Monte Carlo simulation of DNA origami self-assembly

Alexander Michael Cumberworth

Darwin College
University of Cambridge

This thesis is submitted for the degree of Doctor of Philosophy

September 2019

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This thesis does not exceed 60,000 words, including summary/abstract, tables, and footnotes, but excluding table of contents, photographs, diagrams, figure captions, list of figures/diagrams, list of abbreviations/acronyms, bibliography, appendices and acknowledgements. Finally, some of the work presented in this thesis has been published [1]; I was the primary contributor to all aspects of the paper.

Abstract

Monte Carlo simulation of DNA origami self-assembly

Alexander Michael Cumberworth

The optimal design of DNA origami systems that assemble rapidly and robustly is hampered by the lack of a model capable of simulating the self-assembly process that is sufficiently detailed yet computationally tractable. In this thesis, we propose a model for DNA origami that strikes a balance between these two criteria by representing DNA origami systems on a lattice at the level of binding domains. Particular attention is paid to the constraints imposed by the double-helical twist, as they determine where strand crossovers between adjacent helices can occur.

Because of the highly specific types of interaction and the length of the scaffold, standard Monte Carlo simulation methods for polymeric systems are found to be ineffective at sampling the dense, near-assembled states considered here. In order to address the issue of sampling such states, we develop Monte Carlo methods that extend the configurational bias and recoil growth methods, and consider the sampling of scaffold conformations independently from the sampling of staple binding states. We demonstrate the validity of our model and the feasibility of our sampling methods with simulations of a small origami design previously studied with the oxDNA model, as well as with designs that include staples that span longer scaffold segments.

In other self-assembling systems, it is often the case that nucleation barriers control the self-assembly behaviour. We investigate whether there is a nucleation barrier along the self-assembly pathway of DNA origami. Our simulations reveal that for simple systems, stacking interactions govern a nucleation barrier, albeit one that is never prohibitively large relative to thermal fluctuations. These findings may prove useful in the design of DNA origami structures capable of controllable reversible folding for functional purposes and in assisting the optimization of assembly pathways at the design stage.

Contents

Declaration i

Abstract iii

Acknowledgements vii

Glossary ix

1 Introduction 1
- 1.1 Structural DNA nanotechnology 1
- 1.2 DNA origami 3
- 1.3 Self-assembly of DNA origami 6
- 1.4 Modelling DNA origami 9
- 1.5 Issues and approach 14

2 Lattice models of DNA origami 15
- 2.1 State space 15
- 2.2 Potential energy 21
  - 2.2.1 Bonding term 21
  - 2.2.2 Stacking term 28
  - 2.2.3 Steric term 32
- 2.3 Explicit helical axis model 37

3 Simulation methods 39
- 3.1 MCMC simulations for molecular systems 39
- 3.2 MC methods for lattice polymers 41
- 3.3 Move types for DNA origami lattice models 43
  - 3.3.1 General considerations 43
  - 3.3.2 Biased chain regrowth methods 45
- 3.4 Orientation vector rotation moves 47
- 3.5 Staple regrowth moves 48
- 3.6 Staple exchange moves 49

3.7 Scaffold regrowth moves 51
3.7.1 Growth bias 51
3.7.2 Segment selection 53
3.8 Replica exchange 58
3.9 Free-energy calculations 61
3.10 Numerical validation of MCMC move types 64
3.11 Optimization of MCMC parameters 64

4 Feasibility and validity of approach 67
4.1 Motivation 67
4.2 Simulation and analysis methods 67
4.3 Initial parameter selection 68
4.4 Results 69
4.5 Conclusions 83

5 DNA origami and nucleation 87
5.1 Motivation 87
5.2 Simulation and analysis methods 88
5.3 Results 89
5.4 Conclusions 109

6 Conclusions 115

References 121

Acknowledgements

I would like to thank my supervisor, Prof. Daan Frenkel, for providing me with the autonomy to approach scientific problems as I saw fit, and deftly steering me away from unproductive avenues when I went too far off course. I would further like to thank Dr Aleks Reinhardt, who provided invaluable guidance throughout my PhD. Thanks to all the members over the years of office 360, the Downing lunch group, the Frenkel group, and other members of the Chemistry department for the friendly, supportive environment and engaging conversations they provided, especially (listed in alphabetical order) Dr Tine Curk, Dr Gül Güryel, Dr Jerelle Joseph, Wei Kang, Haydn Lloyd, Dr Stefano Martiniani, Lisa Masters, Dr Matthias May, Dr Carl Poelking, Dr Guillem Portella, Dr Bianca Provost, Simon Ramirez Hinestrosa, Xiaoliang Tang, Dr Nick Tito, Dr Charlie Wand, Dr Michael Willatt, Dr Peter Wirnsberger, Rhiannon Zarotiadis, Dr Chao Zhang, and Dr Mengjie Zu. I would also like to thank the Marie Skłodowska-Curie training network, COLLDENSE, for funding my PhD, as well as all the other early-stage researchers for their highly enjoyable company at our many meetings.

Thanks to the friends I made at Darwin College who supported me through this time, especially Charlotte Tumescheit, Melanie Whitfield, James Luis, and Helen Street. A special thanks to Grace Bentham for her support in the final months of my PhD. Thanks are due to my long-suffering housemate, Nadeem Gabbani, for putting up with me. Finally I would like to thank my family, especially my parents, sister, and brother-in-law, whose support has always been unwavering.

Glossary

AFM atomic force microscopy

bp base pair

CAD computer-aided design

CB configurational bias

CT conserved topology

CTMC continuous-time Markov chain

dsDNA double-stranded

fcc face-centred cubic

FRET Förster resonance energy transfer

HP hydrophobic-polar

LFE Landau free energy

MBAR multi-Bennett acceptance ratio

MC Monte Carlo

MCMC Markov chain Monte Carlo

MD molecular dynamics

NN nearest-neighbour

nt nucleotide

PCA principal component analysis

PERM pruned and enriched Rosenbluth sampling

REMC replica exchange Monte Carlo

RMSD root mean square deviation

RG recoil growth

SAW self-avoiding walk

ssDNA single-stranded DNA

US umbrella sampling

WHAM weighted histogram analysis method

1
Introduction

1.1 Structural DNA nanotechnology

Precise control at small length scales, and especially the manufacturing and assembly of structures at these scales, has led to some of the most important technological advances of the last century. This control usually comes from assembly methods that are "top-down": an assembling machine contains the information necessary for the final structure, and adds and subtracts components or material to reach it. Biological systems are also fundamentally based around assembly at small scales, but besides top-down approaches (an example being ribosomal polypeptide synthesis), they also make use of "bottom-up" approaches. In bottom-up assembly, the components contain the information of the final structure, and will self-assemble given sufficient time and appropriate conditions; one example is the folding of a polypeptide chain. Hijacking and mimicking these biological systems provides another route for assembling designed structures at small scales. In contrast to proteins, nucleic acids have highly specific interactions between a small number of monomer types, which makes them particularly amenable for repurposing as a self-assembling material.

The monomers of nucleic acids, nucleotides, are composed of a phosphate group, a 5-carbon sugar, and a nitrogen-containing aromatic group, referred to as a nucleobase, or simply a base. The phosphate group and the sugar are covalently bonded between monomers to form the backbone of the polymer, while the base is attached to the sugar. DNA, best known for its role as the primary carrier of genetic information in cells, is a nucleic acid with four monomers, which vary only by their bases. The four monomers fall into two groups: the purines, adenine (A) and guanine (G), and the pyrimadines thymine (T) and cytosine (C) (Figure 1.1(a)).

The key property of DNA that makes it so useful to cells as an information storage material is that it self assembles into a double helix in which each base pairs to only one other base; specifically A pairs with T and C pairs with G. This process is commonly referred to as hybridization. The base pairing is facilitated by

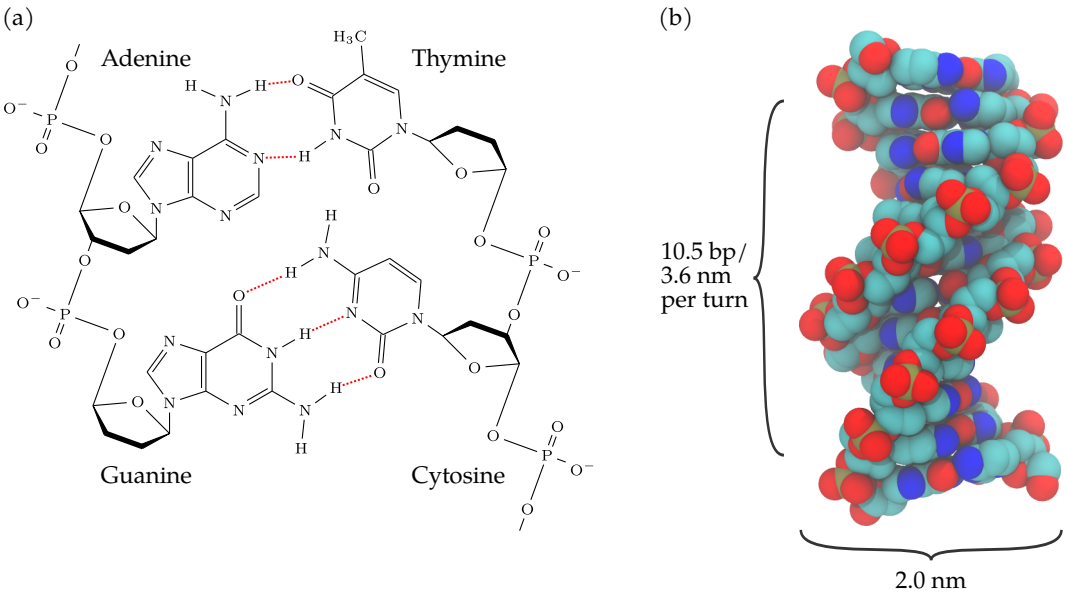

Figure 1.1: Overview of DNA structure. (a) Chemical structure of the nucleotides. Hydrogen bonding is shown in red. (b) Structure of the B-form DNA double helix.

hydrogen bonding and complementary shapes of the bases. However, stacking of the bases between pairs contributes roughly half of the stability of the helix [2], an interaction that involves both an electrostatic and a hydrophobic component. Wile other pairings between free nucleotides may have the same number of hydrogen bonds and a similar strength of base stacking, they are disfavoured because they are not compatible with the geometry of the double helix that is favourable under typical physiological conditions. While more than one form of double helix is possible, under physiological conditions, B-form DNA, shown in Figure 1.1(b), is prevalent.

Structural DNA nanotechnology is the knowledge relating to the creation and use of materials composed of DNA, and the materials themselves, which contrasts with technologies relating to DNA's information-storage capabilities. Because these systems are based on biological molecules, they have the advantage of being naturally compatible with biological systems, although their applications extend far beyond. The idea of using DNA to construct functional structures was initially conceived by Seeman in the early 1980s [3–5], who demonstrated that DNA could be used as a versatile design material to build novel nano-structures. Since then, there has been great interest in pushing the limits of the size and intricacy of the structures that can be designed and assembled with DNA. However, for the following two decades, studies mostly focused on individual structural motifs or periodic 2D or 3D arrays comprising one or several of these motifs (sometimes referred to as the

tile approach) [6]. These methods are limited in the scope of structures that can be realized and, in the case of periodic designs, the necessity of extensive purification and precise stoichiometry to achieve appreciable size and yield. In the last decade, two new approaches have emerged that circumvent these issues: DNA origami and DNA bricks. With the more recent of these, DNA bricks, multiple copies of a single small structural motif, or brick, with distinct sequences assemble into structures in which each brick has a unique, addressable location [7]. While the approach is promising and has been extended to the construction of 3D structures [8], it is still at an early stage in its development.

1.2 DNA origami

It was not until the seminal paper of Rothemund [9], which introduced the DNA origami method, that the potential complexity and applications of structural DNA nanotechnology systems really became apparent and began to be realized. The key idea behind DNA origami is to employ a long single-stranded DNA (ssDNA) 'scaffold' strand that is subsequently folded into its target structure with the hybridisation of a number of designed, shorter 'staple' strands that link selected binding domains on the scaffold strand (Figure 1.2). As the staples and scaffold bind to each other, double helices are formed, which are much more rigid, and so provide the final structure with mechanical stability. The helices are connected to each other by crossovers between individual strands on adjacent helices, typically forming four way junctions, which are known as Holliday junctions [3]. While the two helices that comprise the junctions in isolation are capable of being parallel, and when unconstrained are relatively flexible, the presence of many junctions between helices in the structures, the sequence design of the staples, and the high ionic strength leads to the junctions taking on antiparallel configurations [3]. Because of the double helical twist, junctions between two antiparallel helices will only be possible at certain intervals. While early designs were mostly simple planar structures, it is now possible to design and assemble virtually any connected 3D shape, including dynamic, stimuli-responsive structures with flexible joints [10–16]. The advantages provided by DNA origami have stimulated numerous investigations exploring design and assembly methods, structural and functional classes, and applications in both medical and non-medical fields, of which there have been many reviews [6, 14, 16–49].

DNA origami can act as an addressable platform for precise positioning of molecules, as each staple type can be individually functionalized and has a unique known position in the final structure [19, 20, 50, 51]. Such positioning has a broad

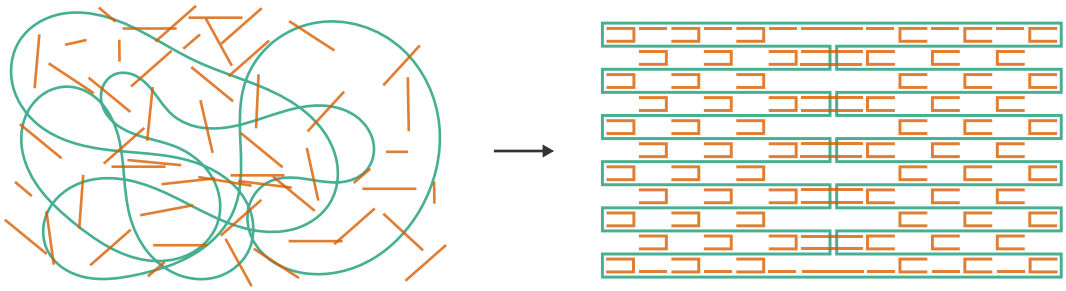

Figure 1.2: A diagram of a DNA origami system in unbound and a bound states. Staples are in orange while the scaffold is in green.[2]

range of applications: chemical sensors (e.g. by positioning RNA probes [52]), nanoreactors (e.g. by positioning enzymes [53, 54]), electronics (e.g. by positioning carbon nanotubes [55] or DNA origami itself on a surface [56]), and photonics (e.g. by positioning gold nanoparticles [57–59]). DNA origami is also of particular interest for drug delivery applications [17, 27]. Staples may be functionalized with compounds that allow for increased cell uptake and targeted delivery, or with the drugs themselves [30, 32]. Similar medicinal effects can also be achieved by using different shapes or including aptamers or immunostimulating sequences in the origami design [22, 30, 32]. Hollow 3D shapes which may function as containers, some even with lock and key lids [60, 61], have also been explored, and again have a clear application in drug delivery. Cancer has proven to be one especially attractive target for DNA origami based-drug delivery and detection systems [27, 62, 63]

While static DNA origami designs already have a broad range of uses, the introduction of dynamic, or mechanical, motifs further increases their possible utility [14]. Such motifs may interface with origami systems, as in the case of a DNA walker moving on a DNA origami substrate [64], or be directly incorporated as part of the origami structure [65]. In some cases they may be designed to be responsive to external electric fields, allowing them to be directly user-controlled [66, 67]. These motifs provide the possibility of creating nanomachines and nanorobots [24, 33]; one study has already demonstrated DNA origami acting as thought-controlled nanorobots in living systems [68]. Finally, systems composed of a variety of origami and non-origami motifs have been proposed that can carry out logical computations on given inputs and carry out an action based on the results [69]; this might be used for the intelligent release of drugs [70] or the creation of a molecular assembly line [71].

[2]The diagram of the assembled state is adapted by permission from Springer Nature Customer Service Centre GmbH: Nature Research, Nature, Guiding the Folding Pathway of DNA Origami, K. E. Dunn, F. Dannenberg, T. E. Ouldridge, M. Kwiatkowska, A. J. Turberfield, and J. Bath, 2015.

The methods for designing and assembling DNA origami themselves have seen significant advancement in the decade since the origami approach was introduced. While the first DNA origami structures were designed by hand, a likely very time-consuming process, Douglas et al. [72] released a computer-aided design (CAD) program for designing both 2D and 3D origami structures. The designs rely on placing the helices in parallel in a regular way to deal with the constraints imposed by the double helical twist on the location of junctions between helices. In the case of 3D structures, honeycomb or square lattices are used, although curves are able to be incorporated by shortening or lengthening some of the helices in the structure to manipulate the internal forces [10, 11, 73]. One important application to come from the development of 3D structures was the creation of nanopores from DNA origami, some of which are able to be inserted into lipid membranes [34, 35, 74, 75]. However, more advanced design methods and algorithms have recently been developed that allow almost arbitrary wireframe structures to be constructed [12, 13, 15, 76, 77], and efforts have been made towards hierarchical self-assembly in which DNA origami structures themselves become the components in larger structures [78–81].

Another design constraint is the scaffold strand itself. While the staples can be synthetically produced, the desired length of the scaffold strand has often precluded its production by the same route. Thus, biotechnological approaches for the synthesis of the scaffold are used: typically, the scaffold is the genome of the M13mp18 bacteriophage, which is a circular 7429 nucleotide strand [82]. Of course, using a single scaffold for all designs is not optimal; alternative scaffolds are being explored by modifying the commonly used viral genome or using other viral genomes as the scaffold, and now even through fully synthetic production of the scaffold [82–91]. With a fully customizable scaffold and careful sequence design considerations, it is now also possible to produce ssDNA origami, in which the scaffold folds without the aid of staples [92].

More practical restrictions on the use of DNA origami relate to its cost, stability, and quality. One approach to decrease the cost has been to avoid the expensive and poorly scaling synthesis of the staple strands by inserting them into the phagemid that the scaffold is produced within and producing them with biotechnological methods [83, 93]. An alternative approach is to use multiple copies of each staple type [94]. The stability of DNA origami under various conditions, especially cation concentration and the presence of nucleases, has been examined and stabilization methods proposed [18, 21]. For sensitive applications, it is important to be able to quantify the quality of the assembled structures and the actual level of addressability

provided; methods have been developed for this purpose [95, 96].

1.3 Self-assembly of DNA origami

One of the advantages of the DNA origami method is that the intensive purification and precise stoichiometry required when attempting to assemble periodic, tile-based structures is no longer necessary [3]. Assembly is often performed by thermally annealing a mixture with an excess of staple strands over several hours or even days [97]. However, Sobczak et al. [98] found that, following a high-temperature denaturing step, it is also possible to assemble structures isothermally. While isothermal assembly had been demonstrated previously, it had replaced the thermal annealing with an inconvienient chemical denaturant annealing step [99]. Isothermal assembly has been shown to be more efficient for a range of designs, with the optimal temperature for this process depending on both the design of the target structure [100] and the conditions [99, 101, 102]. With the addition of a chemical denaturant [101, 102] or the use of mechanical pulling [103], it is even possible to carry out assembly at or near room temperature. While the most common assembly conditions involve the use of a buffer solution that includes magnesium chloride, assembly has been demonstrated with only monovalent ions [104]. Besides potential increases in speed and yield, having multiple conditions available is useful when the staple strands are functionalized to groups sensitive to some of the typical assembly conditions. Other more involved assembly methods have been explored that allow for one-pot assembly of multiple different structures [105] and the ability to trigger assembly with a lock-and-key approach for in vivo situations [106].

The factors that determine the kinetics of origami formation are different from those that determine the formation of an ordered crystal. Cademartiri and Bishop [107] differentiate between two fundamentally different types of self-assembly: the *puzzle* and the *folding* approach. Crystals and many periodic structures typically form via the puzzle mechanism. The information of how and where a given component must bind under the puzzle mechanism is entirely stored in the interactions between components; they move freely through the solution until the correct partners in the correct orientation are encountered. To achieve complex and addressable structures [50], the interactions between components must be highly specific (e.g. in DNA bricks [8]). By contrast, the folding mechanism relies on the fact that some of the components are already covalently bonded to each other, where the covalent bonds are formed by some non-self-assembling process. As an example, protein folding starts from a structure in which the individual amino acids have been bonded

together by a ribosome in a particular sequence that is encoded in the genome. This pre-forming of more permanent bonds by some other process allows complex structures to be encoded with less specific types of interaction between the individual components (e.g. non-covalent interactions between amino acids) because of the additional constraints on the system. With the binding of the staples to specific segments of a scaffold strand and the subsequent folding up of the scaffold strand, DNA origami combines both of these approaches.

The assembly process of DNA origami has been characterized experimentally and through simulation. Experimental characterization can be done for bulk properties, usually with spectroscopic techniques for measuring melting and annealing curves, or for individual structures, with atomic force microscopy (AFM) images of origami structures [108]. Measuring overall assembly can be done by recording UV absorbance, which decreases upon transitioning from ssDNA to double-stranded (dsDNA), or recording the fluorescence of intercalating dyes, which fluoresce more efficiently in dsDNA. Information on the assembly of localized parts of an origami structure can be obtained by using Förster resonance energy transfer (FRET) probes hybridized to staples close to each other in the final structure. The models that have been used to study the assembly process are described in Section 1.4.

An important question about the assembly process is whether it is under kinetic or thermodynamic control, and to what extent this depends on the conditions, assembly protocol and origami design. Hysteresis between the melting and annealing curves is commonly observed [109–114], with annealing occurring at lower temperatures than melting; the effect seems to be stronger in 3D origami structures [111]. Increasing the heating/cooling rate has been found to have a more pronounced effect on annealing than on melting [98, 110], which suggests that for some conditions and designs assembly is slow relative to melting and may involve high free-energy barriers, corresponding to a process under kinetic control. Conversely, Wah et al. [115] examined intermediates of the assembly process with AFM of origamis of a similar design and found that the annealing and melting pathways were largely the reverse of each other, and that the calculated melting and annealing temperatures were in agreement within experimental error, suggesting a thermodynamically controlled process.

Using both AFM measurements and simulations, Dunn et al. [109] examined an origami design for which multiple fully assembled configurations were equally stable and showed that both thermodynamic and kinetic factors could be manipulated to control the outcome of the assembly reaction. They were able to shift the assembly

yield towards specific assembled configurations by modifying the staple designs to increase the stability of those states. A similar shift in assembly yield was also able to be achieved with staple modifications that instead changed the stabilities of some of the intermediate states in the assembly pathways, leaving the stabilities of the assembled configurations unchanged.

The assembly of DNA origami is a cooperative process [98]. This follows from the fact that the melting and annealing curves, as measured with spectroscopic techniques, are narrower than those of the corresponding isolated binding domains. The most obvious form of cooperativity is the increase in local concentration of binding domains when they are brought together by a staple that binds nearby domains to form a loop [109, 116]. However, Dannenberg et al. [110] found that coaxial stacking between staples adjacent to each other on the scaffold may also increase cooperativity of the assembly process. Another form of cooperativity was uncovered by Shapiro et al. [117]. They found that for heterotypic junctions, where the melting temperature of each arm is different, the annealing temperatures of domains in the junction were shifted relative to the free state ones, and that the order of the arms can further shift the annealing temperatures. Both forms of cooperativity would be expected to have primarily a local effect and indeed FRET experiments [111] and simulations [110] have found that excluding a staple from the reaction mixture only affects the binding of nearby staples in the assembled structure. By contrast, AFM studies of assembly pathways found that assembly started at the edges of the structure and proceeded inwards [115, 118]. However, a study using a similar method found the staples were binding to the partially formed structure in an independent manner [119].

Perhaps the most thorough study of the assembly pathway of an origami design, which was a 3D structure based on a honeycomb lattice, is that of Schneider et al. [116]. For each staple type, they ran isothermal assembly reactions where that staple type had FRET probes to measure the kinetics of its binding to the scaffold; additionally, they placed the FRET probes on pairs of staples expected to be in close proximity in the assembled structure to measure the correlation between these staples. By comparing results between two different staple sets with the same topology but different sequences, they found that sequence was not a strong factor in the order of staple binding. The largest factor in correlations between assembly times was the topology, with staples that closed similar loops being well correlated, and longer loops being correlated with later incorporation times. They also found that the geometric position in the final structure was not well correlated with staple binding times; i.e. 'internal' staples did not incorporate more slowly, which they speculate is

because the division between internal and external is not well defined until very late in the assembly process. The staples had on average 5.3 domains, and they found the binding of the termini of individual staples to not be strongly correlated, implying that the pathway is better described by the binding of individual domains, rather than by whole staples.

The highest resolution look at an assembly mechanism is that of Snodin et al. [120]. They ran simulations (see Section 1.4 for a description of their approach) of a small origami system with several temperatures and two staple concentration regimes. They found one of the greatest kinetic hindrances to full assembly was the blocking of staple binding by the binding of another staple of the same type at the other scaffold domain. Misbinding, especially of a bound domain to adjacent domains, also had a substantial effect. However, the results are difficult to interpret, as they simulated assembly fully only a single time, used a model with base-pair averaged binding energies, employed much higher staple strand concentrations than is typical in experiments, and studied a design that lacks long-ranged staple crossovers. It may be that the timescale of blocking resolution is fast enough not to be a limiting factor in experimental assembly conditions.

1.4 Modelling DNA origami

Molecular simulations can be used to gain a better understanding of the factors that influence the thermodynamics and kinetics of assembly and melting of DNA origami structures. A great variety of models of DNA have been developed [121–125]. These span the gamut of particle resolution, ranging from atomistic [124], sub-nucleotide coarse-grained models [124, 126–151], nucleotide coarse-grained models [152–159], single and multiple base-pair coarse-grained models [160–163], to continuum models [164]. Some are intended as general DNA models, while others are designed with specific applications in mind; for example, models of DNA that are more statistical in nature have been designed to gain a better understanding of the fundamentals of hybridization and denaturation of strands [165–172].

Some approaches to modelling DNA origami are intended or only feasible for studying the assembled state. Because most experimental characterization of folded states is in the form of AFM, which requires the origami structure to be adsorbed on a surface, these studies of assembled states provide a much needed glimpse at the structure when free in solution, as well a way to study their mechanical properties. The brute-force approach is to use fully atomistic simulations, which in addition to mechanical properties can also be used to study the dynamics of the assembled

structures [173, 174]. Still, these simulations are time consuming and include much more detail than is necessary if one is interested only in mechanical properties of assembled states. A more popular and successful approach with such a focus has been developed [11, 175, 176], in which DNA double helices are modelled as finite-element elastic rods rigidly connected to other double helices, and single-stranded DNA as a finite element approximation of a freely jointed chain; the system is relaxed from a given initial state to find a force-balanced equilibrium state. A more recent model that takes a more discrete approach to studying mechanical properties has also been developed [177].

Unfortunately, as in protein-folding simulations, direct atomistic simulations of the assembly process are not feasible for all but the shortest sequences. It is therefore necessary to use simplified yet realistic models. In fact, several such models that vary in their level of detail have been introduced. These models can be split into two classes: lower-resolution models that take a statistical approach, and higher-resolution models that are based on physical coarse-grained models of DNA.

In the lower-resolution models, the standard approach is to model DNA origami self-assembly by extending a thermodynamic model of DNA hybridization to account for the entropic effects of folding the scaffold. A naive approach to calculating the free-energy change upon hybridization of two strands would be to sum the free-energy contribution of each base pair, with contributions from A/T base pairs and C/G base pairs, plus some additional free-energy cost of bonding two strands together. However, in addition to the difference in hydrogen bonding between base pairs, there are differences in the interactions (e.g. the base stacking) between different combinations of base pairs. In what is commonly referred to as the nearest-neighbour (NN) approach [167], the two directly adjacent nucleotides are considered to provide sufficient context for calculating an accurate free energy of hybridization contribution, such that

$$\Delta G^{\ominus}_{\mathrm{NN}} = \sum_{i}^{i=n-1} \Delta G^{\ominus}_{\mathrm{NN},\,i} \tag{1.1}$$

where $\Delta G^{\ominus}_{\mathrm{NN}}$ is the standard state NN Gibbs free energy of hybridization, n is the number of base pairs, and $\Delta G^{\ominus}_{\mathrm{NN},\,i}$ is the standard state NN Gibbs free-energy contribution of the pair of base pairs at position i and $i+1$. This results in ten parameters rather than two, corresponding to ten distinct pairings of base pairs. The number of parameters doubles if the temperature dependence of the free energy is desired, as now the enthalpy and entropy are required separately, with $\Delta G^{\ominus}_{\mathrm{NN}} = \Delta H^{\ominus}_{\mathrm{NN}} - T\Delta S^{\ominus}_{\mathrm{NN}}$,

where H°_{NN} is the standard state NN entropy and $\Delta S^{\circ}_{\mathrm{NN}}$ is the standard state NN entropy. One could consider going even further and including the next nearest neighbours, but this was found to lead to little increase in accuracy [167].

To determine the parameters for a NN model, one can measure the melting temperatures for a set of sequences across a range of concentrations, providing the enthalpies and entropies for each sequence, which can then be used in a linear regression. The enthalpies and entropies are usually assumed to be independent of temperature, so the validity of the parameters decreases when the temperature deviates too far from the melting temperatures of the sequences used in the parameterization, although this dependence can be accounted for [178, 179]. These models have also been modified to account for varying salt concentrations [167, 180]. Besides fully hybridized states, NN models can also account for single internal and terminal mismatches, dangling ends, various loops and bulges, and coaxial stacking between separate strands [167]. A number of groups have derived NN parameter sets, but the most successful is set was derived by the Santalucia group [167, 181, 182], which is based on a set of 108 sequences.

With the NN approach as a foundation, Arbona, Aimé and Elezgaray [112–114] modelled the assembly process as a series of equilibrium reactions to calculate the likelihood that a particular staple or individual staple binding domain is bound to its complementary binding domain(s) on the scaffold at a given temperature. To calculate the equilibrium constants of each reaction, they introduce a model of the free-energy change upon binding of each staple, $\Delta G = \Delta G_{\mathrm{NN}} + \Delta G_{\mathrm{top}}$, where ΔG_{NN} is from the NN model and ΔG_{top} is the contribution of the topology of the system in its current bound state. To solve their equations they make a number of assumptions. The first is to assume that staples bind fully and only to the correct domains. The second is to assume that if the probability of one staple forming is greater than another, then that staple will always bind first. Perhaps most fundamental, however, are the assumptions involved in calculating ΔG_{top}, which involves considering the free-energy cost of forming loops when new staples form. They calculate the loop contribution by empirically modulating another term from the NN model that gives the free-energy cost of hybridizing two strands when a segment of one of the strands has extra, non-complementary bases that stick out and form a bulge.

Dannenberg et al. [110] and Dunn et al. [109] instead formulated their model as a continuous-time Markov chain, where the state space is described by the binding states of each staple type in the system, which allows a clearer link to the kinetics of the assembly process. The state space is described by the states of each staple type

in the system, where a two-domain staple can either be unbound, bound to one, the other, or both scaffold domains, or have two copies, one at each domain. The absolute values of the rate constants are made calculable by assuming the reverse rate to be the unbinding rate of an isolated duplex. The continuous-time Markov chain (CTMC) is simulated by selecting a timestep interval and a temperature change rate, and cycling between an initial and final temperature.

As with the previous model, the NN model is used for the basic hybridization free energy, supplemented with a term to include the effects of topology, but here an additional term is added to describe stacking interactions between staple domains on separate staples bound to contiguous segments of the scaffold, ΔG_{stack}. The stacking term is also based on the NN model, but they take the sequence averaged value and multiply it by a parameter they tune during the parameterization of the model. The topology term is taken to be the free-energy change upon breaking and forming loops in the system; the total contribution relative to the fully unbound state is

$$G_s^{\text{top}} - G_{\text{null}}^{\text{top}} = \sum_{L(s)} \Delta G_j^{\text{loop}},$$

where G_s^{top} is the absolute free energy in state s, $G_{\text{null}}^{\text{top}}$ is the absolute free energy in the unbound state, $L(s)$ is the set of loops in state s, and ΔG_j^{loop} is the free energy of forming loop j.

At least for 2D origami designs, the loops can be unambiguously identified; more complex designs require a simplified approach, which they term the local model (in contrast to the full, "global" model), where they assume that only free-energy changes resulting from the formation or dissolution of the smallest loop are relevant. The loop free energies are calculated from the probability that the ends of the loop come together within an arbitrarily small distance r_C when not constrained, $P_{\text{loop}}^{r_C}$,

$$\Delta G^{\text{loop}} = -RT \ln\left(\frac{P_{\text{loop}}^{r_C}}{P_{v^0}^{r_C}}\right),$$

where $P_{v^0}^{r_C}$ is the probability that the ends come together in two unbound strands in an ideal system with volume v^0. By assuming the probability distribution can be described by that of a freely jointed chain and integrating, they obtain

$$\Delta G^{\text{loop}} = RT\gamma \ln \frac{C}{E\left[r^2\right]_{\text{loop}}},$$

where γ and C are parameterized constants, and $E\left[r^2\right]_{\text{loop}}$ is the mean squared distance between the two ends.

These approaches allow the assembly process to be simulated in under an hour on current computers, and have led to some important insights into the self-assembly process (see Section 1.3). Nevertheless, the efficiency advantage that these statistical models provide comes at the price of having no explicit geometric representation of the system, and making fairly strong assumptions about the entropic changes that occur during assembly. Furthermore, in Dunn et al. [109], an ad hoc exclusion algorithm is used to reject configurations that are not on a pre-defined folding path as a proxy for steric constraints. Finally, these models ignore the possibility that staples may bind (albeit less strongly) to incorrect binding domains.

The higher resolution approach involved the use of a more general coarse-grained model of DNA known as oxDNA that can be simulated with Monte Carlo (MC) or molecular dynamics (MD) methods [183–186]. This model is coarse-grained to the level of nucleotides, which are rigid and resolve the backbone from the base in the potential. The potential includes a spring potential for the backbone, excluded volume interactions for the backbone and base, as well as hydrogen bonding base-stacking interactions that have an angular component. They parameterized it by matching to structural, thermodynamic, and mechanical properties, with a trial-and-error approach. The model has been successfully used to study mechanical properties of DNA origami structures [187], as well as their stability upon force-induced unravelling [188]. Perhaps most ambitiously Snodin et al. [120] ran simulations of the self-assembly of a small DNA origami, the results of which we discussed in Section 1.3. They were able to capture a full assembly event in unbiased simulations of the system with their model, which allowed them to study the process in unprecedented detail. However, because of the level of detail that oxDNA provides, these simulations of a small origami design with only short loops present in the final structure took several months on a cluster with GPU acceleration. Moreover, they found it necessary to use staple concentrations in excess of those typically used in experimental assembly conditions. This is not just a matter of speeding up the kinetics: such high concentrations shift the equilibrium between free and bound staple strands towards the bound states.

1.5 Issues and approach

Although the rules for designing a DNA origami system with a particular final structure are well understood, our understanding of the precise assembly thermodynamics and kinetics (i.e. the order and cooperativity of staple binding) is much more limited. Yet such understanding is potentially very useful for designing origami structures that fold most efficiently into their target structure. Given the range of possible applications of DNA origamis, improving the speed and yield of their assembly may have significant practical use.

While there has certainly been progress in understanding fundamental aspects of DNA origami self-assembly, there are contradictions in the findings, and limitations in the approaches taken. Experimental studies of assembly with AFM can examine individual intermediate and final origami structures [115, 119, 189], but they are fundamentally limited to simple, planar designs. OxDNA is useful for studying the dynamics of DNA nanostructures, but it is too detailed for studying the assembly process: even with the previously mentioned unrealistic conditions chosen to speed up the process, the simulations took months running on GPUs to simulate a single assembly event [120]. The statistical models are certainly computationally feasible, but they lack an explicit model of the geometry of the system, instead making very strong assumptions about the nature of cooperativity in the assembly process.

In Chapter 2, a model is proposed that is intended to bridge the gap between the detail of the oxDNA model and the efficiency of the statistical models for the study of DNA origami self-assembly. In Chapter 3, methods for sampling the configuration space defined by our proposed model are developed, which specifically address the challenges of sampling near-assembled and assembled states of DNA origami systems. In Chapter 4, we test the feasibility of our approach by simulating three different small test systems, and compare our results to what we expect as a test of the validity of the model. Finally, in Chapter 5, we examine the role of nucleation in the self-assembly of DNA origami.

2
Lattice models of DNA origami

2.1 State space

To realize our goal of developing a computationally feasible model of DNA origami self-assembly, we chose to use a lattice representation in order to increase computational efficiency by reducing configuration space. The use of a lattice representation for structural-DNA-nanotechnology systems has good precedent: a lattice model of DNA bricks was remarkably effective [190–192], and unexpectedly yielded near quantitative agreement with experimental measurements of the nucleation kinetics [193]. Apart from this model, a number of other lattice models of nucleic acids have been developed previously, although of those that allow for changes in the hybridization state, the particle resolution is typically at the level of nucleotides or finer, and they are mainly aimed at understanding ssDNA dynamics and hybridization thermodynamics and kinetics [155, 159, 165, 166, 194], or in one case predicting RNA folded states with a scoring function [195]. The goals of lattice models of protein folding [196, 197] overlap more with the goals of the model developed here, but the structural details are sufficiently different that again there is not a model that can be easily modified to suit our needs.

One of the most fundamental choices to be made in the design of the model is the level of resolution of the molecules. In the context of DNA origami, a 'binding domain' is defined as a segment of an individual DNA chain that, in the final assembled state, is fully bound to another, complementary segment of DNA. In our model, we choose binding domains as the basic unit and represent such binding domains as particles on a lattice. The definition of a binding domain in the model may differ from the definition of a binding domain for a given design, as here the binding domains must be all approximately the same length, or number of residues. If a design has binding domains whose lengths are integer multiples of each other, the binding domain as defined in our model will be the smallest binding domain in the design, with larger binding domains in the design being represented by multiple

binding domains in the model. Designs commonly vary the length of the binding domains by one or two residues to reduce internal stresses in the final structure, and these may be represented in our model as a single binding domain as the total change in length is relatively small. However, if the binding domains substantially in length and are not integer multiples of each other, it may not be feasible to represent the design with our model. Finally, contiguous binding domains on a given chain are constrained to occupy adjacent lattice sites.

Initially, four lattices were considered as possibilities for the model: simple cubic (coordination number of 6), face-centred cubic (fcc) (coordination number of 12), high coordination cubic [155] (coordination number of 26), and the bond fluctuation model [198, 199] (coordination number of 108). The latter two provide a way of extending the coordination number of a lattice by representing the occupancies with multiple lattice sites, and allowing a small amount of variation in the bond length. These approaches allow for a finer description of space, but at the cost of extra complexity of the model. The fcc lattice has the issue of having variable angles between neighbouring lattice sites. Because many origami designs have angles between helical axes that involve only angles that are multiples of 90° in the final structure, the simple cubic lattice, while having a relatively low coordination number, gives a reasonable approximation of space for our purposes.

The most common way of describing an association between two units in nucleic acid lattice models is to have them interact when on adjacent lattice sites [155, 159, 165, 190, 195], possibly with a particular orientation of some conformational degrees of freedom [159]. However, this description makes inclusion of the double-helical twist difficult, which, as will be discussed later, is critical to modelling DNA origami. Therefore, in our model two domains can hybridize only when sharing the same lattice site, a choice also made by Everaers et al. [166] and Causo et al. [194] in their models of DNA hybridization (both of which also happened to use a simple cubic lattice). The lattice sites can have an occupancy of zero (unoccupied), one (unbound), or two (bound or misbound), where the number indicates how many domains are present at that site (Figure 2.1). Bound states are defined to be only those in which the two binding domains occupying the same lattice site have fully complementary sequences; if the sequences are not fully complementary, it is defined as a misbound state.

The primary challenge in designing a model at this level of resolution for the simulation of DNA origami self-assembly is to account for the constraints imposed by the double-helical twist on the structure of the system. There are two aspects

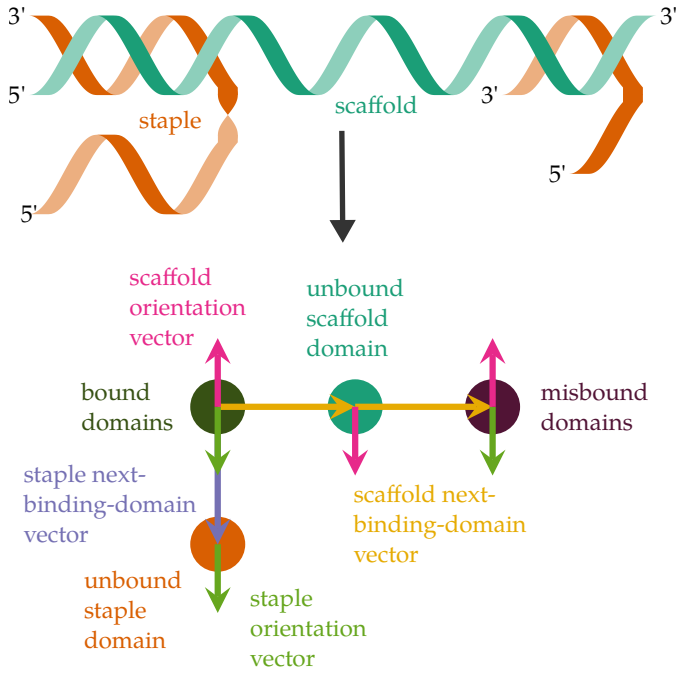

Figure 2.1: Schematic illustrations of the basic elements of the model. A cartoon helix representation is given on the top, and on the bottom is a representation of the same configuration in our model. There is one scaffold with three binding domains, one staple with two binding domains, and one staple with a single binding domain.

to this challenge. First, we need constraints to restrict where a strand crossover can occur between parallel helices. This is necessary because the strands of two adjacent parallel helices are only in a position compatible with a strand crossover at certain intervals of base pairs along the helices. Second, some way of transmitting information on the current phase of the helical twist along adjacent binding domains in the same helix is needed.

By associating an orientation unit vector with each binding domain, it is possible to create a set of rules that meets both requirements. In arguing for the form of the model, we will refer to diagrams of an idealized double-helical structure of DNA, rather than a fully atomistic model, which is sufficient for the level of detail we are targeting in the design of our model. In a bound or misbound state, we define the orientation vector as the vector which points out orthogonally from the helical axis to the position of the strand at the end of the helix in the current binding domain. In the case of a scaffold chain, the positive direction is defined as 5′ to 3′, while in the case of a staple chain, it is defined as 3′ to 5′.

Suppose two particles in our model occupy a given lattice site. According to the above definition of an orientation vector, if the lattice site is in a bound or misbound

state, the orientation vectors of the two binding domains must add up to zero. As an example, consider a system with two fully complementary pairs of 16-nucleotide (nt) binding domains, as in Figure 2.2(a)(i). For visualization purposes, it is convenient to represent a point on the lattice as corresponding to the centre of double helix. At the end of the first binding domain, the scaffold strand (teal) is at the bottom of the helix, and so the orientation vector for the binding domain of the scaffold strand (pink arrow) points downwards from the centre line. By contrast, the staple strand (orange) is at the top of the helix at the same point, and so the orientation vector for the binding domain of the staple strand (green arrow) points upwards from the centre line. In an unbound state, the direction of the orientation vector is uniformly distributed.

The orientation vector thus clearly contains information about the current phase of the twist at the end of the helix in the binding domain. In order for the model to be consistent with helical geometry, which here is assumed to correspond to B-DNA, when two adjacent binding domains are in the same helix, the dihedral angle between the planes defined by the orientation vectors and the vector connecting the two domains must be determined by the number of turns of the helix between them (Figure 2.2(b)). We shall refer to the unit vector that connects two binding domains as the 'next-binding-domain vector', since for a given binding domain, it points to the next binding domain along the chain. It is important to differentiate the next-binding-domain vector from a vector that is parallel to the helical axis. The version of the model that is used in the simulations presented in Chapter 4 and Chapter 5 does not define an explicit helical axis, a decision made to reduce simulation time by reducing configurational space.

Because there is no explicit helical-axis vector in the model, a single pair of bound domains will only implicitly define the helical axis to lie within a plane (Figure 2.2(d)). The helical axis is not resolved until an adjacent binding domain enters a bound state in the same helix. If a binding domain contiguous to one of two bound domains enters a bound state that is not in the same helix, then the helical axes of the first bound pair and the new bound pair will become more restricted in a configuration dependent way, but will not be fully resolved. This will be discussed further in Section 2.2.2 and Section 2.2.3. We briefly consider an alternative model in which an explicit helical axis is defined in Section 2.3.

