7561, pp. 441–444, 2015.
- [51] Y. Ke, S. M. Douglas, M. Liu, J. Sharma, A. Cheng, A. Leung, Y. Liu, W. M. Shih, and H. Yan, "Multilayer dna origami packed on a square lattice," *Journal of the American Chemical Society*, vol. 131, no. 43, pp. 15 903–15 908, 2009.
- [52] Y. Ke, N. V. Voigt, K. V. Gothelf, and W. M. Shih, "Multilayer dna origami packed on hexagonal and hybrid lattices," *Journal of the American Chemical Society*, vol. 134, no. 3, pp. 1770–1774, 2012.
- [53] J.-P. J. Sobczak, T. G. Martin, T. Gerling, and H. Dietz, "Rapid folding of dna into nanoscale shapes at constant temperature," *Science*, vol. 338, no. 6113, pp. 1458–1461, 2012.
- [54] J. Song, J.-M. Arbona, Z. Zhang, L. Liu, E. Xie, J. Elezgaray, J.-P. Aime, K. V. Gothelf, F. Besenbacher, and M. Dong, "Direct visualization of transient thermal response of a dna origami," *Journal of the American Chemical Society*, vol. 134, no. 24, pp. 9844–9847, 2012.
- [55] L. J. Wah, C. David, S. Rudiuk, and D. Baigl, "Observing and controlling the folding pathway of DNA origami at the nanoscale," *ACS NANO*, 2016.
- [56] X. Wei, J. Nangreave, S. Jiang, H. Yan, and Y. Liu, "Mapping the thermal behavior of dna origami nanostructures," *Journal of the American Chemical Society*, vol. 135, no. 16, pp. 6165–6176, 2013.
- [57] F. Schneider, N. Möritz, and H. Dietz, "The sequence of events during folding of a dna origami," *Science advances*, vol. 5, no. 5, eaaw1412, 2019.
- [58] K. E. Dunn, F. Dannenberg, T. E. Ouldridge, M. Kwiatkowska, A. J. Turberfield, and J. Bath, "Guiding the folding pathway of DNA origami," *Nature*, vol. 525, no. 7567, pp. 82–86, 2015.

- [59] F. Dannenberg, K. E. Dunn, J. Bath, M. Kwiatkowska, A. J. Turberfield, and T. E. Ouldridge, “Modelling DNA Origami Self-Assembly at the Domain Level,” *arXiv*, 2015. eprint: 1509.03066.
- [60] Y. Ke, L. L. Ong, W. M. Shih, and P. Yin, “Three-Dimensional structures Self-Assembled from DNA bricks,” *Science*, vol. 338, no. 6111, pp. 1177–1183, 2012.
- [61] B. Wei, M. Dai, and P. Yin, “Complex shapes self-assembled from single-stranded DNA tiles,” *Nature*, vol. 485, no. 7400, pp. 623–626, 2012.
- [62] B. Wei, M. Dai, C. Myhrvold, T. Ke, R. Jungmann, and P. Yin, “Design space for complex dna structures,” *Journal of the American Chemical Society*, vol. 135, no. 48, pp. 18 080–18 088, 2013.
- [63] Y. Ke, L. L. Ong, W. Sun, J. Song, M. Dong, W. M. Shih, and P. Yin, “Dna brick crystals with prescribed depths,” *Nature chemistry*, vol. 6, no. 11, pp. 994–1002, 2014.
- [64] L. L. Ong, N. Hanikel, O. K. Yaghi, C. Grun, M. T. Strauss, P. Bron, J. Lai-Kee-Him, F. Schueder, B. Wang, P. Wang, *et al.*, “Programmable self-assembly of three-dimensional nanostructures from 10,000 unique components,” *Nature*, vol. 552, no. 7683, pp. 72–77, 2017.
- [65] D. Frenkel, “Order through entropy,” *Nature materials*, vol. 14, no. 1, pp. 9–12, 2015.
