

Lattice models and Monte Carlo methods for simulating DNA origami self-assembly

Alexander Cumberworth,¹ Aleks Reinhardt,¹ and Daan Frenkel¹

Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom

The optimal design of DNA origami systems that assemble rapidly and robustly is hampered by the lack of a model for self-assembly that is sufficiently detailed yet computationally tractable. Here, we propose a model for DNA origami that strikes a balance between these two criteria by representing these systems on a lattice at the level of binding domains. The free energy of hybridization between individual binding domains is estimated with a nearest-neighbour model. Double helical segments are treated as rigid rods, but we allow flexibility at points where the backbone of one of the strands is interrupted, which provides a reasonably realistic representation of partially and fully assembled states. Particular attention is paid to the constraints imposed by the double helical twist, as they determine where strand crossovers between adjacent helices can occur. To improve the efficiency of sampling configuration space, we develop Monte Carlo methods for sampling scaffold conformations in near-assembled states, and we carry out simulations in the grand canonical ensemble, enabling us to avoid considering states with unbound staples. We demonstrate that our model can quickly sample assembled configurations of a small origami design previously studied with the oxDNA model, as well as a design with staples that span longer segments of the scaffold. The sampling ability of our method should allow for good statistics to be obtained when studying the assembly pathways, and is suited to investigating in particular the effects of design and assembly conditions on these pathways and their resulting final assembled structures.

I. INTRODUCTION

In the early 1980s, Seeman¹ demonstrated that DNA, the carrier of genetic information, could also 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. But it was not until the seminal paper of Rothemund,² which introduced the DNA origami method, that the complexity of structural DNA nanotechnology systems really took off. 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 by hybridising a number of designed, shorter ‘staple’ strands that link selected binding domains on the scaffold strand.² A fully assembled origami structure is largely made up of parallel double helices with strand crossovers between adjacent parallel helices, which provide structural stability.

While early designs were mostly simple planar structures, it is now possible to design and assemble virtually any connected 3D shape, even structures with flexible joints.³ This variety of design, and the ability to functionalize the staple strands individually with other molecules, opens up DNA origami to applications including chemical sensors,⁴ nanoreactors,⁵ electronic devices⁶ and drug delivery vehicles.⁷ 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.

The factors that determine the kinetics of origami formation are different from those that determine, say, the formation of an ordered crystal. In fact, Cademartiri and Bishop⁸ differentiate between two fundamentally different types of self-assembly: the *puzzle* mechanism and the *folding* mechanism. 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,⁹

the interactions between components must be highly specific (e.g. in DNA bricks¹⁰). 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 of DNA origami is a cooperative process.¹¹ 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.¹² However, Dannenberg *et al.*¹³ found that coaxial stacking between staples adjacent to each other on the scaffold may also increase cooperativity of the assembly process. Both forms of cooperativity would be expected to have primarily a local effect and indeed FRET experiments¹⁴ and simulations¹³ have found that excluding a staple from the reaction mixture only affects the binding of nearby staples in the assembled structure.

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,^{12–15} with annealing occurring at lower temperatures than melting; the effect seems to be stronger in 3D origami structures.¹⁴ Increasing the heating/cooling rate has been found to have a more pronounced effect on annealing than on melting,¹³ 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.*¹⁶ examined intermediates of the assembly process with atomic force microscopy (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.*¹² 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 modification that instead changed the stabilities of some of the intermediate states in the assembly pathways, leaving the stabilities of the assembled configurations unchanged.

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. 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. As we argue below, all existing approaches have their drawbacks, either providing too little information about the pathway, or being too slow for large origamis.

In the lower-resolution models, the standard approach is to model DNA origami self-assembly by extending the SantaLucia Jr. and Hicks¹⁷ nearest-neighbour (NN) model of DNA hybridization to account for the entropic effects of folding the scaffold. Using this approach, Arbona, Aimé and Elezgaray¹⁵ modeled 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. Dannenberg *et al.*¹³ and Dunn *et al.*¹² 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. These approaches allow the assembly process to be simulated in under an hour on current computers. 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, these models ignore the possibility that staples may bind (albeit less strongly) to incorrect binding domains.

In the higher resolution approach, Snodin *et al.*¹⁸ used a coarse-grained model of DNA known as oxDNA^{19–22} to model the self-assembly of a small DNA origami. 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.

Here, we bridge the gap between these two approaches for simulating the assembly of DNA origami by introducing a model that aims to combine the low computational cost associated with the SantaLucia model with the structural information provided by the oxDNA model.

II. MODEL

A. Model description

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 represent such binding domains as particles on a lattice. We chose a lattice representation in order to increase computational efficiency by reducing configuration space. Further, the use of a lattice representation for structural DNA nanotechnology systems has good precedent: a lattice model of DNA bricks was remarkably effective,²³ and unexpectedly yielded near quantitative agreement with experimental measurements of the nucleation kinetics.²⁴ For simplicity, in this work we consider origami designs in which the angles between helical axes involve only angles that are multiples of 90° in the final structure, so we chose to work with a simple cubic lattice. Contiguous binding domains on a given chain are constrained to occupy adjacent lattice sites. 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 (fig. 1(a)). 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.

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 SantaLucia NN model,¹⁷ which are a function of both temperature and salt 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.^{23,25}

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 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. Suppose two particles in our

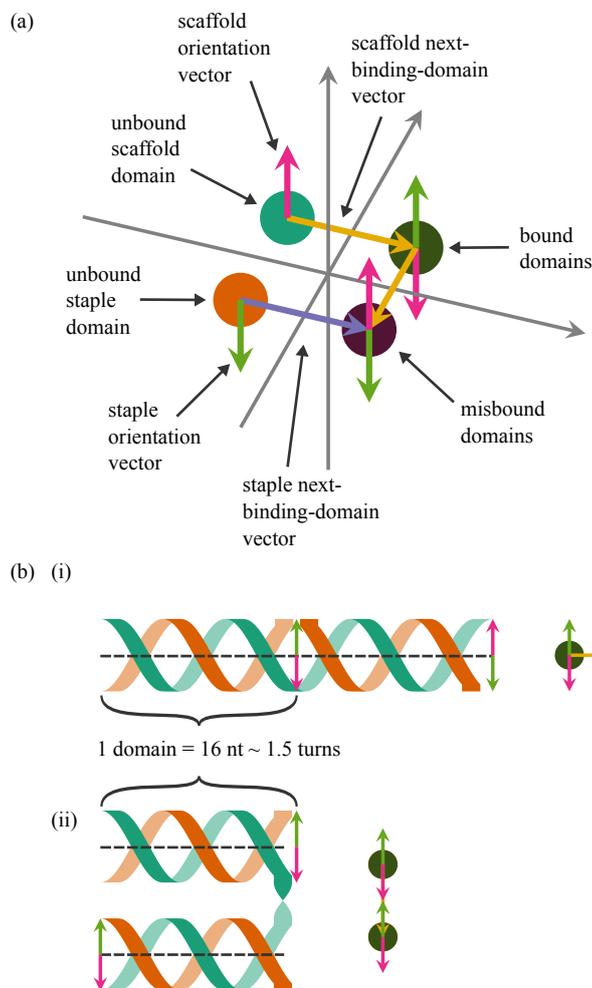


Figure 1. Schematic illustrations of the basic elements of the model. (a) Occupied lattice states. There is one scaffold with three binding domains, one staple with two binding domains, and one staple with a single binding domain. All vectors shown are unit vectors. (b) Two binding domains of a scaffold and a staple in both an idealized cartoon helix representation and the corresponding lattice-model representation. The binding domains are 16-nt long, 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; this particular kinked configuration is also that of a strand cross-over between two adjacent parallel helices.

