Structural, mechanical and thermodynamic properties of a coarse-grained DNA model

Thomas E. Ouldridge[1], Ard A. Louis[1], and Jonathan P. K. Doye[2]
[1] *Rudolf Peierls Centre for Theoretical Physics, 1 Keble Road, Oxford, UK OX1 3NP, UK*
[2] *Physical & Theoretical Chemistry Laboratory, Department of Chemistry, University of Oxford, South Parks Road, Oxford, OX1 3QZ, UK*
(Dated: September 24, 2010)

We explore in detail the structural, mechanical and thermodynamic properties of a coarse-grained model of DNA similar to that introduced in Ref. 1. Effective interactions are used to represent chain connectivity, excluded volume, base stacking and hydrogen bonding, naturally reproducing a range of DNA behaviour. We quantify the relation to experiment of the thermodynamics of single-stranded stacking, duplex hybridization and hairpin formation, as well as structural properties such as the persistence length of single strands and duplexes, and the torsional and stretching stiffness of double helices. We also explore the model's representation of more complex motifs involving dangling ends, bulged bases and internal loops, and the effect of stacking and fraying on the thermodynamics of the duplex formation transition.

PACS numbers: 87.14.gk,87.15.A-,34.20.Gj

I. INTRODUCTION

A single-stranded molecule of DNA (ssDNA) consists of a chain of alternating sugar and phosphate groups.[2] Attached to each sugar is a base, alanine (A), thymine (T), cytosine (C) or guanine (G). Bases are inherently planar, and their tendency to from coplanar stacks and undergo hydrogen-bonding leads to the formation of double-stranded helices (dsDNA). The canonical Watson-Crick base pairs (bp), C-G and A-T, are called complementary base pairs because they form the most stable hydrogen bonds.

The different base identities, along with the rules of complementarity, allow information to be encoded into the single strands.[3] In nature, this allows both strands of a double helix to carry the genetic information required for life. Recently, this information-carrying property has been harnessed in nanotechnology. A set of single strands can be designed with a pattern of complementarity that specifies a certain 2- or 3-dimensional structure (usually formed from branched double-helices) as the global free-energy minimum of the system. Strands can then be mixed and self-assemble, provided the sequences are well designed. When combined with its structural properties (dsDNA is stiff on the nanoscale, with a persistence length of around 50 nm or 150 bp,[4] and ssDNA has the flexibility to act as hinges between duplex sections), such selective interactions make DNA an ideal material for nanoscale self-assembly.

The self-assembly of short strands (oligonucleotides) was first demonstrated by the Seeman lab, who created a four-armed junction.[5] Junctions of this type, and more complex motifs,[6,7] have been used to create lattices[8,9] and ribbons.[7] 3-dimensional structures have also been realized: initially, the Seeman group constructed a cube[10] and a truncated octahedron[11] in several discrete stages. Polyhedral cages that rapidly form as solutions of oligonucleotides are cooled have since been developed.[12–16] These examples illustrate the potential of DNA as a material for controllable nanoscale self-assembly.

An alternative approach to self-assembly, DNA origami, was recently developed by Rothemund.[17] In this case, a long single strand is folded into a desired structure by short "staple" strands, allowing the assembly of an enormous range of 2-dimensional structures. This approach has been extended to three dimensions, either by linking together 2-dimensional sheets,[18] or by using the twist of DNA to form inherently 3-dimensional folded strands.[19] Additional methods of 3-dimensional self-assembly are possible: self-interactions within a single strand have been used to create a tetrahedron,[20] and other structures have been created from pre-assembled components involving DNA and other organic molecules.[21,22]

DNA nanotechnology is not limited to the self-assembly of static structures, as hybridization can also be used to drive nanodevices.[23] Such devices typically undergo structural changes due to duplex formation or toehold-mediated strand displacement (wherein a strand in a partially formed duplex is replaced by a strand which can form a more complete duplex). The earliest designs, such as the "tweezers" of Yurke *et al.*,[24] required sequential addition of single strands to force a system through a conformational cycle, or along a track.[25,26] The use of enzyme-facilitated hydrolysis,[27] or fuel in metastable states such as single-stranded hairpins,[28] has allowed the design of autonomous devices. The selectivity of DNA binding has also been used to perform simple logic operations,[29] offering the potential for "intelligent" nanostructures or devices, which respond to certain features of their environment.