- [66] W. M. Jacobs, A. Reinhardt, and D. Frenkel, “Theoretical prediction of free-energy landscapes for complex self-assembly,” *The Journal of chemical physics*, vol. 142, no. 2, p. 021 101, 2015.
- [67] A. Reinhardt and D. Frenkel, “Numerical evidence for nucleated self-assembly of dna brick structures,” *Physical review letters*, vol. 112, no. 23, p. 238 103, 2014.
- [68] A. Reinhardt, C. P. Ho, and D. Frenkel, “Effects of co-ordination number on the nucleation behaviour in many-component self-assembly,” *Faraday Discussions*, vol. 186, pp. 215–228, 2016.
- [69] A. Reinhardt and D. Frenkel, “Dna brick self-assembly with an off-lattice potential,” *Soft Matter*, vol. 12, no. 29, pp. 6253–6260, 2016.
- [70] P. Fonseca, F. Romano, J. S. Schreck, T. E. Ouldridge, J. P. Doye, and A. A. Louis, “Multi-scale coarse-graining for the study of assembly pathways in dna-brick self-assembly,” *The Journal of Chemical Physics*, vol. 148, no. 13, p. 134 910, 2018.
- [71] J. Sponer, J. E. Sponer, A. Mladek, P. Banas, P. Jurecka, and M. Otyepka, “How to understand quantum chemical computations on dna and rna systems,” *Methods*, vol. 64, no. 1, pp. 3–11, 2013.
- [72] C. A. Laughton and S. A. Harris, “The atomistic simulation of dna,” *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1, no. 4, pp. 590–600, 2011.

- [73] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules j. am. chem. soc. 1995, 117, 5179- 5197," *Journal of the American Chemical Society*, vol. 118, no. 9, pp. 2309–2309, 1996.
- [74] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, *et al.*, "Charmm: The biomolecular simulation program," *Journal of computational chemistry*, vol. 30, no. 10, pp. 1545–1614, 2009.
- [75] J. Yoo and A. Aksimentiev, "In situ structure and dynamics of dna origami determined through molecular dynamics simulations," *Proceedings of the National Academy of Sciences*, vol. 110, no. 50, pp. 20 099–20 104, 2013.
- [76] C. Maffeo, J. Yoo, and A. Aksimentiev, "De novo reconstruction of dna origami structures through atomistic molecular dynamics simulation," *Nucleic acids research*, vol. 44, no. 7, pp. 3013–3019, 2016.
- [77] M. F. Hagan, A. R. Dinner, D. Chandler, and A. K. Chakraborty, "Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in dna," *Proceedings of the National Academy of Sciences*, vol. 100, no. 24, pp. 13 922–13 927, 2003.
- [78] M. Rubinstein, *Polymer Physics*. Oxford University Press, 2003.
- [79] O. Kratky and G. Porod, "Röntgenuntersuchung geloster fadenmoleküle," *Recueil des Travaux Chimiques des Pays-Bas*, vol. 68, no. 12, pp. 1106–1122, 1949.
- [80] D.-N. Kim, F. Kilchherr, H. Dietz, and M. Bathe, "Quantitative prediction of 3d solution shape and flexibility of nucleic acid nanostructures," *Nucleic acids research*, vol. 40, no. 7, pp. 2862–2868, 2012.
- [81] S. M. Douglas, A. H. Marblestone, S. Teerapittayanon, A. Vazquez, G. M. Church, and W. M. Shih, "Rapid prototyping of 3d dna-origami shapes with cadnano," *Nucleic acids research*, vol. 37, no. 15, pp. 5001–5006, 2009.
- [82] J. SantaLucia Jr, "A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics," *Proceedings of the National Academy of Sciences*, vol. 95, no. 4, pp. 1460–1465, 1998.
- [83] J. SantaLucia Jr and D. Hicks, "The thermodynamics of dna structural motifs," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 33, pp. 415–440, 2004.