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-nt binding domains, as in fig. 1(b)(i). At the end of the first (leftmost) 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

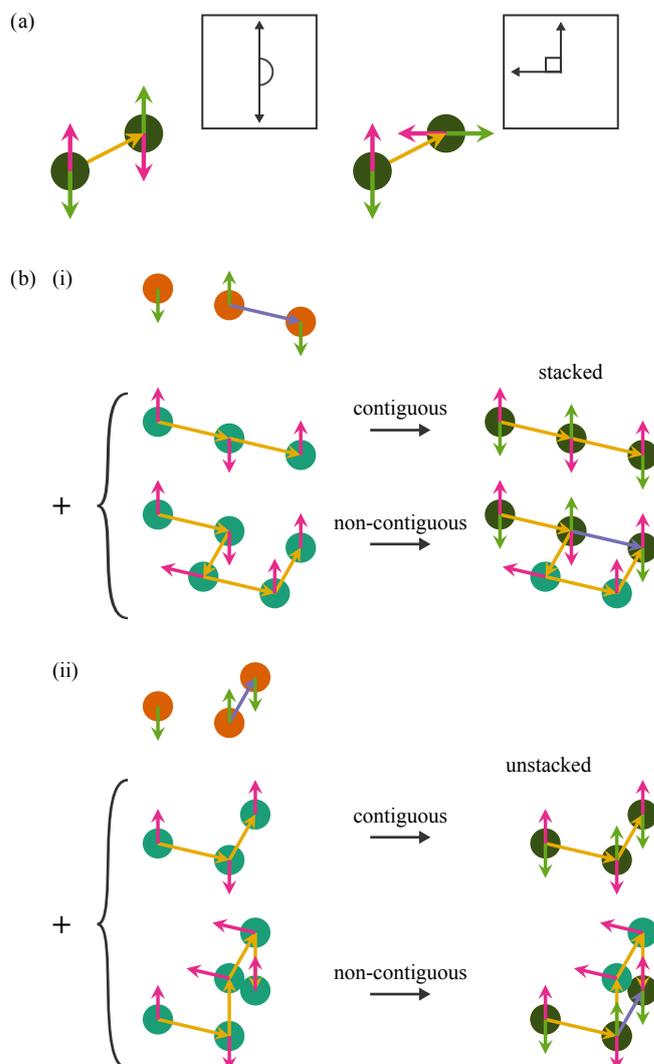


Figure 2. Helical stacking in the model. (a) Orientation vectors and helical phase. The boxes are projections of the scaffold orientation vectors of the two binding domains onto a plane normal to the next-binding-domain vector, with the dihedral angle indicated. Left: the orientation vectors are consistent 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. (b) Three bound domain pairs in which each pairing of contiguous domains along the strand is in a stacked configuration (all binding domains are 16 nt in length). The staples and scaffold are shown separately for clarity, and a scaffold configuration is given for both the case in which there is and the case in which there is not one contiguous segment of a chain forming the three bound domain pairs. (i) Stacked triplet configurations. (ii) Unstacked triplet configurations.

orientation vectors and the vector connecting the two domains must be determined by the number of turns of the helix between them (fig. 2(a)). 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. 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’.

The dihedral angle that we expect depends on the length of the binding domains. For example, in the case we considered above (fig. 1(b)), each binding domain corresponds to 1.5 turns of the helix. If two adjacent bound domains are in the same helix, we therefore expect a dihedral angle of 180° ; the orientation vectors of the scaffold (pink) and staple (green) strands in fig. 1(b)(i) must therefore alternate in sign when they

are part of the same helix. We also note that if two contiguous domains are in bound states and part of the same helix, then there is an additional ‘stacking’ interaction (see section II B for further discussion of this interaction) that we have not yet accounted for when calculating the NN model hybridization free energy for the two domains separately.

The length of the binding domains is typically well below the persistence length of double stranded DNA, so we need to ensure that helices are rigid in this model. Because there is no explicit helical axis vector in the model, a single binding domain pair will only implicitly define the helical axis to lie within a plane. The helical axis is not resolved until an adjacent binding domain enters a bound state in the same helix. Explicit checks of helical rigidity are thus only necessary once we consider triplets of bound domain pairs. Consider two pairs of adjacent lattice sites with one site in common, where all three sites are in bound states. If for each pair, the two sites are occupied by two contiguous binding domains of a chain, and the two contiguous bound pairs satisfy the same helix geometry, then in order for all three bound domain pairs to be in the same helix, the lattice sites must all lie on a common line (fig. 2(b)(i)). These two pairs of two contiguous binding domains could either be a single contiguous segment of three domains, or they could be two separate segments that are bound to each other at the middle lattice site.

The above considerations do not apply if one of the domains is misbound, because regions of ssDNA, which has a much lower persistence length, are then present between the helical sections. Points in a helix at which there is a break in the backbone of one of the strands, which we will refer to as breakpoints, are also more flexible, and are able to become unstacked to form kinks. We account for the flexibility at breakpoints in our model by allowing such kinks to form. 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 with no stacking interaction (for an example see fig. 1(b)(ii), which also happens to be a crossover configuration). If a triplet of bound domains of the type described above does not meet the conditions necessary for it to be a single helix, then the configuration is likewise considered to have a kink between one of the contiguous binding domain pairs (fig. 2(b)(ii)). These configurations will have one less stacking interaction than configurations in which all three bound domain pairs are in the same helix.

Consider a pair of contiguous bound domains with a kink between them. In reality, the kink 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 two further rules (see appendix 1 for detailed arguments). First, 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 the next-binding-domain vector of the first. Second, if the first binding domain’s next-binding-domain vector is parallel to the first binding domain’s orientation vector but not parallel to the second binding domain’s orientation vector, or if the first binding domain’s next-binding-domain vector is antiparallel to its orientation vector, then the configuration is disallowed. If both of the bound domains with a kink between them have a defined helical axis because of an adjacent bound domain (on either the chain with the kink or on one of the chains bound to the domains forming the kink), a third rule is applied: if the two helical axes are parallel or antiparallel to each other and orthogonal to the next-binding-domain vector between the

binding domains forming the kink, then there is an additional kink present amongst these four bound domains and thus one less stacking interaction. These rules also happen to ensure that the first aspect of the challenge to reproduce the constraints of the double helical twist is met, namely that crossovers only occur at certain intervals between parallel helices.

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 (fig. 3(a)). 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.

In reality, 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 fig. 3(b)(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 fig. 3(b)(ii). However, when a binding domain pair is involved in a double crossover as discussed above, which can occur in binding domains with a length of 16 nt, this will not be captured by the model as currently defined (fig. 3(b)(iii)).

To impose the parallel helices constraint on the system in the case of a double crossover between two binding-domain pairs, one must consider only cases where there are two additional contiguous binding domains on one of the chains involved, one before and one after the pair forming the double crossover, where both have orientation vectors that allow them to be in the same helix as the relevant binding domain they are contiguous with. Then, in order for the pairings before and after the crossover to both be in the same helix, the next-binding-domain vector of the first binding domain and the next-binding-domain vector of the third binding domain must be antiparallel, or equivalently, the first and last binding domains must occupy adjacent lattice sites. If this is not the case, the configuration has one more kink, and thus one less stacking interaction, than configurations where these criteria are met. Further, configurations where the first and last binding domains are on opposing sides of the junction, or more precisely, where the next-binding-domain vectors are parallel, are sterically prohibited, and so are made disallowed configurations in our model.