As discussed above, much of DNA nanotechnology relies either largely or entirely upon B-DNA duplex hybridization from single strands (although other transitions can be exploited, such as the formation of single-stranded "i motif" structures[30]). Furthermore, some biologically relevant behaviour (such as the opening of tran-

sient "bubbles" (stretches of broken bps) within helices and the extrusion of cruciform structures in negatively supercoiled DNA[31]) relies primarily on the properties of single and double strands, and the competition between the two.

Information about the intermediate states in assembly processes, which are often difficult to resolve in experiment yet crucial to the processes as a whole, would aid the design of nanostructures and nanotechnology. Computer modelling, provided it can capture the transition between single- and double-stranded DNA, has the potential to offer significant insight into these systems.

At the most detailed level, atomistic simulations using force fields such as AMBER or CHARM offer an intimate representation of DNA.[32] A large-scale systematic study of the structural properties of short sequences as represented by AMBER has been carried out by the Ascona B-DNA Consortium.[33] Unfortunately, the number of degrees of freedom (including those of the solvating H_2O molecules) prohibits the simulation of large molecules for long periods of time. For example, simulations of double helices (on the scale of 10–20 base pairs) have only recently been extended to time scales of $\sim 1\,\mu$s.[34,35] The use of enhanced sampling techniques has given atomistic simulations some access to hybridization transitions in the smallest duplexes[36] and hairpins,[37,38] although larger systems remain prohibitively expensive to model.

At the other end of the spectrum, continuum models of DNA[39] treat the double helix as a uniform medium. Whilst these approaches can provide important insight into DNA behaviour on long length-scales, they are not constructed to deal with the details of processes involving duplex hybridization or melting.

To gain further insight into hybridization, coarse-grained models, which represent DNA through a reduced set of degrees of freedom with effective interactions, are required. In particular, models whose coarse-grained scale is approximately that of the nucleotide may provide the necessary compromise between resolution and computational speed.

The simplest available coarse-grained models are statistical, neglecting structural and dynamical detail. These models use sequence-dependent parameters that describe the free-energy gain per base pair relative to the denatured state, with extra parameters used for initialization of duplex regions and to describe unpaired sections within the a structure. Among the most popular are the Poland-Scheraga[40] and nearest-neighbour models,[41,42] generally used in the context of polynucleotide and oligonucletide melting, respectively. A particularly important variant of the nearest-neighbour model, which has been shown to reproduce experimental melting temperatures of duplexes ranging from 4–16 bp in length with a standard deviation of 2.3 K, was introduced by SantaLucia and Hicks.[41,42] In this model, the concentrations of oligonucletides A and B, and their duplex AB, are given by:

$$\frac{[AB]}{[A][B]} = \exp\Big(-\beta\big(\Delta H_{AB} - T\Delta S_{AB}\big)\Big), \qquad (1)$$

where the constants ΔH_{AB} and ΔS_{AB} are computed by summing contributions from each nearest-neighbour set of two base pairs, together with terms for helix initiation and various structural features, all of which are assumed to be temperature independent. Such a description, in which ΔH_{AB} and ΔS_{AB} are temperature independent, constitutes a "two-state" model.

Alternatives to these purely statistical models have also been proposed. Everaers *et al.*[43] have suggested a lattice model of DNA explicitly designed to unify nearest-neighbour and Poland-Scheraga models,with the added advantage that some structural information is also preserved. Peyrard-Bishop-Dauxois models[44] represent base pairs through a continuous 1-dimensional coordinate, allowing dynamical simulations of denaturation bubbles in polynucleotide DNA. None of the models discussed, however, provide a sufficiently sophisticated representation of the three-dimensional structure and dynamics of DNA to allow the detailed study of the transitions involved in nanotechnology.