- [84] W. W. Hadiwikarta, J.-C. Walter, J. Hooyberghs, and E. Carlon, "Probing hybridization parameters from microarray experiments: Nearest-neighbor model and beyond," *Nucleic acids research*, vol. 40, no. 18, e138–e138, 2012.
- [85] J. M. Huguet, C. V. Bizarro, N. Forns, S. B. Smith, C. Bustamante, and F. Ritort, "Single-molecule derivation of salt dependent base-pair free energies in dna," *Proceedings of the National Academy of Sciences*, vol. 107, no. 35, pp. 15 431–15 436, 2010.
- [86] J.-M. Arbona, J. Elezgaray, and J.-P. Aime, "Modelling the folding of dna origami," *EPL (Europhysics Letters)*, vol. 100, no. 2, p. 28 006, 2012.

- [87] R. V. Reshetnikov, A. V. Stolyarova, A. O. Zalevsky, D. Y. Panteleev, G. V. Pavlova, D. V. Klinov, A. V. Golovin, and A. D. Protopopova, “A coarse-grained model for dna origami,” *Nucleic acids research*, vol. 46, no. 3, pp. 1102–1112, 2018.
- [88] J. M. Arbona, J.-P. Aimé, and J. Elezgaray, “Modeling the mechanical properties of dna nanostructures,” *Physical Review E*, vol. 86, no. 5, p. 051 912, 2012.
- [89] A. Savelyev and G. A. Papoian, “Chemically accurate coarse graining of double-stranded dna,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 47, pp. 20 340–20 345, 2010.
- [90] J. C. Araque, A. Z. Panagiotopoulos, and M. A. Robert, “Lattice model of oligonucleotide hybridization in solution. i. model and thermodynamics,” *The Journal of chemical physics*, vol. 134, no. 16, 04B616, 2011.
- [91] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, “Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model,” *J. Chem. Phys.*, vol. 134, no. 8, pp. 085 101–1, 2011.
- [92] C. B. Markegard, I. W. Fu, K. A. Reddy, and H. D. Nguyen, “Coarse-grained simulation study of sequence effects on dna hybridization in a concentrated environment,” *The Journal of Physical Chemistry B*, vol. 119, no. 5, pp. 1823–1834, 2015.
- [93] B. E. K. Snodin, F. Randisi, M. Mosayebi, P. Sulc, J. S. Schreck, F. Romano, T. E. Ouldridge, R. Tsukanov, E. Nir, A. A. Louis, and J. P. K. Doye, “Introducing improved structural properties and salt dependence into a Coarse-Grained model of DNA,” *J. Chem. Phys.*, vol. 142, pp. 234 901–1, 2015.
- [94] P. Šulc, F. Romano, T. E. Ouldridge, L. Rovigatti, J. P. K. Doye, and A. A. Louis, “Sequence-dependent thermodynamics of a coarse-grained DNA model,” *J. Chem. Phys.*, vol. 137, pp. 135 101–1, 2012.
- [95] T. E. Ouldridge, P. Sulc, F. Romano, J. P. K. Doye, and A. A. Louis, “DNA hybridization kinetics: Zippering, internal displacement and sequence dependence,” *Nucleic Acids Res*, vol. 41, pp. 8886–1, 2013.
- [96] M. Mosayebi, F. Romano, T. E. Ouldridge, J. P. K. Doye, and A. A. Louis, “The role of loop stacking in the dynamics of DNA hairpin formation,” *J. Chem. Phys. B.*, vol. 118, pp. 14 326–1, 2014.
- [97] N. Srinivas, T. E. Ouldridge, P. Sulc, J. M. Schaeffer, B. Yurke, A. A. Louis, J. P. K. Doye, and E. Winfree, “On the biophysics and kinetics of toehold-mediated DNA strand displacement,” *Nucleic Acids Res*, vol. 41, pp. 10 641–1, 2013.
- [98] J. S. Schreck, F. Romano, M. H. Zimmer, A. A. Louis, and J. P. Doye, “Characterizing dna star-tile-based nanostructures using a coarse-grained model,” *ACS nano*, vol. 10, no. 4, pp. 4236–4247, 2016.