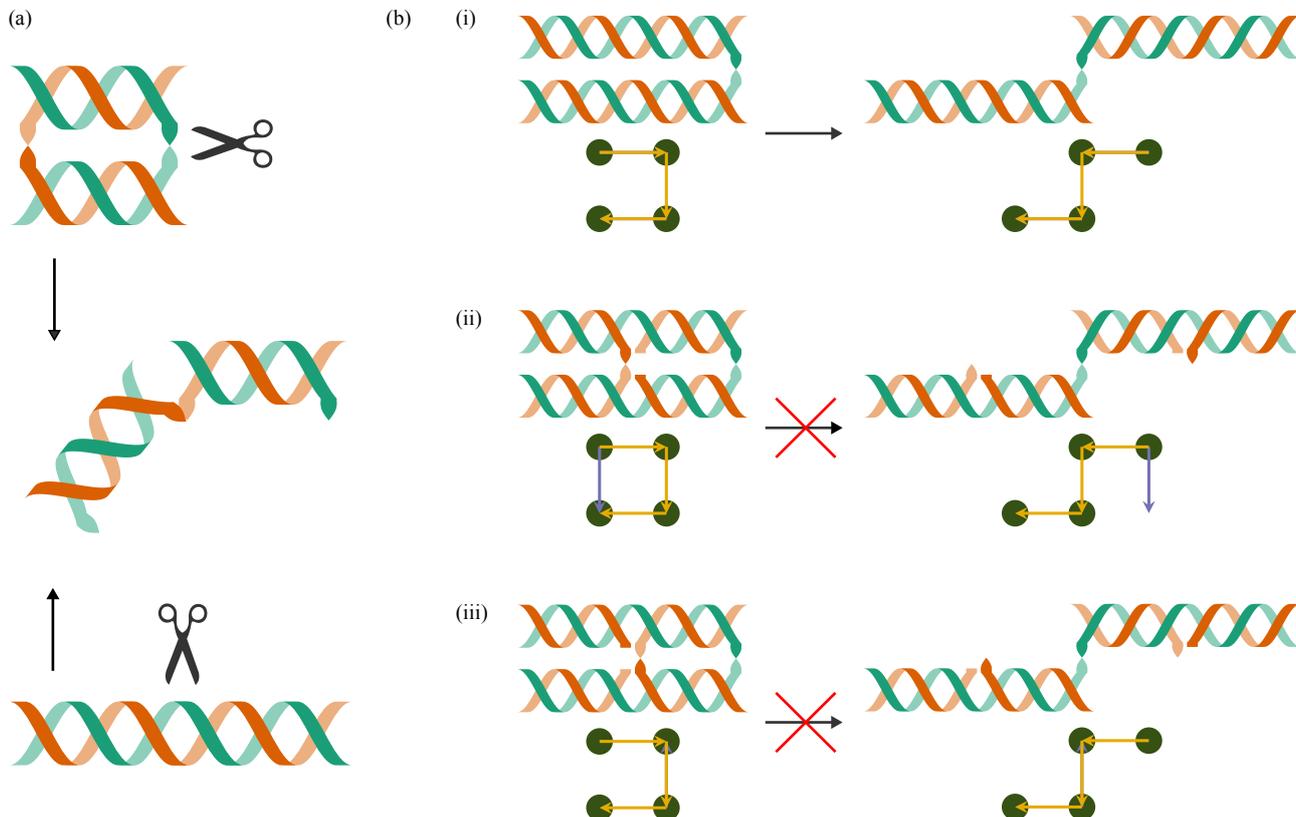


Figure 3. Strand crossovers between helices involving 16-nt long binding domains. For clarity, the orientation vectors have been omitted. (a) Doubly contiguous binding domains in bound states. (b) Crossovers between two adjacent parallel helices with a four-binding-domain scaffold. Red crosses over transformation arrows imply that the transformation cannot be achieved because a covalent bond would need to be broken. (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. 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 nmol L^{-1} , which is the value we use here. When mapping this concentration to the chemical potential for simulations in the grand canonical ensemble, in order to be consistent with hybridization experiments, we assume that we can write the staple-strand chemical potential to within a constant as $\mu_i = k_B T \ln(C/C^\circ)$, where $C^\circ = 1 \text{ mol L}^{-1}$. Assembly is typically carried out in solutions with significant amounts of dissolved salts, so we set the monovalent cation concentration to be 0.5 mol L^{-1} , which is the same concentration as that used by Snodin *et al.*¹⁸ in their oxDNA simulations. As we are running replica exchange Monte Carlo (REMC) simulations (section III), we do not have to select a single temperature; the range is selected to encompass the region of the relevant order parameter curves with the steepest gradient, approximately centred on the melting temperature.

The physical origin of the stacking interactions is the attractive π - π interactions between the aromatic rings of the bases of adjacent nucleotides. Since stacked states have a constrained geometry, they are entropically disfavoured. 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 (fig. S1). We selected a value of $-1000 k_B K$, which gave a free-energy difference that roughly matched experimentally measured values²⁶ of $\sim -4 \text{ kJ mol}^{-1}$.

In the next section, we describe the Monte Carlo scheme that we have used to simulate origami self-assembly. It turns out that, in order to achieve good efficiency, we need to use a combination of various more advanced sampling schemes that we describe below. However, readers who are primarily interested in the results of our simulations may prefer to go directly to section IV.

III. MONTE CARLO METHODS

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 canonical

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.

In order to increase sampling efficiency, we use REMC.²⁷ This advanced sampling method involves running multiple replicas of the simulation that differ only with respect to a few of the simulation parameters. 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. 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 variable that distinguishes different replicas. 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.²⁸

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. To meet the challenge, we developed several Monte Carlo (MC) move types, which we outline in the following subsections. While they are described as applied to the current DNA origami model, these move types are applicable in general to polymer lattice models.

A. Monte Carlo move types

All move types described below are dynamic (i.e. the move types use the current configuration to generate a trial configuration, thus producing configurations that are correlated but not necessarily in a way that mimics the dynamics of real origami assembly) and obey detailed balance. Apart from those move types that only modify orientation vectors, they involve sequential growth of a set of binding domains. If the move involves regrowth of binding domains already present in the system volume, they are first unassigned (i.e. the lattice sites become unoccupied) before regrowth begins. The order in which they are grown depends on the specifics of the move type, but 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. Growth requires selection of a position on a neighbouring lattice site and an orientation vector. The position can be decomposed into the sum of the position of the lattice site being grown from and a unit vector, which we refer to as the position difference vector.

Growth is done with three main variants: symmetric, configurational bias (CB) and recoil growth (RG). In the symmetric variants, 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 probability,

$$p_{\text{acc}}(\mathbf{x} | \mathbf{y}) = \min[1, e^{-\beta\Delta U(\mathbf{x},\mathbf{y})}], \quad (1)$$

where $\Delta U(\mathbf{x}, \mathbf{y})$ is the change in energy from the old configuration \mathbf{y} to the new configuration \mathbf{x} , and β is the inverse thermodynamic temperature.

In the CB variants,²⁹ 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

$$p_i^{\text{trial}} = \frac{e^{-\beta\Delta\varepsilon_{i,j}}}{\sum_{j'}^k e^{-\beta\Delta\varepsilon_{i,j'}}}, \quad (2)$$

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 $\Delta\varepsilon_{i,j}$ is the energy change 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

$$p_{\text{acc}}(\mathbf{x} | \mathbf{y}) = \min\left[1, \frac{W_{\text{new}}}{W_{\text{old}}}\right], \quad (3)$$

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

$$W = \prod_i^n w_i = \prod_i^n \left(\sum_{j=1}^k e^{-\beta\Delta\varepsilon_{i,j}} \right). \quad (4)$$

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

Finally, in the RG variants,^{30,31} 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

$$p_{i,j}^{\text{open}} = \min[1, e^{-\beta\Delta\varepsilon_{i,j}}]. \quad (5)$$

If the configuration is chosen to be closed, another is proposed, up to a total of k_{max} configurations. 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

$$W = \prod_i^n w_i = \prod_i^n \left(\frac{m_{i,j}}{p_{i,j}^{\text{open}}} \right). \quad (6)$$

Then, the move is accepted with

$$p_{\text{acc}}(\mathbf{x} | \mathbf{y}) = \min \left[1, e^{-\beta \Delta U(\mathbf{x}, \mathbf{y})} \frac{W_{\text{new}}}{W_{\text{old}}} \right]. \quad (7)$$

A super-detailed balance argument is used by Consta *et al.*³⁰ to show that detailed balance is obeyed with this acceptance probability.