To study the processes involved in nucleic acid structure formation, a fully 3-dimensional, dynamical, coarse-grained model is required. "Rigid base-pair" models, in which undeformable base pairs are the fundamental unit, have been used to study perturbations to DNA such as those induced by enzymes.[45] By definition, such models cannot represent the transition from single strands to duplexes, and hence are inappropriate for the study of assembly processes. Lankas *et al.*[46] directly compared rigid base-pair and rigid base models that were parameterized to reproduce positional time-series that were generated from atomistic simulations of B-DNA. Interestingly, they found that the rigid base models, in which the base pairs are deformable and nucleotides are the essential unit of simulation, generated a more local representation of the interactions than rigid base-pair models did, suggesting that the base-pairs are a more appropriate level of description for structural and mechanical properties of B-DNA.

Several rigid base models, and others in which each nucleotide has stiff internal degrees of freedom, have been proposed in the last decade. These models represent nucleotides by several interaction sites, and can be divided into two kinds. Firstly, some modellers parameterize their effective force fields by direct comparison with either atomistic simulations or data from crystal structures. An alternative is to take a more heuristic approach, designing force fields to provide a reasonable description of a range of large-scale properties (such as melting temperatures of helices) when compared to experiment: these two approaches could be described as "bottom-up" and "top-down", respectively.

Bottom-up approaches have been used to study RNA nanostructures,[47] the response of DNA minicircles to

supercoiling,[48,49] the behaviour of B-DNA over a range of conditions,[50] binding of DNA to the nucleosome[51] and the properties of the resultant model as a function of parameterization.[52] Although systematically coarse-graining removes some of the arbitrary choices in designing a minimal model, there are drawbacks. Firstly, the resultant force-field will be biased towards the structures with which it was parameterized: in particular, equilibrium duplex structures are often the primary source of information, and hence single-stranded behaviour is not necessarily well reproduced. Perhaps more significantly, the transition between ssDNA and dsDNA may be poorly represented: indeed, none of the bottom-up approaches described above have been used to investigate melting transitions in a rigorous way, with the focus being largely on structural properties. Secondly, "representability problems"[53] mean that careful fitting to distribution functions will not necessarily reproduce thermodynamic properties in a reliable fashion.[54] Finally, it is not yet known how accurate atomistic simulations are in reproducing the duplex hybridization transition.

All coarse-grained models represent a compromise, and an appropriate model must be chosen for the investigation at hand. Current examples of bottom-up approaches are well-suited to studying fluctuations in the vicinity of the equilibrium structure in question. By contrast, top-down approaches appear to lend themselves to the study of larger changes, particularly assembly transitions. Top-down approaches have been used to study duplex denaturation,[55] hairpin formation,[56,57] RNA folding[58,59] and mechanical unfolding,[60,61] Holliday junction formation,[62] duplex thermodynamics[63,64] and overstretching.[65]

For this paper we are mainly concerned with developing a model that can treat the formation of complexes involving single strands and B-DNA, with the particular goal of describing processes that are relevant to the self-assembly of DNA nanostructures and the dynamics of nanodevices,[1] but also with a view towards biological applications. We thus require a good representation of the structural, mechanical and thermodynamic properties of both single and double stranded DNA.

An important property to reproduce is the tendency of consecutive bases tend to form coplanar stacks, with an average separation of about 3.4 Å,[66], which is shorter than the equilibrium separation of phosphates (along the backbone) of approximately 6.5 Å.[67] The difference between the two length-scales helps determine the shape of B-DNA, which forms a double helix to exploit the stacking interactions. Helicity can also play an important role in the kinetics of assembly, in particular leading to frustration of bonding when strands are topologically constrained.[68]

The two length-scales also mean that single strands are ordered in a helical structure at low temperatures. At higher temperatures, where entropy dominates, they are disordered and significantly less stiff.[2,69] Such unstacked strands are extremely flexible relative to duplexes, permitting the formation of DNA structures which involve sharply bent single-stranded regions, such as hairpins. Furthermore, stacking has significant consequences for the thermodynamics and kinetics of assembly (the role of stacking in the thermodynamics of duplex formation is discussed in Section III B 3).