The dihedral angle that we expect depends on the length of the binding domains. For example, in the case we considered above (Figure 2.2(a)), each binding domain corresponds to 1.5 turns of the helix. If two adjacent bound domains are in the

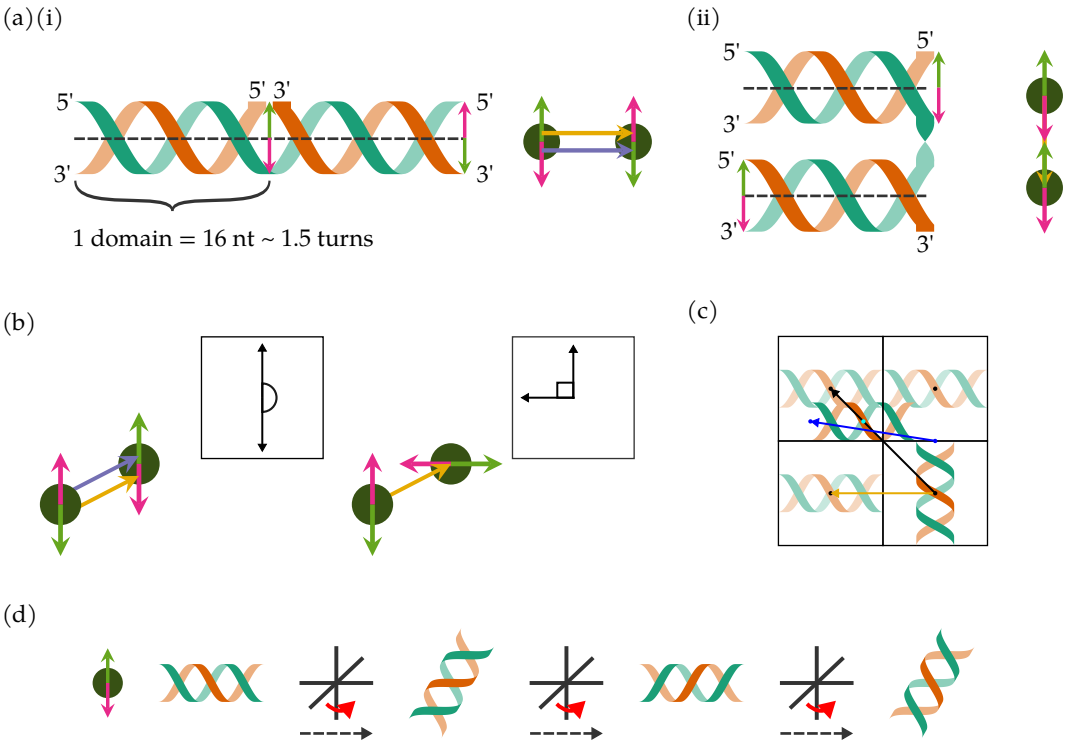

Figure 2.2: Representation of helices in the model. (a) Two 16-nt long binding domains, which in B-form DNA corresponds to about 1.5 turns of the helix. In (i), both binding domains are part of the same helix, while in (ii), they are part of separate helices with a kink between them. (b) Orientation vectors and helical phase. The boxes are projections of the scaffold orientation vectors of the two binding domains onto a plane normal to the next-binding-domain vector, with the dihedral angle indicated. Left: the orientation vectors are consistent with two stacked 16-nt (1.5 turn) binding domains. Right: the orientation vectors are consistent with two stacked 8-nt (0.75 turn) binding domains. (c) Physical interpretation of the next-binding-domain vector when considering two orthogonal helices. The helix of the first binding domain is drawn on the bottom right lattice site, while the helix of the second binding domain is drawn after rotation, as well as centered on the remaining three lattice sites (at 50% opacity). Black points are place at the centre of each lattice site. The black vector points from the centre of the first binding domain to the centre of the top left lattice site, which happens to pass through the centre of the second binding domain (marked by a teal point). The blue vector points from the end of the first binding domain to the end of the second binding domain. The yellow vector is the result of coarse-graining the blue vector to the next-binding-domain vector of our model. (d) A single pair of bound domains only constrains the helical axis to a plane.

same helix, we therefore expect a dihedral angle of 180°; the orientation vectors of the scaffold (pink) and staple (green) strands in Figure 2.2(a)(i) must therefore alternate in sign when they are part of the same helix. More generally, because we are restricted to a simple cubic lattice, all binding domains that may be modelled fall into

four classes, based on the dihedral angle prescribed by the number of turns. These classes differ by the fraction of a turn that remains after the n whole turns that make up the double helix for a binding domain in a hybridized state. We refer to these as whole-turn binding domains (remainder of zero), quarter-turn binding domains, half-turn binding domains, and three-quarter-turn binding domains. However, whole-turn binding domains are not useful in origami design, at least using our definition of binding domains, as they do not allow for crossovers between parallel helices, so we do not consider them further.

The mapping between the model geometry and more detailed representations is not intended to be exact. There are two related issues that should be mentioned here. First, parallel helices with crossovers between them are represented as being on adjacent domains, yet the distance between the centres of helices along the same helix and across parallel helices will in general not be the same. For small binding domains, these distances are approximately the same, but for longer binding domains, the approximation will become progressively worse. We consider 8-base pair (bp) and 16-bp long binding domains in this work; an 8-bp binding domain is about 2.7 nm long, while a 16-bp is about 5.5 nm long. Using B-form DNA geometry and assuming approximately 1 nm spacing between parallel helices in an origami structure [9], this gives a distance of 3 nm between the centres of parallel helices. For the level of detail of this model, these ratios are deemed acceptable.

The second involves the physical meaning of the next-binding-domain vector. A simple way to visualize this vector is to have it point from the centre of the first binding domain to the centre of the second binding domain. While this is sufficient for unbound domains and binding domains that form a single helix, the mapping does not always work for bound domains that are not in the same helix. For two orthogonal helices, the vector would point approximately towards a lattice site diagonal to the current site, yet we only allow contiguous binding domains to be on adjacent lattice sites. To resolve this, we first note that we simply need a reasonable way to bin configurations represented at a higher level of detail to our lattice model in a consistent way. Thus, when binning configurations with orthogonal helices, the next-binding-domain vector is defined to point from the end of the first helix to the end of the second helix (Figure 2.2(c)). Because the assembled structures do not involve these configurations, we deem this approximation acceptable as the exact geometry of the partially assembled states is unlikely to be critical for purposes of our modeling. Finally, for two bound domains that are in parallel helices, as in Figure 2.2(a)(ii), it is sufficient to consider the next-binding-domain vector as pointing

from the centre of the first to the centre of the second.

2.2 Potential energy

The next step in developing the model is to define a potential energy function that captures the relevant details of these systems. The potential energy is composed of three primary terms,

$$U = U_{\text{bond}} + U_{\text{stack}} + U_{\text{steric}}, \tag{2.1}$$

where U_{bond} is the contribution from bonding between binding domains, U_{stack} is the contribution from base stacking between binding domains, and U_{steric} is the contribution from steric interactions. Both U_{stack} and U_{steric} are composed of further subterms that depend on two, three, or four pairs of bound domains.

2.2.1 Bonding term

To account for the fact that DNA strands bond with one another, we compute an energy of interaction for all bound or misbound lattice sites. This energy of interaction in our model is taken to be the hybridization free energy of the two strands that occupy the same lattice site, accounting not only for the energy of bonding, but also, in a coarse-grained way, for the entropy of hybridizing two molecules. Consequently, the interaction energies in our model are strongly temperature dependent. We compute the hybridization free energies associated with bound and misbound states using the unified-NN model [167, 181, 182], which are a function of both temperature and salt concentration. Here, we consider calculations only for fully hybridized segments, the relevant terms for which are

$$\Delta G^{\ominus}_{\text{NN}} = \Delta G^{\ominus}_{\text{initiation}} + \Delta G^{\ominus}_{\text{symmetry}} + \sum \Delta G^{\ominus}_{\text{stack}} + \Delta G^{\ominus}_{\text{AT terminal}}. \tag{2.2}$$

$\Delta G^{\ominus}_{\text{initiation}}$ is a sequence-independent initiation free energy, while $\Delta G^{\ominus}_{\text{AT terminal}}$ is a term to account for having a terminal AT pair, if applicable. $\Delta G^{\ominus}_{\text{symmetry}}$ accounts for palindromic sequences, but since staples are designed never to be palindromic to prevent self binding, this term can be ignored. Finally $\Delta G^{\ominus}_{\text{stack}}$ is the stacking-free-energy term, which is calculated for all (overlapping) pairs along the sequence. The parameters of the unified-NN model as given in [167] assume a standard state amount concentration of 1 M. Parameters are provided for both the enthalpic and entropic contribution, allowing for inclusion of the temperature dependence ($\Delta G^{\ominus}_{\text{NN}} = \Delta H^{\ominus}_{\text{NN}} - T\Delta S^{\ominus}_{\text{NN}}$), where T is the temperature. Sodium ion dependence

can also be taken into account through an empirical relation,

$$\Delta S^{\ominus}_{\mathrm{NN}}([\mathrm{Na}^+]) = \Delta S^{\ominus}_{\mathrm{NN}} + \frac{0.368N}{2} \ln\left(\frac{[\mathrm{Na}^+]}{[\ominus]}\right), \tag{2.3}$$

where $\Delta S^{\ominus}_{\mathrm{NN}}$ is the standard-state entropy at 1 M NaCl, N is the number of phosphate groups in the DNA strand, $[\mathrm{Na}^+]$ is the sodium ion amount concentration, and $[\ominus]$ is the standard state amount concentration. In the case of partially complementary sequences, the hybridization free energy is approximated by the predicted free energy for the longest contiguous complementary sequence of the pair; this approximation has been shown to work well when simulating DNA bricks [190, 191, 200].

To formally map the hybridization free energies from the unified-NN model to the interaction energy of two binding domains in our model, we follow the approach of Reinhardt and Frenkel [192]. Here, we must consider two different types of reactions: intermolecular, in which a binding domain on a free staple hybridizes to a binding domain in a scaffold system, where a scaffold system refers to a single scaffold strand and any staples bound directly or indirectly to it, and intramolecular, in which two binding domains in a scaffold system hybridize. In the first case, we have a bimolecular reaction with an equilibrium constant K_{b} of the hybridization reaction between a staple binding domain A and a scaffold binding domain B,

$$\mathrm{A} + \mathrm{B} \rightleftharpoons \mathrm{AB}, \tag{2.4}$$

$$K_{\mathrm{b}} = \frac{[\mathrm{AB}][\ominus]}{[\mathrm{A}][\mathrm{B}]} = \frac{C_{\mathrm{AB}}C^{\ominus}}{C_{\mathrm{A}}C_{\mathrm{B}}} = \mathrm{e}^{-\beta\Delta G^{\ominus}_{\mathrm{NN}}}, \tag{2.5}$$

where C_x is the number density (i.e. number concentration) of x, $C^{\ominus} = \mathrm{N_A}[\ominus]$ is the standard state number density, $\mathrm{N_A}$ is Avogadro's constant, and $\beta = 1/k_{\mathrm{B}}T$, with T being the temperature, and k_{B} being the Boltzmann constant. In equilibrium,

$$\mu_{\mathrm{A}} + \mu_{\mathrm{B}} = \mu_{\mathrm{AB}}, \tag{2.6}$$

where μ_x is the chemical potential of x.

To a first approximation, we can assume that the binding domains behave ideally with respect to each other, allowing the partition functions of the binding domains in our model, $Q_x(N_x, V, T)$, to be expressed in terms of the internal partition functions of the binding domains, $q_x(V, T)$,

$$Q_x = \frac{V_{\mathrm{L}}^{N_x}}{N_x!} q_x^{N_x}, \tag{2.7}$$

where V_{L}, the system lattice volume, is a dimensionless quantity as it represents a sum over all the lattice sites in the system, N_x is the number of x present in V_{L}, and the function notation on the partition functions has been dropped for notational simplicity. Then using the relation

$$\mu_x = -k_{\mathrm{B}}T\left(\frac{\partial \ln Q_x}{\partial N_x}\right)_{V,T} \tag{2.8}$$

and Stirling's approximation, the chemical potential can be written in terms of the internal partition function,

$$\mu_x = k_{\mathrm{B}}T\ln\left(\frac{\rho_x}{q_x}\right), \tag{2.9}$$

where ρ_x is the lattice number density of x. The lattice number density can be related to the number density with

$$\rho_x = a^3 C_x, \tag{2.10}$$

where a is the lattice constant, which has units of length.

Plugging in Equation (2.9) and Equation (2.10) to Equation (2.6) and rearranging, we get

$$\frac{C_{\mathrm{AB}}}{C_{\mathrm{A}}C_{\mathrm{B}}} = a^3\frac{q_{\mathrm{AB}}}{q_{\mathrm{A}}q_{\mathrm{B}}}. \tag{2.11}$$

Comparing to Equation (2.5), we can multiply through by $C^{\ominus}$ to obtain

$$\frac{q_{\mathrm{AB}}}{q_{\mathrm{A}}q_{\mathrm{B}}} = \frac{\mathrm{e}^{-\beta\Delta G^{\ominus}_{\mathrm{NN}}}}{a^3 C^{\ominus}}. \tag{2.12}$$

Given that each binding domain has an orientation vector with six possible configurations, the internal partition functions become

$$q_{\mathrm{A}} = q_{\mathrm{B}} = 6, \quad q_{\mathrm{AB}} = 6\mathrm{e}^{-\beta\epsilon_{\mathrm{b}}}, \tag{2.13}$$

where ϵ_{b} is the bimolecular binding domain interaction energy of our model. Plugging in Equation (2.13) to Equation (2.11), we can solve for ϵ_{b},

$$\frac{\mathrm{e}^{-\beta\epsilon_{\mathrm{b}}}}{6} = \frac{\mathrm{e}^{-\beta\Delta G^{\ominus}_{\mathrm{NN}}}}{a^3 C^{\ominus}} \tag{2.14}$$

$$\rightarrow \epsilon_{\mathrm{b}} = \Delta G^{\ominus}_{\mathrm{NN}} + k_{\mathrm{B}}T\ln\left(a^3 C^{\ominus}\right) - k_{\mathrm{B}}T\ln 6. \tag{2.15}$$

For the second case, where we are considering two binding domains already in the scaffold system, we have a unimolecular reaction with an equilibrium constant K_{u} of the hybridization reaction between a system with two unbound domains C

and a system with the two binding domains hybridized D,

$$\mathrm{C} \rightleftharpoons \mathrm{D}, \tag{2.16}$$

$$K_\mathrm{b} = \frac{[\mathrm{C}]}{[\mathrm{D}]} = \frac{C_\mathrm{C}}{C_\mathrm{D}} = \frac{\rho_\mathrm{C}}{\rho_\mathrm{D}} = \mathrm{e}^{-\beta \Delta G^{\ominus}_{\mathrm{NN,\,u}}}, \tag{2.17}$$

where $\Delta G^{\ominus}_{\mathrm{NN,\,u}}$ is the unimolecular unified-NN standard Gibbs free energy of hybridization. Because the $\Delta G^{\ominus}_{\mathrm{initiation}}$ term captures the translational entropy cost of combining two free strands into one [167], we will assume that

$$\Delta G^{\ominus}_{\mathrm{NN,\,u}} = \Delta G^{\ominus}_{\mathrm{NN}} - \Delta G^{\ominus}_{\mathrm{initiation}}. \tag{2.18}$$

We also assume that the scaffold systems act ideally with respect to each other, but now the treatment of the internal partition function is more complicated. To make the problem tractable, we treat the binding domains within the scaffold system as being independent if they are not hybridized to each other. Then, as before, we only need to consider the internal partition functions of the two unbound domains and the hybridized binding domains, allowing us to solve for the unimolecular binding domain interaction energy ϵ_u,

$$\frac{\rho_\mathrm{D}}{\rho_\mathrm{C}} = \frac{q_\mathrm{D}}{q_\mathrm{C}} = \frac{q_\mathrm{AB}}{q_\mathrm{A} q_\mathrm{B}} = \frac{\mathrm{e}^{-\beta \epsilon_\mathrm{u}}}{6} = \mathrm{e}^{-\beta \Delta G^{\ominus}_{\mathrm{NN,\,u}}} \tag{2.19}$$

$$\rightarrow \epsilon_\mathrm{u} = \Delta G^{\ominus}_{\mathrm{NN,\,u}} - k_\mathrm{B} T \ln 6. \tag{2.20}$$

Because this is a unimolecular reaction, $\Delta G^{\ominus}_{\mathrm{NN,\,u}}$ is not dependent on the standard state concentration, and so the model interaction energy does not need to have a term with the standard state concentration to be independent of changes to the standard state.

In order to calculate a chemical potential of a staple from a given concentration, we assume staples act ideally when in solution. The canonical partition function for the staples of type i is

$$Q_i = \frac{(q_i V_\mathrm{L})^{N_i}}{N!} = \frac{\left(6^{2n_i - 1} V_\mathrm{L}\right)^{N_i}}{N_i!}, \tag{2.21}$$

where N_i is the number of staples of strand i, and n_i is the number of binding domains that the staple strand comprises. The chemical potential of staple strand i is then

$$\mu_i = k_\mathrm{B} T \left[\ln\left(a^3 C_i\right) - (2n_i - 1) \ln 6\right], \tag{2.22}$$

where k_B is Boltzmann's constant.

If we derive the melting temperature of the model for a single-binding-domain scaffold strand, we can compare it to the melting temperature of the unified-NN model assuming a two state reaction to verify the derivation of our potential. The average occupancy can be calculated exactly for this system, and it does not require the more complicated terms of the model detailed in the next sections. We will calculate this value using the grand ensemble, where the system volume is the number of scaffold binding domains. The grand partition function is

$$\Xi(\mu, V, T) = \sum_{N=0}^{1} e^{\beta\mu N} \sum_i e^{-\beta U_i} \tag{2.23}$$

$$= \sum_i e^{-\beta U_i} + e^{\beta\mu} \sum_i e^{-\beta U_i} \tag{2.24}$$

$$= 6 + \frac{6^2 e^{\beta\mu} e^{-\beta\Delta G^{\ominus}_{\mathrm{NN}}}}{a^3 C^{\ominus}}, \tag{2.25}$$

where the inner sum in the first line and the sums in the second line are over states with N bound staples with potential energy U_i, and in the third line we have plugged in Equation (2.15) and simplified. The average occupancy is

$$\langle s \rangle = \frac{1}{\Xi} \sum_N e^{\beta\mu N} \sum_i s e^{-\beta U_i} \tag{2.26}$$

$$= \frac{6 e^{\beta\mu} e^{-\beta\Delta G^{\ominus}_{\mathrm{NN}}}}{a^3 C^{\ominus} + 6 e^{\beta\mu} e^{-\beta\Delta G^{\ominus}_{\mathrm{NN}}}}, \tag{2.27}$$

where s is 0 when the single scaffold binding is unbound and 1 when it is bound. At the melting temperature, T_m, the average occupancy of the scaffold lattice site by a staple binding domain is $1/2$, thus

$$1 = \frac{e^{\beta\mu} 6 e^{-\beta\Delta G^{\ominus}_{\mathrm{NN}}}}{a^3 C^{\ominus}} \tag{2.28}$$

$$1 = \frac{C}{C^{\ominus}} e^{-\beta(\Delta H^{\ominus}_{\mathrm{NN}} - T\Delta S^{\ominus}_{\mathrm{NN}})} \tag{2.29}$$

$$\rightarrow T_m = \frac{\Delta H^{\ominus}_{\mathrm{NN}}}{k_B \ln\left(\frac{C}{C^{\ominus}}\right) + \Delta S^{\ominus}_{\mathrm{NN}}}, \tag{2.30}$$

where in the second line we have used Equation (2.22) with $n = 1$ and simplified.

The melting temperature for the unified-NN model assuming a two state reaction can be derived directly. If when considering Equation (2.4) we let A be the staple and B be the scaffold, and let $[A]_T >= [B]_T$, where the subscript denotes the total concentration (i.e. including the staple and scaffold binding domains when they are

in the bound AB state), and if we consider an initial state with all B being bound in the AB form, then in equilibrium we have

$$[A] = [A]_{\mathrm{T}} - [B]_{\mathrm{T}} + x, \quad [B] = x, \quad [AB] = [B]_{\mathrm{T}} - x, \tag{2.31}$$

where x is the change in concentration. At the melting temperature, [B] = [AB]; plugging in this and Equation (2.31) to Equation (2.5) and rearranging gives

$$K = \frac{[A]_{\mathrm{T}}}{[\ominus]} - \frac{[B]_{\mathrm{t}}}{2[\ominus]} = \mathrm{e}^{-\beta \Delta G^{\ominus}_{\mathrm{NN}}} \tag{2.32}$$

$$\rightarrow T_{\mathrm{m}} = \frac{\Delta H^{\ominus}_{\mathrm{NN}}}{k_{\mathrm{B}} \ln\left(\frac{[\mathrm{A}]_{\mathrm{T}}}{[\ominus]} - \frac{[\mathrm{B}]_{\mathrm{T}}}{2[\ominus]}\right) + \Delta S^{\ominus}_{\mathrm{NN}}}, \tag{2.33}$$

$$\simeq \frac{\Delta H^{\ominus}_{\mathrm{NN}}}{k_{\mathrm{B}} \ln\left(\frac{[\mathrm{A}]}{[\ominus]}\right) + \Delta S^{\ominus}_{\mathrm{NN}}}, \tag{2.34}$$

$$= \frac{\Delta H^{\ominus}_{\mathrm{NN}}}{k_{\mathrm{B}} \ln\left(\frac{C_{\mathrm{A}}}{C^{\ominus}}\right) + \Delta S^{\ominus}_{\mathrm{NN}}}, \tag{2.35}$$

where the asymptotic equality follows when $[\mathrm{A}]_{\mathrm{T}} >> [\mathrm{B}]_{\mathrm{T}}$, which we will assume in our model (see Section 3.3 for further discussion of this assumption). Comparing this with Equation (2.30), we see that the melting temperatures agree.

While the model melting temperature agrees with the unified-NN two-state melting temperature for a single binding domain scaffold, it will not hold for anything longer because of the oversimplified assumption of the internal partition function of the system binding domains being independent. The internal partition function of the system is highly non-trivial, so the best we can do is use an mean field approach to give an average difference of the logarithm of the partition function with a change in the binding state of the system. It is important to keep in mind, however, that even if we could calculate the ratio of the partition functions exactly in order to correct the unified-NN hybridization free energy for every possible hybridization reaction, we would be creating a model in which the individual binding domains hybridize with the same statistics as the unified-NN model. This is not the expected behaviour for DNA origami binding domains. It is precisely the deviation from the NN model in these internal hybridization reactions that we are interested in studying. It is here that the advantage of using a model with a physical basis over the more statistical models discussed in Section 1.4 becomes apparent, as these deviations, which are entropic in nature, are naturally present up to some constant, and so the entropy differences will be roughly captured.

To understand why some correction is still needed, consider that while the coop-

erativity involved in DNA origami self-assembly is expected to change the overall slope of a curve of an order parameter against temperature, the curve should not be shifted to overall higher or lower melting temperatures relative to a pure unified-NN. If a fully assembled state has only one allowed configuration, then without further modification, the model as defined will have a melting temperature that is dependent on the choice of binding domain size. Consider a particular design represented in two different ways, where the second has binding domains defined as being twice as small as the first. Because in both cases the assembled state has just one configuration, the loss in entropy will be twice as large for the second system, which will shift the assembly to lower temperatures.

The above mentioned mean field correction can allow for such an overall correction. However, fully assembled states will not in general have only one configuration in our model. The number of states available will depend on how the fully assembled state is defined, how the remaining terms of the potential are defined, and on the specific design being considered. We can begin by considering the most extreme case, where there is only one configuration available in the bound state to give an upper bound on the absolute value of the correction. If each binding domain has six relative positions and six orientation vectors, then upon hybridization of two binding domains,

$$q_{\mathrm{A}} = q_{\mathrm{B}} = 6^2, \quad q_{\mathrm{AB}} = \mathrm{e}^{-\beta\epsilon_{\mathrm{u}}} \tag{2.36}$$

$$\frac{\rho_{\mathrm{D}}}{\rho_{\mathrm{C}}} = \frac{q_{\mathrm{D}}}{q_{\mathrm{C}}} = \frac{q_{\mathrm{AB}}}{q_{\mathrm{A}}q_{\mathrm{B}}} = \frac{\mathrm{e}^{-\beta\epsilon_{\mathrm{u}}}}{6^4} = \mathrm{e}^{-\beta\Delta G^{\ominus}_{\mathrm{NN,\,u}}} \tag{2.37}$$

$$\rightarrow \epsilon_{\mathrm{u}} = \Delta G^{\ominus}_{\mathrm{NN,\,u}} - 4k_{\mathrm{B}}T\ln 6. \tag{2.38}$$

For binding of a staple to a partially assembled scaffold, we will have a different expression, as the first binding event is a change in absolute position rather than relative position,

$$q_{\mathrm{A}} = 6, \quad q_{\mathrm{B}} = 6^2, \quad q_{\mathrm{AB}} = \mathrm{e}^{-\beta\epsilon_{\mathrm{b}}} \tag{2.39}$$

$$\frac{q_{\mathrm{AB}}}{q_{\mathrm{A}}q_{\mathrm{B}}} = \frac{\mathrm{e}^{-\beta\epsilon_{\mathrm{b}}}}{6^3} = \frac{\mathrm{e}^{-\beta\Delta G^{\ominus}_{\mathrm{NN}}}}{a^3C^{\ominus}} \tag{2.40}$$

$$\rightarrow \epsilon_{\mathrm{b}} = \Delta G^{\ominus}_{\mathrm{NN}} + k_{\mathrm{B}}T\ln\left(a^3C^{\ominus}\right) - 3k_{\mathrm{B}}T\ln 6. \tag{2.41}$$

Finally, special consideration must be made for the overall system's rotational entropy, which is not lost in the final assembled state. If one considers the first three binding domains of the scaffold, the second will always have six relative positions

to the first by rotation the entire system, and the third will always have 4 positions relative to the second by rotation the entire system around the bond axis between the first and the second binding domains. There is also no relative positional entropy to lose when binding the first scaffold domain. Thus, for the first staple to bind to the scaffold, we add $2k_BT\ln 6$ to ϵ_b, while for the second scaffold domain to bind, whether to another binding domain on the first staple or to a new staple, we add $k_BT\ln 6 - k_BT\ln 2 = k_BT\ln 3$ to either ϵ_u or ϵ_b, respectively.

2.2.2 Stacking term

If two contiguous domains are in bound states and part of the same helix, then there is an additional stacking term that we have not yet accounted for when calculating the unified-NN model hybridization free energy for the two domains separately. There are two cases to consider at the level of resolution of our model. The first case is that there are only two strands involved with each having two contiguous binding domains, in which case, as discussed in Section 2.1, there is only one allowed configuration for the orientation vectors involved (Figure 2.3(a)(i)). Here, it is reasonable to add the $\Delta G^{\ominus}_{\text{stack}}$ corresponding to the nucleotides involved to Equation (2.20), and the mean field entropy correction discussed in Section 2.2.1 would now be a local description for the binding of the second pair of domains.

The second case is that only one pair of binding domains is contiguous (Figure 2.3(a)(ii)–(iv)). This could involve one strand ending and another beginning (this is sometimes referred to as a nick in the backbone of one of the strands forming the helix), or one or both may continue and possibly even be a part of another helix via a crossover junction. We will refer to the point at which any of these situations occur as breakpoints. Breakpoints are able to become unstacked to form kinks. In the model, if the orientation vectors of a pair of contiguous bound domains do not have a configuration prescribed by the helical geometry, they are considered to have a kink and are treated as two separate helices (for an example see Figure 2.2(a)(ii)). In other words, the pair of bound domains can either be in a stacked configuration or a kinked configuration, and the rules that were laid out for a pair of bound domains being part of the same helix can also be referred to as the rules for being stacked. In contrast to the first case in which there is no breakpoint, the stacking and domain binding events are independent, and a separate stacking energy term is required, ϵ_s (Figure 2.3(a)(iii) and (iv)). As this is actually a free energy describing the difference between being stacked and kinked, it will in principle be different from the $\Delta G^{\ominus}_{\text{stack}}$ term. We discuss the parameterization of this term in Section 4.3.

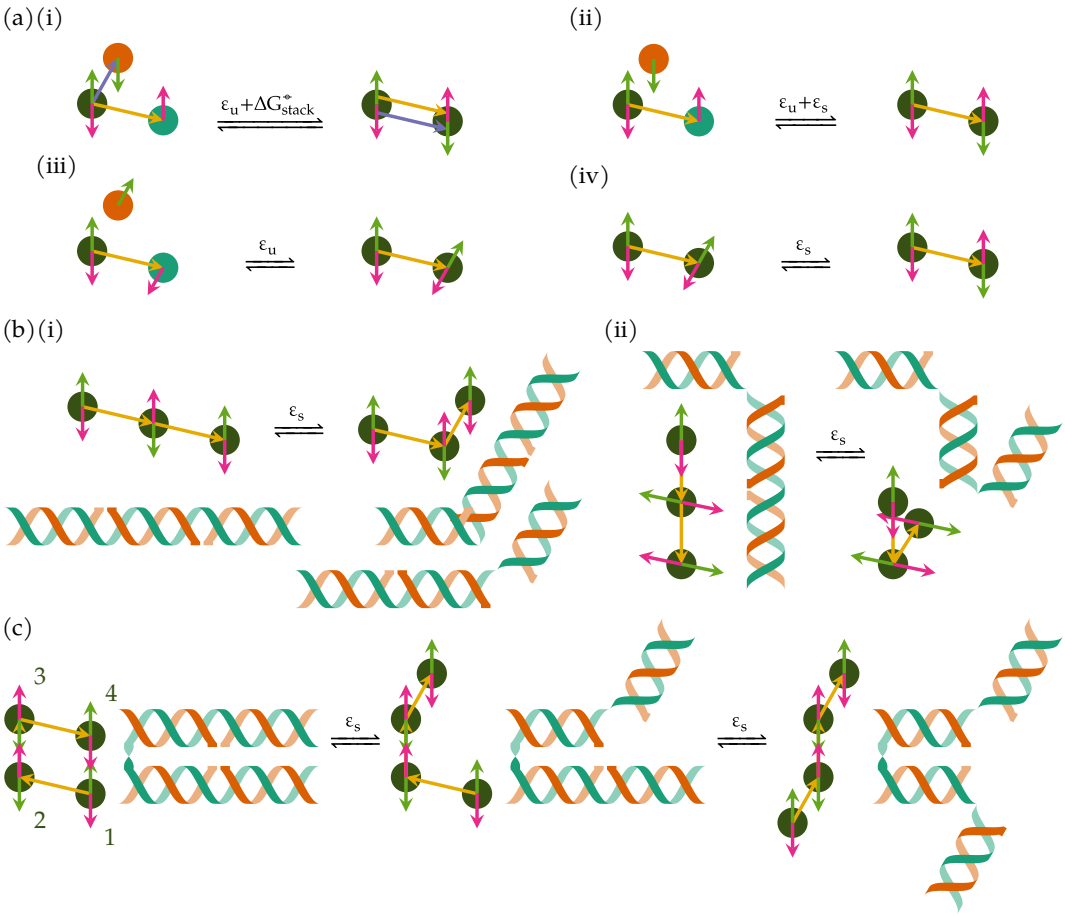

Figure 2.3: Helical stacking in the model. (a) Helical stacking with two bound domain pairs. In (i), there are two pairs of contiguous binding domains bound to each other, such that a single helix is formed. In (ii)–(iv), only one pair of binding domains is contiguous, allowing for unstacked, or kinked, configurations to form. While not drawn, we assume here that the staple domain is part of a staple that is bound by another one of its domains to the scaffold system. While stacking and bonding can occur in one step as in (ii), it is also possible for bonding to occur first, as in (iii), followed by stacking (iv). (b) Helical stacking with three bound-domain pairs. In (i), two model configurations are shown that are both pairwise stacked, but the configuration on the right has one less stacking interaction. In (ii), two model configurations are shown with just one pairwise stack, but the configuration on the right has one less stacking interaction. (c) Helical stacking with four bound-domain pairs. All three configurations have pairwise stacks, but the configuration in the middle has only one stacking interaction, and the one on the right has none.

Whether or not two pairs of bound domains are stacked or kinked cannot always be determined by considering pairwise interactions in our model. This is because the helical axis is defined implicitly, as discussed in Section 2.1. Consider three pairs of bound domains that occupy adjacent lattice sites, in which at least one of the strands

is contiguous between adjacent lattice sites. There are multiple configurations for which the pairwise stacking rule is obeyed for both pairs of bound domains, two of which are shown in (Figure 2.3(b)(i)). However, all but one of these configurations involve a right angle bend in the helix. The length of the binding domains is typically well below the persistence length of double-stranded DNA, so configurations with such sharp turns are extremely unlikely with no external force. Therefore, these configurations must have a kink at one of the breakpoints, and are in fact composed of two separate helices. Such configurations will then only have one of the two stacking interaction energies. If there are two breakpoints, there is ambiguity in which breakpoint is actually kinked, and so the stacking interaction in our model is not entirely local (Figure 2.3(b)(i)).

If we consider the definition of the next-binding-domain vector as given in Section 2.1, we note that there are configurations for two bound-domain pairs with orthogonal helices that also map to model configurations that are defined as being stacked. Configurations only map to one model configuration, so in effect these configurations are not included in the model. We have made this choice to allow stacking to be defined between pairs of bound domains without adding additional degrees of freedom to the model. Once there are more than two bound domains involved these configurations are included by applying the stacking term described above and shown in Figure 2.3(b)(i). However, the model has no way to represent both of these kinds occurring at the same time. Considering that there are many ways for the model to represent kinked configurations, this exclusion seems reasonable to allow for a simpler model.

For configurations in which the second binding-domain's helix is orthogonal to the first binding-domain's helix, it is not possible for a third binding domain to form a stacked configuration with the second binding domain and for that resulting helix to be in the same plane as the first. However, without any further terms, it is possible to construct model configurations in which the this is the case; an example is given in Figure 2.3(b)(ii). In order to make the model consistent, an additional term could be applied such that these configurations, while containing a pairwise stacked configuration, would be defined to have no stacking interaction. However, for quarter- and three-quarter-turn binding domains, one of the model configurations between two bound-domain pairs can be mapped to from either a helix that is orthogonal or parallel to the first binding-domain's helix (see Section 2.2.3 for further discussion of all sterically allowed configurations and their mappings). Because the configurations involving stacked parallel helices with crossovers are critical to origami

designs, it is important not apply a term that prevents stacking of additional binding domains. A simple solution is to not remove the pairwise stacking interaction in configurations in which this ambiguity exists; these are configurations in which the first binding-domain's helical axis is equal to its orientation vector (again see Section 2.2.3 for further discussion).[1]

For configurations in which the second binding-domain's helix is parallel to the first (i.e. those which are involved in crossovers, which are important to represent correctly to model assembled configurations), there is additional complexity in determining whether configurations are stacked or kinked. These configurations map to model configurations in which the first binding domain's next-binding-domain vector is equal to its orientation vector. When both bound-domain pairs on either side of the breakpoint have another bound-domain pair that is contiguous to at least one of the involved strands, it is possible to construct model configurations which have two pairwise stacks that map to configurations that have only one or no stacked bound domain pairs. Using the indices in Figure 2.3(c), if the first binding-domain's next-binding-domain vector is antiparallel to the third binding-domain's next-binding-domain vector, then there are two stacking interactions. If they are orthogonal, then there is one stacking interaction. If the first binding-domain's next-binding-domain vector is parallel to the third binding-domain's next-binding-domain vector, then there are no stacking interactions.[2]

[1] In simulations presented in Chapter 4 and Chapter 5, we use a version of the model where we use a less restrictive rule that applies to four pairs of bound domains: if both of the bound domains with a breakpoint between them have a defined helical axis (which requires one additional pair of bound domains for each), the orientation vector of the first binding domain forming the kink is not equal to its next-binding-domain vector, and the two helical axes are parallel to each other and orthogonal to the next-binding-domain vector between the binding domains forming the kink, then there must be an additional kink present that these four bound domains are involved in and thus one less stacking interaction. This turns out to be the least restrictive rule to prevent unwanted unphysical fully assembled configurations, but it can only be justified by the fact that it prevents such configurations.

[2] In simulations presented in Chapter 4 and Chapter 5, we use a version of the model where we apply a less restrictive version of this term only to configurations that involve a double crossover between a pair of bound domains. In this version of the term, if the first binding-domain's next-binding-domain vector is antiparallel to the third binding-domain's next-binding-domain vector, then there are two stacking interactions; otherwise, there is one. Further, configurations where the first binding-domain's next-binding-domain vector is parallel to the third binding-domain's next-binding-domain vector are sterically prohibited (see Section 2.2.3 for discussion of sterically prohibited configurations).

2.2.3 Steric term

In describing the steric terms, we refer to 'constraints' and 'rules', but it should be understood that formally we are defining configurations that obey these constraints or rules as having a potential energy of zero for the term in question, and all others as having a potential energy of infinity. In principle these could also form a part of the definition of state space, as was done for the rule that orientation vectors on bound domains must be opposing, but we have found it more convenient to define these as part of the potential.

Because it is directly related to the discussion of stacked configurations above, we again consider three bound domain pairs on adjacent lattice sites. In particular, we consider the case in which there are no breakpoints, which occurs when there are two triplets of binding domains that are contiguous on the same strand. Again, there are multiple configurations for which the pairwise stacking rule is obeyed for both pairs of bound domains. Unlike the case when there is at least one breakpoint, it is not possible for there to be a kink to allow for the configurations that have a right angle bend. Therefore, all but the configuration in which there is no bend in the helix are disallowed.

Consider a pair of contiguous bound domains with a breakpoint. In reality, the breakpoint will not allow for all possible relative orientations of the two binding domains. By considering transformations to the two helical domains and making simple steric arguments, we can introduce further rules to account for this. Because the model is already so coarse, the particular choices made are unlikely to have much effect beyond changing the entropic balance between bound and unbound states, unless they affect whether crossovers are only able to occur where they are allowed. Thus it is sufficient to base our steric arguments on considerations of idealized cartoon helices. The correct entropic balance could be restored by considering a correction factor to the free energies of hybridization that could be determined by comparison to experiment, although we do not do so in this work.

Another consideration in constructing these terms is that by using more constrained potentials, sampling can become more difficult because the free-energy landscape becomes more rough. For example, in the extreme case of not allowing kinked configurations, which was the form the model originally took, sampling was very difficult as typically to rearrange the structure, the domains had to unbind and rebind. Therefore, the guiding principle in constructing the potential for kinked configurations was to ensure that crossovers between parallel helices only occur at the correct intervals, to make the partially assembled structures as unconstrained as

possible, and to achieve what physical realism we can with the steric arguments.

For all unique configurations involving two bound-domain pairs that have a breakpoint between them, all sterically allowed model configurations are illustrated in Figure 2.4(a) and (c) and all pairwise sterically prohibited configurations are illustrated in Figure 2.4(e). For helix cartoon configurations drawn in Figure 2.4, a 16-nt half-turn binding domain is used, but the general arguments here hold for all three binding-domain classes. To understand which configurations are possible, we must consider a number of rotations of the second binding-domain's helix relative to the first binding-domain's helix. Beginning from a stacked configuration, consider rotating the second binding-domain's helix around an axis parallel to the helical axis but displaced to the outside of the helix to produce the configurations in Figure 2.4(a)(ii)–(iv). We will refer to this as the first rotation axis. This allows for configurations in which the next-binding-domain vector of the first binding domain is perpendicular to the orientation vector of both the first and the second binding domains.

Following this first rotation with further rotations of the second binding-domain's helix around an axis parallel to the orientation vector of the first will not lead to any new relative orientations of the second binding-domain's orientation vector because of our definition of mapping binding domains that form orthogonal helices to lattice sites (see Section 2.1). We will refer to this as the second rotation axis. An example of the resulting cartoon helix configuration after rotating in one direction is shown below the first cartoon helix configuration in Figure 2.4(a)(ii)–(iv). There are no cartoon helix configurations that map to model configurations in which the second binding-domain's orientation vector is parallel or antiparallel to the next-binding-domain vector of the first. Thus, in our model, if the first binding-domain's next-binding-domain vector is perpendicular to its orientation vector, the second binding domain's orientation vector must also be perpendicular to next-binding-domain vector of the first.

Rotations around an axis perpendicular to the two previously mentioned rotation axes can lead to configurations in which the first binding domain's next-binding-domain vector is parallel or antiparallel to its orientation vector. We will refer to this as the third rotation axis. If a quarter turn is made such that first domain's next-binding-domain vector is parallel with its orientation vector, it leads to configurations with steric clashes, which is illustrated in Figure 2.4(b). A further quarter turn will only lead to worsening the steric clashes. Thus, configurations in which the first binding domain's next-binding-domain vector is antiparallel with its orientation

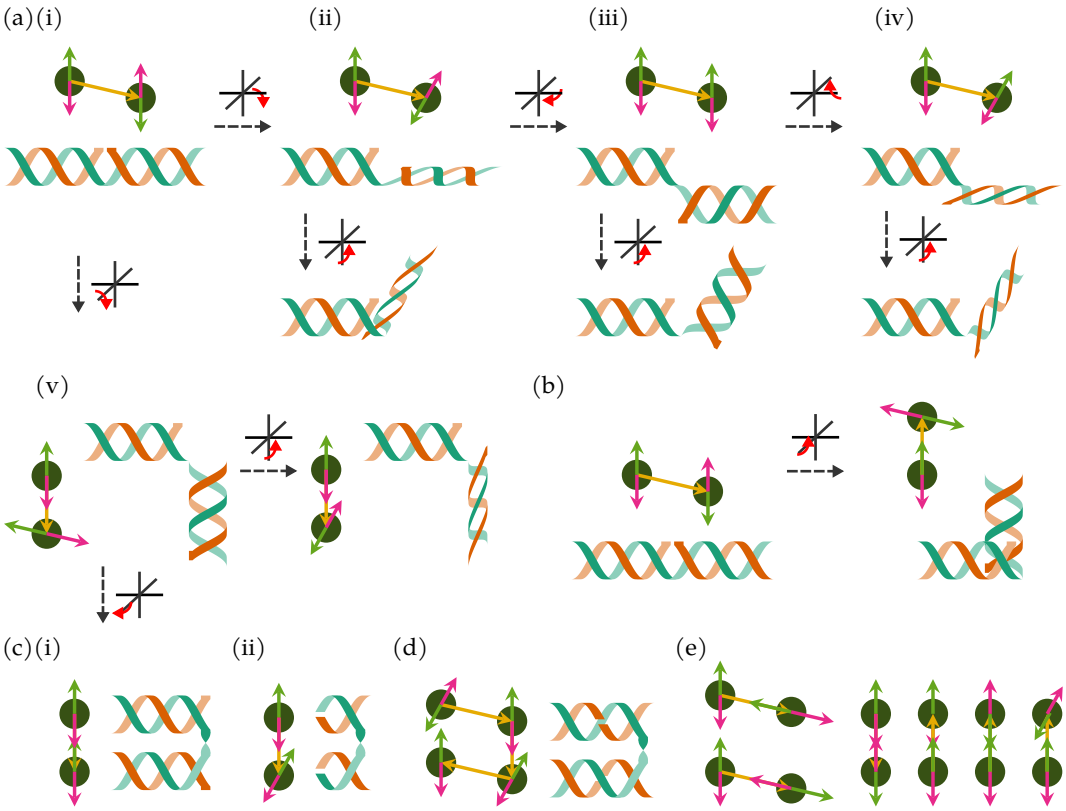

Figure 2.4: All twelve unique configurations of two bound-domain pairs with a breakpoint. (a) Five configurations that are sterically allowed for all binding-domain classes. Cartoon helix representations are shown for a 16-nt half-turn binding domain. (b) An example of a transformation that results in a sterically prohibited configuration. (c) A half-turn around the third rotation axis produces an additional allowed model configuration for half-turn binding domains (i), but not for quarter- or three-quarter-turn binding domains (ii). In (ii) a cartoon helix representation for an 8-nt three-quarter-turn binding domain is shown. (d) Sterically allowed configuration for an 8-nt three-quarter-turn binding domain with two pairs of stacked bound domain pairs connected via a crossover. (e) All six sterically prohibited configurations.

vector are entirely disallowed. Returning to the original stacked configuration and making a quarter turn around the third rotation axis in the opposite direction will lead to a configuration in which the second binding-domain's helical axis is orthogonal to the first Figure 2.4(a)(v). Unlike the previous orthogonal helix configurations, this configuration maps to a new allowed model configuration. Rotations of this configuration around the second rotation axis will configurations that map to the same model configuration.[3]

[3]In simulations presented in Chapter 4 and Chapter 5, we use a version of the model where for the half-turn binding domain systems, the second binding-domain's orientation was constrained to be parallel to the first binding-domain's orientation vector in these particular kinked configurations.

A further quarter turn around the third rotation axis leads to configurations in which the second binding-domain's helical axis is parallel to the first binding-domain's helical axis Figure 2.4(c). Such configurations are those that allow for crossovers between parallel helices, which are prevalent in the assembled state. The second binding-domain's orientation vector will depend on the length of the binding domain in these configurations. In general, relative to the first binding-domain's orientation vector, it will have the dihedral angle prescribed by its length along the next-domain vector parallel to the helical axis, followed by a flip in the plane normal to the first binding-domain's orientation vector. In the case of the half-turn binding domains, the second binding-domain's orientation vector will be parallel to the first binding-domain's orientation vector, producing a unique model configuration.

For quarter- and three-quarter-turn binding domains this will result in configurations that map to the model configuration in Figure 2.4(a)(v), respectively, so no new model configurations are produced, and neither configurations where the second binding domain's orientation vector is parallel or antiparallel to the first binding-domain's next-binding-domain vector are sterically allowed. That the crossover model configuration has more than one cartoon helix configuration that maps to it means there is a loss of information about the helical phase. Both the second cartoon helix of Figure 2.4(a)(v) and a cartoon helix configuration in which the second binding domain is rotated a half turn around the second rotation axis map to his model configuration, but only one correctly describes the crossover configuration. To deal with this will introduce an additional term that applies to configurations in which both bound-domain pairs are stacked with an adjacent bound-domain pair. Then, all four orientation vectors must be in the same configuration as they would be if all four bound domain pairs were stacked in a single helix (Figure 2.4(d)). This term is only critical if crossovers occur such that the final structure is not planar, as otherwise the information loss on the phase has no effect on the assembled structures.[4]

When there is more than one crossover between two helices, the helices become much more restricted in the configurations they are able to take relative to each other (compare Figure 2.5(a)(i) to (ii)). In particular, they will be forced to be roughly parallel. This is naturally captured by the model when there are crossovers between more than one set of binding domain pairs on two separate helices, as seen in Figure 2.5(b)(ii). However, when a single binding domain pair has a double crossover, something which can occur with half-turn binding domains, this will not be captured by the model as currently defined.

[4]In the simulations presented in Chapter 5, we use a version of the model where the three-quarter-turn binding domains do not have this term. All designs simulated have planar assembled states.