1. Orientation vector 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 eq. (1).

2. 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 III A 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 $1/b_j$, where b_j is the number of binding domains on staple j (mis)bound to other chains. This comes from the selection of a binding domain to grow out from. As b_j can change between the current and trial configuration, there is an additional factor of $b_j^{\text{old}}/b_j^{\text{new}}$ in the acceptance probability. In the case of CB, this gives

$$p_{\text{acc}}(\mathbf{x} | \mathbf{y}) = \min \left[1, \frac{W_{\text{new}}}{W_{\text{old}}} \times \frac{b_j^{\text{old}}}{b_j^{\text{new}}} \right]. \quad (8)$$

3. 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 $1/n_j$ in the trial probability, where n_j is the length of staple j . Because when inserting into a lattice site with an unbound domain there is only one orientation that will not lead to rejection, this orientation can be selected immediately. The staple is then grown out from this binding domain, which gives a further factor of $1/6^{2(n_j-1)}$ to the trial probability. However, states which involve binding of multiple binding domains to other chains will be over-counted with the current scheme, as there are b_j ways to grow these configurations. This can be corrected by multiplying the trial probability by a factor of b_j . Altogether, the trial probability of insertion for staple type i is

$$p_{\text{trial}}(\mathbf{x}, N_i + 1 | \mathbf{y}, N_i) = \frac{b_j}{6^{2(n_j-1)} n_j V_{\text{sys}}}, \quad (9)$$

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 III A 2), which gives a trial probability of

$$p_{\text{trial}}(\mathbf{x}, N_i - 1 | \mathbf{y}, N_i) = \frac{1}{N_i}. \quad (10)$$

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

$$p_{\text{acc}}(\mathbf{x}, N_i + 1 | \mathbf{y}, N_i) = \min \left[1, \frac{6^{2(n_j-1)} n_j V_{\text{sys}} e^{\beta \mu_i} e^{-\beta \Delta U(\mathbf{x}, \mathbf{y})}}{b_j (N_i + 1)} \right], \quad (11)$$

while for deletion it is

$$p_{\text{acc}}(\mathbf{x}, N_i - 1 | \mathbf{y}, N_i) = \min \left[1, \frac{b_j N_i}{6^{2(n_j-1)} n_j V_{\text{sys}}} e^{-\beta \mu_i} e^{-\beta \Delta U(\mathbf{x}, \mathbf{y})} \right], \quad (12)$$

where μ_i is the chemical potential of staple type i .

4. Scaffold regrowth moves

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 a similar number of (mis)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.³² 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, $\chi_l(\Delta\mathbf{r}_{l,j}, n_{l,i})$, that is unity if walks remain and zero otherwise, where $\Delta\mathbf{r}_{l,j}$ 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, $\chi_B(s)$, 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. Further, 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

$$P_i^{\text{trial}} = \frac{e^{-\beta\Delta\varepsilon_{i,j}} \chi_B(s_j) \prod_l \chi_l(\Delta\mathbf{r}_{l,j}, n_{l,i})}{\sum_{j'}^k e^{-\beta\Delta\varepsilon_{i,j'}} \chi_B(s_{j'}) \prod_{l'} \chi_{l'}(\Delta\mathbf{r}_{l',j'}, n_{l',i'})}, \quad (13)$$

and the Rosenbluth weight is

$$w_i = \sum_{j=1}^k e^{-\beta\Delta\varepsilon_{i,j}} \chi_B(s_j) \prod_l \chi_l(\Delta\mathbf{r}_{l,j}, n_{l,i}). \quad (14)$$

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,

$$P_{i,j}^{\text{open}} = \min[1, e^{-\beta\Delta\varepsilon_{i,j}} \chi_B(s_j) \prod_l \chi_l(\Delta\mathbf{r}_{l,j}, n_{l,i})]. \quad (15)$$

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. 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.

If the scaffold segment was regrown fully before regrowing any of the staples to be regrown, it would result in binding domains on the scaffold being used in endpoint constraints for the staples to be regrown. 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. 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.

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 possible total number of binding domains to regrow across all segments is chosen with uniform probability, where the range from which the selection 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 scaffold 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 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 possible 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 possible 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 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 possible 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 segments of scaffold binding domains does not depend on the conformation or on whether or not binding domains are mis-bound to other binding domains on the same chain, the move type obeys detailed balance.

In appendix 2, we provide the details of our numerical validation, and in appendix 3, we discuss the optimization of the move sets.

IV. RESULTS AND DISCUSSION

To test the efficacy of our model, we ran simulations of a 24-binding-domain scaffold system previously studied with the oxDNA model¹⁸ (fig. 4). This system has 12 staple types that bind to the scaffold, each with two 16-nt binding domains. The simulations were run 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 (mis)bound to the scaffold, and the number of bound domain pairs that have formed. As can be seen in fig. 5(a), at low temperatures the

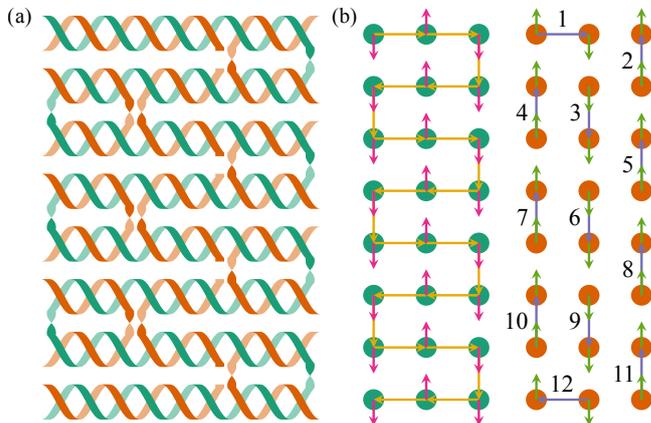


Figure 4. Schematic representations of the 24-binding-domain scaffold system. (a) Helical cartoon representation of the system in an assembled, planar configuration. (b) Representation of the system with the proposed model. The scaffold (left) and staples (right, numbered) are shown in the assembled, planar configuration, but for clarity have been drawn separately.

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 fig. 5(b), we show a typical assembled configuration. Unlike its schematic representation in fig. 4, 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.¹⁸

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 fig. 5, 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 time series of the number of stacked binding domain pairs (fig. 6) reveals that, while the average number of stacked binding domain pairs is below that of a fully stacked assembled 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. To give an idea of how the simulations proceed, we have compiled a short video of a serial simulation of the assembly at 335 K (see supplementary information (SI)).

The efficiency of our model and sampling methods allows us to run simulations across a range of assembly conditions and design parameters. While 100 nmol L^{-1} is a typical value for staple concentrations, the concentration used for a particular assembly protocol commonly varies from tens to hundreds of nmol L^{-1} . 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 nmol L^{-1} to 1 mmol L^{-1} in intervals of factors of 10. As can be seen in fig. 7, at low temperatures, from 1 nmol L^{-1} to $1 \text{ } \mu\text{mol L}^{-1}$, the order parameters indicate that the assembled

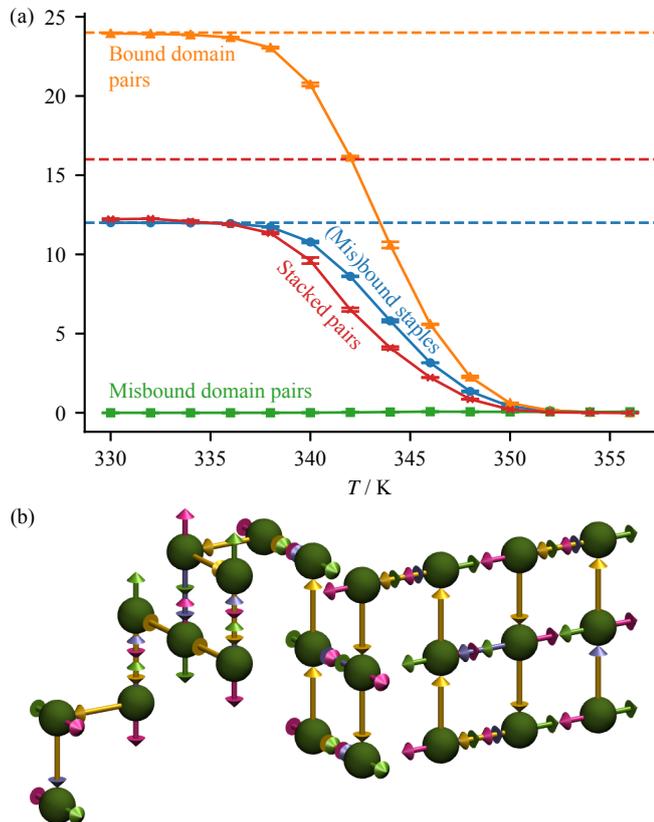


Figure 5. Mean order parameters as a function of system temperature and a typical assembled configuration of a 24-binding-domain scaffold system. Simulations were run with a staple concentration of 100 nmol L^{-1} , a monovalent cation concentration of 0.5 mol L^{-1} , and a stacking energy of $-1000 \text{ k}_B \text{ K}$. (a) Mean order parameters plotted against temperature. The dashed lines correspond to the expected order parameter values in the assembled or fully stacked assembled configurations. The error bars represent the standard error in the means across three independent simulations. (b) An assembled configuration at 330 K.

state is the prevalent structure, with the melting temperature shifting to higher values as the concentration is increased. However, at $10 \text{ } \mu\text{mol L}^{-1}$, 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 type of staple (mis)bound to the scaffold. This situation is referred to as ‘blocking’¹⁸ 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.*,¹⁸ who speculated that they seemed to be in a range in which blocking was somewhat favourable. Blocking becomes substantially more prominent at $100 \text{ } \mu\text{mol L}^{-1}$, 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 $100 \text{ } \mu\text{mol L}^{-1}$, misbinding begins to play a significant role, and becomes even more substantial in the mmol L^{-1} regime.