For complex assembly processes involving several interactions, it is important not only to correctly reproduce properties like melting temperatures, but also the experimentally measured transition widths so that certain features such as hierarchical assembly are preserved. More generally, the widths of the transitions determine the response of melting temperatures to concentration changes (Section III B 2). Finally, a reasonable representation of the elastic properties of DNA is important if the model is to be used to study systems involving DNA under stress, such as minicircles.[70]

Whereas the many other top-down models in the literature each have their strengths and weaknesses, we believe that none are currently optimized for the particular suite of properties that we desire to accurately reproduce. For example, most have either ignored the stacking transition of single strands[56,57,62] or enforced helicity largely through dihedral and angular potentials imposed on the backbone of a single strand.[55,63–65] In addition, where it was considered, the melting transition in previous models was generally significantly wider than experimentally reported.[57,62–64] In Ref. 1 we briefly introduced a model designed to represent ssDNA, B-DNA and the transition between them, and demonstrated its utility for nanodevices by simulating a full cycle of DNA tweezers.[24] We should note that the model is fitted at a fixed salt concentration, and does not distinguish between the strength of A-T and C-G base pairs.

The aim of the current paper is to give a detailed description of our modeling approach. In Section II, we present a slightly modified version of the model that appeared in Ref. 1, and discuss its philosophy, parameterization and simulation. The model's representation of DNA behaviour is presented in Section III. We first discuss model DNA structure and thermodynamics (Sections III A & III B), before considering it's mechanical properties (Section III C) and the representation of certain motifs such as hairpins (Section III D). Finally, we include an extensive discussion of the strengths and weaknesses of our approach in Section IV. The supporting appendices include a detailed representation of our model potential (Appendix A), a statistical model for stacking (Appendix B) and a statistical model for duplex formation that explicitly accounts for the effects of stacking and fraying (Appendix C).

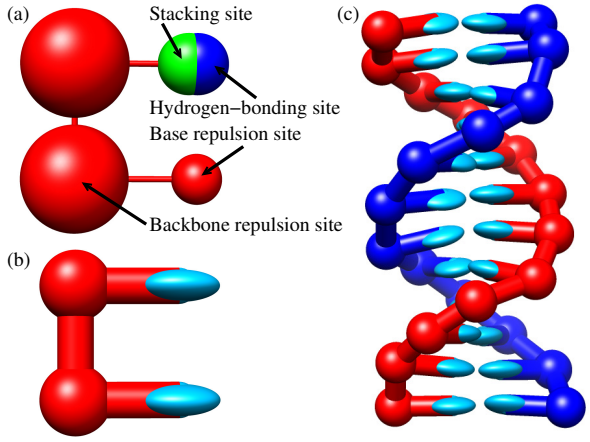

FIG. 1: (a) Model interaction sites. For clarity, the stacking/hydrogen-bonding sites are shown on one nucleotide and the base excluded volume on the other. The sizes of the spheres correspond to interaction ranges: two repulsive sites interact with a Lennard-Jones σ (Appendix A) equal to the sum of the radii shown (note that the truncation and smoothing procedure extends the repulsion slightly beyond this distance (Appendix A)). The distance at which hydrogen-bonding and stacking interactions are at their most negative is given by the diameter of the spheres. Visualization was found to be clearer with nucleotides depicted as in (b), with the subfigures (a) and (b) representing identical nucleotides on the same scale. The ellipsoidal bases allow a representation of the planarity inherent in the model, with the shortest axis corresponding to the base normal. (c) A 12 bp duplex as represented by the model.

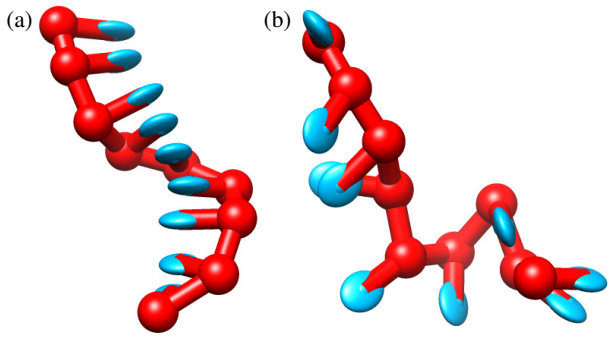

FIG. 2: Two possible configurations of a 9-base strand at 333 K. (a) All neighbours stacked to form a right-handed helix. (b) Most neighbours unstacked, giving a flexible, disordered strand.