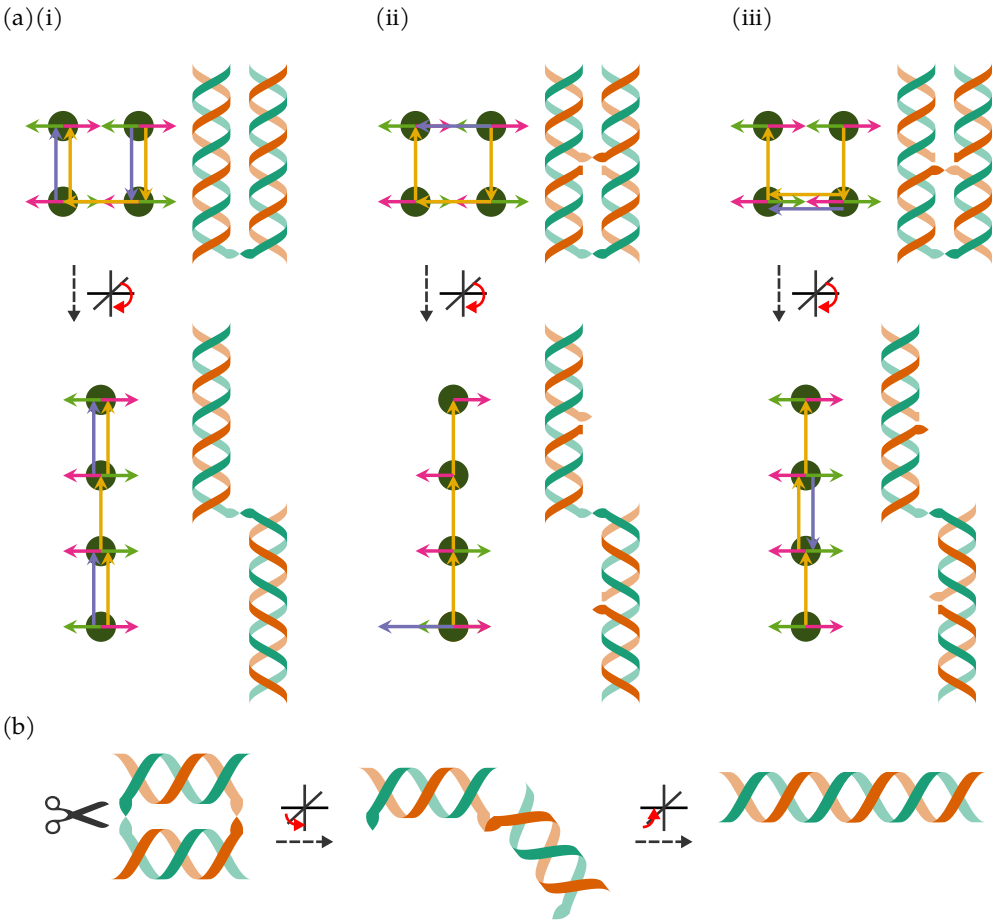

Figure 2.5: Strand crossovers between helices involving 16-nt long binding domains. (a) Crossovers between two adjacent parallel helices with a four-binding-domain scaffold. (i) Helices with a single crossover. (ii) Helices with two crossovers on separate binding domains. (iii) Helices with crossovers on the same binding domain. (b) Doubly contiguous binding domains in bound states.

Consider two adjacent lattice sites in bound states, where at least one pair of binding domains are contiguous. If the other pair of binding domains are not contiguous and in the same helix, their orientation vectors will still satisfy the prescribed helical angle because of the requirement of their orientation vectors to be opposing those of the strand that has two contiguous binding domains in that helix. However, the case in which both pairs of binding domains are contiguous requires further consideration. In reality, if the combined sequence of the two binding domains on one chain is together the reverse complement of the combined sequence of the two binding domains on the other chain, then the only way for all binding domains to be bound to each other is if there is only one helix. If instead the binding domains on one chain must be swapped to make the whole two-binding-domain sequence

the reverse complement of the other whole two-binding-domain sequence, then the only way for all binding domains to be bound to each other is if there are two parallel helices with both strands crossing over. As a concrete illustration, one of the chains would have to be cut and glued to its other end to transition between these two configurations (Figure 2.5(b)). Thus, the model constrains pairs of contiguous complementary binding domains bound to each other to be in the same helix if they are the full reverse complements of each, and to be crossing over if not.

2.3 Explicit helical axis model

One of the issues with an implicitly defined helical axis is that the potential energy function becomes somewhat convoluted, making it difficult to visualize and think about intuitively. Further, the terms that involve three and four binding domains require lengthy checks to determine whether the configuration is sterically prohibited and to calculate the total stacking energy, offsetting the gain in sampling efficiency that comes from a reduced number of degrees of freedom. The divorce of the number of stacked binding domains at breakpoints and thus the total stacking energy from pairwise stacked binding domains is particularly non-intuitive, and with the model as defined above, it is possible to construct configurations which have negative numbers of stacked pairs.

Because of the above considerations, here we speculatively outline a possible alternative model with an explicit helical axis for demonstrative purposes. The state space is expanded by introducing an additional vector to each binding domain, referred to as helical-axis vector. In the bound state, as the name suggests, it corresponds to the helical axis, and points in the positive direction of the chain. In the unbound state, in analogy with the orientation vector, it is uniformly distributed. Because the positive direction along a chain is defined in opposition between the scaffold and staple strands, when two binding domains are hybridized to each other, their helical-axis vectors must be equal. For simplicity, we apply the same rule to misbinding, as while in principle the helical axis vectors should be antiparallel if same-chain misbinding occurs, in practice it would not affect the results.

The mappings between physical configurations and model configurations for both the orientation vector and the next-binding-domain vector can be made more explicit. To be consistent with the interpretation of the orientation vector as pointing to the position of the strand at the end of binding domain orthogonally out from the helical axis, the helical-axis vector and the orientation vector are constrained to be perpendicular when the domains are in bound states. It is no longer necessary to

map configurations of two bound-domain pairs with either parallel or orthogonal helical axes to the same model configuration, as with an explicit helical axis they can be differentiated. If we follow the same guideline of defining the next-binding-domain vector to point from the end of first binding-domain's helical axis to the end of the second binding-domain's helical axis, then configurations in which the second binding-domain's helical axis-vector is orthogonal to the first will map to model configurations in which the first binding-domain's next-binding-domain vector is parallel to the second binding-domain's helical-axis vector.

The bonding term would mostly remain the same, with just the terms relating to the additional loss of degrees of freedom by restricting the helical axis vector in addition to the orientation vector in bound states. The stacking term would be substantially simplified, as the more complicated terms involving three and four bound domain pairs would be unnecessary. A stacked interaction between two bound domain pairs would be defined as one in which the orientation vectors have the dihedral angle prescribed by the binding domain length, as before, but now additionally one in which the helical-axis vectors are parallel. The same arguments can be made for determining sterically allowed and prohibited configurations for two bound domain pairs with a breakpoint. Like with the stacking term, the steric term can be made fully pairwise. In particular, the ambiguity that existed for crossover configurations involving quarter- and three-quarter-turn binding domains no longer exists, as crossover configurations would now map to a unique model configuration, making the term that required four bound-domain pairs unnecessary. The same constraints apply to two pairs of bound domains which involve two strands with contiguous binding domains.

3
Simulation methods

3.1 MCMC simulations for molecular systems

To explore the configuration space defined by the model described in Chapter 2, MC simulations can be used. The fundamental idea behind MC simulations is to use a random sampling from a given space with a known probability distribution function in order to estimate quantities of interest, typically expectation values of functions of that space. In the context of statistical mechanics, these expectation values are often thermodynamic or kinetic quantities that can be defined as averages over the configuration space of a given system as a function of the control variables of the selected statistical ensemble. In the simplest approach to MC simulation, values of the observable of interest $g(\vec{x})$ are calculated with points drawn uniformly from configuration space and multiplied by the values of the relevant distribution function $p(\vec{x})$ to estimate the expectation value of the observable,

$$\langle g(\vec{x}) \rangle = \int \mathrm{d}\vec{x}\, g(\vec{x}) p(\vec{x}) \approx \frac{1}{N} \sum_{i}^{N} g(\vec{x}_i) p(\vec{x}_i), \tag{3.1}$$

where N is the number of samples. In practice, however, this sampling is often highly inefficient as for most realistic systems the majority of the selected configurations will have very low associated probabilities [201]. Further, for any system in which MC simulations are needed to estimate observables, the distribution function will only be known up to a multiplicative constant. The calculation of the constant, known in the context of statistical mechanical distribution functions as the partition function, would itself require an integration over the entirety of configuration space.

Fortunately, there are a vast number of approaches to achieve better sampling in many different contexts [202]. For the purposes of molecular simulation, Markov chain Monte Carlo (MCMC) [203, 204] provides a route to focus sampling on more significant regions of configuration space; it has also been referred to as importance sampling [201, 205] (although this differs from the approach referred to as impor-

tance sampling in the statistics literature [202]). The main idea is to use a Markov chain that is more likely to stay in and move towards subsets of configuration space that have relatively higher weights. It can be shown that if the condition of detailed balance is obeyed,

$$p(\vec{x} \mid \vec{y})p(\vec{y}) = p(\vec{y} \mid \vec{x})p(\vec{x}), \tag{3.2}$$

where $p(\vec{x} \mid \vec{y})$ is the conditional probability of selecting configuration $\vec{x}$ given the current configuration $\vec{y}$, then the sampled states will still be distributed according to $p(\vec{x})$.

The conditional probabilities can be decomposed into two separate probabilities: $p_{\text{trial}}(\vec{x} \mid \vec{y})$, the probability of generating $\vec{x}$ given $\vec{y}$, and $p_{\text{acc}}(\vec{x} \mid \vec{y})$, the probability of accepting a generated $\vec{x}$ given $\vec{y}$, such that $p(\vec{x} \mid \vec{y}) = p_{\text{trial}}(\vec{x} \mid \vec{y})p_{\text{acc}}(\vec{x} \mid \vec{y})$. In the classic Metropolis algorithm [203], $p_{\text{trial}}(\vec{x} \mid \vec{y})$ is set to be symmetric (i.e. $p_{\text{trial}}(\vec{x} \mid \vec{y}) = p_{\text{trial}}(\vec{y} \mid \vec{x})$), and the acceptance probability written as

$$p_{\text{acc}}(\vec{x} \mid \vec{y}) = \min\left[1,\ \frac{p(\vec{x})}{p(\vec{y})}\right]. \tag{3.3}$$

The more general case in which $p_{\text{trial}}(\vec{x} \mid \vec{y})$ is not symmetric is referred to as the Metropolis–Hastings algorithm [204]. The second issue, of not *a priori* knowing the partition function, is also solved, as the acceptance probability is now dealing with relative rather than absolute probabilities. In the canonical ensemble, where $p(\vec{x}) \propto \mathrm{e}^{-\beta U(\vec{x})}$, with $U(\vec{x})$ being the potential energy of configuration $\vec{x}$, the acceptance probability becomes

$$p_{\text{acc}}(\vec{x} \mid \vec{y}) = \min\left[1,\ \mathrm{e}^{-\beta \Delta U(\vec{x},\vec{y})}\right]. \tag{3.4}$$

To run a MCMC simulation, definitions of $p_{\text{trial}}(\vec{x} \mid \vec{y})$, usually referred to as move types, are needed. An example of a simple move type on a lattice would be to select a particle in the system and propose a new configuration in which it occupies a neighbouring lattice site, with each selection carried out according to a uniform probability distribution. Clearly the reverse move is equally probable, and so $p_{\text{trial}}(\vec{x} \mid \vec{y})$ is symmetric, and thus detailed balance is obeyed. Of course, with molecular systems that have internal degrees of freedom, more complex move types are needed, and in a given simulation, several different move types may be used to increase sampling efficiency.

3.2 MC methods for lattice polymers

Polymers provide unique challenges to achieving efficient sampling. Many lattice polymer models can be seen as extensions of a simple self-avoiding walk (SAW), for which many MC sampling methods have been developed [206–208].

Sampling methods developed for lattice polymers can be broadly classified as being static or dynamic [201]. Static methods generate new polymer configurations from scratch at each step, thus producing a series of uncorrelated samples. Dynamic methods are those in which the current configuration is used to generate the next configuration, thus producing configurations that are correlated but not necessarily in a way that mimics the real dynamics of the system. Typically these dynamic sampling methods are in fact MCMC methods with varying $p_{\text{trial}}(\vec{x} \mid \vec{y})$ definitions, or move types.

The static method of simple regrowth of the entire chain is one of simplest methods of sampling polymer configurations, and is easily applied to the DNA origami model developed here. Clearly, however, a large number of the regrowths proposed will include overlaps with other units of the chain. In the case of our origami lattice model, these overlaps will typically result in unfavourable misbinding or steric clashes when a binding domain overlaps with two binding domains bound to each other.

To deal with the attrition problem of sampling longer polymers, methods were developed that allowed the growth of the chain to be biased such that it avoids overlaps and tends towards more energetically favourable configurations in models that include additional interactions. The first of these was the Rosenbluth sampling method [209], which biases the growth of each polymer unit towards more favourable configurations, and records the biases, referred to as Rosenbluth weights, for later reweighting of the resulting configurations. This method was improved upon in the pruned and enriched Rosenbluth sampling (PERM) method [210] and its extensions [211], which not only reduces the attrition problem, but additionally focuses sampling on those configurations that have higher Rosenbluth weights (i.e. those that will contribute significantly to the estimation of expected values of observables). This is achieved by considering multiple configurations simultaneously, and enriching those with high Rosenbluth weights while pruning those with low Rosenbluth weights.

It is often advantageous to take advantage of the benefits of MCMC when sampling polymer configurations. Broadly, MCMC methods for polymers can be split

into those that employ local move types, which can approximate the system dynamics, and those that employ non-local moves. The classic local move types are often referred to as the generalized Verdier–Stockmayer move set [212, 213], which allows the rearrangement of kinks by flipping one unit of the polymer (the end flip and corner flip move types) or two at once (the crankshaft flip move type). An alternative approach that can still approximate polymer dynamics is the use of reptation move types [214], which in the simplest "slithering snake" form involves the removal and addition of a unit from one end to the other. However, these methods are generally non-ergodic [215], and are typically not as efficient as approaches that are not constrained to approximate polymer dynamics.

Perhaps the simplest type of non-local move is the pivot [216], in which a unit is selected to act as a rotation point for the rest of the system. While this can significantly speed up sampling relative to local moves, in denser systems or those with strong interactions, the method becomes ineffective as it will tend to propose configurations with large unfavourable changes in energy. The pull move type, a non-local move type that is conceptually similar to reptation, can still provide some approximation of the kinetics while also being effective in contexts where the pivot move type is not [217, 218]. For dense systems, whether in a polymer melt or within a single compact polymer, a more efficient route can be to break and reform bonds [219]. A combination of the pull move type and this bond-rebridging move type was found to be highly effective for the hydrophobic-polar (HP) protein folding lattice model [220, 221]. Some rather inventive methods have been developed specifically for polymer melts, such as the wormhole move type [222] and the use of an additional spatial dimension [223].

For sampling individual polymers in compact configurations, however, those methods that extend the above discussed Rosenbluth sampling to an MCMC framework are arguably the most effective. One such approach is configurational bias (CB) [201, 224]. Like Rosenbluth sampling, the selection of the growth of each polymeric unit is biased, but rather than reweighting the samples at the end of the simulation upon calculating ensemble averages, the Metropolis–Hastings algorithm (see Section 3.3) is used to allow direct sampling of the desired probability distribution. Because it is not necessary to update the whole system in MCMC moves, this method can be used to partially regrow a chain, and the bias applied during growth may be modified to suit a given sampling problem.

Still, in very compact configurations, this method can fail by reaching dead-ends, where there are no acceptable configurations available for the next unit. A variation

was developed to combat this problem, referred to as recoil growth (RG) [225, 226], in which the growth of the chain can "recoil" to the previous unit if it hits a dead-end. Similarly to CB, the biasing function used is able to be modified to suit the challenges of a given system. A variant of PERM that allows it to be used as a move type in MCMC simulations was developed [227], but it was found to be no better CB.

3.3 Move types for DNA origami lattice models

3.3.1 General considerations

The details of the behaviour of staple strands in the assembly process are of interest solely when they are (mis)bound to a scaffold strand. Only the availability of the staples for binding to scaffold strands is relevant; this availability is determined by the initial staple concentrations, the binding of staples to scaffolds, and the binding of staples to other staples. Because in a typical assembly protocol, the staples are present in excess of the required stoichiometry, it can be assumed to a first approximation that the free staple concentrations are constant over the course of the assembly process. Further, at the temperatures relevant to assembly, staple–staple binding should not be a significant factor because the staples are not designed to bind to each other. The sampling of states with free staples can be avoided by running the simulations in the grand ensemble, in which we fix the chemical potential of the staples rather than their number. While staple–staple binding is not favourable overall, because of the local increase in concentration of staples at the scaffold, we do allow staple–staple binding to occur. Thus in states with no free staples, the staples can either be (mis)bound directly to the scaffold or (mis)bound indirectly via binding to a staple already (mis)bound to the system.

The strong and specific interactions of the model and the need to sample states with different numbers of (mis)bound staples makes efficient sampling challenging. Because the model presented in Chapter 2 can share a lattice site with one other binding domain of the same chain, it cannot be classed as an interacting SAW, although it does become self avoiding in partially assembled states. It can also be useful to consider the nature of the system for a given set of (mis)bound domain pairs, at which point it becomes an interacting walk with branches and loops.

While it would be desirable to reproduce dynamics, our top priority is efficiency, which has guided our choice of sampling algorithms. Thus we immediately dismissed the classic general Verdier–Stockmayer and reptation move types. Additionally, static move types in general seem unlikely to be successful. Consider a scaffold

binding domain grown early in the algorithm that needs to be linked by a staple to a scaffold binding domain grown much later; such a configuration becomes extremely unlikely to be proposed, even in the often highly effective PERM method. For such situations, the ability to do partial chain updates with an MCMC approach would be necessary.

In the bond-rebridging move type with a heteropolymer model, it is necessary to relabel the polymer units in order to achieve the same sequence after breaking and reforming bonds. While the move type was found to be effective in protein folding simulations of the HP lattice model, the HP model has only two polymer unit types, and interactions are between adjacent lattice sites [228]. The highly specific interactions present in our model make it highly likely that large unfavourable changes in the energy will occur upon relabeling the binding domains. Pull moves, which were also found to be effective with the HP model would likely also fail to sample effectively in near assembled states, as they would tend to break many bound pairs, again leading to large unfavourable energy changes. The system is composed of multiple polymers that are quite dense in near-assembled states, but because they bind by sharing a lattice site, and because the staple strands are short, it is not comparable to polymer melts. Thus we did not consider specialized move types for polymer melts.

Based on the above considerations, we decided to focus on methods that involve sequential growth of a set of binding domains in an MCMC framework that allow us to apply custom biases to the trial generation probability of each binding domain. The general outline of a chain regrowth move is as follows.

1. Select an available move type according to a predetermined distribution.
2. Select a set of binding domains and the order in which to regrow them.
3. Unassign the selected set of binding domains.
4. For the binding domain to be regrown, select a new position and orientation.
5. Repeat the previous step until all binding domains selected have been regrown.
6. Calculate the acceptance probability according to the selected biasing scheme.
7. Accept or reject the move with the probability calculated in the previous step.
8. If accepted, return to the step 1.
9. If rejected, revert to the previous configuration, then return to step 1.

For step 2, the order of regrowth will always be selected such that either an adjacent binding domain on that chain is set, restricting the positions to be chosen from in step 4 to the six neighbour sites, or the binding domain is bound to a domain that is set, determining its position. Unassigning a binding domain is essentially setting the position and orientation vectors to be undefined, and setting the state of the lattice

site to unoccupied. The selected position on a neighbour site can be decomposed into the sum of the position vector of the lattice site being grown from and a unit vector, which we refer to as the position difference vector. The way in which binding domain are selected for regrowth in step 2, and the bias used for selecting position difference vectors and orientation vectors is what differentiates move types. In the simplest case, the set of binding domains to be grown forms a contiguous segment of a single strand, and the binding domains are grown according to their order in the strand. More complex move types may involve multiple segments from the same strand, or segments from multiple strands, where growth of binding domains from a given segment in the stack may be interrupted by growth of binding domains from other segments.

3.3.2 Biased chain regrowth methods

We develop move types in which chain regrowth is primarily done with either CB or RG. We contrast these biased variants with an unbiased variant, which we refer to as symmetric regrowth, which for convenience we use in Section 3.6. While they are described as applied to the current DNA origami model, these move types are applicable in general to polymer lattice models. For symmetric regrowth, the position difference vectors and orientation vectors are chosen with uniform probability from the set of all unit vectors. As generation of configurations is symmetric, the trial generation probabilities of binding-domain growth for the forward and reverse moves will cancel. Thus the move is accepted according to the classic canonical Metropolis acceptance criterion (Equation (3.4)).

In the CB variants [224], the selection of a new configuration for each binding domain is biased by the associated energy change, such that the trial generation probability of each binding domain is

$$p_i^{\text{trial}} = \frac{\mathrm{e}^{-\beta\epsilon_{i,j}}}{\sum_{j'}^{k} \mathrm{e}^{-\beta\epsilon_{i,j'}}}, \tag{3.5}$$

where the sum is over the number of possible configurations k, which here is the number of neighbouring lattice sites times the number of possible orientation vectors (thus $k = 36$), and $\epsilon_{i,j}$ is the energy of setting binding domain i to have configuration j after having grown out all previous binding domains. As the trial generation probability is no longer symmetric, the acceptance probability will have additional

terms that account for this. Rearranging and grouping these terms gives

$$p_{\mathrm{acc}}(\vec{x} \mid \vec{y}) = \min\left[1,\ \frac{W_{\mathrm{new}}}{W_{\mathrm{old}}}\right], \tag{3.6}$$

where the Rosenbluth weight W is defined in terms of the Rosenbluth weights of each of the n binding domains grown, w_i, as

$$W = \prod_{i}^{n} w_i = \prod_{i}^{n}\left(\sum_{j=1}^{k} \mathrm{e}^{-\beta\epsilon_{i,j}}\right). \tag{3.7}$$

W_{old} is calculated by growing the old configuration and calculating w_i at each step.

In the RG variants [225, 226], if growth becomes stuck, the binding domains that were previously set can be unassigned, allowing them to be regrown in a different configuration. The growth of each binding domain involves selecting a configuration with uniform probability and choosing whether to consider the configuration 'open' or not according to some probability distribution. If it is chosen to be open, the configuration is selected for use and growth of the next binding domain proceeds. The probability of a configuration being open can be defined as needed; one possibility is

$$p_{i,j}^{\mathrm{open}} = \min\left[1,\ \mathrm{e}^{-\beta\epsilon_{i,j}}\right]. \tag{3.8}$$

If the configuration is chosen to be closed, another is proposed, up to a total of k_{max} configurations. This number of configurations to test, k_{max}, is a parameter that may be freely adjusted to improve efficiency. If no open configurations result, growth recoils to the previously set binding domain and testing continues for choosing open configurations where it left off. Recoiling can occur l_{max} times, or until all binding domains being grown have been unassigned, which if reached will result in the move being rejected.

To calculate the acceptance probability, the number of available configurations $m_{i,j}$ at each binding-domain i in the selected configuration j in both the new and old configuration must be determined. A binding-domain configuration is considered available if there is at least one open configuration for the next l_{max} binding domains to be grown, or for all the remaining binding domains to be grown if this is less than l_{max}. For each binding domain in the grown segment, one available configuration is already known; checking for available configurations continues until a total of k_{max}

configurations have been tested. The RG weights are defined as

$$W = \prod_{i}^{n} w_i = \prod_{i}^{n} \left(\frac{m_{i,j}}{p_{i,j}^{\text{open}}} \right). \tag{3.9}$$

Then, the move is accepted with

$$p_{\text{acc}}(\vec{x} \mid \vec{y}) = \min\left[1,\ \mathrm{e}^{-\beta \Delta U(\vec{x},\vec{y})} \frac{W_{\text{new}}}{W_{\text{old}}}\right]. \tag{3.10}$$

A super-detailed balance argument is used by Consta et al. [225] to show that detailed balance is obeyed with this acceptance probability.

Below, we describe the four classes of move types we developed for sampling our lattice model of DNA origami, which we also outline in Figure 3.1. In Section 3.10, we provide the details of our numerical validation, and in Section 3.11, we discuss the optimization of the move sets.

3.4 Orientation vector rotation moves

The first step in an orientation vector move consists of selecting a binding domain in the system and generating a new orientation vector, both with uniform probability. If the binding domain is in an unbound state, the change in energy upon a change in the orientation vector is zero, so the acceptance probability will be unity. If the binding domain is in a (mis)bound state, the orientation vector of the partner binding domain will also be modified in the trial configuration to be the additive inverse of the proposed orientation vector of the selected binding domain. This is then accepted according to Equation (3.4).

An example orientation rotation move is shown in Figure 3.1. In this move, scaffold-binding-domain 2 is selected. A new orientation vector direction is proposed, in which it now points up. Because it is in a bound state, the change in energy is zero, so the proposed configuration is accepted.

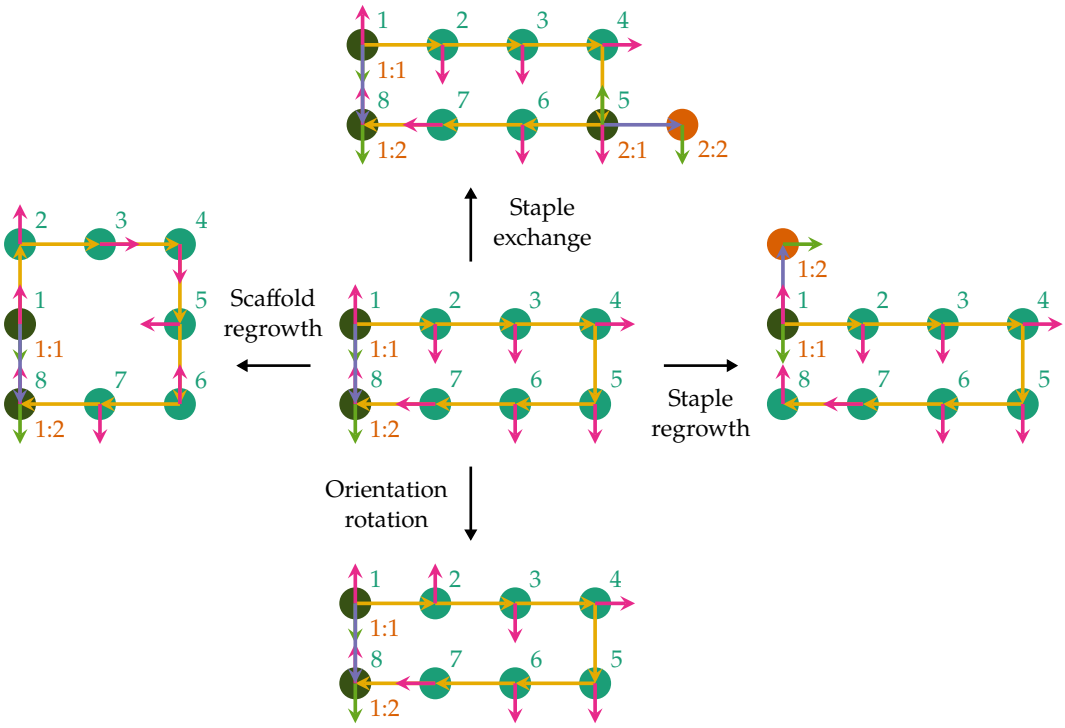

Figure 3.1: Outline of the four classes of move types developed in this thesis. An example move type of each class is applied to the configuration in the center to produce the configurations on the periphery. Scaffold indices are placed on the upper right of the binding domain while staple indices are placed on the bottom right of the binding domain, with an index for the staple itself on the left of the colon, and an index for the binding domain on the right. For simplicity all configurations only use two dimensions. A full legend for all diagram elements is provided in Figure 2.1.

3.5 Staple regrowth moves

Staple regrowth consists of first either selecting a staple in the system with uniform probability or rejecting the move if no staples are present. If this staple is a connecting staple, that is, a staple that if removed would leave a network of staples that has no connection to the scaffold, the move is rejected immediately. Otherwise, one of the binding domains on the selected staple that is in a (mis)bound state is selected to act as a point from which the remainder of the staple will be grown out from. Then the staple is grown out in both directions with the CB method, although in principle any of the growth schemes discussed in Section 3.3 may be used.

This scheme introduces an asymmetry into the generation of trial configurations. The probability of generating a trial configuration involves a factor of $1/b$, where b is the number of binding domains on the regrown staple (mis)bound to other

chains. This comes from the selection of a binding domain to grow out from. As b can change between the current and trial configuration, there is an additional factor of $b_{\text{old}}/b_{\text{new}}$ in the acceptance probability. In the case of CB, this gives

$$p_{\text{acc}}(\vec{x} \mid \vec{y}) = \min\left[1,\ \frac{W_{\text{new}}}{W_{\text{old}}} \times \frac{b_{\text{old}}}{b_{\text{new}}}\right]. \tag{3.11}$$

An example staple regrowth move is shown in Figure 3.1. As there is only one staple present, staple 1 is selected for regrowth. Removing this staple would clearly not leave any staples unbound to the system, so the move goes ahead. As both staple binding domains are in bound states, a random selection must be made to determine which will be used as the starting point for regrowth. Binding-domain 1:1 is selected, and so binding-domain 2:1 is grown out with CB. The new position is unbound, so for calculating the acceptance probability we would use $b_{\text{old}} = 2$ and $b_{\text{new}} = 1$.

3.6 Staple exchange moves

Staple exchange starts with a uniform random selection of either a staple insertion or staple deletion move. Then, in either case, a staple type is selected with uniform probability. While CB and RG variants could be used, we only use the symmetric scheme for binding-domain growth in the insertion move type. Clearly though, the trial configuration generation probabilities of the forward and reverse moves will not cancel, as there are many ways to insert and grow a given staple, but just one way to remove it.

In the case of an insertion move, a lattice site in the system volume, V_{sys}, which is defined as all the lattice sites occupied by at least one binding domain in the system, is selected with uniform probability to insert the first binding domain of the staple into, leading to a factor of $1/V_{\text{sys}}$ in the trial configuration generation probability. A binding domain on the staple being inserted is then selected with uniform probability to grow from, leading to an additional factor of $1/n_i$ in the trial probability, where n_i is the length of staple type i. The staple is then grown out from this binding domain, which gives a further factor of $6^{-2(n_i-1)}$ to the trial probability. However, states that involve binding of multiple binding domains to other chains will be over-counted with the current scheme, as there are b ways to grow these configurations. This can be corrected by multiplying the trial probability by a factor of b. Altogether, the trial

probability of insertion for staple type i is

$$p_{\text{trial}}(\vec{x}, N_i + 1 \mid \vec{y}, N_i) = \frac{b}{6^{2(n_i-1)} n_i V_{\text{sys}}}, \tag{3.12}$$

where N_i is the number of staples of type i.

For a deletion move, a staple of the selected type in the system is selected with uniform probability and removed if it is not a connector (see Section 3.5), which gives a trial probability of

$$p_{\text{trial}}(\vec{x}, N_i - 1 \mid \vec{y}, N_i) = \frac{1}{N_i}. \tag{3.13}$$

Because the number of staples is changing, the probability of being in a particular state is given by the grand ensemble probability distribution. Using the above trial probabilities, the acceptance probability for insertion of staple type i is

$$p_{\text{acc}}(\vec{x}, N_i + 1 \mid \vec{y}, N_i) = \min\left[1, \frac{6^{2(n_i-1)} n_i V_{\text{sys}}}{b(N_i + 1)} \mathrm{e}^{\beta\mu_i} \mathrm{e}^{-\beta\Delta U(\vec{x},\vec{y})}\right], \tag{3.14}$$

while for deletion it is

$$p_{\text{acc}}(\vec{x}, N_i - 1 \mid \vec{y}, N_i) = \min\left[1, \frac{bN_i}{6^{2(n_i-1)} n_i V_{\text{sys}}} \mathrm{e}^{-\beta\mu_i} \mathrm{e}^{-\beta\Delta U(\vec{x},\vec{y})}\right], \tag{3.15}$$

where μ_i is the chemical potential of staple type i. If we substitute in Equation (2.22) for the chemical potential, then for insertion,

$$p_{\text{acc}}(\vec{x}, N_i + 1 \mid \vec{y}, N_i) = \min\left[1, \frac{n_i V_{\text{sys}}}{6a^3 C_i b(N_i + 1)} \mathrm{e}^{-\beta\Delta U(\vec{x},\vec{y})}\right], \tag{3.16}$$

while for deletion,

$$p_{\text{acc}}(\vec{x}, N_i - 1 \mid \vec{y}, N_i) = \min\left[1, \frac{6a^3 C_i b N_i}{n_i V_{\text{sys}}} \mathrm{e}^{-\beta\Delta U(\vec{x},\vec{y})}\right]. \tag{3.17}$$

It seems that the lattice constant must be determined. However, if we note that $\Delta U(\vec{x}, \vec{y})$ will have exactly one ϵ_{b}, and so by substituting in Equation (2.15), the lattice constant will cancel in both acceptance probabilities.

An example staple regrowth move is shown in Figure 3.1. Here, a staple insertion move is selected. We have not defined the staple types for this system, so here we will simply say that a staple type is selected that binds to scaffold-binding-

domain 5 with staple-binding-domain 1 and to scaffold-binding-domain 6 with staple-binding-domain 2. An insertion site must be randomly selected; here scaffold-binding-domain 5 is selected. The binding domain to insert of the selected staple type must be chosen; here staple-binding-domain 1:1 is chosen. This happens to lead to a configuration in which staple and scaffold binding domains are complementary, but because the insertion site and insertion staple binding domain are chosen with uniform probability, it is much more common to propose a move in which they are not complementary. The staple has just one binding domain to be grown out, which is done with symmetric growth. Symmetric growth means that even though staple-binding-domain 1:2 is complementary to scaffold-binding-domain 4, and thus energetically favourable, trial moves are not biased towards this configuration; here an unbound configuration is proposed for the staple-binding-domain 2. To calculate the acceptance probability, we would use that for this move we have $n_i = 2$, $b = 1$, and $V_{\text{sys}} = 8$.

3.7 Scaffold regrowth moves

3.7.1 Growth bias

A seemingly straightforward way to sample scaffold conformational states would be to select a segment of the scaffold and regrow these binding domains and any (mis)bound staples. However, even with advanced polymer growth schemes like CB and RG, if the scaffold segment to be regrown is in a near assembled state, the proposed configurations will rarely have as many bound domains, and will thus be of a substantially less favourable energy. Hence such moves will thus almost always be rejected. To address this, we have chosen to keep the sampling of binding states and scaffold conformational states separate by developing variants of CB and RG that allow the binding state of the system to be left unchanged when regrowing parts of the system, leaving sampling of binding states to the staple exchange and regrowth moves. Such a separation also simplifies the calculation of the trial generation probabilities by removing the asymmetries involved with changing binding states. If the system is considered as a network where (mis)bound domain pairs act as nodes, these moves can be thought of as holding the network topology constant, and are thus referred to as conserved topology (CT) moves.

Fixed-end CB is a scheme that allows polymers to be grown to a predetermined endpoint [229]. This works by introducing a further bias into the selection of configurations for the growth of each polymer unit. The bias is the number of ideal

random walks from the trial polymer unit's position to the endpoint position, given the number of polymer units remaining to be grown. Importantly, if a configuration for the polymer unit currently being grown has no ideal random walks available to reach the endpoint, the configuration will have zero probability of being proposed.

Here, we use a similar idea but extend it to allow multiple endpoints per segment, and growth of multiple segments on possibly multiple chains. When a move involves growing multiple segments, each can have its own set of endpoint constraints. Once a particular endpoint is reached, the associated endpoint constraint has been satisfied, and so becomes inactive. If a binding domain that is to serve as an endpoint must also be grown, the endpoint constraint is inactive until the associated binding-domain's configuration has been set.

Because of these cases of endpoint positions being set during growth, the number of ideal random walks can no longer be directly used in the bias. This is because the initial number of ideal random walks for such endpoints could differ between the old and new configurations, and would thus not cancel when taking the ratio of the Rosenbluth weights for the old and new configuration, as it does in the original method. Instead, we use an indicator function, $\chi_{\mathrm{I}}\left(\Delta\vec{r}_{l,j}, n_{l,i}\right)$, that is unity if walks remain and zero otherwise, where $\Delta\vec{r}_{l,j}$ is the difference vector between the trial position of configuration j and the position of endpoint l, and $n_{l,i}$ is the number of binding domains remaining to be grown between binding domain i and the binding domain of endpoint l. Whether walks remain or not can be determined by checking if the sum of the absolute values of the components of the position difference vector is greater than or equal to the number of binding domains remaining to be grown.

While the endpoint constraints ensure that the system will still have the (mis)-bound pairs it began with (with the exception of same-chain misbinding; see following discussion), they do not prevent new pairings from forming. To prevent new pairings, another indicator function, $\chi_{\mathrm{B}}(s)$, of lattice site s can be used. This function is unity if the lattice site is unoccupied, the position of an endpoint of an active endpoint constraint on the segment being grown, or occupied by another binding domain of the chain currently being grown, and zero otherwise. We allow misbinding between binding domains on the same chain because the staple exchange and regrowth moves will not allow sampling of states involving scaffold binding domains misbinding with themselves. Because we allow these misbound pairings to form, we must also allow them to unform, and so they are not used by endpoint constraints. Because misbinding interactions are relatively weak, decreasing the number of misbound pairs will not typically lead to large unfavourable energy changes. Fur-

ther, because they are by definition not as specific as fully bound pairs, there are many ways to propose moves that have the same number of misbound pairs. Finally, we note that changing the number of intra-chain misbound pairs will not introduce asymmetry into the trial probability, as there is no selection of a domain to grow out from involved.

The trial probability of selecting a configuration for binding domain i is now

$$p_i^{\text{trial}} = \frac{\mathrm{e}^{-\beta\epsilon_{i,j}}\chi_{\mathrm{B}}\left(s_j\right)\prod_l \chi_{\mathrm{I}}\left(\Delta\vec{r}_{l,j}, n_{l,i}\right)}{\sum_{j'}^{k}\mathrm{e}^{-\beta\epsilon_{i,j'}}\chi_{\mathrm{B}}\left(s_{j'}\right)\prod_{l'}\chi_{\mathrm{I}}\left(\Delta\vec{r}_{l',j'}, n_{l',i'}\right)}, \tag{3.18}$$

and the Rosenbluth weight is

$$w_i = \sum_{j=1}^{k}\mathrm{e}^{-\beta\epsilon_{i,j}}\chi_{\mathrm{B}}\left(s_j\right)\prod_l \chi_{\mathrm{I}}\left(\Delta\vec{r}_{l,j}, n_{l,i}\right). \tag{3.19}$$

A similar modification can be made to the RG scheme for growing binding domains to construct a CT variant. The modification is made to the probability of a configuration being open,

$$p_{i,j}^{\text{open}} = \min\left[1,\ \mathrm{e}^{-\beta\epsilon_{i,j}}\chi_{\mathrm{B}}\left(s_j\right)\prod_l \chi_{\mathrm{I}}\left(\Delta\vec{r}_{l,j}, n_{l,i}\right)\right]. \tag{3.20}$$

3.7.2 Segment selection

The CT move types, whether CTCB or CTRG, begin with the selection of a segment or segments of the scaffold to regrow. Then, of the set of staples that are involved in the network of staples (mis)bound to the selected scaffold segment(s), it must be determined which will be regrown and which will act as endpoints. Here, if a set of staples is involved in a network that includes scaffold binding domains external to the selected segment(s), the staples will remain in their current configuration, with those that are (mis)bound to the selected scaffold segment acting as endpoints for its regrowth. These are referred to as external staples. If a staple is not involved in such a network, it is regrown with the scaffold, with endpoints for the required endpoint constraints being determined during regrowth. These are referred to as internal staples.

If the scaffold segment was regrown fully before regrowing any of the internal staples, it would result in binding domains on the scaffold being used in endpoint constraints for the internal staples. However, this would be less effective than regrowing the staples first such that the endpoints were instead on the staples and

used by endpoint constraints applied to regrowth of the scaffold, as typically the staples are only two or three binding domains and so often have no way of reaching an endpoint on a scaffold binding domain. Thus, as the scaffold is regrown, if there is a fully unset staple to be (mis)bound to the binding domain that has just been set, regrowth of the current chain will be put on hold to regrow this staple. This may also happen while regrowing a staple, in a recursive manner. If a binding domain on the chain being regrown is to be (mis)bound to an unassigned binding domain on a chain already in the process of being regrown, an endpoint constraint is set up for this other chain (typically the scaffold).

There are many ways to select the set of scaffold segments for regrowth. We use two variants: single- and multiple-segment selection. In the single segment, or contiguous scaffold regrowth, variant, the segment is selected such that the distribution of lengths is uniform, where the range of possible lengths is a parameter of the move type. To create a segment, a uniformly random scaffold binding domain is selected from the set of all scaffold binding domains to act as the seed binding domain from which to create the segment. While the seed itself is not regrown, for a staple to be bound to it will count towards it being an internal staple, and such an internal staple would be the first thing to regrow. A direction with which to add binding domains to the segment is then selected with uniform probability. Binding domains are added until either the selected segment length or the end of the chain is reached. If the end of the chain is reached, segments will begin to be added from the other side of the seed binding domain.

An example scaffold regrowth move is shown in Figure 3.1, but as scaffold regrowth moves involve many steps, it has been expanded upon in Figure 3.2 to demonstrate a contiguous scaffold regrowth move. First, scaffold-binding-domain 1 is selected as the seed binding domain. The forward direction is selected for regrowth, although as it is the first binding domain of the chain, this is inconsequential for the above outlined regrowth algorithm. Then, the number of scaffold binding domains to be regrown is selected to be eight, which here is the entire chain. This means there will be no endpoint constraint on scaffold regrowth to ensure an unbroken chain. To determine which staples will be regrown is straightforward, as there is only staple in the system, which must be internal.

To begin regrowth, the binding domains to regrow are first unassigned, shown in Figure 3.2(b). Regrowth biasing can be done with either CB or RG. As it is bound to the seed binding domain, staple-binding-domain 1:1 becomes the first to be set in (c). In (d), staple-binding-domain 1:2 is then grown out, and will now act as

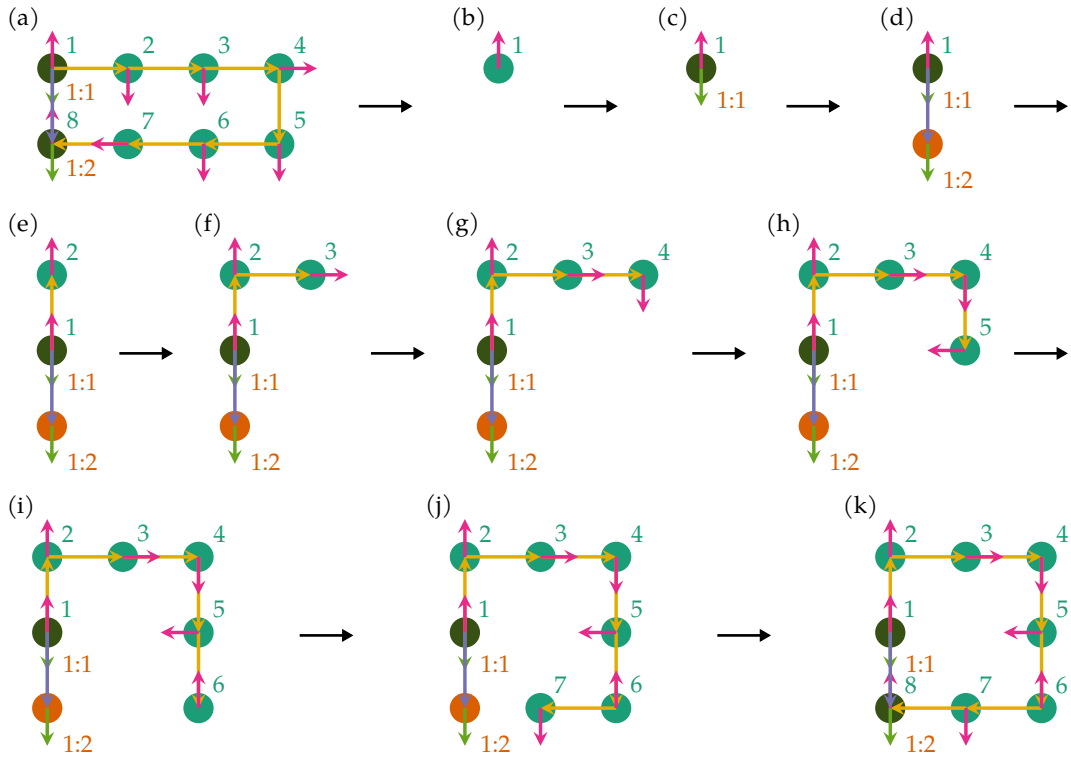

Figure 3.2: Example of a contiguous scaffold regrowth move. The same index labeling scheme is used as in Figure 3.1. A full legend for all diagram elements is provided in Figure 2.1.

an endpoint constraint for the scaffold chain, as this binding domain is bound to scaffold-binding-domain 8. The selection of the position of scaffold-binding-domain 2 and 3 in (e) and (f), respectively, is unaffected by this endpoint constraint, but for regrowing scaffold-binding-domain 4 in (g), some positions, for example the position above scaffold-binding-domain 3, are prohibited by the endpoint constraint. For scaffold-binding-domain 5, there are only two possibilities that can allow the endpoint constraint to be achieved: misbinding with scaffold-binding-domain 4, or as chosen in (h), directly below scaffold-binding-domain 4. There are also only two possibilities for scaffold-binding-domain 6, one of which is selected in (i). Finally, the positions for scaffold-binding-domains 7 and 8 are fully determined by the endpoint constraint, shown in (j) and (k), respectively.

In the multiple segment, or non-contiguous scaffold regrowth, variant, the intention is to allow the selection of binding domains for regrowth to be able to jump at points where two scaffold binding domains are adjacent due to a linking staple. For each move, a maximum total number of binding domains to regrow across all segments is chosen with uniform probability, where the range from which the selec-

tion is made is a parameter of the move type. For each individual segment that is created, a maximum segment length is selected with uniform probability, where the range from which the length is selected is another parameter of the move type. The addition of binding domains to a given segment proceeds until either the maximum segment length is reached, the maximum regrowth length is reached, the end of the chain is reached, or the binding domain following the binding domain being considered for addition to the segment is already part of another segment to regrow. This last condition is to simplify the construction of endpoint constraints.