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 fig. 8, the average number of stacked binding domain pairs changes quite dramatically when the stacking energy is halved

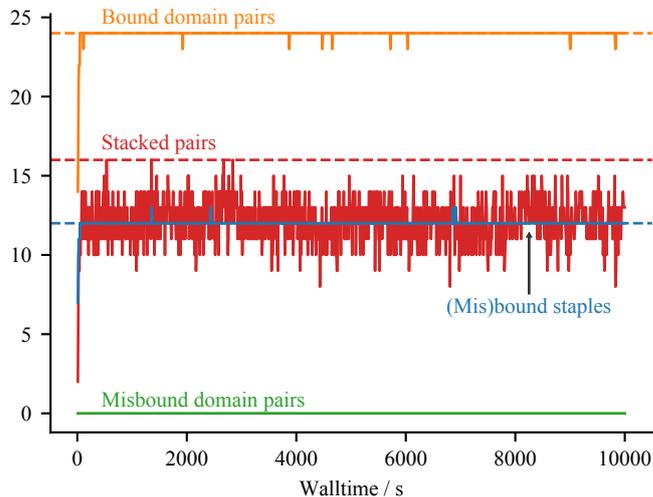


Figure 6. Order parameter time series for a 330 K replica of a REMC simulation of a 24-binding-domain scaffold. The simulations were run on a single node of a commodity cluster. The dashed lines correspond to the expected order parameter values in the assembled or fully stacked assembled configurations.

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 exceeds that expected in the assembled state. It seems that if the stacking energy is sufficiently favourable, the sequence specificity of hybridization can become overshadowed by the non-specific 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 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 fig. S2). 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; however, since the effect is relatively small given our model’s intended level of accuracy, we have not included such corrections in our model at present.

In order to see how the thermodynamic behaviour of individual staples is affected 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 fig. 9, 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 are those that have no crossovers (same helix; see

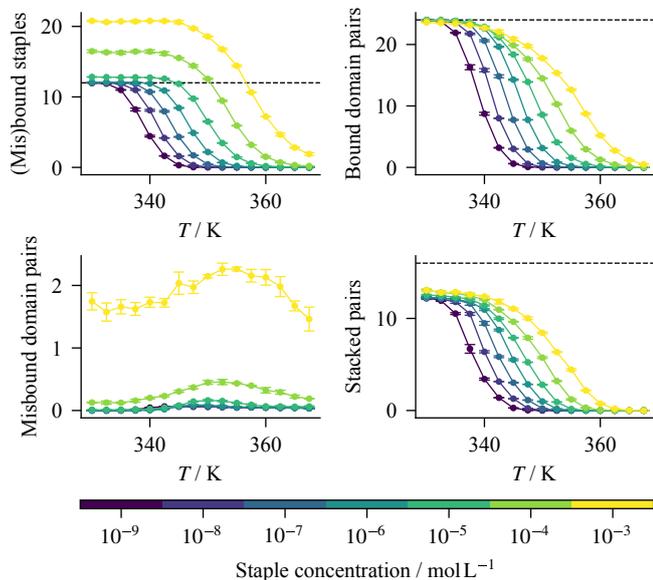


Figure 7. Mean order parameters plotted against temperature for a range of staple concentrations. The dashed lines correspond to the expected order parameter values in the assembled or fully stacked assembled configurations. The error bars represent the standard error in the means across three independent simulations. Simulations were run with a monovalent cation concentration of 0.5 mol L^{-1} and a stacking energy of $-1000 k_B K$.

fig. 4(b), staples 1 and 12), followed by those that span two scaffold binding domains (span-2; see fig. 4(b), staples 3, 6, and 9), followed by those that span no scaffold binding domains (span-0; see fig. 4(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 involve 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 it is expected that these double crossover bound domains pairs will have a lower melting temperature than those with just one strand crossover. 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 fig. 4(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 fig. 4(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

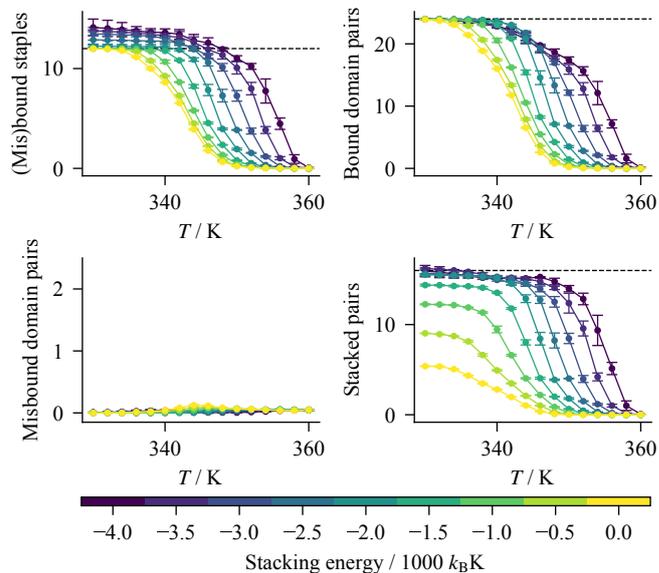


Figure 8. Mean order parameters plotted against temperature for a range of stacking energies. The dashed lines correspond to the expected order parameter values in the assembled or fully stacked assembled configurations. The error bars represent the standard error in the means across three independent simulations. Simulations were run with a staple concentration of 100 nmol L^{-1} and a monovalent cation concentration of 0.5 mol L^{-1} .

system with such a favourable stacking energy (fig. 8) suggests that this would make the system on average nearly planar, which contradicts the simulations of Snodin *et al.*¹⁸ 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.

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 (fig. S3). The fully stacked assembled state involves three parallel helices composed of seven binding domains each; the design is a subset of the rectangular tile design presented in the original DNA origami paper² (the top three rows on the left side of the seam). The assembly of this system is further complicated by the presence of single-domain staples, which are expected not to bind until significantly lower 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 fig. 10. As with the 24-binding-domain 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.

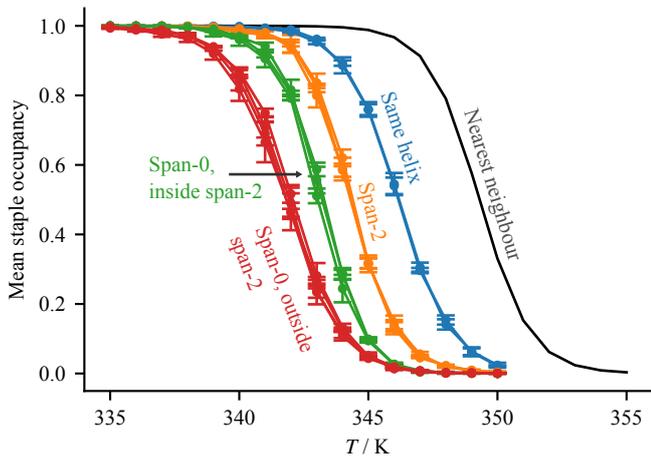


Figure 9. Mean staple occupancy plotted against temperature for simulations with averaged hybridization free energies. The NN values were calculated directly with the averaged hybridization free energies by assuming the staple strands are in excess of the scaffold strands. The error bars represent the standard error in the means across three independent simulations. Simulations were run with the same parameters as the simulations referred to in fig. 5.