II. METHODS

A. The model

1. *Philosophy of the model*

In designing a model, we have aimed to embed the thermodynamics of transitions involving ssDNA and dsDNA (in the most common B-form) into a 3-dimensional, dynamical, coarse-grained representation that provides a reasonable representation of structural and thermodynamic properties. This ambition naturally coincides with a top-down approach. We are not primarily concerned with the chemical details of interactions, but rather their net effect with regard to the properties of DNA. In addition, we have attempted to capture these properties by using only pairwise excluded volume, backbone connectivity, hydrogen-bonding, stacking and cross-stacking interactions (with no explicitly length- or loop size-dependent potentials[58,63,64]). The model we present here is a slightly modified version of that which appeared in Ref. 1, with the changes improving the representation of dsDNA flexibility and making the potential continuous and differentiable, allowing simulation methods which require forces, such as Langevin dynamics.[71]

An additional consideration in model design is the need for computational efficiency (if assembly transitions of complex structures are to be simulated). In our model, all interactions are pairwise (i.e., only involve two nucleotides, which are taken as rigid bodies). This pairwise character allows us to make efficient use of cluster-move Monte Carlo (MC) algorithms,[72] which facilitate relaxation on all length-scales in a bound structure, and allow a much larger typical step size than possible in Langevin dynamics.

Our model consists of rigid nucleotides, illustrated in Fig. 1. The three interaction sites lie in a line, with the base stacking and hydrogen-bonding/base excluded volume sites separated from the backbone excluded volume site by 6.3 Å and 6.8 Å, respectively. The orientation of bases is specified by a normal vector, which gives the notional plane of the base: the relative angle of base planes is used to modulate interactions (rather than through the use of off-axis sites).

2. *The potential*

In this section we present an overview of the potential. Further details are given in Appendix A. Model nucleotides interact in a pairwise fashion with other nucleotides in the system. Interactions between nearest-neighbours (nn) on a strand are distinct from all others, allowing for strand connectivity and stacking. The potential can therefore be written as a sum over nn pairs, and a sum over all others:

$$\begin{aligned} V &= \sum \left(V_{backbone} + V_{stack} + V'_{exc} \right) \\ &+ \sum_{\text{other pairs}}^{\text{nn}} \left(V_{HB} + V_{c\_stack} + V_{exc} \right). \end{aligned} \tag{2}$$

$V_{backbone}$ is a finitely extensible non-linear elastic (FENE) spring (see Appendix A), with an equilibrium length of 6.4 Å, representing the covalent bonds which hold nucleotides in a strand together.

V_{stack} represents the tendency of bases to form coplanar stacks: it is a smoothly cut-off Morse potential between base-stacking sites, with a minimum at 3.4 Å. It

is modulated by angular terms which favour the alignment of normal vectors, and the alignment of the normal vectors with the vector between stacking sites. As such, the interaction encourages coplanar stacks, separated by a shorter distance than the equilibrium backbone length, leading to helical structures. Right-handed helices are imposed through an additional modulating factor which reduces the interaction to zero for increasing amounts of left-handed twist.

V_{exc} and V'_{exc}, representing the excluded volume of nucleotides, prevent the crossing of chains and provide stiffness to unstacked single strands. The lack of explicit angular or dihedral potentials along the backbone allows single strands to be extremely flexible. For non-nearest neighbours, smoothly cut-off (and purely repulsive) Lennard-Jones interactions are included between all repulsion sites on the two nucleotides. For nearest neighbours, the backbone/backbone site interaction is not included because the distance between sites is regulated by the FENE spring.