Segment creation begins with the selection of a seed binding domain and direction from which to add binding domains to the segment in the same manner as with the contiguous scaffold regrowth variant. As with contiguous regrowth, a staple bound to a seed domain would be the first thing to be regrown. As binding domains are added to the segment (and all subsequent segments created), if a binding domain is bound to a staple that is bound to another binding domain of the scaffold that is neither already in a segment to regrow nor contiguous with a scaffold binding domain that is in a segment to regrow, it is added to a queue of potential segment seed binding domains. Once addition of binding domains to the segment has been terminated, if the maximum total number of binding domains to regrow has not been reached, a new segment is created with a binding domain from the front of the aforementioned queue. The direction from which to proceed is selected as with the first segment. Once addition of binding domains to this segment has been terminated, and if the maximum total number of binding domains to regrow has not been reached, a segment beginning from the binding domain in the opposite direction of the previous segment seed will be used as the seed for a new segment, if it exists and if the following binding domain is not already part of another segment to regrow. Once addition of binding domains to this segment has been terminated, the steps after initial segment creation are repeated until either the queue is empty or the maximum total number of binding domains to regrow is reached. The segments are regrown in the order in which they were created. Because the probability of selecting a particular set of segments of scaffold binding domains does not depend on the conformation or on whether or not binding domains are misbound to other binding domains on the same chain, the move type obeys detailed balance.

An example of a non-contiguous scaffold regrowth move is given in Figure 3.3. First, the maximum total number of binding domains to regrow is selected to be seven. The seed binding domain is selected to be scaffold-binding-domain 5, and for the first segment, the direction is chosen to be forward and the maximum length three.

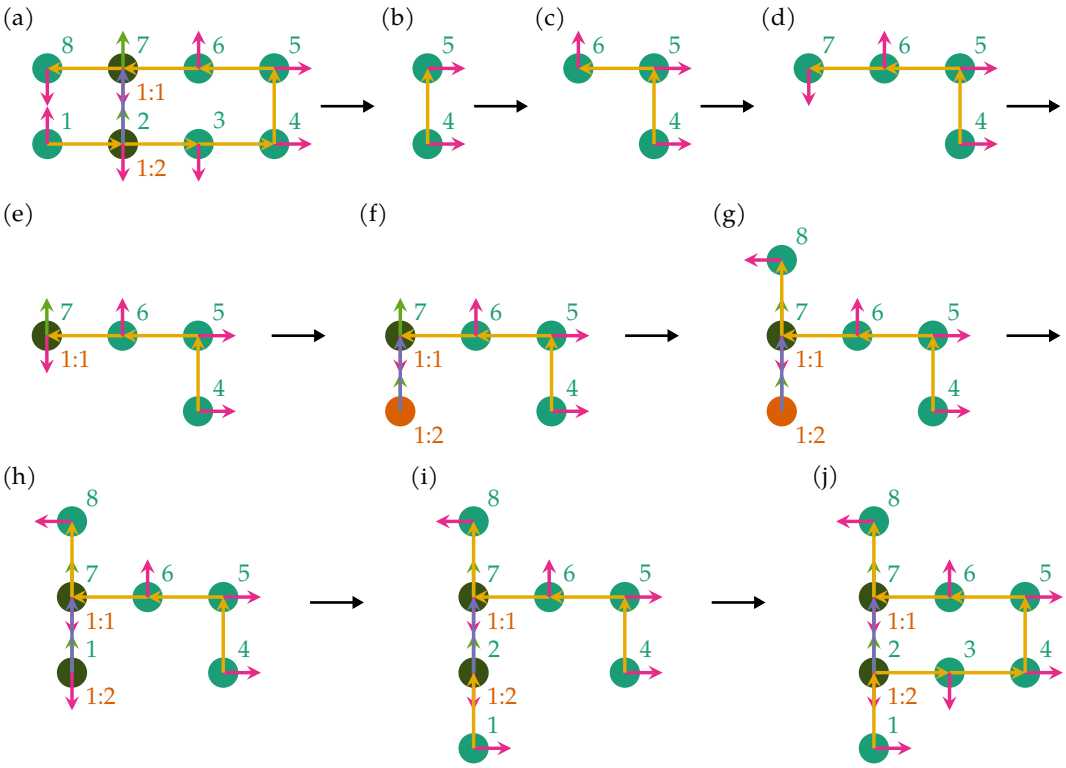

Figure 3.3: Example of a non-contiguous scaffold regrowth move. The same index labeling scheme is used as in Figure 3.1. A full legend for all diagram elements is provided in Figure 2.1.

While adding scaffold-binding-domains 6, 7, and 8 to the first segment, scaffold-binding-domain 2 is added to the queue of potential segment seed binding domains. With the maximum length of the first segment having been reached, the segment seed in the queue is used to begin a new segment. The negative direction, and a maximum segment length of two, is selected. Because the seed binding domain must be regrown, it is included in the count of the segment length, so only scaffold-binding-domain 1 is additionally added to this segment. The total number of binding domains to be regrown is still less than six, so another segment is added, and binding domains added by moving in the opposite direction as the previous segment out from that seed binding domain. A maximum segment size of three is selected, but since only scaffold-binding-domain 3 can be added, the segment length is only one. Finally, while the total number of binding domains to be regrown is only six, the queue of potential seed binding domains is empty, so selection of scaffold segments ends.

With all the segments selected, regrowth can begin. Again, either CB or RG can be used for regrowth. From Figure 3.3(b) to (d), the first two binding domains of

the first segment are grown. Then, the staple bound to scaffold binding domain 7 is grown out in (e) and (f), before returning to the first segment to grow the final binding domain in (g). In (h) and (i) the second segment is grown out from staple-binding-domain 1:2. So far, there have been no endpoint constraints, but for the growth of the single binding domain in the third segment, there is an endpoint constraint enforcing that it be adjacent to scaffold-binding-domain 4, which determines its position.

3.8 Replica exchange

In order to increase sampling efficiency, we use a method that has come to be known as parallel tempering, or replica exchange Monte Carlo (REMC) [230–234]. This advanced sampling method involves running multiple replicas of the simulation that differ with respect to a simulation control parameter, most commonly temperature, which we refer to as exchange variables. An additional move type is constructed in these simulations, which in the simplest form of REMC involves an exchange attempt at a random step interval of the configurations of a random pair of replicas, provided that the replicas are adjacent with respect to the exchange variables. REMC alone would not be ergodic, as the initial set of configurations would simply be swapped between the replicas, so it must be used in conjunction with an ergodic move set. The idea is that by selecting exchange variables that control the heights of barriers along the free-energy landscape, fluctuations that occur when the barriers are smaller can be passed to the replicas with larger barriers. In a more general form known as Hamiltonian REMC, the Hamiltonian itself can be exchanged [235–237]. It is also possible to carry out multi-dimensional REMC, in which multiple exchange variables are used [235, 236].

While we are primarily interested in running the simulations across a range of temperatures, here we must also use a range of Hamiltonians because of the temperature dependence of the NN model hybridization free energies. Further, because we are in the grand ensemble but would like to keep the staple concentration constant across the replicas, we must also exchange the staple chemical potential. Because they are both functions of the temperature, this is still one-dimensional REMC.

Considering each replica as its own simulation, a REMC swap move can be considered as two separate moves for the selected replica pair, i and j. For replica i the detailed balance condition is

$$p(\vec{y};\mu_i,T_i,\mathcal{H}_i \mid \vec{x};\mu_i,T_i,\mathcal{H}_i)p(\vec{x};\mu_i,T_i,\mathcal{H}_i) \\ = p(\vec{x};\mu_i,T_i,\mathcal{H}_i \mid \vec{y};\mu_i,T_i,\mathcal{H}_i)p(\vec{y};\mu_i,T_i,\mathcal{H}_i), \quad (3.21)$$

while for replica j it is

$$p\big(\vec{x};\mu_j,T_j,\mathcal{H}_j \mid \vec{y};\mu_j,T_j,\mathcal{H}_j\big)p\big(\vec{y};\mu_j,T_j,\mathcal{H}_j\big) \\ = p\big(\vec{y};\mu_j,T_j,\mathcal{H}_j \mid \vec{x};\mu_j,T_j,\mathcal{H}_j\big)p\big(\vec{x};\mu_j,T_j,\mathcal{H}_j\big). \quad (3.22)$$

For the swap to be accepted, both of these individual moves must occur, so the total transition probability is the product of the two individual transition probabilities,

$$p\big(\vec{y};\mu_i,T_i,\mathcal{H}_i \wedge \vec{x};\mu_j,T_j,\mathcal{H}_j \mid \vec{x};\mu_i,T_i,\mathcal{H}_i \wedge \vec{y};\mu_j,T_j,\mathcal{H}_j\big) \\ = p(\vec{y};\mu_i,T_i,\mathcal{H}_i \mid \vec{x};\mu_i,T_i,\mathcal{H}_i)p\big(\vec{x};\mu_j,T_j,\mathcal{H}_j \mid \vec{y};\mu_j,T_j,\mathcal{H}_j\big) \quad (3.23)$$

We can then rewrite the transition probability in terms of the acceptance and trial probabilities. The generation of a trial configuration for a replica is essentially taking a configuration from the equilibrium ensemble for its selected pair for that move, so

$$p_{\text{trial}}(\vec{y};\mu_i,T_i,\mathcal{H}_i \mid \vec{x};\mu_i,T_i,\mathcal{H}_i) = p\big(\vec{y};\mu_j,T_j,\mathcal{H}_j\big). \quad (3.24)$$

If we combine Equations (3.21) to (3.24), and rearrange to solve for the acceptance probability, then

$$\begin{aligned} &p_{\text{acc}}\big(\vec{y};\mu_i,T_i,\mathcal{H}_i \wedge \vec{x};\mu_j,T_j,\mathcal{H}_j \mid \vec{x};\mu_i,T_i,\mathcal{H}_i \wedge \vec{y};\mu_j,T_j,\mathcal{H}_j\big) \\ &= \min\left[1,\ \frac{p(\vec{y};\mu_i,T_i,\mathcal{H}_i)p\big(\vec{x};\mu_j,T_j,\mathcal{H}_j\big)}{p\big(\vec{y};\mu_j,T_j,\mathcal{H}_j\big)p(\vec{x};\mu_i,T_i,\mathcal{H}_i)}\right] \\ &= \min\left[1,\ \frac{\left(\mathrm{e}^{\beta_i\mu_i N_{\vec{y}}}\mathrm{e}^{-\beta_i\mathcal{H}_i(\vec{y})}\right)\left(\mathrm{e}^{\beta_j\mu_j N_{\vec{x}}}\mathrm{e}^{-\beta_j\mathcal{H}_j(\vec{x})}\right)}{\left(\mathrm{e}^{\beta_j\mu_j N_{\vec{y}}}\mathrm{e}^{-\beta_j\mathcal{H}_j(\vec{y})}\right)\left(\mathrm{e}^{\beta_i\mu_i N_{\vec{x}}}\mathrm{e}^{-\beta_i\mathcal{H}_i(\vec{x})}\right)}\right] \\ &= \min\left[1,\ \mathrm{e}^{\Delta_{\mathrm{r}}(\beta\Delta_{\mathrm{c}}\mathcal{H})-\Delta_{\mathrm{r}}(\beta\mu)\Delta_{\mathrm{c}}N}\right] \end{aligned} \quad (3.25)$$

where Δ_{r} is a difference operator between replicas i and j (e.g. $\Delta_{\mathrm{r}}(ab) = a_j b_j - a_i b_i$) and Δ_{c} is a difference operator between configurations $\vec{x}$ and $\vec{y}$ (e.g. $\Delta_{\mathrm{c}} a = a_{\vec{y}} - a_{\vec{x}}$).

From Equation (3.25), it can be seen that to calculate the acceptance probability for a given swap, it is necessary to calculate the energy of both configurations with both Hamiltonians. If we expand the Hamiltonian in the first term of the exponential in Equation (3.25) in terms of the enthalpy and entropy of the model, we can simplify

this calculation such that

$$\begin{aligned}\Delta_{\mathrm{r}}(\beta\Delta_{\mathrm{c}}\mathcal{H}) &= \frac{1}{T_j}\left(\Delta H_{\mathrm{total}}(\vec{y}) - T_j\Delta S_{\mathrm{hyb}}(\vec{y}) - \Delta H_{\mathrm{total}}(\vec{x}) + T_j\Delta S_{\mathrm{hyb}}(\vec{x})\right) \\ &\quad - \frac{1}{T_i}\left(\Delta H_{\mathrm{total}}(\vec{y}) - T_i\Delta S_{\mathrm{hyb}}(\vec{y}) - \Delta H_{\mathrm{total}}(\vec{x}) + T_i\Delta S_{\mathrm{hyb}}(\vec{x})\right) \\ &= \frac{\Delta H_{\mathrm{total}}(\vec{y}) - \Delta H_{\mathrm{total}}(\vec{x})}{T_j} - \frac{\Delta H_{\mathrm{total}}(\vec{y}) - \Delta H_{\mathrm{total}}(\vec{x})}{T_i} \\ &= \Delta_{\mathrm{c}}(\Delta H_{\mathrm{total}})\Delta_{\mathrm{r}}\beta, \end{aligned} \tag{3.26}$$

where $\Delta H_{\mathrm{total}}(\vec{x}) = \Delta H_{\mathrm{hyb}}(\vec{x}) + \Delta H_{\mathrm{stack}}(\vec{x})$, with $\Delta H_{\mathrm{hyb}}(\vec{x})$, $\Delta H_{\mathrm{stack}}(\vec{x})$, and $\Delta S_{\mathrm{hyb}}(\vec{x})$ being the hybridization enthalpy, stacking energy, and hybridization entropy, respectively, for the selected model and system in configuration $\vec{x}$. This allows us to use the values for these enthalpies and entropies that we update at each step without a full recalculation for different temperatures.

In addition to what is essentially REMC with temperature as the independent exchange variable, we also consider a true Hamiltonian REMC with a multiplier on the stacking energy acting as an exchange variable. This may allow the system to cross barriers between states with different stacked pairs by providing a route that avoids the binding and unbinding of domains and staples that using temperature as the exchange variable may lead to. In this case, the first term of the exponential in Equation (3.25) instead becomes

$$\Delta_{\mathrm{r}}(\beta\Delta_{\mathrm{c}}\mathcal{H}) = \Delta_{\mathrm{c}}\left(\Delta H_{\mathrm{hyb}}\right)\Delta_{\mathrm{r}}\beta + \Delta_{\mathrm{c}}E_{\mathrm{stack}}\Delta_{\mathrm{r}}(m\beta), \tag{3.27}$$

where m is the stacking energy multiplier. However if we are only exchanging the stacking multiplier, the right-hand-side of Equation (3.25) simplifies further to

$$\min\left[1,\ \mathrm{e}^{\Delta_{\mathrm{r}}(\beta\Delta_{\mathrm{c}}\mathcal{H})-\Delta_{\mathrm{r}}(\beta\mu)\Delta_{\mathrm{c}}N}\right] = \min\left[1,\ \mathrm{e}^{\beta\left(\Delta_{\mathrm{c}}\left(\Delta H_{\mathrm{hyb}}\right)+\Delta_{\mathrm{c}}E_{\mathrm{stack}}\Delta_{\mathrm{r}}m\right)}\right]. \tag{3.28}$$

To improve parallelization, the variant of REMC used here attempts an exchange at a set step interval, and alternates between attempting an exchange between all even pairs and all odd pairs of replicas, where pairs are numbered with the index of the first replica in the pair along the exchange variable. While this variant of REMC no longer obeys detailed balance, it can be shown to obey total balance [238], and has been found to be relatively efficient compared to other exchange schemes [239]. We also consider a 2D REMC scheme, in which case we alternate between both the exchange variables and the even/odd pairings within each exchange variable.

3.9 Free-energy calculations

Calculating free energies presents additional challenges to those of calculating expectations of system properties that can take on instantaneous values for a given configuration (i.e. mechanical properties). This is because they depend on the logarithm of the partition function, and so require a more thorough sampling of phase space, as to estimate the partition function the integrated total weighted volume of phase space must be estimated. Fortunately as only free-energy differences are meaningful, we are typically only calculating ratios of partition functions, which alleviates the issue somewhat.

Apart from free-energy differences of a given system in different thermodynamic states, we may also be interested in comparing free energies between two different systems, or partitioning configuration space for a given system in a given thermodynamic state and considering free energy differences between these partitions, which typically correspond to distinct phases or macrostates. To partition configuration space, it is often useful to define some coarse-grained variable(s), which we will refer to as order parameters, although in other contexts they are referred to as collective variables or reaction coordinates [240]. The free-energy differences along these order parameters can then be calculated, which we will refer to as Landau free energys (LFEs), but they have also been described as potentials of mean force.

If one considers an extended ensemble that includes all macrostates of interest (whether thermodynamic, chemical, or configurational), the extended partition function is simply the product of the individual partition functions, and the probability of being in a particular macrostate is the ratio of that macrostate's partition function to the extended partition function. Clearly, then, the free-energy difference between any two macrostates can be estimated by the relative probabilities of the two states, which may be done by keeping track of the number of times a simulation is in a given macrostate. Thus any simulation that provides samples from the extended ensemble that can be used to compare weights of various macrostates can be used to calculate free-energy differences.

However, typically specialized methods are required to achieve better convergence. A huge variety of methods have been developed for many different contexts, but in this thesis we are interested in those methods applicable to calculating LFEs of polymers in a dilute solution phase. The REMC method described above is one approach that helps sample configuration space more broadly if the exchange variables provide a good proxy for sampling across the relevant range of the order parameters

of interest. Combined with a method to reweight the replicas to the thermodynamic state of interest to harvest all available information (discussed below), this can be an effective method for calculating LFEs.

However, if for some values of the order parameters the relevant associated subset of configuration space has very low weights, and changing the control variables does not allow this subset to be sampled well, as for example is the case when studying nucleation barriers, it may be necessary to consider methods that sample from a distribution other than that of the ensemble of interest. Typically, such methods attempt to flatten the free energy surface by using a biasing weight or potential such that all relevant states can be sampled with equal weight. If the biasing potential is an additional term in the Hamiltonian, then the biasing potential for an order parameter q that will give a uniform distribution is

$$\Phi(q) = k_{\mathrm{B}} T \ln p(q), \tag{3.29}$$

where $p(q)$ is the probability distribution of the order parameter in the given ensemble. A disadvantage of such biasing methods is that $p(q)$ is not known, and must be estimated in an iterative manner. One approach to estimating and applying this biasing weight in the calculation of free energies is known as umbrella sampling (US) [241, 242]. We use a simplified version of a multi-windowed adaptive US scheme [243] (see below for definition of a window).

In this scheme, one or more order parameters are selected, and the relevant ranges are partitioned into bins, although here all order parameters are integer valued and encompass a relatively small range, making binning unnecessary; however, we will still refer to each combination of order parameter values as a bin. As the simulation proceeds, a histogram is built by counting the number of configurations that fall into each bin. After a set number of steps, the histogram is reweighted with the current bias weights to estimate $\Phi(q)$, which is then used as the bias weight in the next round. To improve convergence during the early stages in which some bins may have very few samples and thus lead to poor estimates of $\Phi(q)$, a maximum change in the bias weight is enforced.

To improve parallelization, rather than running a single simulation with the goal of achieving uniform sampling across the whole range of order parameters, multiple windows can be defined that cover only a subset of the range. These windows have a further unvarying bias that prevents the simulation from sampling states outside the window; here a simple step function is used where the bias is zero inside the window range, and is determined by a linear (i.e. affine) function outside window

that slopes towards the region of zero bias. In order to reconstruct a single LFE at the end of the simulation, the windows must overlap so that a single histogram may be constructed for the set of windows.

Ferrenberg and Swendsen [244] developed a method that allows the data from simulations from multiple states to be reweighted to the state of interest, which was further extended by Kumar et al. [245] into a widely used method known as the weighted histogram analysis method (WHAM). This method involves self-consistently solving a set of equations to provide the partition functions up to a multiplicative constant, allowing for free-energy differences between the states to be calculated. This method does not provide an estimate of the variance, so another method such as bootstrapping must be used. The multi-Bennett acceptance ratio (MBAR) method [246] was also developed to allow data from multiple simulations to be combined, although in contrast to WHAM, it does not require the data to be binned to form histograms and provides an estimate of the variance. Both WHAM and MBAR may be used to combine and reweight the results of multi-window US simulations, or to reweight REMC simulations to take advantage of data at states other than those of interest. We chose to use the MBAR method for both US and REMC simulation analysis.

The MBAR method requires a series of independent samples as input. We use the method of Chodera et al. [247] to estimate the statistical inefficiency, which allows an uncorrelated subset of the samples generated by the simulations to be extracted. Convergence of expectation values can be achieved with less data by discarding samples from the start of the simulation, when there are often highly atypical configurations. We use an automated method for determining the equilibration, or burn-in, steps [248]. We used a freely available software package, pymbar, for the MBAR, statistical inefficiency, and automated equilibration detection time calculations. For the calculation of burn-in steps and statistical inefficiency, we use a form of the reduced potential as the input series [246],

$$u_i(\vec{x}) = \beta_i \left(U_i(\vec{x}) + \mu_i N(\vec{x}) \right), \tag{3.30}$$

where the indices refer to the particular state being considered. The reduced potential is a suitable choice because it is a relatively general measure of relevant fluctuations in the system, and because it is also a required input of the MBAR method.

3.10 Numerical validation of MCMC move types

To examine the validity of the move types, simulations of a simple system were run and compared to exact results. The system used consists of a four-binding-domain scaffold and two staples, in which one of the staples links the terminal scaffold binding domains in the assembled state, as can be seen in Figure 3.4(b). The exact results were calculated by taking the ensemble averages across all configurations that have at most four total staples or two staples of a given type, which were determined with a recursive enumeration algorithm. The move set was nearly the same as the temperature REMC move set described in the following section, with the exception that the maximum number of scaffold domains to regrow is set to four. The MBAR method was used to calculate the expectation values for selected order parameters. The average number of bound-domain pairs, the average number of (mis)bound staples, the average number of misbound-domain pairs, and the average number of stacked-binding-domain pairs for both the simulation and enumeration results are plotted in Figure 3.4(a), which clearly shows that the two approaches agree within sampling error.

3.11 Optimization of MCMC parameters

A given move set has many free parameters and cannot be fully optimized without extensive efforts. For the move sets used in this thesis, which have an orientation rotation move type, a staple exchange move type, a CB staple regrowth move type, a contiguous CTRG scaffold regrowth move type, a non-contiguous CTRG scaffold regrowth move type, and a REMC exchange move type, there are 12 adjustable parameters. Further, the optimal parameters will be different depending on the system being simulated, as well as the simulation conditions and model parameters. We instead undertake only a small amount of optimization to avoid wasting effort on the diminishing returns associated with more thorough optimization. The strategy was to optimize parameters in isolation by assuming that the dependency in optimal value of a given parameter on the others is small.

The individual parameters were optimized primarily by running simulations on the 24-binding-domain scaffold system (see Figure 4.2), although some additional optimization was performed on the 56-binding-domain scaffold system (see Figure 4.11). To optimize the parameters, we considered the acceptance frequencies (and in the case of scaffold regrowth moves, acceptance frequencies as a function of the number of scaffold domains being regrown), the mean times to the first as-

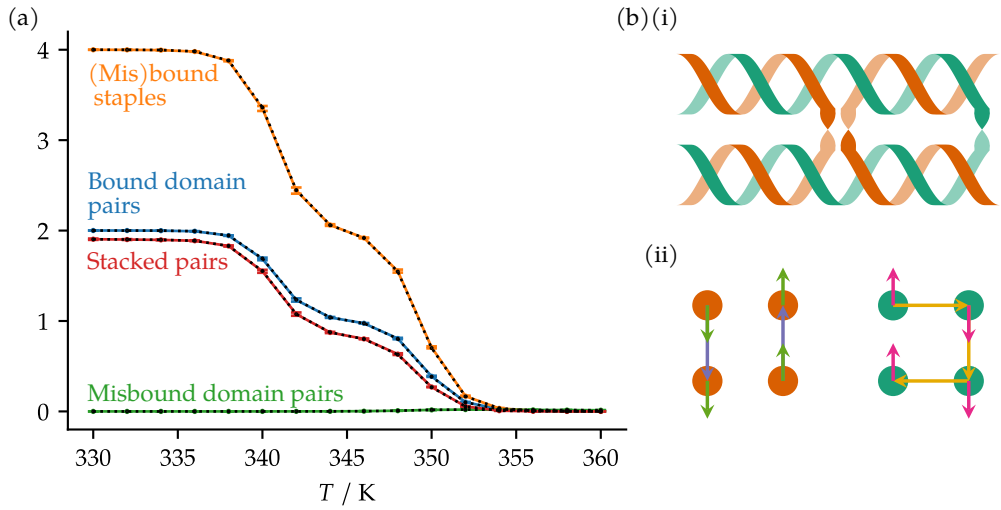

Figure 3.4: Numerical validation of the move types and their implementation. (a) Mean order parameters plotted against temperature. The exact result for each order parameter is plotted in dashed black lines. The error bars represent the standard error as calculated with the MBAR method, using data from ten independent simulation. Simulations were run with the same parameters as the simulations referred to in Figure 4.3. (b) Representation of the four-binding-domain scaffold system used. (i) Helical cartoon representation of the system in a fully stacked assembled configuration. (ii) Representation of the system with the implicit helical model. The scaffold (left) and staples (right) are shown in the assembled, planar configuration, but for clarity have been drawn separately. A full legend for all diagram elements is provided in Figure 2.1.

sembled state and the first fully stacked assembled state, expectation values of order parameters, and the effective sample size [247] (see Figure 3.5 for an example of selecting the RG move type parameters). Based on this, for both the contiguous and non-contiguous CTRG scaffold regrowth move types, we chose a maximum of one recoil, a maximum of 36 (all possible) configurations to be attempted at each growth step, and a maximum of 12 total scaffold binding domains to attempt to regrow. For the non-contiguous CTRG move type, we chose a maximum of two scaffold binding domains to attempt to regrow per segment. For the ratio of move type frequencies, we chose orientation rotation, staple exchange, staple regrowth, contiguous CTRG scaffold regrowth, and non-contiguous CTRG scaffold regrowth moves in a ratio of $2:1:1:1:1$. Finally, we make an exchange attempt between replicas every 100 steps. For many of the parameters, there was a substantial range in which the sampling efficiency was very similar, so many of the values given above could well have been chosen differently.

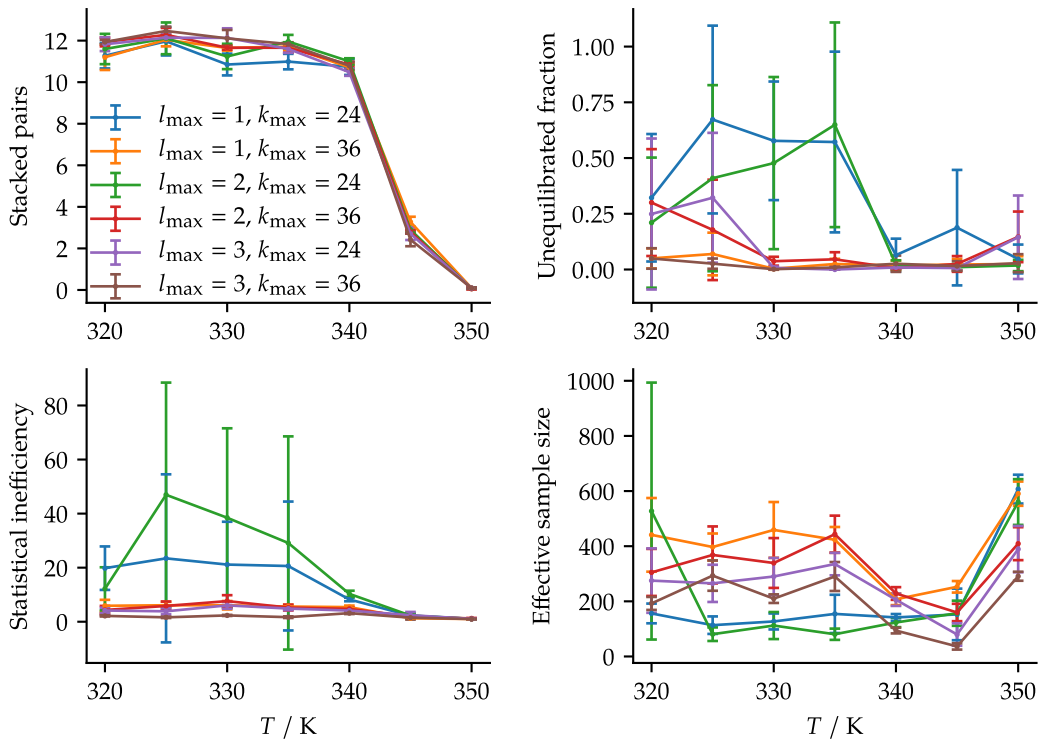

Figure 3.5: Example of analysis used to optimize move set parameters. Independent simulations were run for each temperature, which were analyzed separately (i.e. without the MBAR method). The error bars represent the standard error in the means across three independent simulations. An older version of model with an unoptimized stacking energy was used in these simulations.

4
Feasibility and validity of approach

4.1 Motivation

In the present chapter, we run simulations of the model described in Chapter 2. The goal is to test whether the model is capable of representing assembled states in a way that matches our expectations. Additionally, we want to test whether the simulation methods described in Chapter 3 are able to efficiently sample assembled states. We also provide some simple examples of the types of thermodynamic analyses possible with the model and simulation methods.

4.2 Simulation and analysis methods

The simulations are run with Hamiltonian REMC in the grand ensemble as described in Chapter 3, with staple concentration held constant across the replicas. The move set and associated parameters are those described in Section 3.11. We calculated the free-energy differences that are plotted in Figure 4.1 by directly using the ratio of stacked to unstacked states. All expectation values for the 24- and 21-binding-domain scaffold systems were calculated by taking simple means at the thermodynamic state in question; standard errors were calculated from means taken for each independent simulation run in the same state. For the 56-binding-domain scaffold system, the expectation values and standard errors were calculated with the MBAR method discussed in Section 3.9, which combines the data from all thermodynamic states and independent simulations in a single analysis.

4.3 Initial parameter selection

There are three main parameters related to the assembly conditions that must be chosen in order to run a simulation: the staple concentration, the salt concentration and the temperature. Here, we set all staples to have the same amount concentration C. While staple concentrations used vary from study to study, typical values are around 100 nM, which is the value we use here. Assembly is typically carried out in solutions with significant amounts of dissolved salts, so we set the monovalent cation concentration to be 0.5 M, which is the same concentration as that used by Snodin et al. [120] in their oxDNA simulations.

As we are running REMC simulations, we do not have to select a single temperature; the selection of the range is guided by two principles. The first is to encompass the region of the relevant order parameter curves with the steepest gradient, approximately centred on the melting temperature. The second is to keep the acceptance probabilities of replica swaps roughly uniform between each adjacent pair. In the selection of the stacking multiplier for the 2D REMC simulations, the first principle does not apply; in this case the states with other stacking multipliers are only used to improve sampling.

The physical origin of the stacking interactions involves a favourable electrostatic component and an entropic penalty. However, because some of the entropic component of the stacking free energy is accounted for by constraining the orientation vectors in bound states within our model, we assume the stacking interaction parameter is entirely energetic and treat the stacking interaction as a single temperature-independent tunable parameter. To select its value, we ran short serial simulations of a two-binding-domain scaffold with two single-binding-domain staples at a temperature below the melting temperature and calculated the free-energy difference between the stacked and unstacked states for a range of stacking energies (Figure 4.1). We selected a value of $-1000\,\mathrm{k_B\,K}$, which gave a free-energy difference that roughly matched experimentally measured values [249, 250] of $\sim\!-5\,\mathrm{kJ\,mol^{-1}}$.

A few of the model parameters are different those described in Chapter 2. First, we used a different expression for chemical potential than Equation (2.22) that did not account for the internal staple degrees of freedom. We assumed that we could write the staple-strand chemical potential to within a constant as $\mu_i = k_\mathrm{B}T\ln(C/C^{\ominus})$, where $C^{\ominus} = 1\,\mathrm{M}$. In Section 4.5, we argue that while this will affect the results in a quantitative way, the qualitative results of this chapter will hold. For the unimolecular binding reactions, we included $\Delta G^{\ominus}_{\mathrm{initiation}}$ in ϵ_u, essentially using ϵ_b for

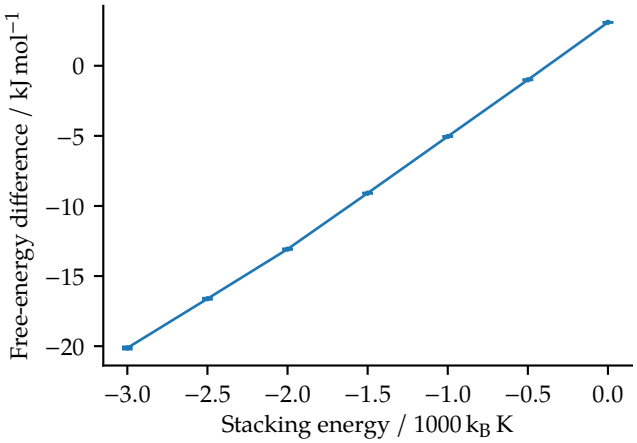

Figure 4.1: Stacking free energy as a function of stacking energy calculated from simulations of a two-binding-domain scaffold system with two one-binding-domain staples. The simulations were run at 320 K, which is below the melting temperature. The error bars represent the standard error in the means across three independent simulations. Simulations were run with a staple concentration of 100 nM and a monovalent cation concentration of 0.5 M.

all binding energies. The contribution of $\Delta G^{\ominus}_{\text{initiation}}$ to $\Delta G^{\ominus}_{\text{NN}}$ is small, and so this is unlikely to have much of an impact on the results. We also do not include the $k_B T \ln 6$ term that appears in ϵ_s from the loss of the orientation vector degrees of freedom upon binding. While this will cause a small shift in the assembly curves, it will not affect the qualitative results presented here. Finally, we do not apply any mean field corrections for changes in the model's degrees of freedom as it assembles.

4.4 Results

To test the efficacy of our model, we ran simulations of a 24-binding-domain scaffold system previously studied with the oxDNA model [120] (Figure 4.2). This system has 12 staple types that bind to the scaffold, each with two 16-nt binding domains. The REMC simulations were run with temperature as the independently controlled exchange variable with 16 replicas in under three hours of walltime on a commodity cluster. To determine the extent of assembly, we look at two order parameters: the number of staples in the system, whether bound or misbound (henceforth referred to as (mis)bound), and the number of bound domain pairs that have formed. As can be seen in Figure 4.3(a), at low temperatures the system has the number of (mis)bound staples and the number of bound domain pairs expected in the assembled state. The error bars are quite narrow, which gives us confidence that the averages have converged. In Figure 4.3(b), we show a typical assembled configuration. Unlike

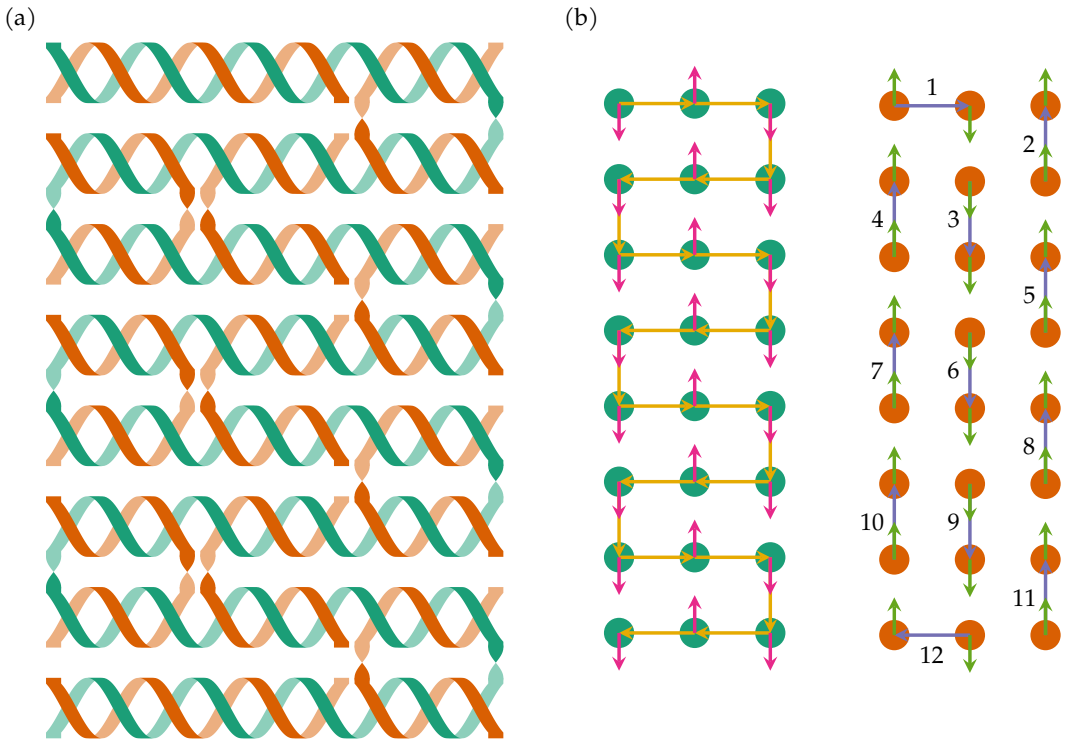

Figure 4.2: Schematic representations of the 24-binding-domain scaffold system. (a) Helical cartoon representation of the system in a fully stacked assembled configuration. (b) Representation of the system with the lattice model. The scaffold (left) and staples (right, numbered) are shown in the fully stacked assembled configuration, but for clarity have been drawn separately. A full legend for all diagram elements is provided in Figure 2.1.

its schematic representation in Figure 4.2, the conformation of the scaffold is not planar. That a typical assembled configuration is not a well-ordered planar state is reasonable because the scaffold is relatively unconstrained by staple crossovers, in part because the crossovers that occur connect only relatively close segments of the scaffold. Moreover, non-planar configurations were also found to be typical of the assembled state of the same system in oxDNA simulations [120].

We can examine the extent of the assembled state's structural disorder by looking at the number of stacked binding domain pairs. In the assembled state, the planar configuration is also the configuration that maximizes the number of stacked binding domain pairs. As can be seen in Figure 4.3, the average value of this order parameter converges to a value that is well below the fully stacked assembled configuration at the lower temperatures. Nevertheless, an examination of a step series of the number of stacked binding domain pairs (Figure 4.4) reveals that, while the average number of stacked binding domain pairs is below that of a fully stacked assembled

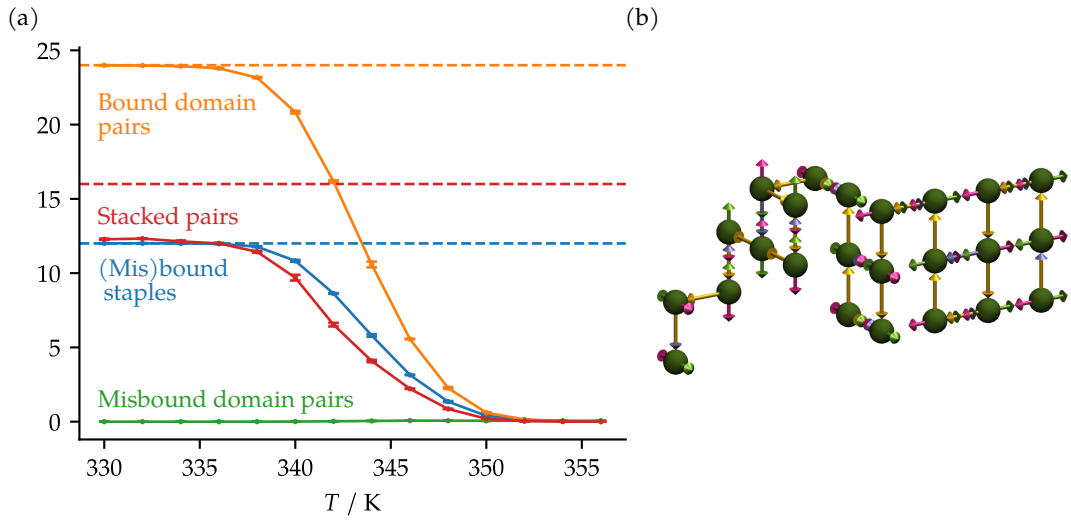

Figure 4.3: Mean order parameters as a function of system temperature and a typical assembled configuration of a 24-binding-domain scaffold system. Simulations were run with a staple concentration of 100 nM, a monovalent cation concentration of 0.5 M, and a stacking energy of $-1000\,k_B\,K$. (a) Mean order parameters plotted against temperature. The black dashed lines correspond to the expected order parameter values in the fully stacked assembled configuration. The error bars represent the standard error in the means across three independent simulations. (b) An assembled configuration at 330 K.

system, the simulation does sample such configurations. The fact that the simulations generate such configurations and the large degree of fluctuation in the number of stacked domain pairs gives us confidence that the simulation methods are able to sample origami configurations effectively even in near- and fully assembled states. Simulations generally result in full assembly within about one hundred seconds of walltime, and fully stacked assembled configurations within an hour.

The efficiency of our model and sampling methods allows us to run simulations across a range of assembly conditions and design parameters. While 100 nM is a typical value for staple concentrations, the concentration used for a particular assembly protocol commonly varies from tens to hundreds of nM. To see how staple concentration affects the assembly of this system within and beyond the ranges found in experimental conditions, we ran simulations with staple concentrations from 1 nM to 1 mM in intervals of factors of 10. As can be seen in Figure 4.5, at low temperatures, from 1 nM to 1 μM, the order parameters indicate that the assembled state is the prevalent structure, with the melting temperature shifting to higher values as the concentration is increased. However, at 10 μM, the average number of (mis)bound staples exceeds 12 and the number of misbound domain pairs is near zero, indicating that at least some of the configurations now have two of the same

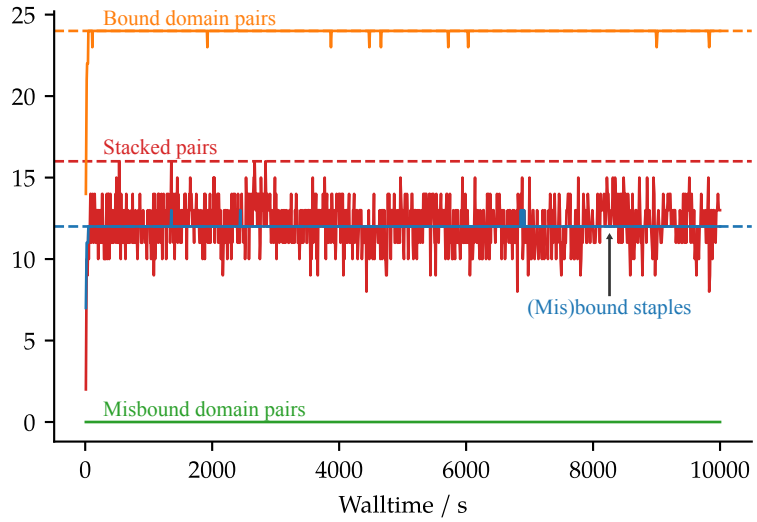

Figure 4.4: Order parameter step series for a 330 K replica of a REMC simulation of a 24-binding-domain scaffold system. The simulations were run on a single node of a commodity cluster. The black dashed lines correspond to the expected order parameter values in the fully stacked assembled configuration.

type of staple bound to the scaffold. This situation is referred to as 'blocking' [120] because such staples prevent each other from fully binding to the scaffold. This is also approximately the staple concentration used in the simulations of Snodin *et al.* [120], who speculated that they seemed to be in a range in which blocking was somewhat favourable. Blocking becomes substantially more prominent at 100 μM, at which there are now significant contributions from configurations that have blocked staples for more than one staple type. The number of stacked binding domain pairs also significantly increases at such high staple concentrations. The reason for this behaviour is that with multiple staples of the same type bound to the system, there will be fewer crossovers, which can allow for longer segments of stacked helices. Also at 100 μM, misbinding begins to increase, and becomes even more substantial in the millimolar regime. This can occur as the staples are able to misbind to staples already bound to the scaffold.