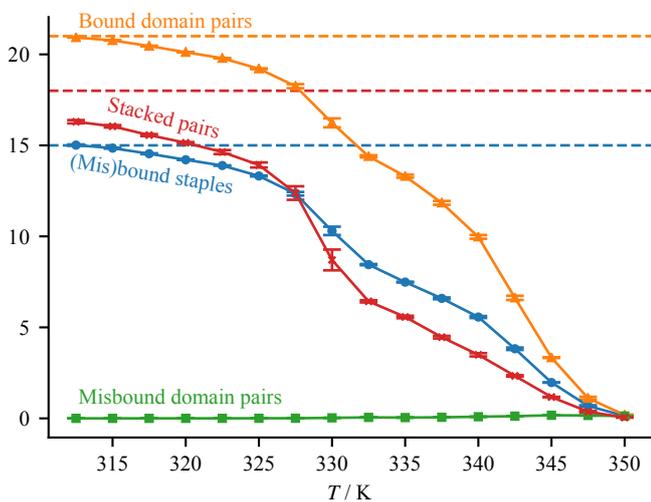


Figure 10. Mean order parameters plotted against temperature for a 21-binding-domain scaffold. The dashed lines correspond to the expected order parameter values in the assembled or fully stacked assembled configurations. The error bars represent the standard error in the means across three independent simulations. Simulations were run with the same parameters as the simulations referred to in fig. 5.

V. 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 Monte Carlo 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. 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.

It has been found experimentally that the stacking free energy is sequence specific²⁶ and depends on both temperature and salt concentration. It has been observed to range from below -10 kJ mol^{-1} to slightly above 1 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 the results section, 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 modeling of some of the entropic contribution to the stacking free energy, the sequence specificity and salt dependence is 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.

Here, we have only considered binding domains that are 16 nt in length. Because of the level of flexibility we assume at the crossovers, it is possible for some fully stacked configurations with only 16 nt to be non-planar. However, to model structures that are intended to be assembled into explicitly 3D structures, it will be necessary to extend the model to other lengths of binding domains. This will primarily entail constructing a stacking potential for each binding domain length considered. We intend to pursue this in future work.

Finally, we have used 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 Hamiltonian to weight kinked configurations based on their frequency in the higher resolution simulations.

However, because these improvements would only be feasibly applicable to at most four-body interactions, it would still be possible to draw allowed and even energetically favourable model configurations whose physical interpretations are in fact non-physical (see discussion in appendix 1). If an even more accurate model were desired, another vector could be introduced to each binding domain that would explicitly represent the helical axis. 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.

VI. ACKNOWLEDGEMENTS

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 642774 (ETN-COLLDENSE). D.F. acknowledges support from a UKIERI grant: DST-UKIERI-2016-17-0190.

REFERENCES

- ¹N. C. Seeman, "Nucleic acid junctions and lattices," *J. Theor. Biol.* **99**, 237 (1982).
- ²P. Rothmund, "Folding DNA to create nanoscale shapes and patterns," *Nature* **440**, 297 (2006).
- ³S. 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 (2009); 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 (2011); 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 (2013); 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 (2015); C. E. Castro, H.-J. Su, A. E. Marras, L. Zhou, and J. Johnson, "Mechanical design of DNA nanostructures," *Nanoscale* **7**, 5913 (2015); 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 (2015).
- ⁴Y. 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 (2008).
- ⁵V. Linko, M. Erikainen, and M. A. Kostianen, "A modular DNA origami-based enzyme cascade nanoreactor," *Chem. Commun.* **51**, 5351 (2015); M. Liu, J. Fu, C. Hejlesen, Y. Yang, N. W. Woodbury, K. Gothelf, Y. Liu, and H. Yan, "A DNA tweezer-actuated enzyme nanoreactor," *Nat. Commun.* **4**, 2127 (2013).
- ⁶H. T. Maune, S.-p. Han, R. D. Barish, M. Bockrath, W. A. Goddard, III, P. W. K. Rothmund, and E. Winfree, "Self-assembly of carbon nanotubes into two-dimensional geometries using DNA origami templates," *Nat. Nanotechnol.* **5**, 61 (2010).
- ⁷V. Linko, A. Ora, and M. A. Kostianen, "DNA nanostructures as smart drug-delivery vehicles and molecular devices," *Trends Biotechnol.* **33**, 586 (2015).
- ⁸L. Cademartiri and K. J. M. Bishop, "Programmable self-assembly," *Nat. Mater.* **14**, 2 (2015).
- ⁹D. Frenkel, "Order through entropy," *Nat. Mater.* **14**, 9 (2015).
- ¹⁰Y. Ke, L. L. Ong, W. M. Shih, and P. Yin, "Three-dimensional structures self-assembled from DNA bricks," *Science* **338**, 1177 (2012).
- ¹¹J.-P. J. Sobczak, T. G. Martin, T. Gerling, and H. Dietz, "Rapid folding of DNA into nanoscale shapes at constant temperature," *Science* **338**, 1458 (2012).
- ¹²K. E. Dunn, F. Dannenberg, T. E. Ouldrige, M. Kwiatkowska, A. J. Turberfield, and J.-P. Aimé, "Guiding the folding pathway of DNA origami," *Nature* **525**, 82 (2015).
- ¹³F. Dannenberg, K. E. Dunn, J. Bath, M. Kwiatkowska, A. J. Turberfield, and T. E. Ouldrige, "Modelling DNA origami self-assembly at the domain level," *J. Chem. Phys.* **143**, (2015).
- ¹⁴X. Wei, J. Nangreave, S. Jiang, H. Yan, and Y. Liu, "Mapping the thermal behavior of DNA origami nanostructures," *J. Am. Chem. Soc.* **135**, 6165 (2013).
- ¹⁵J.-M. Arbona, J.-P. Aimé, and J. Elezgaray, "Cooperativity in the annealing of DNA origamis," *J. Chem. Phys.* **138**, (2013); J.-M. Arbona, J. Elezgaray, and J.-P. Aimé, "Modelling the folding of DNA origami," *Europhys. Lett.* **100**, 28006 (2012); J.-M. Arbona, J.-P. Aimé, and J. Elezgaray, "Folding of DNA origamis," *Front. Life Sci.* **6**, 11 (2012).
- ¹⁶J. 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 (2016).
- ¹⁷J. SantaLucia Jr. and D. Hicks, "The thermodynamics of DNA structural motifs," *Annu. Rev. Biophys. Biomol. Struct.* **33**, 415 (2004).
- ¹⁸B. E. K. Snodin, F. Romano, L. Rovigatti, T. E. Ouldrige, A. A. Louis, and J. P. K. Doye, "Direct simulation of the self-assembly of a small DNA origami," *ACS Nano* **10**, 1724 (2016).
- ¹⁹B. E. K. Snodin, F. Randisi, M. Mosayebi, P. Šulc, J. S. Schreck, F. Romano, T. E. Ouldrige, 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).
- ²⁰P. Šulc, F. Romano, T. E. Ouldrige, 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).
- ²¹T. E. Ouldrige, 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).
- ²²T. E. Ouldrige, A. A. Louis, and J. P. K. Doye, "DNA nanotweezers studied with a coarse-grained model of DNA," *Phys. Rev. Lett.* **104**, 178101 (2010).
- ²³A. Reinhardt and D. Frenkel, "Numerical evidence for nucleated self-assembly of DNA brick structures," *Phys. Rev. Lett.* **112**, 238103 (2014); 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 (2015); A. Reinhardt and D. Frenkel, "DNA brick self-assembly with an off-lattice potential," *Soft Matter* **12**, 6253 (2016).
- ²⁴M. Sajfutdinov, 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 (2018).
- ²⁵H. K. Wayment-Steele, D. Frenkel, and A. Reinhardt, "Investigating the role of boundary bricks in DNA brick self-assembly," *Soft Matter* **13**, 1670 (2017).
- ²⁶E. Protozanova, P. Yakovchuk, and M. D. Frank-Kamenetskii, "Stacked-unstacked equilibrium at the nick site of DNA," *J. Mol. Biol.* **342**, 775 (2004); 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 (2006).
- ²⁷E. Marinari and G. Parisi, "Simulated tempering: A new Monte Carlo scheme," *Europhys. Lett.* **19**, 451 (1992).
- ²⁸V. I. Manousiouthakis and M. W. Deem, "Strict detailed balance is unnecessary in Monte Carlo simulation," *J. Chem. Phys.* **110**, 2753 (1999).
- ²⁹J. I. Siepmann and D. Frenkel, "Configurational bias Monte Carlo: A new sampling scheme for flexible chains," *Mol. Phys.* **75**, 59 (1992).
- ³⁰S. Consta, N. B. Wilding, D. Frenkel, and Z. Alexandrowicz, "Recoil growth: An efficient simulation method for multi-polymer systems," *J. Chem. Phys.* **110**, 3220 (1999).
- ³¹S. 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 (1999).
- ³²M. Dijkstra, D. Frenkel, and J.-P. Hansen, "Phase separation in binary hard-core mixtures," *J. Chem. Phys.* **101**, 3179 (1994).
- ³³D.-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 (2012).
- ³⁴J. 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 (2007).