V_{HB}, representing the hydrogen bonds which lead to base pairing, is a smoothly cut-off Morse potential between hydrogen-bonding sites, modulated by angular terms which favour the anti-alignment of normal vectors and a co-linear alignment of all four backbone and hydrogen-bonding sites. V_{HB} is set to zero unless the two bases are complementary (A-T or G-C). Together with V_{stack}, V_{HB} causes the formation of anti-parallel, right-handed double helices for complementary strands.

V_{c_stack} represents cross-stacking interactions between a base in a base pair and nearest-neighbour bases on the opposite strand, providing additional stabilization of the duplex.[73,74] We incorporate it through smoothed, cut-off quadratic wells, modulated by the alignment of base normals and backbone-base vectors with the separation vector in such a way that its minimum is approximately consistent with the structure of model duplexes.

Our model currently neglects some features of DNA. Although it incorporates sequence specificity (in that only A-T and C-G hydrogen bonds are possible), there is no sequence dependence in the potentials for either stacking, cross-stacking, hydrogen-bonding or excluded volume. We have made the simplifying assumption that non-complementary base pairs have zero attraction, and also neglected the possibility of alternative base-pair geometries (such as Hoogsteen[2]).We also have no explicit electrostatic interaction in the model, which may be expected to be important as bare ssDNA has a charge of $-e$ per base. For this reason, we fit to experimental data (where possible) at $[Na^+] = 500\,\mathrm{mM}$, where electrostatic properties are strongly screened. Indeed, at these ionic concentrations, the Debye screening length is approximately 4.3 Å, smaller than the excluded volume diameter for backbone-backbone interactions in our model ~ 6 Å. At the shortest distances allowed by the steric interactions, charges would have an energy of $\sim 2\,kT$ in a Debye-Huckel approximation. Other authors have attempted to explicitly include a Debye-Huckel term,[63,64] but also included a salt-dependent, medium-range attraction between strands in monovalent salt to facilitate hybridization, the physical origin of which is unclear.

Many of the simplifications in our model were made to reduce the number of possible parameters. For example, sequence dependence would give 16 combinations of stacking pairs, each pair requiring several parameters to describe their interaction. We also felt that, as an initial step in modelling, it was important to obtain a good physical representation of the underlying properties of DNA assembly (such as the generic dependence of melting temperature on length), before we incorporated sequence specific or low salt effects. Furthermore, some generic effects may be obscured by sequence-specific terms (for instance, free-energy profiles such as Figure 5 would have sequence-dependent fluctuations overlying the general trend).

3. Parameterization of interactions

Parameterizing such a model is a non-trivial process, as it involves a compromise between the representation of various aspects of DNA. In particular, a given parameter may influence a wide range of properties and it is difficult to design a simple metric to compare the reproduction of thermodynamic and mechanical DNA behavior. In our case, lengths were initially chosen to give our approximate B-DNA geometry. Interaction strengths and widths were then altered to give a description of the thermodynamics of stacking and duplex formation close to those in Ref. 75 and Ref. 42, respectively (for comparison to Ref. 42 we used an 'average base pair' – see Section III B 2). Finally, structural properties on long length-scales were checked, and widths of potentials and modulating factors adjusted, as potential width determines structural stiffness. Several iterations of this cycle were performed until the current parameter set was found .

In general, the interaction energy in a coarse-grained model should be interpreted as a free energy, as it incorporates a number of implicit degrees of freedom,[43] and thus it is plausible that interaction strengths could be temperature dependent. To reduce free parameters, we have avoided this temperature dependence except for the case of the the stacking strength. We found that it was difficult to design a stacking transition with an entropy as small as required (see Section III B 1) whilst maintaining an appropriate stiffness for dsDNA. Our stacking strength parameter has therefore been taken to be linearly dependent on temperature (see Appendix A: over the range 270-370 K, the stacking strength increases by $\sim 6\%$), in effect reducing the entropy cost of the transition.