- [99] R. Sharma, J. S. Schreck, F. Romano, A. A. Louis, and J. P. Doye, “Characterizing the motion of jointed dna nanostructures using a coarse-grained model,” *ACS nano*, vol. 11, no. 12, pp. 12 426–12 435, 2017.
- [100] T. E. Ouldridge, A. A. Louis, and D. J. P. K., “DNA nanotweezers studied with a coarse-grained model of DNA,” *Phys. Rev. Lett.*, vol. 104, pp. 178 101–1, 2010.

- [101] T. E. Ouldridge, R. L. Hoare, A. A. Louis, J. P. Doye, J. Bath, and A. J. Turberfield, "Optimizing dna nanotechnology through coarse-grained modeling: A two-footed dna walker," *ACS nano*, vol. 7, no. 3, pp. 2479–2490, 2013.
- [102] P. Sulc, T. E. Ouldridge, F. Romano, J. P. K. Doye, and A. A. Louis, "Simulating a burnt-bridges dna motor with a coarse-grained dna model," *Natural Computing*, vol. 13, pp. 535–1, 2014.
- [103] B. E. Snodin, J. S. Schreck, F. Romano, A. A. Louis, and J. P. Doye, "Coarse-grained modelling of the structural properties of dna origami," *Nucleic acids research*, vol. 47, no. 3, pp. 1585–1597, 2019.
- [104] B. E. K. Snodin, F. Romano, L. Rovigatti, T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, "Direct simulation of the self-assembly of a small DNA origami," *ACS Nano*, vol. 10, pp. 1724–1, 2016.
- [105] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *Journal of computational physics*, vol. 22, no. 4, pp. 403–434, 1976.
- [106] M. E. Newman and G. T. Barkema, *Monte Carlo methods in statistical physics*. Clarendon Press, 1999.
- [107] D. Landau and K. Binder, *A guide to Monte Carlo simulations in statistical physics*. Cambridge university press, 2021.
- [108] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The journal of chemical physics*, vol. 21, no. 6, pp. 1087–1092, 1953.
- [109] S. Whitlam and P. L. Geissler, "Avoiding unphysical kinetic traps in monte carlo simulations of strongly attractive particles," *The Journal of Chemical Physics*, vol. 127, no. 15, p. 154 101, 2007.
- [110] S. Whitlam, E. H. Feng, M. F. Hagan, and P. L. Geissler, "The role of collective motion in examples of coarsening and self-assembly," *Soft matter*, vol. 5, no. 6, pp. 1251–1262, 2009.
- [111] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling," *Journal of Computational Physics*, vol. 23, no. 2, pp. 187–199, 1977.
- [112] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method," *Journal of computational chemistry*, vol. 13, no. 8, pp. 1011–1021, 1992.
- [113] M. Andrec, *The weighted histogram analysis method (wham)*, 2010.
- [114] A. M. Ferrenberg and R. H. Swendsen, "New monte carlo technique for studying phase transitions," *Physical review letters*, vol. 61, no. 23, p. 2635, 1988.
- [115] B. Najafi, K. G. Young, J. Bath, A. A. Louis, J. P. Doye, and A. J. Turberfield, "Characterising dna t-motifs by simulation and experiment," *arXiv:2005.11545*, 2020.
- [116] K. G. Young, B. Najafi, W. M. Sant, S. Contera, A. A. Louis, J. P. Doye, A. J. Turberfield, and J. Bath, "Reconfigurable t-junction dna origami," *Angewandte Chemie*, vol. 132, no. 37, pp. 16 076–16 080, 2020.

- [117] M. C. Engel, D. M. Smith, M. A. Jobst, M. Sajfutdinow, T. Liedl, F. Romano, L. Rovigatti, A. A. Louis, and J. P. Doye, "Force-induced unravelling of dna origami," *ACS nano*, vol. 12, no. 7, pp. 6734–6747, 2018.