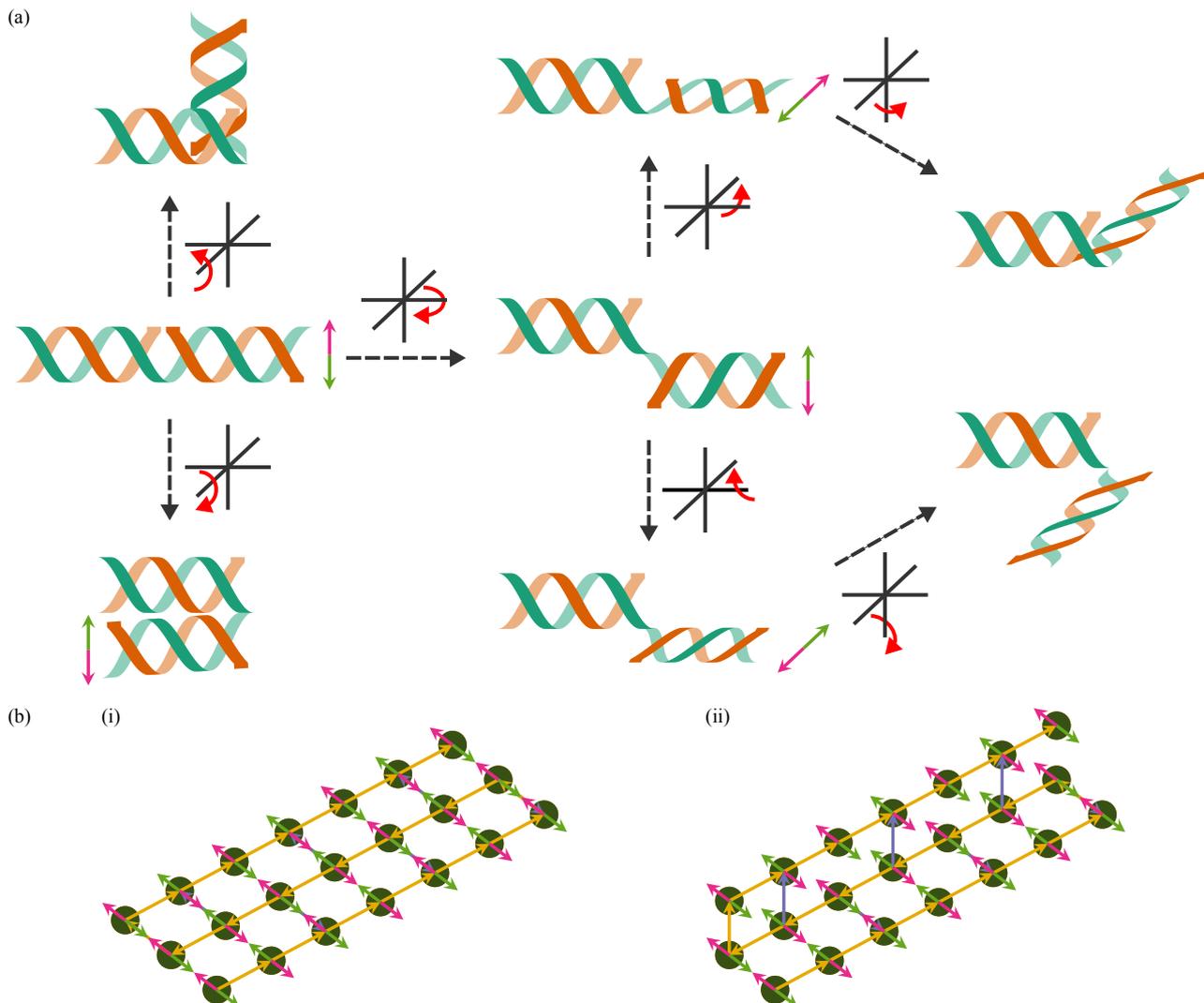


Figure 11. Model diagrams illustrating justification for the kink rules. (a) Kinks at a strand breakpoint in a helix involving 16-nt binding domains. Dashed arrows show the direction of the transformation described by the associated axes and red rotation direction arrow. The first column shows the starting helical conformation and two conformations resulting from rotation around an axis perpendicular to both the helical axis and the first domain's orientation vector. The second column shows three conformations resulting from rotation around an axis parallel to the helical axis but displaced to the outside of the helix. The third column shows two conformations resulting from taking conformations from the second column and applying a rotation around an axis perpendicular to the helical axis and parallel to the first domain's orientation vector. (b) An assembled 21-binding-domain scaffold system in a planar (i) and kinked configuration (ii). The scaffold and staples are shown separately in fig. S3.

APPENDIX

1. Justification for allowed kinked configurations

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 (fig. 11). To begin, consider rotating the second binding domain's helix around an axis parallel to the helical axis but displaced to the outside of the helix. This allows for configurations in which the next-binding-domain vector of the first binding domain is

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 to its orientation vector. In the assembled state these configurations are referred to as strand crossovers. Importantly, these configurations all have an orientation vector of the second binding domain that is equal to the orientation vector of the first binding domain. (This also happens to be necessary if strand crossovers are to be restricted to certain intervals of base pairs)

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 course the direction of the next-binding-domain vector will also be similarly rotated. To make the model consistent with these arguments, 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.

Therefore we constrain our model to disallow kinked configurations in which the first binding domain's next-binding-domain vector is parallel to its orientation vector but not parallel to the second binding domain's orientation vector. Configurations in which the next-binding-domain vector is antiparallel to the orientation vector of the first binding domain can also be disallowed because of steric clashes between the two binding domains.

Once we start considering kinks in which one or the other

of the bound domains has its helical axis defined by virtue of the presence of an adjacent bound domain (on its chain or on one of the bound domain chains), determining which model configurations have sterically prohibited interpretations becomes even more complicated. With the current model, it is not possible to prevent every sterically inhibited configuration without prohibitively complicated potentials. However, only one additional rule is required to ensure that crossovers between parallel helices occur with the desired geometry. In contrast to the rules involving only two bound domains forming a kink, the rule prohibits the underlying physical configuration by ensuring an additional kink is present, rather than directly prohibiting a model configuration. If both of the bound domains with a kink between them have a defined helical axis, and if the two helical axes are parallel or antiparallel 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.