Because our choice of the stacking energy was somewhat crudely determined, we ran further simulations with a range of stacking energies to see how strong an effect the choice can have on the thermodynamic assembly behaviour. According to Figure 4.6, the average number of stacked binding domain pairs changes quite dramatically when the stacking energy is halved or doubled. The melting temperature is also shifted, although not as dramatically. When the stacking energy is doubled, the number of stacked binding domain pairs plateaus at nearly the value expected in the planar state. However, beyond this, the average number of (mis)bound staples

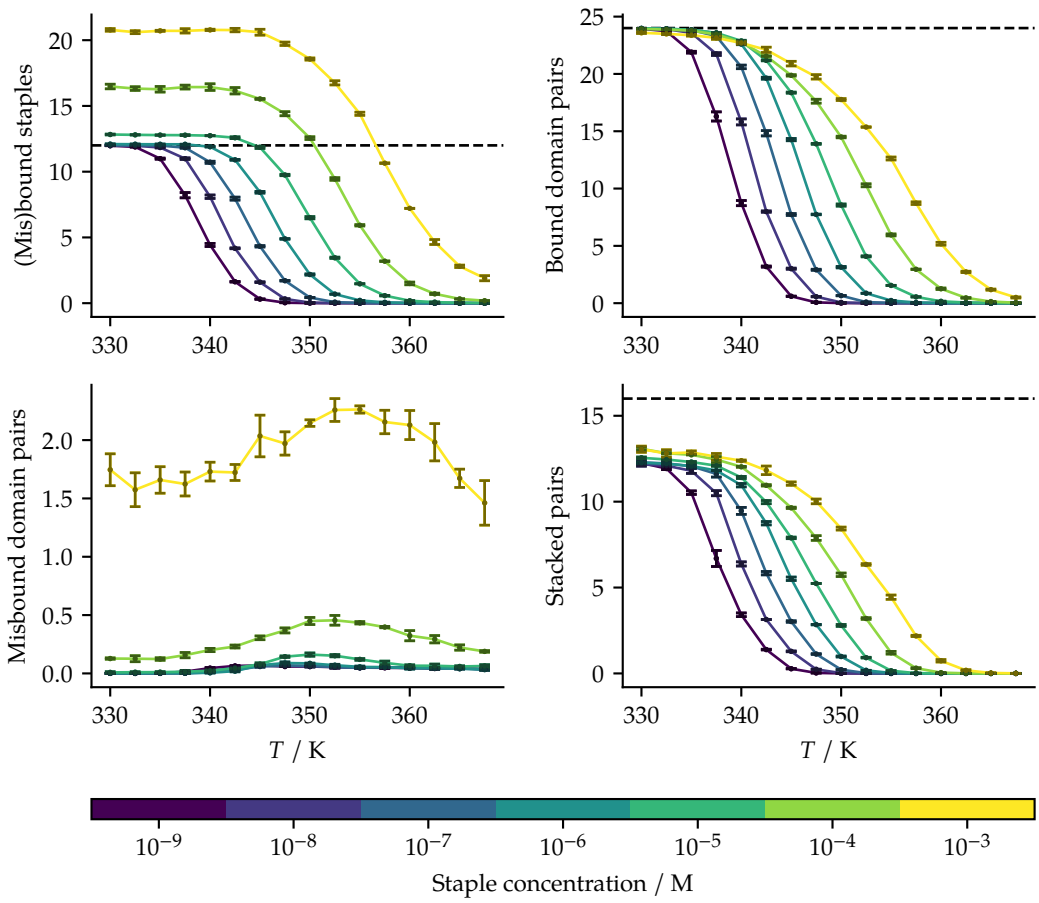

Figure 4.5: Mean order parameters from simulations of the 24-binding-domain scaffold system plotted against temperature for a range of staple concentrations. The black dashed lines correspond to the expected order parameter values in the fully stacked assembled configuration. The error bars represent the standard error in the means across three independent simulations. Simulations were run with a monovalent cation concentration of 0.5 M and a stacking energy of $-1000\,k_B\,K$.

exceeds that expected in the assembled state. At high enough values of the stacking energy, even the number of stacked pairs begins to exceed that expected in the assembled state. This is possible because if the double crossover binding domains at the edges bind two copies of the same staple, they will be able to form an additional stack as part of a single helix, rather than crossing over, as they would in the desired assembled state. It seems that if the stacking energy is sufficiently favourable, the entropic cost of binding multiple copies of the same staple can become overshadowed by the stacking energy. While our choice of stacking energy is in the right range, because of the sensitivity of the average number of stacked binding domain pairs to this value, if we want to make a more direct comparison between our model and real experiments, the stacking energy should be tuned such that the ratio of planar

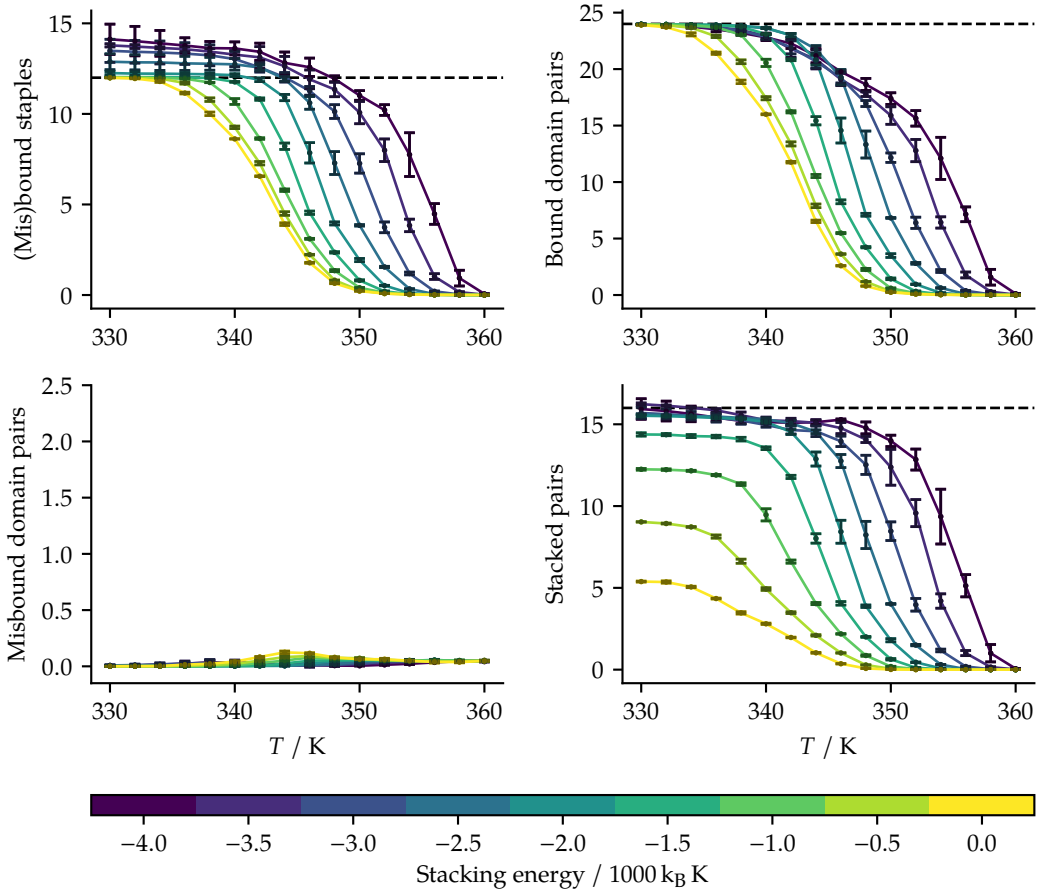

Figure 4.6: Mean order parameters from simulations of the 24-binding-domain scaffold system plotted against temperature for a range of stacking energies. The black dashed lines correspond to the expected order parameter values in the fully stacked assembled configuration. The error bars represent the standard error in the means across three independent simulations. Simulations were run with a staple concentration of 100 nM and a monovalent cation concentration of 0.5 M.

to non-planar configurations matches experimental values.

While salt concentration can also play a role in the self-assembly behaviour, its primary effect is to shift the melting temperature slightly (see Figure 4.7). We have only included monovalent cation dependence in our version of the NN model, but non-monovalent cations, particularly Mg^{2+}, are commonly used in experimental set-ups. Such ions can be accounted for in a crude manner by simply increasing the effective monovalent cation concentration. It is possible to use more general corrections to account for such ions within the NN model [180]; however, since the effect is relatively small given our model's intended level of accuracy, we have not included such corrections at present.

In order to see how the thermodynamic behaviour of individual staples is affected

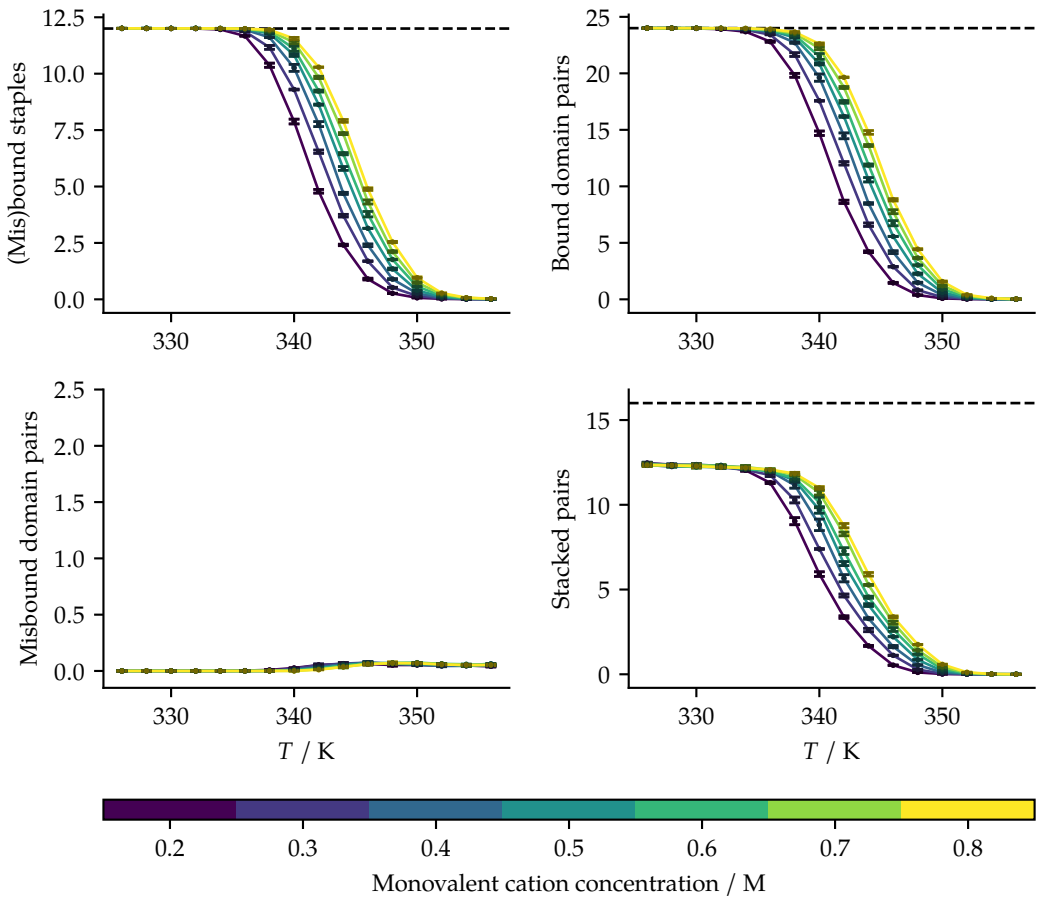

Figure 4.7: Mean order parameters from simulations of the 24-binding-domain scaffold system plotted against temperature for a range of sodium ion concentrations. The black dashed lines correspond to the expected order parameter values in the fully stacked assembled configuration. The error bars represent the standard error in the means across three independent simulations. Simulations were run with a staple concentration of 100 nM and a stacking energy of $-1000\,k_B\,K$.

by the scaffold, we ran simulations of the same system, but with the sequence specific hybridization free energies replaced by their average value. We computed two different averages: one over all the bound pairs and one over all the misbound pairs. In Figure 4.8, the mean staple occupancy curves are plotted for all staples in the system. In general, staples with two binding domains can be classed by the number of scaffold binding domains that are spanned by the staple binding domains in the assembled structure. In the target structure we are assembling here, there are those that span zero and two scaffold binding domains, as well as those that do not have a crossover at all.

The curves of the individual staples turn out to be grouped by these structural classifications. Those with the highest melting temperatures, which we define as the

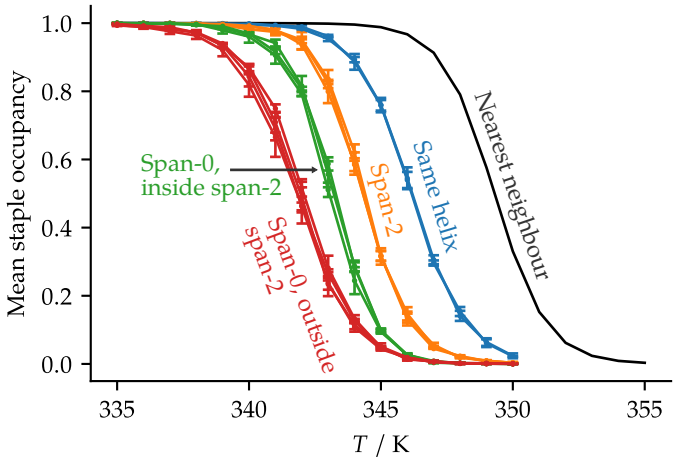

Figure 4.8: Mean staple occupancy from simulations of the 24-binding-domain scaffold system plotted against temperature for simulations with averaged hybridization free energies. The NN values were calculated directly with the averaged hybridization free energies by assuming the staple strands are in excess of the scaffold strands. The error bars represent the standard error in the means across three independent simulations. Simulations were run with the same parameters as the simulations referred to in Figure 4.3.

midpoint of the mean occupancy curves, are those that have no crossovers (same helix; see Figure 4.2(b), staples 1 and 12), followed by those that span two scaffold binding domains (span-2; see Figure 4.2(b), staples 3, 6, and 9), followed by those that span no scaffold binding domains (span-0; see Figure 4.2(b), staples 2, 4, 5, 7, 8, 10, and 11). The staples with no crossovers are expected to be the most stable, as they have an extra stacking interaction between the two domains compared to the staples that have a crossover. Further, these staples happen to occur at the termini of the scaffold, so the rigidity that they introduce is placed at a point that will restrict the configuration of the scaffold the least. However, these staples are still shifted to lower melting temperatures relative to the pure NN curve.

The staples that span no scaffold domains are edge staples, and thus have half as many potential stacking interactions available to them. These staples are also involved in two crossovers, one with the staple strand, and one with the scaffold strand. A double crossover will restrict the configuration of the scaffold more than a single crossover, so, combined with reduced stacking interactions, it is expected that these double crossover bound domains pairs will have a lower melting temperature than those staples that span two scaffold domains.

The curves of the staples involving double crossovers are further split into two distinct groups. The staples with higher melting temperatures turn out to be those that are within the span of a staple that spans two scaffold binding domains (span-0,

inside span-2; see Figure 4.2(b), staples 4, 7, and 10), while the staples with the lower melting temperature are those that are not within the span of any other staples (span-0, outside span-2; see Figure 4.2(b), staples 2, 5, 8, and 11). Again, this seems reasonable because the staples that span two scaffold domains will already restrict the scaffold, such that there is a smaller entropic penalty for the staples within their span.

This analysis suggests another possible way of selecting the stacking energy: we could choose a value at which the mean staple occupancy curve of a two-binding-domain helix with no breaks in the backbone overlaps with the NN mean staple occupancy curve. Simulations of such a system reveal that the stacking energy would need to be approximately double that of the value that we selected via the comparison of stacking free-energy differences between our model and experiment. However, the number of stacked binding domain pairs for a system with such a favourable stacking energy (Figure 4.6) suggests that this would make the system on average nearly planar, which contradicts the simulations of Snodin *et al.* [120]. This suggests that the entropy differences between pairs of bound-domain pairs with an intact backbone and pairs without are not as large as they should be, and so it may be that one stacking term should be used for staples that bind to two contiguous scaffold binding domains, and another for all other pairs. Nevertheless, for the level of accuracy this model is designed for, it may be sufficient to choose a single stacking energy that is optimal for staples that are involved in crossovers, as staples that bind contiguously to the scaffold at multiple binding domains to form a single helix are uncommon. It may also be sufficient to use the mean field approach outlined in Chapter 2 for a staple that binds two contiguous scaffold domains to form a single helix.

The 24-binding-domain scaffold system contains staples that span at most two scaffold binding domains, which are relatively short spans compared to typical origami structures. To test whether our sampling methods are able to handle a system with staples that span longer regions of the scaffold, we also ran simulations of a 21-binding-domain scaffold system with staples that span 0, 2, 4, 6, 8, 10 and 12 scaffold binding domains (Figure 4.9). The fully stacked assembled state comprises three parallel helices composed of seven binding domains each; the design is a subset of the tile design of Dunn et al. [109] (the top three rows on the left side of the seam).[1] The assembly of this system is further complicated by the presence of single-binding-domain staples, which are expected not to bind until significantly lower

[1]We removed the binding domains of staples that are complementary to a scaffold binding domain outside of the included subset, but otherwise use the sequences as given in [109].

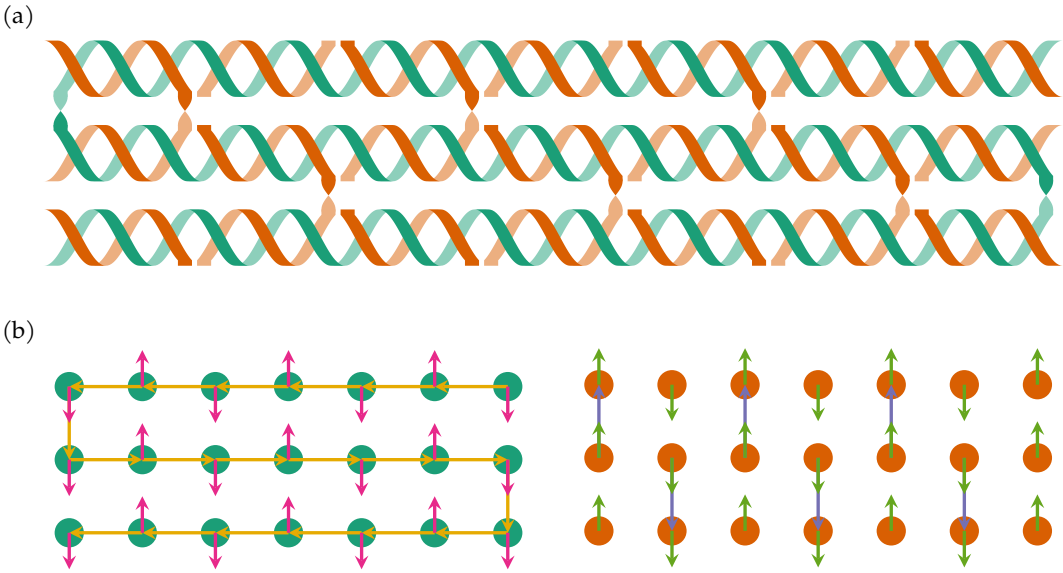

Figure 4.9: Schematic representations of the 21-binding-domain scaffold system. (a) Helical cartoon representation of the system in a fully stacked assembled configuration. (b) Representation of the system with the lattice model. The scaffold and staples are shown in assembled configurations, but for clarity have been drawn separately. A full legend for all diagram elements is provided in Figure 2.1.

temperatures than the two-binding-domain staples. The simulations were again run for under three hours of walltime on a commodity cluster. The relevant order parameters as a function of temperature are plotted in Figure 4.10. As with the 24-binding-domain scaffold system, at low temperatures, the system is assembled but not fully stacked. The order parameter curves now display two distinct regions and do not approach the assembled state values until significantly lower temperatures, as expected.

Since converged values were able to be obtained from both systems in only a few hours, we decided to further increase the complexity of the design along with the total number of binding domains and staple types involved. We again took a subset of the tile design of [110], this time considering the top 4 rows from both halves such that some of the seam staples were included (see Figure 4.11). The seam staples are those that connect the two mirrored halves of the design. This system has a 56-binding-domain scaffold with 34 staple types.[2] REMC simulations were run, this time with 18 replicas; the means are plotted in Figure 4.12. Unlike the previous two systems, convergence was not achieved in 3 hours. In fact, it took over a week

[2]As with the 21-binding-domain scaffold system, we removed the binding domains of staples that are complementary to a scaffold binding domain outside of the included subset, and additionally modified the sequences of some of the single-binding-domain staples to compress the temperature range over which assembly occurs.

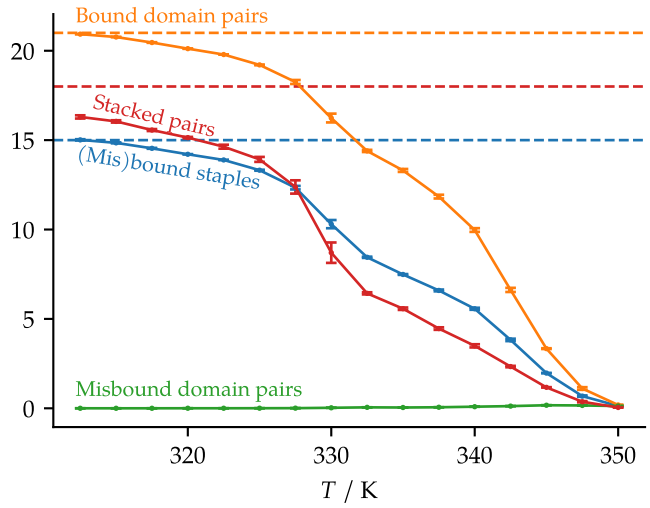

Figure 4.10: Mean order parameters from simulations of the 21-binding-domain scaffold system plotted against temperature. The black dashed lines correspond to the expected order parameter values in the fully stacked assembled configuration. The error bars represent the standard error in the means across three independent simulations. Simulations were run with the same parameters as the simulations referred to in Figure 4.3.

of walltime before all three independent simulations had sampled fully stacked states. This seems to imply a very non-linear scaling in the sampling efficiency with increasingly large designs. It is unlikely that there is a simple scaling law for the convergence time needed by our algorithm as a function of DNA origami design size, as the design itself likely has a large effect on the sampling efficiency, and more complex designs are only possible with larger origamis.

In contrast to the time required to sample a fully stacked assembled state, the time to sample any fully assembled state was around a day of sampling time, which is nearly an order of magnitude smaller. It would seem that the slowest timescale is the sampling of stacked states. While an individual stacked pair does not contribute a substantial amount to the stability at these temperatures, if many pairs must be unstacked and restacked to transition between different relevant stacked states, then there would be substantial barriers to sampling these states. Therefore, to see if we could increase sampling efficiency, or at least reduce the walltime required to achieve sufficient sampling, we employed 2D REMC, with a multiplier on the stacking energy acting as the second independent exchange variable.

We ran the 2D REMC simulations with the same temperature range, but added in 10 different stacking multipliers ranging between zero and one, for a total of 180 replicas. Because of the number of threads required for a single simulation, we

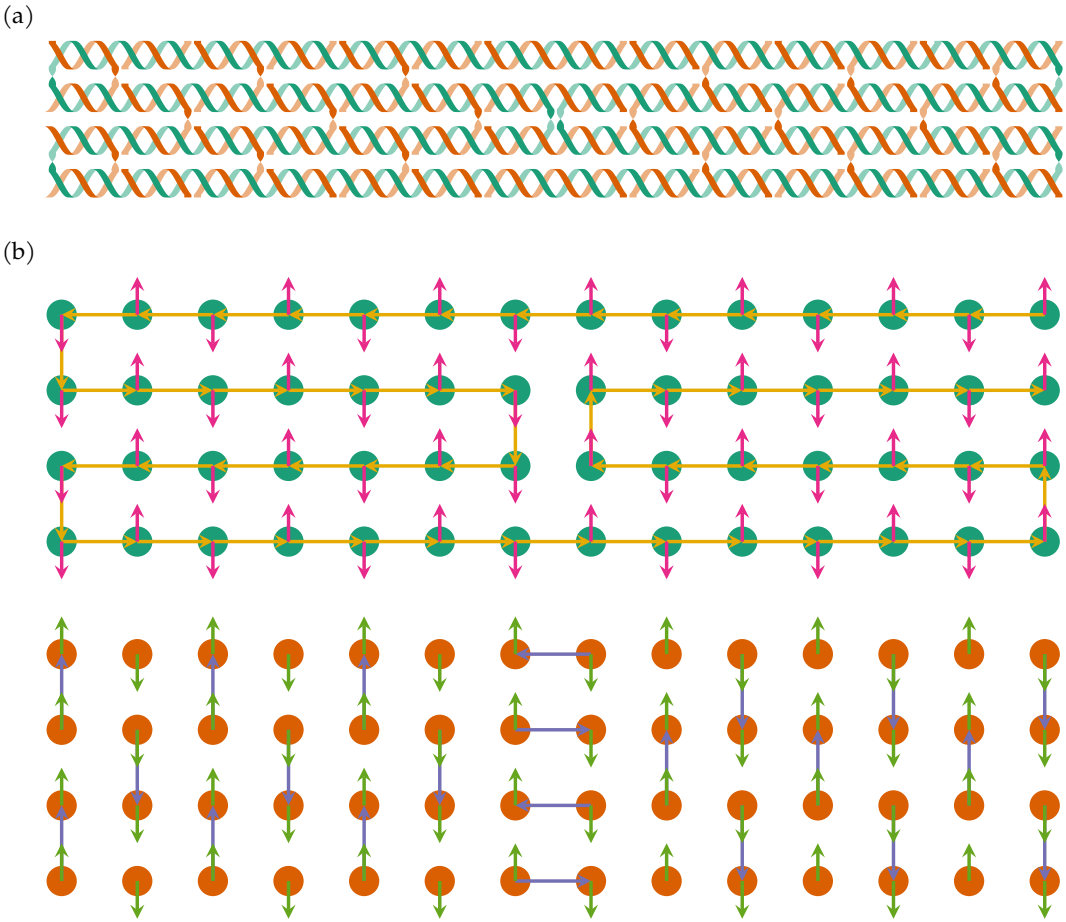

Figure 4.11: Schematic representations of the 56-binding-domain scaffold system. (a) Helical cartoon representation of the system in a fully stacked assembled configuration. (b) Representation of the system with the lattice model. The scaffold and staples are shown in assembled configurations, but for clarity have been drawn separately. A full legend for all diagram elements is provided in Figure 2.1.

ran only two independent simulations. The first fully stacked state was sampled within two days in the first and within four days in the second. While this is an improvement in terms of walltime, in terms of total resources it is less efficient.

We continued the first simulation for a little under 6 days in order to obtain a well converged sample. As can be seen in Figure 4.12, while the number of (mis)bound staples and the number of bound domain pairs are in good agreement, the number of stacked pairs is higher in the 2D REMC simulations. The convergence should be better in the 2D REMC simulations with respect to the number of stacked pairs because of the inclusion of states with weaker stacking interactions, and so we assume that the result reveals a lack of convergence in the 1D REMC simulations. It is unclear how much longer the 1D REMC simulations would need to be run to

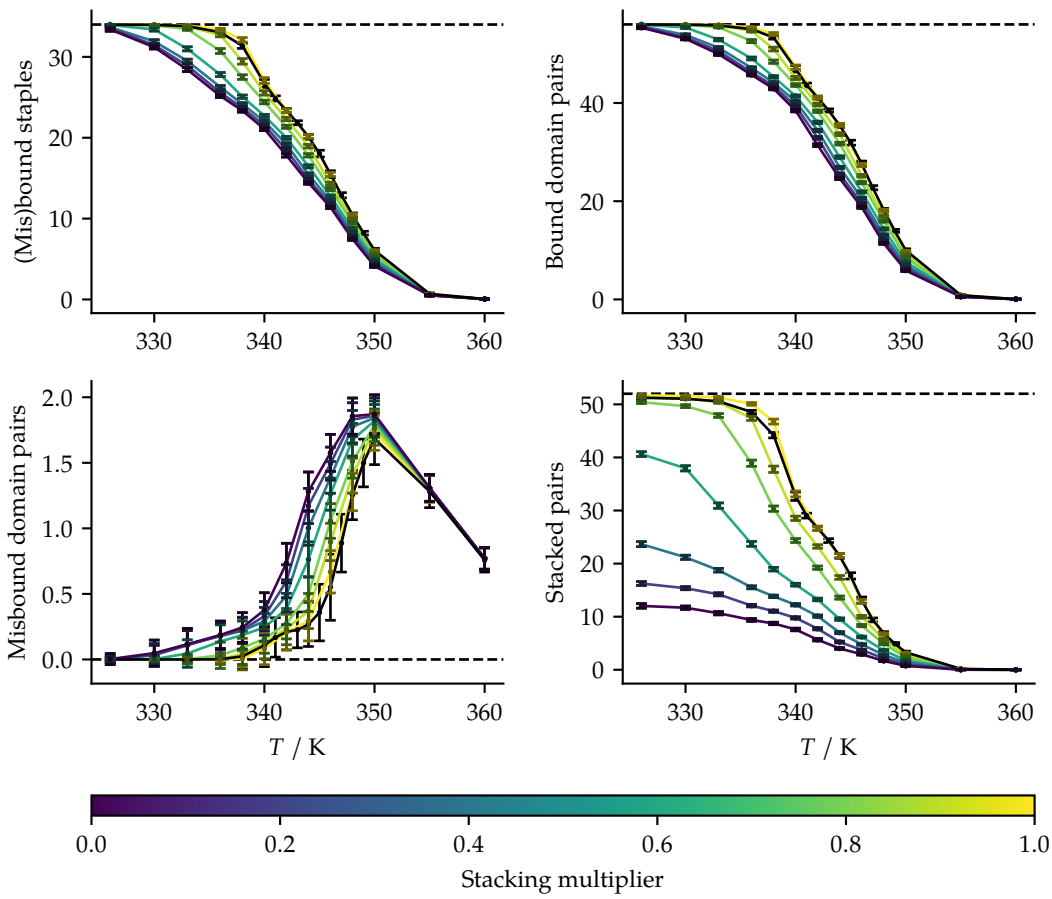

Figure 4.12: Mean order parameters from simulations of the 56-binding-domain scaffold system plotted against temperature. The black solid lines correspond to the 1D REMC simulations, while the coloured lines correspond to the 2D REMC simulations. The black dashed lines correspond to the expected order parameter values in the fully stacked assembled configuration. The error bars represent the standard error as calculated with the MBAR method, using data from one (2D REMC) or three (1D REMC) independent simulations. Simulations were run with the same parameters as the simulations referred to in Figure 4.3.

achieve the same level of convergence, but this underlines the difficulty in not only achieving good sampling, but determining whether sampling has converged at all. Thus, while the 2D REMC did not improve initial sampling of fully stacked states, it may still be of use to achieving properly converged samples.

A 2D REMC simulation of the 56-binding-domain scaffold system with a uniform potential was run to carry out a similar analysis of the thermodynamics of individual staple types as was done for the 24-binding-domain scaffold system. The same two averages as before were calculated for use by the uniform potential, but with the sequences from this system. The mean staple occupancies as a function of temperature

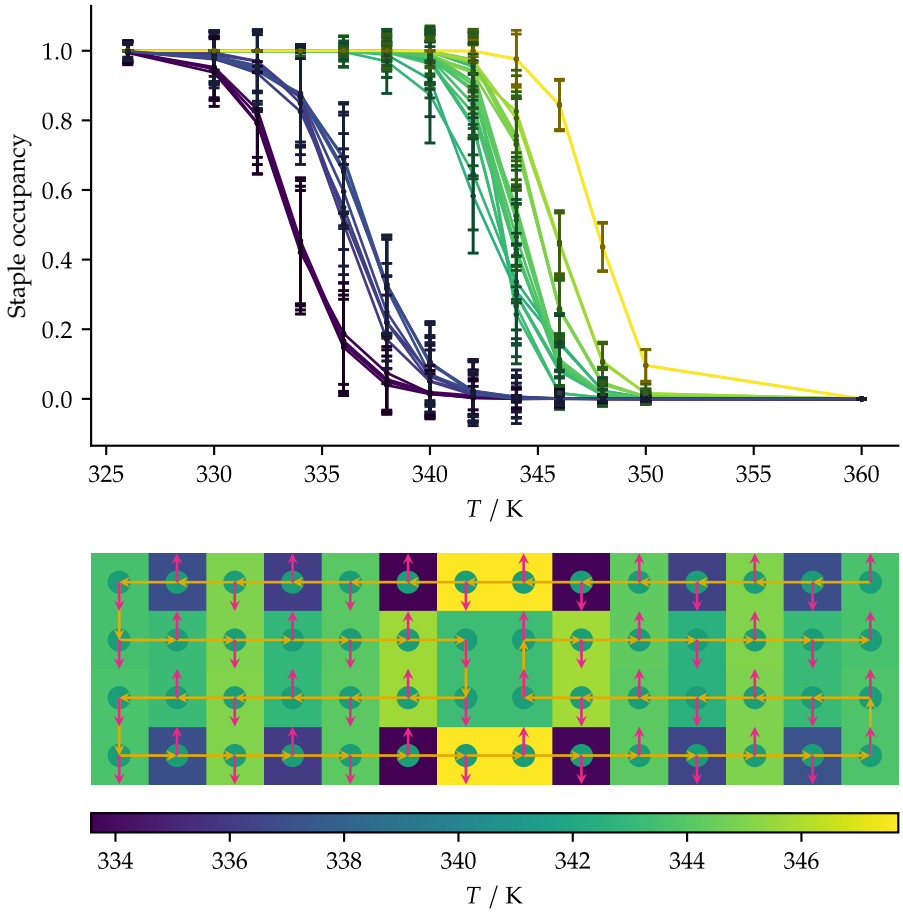

Figure 4.13: Mean staple occupancy (top) and melting temperatures (bottom) from simulations of the 56-binding-domain system with averaged hybridization free energies. The mean staple occupancy curves are coloured according to the melting temperatures. The melting temperatures are plotted on a grid corresponding to the staple locations in the fully stacked assembled state. The error bars represent the standard error as calculated with the MBAR method, using data from one independent simulation. For reference, the scaffold in a fully stacked assembled configuration has been superimposed on the melting temperature heat map; see Figure 4.11 for the corresponding staple configurations. Simulations were run with the same parameters as the simulations referred to in Figure 4.3.

are plotted in Figure 4.13. Here, there are many more staple types and environments, so instead of classifying the staple types structurally as before, we simply group them by their melting temperature, which we estimate by interpolation.

There are four clear groupings of staple types by melting temperature, and as can be seen from the schematic plot of the melting temperatures, there is a clear symmetry between the left and right half, as anticipated. The most stable are the staples that have no crossovers, and are essentially one binding domain that is twice as long as the rest. The least stable are, as expected, the single-binding-domain staples. However,

it is surprising that the four single-binding-domain staples that are adjacent to the most stable staples are less stable than the other eight single-binding-domain staples. This does not seem to be explainable with simple topological considerations; we can only speculate that this is caused by subtleties in the model potential or to a lack of sufficient sampling.

The two-binding-domain staples that have a crossover fall into the remaining cluster, and the errors in the curves overlap amongst many of them. However, there are still a few observations that can be made. First, the two staples that have the highest melting temperatures are actually the ones that span the second fewest scaffold binding domains. The explanation for the staples spanning the fewest scaffold binding domains not being the most stable is similar to that given for the same observation in the 24-binding-domain system: the staples that span the fewest scaffold domains are edge staples and so have half as many stacking interactions available, and also are double crossover staples, so are somewhat more constrained and thus incur a higher entropic penalty upon binding. Second, the seam staples, which span 26 scaffold binding domains – far more than any other staple type – are in the middle of the cluster, which naively would be unexpected given the much larger entropic cost of closing such a large loop. That they have a similar melting temperature to the other two-binding-domain staples is an indication of cooperativity playing a role in staple binding: the binding of other staples in the system reduces the size of the loop that these staples must close, and unlike any other staple in the system, it is effectively the same loop that both seam staples are closing.

4.5 Conclusions

We have introduced a model and sampling methods for simulating DNA origami self-assembly that is computationally feasible, yet includes the structural information most relevant to the assembly process. We demonstrated that small origamis can be sampled efficiently enough to achieve good statistics for not only one particular set of assembly conditions and design parameters, but for a range of values of these variables. It is difficult to predict how the approach will scale with system size, as we expect this may be highly dependent on the specifics of the origami design. However, even if simulating the self-assembly of very large systems may not yet be tractable, much insight can be gained from studying smaller origami system.

For example, we can use the model to study thermodynamic properties of origami designs, such as the relative stability of staples, the types and degree of staple binding cooperativity, or the effects of scaffold routing and loop closure on the cost of staple

binding. Because we use MC simulations to sample configuration space, we cannot directly study dynamical quantities. However, we can calculate free-energy barriers along selected order parameters, which in turn could be used to estimate relative rates between different assembly pathways. Such calculations would allow us to pursue questions relating to the kinetics of assembly, such as whether there is a nucleation barrier, and how it depends on assembly conditions and staple design (which we address in Chapter 5). We may also be able to shed some light on whether and why hysteresis occurs for a given design and set of assembly conditions.

There are several caveats to our approach. We assume that the staples are always in excess of the scaffold. If that were not the case, the assumption that the free staple concentration remains constant regardless of the degree of assembly would become less convincing. One solution may be to reduce the free staple concentration relative to the total staple concentration based on the average number of staples (mis)bound to the scaffold. Of course, because simulations must be run to determine the average staple occupancy on the scaffold, this would require an initial guess and subsequent iterations to converge to a consistent value. An alternative solution may be to make the free staple concentration a function of the number of staples currently bound to the system. While not ideal, it may be sufficient for the level of accuracy the model is intended to provide.

In the derivation of the model in Chapter 2, we make a different assumption about the relationship between the chemical potential and the staple concentration that takes into account the orientation degrees of freedom present on each binding domain. For systems that have only one staple length, this has the effect of shifting the staple concentration to higher values for a given chemical potential. In the case of the simulations presented in this chapter, the staple concentration used would become an order of magnitude larger. While for the 24-binding-domain scaffold system this would change the results presented here quantitatively, the qualitative interpretations do not change. However, because the 21- and 56-binding-domain scaffold systems have single binding domain staples, it does have the effect of narrowing the temperature range over which assembly occurs, an effect we deem more important than an overall shift.

It has been found experimentally that the stacking free energy is sequence-specific [249, 250] and depends on both temperature and salt concentration. It has been observed to range from below $-10\,\mathrm{kJ\,mol^{-1}}$ to slightly above $1\,\mathrm{kJ\,mol^{-1}}$ (i.e. for some sequences and conditions, stacking is slightly disfavoured), which corresponds to a stacking energy range in our model of half to double the chosen

value. As discussed in Section 4.4, the mean number of stacked binding domain pairs shifts substantially over this range of stacking energies. While some of the temperature dependence is taken into account here by the explicit modelling of some of the entropic contribution to the stacking free energy, the sequence specificity and salt dependence are not accounted for. For this pilot study, a roughly selected constant value is sufficient to demonstrate that the model is reasonable, but in future studies, we may also consider using sequence-specific salt-dependent stacking energies for more accurate predictions for a particular design.

Finally, in Chapter 2 we used relatively simple arguments to support our choices of which kinked configurations are to be allowed and which are to be disallowed within our model. It may well be that different choices could improve both the reproduction of the balance of the energy/entropy trade-off of stacked assembled configurations and the structural accuracy of the model. Using a more detailed DNA model such as oxDNA, one could run simulations of helices with breaks in the backbone or simulations of helices with crossovers in order to provide a more detailed reference point from which to determine which configurations are sensible to allow. Furthermore, another term could be introduced into the potential to weight kinked configurations based on their frequency in the higher resolution simulations. If an even more accurate model were desired, we could consider implementing the explicit helical axis model of Section 2.3. Such a model would allow us to control more finely the level of flexibility afforded to kinked segments in the structure and would make it more straightforward to prevent some of the non-physical configurations without introducing further many-body interactions.

While such modifications may improve the accuracy of the model, they would also be costly in both development time and simulation time, and we do not expect they would fundamentally alter the results, but rather may incrementally improve them. For studying fundamental aspects of DNA origami self-assembly, we believe that such expensive incremental improvements are likely to be of marginal use. We are therefore hopeful that the use of our model will be able to yield both fundamental and practical insights into the thermodynamics and kinetics of DNA origami self-assembly.

5
DNA origami and nucleation

5.1 Motivation

While there is much practical knowledge on how to optimize the assembly of DNA origamis, an understanding of the underlying physical reasons—such as the nature of any free-energy barriers to assembly and how they change with assembly conditions—is lacking. Many studies on DNA origami have found hysteresis between melting and annealing as the temperature is varied [98, 109–115, 117, 251], where the annealing curves tend towards the melting curves when the reactions are carried out over a longer time [98, 110]. This suggests that the temperature-ramp assembly process is out of equilibrium, and that there are significant free-energy barriers present.

One class of barrier to consider are nucleation barriers to staple binding. These barriers can be split further into two classes: those that inhibit staple binding, and those that inhibit the scaffold from folding up to its designed shape. The analogy to crystal nucleation is more clear with the first class, while the second can be related to the less well defined idea of nucleation in protein folding [252]. It has been suggested that the melting–annealing hysteresis could be attributed to a nucleation barrier to staple binding [98, 116, 253], but no concrete evidence has been given to show that such a barrier exists.

Perhaps the most closely related work has been studies of the self-assembly of "DNA brick" structures [7, 8], which consist of a large number of short unique strands that assemble in the absence of any longer scaffold strand. DNA-brick self-assembly does entail a nucleation barrier, which plays an important role in allowing error-free assembly of these many-component systems [190]. The origins of this barrier and details on the pathways taken have been studied in some depth, uncovering its very non-classical behaviour, which has the potential to inform future rational designs of these systems such that they have favourable assembly kinetics [191–193, 200, 254]. However, while the DNA origami assembly process has been modelled with success

already [109, 110, 120], and has even benefited experimental work [255], none of these efforts specifically examined the role of nucleation or calculated the associated free-energy landscapes.

Here, we use the lattice model of DNA origami described in Chapter 2 to determine under what conditions nucleation barriers affect the assembly process. We simulate three systems with a sequence-specific potential and initially find no nucleation barriers to assembly. We then focus on a set of three additional systems with a sequence-averaged potential that allows us to vary the number of binding domains per staple. The simulations of these systems reveal a nucleation barrier that, while significant, is not prohibitively large relative to thermal fluctuations.

5.2 Simulation and analysis methods

We again run Hamiltonian REMC simulations in the grand ensemble with the staple concentration held constant across the replicas, as described in Chapters 3 and 4. Here we consider systems with staples that have up to four binding domains.

We use the same default assembly condition parameters and stacking energy as discussed in Section 4.3. For some of the REMC simulations that used temperature as the independently controlled exchange variable, the temperatures are generated in an iterative, automated fashion. In these simulations, the temperatures are initially spaced uniformly, but at regular intervals the averages of a selected order parameter are calculated and used to select a new temperature set. The new temperature set is selected with a simple linear interpolation such that the averages of the selected order parameter will be spaced uniformly, with the exception of a slightly higher density of temperature points at the upper and lower temperature range. In addition to the use of temperature as an exchange variable for the REMC simulations, we also separately use a multiplier on the stacking energy as an exchange variable. The stacking multipliers could be iteratively updated in a similar manner, although here we use a uniform spacing.

Like the simulations of Chapter 4, some of the model parameters are different from those in Chapter 2. Unlike in Chapter 4, we do use Equation (2.22) for the chemical potential. However, like Chapter 4, we included $\Delta G^{\ominus}_{\text{initiation}}$ in ϵ_{s}. The $k_{\text{B}}T \ln 6$ term was included for ϵ_{b}, but not for ϵ_{u}. Again, while such differences will cause a small shift in the assembly curves, they will not affect the qualitative results presented here. We also do not apply any mean field corrections for changes in the model's degrees of freedom as it assembles.

5.3 Results

To begin our investigation of nucleation barriers in DNA origami self-assembly, we ran REMC simulations of three small designs and calculated LFEs for several order parameters. The first is a 24-binding-domain scaffold system with 12 staple types, each with two binding domains (Figure 4.2). The second system is a 21-binding-domain scaffold system with 7 two-binding-domain staple types and 7 single-binding-domain staple types (Figure 4.9). The third is a 56-binding-domain scaffold system with 22 two-binding-domain staple types and 12 single-binding-domain staple types (Figure 4.11). The 21- and 56-binding-domain scaffold systems are subsets of the monomer tile design used by Dunn et al. [109]; the 21-binding-domain scaffold design is taken from the top three helices on the left side of the seam, while the 56-binding-domain scaffold design is taken from the top four helices (including both sides of the seam). We have previously simulated these systems with our model in Chapter 4, and the 24-binding-domain scaffold system has also been simulated with the oxDNA model [120].[1]

To quantify the progress of the assembly reaction, we must select order parameters along which we can construct a free-energy landscape. The first order parameter we considered was the number of (mis)bound staples (defined as in Section 4.4). This is perhaps the closest analogue of the order parameter used in simulation studies of DNA bricks, i.e. the number of bricks in the largest cluster [190, 192, 200, 256]. We calculated the LFEs along this order parameter for each system at each temperature (Figure 5.1) As the temperature decreases, the number of bound staples shifts towards higher values. At the lowest temperatures, the favoured state has the number of bound staples expected in the assembled state. Most importantly, for all systems at all temperatures, the free energy is always downhill to the favoured state, implying that there is no nucleation barrier to assembly, in contrast to the self-assembly of DNA bricks [190].

One problem with the above order parameter is that it cannot be used to determine whether the system is fully assembled, and it does not directly show the extent to which the scaffold is correctly folded. By way of example, suppose that a copy of each staple type is bound to a completely unfolded scaffold. This would be the case if each staple was bound at only one of its binding domains; clearly, such a configuration is far from assembled, but cannot be differentiated from an assembled state with this order parameter. A related order parameter which circumvents this

[1] The 56-binding-domain-scaffold system differs slightly from that of Chapter 4 in that the sequences of the single-binding-domain staples were not modified to compress the assembly temperature range.