- [118] S. Hamada and S. Murata, "Substrate Assisted assembly of interconnected Single Duplex DNA nanostructures," *Angew. Chem. Int. Ed*, vol. 121, pp. 6952–6955, 2009.
- [119] M. Li, H. Zuo, J. Yu, X. Zhao, and M. C, "One DNA strand homo-polymerizes into defined nanostructures," *Nanoscale*, vol. 9, pp. 10 601–10 605, 2017.
- [120] X. Li, C. Zhang, C. Hao, C. Tian, G. Wang, and C. Mao, "DNA polyhedra with t-linkage," *ACS Nano*, vol. 6, pp. 5138–5142, 2012.
- [121] H. Zuo, S. Wu, M. Li, Y. Li, W. Jiang, and C. Mao, "A case study of the likes and dislikes of DNA and RNA in SelfAssembly," *Angew. Chem. Int. Ed*, vol. 54, pp. 15 118–1, 2015.
- [122] A. Tandon, S. Kim, Y. Song, H. Cho, S. Bashar, J. Shin, T. H. Ha, and S. H. Park, "Calculation of pi and classification of self-avoiding lattices via DNA configuration," *Nat. Sci. Rep*, vol. 9, no. 1, p. 2252, 2019.
- [123] J. S. Schreck, T. E. Ouldrige, F. Romano, A. A. Louis, and J. P. K. Doye, "Characterizing the bending and flexibility induced by bulges in DNA duplexes," *J. Chem. Phys*, vol. 142, no. 16, p. 165 101, 2015.
- [124] B. N. Macchion, R. Strömberg, and L. Nilsson, "Analysis of the Stability and Flexibility of RNA Complexes Containing Bulge Loops of Different Sizes," *J. Biomol. Struct. Dyn*, vol. 26, no. 2, pp. 163–173, 2008.
- [125] T. E. Ouldrige, A. A. Louis, and J. P. Doye, "Extracting bulk properties of self-assembling systems from small simulations," *Journal of Physics: Condensed Matter*, vol. 22, no. 10, p. 104 102, 2010.
- [126] R. Wing, H. Drew, T. Takano, C. Broka, S. Tanaka, K. Itakura, and R. E. Dickerson, "Crystal structure analysis of a complete turn of b-dna," *Nature*, vol. 287, no. 5784, pp. 755–758, 1980.
- [127] K. G. Young, *Folding pathways of dna nanostructures*, eng, 2018.
- [128] K. W. Plaxco, K. T. Simons, and D. Baker, "Contact order, transition state placement and the refolding rates of single domain proteins," *Journal of molecular biology*, vol. 277, no. 4, pp. 985–994, 1998.
- [129] T. Tørring, N. V. Voigt, J. Nangreave, H. Yan, and K. V. Gothelf, "Dna origami: A quantum leap for self-assembly of complex structures," *Chemical Society Reviews*, vol. 40, no. 12, pp. 5636–5646, 2011.
- [130] H. K. Subramanian, B. Chakraborty, R. Sha, and N. C. Seeman, "The label-free unambiguous detection and symbolic display of single nucleotide polymorphisms on dna origami," *Nano letters*, vol. 11, no. 2, pp. 910–913, 2011.
- [131] H. Zhang, J. Chao, D. Pan, H. Liu, Y. Qiang, K. Liu, C. Cui, J. Chen, Q. Huang, J. Hu, *et al.*, "Dna origami-based shape ids for single-molecule nanomechanical genotyping," *Nature communications*, vol. 8, no. 1, pp. 1–7, 2017.
- [132] J. Elbaz, P. Yin, and C. A. Voigt, "Genetic encoding of dna nanostructures and their self-assembly in living bacteria," *Nature communications*, vol. 7, no. 1, pp. 1–11, 2016.

- [133] [Online]. Available: <https://github.com/behnamn90/DLM>.