There are two main possible reasons why this temperature dependence of the interaction parameters is required in our model. Firstly, it may be that it is an intrinsic property of the stacking interaction. In particular, stacking is thought to be partially a result of hydrophobic

effects,[2,76] and hence might be expected to be temperature dependent in any model without explicit water. Secondly, it may be that the coarse-graining leads to an overestimation of the entropy of the unstacked state relative to the stacked state, which then needs to be compensated by a temperature dependence in the interaction parameters. In particular, in order to replicate the flexibility of single strands, we impose no restriction on the conformation of the backbone-backbone, backbone-base and base normal vectors except for excluded volume. This lack of constraints is certainly a significant simplification, and will allow some conformations that would likely be excluded by specific steric clashes in a finer-grained model (such specific geometric effects would be exceedingly difficult to reproduce in a bead-spring model such as ours). We deem this likely overestimate of available configurations to be an acceptable price to pay for the flexibility of ssDNA necessary for hairpins and nanostructures.

B. Simulation technique

The results reported in this paper were obtained using the Virtual-Move Monte-Carlo (VMMC) algorithm developed by Whitelam and Geissler[72], which allows efficient MC simulation of strongly bound systems. The algorithm takes a selected single-particle move, as with conventional MC algorithms, and then grows a cluster from connected particles according to energy changes associated with the move. The algorithm combines collective motion with the large step sizes of MC (allowing quicker decorrelation and hence equilibration).

The combination of coarse-graining and an efficient MC algorithm provides access to processes on long time scales. To indicate simulation efficiency, we considered the formation of a 4 bp duplex at its melting temperature, in a periodic box of side length 17 nm (effective concentration 0.34 mM). A recent study[36] considered a similar system using an atomistic description with continuous solvent. In the atomistic case, sophisticated sampling techniques (replica exchange molecular dynamic and umbrella sampling) were required to provide data for the transition, which was the sole focus of the study. For our model, ~ 8 complete binding and unbinding cycles per hour were observed for an unbiased simulation (i.e., one without enhanced sampling) performed on a single CPU core.

In order to obtain good statistics for the melting transitions, umbrella sampling[77] simulations were performed at around the melting temperature and the results extrapolated to other temperatures using single-histogram re-weighting.[78] The number of bases with negative hydrogen-bonding energy was taken as a discrete order parameter for the reaction, $Q(\mathbf{x}^N)$ (with $\mathbf{x}^N$ representing the coordinates of the system), and the simulations were performed using the biasing weight $\exp(\beta W(Q))$, with $W(Q)$ chosen iteratively to make the partial partition functions

$$Z_Q^{biased} = \int \mathrm{d}\mathbf{x}^N \exp\left(-\beta(V(\mathbf{x}^N) - W(Q'(\mathbf{x}^N))\right) \delta_{Q,Q'} \tag{3}$$

approximately constant in Q. $W(Q)$ is chosen to flatten free-energy barriers, encouraging the simulation to visit rarely sampled states, thereby increasing the frequency of barrier crossing and improving statistics. We extract the unbiased partition functions using:

$$Z_Q^{unbiased} = Z_Q^{biased} / \exp(\beta W(Q)). \tag{4}$$

Simulation efficiency precluded the need for multiple umbrella windows for the study of duplex formation, and the accuracy of single-histogram re-weighting was checked for 15 bp duplexes, for which no systematic error over the range of extrapolation was found. Simulations of duplex formation were performed using two strands in a periodic box. Such simulations show strong finite-size effects due to the neglect of concentration fluctuations. These effects can be corrected for using the formalism of Ref. 79, allowing the extraction of bulk bonding probabilities.

III. RESULTS

A. Basic structure

The model is specifically designed to allow an approximate representation of B-DNA in its double-stranded state. The relative sizes of the equilibrium backbone separation and ideal stacking distance lead to a pitch of 10.34 bp per turn at 296.15 K (23°C, approximately room temperature) similar to experimental estimates of 10–10.5.[2,31] Our model length scale is chosen so that the average rise per bp at room temperature is equal to 3.4 Å,[66] which results in a helix with a radius (taken as the furthest extent of the excluded volume) of 11.5 Å, comparable to the experimental value of 11.5–12 Å.[66,80]