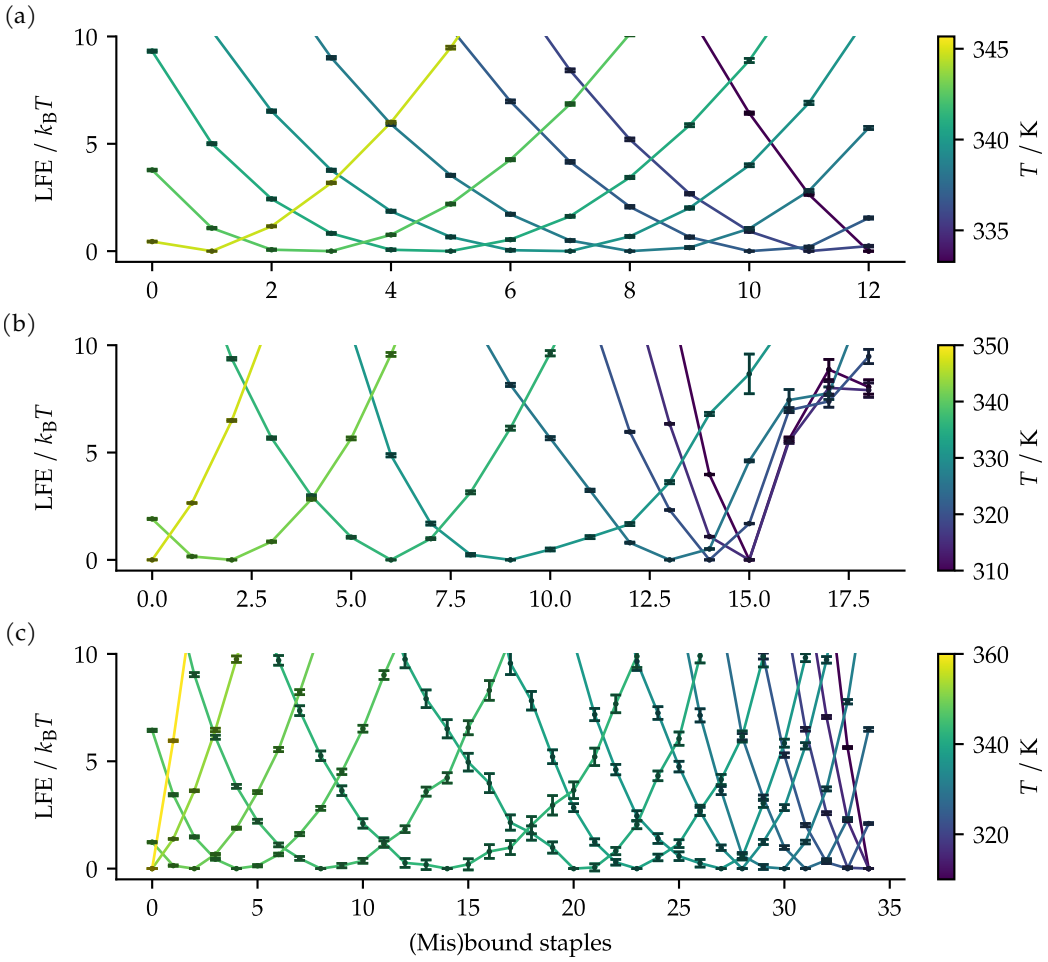

Figure 5.1: LFEs along the number of (mis)bound staples. (a) 24-binding-domain scaffold system. (b) 21-binding-domain scaffold system. (c) 56-binding-domain scaffold system. Only half of the number of temperatures used in the REMC simulations are plotted here for clarity. The error bars represent the standard error as calculated with the MBAR method, using data from three independent temperature REMC simulations.

issue is the number of fully bound staples, where a fully bound staple is one in which all of its binding domains are bound to the correct scaffold binding domains. In Figure 5.2 we have plotted the LFEs as a function of this order parameter. However for all three systems there is again no barrier, and overall the curves are very similar to those of the previous order parameter.

In order to achieve a higher resolution view of the binding of each staple, we can use the total number of bound domains as an order parameter. The associated LFEs plotted in Figure 5.3 are more jagged, and both the 21- and 24-binding-domain scaffold systems show small peaks at odd numbers near the bottom of the LFE curves at a given temperature. This is consistent with the second binding domain of a staple

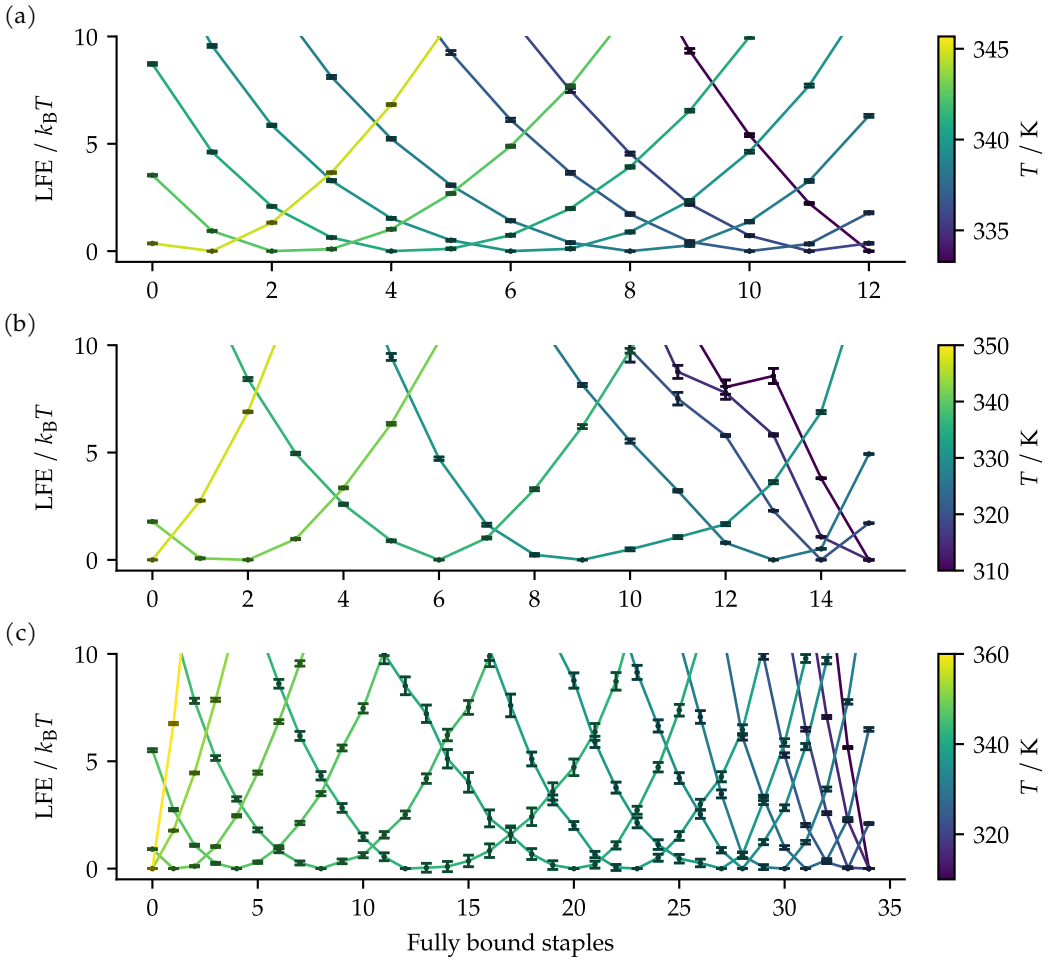

Figure 5.2: LFEs along the number of fully bound staples. (a) 24-binding-domain scaffold system. (b) 21-binding-domain scaffold system. (c) 56-binding-domain scaffold system. Only half of the number of temperatures used in the REMC simulations are plotted here for clarity. The error bars represent the standard error as calculated with the MBAR method, using data from three independent temperature REMC simulations.

having a lower entropic cost of binding than the first binding domain of a staple, and with a small easily surmountable barrier for staples that are near their individual melting points. That the jaggedness of the curves decreases in the 21- and even further with the 56-binding-domain scaffold systems is likely related to the presence of single-binding-domain staples, which could wash out the jaggedness caused by the two-binding-domain staples.

Yet another possibly relevant view of the assembly process may be gained by considering the number of stacked binding domain pairs, which can vary independently and substantially for given values of the previously considered order parameters. While less regular than the number of (mis)bound and fully bound staples, the

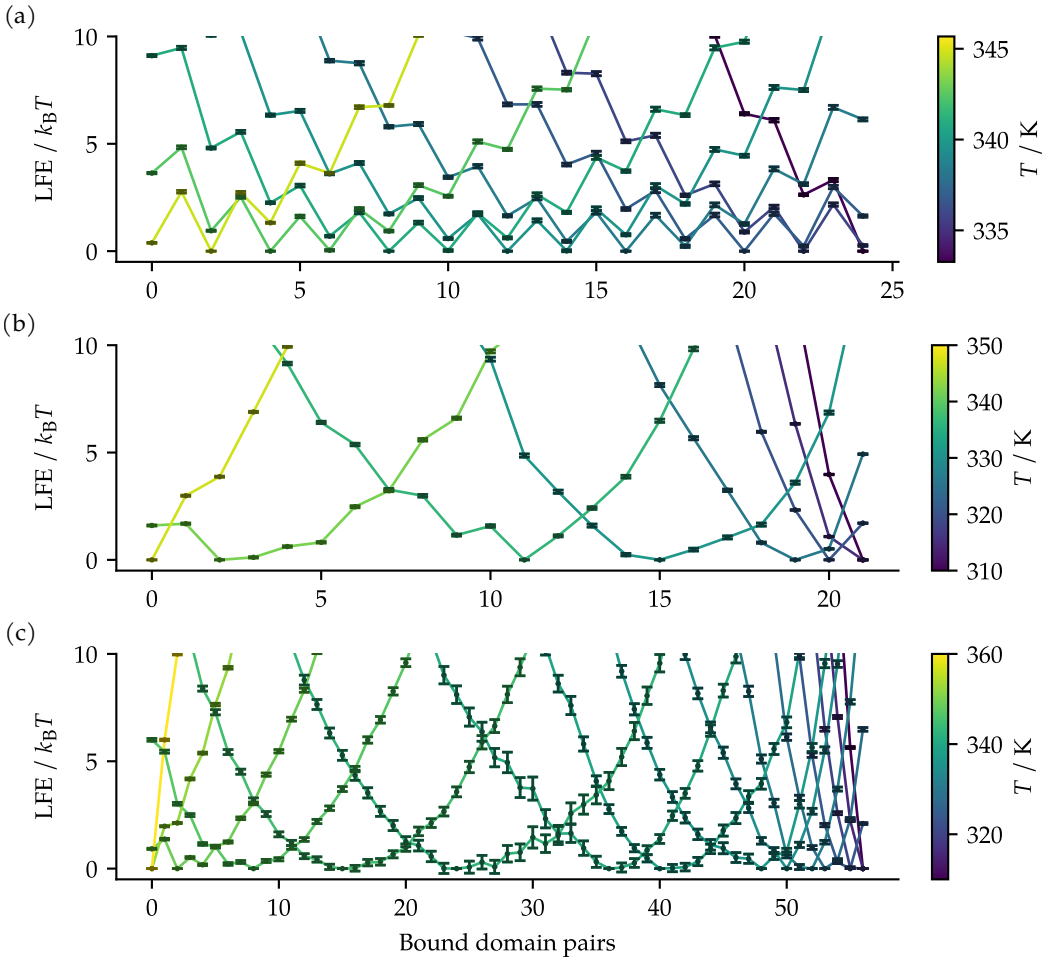

Figure 5.3: LFEs along the number of bound domains. (a) 24-binding-domain scaffold system. (b) 21-binding-domain scaffold system. (c) 56-binding-domain scaffold system. Only half of the number of temperatures used in the REMC simulations are plotted here for clarity. The error bars represent the standard error as calculated with the MBAR method, using data from three independent temperature REMC simulations.

LFEs (Figure 5.4) are also downhill to the favoured state, with only a few small exceptions. The most prominent exception is the peak at 50 stacked pairs in the 56-binding-domain scaffold system. This does not represent a barrier along relevant assembly pathways because in highly stacked states the number of stacked pairs can increase by more than one in a single move. This jaggedness is not seen at lower numbers of stacked pairs because there are more configurations in which the binding of a domain can increase the number of stacked pairs by just one.

In a simulation study using a model that was originally created to study DNA bricks, it was found that increasing the coordination number of the assembly units increased the barrier height [256]. Typical DNA bricks have a coordination number

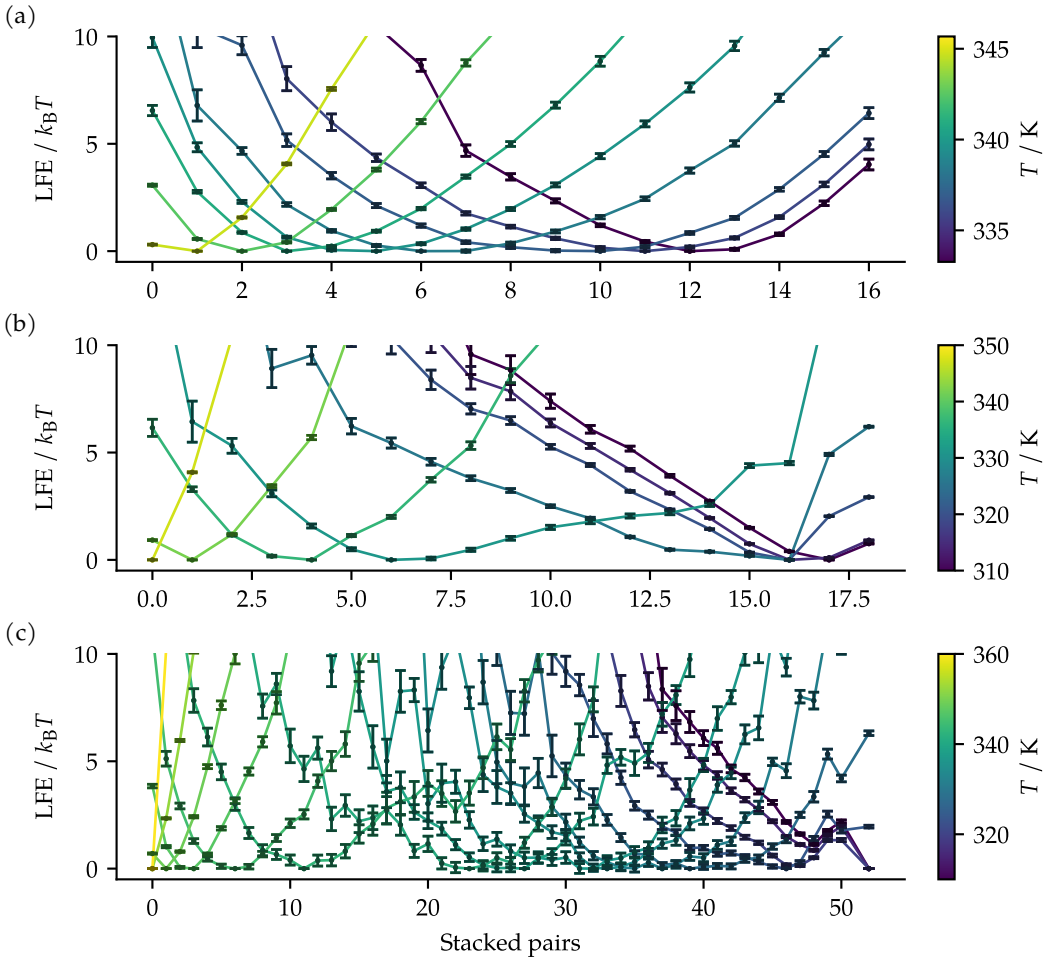

Figure 5.4: LFEs along the number of stacked pairs. (a) 24-binding-domain scaffold system. (b) 21-binding-domain scaffold system. (c) 56-binding-domain scaffold system. Only half of the number of temperatures used in the REMC simulations are plotted here for clarity. The error bars represent the standard error as calculated with the MBAR method, using data from three independent temperature REMC simulations.

of four, while the DNA origami designs simulated so far have at most a coordination number of two, so to test whether the same principle might apply in the context of DNA origamis, we increased the number of binding domains per staple as a way of increasing the coordination number. To test whether changing the number of binding domains per staple can lead to a barrier in a systematic manner, we designed a set of systems that have the maximum number of crossovers possible for a system with a given number of staple types and helices in the fully stacked assembled structure (Figure 5.5). In the fully stacked assembled state of these designs, the scaffold forms a series of rows, each of which consists of a single helix. At each column, a single staple crosses over all helices formed by the scaffold, and thus the

number of binding domains per staple corresponds to the number of rows in the design. We use two averaged values for the entropies and enthalpies of hybridization: one for bound states and one for misbound states. The average values for the bound states were calculated by using a value for each NN pair which was averaged over all 10 possible nucleotide pairings,[2]while the average values for the misbound pairings was taken to be those used in Chapter 4 for the 56-binding-domain scaffold system. Taking a simple mean over all misbound pairs may not give a good representation of misbinding, as a small number of much more favourable pairings may make a much larger contribution than the majority of pairings; however, we have chosen to start with this very simple model for simplicity, and use the results of the previous chapter, which find very little misbinding when using real sequences.

We began with systems that had 9 binding domains per row, and considered two-, three-, and four-row variants. Before examining the LFEs, it is instructive to examine the mean order parameter values across a range of temperatures (Figure 5.6). While we expect the systems to assemble over quite a narrow temperature range given that all binding domains have the same hybridization free energy, these systems have an impressively narrow range within which they transition, with both the three- and four-row systems going between entirely unbound and entirely bound within less than 1 K. For comparison, we have plotted in the same figure the curves for the number of (mis)bound and fully bound staples that result from assuming that all binding domains act independently of each other. We have recentred them to the simulation curves at the temperature at which the order parameter is halfway between zero and its value in the assembled state (henceforth referred to as the halfway temperature) in order to give a better comparison of the differences in transition temperature ranges. The much narrower range of the simulation curves is a clear sign of cooperativity in the system.

In Figure 5.7 the temperature curves and halfway temperatures are plotted for the staple states of each staple type for each of the three systems. In all systems, the staple with the lowest halfway temperature is always an edge staple, which is consistent with both that these staples additionally contain scaffold crossover(s) and that they are only able to form half as many stacks as the other staples. Outside of these edge staples, the two-row system has a small but significant shift in halfway temperatures, going from higher on the right to lower on the left, with the edge staple on the right side having an intermediate halfway temperature. This is consistent with an

[2]A small error was discovered in the calculation of this average after running all simulations which resulted in a slightly more favourable average value being used. This is not expected to have a qualitative impact on the results presented here.

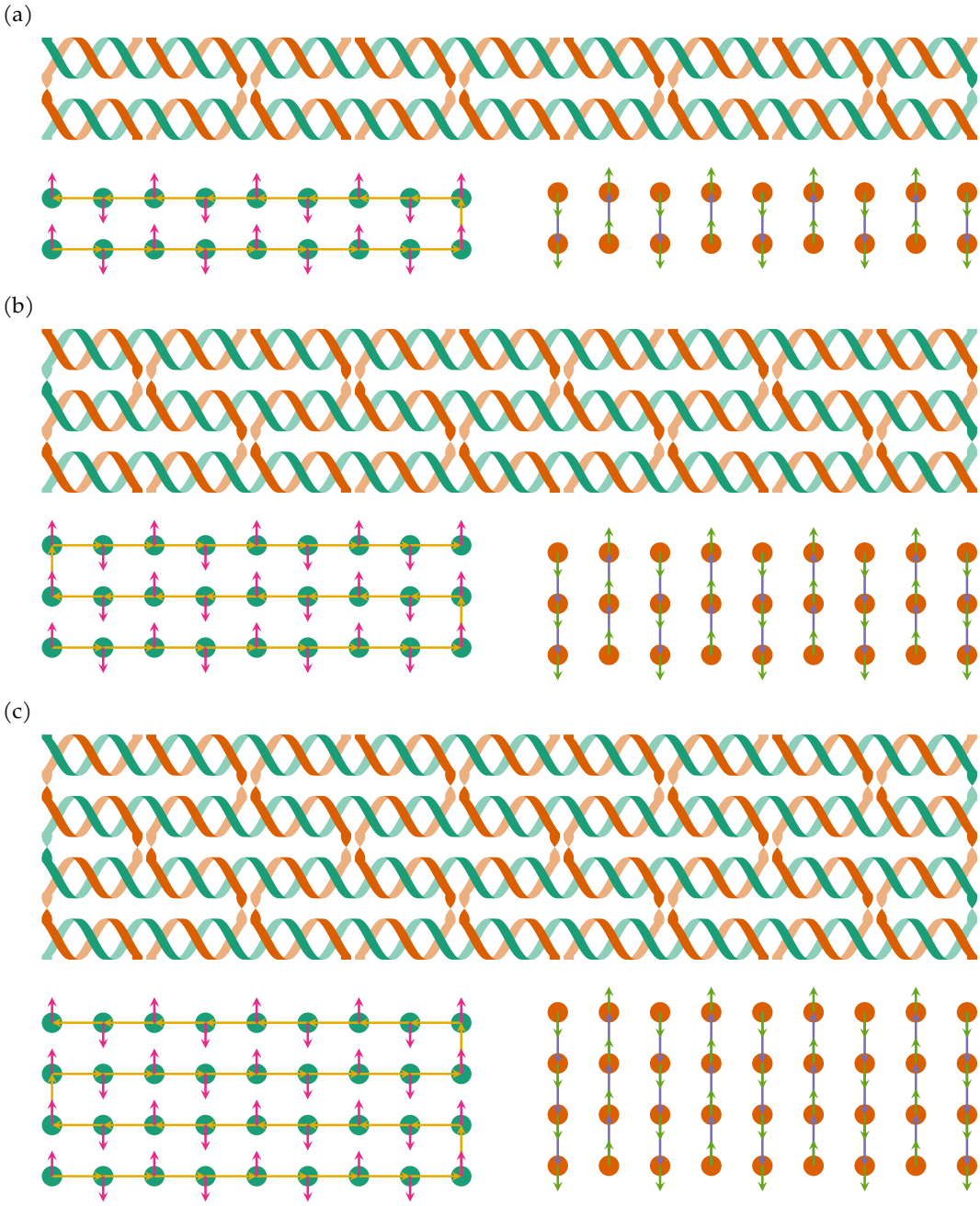

Figure 5.5: Schematic representations of the maximum crossover systems. (a) Two-row system. (b) Three-row system. (c) Four-row system. For each system, the helical cartoon representation of the system in a fully stacked assembled configuration is given above, and the lattice model representation of the system in the same configuration is given below. The scaffold (left) and staples (right) have been drawn separately for clarity. A full legend for all diagram elements is provided in Figure 2.1.

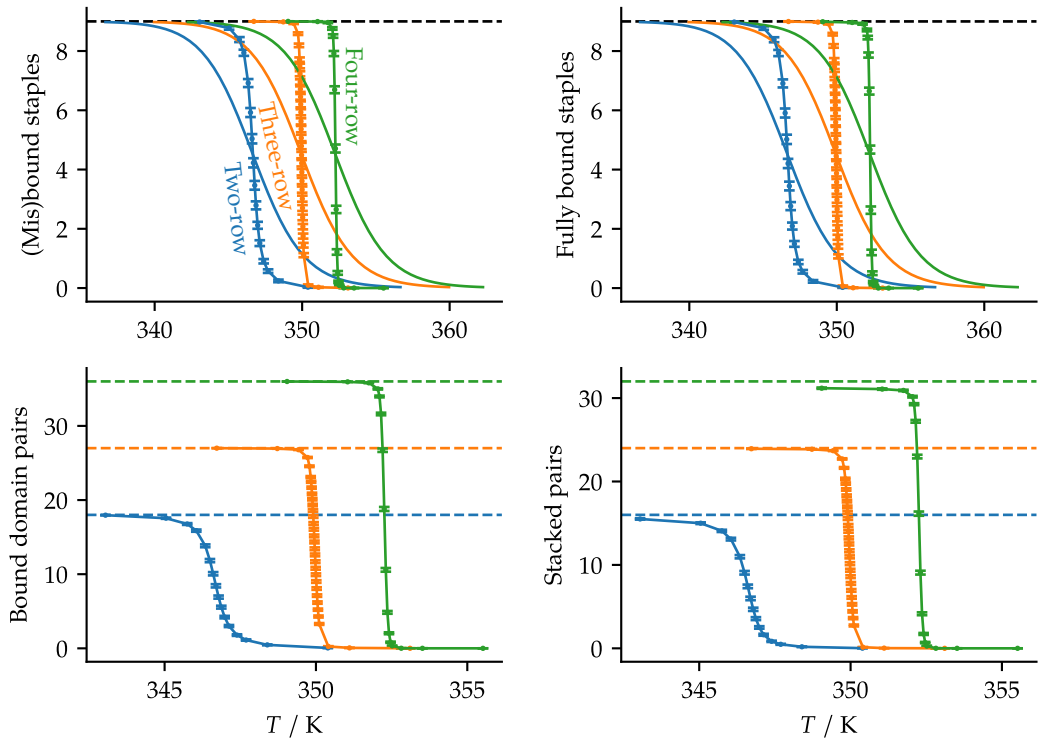

Figure 5.6: Mean order parameters plotted across a range of temperatures for the two-, three-, and four-row systems. The dashed lines correspond to the expected order parameters in the fully stacked assembled state. The solid lines without markers present in the top two plots represent the shifted curves that would result if all domains were isolated from each other. The error bars represent the standard error as calculated with the MBAR method, using data from three independent temperature REMC simulations.

increasing entropic cost as the loop size increases from right to left. However, in the three- and four-row systems, outside of the edge staples, the halfway temperatures are nearly identical. This seems reasonable for the three-row system given that it has two loops that increase in size in opposing directions, the effects of which should cancel each other. In the four-row system, such uniformity is less expected. There are two identical loops that increase in size from right to left and one loop that increases in size from right to left, which should give an overall entropic cost of loop formation that increases from right to left. This suggests that there is an increase in cooperativity relative to the two-row system.

Consistent with the previous simulations, the LFEs along the number of (mis)-bound staples and the number of fully bound staples (Figure 5.8) for the two-row system are downhill to the favoured state at all temperatures, although now the favoured state is either a fully bound or a fully misbound state. By contrast, the three- and four-row systems have a clear barrier to assembly that is apparent along

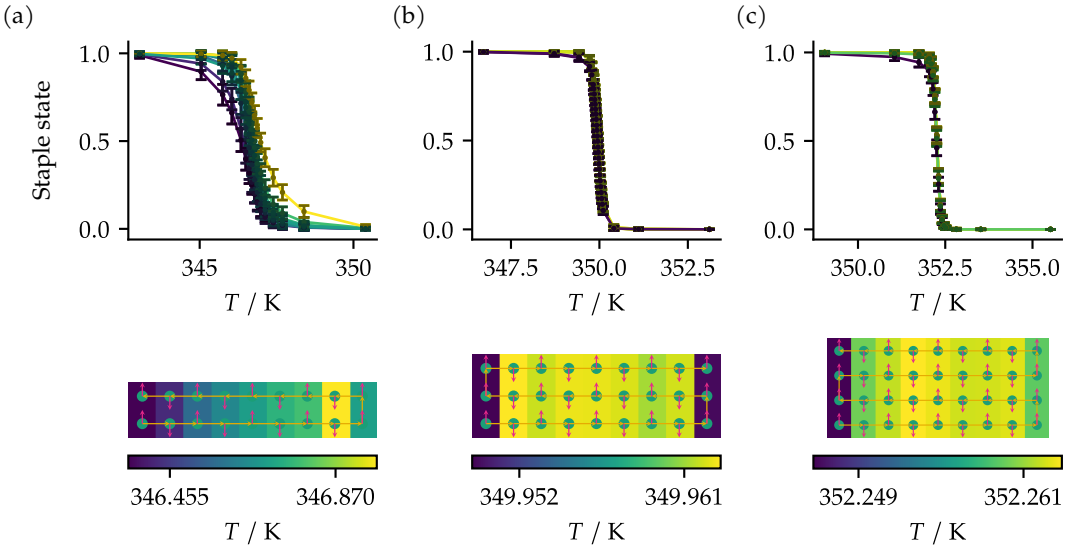

Figure 5.7: Mean staple states plotted across a range of temperatures (above) and mean staple state halfway temperatures plotted as a heat map (below). The mean staple state halfway temperatures are calculated from the curves by simple interpolation and plotted on a grid corresponding to the staple locations in the fully stacked assemble state. For reference, the scaffold in a fully stacked assembled configuration has been superimposed on the halfway temperature heat map; see Figure 5.5 for the corresponding staple configurations. (a) Two-row system. (b) Three-row system. (c) Four-row system. The error bars represent the standard error as calculated with the MBAR method, using data from three independent temperature REMC simulations.

both of these order parameters. If we define the staple melting temperature as the point at which the local minima on either side of the barrier of the number of fully bound staples are equal, we can use the MBAR method [246] to reweight the configurations for an arbitrary thermodynamic state and thus iteratively solve for this temperature. This assumes that there is good overlap in the distribution at this state and states that have been sampled, which should be the case here, given that the REMC simulations span a range of temperatures or stacking energies. The melting temperatures calculated using the number of (mis)bound staples are very similar to those calculated using the number of fully bound staples.

The sharp response to temperature that these systems displayed provided an additional challenge to the simulations in the selection of the temperatures. Selection of a temperature range that has good coverage of the transition region by hand becomes infeasible, and so an automated iterative approach to selecting temperatures was employed, as described in Section 5.2. While this was sufficient for the above analysis, the LFEs proved to be more sensitive to insufficient sampling, which seemed

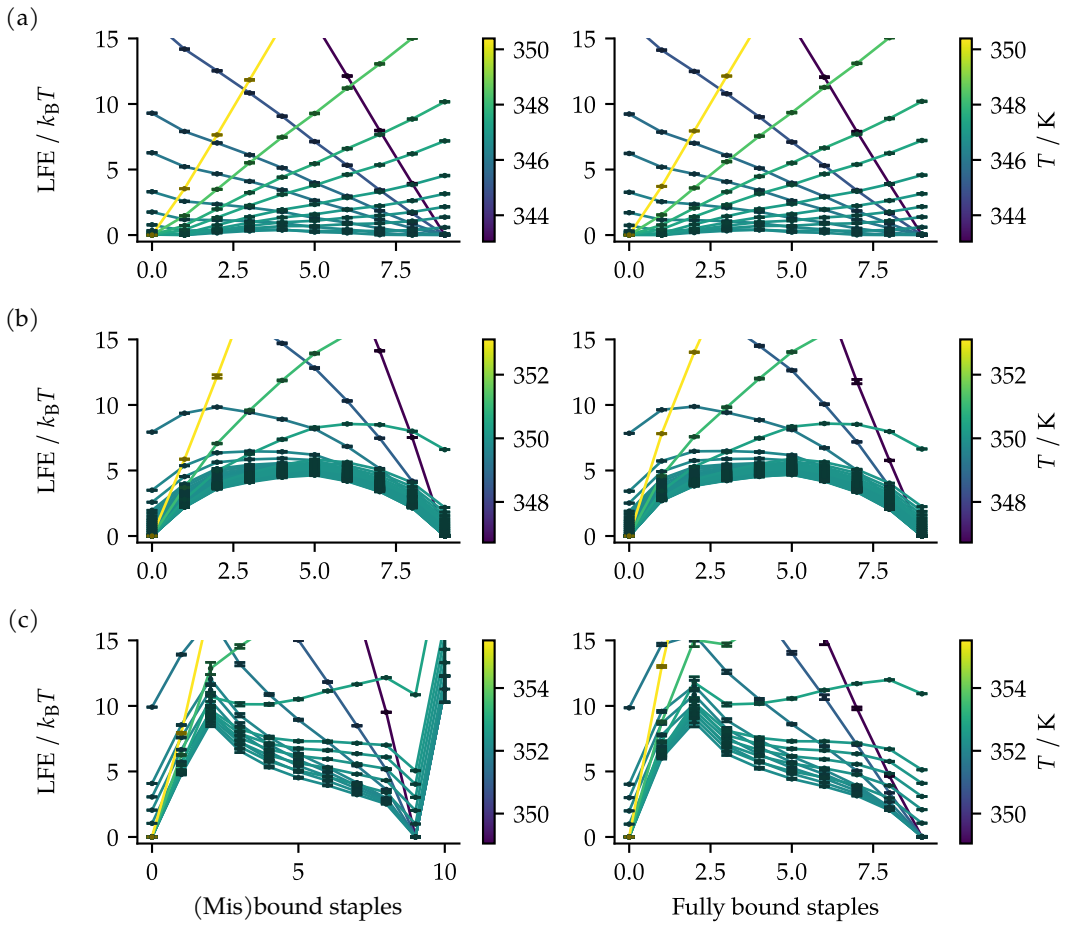

Figure 5.8: LFEs along the number of (mis)bound and the number of fully bound staples. (a) Two-row system. (b) Three-row system. (c) Four-row system. The error bars represent the standard error as calculated with the MBAR method, using data from three independent temperature REMC simulations.

in particular to stem from sampling of states with different stacked pair combinations, something which was touched upon in Chapter 4.

We began our search for a better sampling procedure by again considering 2D REMC with both the temperature and a stacking multiplier as exchange variables, but found that the sharp transitions presented by these systems made this an inviable approach. This is because the transition is not only sensitive to temperature, but also to the stacking multiplier. Even a small change in the stacking multiplier can lead to a sufficiently large change in the average number of bound staples that the swap probabilities become very small. Further, without using an excessive number of replicas, the temperature range will either be optimized for a single stacking multiplier and give poor coverage along the others, or provide mediocre coverage for all stacking multipliers.

The approach which seemed to provide the best sampling was to use only the stacking multiplier as the exchange variable. Here, the idea is to use temperature REMC simulations to estimate the melting temperature, which is used as the initial temperature for the stacking multiplier REMC simulations. The temperature is then iteratively updated, with the temperature for each new run being the estimated melting temperature from the current run. The melting temperatures estimated with this approach were higher than those estimated from the temperature REMC simulations, which is consistent with a broader set of configurations in near-assembled and assembled states, which would provide entropic stabilization. This broader set of configurations may not differ in the examined order parameters, but can contain different combinations of locally stacked pairs.

The LFEs calculated with this approach are plotted in Figure 5.9. The barrier for the three-row system occurs nearly halfway to the fully assembled state, at 4 (mis)bound or fully bound staples, with a magnitude of between 5 to 6 $k_B T$. The two order parameters are nearly identical, which suggests that staples bind fully if they are bound at all. In contrast to the three-row system, the four-row system peaks at two (mis)bound or fully bound staples with a magnitude of around 8 $k_B T$, and the LFEs are qualitatively different between these two order parameters. Along the number of (mis)bound staples, the LFE peaks sharply at two, while along the number of fully bound staples, it has nearly peaked by one and only increases marginally at two. In contrast to the three-row system, this suggests that there is a tendency for staples not to bind fully in the four-row system, at least when only a small number of staples are present.

We can examine the behaviour of staple binding in more detail by again examining the LFEs along the number of bound domain pairs. We focus our investigation on the LFEs at the melting temperature defined above (Figure 5.9), using the melting temperature calculated with the number of fully bound staples. In both the three- and four-row systems, there is clearly a barrier reminiscent of the one observed along the bound staple order parameters. In all three systems, there are again peaks at regular intervals, occurring with a frequency equal to the number of binding domains in the staple. This is consistent with a barrier to the initial binding of a staple to the scaffold, followed by favourable binding of the remaining binding domains of the staple.

However, in the three- and four-row systems, the pattern is broken at lower values of the order parameter. In the three-row system, the LFE is higher at three bound domains than at two, which is consistent with the binding of the third staple

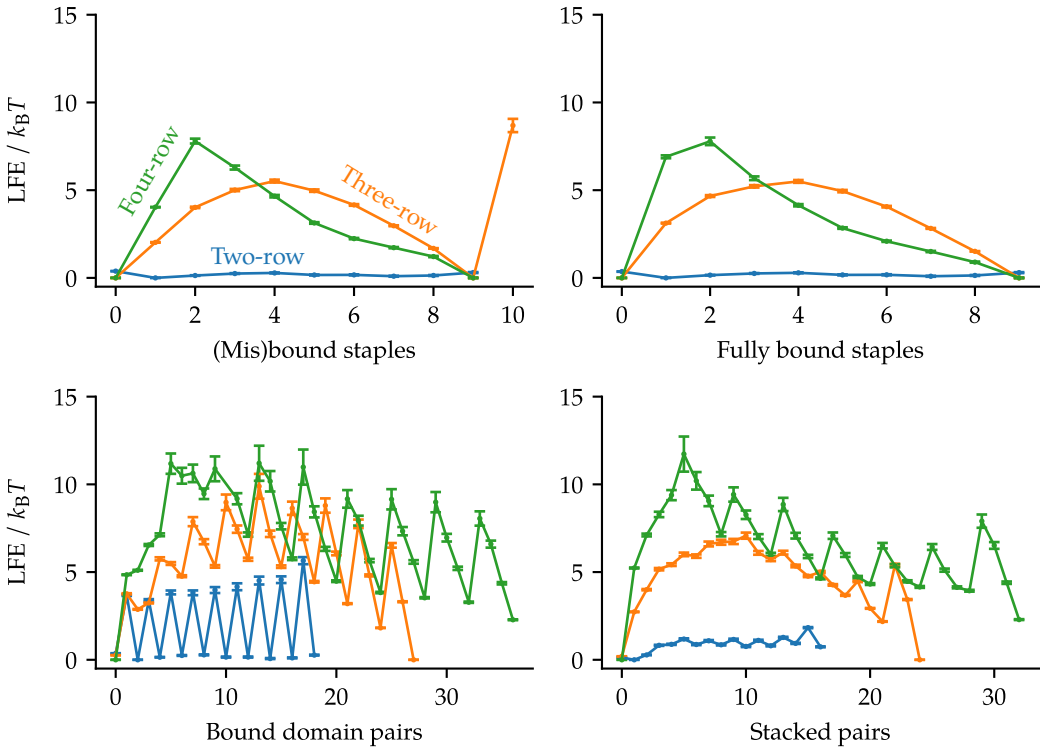

Figure 5.9: LFEs at melting temperatures for the two-, three-, and four-row systems. The error bars represent the standard error as calculated with the MBAR method, using data from three independent stacking multiplier REMC simulations.

domain closing a longer loop than the second binding domain. In the binding of subsequent staples, the entropic cost of closing the longer loop is reduced by the binding of previous staples in a cooperative manner. In the four-row system, the LFE increases until five domains are bound. If the first staple to bind is that with the lowest entropic cost, then there are now two smaller loops and one larger loop to be closed to fully bind the staple. That the closing of the smaller loops is now outweighing the gain in hybridization free energy could be explained by the higher melting temperature of the four-row system, which would increase the entropic cost.

The LFEs along the number of stacked pairs also show an overall barrier in addition to some smaller, regular peaks. The smaller peaks along this order parameter become more exaggerated as the number of bound domains increases. These peaks cannot be entirely related to the peaks seen along the number of bound domains; in a fully stacked structure, unbinding of one domain would lead to the loss of two stacked pairs. To explain these peaks, consider the final peak of the four-row system. When the number of stacked pairs is four less than the maximum possible number, a kink can form in the structure without any domains being unbound. This means

that there are many more possible structures with a more favourable energy at this number of stacked binding domain pairs than at one, two, or three fewer stacked binding domain pairs than in the fully stacked state. However, as discussed above, the number of stacked pairs can change by more than one in a single step, so these barriers do not necessarily have to be crossed in order to reach the fully stacked state. An analogous explanation holds for the other systems.

To further test the validity of these results, we ran US simulations of the three-row system. While we could have used a similar approach to the stacking multiplier REMC for iteratively updating the simulation temperature, the increased walltime required for running adaptive multi-window US made this impractical. Instead, we use the melting temperature estimated from the temperature REMC simulations. We applied the bias along the number of (mis)bound staples order parameter. We did find a barrier (Figure 5.10), confirming the qualitative result found above, but the LFEs indicate that the simulation is being run above the melting temperature, yet the simulation temperature is below the melting temperature found in the stacking multiplier REMC. Because we expect the sampling of near and fully assembled states to be more difficult than unassembled states, and because poor sampling of near and fully assembled states would tend to lead to underestimating their stability, this result implies that the US simulation is not equilibrated.

In an attempt to improve sampling convergence, we instead used the number of bound domains as the order parameter along which to apply the bias. While this improved sampling along this order parameter, it did not result in better agreement with respect to the melting temperature. We also considered using the number of stacked pairs as the order parameter on which we apply the bias, but it seems unlikely this would provide much additional benefit, as the issue seems to be sampling different states with the same number of stacked pairs, rather than sampling a range of states across the stacked pairs order parameter. For US to be effective here, it would likely need to be combined with REMC at each window, or between windows. We decided to focus on the REMC methods we developed for the remainder of the study.

Turning back from considerations of sampling, we can further investigate the pathways taken by calculating expectations of individual staple states for set values of the number of fully bound staples at the melting temperature. In the case of the two-row system, the staples tend to bind in the order the halfway temperature predicts: from shorter to longer loops formed with the scaffold upon fully staple binding, and edge staples last (Figure 5.11). The three-row system has diffuse

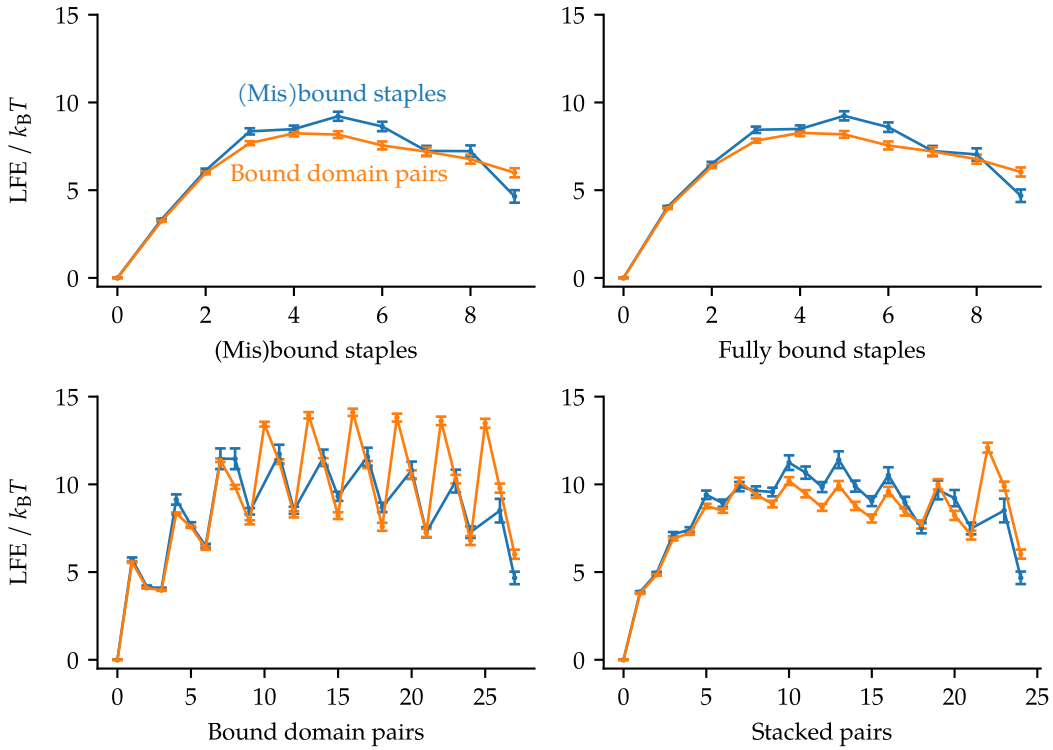

Figure 5.10: LFEs calculated from US simulations. The error bars represent the standard error as calculated with the MBAR method, using data from a single US simulation.

averages across the staple types until five fully bound staples, which is one more than at the barrier peak, at which point the central staple is nearly always bound (Figure 5.11). As the number of fully bound staples increases beyond five, the staples become more bound moving out from the centre. Unlike the two-row system, this is not clear from the temperatures of the staples, as they are nearly identical across the staple types. The behaviour of the four-row system is similar to the two-row system in that staples first bind on the right where the loop closure cost is lower and progress towards positions that form more costly loops. However, unlike the two-row system, the staples tend to bind fully after the first three have bound, which also contrasts with the nearly indistinguishable temperatures of each staple type.

Cooperative behaviour of staples and binding domains can occur via three routes: closing of scaffold loops, stacking with other binding domains adjacent in the same helix, and, for cooperativity within a single staple, binding one of the staple domains to the scaffold. The last route is demonstrated by all systems, and while it can lead to a barrier in the initial binding of each staple, it cannot explain the barrier we have observed along the order parameters based on the number of staples in the system. The first route, the closing of loops, could plausibly lead to a nucleation barrier.

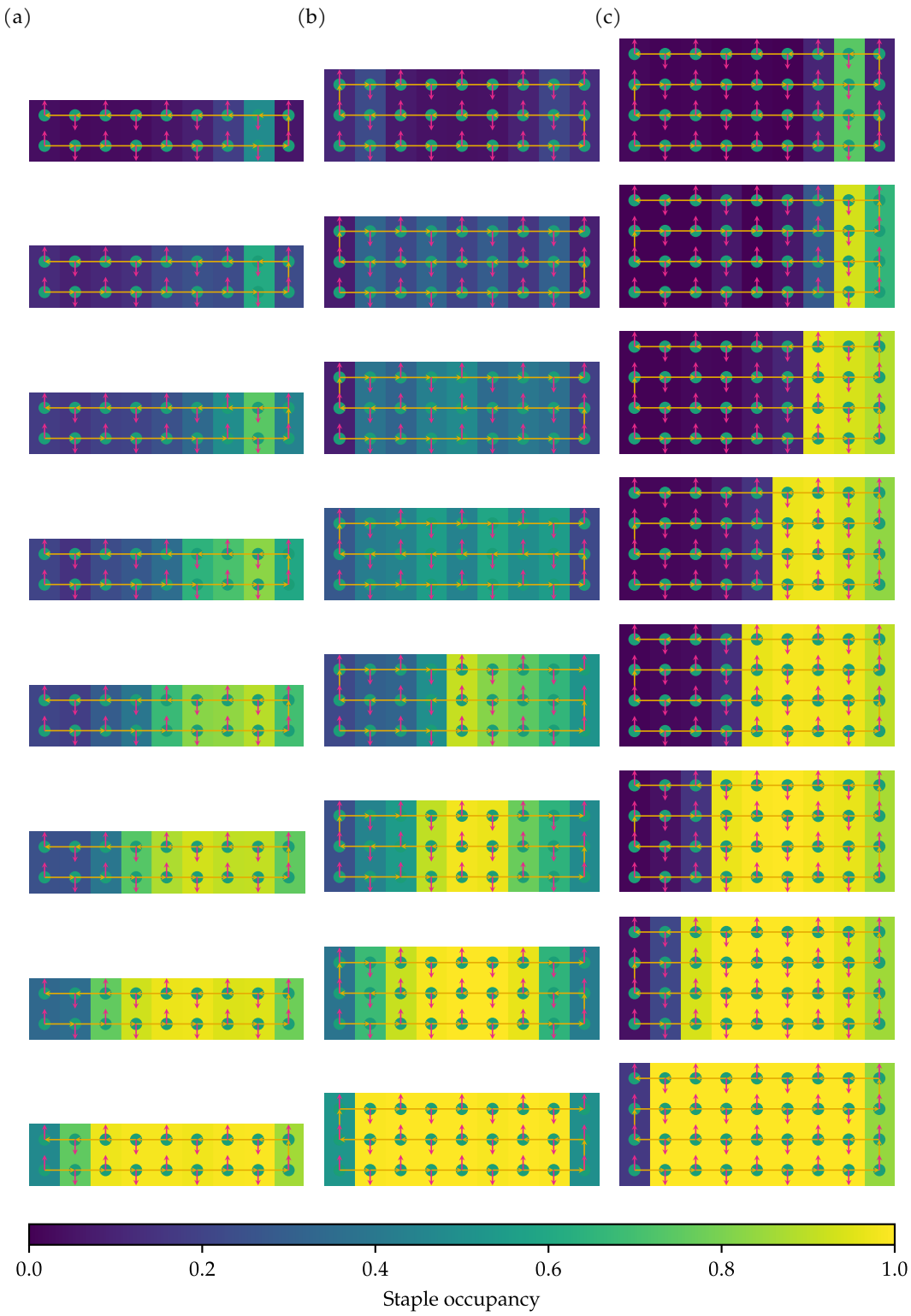

Figure 5.11: Expectation values for staple states at slices along the number of fully bound staples. The number of fully bound staples of the slices begins at 1 and increases from top to bottom to 8. (a) Two-row system. (b) Three-row system. (c) Four-row system.