The reason why these configurations are not allowed is not clear when using the idealized cartoon helix representation. It is best explained by considering parallel helices with multiple crossovers. Consider a 21-binding-domain scaffold that is designed to assemble into three parallel helices with crossovers between them (fig. S3). Without this rule, fig. 11(b)(i) and fig. 11(b)(ii) would be considered energetically equivalent, yet we know that the geometry of fig. 11(b)(ii) is unfavourable. If it were not, the classic tile structures (this 21-binding-domain system is a subset of the rectangular tile of ref. 2) would lack rigidity in solution, which is not the case.³³

2. Numerical validation of move types

To examine the validity of the move types, simulations of a toy 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 fig. 12(a)). 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 that used to run the simulations presented in the results section with the exception that the maximum number of scaffold domains to regrow is set to four. 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 fig. 12, which clearly shows that the two approaches agree within sampling error.

3. Optimization of move sets

A given move set has many free parameters and cannot be fully optimized without extensive efforts. For the move set used in the current study, which has 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

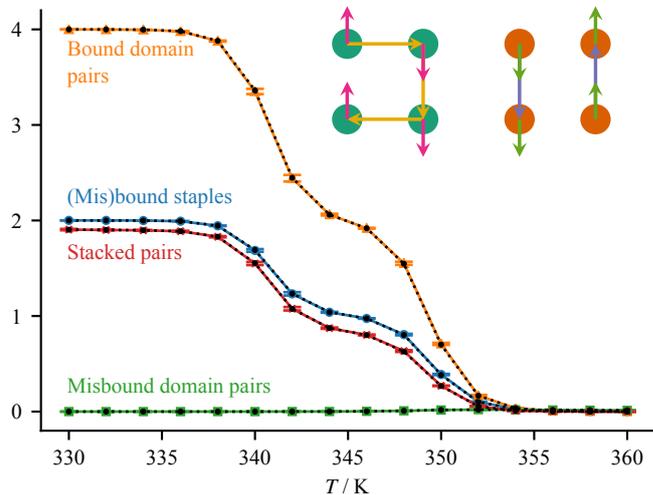


Figure 12. Numerical validation of the move types and their implementation. Plotted are the mean order parameters against temperature. The exact result for each order parameter is plotted in dashed black lines. The error bars on the simulation results represent the standard error in the means across ten independent simulations. Simulations were run with the same parameters as the simulations referred to in fig. 5. Above the curves is a schematic representation of the four-binding-domain scaffold system used. The scaffold and staples are shown in assembled configurations, but for clarity have been drawn separately.

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 by running simulations on the 24-binding-domain scaffold system, from which we determined the mean times to the first assembled state and to the first fully stacked assembled state, as well as estimated the effective sample size.³⁴ 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.

SUPPLEMENTARY FIGURES

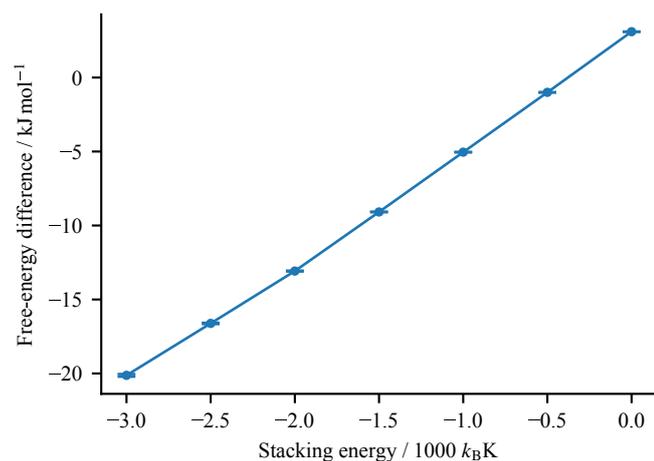


Figure S1. Stacking free energy as a function of stacking energy calculated from simulations of a two-binding-domain scaffold system with two one-binding-domain staples. The simulations were run at 320 K, which is below the melting temperature. The error bars represent the standard error in the means across three independent simulations. Simulations were run with a staple concentration of 100 nmol L^{-1} and a monovalent cation concentration of 0.5 mol L^{-1} .

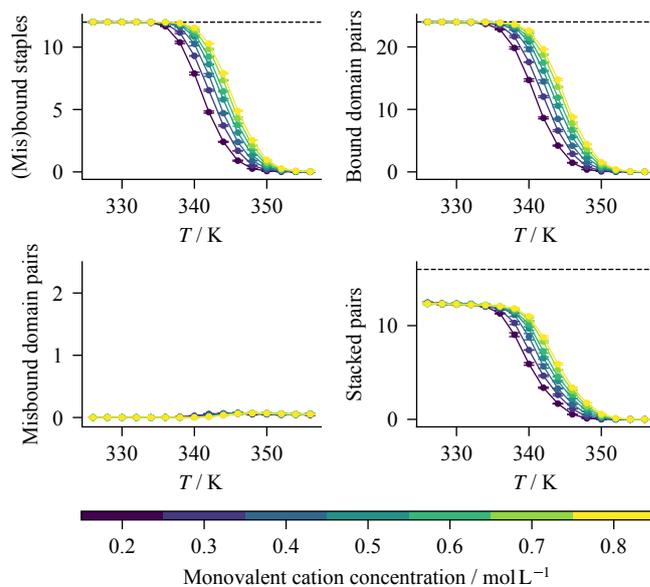


Figure S2. Mean order parameters from simulations of the 24-binding-domain system plotted against temperature for a range of sodium ion concentrations. The dashed lines correspond to the expected order parameter values in the fully assembled or fully stacked configurations. The error bars represent the standard error in the means across three independent simulations. Simulations were run with a staple concentration of 100 nmol L^{-1} and a stacking energy of $-1000 k_B K$.

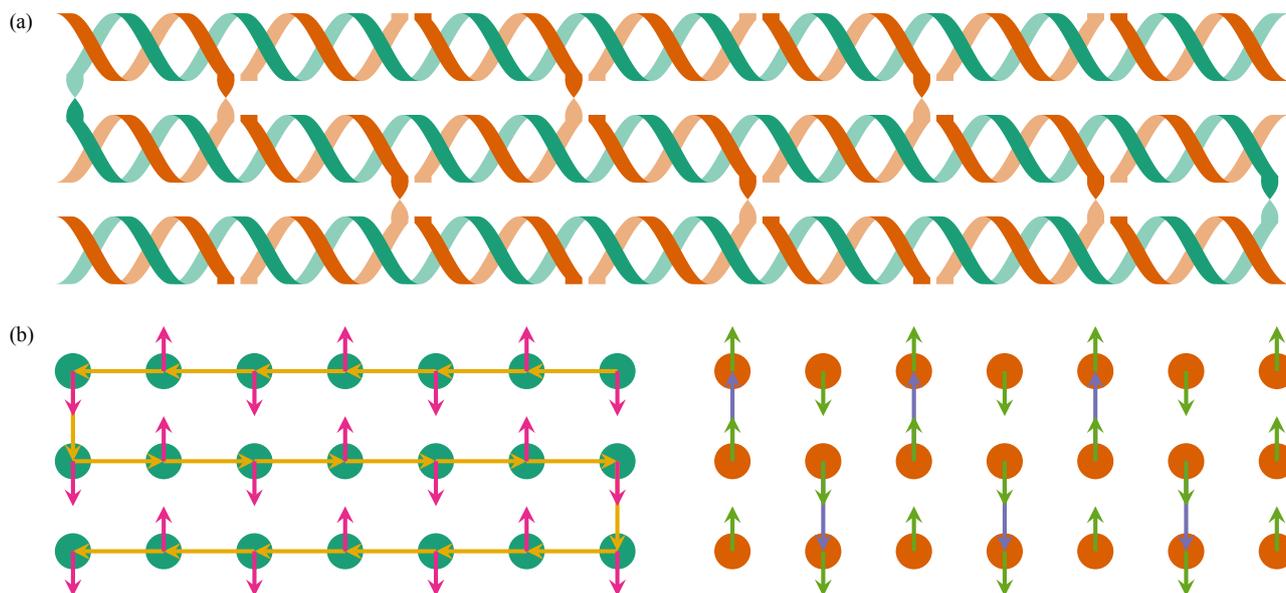


Figure S3. Schematic representations of the 21-binding-domain system. (a) Helical cartoon representation of the system in a fully assembled, planar configuration. (b) Representation of the system with the proposed model. The scaffold and staples are shown in assembled configurations, but for clarity have been drawn separately.