- [134] P. Lecca, I. Laurenzi, and F. Jordan, *Deterministic versus stochastic modelling in biochemistry and systems biology*. Elsevier, 2013.
- [135] R. Erban and S. J. Chapman, *Stochastic modelling of reaction–diffusion processes*. Cambridge University Press, 2020, vol. 60.
- [136] N. G. Van Kampen, *Stochastic processes in physics and chemistry*. Elsevier, 1992, vol. 1.
- [137] S. K. Hahl and A. Kremling, “A comparison of deterministic and stochastic modeling approaches for biochemical reaction systems: On fixed points, means, and modes,” *Frontiers in genetics*, vol. 7, p. 157, 2016.
- [138] S. Ilie, W. H. Enright, and K. R. Jackson, “Numerical solution of stochastic models of biochemical kinetics,” *Canadian Applied Mathematics Quarterly*, vol. 17, no. 3, pp. 523–554, 2009.
- [139] P. Lecca, F. Bagagiolo, and M. Scarpa, “Hybrid deterministic/stochastic simulation of complex biochemical systems,” *Molecular Biosystems*, vol. 13, no. 12, pp. 2672–2686, 2017.
- [140] A. P. J. Jansen, *An introduction to kinetic Monte Carlo simulations of surface reactions*. Springer, 2012, vol. 856.
- [141] J. Santalucia, “A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics,” 1998, pp. 1460–1465. [Online]. Available: www.pnas.org..
- [142] J. SantaLucia and D. Hicks, *The thermodynamics of dna structural motifs*, 2004.
- [143] S. J. Schroeder and D. H. Turner, “Optical melting measurements of nucleic acid thermodynamics,” *Methods in enzymology*, vol. 468, pp. 371–387, 2009.
- [144] B. Yurke and A. P. Mills, “Using dna to power nanostructures,” 2003.
- [145] D. U. Zhang, A. J. Turberfield, B. Yurke, and E. Winfree, “Engineering entropy-driven reactions and networks catalyzed by dna,” *Science*, vol. 318, pp. 1121–1125, 5853 Nov. 2007.
- [146] D. Y. Zhang and E. Winfree, “Control of dna strand displacement kinetics using toehold exchange,” *Journal of the American Chemical Society*, vol. 131, pp. 17 303–17 314, 47 Dec. 2009.
- [147] J. G. Wetmur and N. Davidson, “Kinetics of renaturation of dna,” *Journal of molecular biology*, vol. 31, no. 3, pp. 349–370, 1968.
- [148] G. He, J. Li, H. Ci, C. Qi, and X. Guo, “Direct measurement of single-molecule dna hybridization dynamics with single-base resolution,” *Angewandte Chemie*, vol. 128, no. 31, pp. 9182–9186, 2016.
- [149] V. A. Vasiliskov, D. V. Prokopenko, and A. D. Mirzabekov, “Parallel multiplex thermodynamic analysis of coaxial base stacking in dna duplexes by oligodeoxyribonucleotide microchips,” 2001, pp. 2303–2313.
- [150] S. Neidle and M. Sanderson, *Principles of nucleic acid structure*. Academic Press, 2021.

- [151] M. Murphy, I. Rasnik, W. Cheng, T. M. Lohman, and T. Ha, “Probing single-stranded dna conformational flexibility using fluorescence spectroscopy,” *Biophysical journal*, vol. 86, no. 4, pp. 2530–2537, 2004.
- [152] A. Cumberworth, D. Frenkel, and A. Reinhardt, “Simulations of dna-origami self-assembly reveal design-dependent nucleation barriers,” *Nano Letters*, Sep. 2022.
- [153] J. M. Arbona, J. Elezgaray, and J. P. Aimé, “Modelling the folding of dna origami,” *EPL*, vol. 100, 2 Oct. 2012.
- [154] J. M. Majikes, J. A. Nash, and T. H. LaBean, “Competitive annealing of multiple dna origami: Formation of chimeric origami,” *New Journal of Physics*, vol. 18, no. 11, p. 115 001, 2016.