If the cost of forming a large loop was sufficiently high such that multiple staples were required to bind to make loop closure thermodynamically favourable, then there would be a critical number of staples after which further binding of staples involving this loop could be downhill in free energy. However, this would require the staples that initially close the loop to bind substantially more strongly than staples that could alternatively close the loop by in a zip-like fashion, taking a pathway where the shortest loops are closed first, for this assembly pathway to be viable. The sequence design and scaffold routing requirements for such a barrier to be present seem unlikely to be common in origami designs. Further, this cannot be the case here, as we are using uniform hybridization entropies and enthalpies. This type of cooperativity likely does contribute to the sharp transitions observed through the zip-like mechanism, however.

We therefore focused our investigation on the stacking of binding domains adjacent along the same helix. We began by setting the stacking energy to zero and running simulations of the three-row system. The LFEs reveal that the barrier involving the number of bound staples is effectively no longer present, being less than 1 $k_B T$ (Figure 5.12). Unlike the two-row system, the LFEs are not flat; at the melting temperature, there is a relatively flat region at lower values of the order parameter before a substantial increase. The LFEs look similar at one-quarter of the standard stacking energy, while at half the standard stacking energy the curve becomes effectively flat; not until the stacking energy reaches three-quarters of the reference value does the barrier become apparent again.

If the barrier is controlled by the stacking energy, then increasing the number of binding domains per staple may cause the appearance of a barrier simply due to an increase in the number of stacking interactions possible per staple. A further test of this would be to see if increasing the stacking energy can lead to a barrier in the two-row system. We again ran simulations for a range of stacking energies, this time from the original stacking energy to twice its value. As can be see in Figure 5.13, while negligible, a small barrier appears even at the first increment of the stacking energy (1.25 times the original value), and then increases monotonically.

When a fluctuation occurs where several staples bind concurrently in such a way that they can stack with each other, the energetic gain is sufficient to overcome the entropic cost of binding at a temperature that is higher than it would be for a given staple in isolation. The stronger the stacking per staple, whether by higher stacking at each domain or by having more domains to stack per staple, the higher the temperature at which a cluster of staples is able to bind relative to the staples

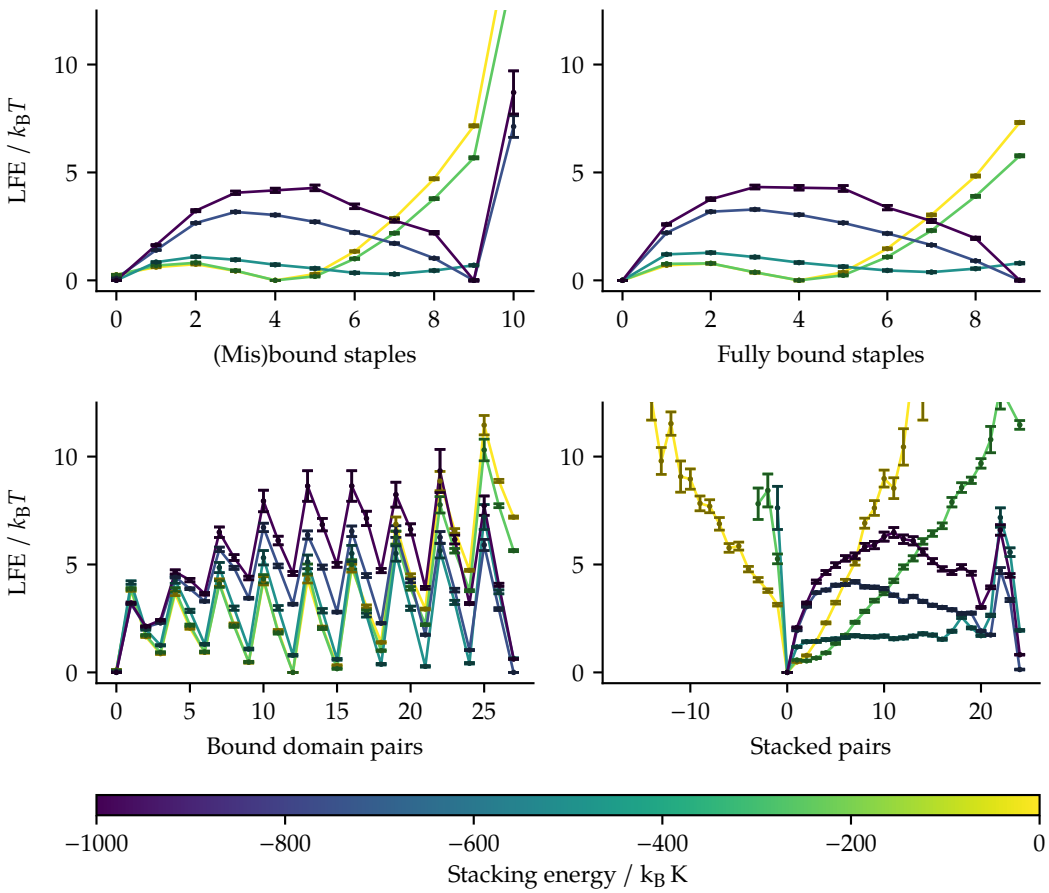

Figure 5.12: LFEs at melting temperatures for the three-row system for a range of stacking energies. The error bars represent the standard error as calculated with the MBAR method; in the case of the simulations with a stacking energy of 0, −250, and −500 k_B K, data from three independent temperature REMC simulations were used, while in the case of the simulations with a stacking energy of −750 and −1000 k_B K, data from three independent stacking multiplier REMC simulations were used.

in isolation. The per-domain stacking could be controlled by changing the salt concentrations, by modifying the sequence pairings that occur at breakpoints, or even by using modified nucleobases which have different stacking interactions.

This increased temperature difference also leads to a higher barrier, as the fluctuation needed for a given staple to bind has a higher entropic cost. While the barrier in the number of total bound staples is most clear along either the number of (mis)bound or the number of fully bound staples, the total barrier, which includes the initial cost of binding a staple to the scaffold, is more accurately seen along the number of bound domain pairs. Even considering the highest barrier along the number of bound domains, it is still of a magnitude that is surmountable in a fraction of the time required in typical self-assembly protocols for DNA origami.

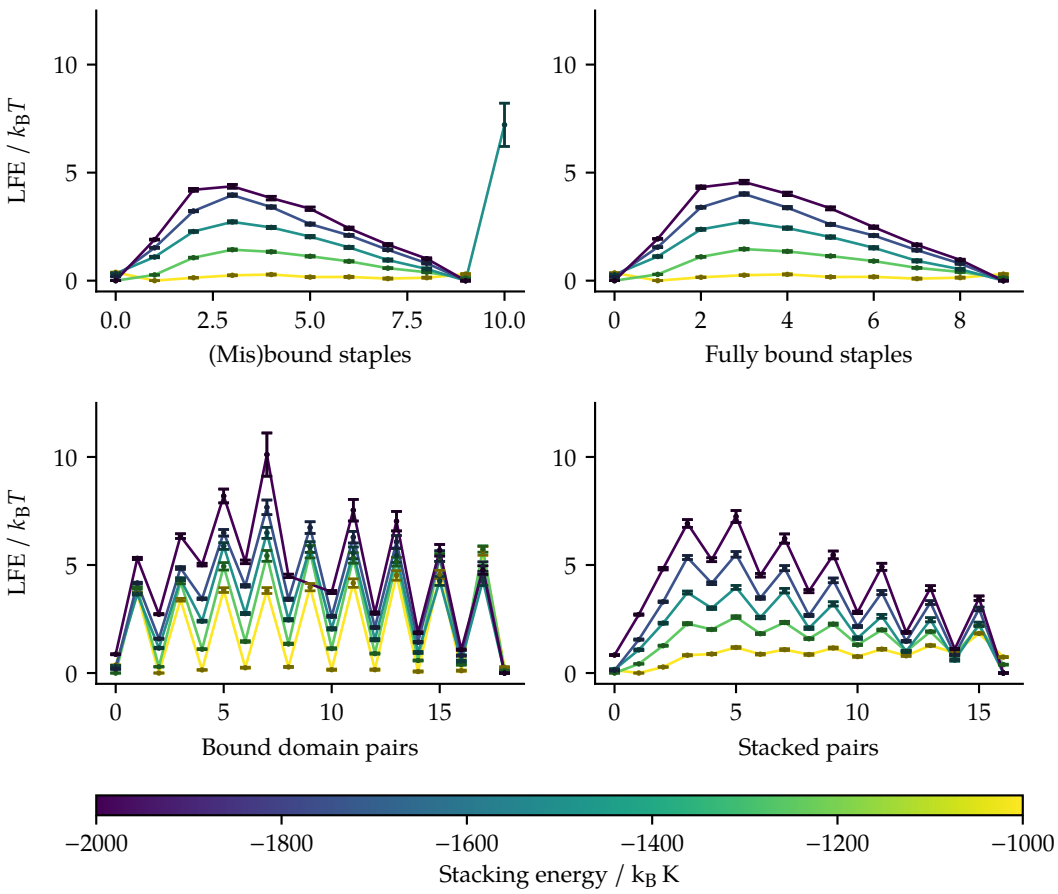

Figure 5.13: LFEs at melting temperatures for the two-row maximum crossover system for a range of stacking energies. The error bars represent the standard error as calculated with the MBAR method; in the case of the simulations with a stacking energy of −1000 and −1250 k_B K, data from three independent temperature REMC simulations were used, while in the case of the simulations with a stacking energy of −1500, −1750, and −2000 k_B K, data from three independent stacking multiplier REMC simulations were used.

From these considerations, one conclusion we may arrive at is that increased heterogeneity in the individual staple hybridization free energies may lead to lower barriers. With sufficiently wide spacings in the melting temperatures, the stacking energy will be insufficient to allow for multiple staples binding such that they stack with each other to overcome the entropic cost of binding. As a first step to testing whether these barriers exist with a non-uniform hybridization potential, we simulated the 21-binding-domain scaffold system again, but with the stacking energy doubled. As can be seen in Figure 5.14, a local maximum does appear in the LFEs, although it is too small to be called a barrier. Having four binding domains per staple is common, so the 21-binding-domain scaffold system, and all the systems we

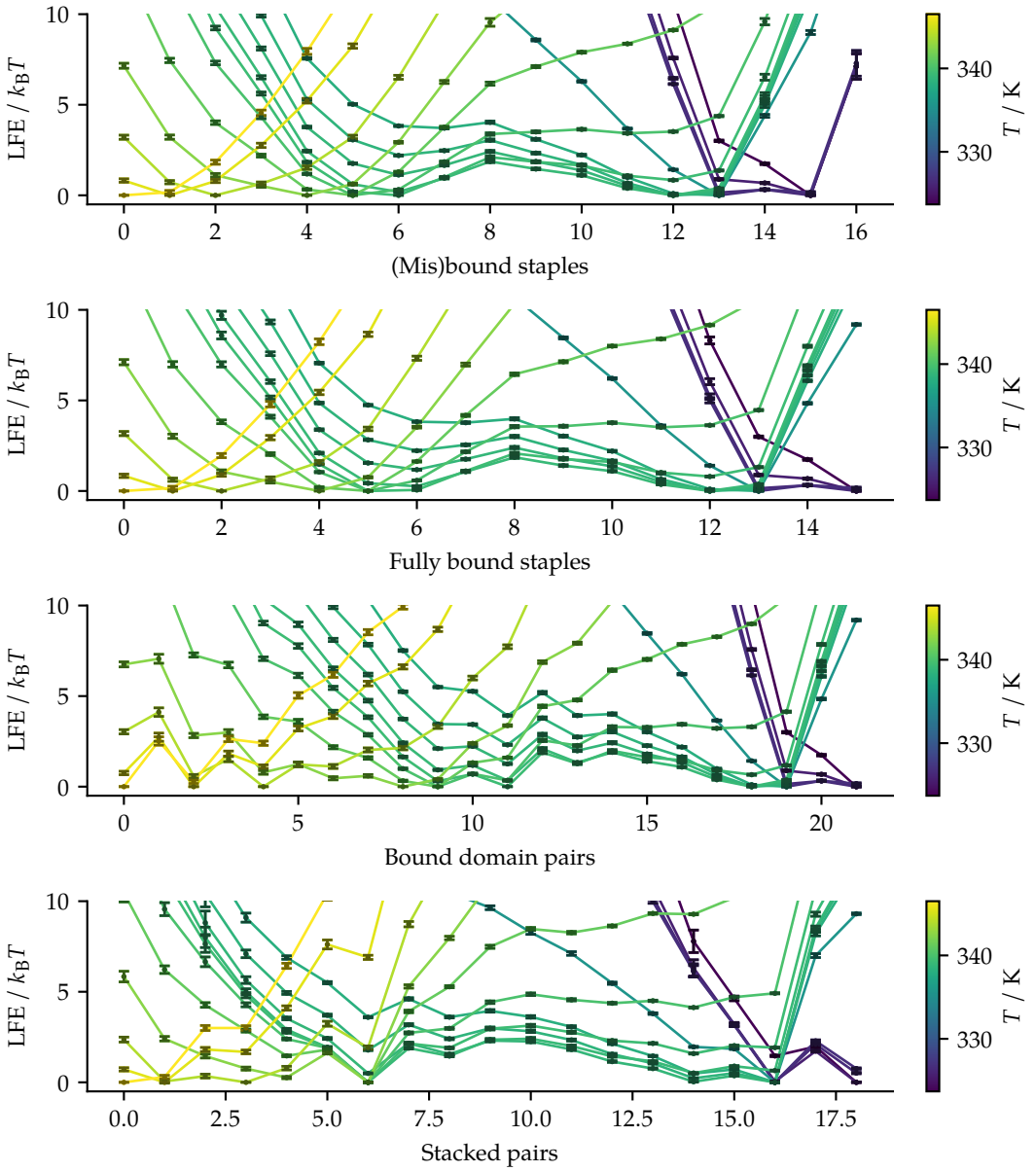

Figure 5.14: LFEs calculated for the 21-binding-domain scaffold system with double the default stacking energy ($-2000\,k_B\,K$). The error bars represent the standard error as calculated with the MBAR method, using data from three independent temperature REMC simulations.

simulate here with NN hybridization free energies, are not a representative sample of DNA origami designs. Perhaps the next step in systematically examining the link between staple binding heterogeneity and nucleation barriers would be to draw a set of hybridization free energies from a normal distrubution to create multiple staple sets, and calculate the average barrier height as a function of the variance of the distribution.

Because our selection of 9 binding domains per row was arbitrary, we ran simulations of the three-row system with 3, 5, and 7 binding domains per row. The LFEs have been plotted in Figure 5.15. The barrier in the number of (mis)bound and fully bound staples increases with the number of binding domains per row, while the location of the barrier increases from 2 with 3 binding domains per row, to 3 with 5 binding domains per row, and finally to 4 with 7 binding domains per row, which is also the location of the peak with 9 binding domains per row. In contrast to the monotonically increasing barrier height in staple-based order parameters, the barrier height along the number of bound domain pairs is quite similar between the systems, especially between 7 and 9 binding domains per row, with a greater difference appearing in the troughs. The melting temperatures of the 5- and 7-binding-domains-per-row systems were nearly identical, differing by only 0.01 K, while the 9-binding-domains-per-row system's melting temperature was approximately 1.5 K higher. Because these systems were smaller, we assumed that the iterative stacking multiplier REMC we used above would not be necessary to achieve good sampling, so the difference in melting temperature could be a sampling issue. It seems possible that once a certain number of binding domains are present per row, a 'bulk phase' is entered in which the nucleation barrier remains largely the same, where there are enough staple types to form a critical 'nucleus', and effects of the edge staples become insignificant. However, we would need to run further simulations on systems with more binding domains per row to confirm this.

One potential criticism of our findings is that they may strongly depend on the details of the model. As a final test of the validity of our results, we use a different representation of the three-row system, with 8 nt binding domain lengths, instead of 16 nt, which requires the use of the three-quarter-turn binding domain, rather than the half-turn binding domain (see Chapter 2 for the differences between these domain types). The first point of interest is that this change in representation of the system results in a very large shift of the expectation values of the order parameters towards lower temperatures (Figure 5.16). Such a shift can be explained as the 8 nt representation having a larger entropic cost to assembly than the 16 nt, which is consistent with the 8 nt having more states available to it in an unassembled state. In addition to including the $k_B T \ln 6$ for ϵ_u, a mean field correction would be needed to better match experimental curves, as discussed in Section 2.2.1. However, if we consider the LFEs (Figure 5.17), they agree qualitatively with the 16 nt representation simulations. The two-row system is effectively downhill to either the fully assembled state or the unassembled state, while the three-row system has a small barrier on

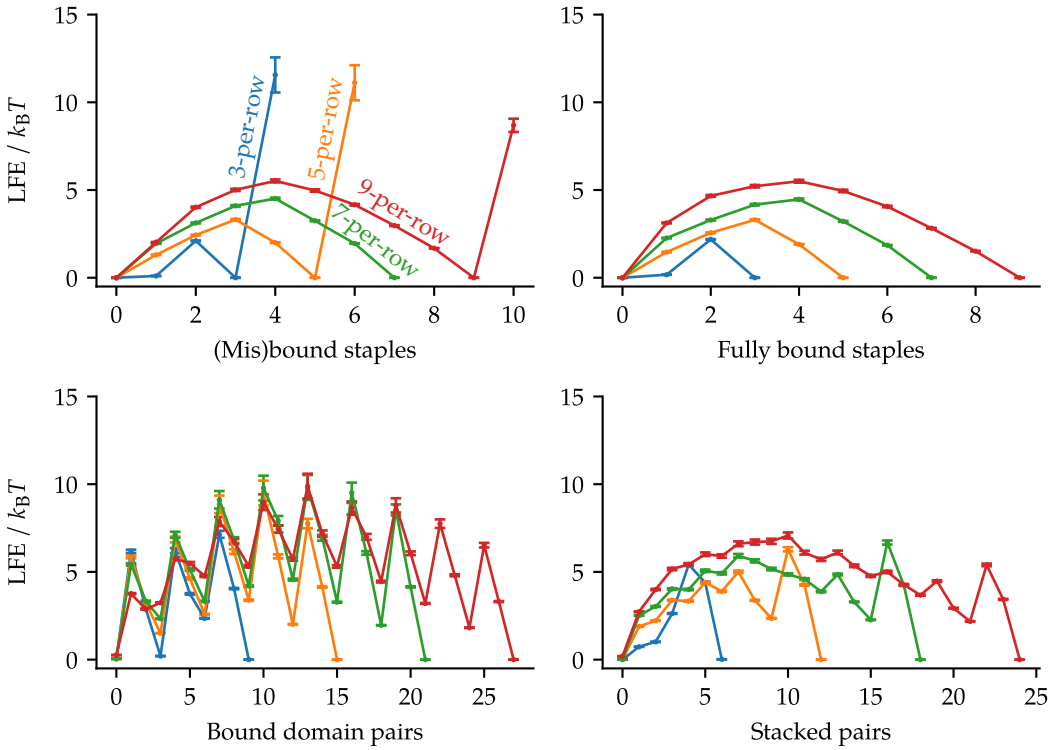

Figure 5.15: LFEs at melting temperatures for varying number of binding domains per row in the three-row system. The error bars represent the standard error as calculated with the MBAR method, using data from three independent temperature REMC simulations (the 9-binding-domains-per-row system uses the previously presented data from three independent stacking multiplier REMC simulations).

the order of 5 $k_B T$.

5.4 Conclusions

We have used the model developed in Chapter 3 to investigate whether nucleation barriers exist and are relevant to the self-assembly of DNA origami. We have found that nucleation barriers are dependent on the co-axial stacking of staples to each other when adjacent on the same helix. Changing either the strength of the individual stacking interactions or the number of stacking interactions per staple (by increasing the number of binding domains per staple), we were able to control, and even completely remove, the barrier. With the standard stacking energy, the size of the barrier is not overly large relative to thermal fluctuations, and is thus easily surmountable.

The small barriers, and especially the total lack of a staple binding nucleation barrier in systems with staples that have only two binding domains, may be useful

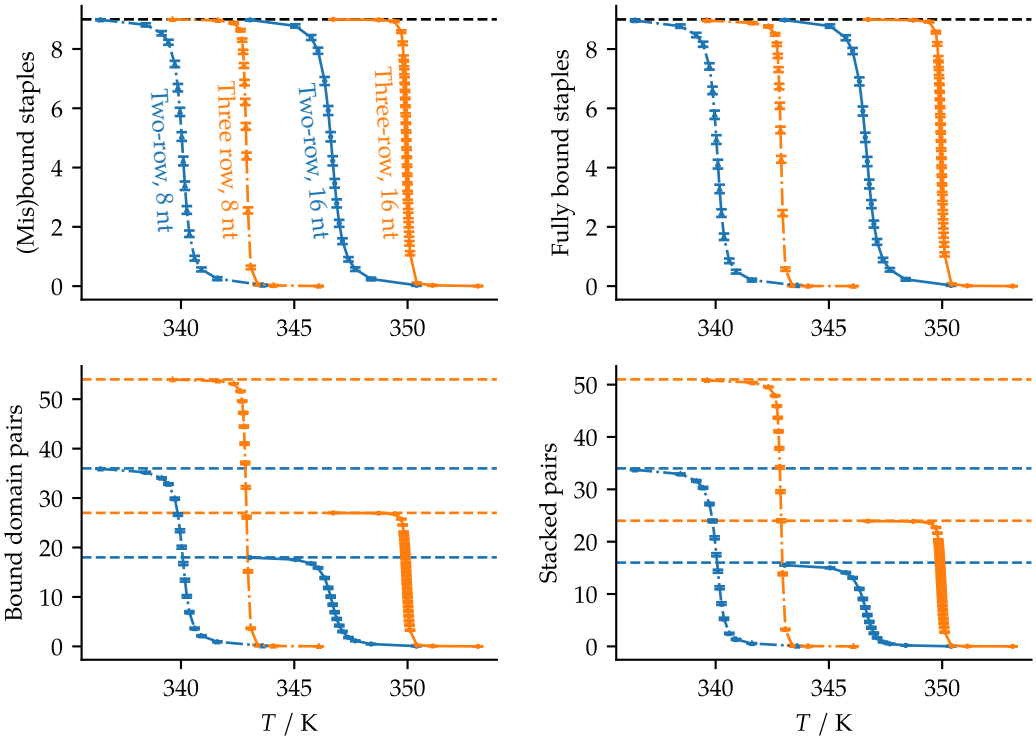

Figure 5.16: Mean order parameters plotted across a range of temperatures for the two- and three-row systems with three-quarter-turn and half-turn binding domain representations. The dashed lines correspond to the expected order parameters in the fully stacked assembled state. The error bars represent the standard error as calculated with the MBAR method, using data from three independent temperature REMC simulations.

in a number of applications. The nature of the barriers would allow for a reversible change in state, and thus the origami may be switched between a bound and unbound configuration by changing solution conditions for functional purposes. The high degree of cooperativity in the maximal crossover system and the associated narrow temperature range over which these systems transition between is an additional property that may be of interest. One example of where these properties could be desirable is in creating molecular-scale thermometers, somewhat analogously to DNA origami being used as an in situ ruler [257]. However, this sensitivity may be extend or be extended (say by functionalizing the staples with the appropriate species) to other system conditions to act as a more general molecular sensor; these conditions may include pH, ionic strength, or concentration of various species. Because increasing the number of binding domains per staple increases the level of cooperativity and the narrowness of the transition range, the precision of these sensors may be tuned by the number of rows in the system.

Because of current limitations of the sampling methods, sampling efficiency

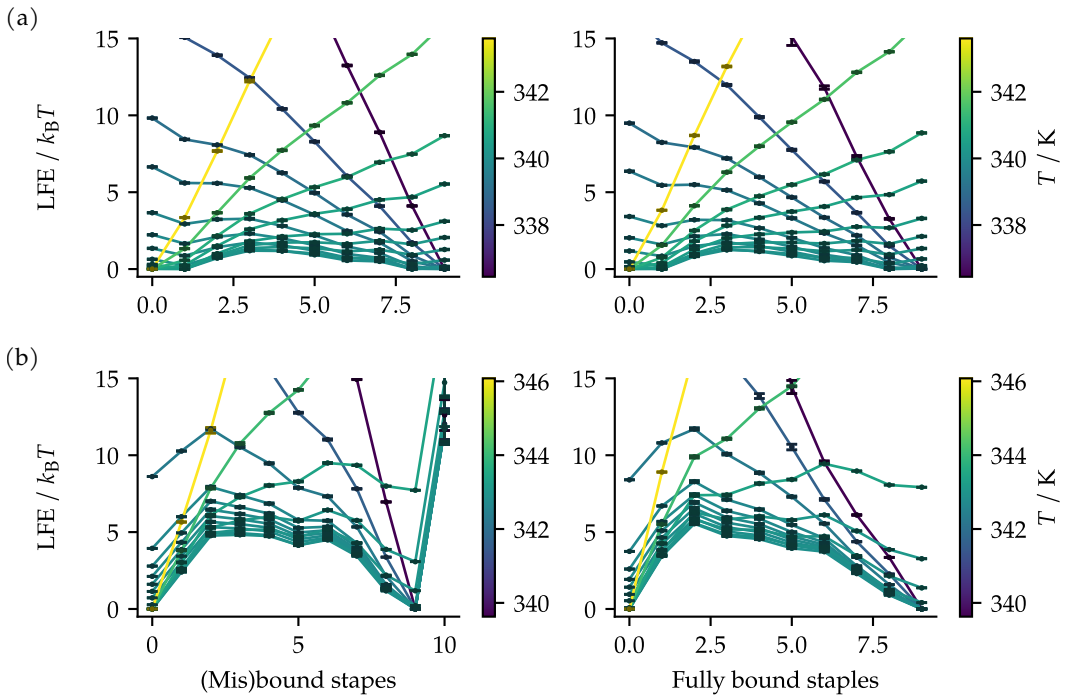

Figure 5.17: LFEs along the number of (mis)bound and the number of fully bound staples with three-quarter-turn binding domains. (a) Two-row system. (b) Three-row system. (c) Four-row system. The error bars represent the standard error as calculated with the MBAR method, using data from three independent temperature REMC simulations.

does not scale well with system size, and so we have been limited to relatively small designs in our study. It may be that with significantly more complex scaffold routing, other barriers may present themselves which have large enough magnitudes to be relevant to the assembly protocol. However, many commonly used origami designs are simply scaled up versions of what we have considered here, with more crossovers and some longer loops. We speculate that these results will hold for such larger designs, given that the barrier height scales with the per-staple stacking strength, rather than a global measure of the origami size.

Even if it is the case that our results are valid only for small origamis with a narrow range of hybridization free energies, they are still relevant to realizable designs and applications. While it is true that most origami designs employ larger scaffolds than those used here, that has been in part because of the convenience and availability of the M13mp18 phagemid. Smaller scaffolds have been developed [83, 85, 87, 88, 90, 91], the most recent of which additionally provides software and experimental protocols for the efficient creation of scaffolds of a range of sizes with minimal sequence constraints and with various user-determined desired properties [83]. We

speculate that the use of smaller scaffolds may become more popular with these innovations. Further, the use of custom scaffolds would allow for the creation of the highly cooperative maximum crossovers designs with a uniform hybridization potential that we simulated here by allowing for binding domains with a narrow distribution of hybridization free energies.

In studies such as these, poor selection of order parameters can not only misrepresent the barriers involved in assembly, but can also hinder efficient sampling if biased methods (e.g. US) are used. In addition to the order parameters considered here, we also consider order parameters more directly related to the configuration of the scaffold, including the radius of gyration, the end-to-end distance, the root mean square deviation (RMSD) of the scaffold relative to a fully stacked assembled configuration, and sums of the scaffold's internal distance matrix, including sums exclusively of distances of scaffold domains that are spanned by staples; however, none of these order parameters provided any additional insight. We also considered calculating LFE difference between staple states and binding domain states as a function of the total number of bound staples or domains, but these proved difficult to interpret because of insufficient statistics.

In the case of the 56-binding-domain scaffold system, we explored other order parameters by applying principal component analysis (PCA) on a vector of the occupancies of each scaffold site, the staple type states, and the scaffold stacked pairs (which does not sum to the number of stacked pairs order parameter, as this only considers pair stacking rules). The top two components were then combined in various ways, and clustering applied to identify macrostates. However, no clear or meaningful macrostates were discovered with the components examined. We do not rule out the possibility of utility in such an approach that was overlooked, but we concluded that it was unlikely to be fruitful and did not pursue it further by examining LFEs along the various components or applying the same approach to other systems.

Because of both the coarseness of our model and the fact that we use non-dynamic MC simulations, we cannot estimate absolute assembly rates. It may be that even with the small systems that have a small nucleation barrier or none at all, the assembly rate is slow enough to prevent effective use of the in principle reversible assembly. Such slow assembly could be caused by details finer than what we include here, and the rates of rearranging and aligning helices. If staples bind to multiple places on the scaffold concurrently, the rearrangement times may become quite slow. These jammed configurations could be avoided by designing the staple's hybridization free

energies to bind at higher temperatures to one particular location on the scaffold, and relying on cooperative effects to thermodynamically stabilize the subsequent binding of staples on adjacent parts of the scaffold. However, if one is interested in creating systems that can assemble in a narrow temperature range, an alternative approach to achieving this would be to design a nucleation barrier by using more binding domains per staple to increase the amount of stacking.

One possible difference between the self-assembly behaviour of DNA origami and DNA bricks is their propensity for aggregating in such a way as to prevent full assembly. In studies of DNA bricks, it was found that at lower temperatures incidental interactions led to aggregation of partially assembled structures, creating a rugged free energy landscape that inhibits the assembly process [190, 191, 193]. Our approach cannot directly simulate such aggregation in DNA origami systems because it does not include free staples or other scaffolds. However, some insight may still be gleaned from our simulations. Here, the LFEs along the number of bound domains are almost always downhill after the binding of the first domain of a staple. This would seem to imply that the staples tend to bind fully and have fewer unhybridized segments available. This could make DNA origami less prone to aggregation, as the partially assembled structures have fewer possibilities for incidental interactions with each other.

In summary, our results reveal that DNA origami self-assembly exhibits fundamentally different nucleation behaviour from DNA bricks self-assembly, and that it is possible to control the size of the barrier with the staple design, and even to eliminate it. We hope that our findings will prove useful in the creation of small DNA origami designs that utilize reversible folding for functional purposes, and more broadly to assist in the design of DNA origamis with desirable assembly pathway characteristics.

6
Conclusions

The self-assembly of DNA origami is a process that is of both practical and theoretical interest. We have introduced a lattice model that includes a level of detail relevant to uncovering the unique elements of the self-assembly mechanism, while ignoring further detail to allow for efficient simulation. The model is novel in the way it represents the helical twist constraints, with potential energy terms that are a function of the orientation vectors associated with the binding domains. Variations on the potential for different binding domains lengths are provided such that all designs able to represented on a simple cubic lattice are able to be modeled.

Efficient simulation of partially and fully assembled states required specialized MC methods. Two main ideas underly our approach. The first is that simulations are run in the grand ensemble with just a single copy of the scaffold, which allows staples not bound to the scaffold to be ignored. The second is that the sampling of staple binding states is done separately from the sampling of scaffold configurations, which both increases acceptance frequencies of moves and simplifies the development of new move types. New move types were developed that extend CB and RG for regrowing what can be seen generally as branched and looped lattice polymers, given that the staple states are held constant.

We have demonstrated that our model and methods are able to effectively sample assembled states, including fully stacked states, when starting from a fully unbound scaffold. The example thermodynamic analyses that we performed on the test systems, which included a system also simulated with the oxDNA model, largely matched our expectations. We found that staple blocking, where two copies of the same staple type bind to the scaffold and block each other from fully binding, was thermodynamic in origin once the staple concentration was sufficiently high.

We used the model to study the kinetics of origami self-assembly by examining barriers to assembly, and in particular investigated whether nucleation barriers were present and, if so, whether their role was significant. Systems with only two binding domains per staple were found to be downhill in free energy along all order

parameters tested. However, when we increased the number of binding domains per staple to three and four, a nucleation barrier appeared, albeit one easily surmountable with typical thermal fluctuations. The barrier was found to be primarily caused by the coaxial stacking between staples adjacent on the same helix. Increasing the number of binding domains per staple was one way of increasing the level of stacking per staple; we were also able to reproduce the nucleation barrier in systems with two binding domains per staple by increasing the stacking energy. Finally, the systems we designed to systematically test for nucleation barriers exhibited a very high degree of cooperativity, which could be useful in the creation of molecular sensors.

The model could be improved in a number of ways. We discussed how the reproduction of the stacking behaviour could be improved in Section 4.5 by having more than one stacking parameter depending on the context, and by including sequence specificity. We further discussed the possible pitfalls of assuming a constant staple concentration, and some possible alternatives to making such an assumption. In Section 2.3, we described what may be a more intuitive version of the model, the explicit helical axis model. Aside from being easier to reason about, it may allow for a more accurate representation of kinks and Holliday junctions, and would make the implementation of some of the potential terms that are a function of three or four binding domains be a function of only two binding domains, which could lead to gains in efficiency. However, as we argue in Section 4.5, the gains from such a model seem likely to be marginal over an implicit helical model. On a similar note, an off-lattice analogue could also be of interest to explore, but we note that the off-lattice analogue of the DNA bricks model was largely in agreement with the lattice model [192], which provides some support for a similar result here.

Perhaps the most important improvement that could be made to the model would be to better represent the trade-off between enthalpy and entropy as the system transitions to an assembled state. One approach to doing so would be to introduce additional parameters that could be tuned by comparison to experiment or simulations with a more detailed model. In fact, detailed simulations of a Holliday junction have been performed with the oxDNA model [187], which could be used to set an additional parameter for these junctions in our model. There are also a number of experimental studies that involve global transitions between the two isoforms of the Holliday junction in DNA origami (i.e. which strands continue the helix and which are involved in a strand crossover) [251, 258–261]. These studies additionally propose functional uses for such a global transition between distinct structures provided by the Holliday junction isoform transition, so it would be of

interest to reproduce the transition in our model beyond simply improving the accuracy of the results in other contexts.

The sampling methods likely represent the bottleneck to the applicability of our approach, rather than the accuracy of the model. In particular, the sampling of different stacked states of similar energies was one of the main culprits of the sampling difficulties. Even though individual stacked pair interactions are small, if a configuration is proposed with many fewer stacks, this will will result in a low likelihood of acceptance. To transition between configurations with many stacks, they will have to pass through these energetically unfavourable configurations, as the scaffold regrowth move types will not often propose configurations with as many stacks if they are regrowing many binding domains. While both CB and RG are biased towards lower energy configurations at the growth of each binding domain, a stack requires previous binding domain positions and orientation vectors to already be set a particular way for a stacked configuration to be an option to select from. This issue is similar to the issue that if a naive regrowth scheme were used, i.e. with no endpoint constraints as in the conserved topology move types developed here, moves would not often propose configurations with as many bound domains. Unfortunately, there is obvious analogous splitting of sampling of stacked states and scaffold configurations.

With smaller systems, our methods can effectively produce samples with well converged quantities of interest. However, the scaling seems to be rather poor with system size and complexity. There are a number of improvements that could be made. It was found that with larger systems, 2D REMC with a stacking energy multiplier and the temperature acting as independent exchange variables may provide an advantage over 1D REMC with temperature alone as an exchange multiplier. However, with systems that display high levels of cooperativity and a correspondingly sharp assembly transition, this method failed because of a lack of good coverage of the transition at multiple values of the stacking multiplier without having a prohibitively large number of replicas. One solution may be to vary the stacking multiplier with the staple concentration such that the total number of bound staples remains constant, or to vary the stacking multiplier with a multiplier on the hybridization free energy such that the total number of bound domain pairs remains constant.

For the calculation of LFEs, we found that the multi-window adaptive US scheme used here to be much less effective than the REMC schemes. However, the REMC schemes, at least for calculating LFEs of the highly cooperative systems that displayed nucleation barriers, were very sensitive to the selection of exchange variables with

respect to their sampling efficiency, and required the use of iterative schemes to select these variables. A more robust approach could be to use a form of replica exchange US with exchanges between windows, as this would ensure good sampling across the order parameter space. It could also be of some interest to explore other methods for calculating free energies which involve flattening the free-energy landscape, such as Wang–Landau sampling and its variations [262–265].

It may also be productive to further develop move types for sampling scaffold configurations. One idea that showed promise, but which has not been fully developed, was a move type that we refer to as transformation-linker-regrowth. This move type allowed a segment of the scaffold and any bound staples to be transformed with a series of translations and rotations, without regrowing it. It also involved the selection of linker regions that would be regrown after the transformation was applied. Such a move type allows separate chunks of the system that may be individually in a stacked assembled state to be reoriented with respect to each other without the cost of proposing a regrowth of many binding domains in the same stacked assembled state.

A more general extension would be to include an additional method for regrowing binding domains, in addition to the symmetric, CB, and RG methods. Recently, Boon [266] developed a polymer regrowth scheme that combines the advantages of the CB scheme with the PERM scheme, and can effectively apply a technique known as "waste recycling" that allows rejected configurations to be used in the ensemble averages to reduce statistical noise [267]. Like the RG method, it is useful for avoiding dead ends during regrowth, but compared against RG, it was found to be an order of magnitude more efficient with their test system and implementations. One of the problems with RG in our work was that the cost of 'looking further ahead' by allowing for more recoils, while effective at increasing the acceptance rate of long chain regrowths, came at a very high computational cost, which made it less efficient than using a relatively small maximum number of recoils. It seems plausible that the method of Boon may scale better than RG in this respect, given its use of a PERM-like method for avoiding dead-ends.

Our approach should be generally useful in a variety of contexts. We focused our efforts on systems that had two binding domains per staple, and systems that had staples which crossed over at every opportunity (at least given the resolution of the model). However, there are other ways to design staples, with even the designs in the original paper of Rothemund [9] having motifs that we have not studied. It would be of interest to perform similar analysis to those performed in this thesis on such

systems. It would also be of interest to study systems that are designed to assemble into 3D structures. With the current form of our model, it would be possible to simulate 3D designs that are based on a square lattice, as Ke et al. [73] pioneered. It would also not be difficult to extend the approach to allow for simulation of ssDNA origami [92]. The model might also be extended to include staples that have been functionalized in some way by adding additional terms to the potential that act to occupy lattice sites or even to interact with each other or binding domains in the system.

The model with the half-turn and three-quarter-turn binding domains have been implemented as a computer program which we have released to the public with a permissive license. The software has been designed to not require modification of the source code to run a simulation. Options are specified as command line arguments or in a configuration file with key value pairs, with additional options including system definitions, configurations, biases, order parameters, and move sets being defined in JSON formatted files. Also included are a variety of scripts to analyze and visualize the simulations. We hope that the software will be of some general utility to the community, and that its use by a broader set of researchers will lead to further applications of the model that we have not even considered here.

References

[1] A. Cumberworth, A. Reinhardt, and D. Frenkel, "Lattice models and Monte Carlo methods for simulating DNA origami self-assembly", J. Chem. Phys. **149**, 234905 (2018).

[2] E. T. Kool, "Hydrogen bonding, base stacking, and steric effects in dna replication", Annu. Rev. Biophys. Biomol. Struct. **30**, 1–22 (2001).

[3] N. C. Seeman, *Structural DNA nanotechnology* (Cambridge University Press, 2015).

[4] N. C. Seeman, "Nanomaterials based on DNA", Annu. Rev. Biochem. **79**, 65–87 (2010).

[5] N. C. Seeman, "Nucleic acid junctions and lattices", J. Theor. Biol. **99**, 237–247 (1982).

[6] M. Tintoré, R. Eritja, and C. Fábrega, "DNA nanoarchitectures: steps towards biological applications", ChemBioChem **15**, 1374–1390 (2014).

[7] B. Wei, M. Dai, and P. Yin, "Complex shapes self-assembled from single-stranded DNA tiles", Nature **485**, 623–626 (2012).

[8] Y. Ke, L. L. Ong, W. M. Shih, and P. Yin, "Three-dimensional structures self-assembled from DNA bricks", Science **338**, 1177–1183 (2012).

[9] P. Rothemund, "Folding DNA to create nanoscale shapes and patterns", Nature **440**, 297–302 (2006).

[10] S. M. Douglas, H. Dietz, T. Liedl, B. Hoegberg, F. Graf, and W. M. Shih, "Self-assembly of DNA into nanoscale three-dimensional shapes", Nature **459**, 414–418 (2009).

[11] C. E. Castro, F. Kilchherr, D.-N. Kim, E. L. Shiao, T. Wauer, P. Wortmann, M. Bathe, and H. Dietz, "A primer to scaffolded DNA origami", Nat. Methods **8**, 221–229 (2011).

[12] D. Han, S. Pal, Y. Yang, S. Jiang, J. Nangreave, Y. Liu, and H. Yan, "DNA gridiron nanostructures based on four-arm junctions", Science **339**, 1412–1415 (2013).

[13] E. Benson, A. Mohammed, J. Gardell, S. Masich, E. Czeizler, P. Orponen, and B. Hogberg, "DNA rendering of polyhedral meshes at the nanoscale", Nature **523**, 441–444 (2015).

[14] C. E. Castro, H.-J. Su, A. E. Marras, L. Zhou, and J. Johnson, "Mechanical design of DNA nanostructures", Nanoscale **7**, 5913–5921 (2015).

[15] F. Zhang, S. Jiang, S. Wu, Y. Li, C. Mao, Y. Liu, and H. Yan, "Complex wireframe DNA origami nanostructures with multi-arm junction vertices", Nat. Nanotechnol. **10**, 779–784 (2015).

[16] H. Ijäs, S. Nummelin, B. Shen, M. Kostiainen, and V. Linko, "Dynamic DNA origami devices: from strand-displacement reactions to external-stimuli responsive systems", Int. J. Mol. Sci. **19**, 2114 (2018).

[17]D. Balakrishnan, G. D. Wilkens, and J. G. Heddle, "Delivering DNA origami to cells", Nanomedicine **14**, 911–925 (2019).

[18]H. Bila, E. E. Kurisinkal, and M. M. C. Bastings, "Engineering a stable future for DNA-origami as a biomaterial", Biomater. Sci. **7**, 532–541 (2019).

[19]Y. Dong and Y. Mao, "DNA origami as scaffolds for self-assembly of lipids and proteins", ChemBioChem.

[20]S. Fan, D. Wang, A. Kenaan, J. Cheng, D. Cui, and J. Song, "Create nanoscale patterns with DNA origami", Small **15**, 1805554 (2019).

[21]S. Ramakrishnan, H. Ijäs, V. Linko, and A. Keller, "Structural stability of DNA origami nanostructures under application-specific conditions", Comput. Struct. Biotechnol. J. **16**, 342–349 (2018).

[22]Y. Sakai, M. S. Islam, M. Adamiak, S. C.-C. Shiu, J. A. Tanner, and J. G. Heddle, "DNA aptamers for the functionalisation of dna origami nanostructures", Genes **9**, 571 (2018).

[23]S. Kogikoski, W. J. Paschoalino, and L. T. Kubota, "Supramolecular DNA origami nanostructures for use in bioanalytical applications", TrAC, Trends Anal. Chem. **108**, 88–97 (2018).

[24]M. Endo and H. Sugiyama, "DNA origami nanomachines", Molecules **23**, 1766 (2018).

[25]S. Loescher, S. Groeer, and A. Walther, "3D DNA origami nanoparticles: from basic design principles to emerging applications in soft matter and (bio-)nanosciences", Angew. Chem., Int. Ed. **57**, 10436–10448 (2018).

[26]P. Wang, T. A. Meyer, V. Pan, P. K. Dutta, and Y. Ke, "The beauty and utility of DNA origami", Chem **2**, 359–382 (2017).

[27]A. Udomprasert and T. Kangsamaksin, "DNA origami applications in cancer therapy", Cancer Sci. **108**, 1535–1543 (2017).

[28]F. Hong, F. Zhang, Y. Liu, and H. Yan, "DNA origami: scaffolds for creating higher order structures", Chem. Rev. **117**, 12584–12640 (2017).

[29]A. R. Chandrasekaran, N. Anderson, M. Kizer, K. Halvorsen, and X. Wang, "Beyond the fold: emerging biological applications of DNA origami", ChemBioChem **17**, 1081–1089 (2016).

[30]V. Linko, A. Ora, and M. A. Kostiainen, "DNA nanostructures as smart drug-delivery vehicles and molecular devices", Trends Biotechnol. **33**, 586–594 (2015).

[31]H. Jabbari, M. Aminpour, and C. Montemagno, "Computational approaches to nucleic acid origami", ACS Comb. Sci. **17**, 535–547 (2015).

[32]L. A. Lanier and H. Bermudez, "DNA nanostructures: a shift from assembly to applications", Curr. Opin. Chem. Eng. **7**, 93–100 (2015).

[33]A. Kuzuya and Y. Ohya, "Nanomechanical molecular devices made of DNA origami", Acc. Chem. Res. **47**, 1742–1749 (2014).

[34]N. A. Bell and U. F. Keyser, "Nanopores formed by DNA origami: a review", FEBS Lett. **588**, 3564–3570 (2014).

[35]S. Hernández-Ainsa and U. F. Keyser, "DNA origami nanopores: developments, challenges and perspectives", Nanoscale **6**, 14121–14132 (2014).

[36]I. Bald and A. Keller, "Molecular processes studied at a single-molecule level using DNA origami nanostructures and atomic force microscopy", Molecules **19**, 13803–13823 (2014).

[37]M. Endo and H. Sugiyama, "Single-molecule imaging of dynamic motions of biomolecules in DNA origami nanostructures using high-speed atomic force microscopy", Acc. Chem. Res. **47**, 1645–1653 (2014).

[38]F. Zhang, J. Nangreave, Y. Liu, and H. Yan, "Structural DNA nanotechnology: state of the art and future perspective", J. Am. Chem. Soc. **136**, 11198–11211 (2014).

[39]V. Linko and H. Dietz, "The enabled state of DNA nanotechnology", Curr. Opin. Biotechnol. **24**, 555–561 (2013).

[40]G. Zhang, S. P. Surwade, F. Zhou, and H. Liu, "DNA nanostructure meets nanofabrication", Chem. Soc. Rev. **42**, 2488–2496 (2013).

[41]I. Saaem and T. H. LaBean, "Overview of DNA origami for molecular self-assembly", Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol. **5**, 150–162 (2013).

[42]B. Saccà and C. M. Niemeyer, "DNA origami: the art of folding dna", Angew. Chem., Int. Ed. **51**, 58–66 (2012).

[43]N. Michelotti, A. Johnson-Buck, A. J. Manzo, and N. G. Walter, "Beyond DNA origami: the unfolding prospects of nucleic acid nanotechnology", Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol. **4**, 139–152 (2012).

[44]A. Rajendran, M. Endo, and H. Sugiyama, "Single-molecule analysis using DNA origami", Angew. Chem., Int. Ed. **51**, 874–890 (2012).

[45]O. I. Wilner and I. Willner, "Functionalized DNA nanostructures", Chem. Rev. **112**, 2528–2556 (2012).

[46]J. Fu, M. Liu, Y. Liu, and H. Yan, "Spatially-interactive biomolecular networks organized by nucleic acid nanostructures", Acc. Chem. Res. **45**, 1215–1226 (2012).

[47]T. Tørring, N. V. Voigt, J. Nangreave, H. Yan, and K. V. Gothelf, "DNA origami: a quantum leap for self-assembly of complex structures", Chem. Soc. Rev. **40**, 5636–5646 (2011).

[48]W. M. Shih and C. Lin, "Knitting complex weaves with DNA origami", Curr. Opin. Struct. Biol. **20**, 276–282 (2010).

[49]J. Nangreave, D. Han, Y. Liu, and H. Yan, "DNA origami: a history and current perspective", Curr. Opin. Chem. Biol. **14**, 608–615 (2010).

[50]D. Frenkel, "Order through entropy", Nat. Mater. **14**, 9–12 (2015).

[51]N. V. Voigt, T. Torring, A. Rotaru, M. F. Jacobsen, J. B. Ravnsbaek, R. Subramani, W. Mamdouh, J. Kjems, A. Mokhir, F. Besenbacher, and K. V. Gothelf, "Single-molecule chemical reactions on DNA origami", Nat. Nanotechnol. **5**, 200–203 (2010).

[52]Y. Ke, S. Lindsay, Y. Chang, Y. Liu, and H. Yan, "Self-assembled water-soluble nucleic acid probe tiles for label-free RNA hybridization assays", Science **319**, 180–183 (2008).

[53]V. Linko, M. Eerikainen, and M. A. Kostiainen, "A modular DNA origami-based enzyme cascade nanoreactor", Chem. Commun. **51**, 5351–5354 (2015).

[54]M. Liu, J. Fu, C. Hejesen, Y. Yang, N. W. Woodbury, K. Gothelf, Y. Liu, and H. Yan, "A DNA tweezer-actuated enzyme nanoreactor", Nat. Commun. **4**, 2127 (2013).

[55]H. T. Maune, S.-p. Han, R. D. Barish, M. Bockrath, W. A. Goddard III, P. W. K. Rothemund, and E. Winfree, "Self-assembly of carbon nanotubes into two-dimensional geometries using DNA origami templates", Nat. Nanotechnol. **5**, 61–66 (2010).

[56]R. J. Kershner, L. D. Bozano, C. M. Micheel, A. M. Hung, A. R. Fornof, J. N. Cha, C. T. Rettner, M. Bersani, J. Frommer, P. W. K. Rothemund, and G. M. Wallraff, "Placement and orientation of individual DNA shapes on lithographically patterned surfaces", Nat. Nanotechnol. **4**, 557–561 (2009).

[57]J. Chao, Y. Lin, H. Liu, L. Wang, and C. Fan, "DNA-based plasmonic nanostructures", Mater. Today **18**, 326–335 (2015).

[58]R. Schreiber, J. Do, E.-M. Roller, T. Zhang, V. J. Schüller, P. C. Nickels, J. Feldmann, and T. Liedl, "Hierarchical assembly of metal nanoparticles, quantum dots and organic dyes using DNA origami scaffolds", Nat. Nanotechnol. **9**, 74–78 (2014).

[59]B. Ding, Z. Deng, H. Yan, S. Cabrini, R. N. Zuckermann, and J. Bokor, "Gold nanoparticle self-similar chain structure organized by DNA origami", J. Am. Chem. Soc. **132**, 3248–3249 (2010).

[60]J. Burns, "DNA origami inside-out viruses", ACS Synth. Biol. **7**, 767–773 (2018).

[61]E. S. Andersen, M. Dong, M. M. Nielsen, K. Jahn, R. Subramani, W. Mamdouh, M. M. Golas, B. Sander, H. Stark, C. L. P. Oliveira, J. S. Pedersen, V. Birkedal, F. Besenbacher, K. V. Gothelf, and J. Kjems, "Self-assembly of a nanoscale DNA box with a controllable lid", Nature **459**, 73–76 (2009).

[62]Q. Zhang, Q. Jiang, N. Li, L. Dai, and Q. a. Liu, "DNA origami as an in vivo drug delivery vehicle for cancer therapy", ACS Nano **8**, 6633–6643 (2014).

[63]S. Li, Q. Jiang, S. Liu, Y. Zhang, Y. Tian, C. Song, J. Wang, Y. Zou, G. J. Anderson, J.-Y. Han, Y. Chang, Y. Liu, C. Zhang, L. Chen, G. Zhou, G. Nie, H. Yan, B. Ding, and Y. Zhao, "A DNA nanorobot functions as a cancer therapeutic in response to a molecular trigger in vivo", Nat. Biotechnol. **36**, 258+ (2018).

[64]S. F. J. Wickham, J. Bath, Y. Katsuda, M. Endo, K. Hidaka, H. Sugiyama, and A. J. Turberfield, "A DNA-based molecular motor that can navigate a network of tracks", Nat. Nanotechnol. **7**, 169–173 (2012).

[65]A. E. Marras, L. Zhou, H.-J. Su, and C. E. Castro, "Programmable motion of DNA origami mechanisms", Proc. Natl. Acad. Sci. U. S. A. **112**, 713–718 (2015).

[66]E. Kopperger, J. List, S. Madhira, F. Rothfischer, D. C. Lamb, and F. C. Simmel, "A self-assembled nanoscale robotic arm controlled by electric fields", Science **359**, 296–301 (2018).

[67]F. Kroener, A. Heerwig, W. Kaiser, M. Mertig, and U. Rant, "Electrical actuation of a DNA origami nanolever on an electrode", J. Am. Chem. Soc. **139**, 16510–16513 (2017).

[68]S. Arnon, N. Dahan, A. Koren, O. Radiano, M. Ronen, T. Yannay, J. Giron, L. Ben-Ami, Y. Amir, Y. Hel-Or, D. Friedman, and I. Bachelet, "Thought-controlled nanoscale robots in a living host", PLoS One **11**, 1–8 (2016).

[69]J. Chao, J. Wang, F. Wang, X. Ouyang, E. Kopperger, H. Liu, Q. Li, J. Shi, L. Wang, J. Hu, L. Wang, W. Huang, F. C. Simmel, and C. Fan, "Solving mazes with single-molecule DNA navigators", Nat. Mater. **18**, 273+ (2019).

[70]Y. Amir, E. Ben-Ishay, D. Levner, S. Ittah, A. Abu-Horowitz, and I. Bachelet, "Universal computing by DNA origami robots in a living animal", Nat. Nanotechnol. **9**, 353–357 (2014).

[71]H. Gu, J. Chao, S.-J. Xiao, and N. C. Seeman, "A proximity-based programmable DNA nanoscale assembly line", Nature **465**, 202–U86 (2010).

[72]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 Res. **37**, 5001–5006 (2009).

[73]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", J. Am. Chem. Soc. **131**, 15903–15908 (2009).

[74]N. A. W. Bell, C. R. Engst, M. Ablay, G. Divitini, C. Ducati, T. Liedl, and U. F. Keyser, "DNA origami nanopores", Nano Lett. **12**, 512–517 (2012).

[75]M. Langecker, V. Arnaut, T. G. Martin, J. List, S. Renner, M. Mayer, H. Dietz, and F. C. Simmel, "Synthetic lipid membrane channels formed by designed DNA nanostructures", Science **338**, 932–936 (2012).

[76]W. Wang, S. Chen, B. An, K. Huang, T. Bai, M. Xu, G. Bellot, Y. Ke, Y. Xiang, and B. Wei, "Complex wireframe DNA nanostructures from simple building blocks", Nat. Commun. **10**, 1067 (2019).

[77]R. Veneziano, S. Ratanalert, K. Zhang, F. Zhang, H. Yan, W. Chiu, and M. Bathe, "Designer nanoscale DNA assemblies programmed from the top down", Science **352**, 1534–1534 (2016).

[78]G. Tikhomirov, P. Petersen, and L. Qian, "Fractal assembly of micrometre-scale DNA origami arrays with arbitrary patterns", Nature **552**, 67–71 (2017).

[79]J. Fern, J. Lu, and R. Schulman, "The energy landscape for the self-assembly of a two-dimensional DNA origami complex", ACS Nano **10**, 1836–1844 (2016).

[80]T. Gerling, K. F. Wagenbauer, A. M. Neuner, and H. Dietz, "Dynamic DNA devices and assemblies formed by shape-complementary, non–base pairing 3D components", Science **347**, 1446–1452 (2015).

[81]Y. Fu, D. Zeng, J. Chao, Y. Jin, Z. Zhang, H. Liu, D. Li, H. Ma, Q. Huang, K. V. Gothelf, and C. Fan, "Single-step rapid assembly of DNA origami nanostructures for addressable nanoscale bioreactors", J. Am. Chem. Soc. **135**, 696–702 (2013).

[82]A. R. Chandrasekaran, M. Pushpanathan, and K. Halvorsen, "Evolution of DNA origami scaffolds", Mater. Lett. **170**, 221–224 (2016).

[83]F. Engelhardt, "Custom-size, functional, and durable DNA origami with design-specific scaffolds", ACS Nano **13**, 5015–5027 (2019).

[84]X. Chen, Q. Wang, J. Peng, Q. Long, H. Yu, and Z. Li, "Self-assembly of large DNA origami with custom-designed scaffolds", ACS Appl. Mater. Interfaces **10**, 24344–24348 (2018).

[85]P. M. Nafisi, T. Aksel, and S. M. Douglas, "Construction of a novel phagemid to produce custom DNA origami scaffolds", Synth. Biol. **3**, ysy015 (2018).

[86]K. F. Wagenbauer, C. Sigl, and H. Dietz, "Gigadalton-scale shape-programmable DNA assemblies", Nature **552**, 78–83 (2017).

[87]S. Brown, J. Majikes, A. Martínez, T. M. Girón, H. Fennell, E. C. Samano, and T. H. LaBean, "An easy-to-prepare mini-scaffold for DNA origami", Nanoscale **7**, 16621–16624 (2015).

[88]M. Erkelenz, D. M. Bauer, R. Meyer, C. Gatsogiannis, S. Raunser, B. Saccà, and C. M. Niemeyer, "A facile method for preparation of tailored scaffolds for DNA-origami", Small **10**, 73–77 (2014).

[89]A. N. Marchi, I. Saaem, B. N. Vogen, S. Brown, and T. H. LaBean, "Toward larger DNA origami", Nano Lett. **14**, 5740–5747 (2014).

[90]H. Said, V. J. Schüller, F. J. Eber, C. Wege, T. Liedl, and C. Richert, "M1.3 – a small scaffold for DNA origami", Nanoscale **5**, 284–290 (2013).

[91]E. Pound, J. R. Ashton, H. A. Becerril, and A. T. Woolley, "Polymerase chain reaction based scaffold preparation for the production of thin, branched DNA origami nanostructures of arbitrary sizes", Nano Lett. **9**, 4302–4305 (2009).

[92]D. Han, X. Qi, C. Myhrvold, B. Wang, M. Dai, S. Jiang, M. Bates, Y. Liu, B. An, F. Zhang, H. Yan, and P. Yin, "Single-stranded DNA and RNA origami", Science **358**, eaao2648 (2017).

[93]F. Praetorius, B. Kick, K. L. Behler, M. N. Honemann, D. Weuster-Botz, and H. Dietz, "Biotechnological mass production of DNA origami", Nature **552**, 84+ (2017).

[94]S. Niekamp, K. Blumer, P. M. Nafisi, K. Tsui, J. Garbutt, and S. M. Douglas, "Folding complex DNA nanostructures from limited sets of reusable sequences", Nucleic Acids Res. **44**, e102–e102 (2016).

[95]M. T. Strauss, F. Schueder, D. Haas, P. C. Nickels, and R. Jungmann, "Quantifying absolute addressability in DNA origami with molecular resolution", Nat. Commun. **9**, 1600 (2018).

[96]K. F. Wagenbauer, C. H. Wachauf, and H. Dietz, "Quantifying quality in DNA self-assembly", Nat. Commun. **5**, 3691 (2014).

[97]K. F. Wagenbauer, F. A. S. Engelhardt, E. Stahl, V. K. Hechtl, P. Stömmer, F. Seebacher, L. Meregalli, P. Ketterer, T. Gerling, and H. Dietz, "How we make DNA origami", ChemBioChem **18**, 1873–1885 (2017).

[98]J.-P. J. Sobczak, T. G. Martin, T. Gerling, and H. Dietz, "Rapid folding of DNA into nanoscale shapes at constant temperature", Science **338**, 1458–1461 (2012).

[99]R. Jungmann, T. Liedl, T. L. Sobey, W. Shih, and F. C. Simmel, "Isothermal assembly of DNA origami structures using denaturing agents", J. Am. Chem. Soc. **130**, 10062–10063 (2008).

[100]P. D. Halley, "Low-cost, simple, and scalable self-assembly of DNA origami nanostructures", Nano Res. **12**, 1207–1215 (2019).

[101]Z. Zhang, J. Song, F. Besenbacher, M. Dong, and K. V. Gothelf, "Self-assembly of DNA origami and single-stranded tile structures at room temperature", Angew. Chem., Int. Ed. **52**, 9219–9223 (2013).

[102]A. Kopielski, A. Schneider, A. Csaki, and W. Fritzsche, "Isothermal DNA origami folding: avoiding denaturing conditions for one-pot, hybrid-component annealing", Nanoscale **7**, 2102–2106 (2015).

[103]W. Bae, K. Kim, D. Min, J.-K. Ryu, C. Hyeon, and T.-Y. Yoon, "Programmed folding of DNA origami structures through single-molecule force control", Nat. Commun. **5**, 5654 (2014).

[104]T. G. Martin and H. Dietz, "Magnesium-free self-assembly of multi-layer DNA objects", Nat. Commun. **3**, 1103 (2012).

[105]A. N. Marchi, I. Saaem, J. Tian, and T. H. LaBean, "One-pot assembly of a hetero-dimeric DNA origami from chip-derived staples and double-stranded scaffold", ACS Nano **7**, 903–910 (2013).

[106]A. Rajendran, M. Endo, K. Hidaka, N. Shimada, A. Maruyama, and H. Sugiyama, "A lock-and-key mechanism for the controllable fabrication of DNA origami structures", Chem. Commun. **50**, 8743–8746 (2014).

[107]L. Cademartiri and K. J. M. Bishop, "Programmable self-assembly", Nat. Mater. **14**, 2–9 (2015).

[108]X. Wei, J. Nangreave, and Y. Liu, "Uncovering the self-assembly of DNA nanostructures by thermodynamics and kinetics", Acc. Chem. Res. **47**, 1861–1870 (2014).

[109]K. E. Dunn, F. Dannenberg, T. E. Ouldridge, M. Kwiatkowska, A. J. Turberfield, and J. Bath, "Guiding the folding pathway of DNA origami", Nature **525**, 82 (2015).

[110]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", J. Chem. Phys. **143**, 165102 (2015).

[111]X. Wei, J. Nangreave, S. Jiang, H. Yan, and Y. Liu, "Mapping the thermal behavior of DNA origami nanostructures", J. Am. Chem. Soc. **135**, 6165–6176 (2013).

[112]J.-M. Arbona, J.-P. Aimé, and J. Elezgaray, "Cooperativity in the annealing of DNA origamis", J. Chem. Phys. **138**, 015105 (2013).

[113]J.-M. Arbona, J. Elezgaray, and J.-P. Aimé, "Modelling the folding of DNA origami", Europhys. Lett. **100**, 28006 (2012).

[114]J.-M. Arbona, J.-P. Aimé, and J. Elezgaray, "Folding of DNA origamis", Front. Life Sci. **6**, 11–18 (2012).

[115]J. L. T. Wah, C. David, S. Rudiuk, D. Baigl, and A. Estevez-Torres, "Observing and controlling the folding pathway of DNA origami at the nanoscale", ACS Nano **10**, 1978–1987 (2016).

[116]F. Schneider, N. Möritz, and H. Dietz, "The sequence of events during folding of a DNA origami", Sci. Adv. **5**, eaaw1412 (2019).

[117]A. Shapiro, A. Hozeh, O. Girshevitz, A. Abu-Horowitz, and I. Bachelet, "Cooperativity-based modeling of heterotypic DNA nanostructure assembly", Nucleic Acids Res. **43**, 6587–6595 (2015).

[118]J. M. Majikes, J. A. Nash, and T. H. LaBean, "Search for effective chemical quenching to arrest molecular assembly and directly monitor DNA nanostructure formation", Nanoscale **9**, 1637–1644 (2017).

[119]J. Song, Z. Zhang, S. Zhang, L. Liu, Q. Li, E. Xie, K. V. Gothelf, F. Besenbacher, and M. Dong, "Isothermal hybridization kinetics of DNA assembly of two-dimensional DNA origami", Small **9**, 2954–2959 (2013).

[120]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 **10**, 1724–1737 (2016).

[121]P. D. Dans, J. Walther, H. Gómez, and M. Orozco, "Multiscale simulation of DNA", Curr. Opin. Struct. Biol. **37**, 29–45 (2016).

[122]T. Dršata and F. Lankaš, "Multiscale modelling of DNA mechanics", J. Phys.: Condens. Matter **27**, 323102 (2015).

[123]T. E. Ouldridge, "DNA nanotechnology: understanding and optimisation through simulation", Mol. Phys. **113**, 1–15 (2015).

[124]C. Maffeo, J. Yoo, J. Comer, D. B. Wells, B. Luan, and A. Aksimentiev, "Close encounters with DNA", J. Phys.: Condens. Matter **26**, 413101 (2014).

[125]J. P. K. Doye, T. E. Ouldridge, A. A. Louis, F. Romano, P. Sulc, C. Matek, B. E. K. Snodin, L. Rovigatti, J. S. Schreck, R. M. Harrison, and W. P. J. Smith, "Coarse-graining DNA for simulations of DNA nanotechnology", Phys. Chem. Chem. Phys. **15**, 20395–20414 (2013).

[126]G. S. Freeman, D. M. Hinckley, J. P. Lequieu, J. K. Whitmer, and J. J. de Pablo, "Coarse-grained modeling of DNA curvature", J. Chem. Phys. **141**, 165103 (2014).

[127]N. Korolev, D. Luo, A. P. Lyubartsev, and L. Nordenskiöld, "A coarse-grained DNA model parameterized from atomistic simulations by inverse Monte Carlo", Polymers **6**, 1655 (2014).

[128]T. Cragnolini, P. Derreumaux, and S. Pasquali, "Coarse-grained simulations of RNA and DNA duplexes", J. Phys. Chem. B **117**, 8047–8060 (2013).

[129]O. Gonzalez, D. Petkevičiūtė, and J. H. Maddocks, "A sequence-dependent rigid-base model of DNA", J. Chem. Phys. **138**, 055102 (2013).

[130]Y. He, M. Maciejczyk, S. Ołdziej, H. A. Scheraga, and A. Liwo, "Mean-field interactions between nucleic-acid-base dipoles can drive the formation of a double helix", Phys. Rev. Lett. **110**, 098101 (2013).

[131]D. M. Hinckley, G. S. Freeman, J. K. Whitmer, and J. J. de Pablo, "An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: structure, thermodynamics, and dynamics of hybridization", J. Chem. Phys. **139**, 144903 (2013).

[132]L. E. Edens, J. A. Brozik, and D. J. Keller, "Coarse-grained model DNA: structure, sequences, stems, circles, hairpins", J. Phys. Chem. B **116**, 14735–14743 (2012).

[133]C. W. Hsu, M. Fyta, G. Lakatos, S. Melchionna, and E. Kaxiras, "Ab initio determination of coarse-grained interactions in double-stranded DNA", J. Chem. Phys. **137**, 105102 (2012).

[134]I. Kikot, A. Savin, E. Zubova, M. Mazo, E. Gusarova, L. Manevitch, and A. Onufriev, "New coarse-grained DNA model", Biophysics **56**, 387–392 (2011).

[135]C. Knorowski, S. Burleigh, and A. Travesset, "Dynamics and statics of DNA-programmable nanoparticle self-assembly and crystallization", Phys. Rev. Lett. **106**, 215501 (2011).

[136]M. C. Linak, R. Tourdot, and K. D. Dorfman, "Moving beyond Watson–Crick models of coarse grained DNA dynamics", J. Chem. Phys. **135**, 205102 (2011).

[137]P. D. Dans, A. Zeida, M. R. Machado, and S. Pantano, "A coarse grained model for atomic-detailed DNA simulations with explicit electrostatics", J. Chem. Theory Comput. **6**, 1711–1725 (2010).

[138]A. Morriss-Andrews, J. Rottler, and S. S. Plotkin, "A systematically coarse-grained model for DNA and its predictions for persistence length, stacking, twist, and chirality", J. Chem. Phys. **132**, 035105 (2010).

[139]A. Savelyev and G. A. Papoian, "Chemically accurate coarse graining of double-stranded DNA", Proc. Natl. Acad. Sci. U. S. A. **107**, 20340–20345 (2010).

[140]M. Sayar, B. Avşaroğlu, and K. Alkan, "Twist-writhe partitioning in a coarse-grained DNA minicircle model", Phys. Rev. E **81**, 041916 (2010).

[141]N. B. Tito and J. M. Stubbs, "Application of a coarse-grained model for DNA to homo- and heterogeneous melting equilibria", Chem. Phys. Lett. **485**, 354–359 (2010).

[142]M. Kenward and K. D. Dorfman, "Brownian dynamics simulations of single-stranded DNA hairpins", J. Chem. Phys. **130**, 095101 (2009).

[143]F. Lankas, O. Gonzalez, L. M. Heffler, G. Stoll, M. Moakher, and J. H. Maddocks, "On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations", Phys. Chem. Chem. Phys. **11**, 10565–10588 (2009).

[144]S. Niewieczerzał and M. Cieplak, "Stretching and twisting of the DNA duplexes in coarse-grained dynamical models", J. Phys.: Condens. Matter **21**, 474221 (2009).

[145]E. Sambriski, D. Schwartz, and J. de Pablo, "A mesoscale model of DNA and its renaturation", Biophys. J. **96**, 1675–1690 (2009).

[146]N. B. Becker and R. Everaers, "From rigid base pairs to semiflexible polymers: coarse-graining DNA", Phys. Rev. E **76**, 021923 (2007).

[147]T. A. Knotts, N. Rathore, D. C. Schwartz, and J. J. de Pablo, "A coarse grain model for DNA", J. Chem. Phys. **126**, 084901 (2007).

[148]F. W. Starr and F. Sciortino, "Model for assembly and gelation of four-armed DNA dendrimers", J. Phys.: Condens. Matter **18**, L347 (2006).

[149]H. L. Tepper and G. A. Voth, "A coarse-grained model for double-helix molecules in solution: spontaneous helix formation and equilibrium properties", J. Chem. Phys. **122**, 124906 (2005).

[150]K. Drukker, G. Wu, and G. C. Schatz, "Model simulations of DNA denaturation dynamics", J. Chem. Phys. **114**, 579–590 (2001).

[151]W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock, and V. B. Zhurkin, "DNA sequence-dependent deformability deduced from protein–DNA crystal complexes", Proc. Natl. Acad. Sci. U. S. A. **95**, 11163–11168 (1998).

[152]A. Naômé, A. Laaksonen, and D. P. Vercauteren, "A solvent-mediated coarse-grained model of DNA derived with the systematic newton inversion method", J. Chem. Theory Comput. **10**, 3541–3549 (2014).

[153]A. K. Dasanna, N. Destainville, J. Palmeri, and M. Manghi, "Slow closure of denaturation bubbles in DNA: twist matters", Phys. Rev. E **87**, 052703 (2013).

[154]C. Svaneborg, "LAMMPS framework for dynamic bonding and an application modeling DNA", Comput. Phys. Commun. **183**, 1793–1802 (2012).

[155]J. C. Araque, A. Z. Panagiotopoulos, and M. A. Robert, "Lattice model of oligonucleotide hybridization in solution. I. Model and thermodynamics", J. Chem. Phys. **134**, 165103 (2011).

[156]K. Doi, T. Haga, H. Shintaku, and S. Kawano, "Development of coarse-graining DNA models for single-nucleotide resolution analysis", Philos. Trans. R. Soc., A **368**, 2615–2628 (2010).

[157]F. Trovato and V. Tozzini, "Supercoiling and local denaturation of plasmids with a minimalist DNA model", J. Phys. Chem. B **112**, 13197–13200 (2008).

[158]S. P. Mielke, N. Grønbech-Jensen, V. V. Krishnan, W. H. Fink, and C. J. Benham, "Brownian dynamics simulations of sequence-dependent duplex denaturation in dynamically superhelical DNA", J. Chem. Phys. **123**, 124911 (2005).

[159]M. Sales-Pardo, R. Guimerà, A. A. Moreira, J. Widom, and L. A. N. Amaral, "Mesoscopic modeling for nucleic acid chain dynamics", Phys. Rev. E **71**, 051902 (2005).

[160]X. Li, C. M. Schroeder, and K. D. Dorfman, "Modeling the stretching of wormlike chains in the presence of excluded volume", Soft Matter **11**, 5947–5954 (2015).

[161]G. T. Barkema, D. Panja, and J. M. J. van Leeuwen, "Semiflexible polymer dynamics with a bead-spring model", J. Stat. Mech.: Theory Exp. **2014**, P11008 (2014).

[162]J. G. de la Torre, J. G. H. Cifre, Á. Ortega, R. R. Schmidt, M. X. Fernandes, H. E. P. Sánchez, and R. Pamies, "SIMUFLEX: algorithms and tools for simulation of the conformation and dynamics of flexible molecules and nanoparticles in dilute solution", J. Chem. Theory Comput. **5**, 2606–2618 (2009).

[163]A. K. Mazur, "Kinetic and thermodynamic DNA elasticity at micro- and mesoscopic scales", J. Phys. Chem. B **113**, 2077–2089 (2009).

[164]C. Bustamante, J. Marko, E. Siggia, and S. Smith, "Entropic elasticity of lambda-phage DNA", Science **265**, 1599–1600 (1994).

[165]D. Jost and R. Everaers, "A unified Poland-Scheraga model of oligo- and polynucleotide DNA melting: salt effects and predictive power", Biophys. J. **96**, 1056–1067 (2009).

[166]R. Everaers, S. Kumar, and C. Simm, "Unified description of poly- and oligonucleotide DNA melting: nearest-neighbor, Poland-Sheraga, and lattice models", Phys. Rev. E **75**, 041918 (2007).

[167]J. SantaLucia Jr. and D. Hicks, "The thermodynamics of DNA structural motifs", Annu. Rev. Biophys. Biomol. Struct. **33**, 415–440 (2004).

[168]T. Garel and H. Orland, "Generalized Poland–Scheraga model for DNA hybridization", Biopolymers **75**, 453–467 (2004).

[169]C. Richard and A. Guttmann, "Poland–Scheraga models and the DNA denaturation transition", J. Stat. Phys. **115**, 925–947 (2004).

[170]T. Dauxois, M. Peyrard, and A. R. Bishop, "Entropy-driven DNA denaturation", Phys. Rev. E **47**, R44–R47 (1993).

[171]M. Peyrard and A. R. Bishop, "Statistical mechanics of a nonlinear model for DNA denaturation", Phys. Rev. Lett. **62**, 2755–2758 (1989).

[172]D. Poland and H. A. Scheraga, "Occurrence of a phase transition in nucleic acid models", J. Chem. Phys. **45**, 1464–1469 (1966).

[173]C. Maffeo, J. Yoo, and A. Aksimentiev, "De novo reconstruction of DNA origami structures through atomistic molecular dynamics simulation", Nucleic Acids Res. **44**, 3013–3019 (2016).

[174]J. Yoo and A. Aksimentiev, "In situ structure and dynamics of DNA origami determined through molecular dynamics simulations", Proc. Natl. Acad. Sci. U. S. A. **110**, 20099–20104 (2013).

[175]K. Pan, D.-N. Kim, F. Zhang, M. R. Adendorff, H. Yan, and M. Bathe, "Lattice-free prediction of three-dimensional structure of programmed DNA assemblies", Nat. Commun. **5**, 5578 (2014).

[176]D.-N. Kim, F. Kilchherr, H. Dietz, and M. Bathe, "Quantitative prediction of 3D solution shape and flexibility of nucleic acid nanostructures", Nucleic Acids Res. **40**, 2862–2868 (2012).

[177]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 Res. **46**, 1102–1112 (2018).

[178]I. Rouzina and V. A. Bloomfield, "Heat capacity effects on the melting of dna. 1. general aspects", Biophys. J. **77**, 3242–3251 (1999).

[179]I. Rouzina and V. A. Bloomfield, "Heat capacity effects on the melting of dna.2. analysis of nearest-neighbor base pair effects", Biophys. J. **77**, 3252–3255 (1999).

[180]R. Owczarzy, B. G. Moreira, Y. You, M. A. Behlke, and J. A. Walder, "Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations", Biochemistry **47**, 5336–5353 (2008).

[181]J. SantaLucia, "A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics", Proc. Natl. Acad. Sci. U. S. A. **95**, 1460–1465 (1998).

[182]H. T. Allawi and J. SantaLucia, "Thermodynamics and nmr of internal g·t mismatches in dna", Biochemistry **36**, 10581–10594 (1997).

[183]B. E. K. Snodin, F. Randisi, M. Mosayebi, P. Šulc, 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. **142**, 234901 (2015).

[184]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. **137**, 135101 (2012).

[185]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. **134**, 085101 (2011).

[186]T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, "DNA nanotweezers studied with a coarse-grained model of DNA", Phys. Rev. Lett. **104**, 178101 (2010).

[187]B. E. K. Snodin, J. S. Schreck, F. Romano, A. A. Louis, and J. P. K. Doye, "Coarse-grained modelling of the structural properties of DNA origami", Nucleic Acids Res. **47**, 1585–1597 (2019).

[188]M. C. Engel, D. M. Smith, M. A. Jobst, M. Sajfutdinow, T. Liedl, F. Romano, L. Rovigatti, A. A. Louis, and J. P. K. Doye, "Force-induced unravelling of DNA origami", ACS Nano **12**, 6734–6747 (2018).

[189]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", J. Am. Chem. Soc. **134**, 9844–9847 (2012).

[190]A. Reinhardt and D. Frenkel, "Numerical evidence for nucleated self-assembly of DNA brick structures", Phys. Rev. Lett. **112**, 238103 (2014).

[191]W. M. Jacobs, A. Reinhardt, and D. Frenkel, "Rational design of self-assembly pathways for complex multicomponent structures", Proc. Natl. Acad. Sci. U. S. A. **112**, 6313–6318 (2015).

[192]A. Reinhardt and D. Frenkel, "DNA brick self-assembly with an off-lattice potential", Soft Matter **12**, 6253–6260 (2016).

[193]M. Sajfutdinow, W. M. Jacobs, A. Reinhardt, C. Schneider, and D. M. Smith, "Direct observation and rational design of nucleation behavior in addressable self-assembly", Proc. Natl. Acad. Sci. U. S. A. **115**, E5877–E5886 (2018).

[194]M. S. Causo, B. Coluzzi, and P. Grassberger, "A simple model for DNA denaturation transition", Phys. A **314**, 607–612 (2002).

[195]J. Gillespie, M. Mayne, and M. Jiang, "RNA folding on the 3D triangular lattice", BMC Bioinf. **10**, 1–17 (2009).

[196]S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski, "Coarse-grained protein models and their applications", Chem. Rev. **116**, 7898–7936 (2016).

[197]A. Bechini, "On the characterization and software implementation of general protein lattice models", PLoS One **8**, e59504 (2013).

[198]K. Binder and W. Paul, "Monte Carlo simulations of polymer dynamics: recent advances", J. Polym. Sci., Part B: Polym. Phys. **35**, 1–31 (1997).

[199]I. Carmesin and K. Kremer, "The bond fluctuation method: a new effective algorithm for the dynamics of polymers in all spatial dimensions", Macromolecules **21**, 2819–2823 (1988).

[200]H. K. Wayment-Steele, D. Frenkel, and A. Reinhardt, "Investigating the role of boundary bricks in DNA brick self-assembly", Soft Matter **13**, 1670–1680 (2017).

[201]D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications* (Academic Press, 2002).

[202]G. H. Givens and J. A. Hoeting, *Computational statistics* (John Wiley & Sons, Inc., 2013).

[203]N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines", J. Chem. Phys. **21**, 1087–1092 (1953).

[204]W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications", Biometrika **57**, 97–109 (1970).

[205]D. P. Landau and K. Binder, *A guide to monte carlo simulations in statistical physics* (Cambridge University Press, 2015).

[206]E. J. J. van Rensburg, "Monte Carlo methods for the self-avoiding walk", J. Phys. A: Math. Theor. **42**, 323001 (2009).

[207]K. Binder and W. Paul, "Recent developments in Monte Carlo simulations of lattice models for polymer systems", Macromolecules **41**, 4537–4550 (2008).

[208]K. Kremer and K. Binder, "Monte Carlo simulation of lattice models for macromolecules", Comput. Phys. Rep. **7**, 259–310 (1988).

[209]M. N. Rosenbluth and A. W. Rosenbluth, "Monte Carlo calculation of the average extension of molecular chains", J. Chem. Phys. **23**, 356–359 (1955).

[210]P. Grassberger, "Pruned-enriched Rosenbluth method: simulations of θ polymers of chain length up to 1 000 000", Phys. Rev. E **56**, 3682–3693 (1997).

[211]H.-P. Hsu and P. Grassberger, "A review of Monte Carlo simulations of polymers with PERM", J. Stat. Phys. **144**, 597–637 (2011).

[212]P. H. Verdier and W. H. Stockmayer, "Monte Carlo calculations on the dynamics of polymers in dilute solution", J. Chem. Phys. **36**, 227–235 (1962).

[213]H. J. Hilhorst and J. M. Deutch, "Analysis of Monte Carlo results on the kinetics of lattice polymer chains with excluded volume", J. Chem. Phys. **63**, 5153–5161 (1975).

[214]S. Caracciolo, "Bilocal dynamics for self-avoiding walks", J. Stat. Phys. **100**, 1111–1145 (2000).

[215]N. Madras, "Nonergodicity of local, length-conserving Monte Carlo algorithms for the self-avoiding walk", J. Stat. Phys. **47**, 573–595 (1987).

[216]N. Madras, "The pivot algorithm: a highly efficient Monte Carlo method for the self-avoiding walk", J. Stat. Phys. **50**, 109–186 (1988).

[217]N. Lesh, M. Mitzenmacher, and S. Whitesides, "A complete and effective move set for simplified protein folding", in Proceedings of the seventh annual international conference on research in computational molecular biology, RECOMB '03 (2003), pp. 188–195.

[218]A. D. Swetnam and M. P. Allen, "Improved simulations of lattice peptide adsorption", Phys. Chem. Chem. Phys. **11**, 2046–2055 (2009).

[219]J. M. Deutsch, "Long range moves for high density polymer simulations", J. Chem. Phys. **106**, 8849–8854 (1997).

[220]T. Wüst and D. P. Landau, "Optimized Wang–Landau sampling of lattice polymers: ground state search and folding thermodynamics of HP model proteins", J. Chem. Phys. **137**, 064903 (2012).

[221]A. C. Farris, T. Wüst, and D. P. Landau, "Statistical physics meets biochemistry: Wang–Landau sampling of the HP model of protein folding", Am. J. Phys. **87**, 310–316 (2019).

[222]J. Houdayer, "The wormhole move: a new algorithm for polymer simulations", J. Chem. Phys. **116**, 1783–1787 (2002).

[223]W. Paul and M. Müller, "Enhanced sampling in simulations of dense systems: the phase behavior of collapsed polymer globules", J. Chem. Phys. **115**, 630–635 (2001).

[224]J. I. Siepmann and D. Frenkel, "Configurational bias Monte Carlo: a new sampling scheme for flexible chains", Mol. Phys. **75**, 59–70 (1992).

[225]S. Consta, N. B. Wilding, D. Frenkel, and Z. Alexandrowicz, "Recoil growth: an efficient simulation method for multi-polymer systems", J. Chem. Phys. **110**, 3220–3228 (1999).

[226]S. Consta, T. J. H. Vlugt, J. W. Hoeth, B. Smit, and D. Frenkel, "Recoil growth algorithm for chain molecules with continuous interactions", Mol. Phys. **97**, 1243–1254 (1999).

[227]N. Combe, T. J. H. Vlugt, P. R. ten Wolde, and D. Frenkel, "Dynamic pruned-enriched Rosenbluth method", Mol. Phys. **101**, 1675–1682 (2003).

[228]K. F. Lau and K. A. Dill, "A lattice statistical mechanics model of the conformational and sequence spaces of proteins", Macromolecules **22**, 3986–3997 (1989).

[229]M. Dijkstra, D. Frenkel, and J.-P. Hansen, "Phase separation in binary hard-core mixtures", J. Chem. Phys. **101**, 3179–3189 (1994).

[230]R. H. Swendsen and J.-S. Wang, "Replica Monte Carlo simulation of spin-glasses", Phys. Rev. Lett. **57**, 2607–2609 (1986).

[231]C. Geyer, "Markov-chain Monte-Carlo maximum-likelihood", in Computing science and statistics, 23rd Symposium on the Interface Between Computing Science and Statistics - Critical Applications of Scientific Computing : Biology, Engineering, Medicine, Speech (1991).

[232]M. C. Tesi, "Monte Carlo study of the interacting self-avoiding walk model in three dimensions", J. Stat. Phys. **82**, 155–181 (1996).

[233]K. Hukushima and K. Nemoto, "Exchange Monte Carlo method and application to spin glass simulations", J. Phys. Soc. Jpn. **65**, 1604–1608 (1996).

[234]D. J. Earl and M. W. Deem, "Parallel tempering: theory, applications, and new perspectives", Phys. Chem. Chem. Phys. **7**, 3910–3916 (2005).

[235]Q. Yan and J. J. de Pablo, "Hyper-parallel tempering Monte Carlo: application to the Lennard-Jones fluid and the restricted primitive model", J. Chem. Phys. **111**, 9509–9516 (1999).

[236]Y. Sugita, A. Kitao, and Y. Okamoto, "Multidimensional replica-exchange method for free-energy calculations", J. Chem. Phys. **113**, 6042–6051 (2000).

[237]H. Fukunishi, O. Watanabe, and S. Takada, "On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: application to protein structure prediction", J. Chem. Phys. **116**, 9058–9067 (2002).

[238]V. I. Manousiouthakis and M. W. Deem, "Strict detailed balance is unnecessary in Monte Carlo simulation", J. Chem. Phys. **110**, 2753–2756 (1999).

[239]M. Lingenheil, R. Denschlag, G. Mathias, and P. Tavan, "Efficiency of exchange schemes in replica exchange", Chem. Phys. Lett. **478**, 80–84 (2009).

[240]M. E. Tuckerman, *Statistical mechanics: theory and molecular simulation* (Oxford University Press, 2010).

[241]G. Torrie and J. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling", J. Comput. Phys. **23**, 187–199 (1977).

[242]J. Kästner, "Umbrella sampling", Wiley Interdiscip. Rev.: Comput. Mol. Sci. **1**, 932–942 (2011).

[243]M. Mezei, "Adaptive umbrella sampling: self-consistent determination of the non-Boltzmann bias", J. Comput. Phys. **68**, 237–248 (1987).

[244]A. M. Ferrenberg and R. H. Swendsen, "Optimized Monte Carlo data analysis", Phys. Rev. Lett. **63**, 1195–1198 (1989).

[245]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", J. Comput. Chem. **13**, 1011–1021 (1992).

[246]M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states", J. Chem. Phys. **129**, 124105 (2008).

[247]J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, and K. A. Dill, "Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations", J. Chem. Theory Comput. **3**, 26–41 (2007).

[248]J. D. Chodera, "A simple method for automated equilibration detection in molecular simulations", J. Chem. Theory Comput. **12**, 1799–1805 (2016).

[249]E. Protozanova, P. Yakovchuk, and M. D. Frank-Kamenetskii, "Stacked–unstacked equilibrium at the nick site of DNA", J. Mol. Biol. **342**, 775–785 (2004).

[250]P. Yakovchuk, E. Protozanova, and M. D. Frank-Kamenetskii, "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix", Nucleic Acids Res. **34**, 564–574 (2006).

[251]R. Kosinski, A. Mukhortava, W. Pfeifer, A. Candelli, P. Rauch, and B. Saccà, "Sites of high local frustration in DNA origami", Nat. Commun. **10**, 1061 (2019).

[252]K. Röder, J. A. Joseph, B. E. Husic, and D. J. Wales, "Energy landscapes for proteins: from single funnels to multifunctional systems", Adv. Theory Simul. **2**, 1800175 (2019).

[253]Y. Ke, G. Bellot, N. V. Voigt, E. Fradkov, and W. M. Shih, "Two design strategies for enhancement of multilayer-DNA-origami folding: underwinding for specific intercalator rescue and staple-break positioning", Chem. Sci. **3**, 2587–2597 (2012).

[254]P. Fonseca, F. Romano, J. S. Schreck, T. E. Ouldridge, J. P. K. Doye, and A. A. Louis, "Multiscale coarse-graining for the study of assembly pathways in DNA-brick self-assembly", J. Chem. Phys. **148**, 134910 (2018).

[255]A. E. Marras, L. Zhou, V. Kolliopoulos, H.-J. Su, and C. E. Castro, "Directing folding pathways for multi-component DNA origami nanostructures with complex topology", New J. Phys. **18**, 055005 (2016).

[256]A. Reinhardt, C. P. Ho, and D. Frenkel, "Effects of co-ordination number on the nucleation behaviour in many-component self-assembly", Faraday Discuss. **186**, 215–228 (2016).

[257]C. Steinhauer, R. Jungmann, T. L. Sobey, F. C. Simmel, and P. Tinnefeld, "DNA origami as a nanoscopic ruler for super-resolution microscopy", Angew. Chem., Int. Ed. **48**, 8870–8873 (2009).

[258]P. Shrestha, T. Emura, D. Koirala, Y. Cui, K. Hidaka, W. J. Maximuck, M. Endo, H. Sugiyama, and H. Mao, "Mechanical properties of DNA origami nanoassemblies are determined by Holliday junction mechanophores", Nucleic Acids Res. **44**, 6574–6582 (2016).

[259]J. Song, Z. Li, P. Wang, T. Meyer, C. Mao, and Y. Ke, "Reconfiguration of DNA molecular arrays driven by information relay", Science **357**, eaan3377 (2017).

[260]Y. Cui, R. Chen, M. Kai, Y. Wang, Y. Mi, and B. Wei, "Versatile DNA origami nanostructures in simplified and modular designing framework", ACS Nano **11**, 8199–8206 (2017).

[261]C. Lee, J. Y. Lee, and D.-N. Kim, "Polymorphic design of DNA origami structures through mechanical control of modular components", Nat. Commun. **8**, 2067 (2017).

[262]F. Wang and D. P. Landau, "Efficient, multiple-range random walk algorithm to calculate the density of states", Phys. Rev. Lett. **86**, 2050–2053 (2001).

[263]F. Wang and D. P. Landau, "Determining the density of states for classical statistical models: a random walk algorithm to produce a flat histogram", Phys. Rev. E **64**, 056101 (2001).

[264]D. P. Landau, S.-H. Tsai, and M. Exler, "A new approach to Monte Carlo simulations in statistical physics: Wang–Landau sampling", Am. J. Phys. **72**, 1294–1302 (2004).

[265]S. Singh, M. Chopra, and J. J. de Pablo, "Density of states–based molecular simulations", Annu. Rev. Chem. Biomol. Eng. **3**, 369–394 (2012).

[266]N. Boon, "Efficient configurational-bias Monte-Carlo simulations of chain molecules with "swarms" of trial configurations", J. Chem. Phys. **149**, 064109 (2018).

[267]D. Frenkel, "Speed-up of Monte Carlo simulations by sampling of rejected states", Proc. Natl. Acad. Sci. U. S. A. **101**, 17571–17575 (2004).