If strands are to form a double helix, it is not possible to optimize the stacking interaction, as consecutive stacking sites cannot sit directly above one another. Single strands, however, are not constrained in this way and hence form tighter helices, with a radius approximately 80% that of a duplex, similar to the 70-80% observed for a number of polynucleotide single helices.[80] A pleasing result is that, in order to alleviate the reduction in stacking, hydrogen-bonded bases undergo "propellor twisting" whereby bases in a pair twist in opposite directions in order to better align their stacking centres with adjacent bases in the same strand. Experimentally, propellor twist is seen to vary from around 5° to 15° in GC rich regions and from 15° to 25° in sections with large AT content.[81] In our case we observe an average propellor twist of 21.8° at 296.15 K, which is slightly larger than the average found for biological sequences.

B. Model thermodynamics

1. Single-stranded stacking transition

The attractive stacking interaction between adjacent bases causes single strands to form helical stacks at low temperature, with this order being disrupted as the temperature increases.[2] The literature is divided on both the nature of the attraction and the thermodynamics of the transition. The relative contributions of van der Waals, induction, hydrophobic and permanent multipolar electrostatic interactions remain unclear.[76] There has also been much debate on the cooperativity with which bases stack. Vesnaver and Bresslauer claim that a 13-base strand undergoes a completely cooperative transition between helical and random coil,[82] whereas other authors have inferred essentially completely uncooperative transitions for the individual stacks in poly(C) and poly(A).[83–86] Other groups claim weak to moderate cooperativity, with stacking probability affected by nearby base stacking.[87–89] It is clear, however, that stacking has a large influence on the thermodynamics of double helix formation, as the magnitude of the enthalpy and entropy changes of hybridization increase as the single-stranded state becomes more disordered.[75,82,88,90]

Given the uncertainty in stacking behaviour it is difficult to constrain the model in this regard. For simplicity we compare the model to reported uncooperative stacking (We note that introducing a large degree of cooperativity would require adding internal degrees of freedom to the nucleotide or including next-nearest-neighbour interactions). The study of Holbrook *et al.*[75] is most appropriate, as it deals with heterogeneous strands rather than homopolymers, and hence might be expected to provide a reasonable estimate of the average stacking strength.

To characterize the stacking properties of our model, we simulated oligonucleotides consisting of identical nucleotides (preventing the possibility of hydrogen bonding), and recorded the distribution of the number of neighbours with a stacking interaction stronger than a minimum value[114] as a function of temperature and oligonucleotide length. For each strand length (5–9 and 14 bases), two simulations (to check convergence) were performed at $T = 333\,\mathrm{K}$ for 10^{10} MC simulation steps (a minimum of 7×10^8 steps per nucleotide), and we extrapolated the results to other temperatures using single-histogram reweighting. For a 14-base nucleotide, around 50% of neighbours were found to be stacked at 338 K, with the transition being so broad that around 30% of neighbours remained stacked at 373 K, and 70% were stacked at around 306 K.

The stacking was fitted to a simple statistical model (based on that of Poland and Scheraga for helix formation in biopolymers[91]) which is discussed in detail in Appendix B. The model contains stacking enthalpies[115] and entropies Δh^{st} and Δs^{st}, such that the statistical weight (the contribution to the partition function) of an individual pair of stacked bases is $\exp(-\Delta h^{st}/RT + \Delta s^{st}/R)$ relative to the statistical weight of the unstacked state.[116] In addition, the statistical weight is multiplied by a cooperativity parameter σ for each contiguous run of stacked bases, and an end-effect term w for each stack which involves a base at the end of the strand. If σ and w are unity, each neighbour pair is independent. For $0 < \sigma < 1$, stacking is cooperative, and for $\sigma > 1$ stacking is anti-cooperative. For $0 < w < 1$, end bases are less likely to stack, and for $w > 1$ the opposite is true.

The four parameter model was fitted to data from strands of length $5 - 9$ bases, over a temperature range of $320 - 352\,\mathrm{K}$, giving:

$$\begin{aligned} \Delta h^{st} &= -5.55\,\mathrm{kcal\,mol^{-1}}, \\ \Delta s^{st} &= -16.0\,\mathrm{cal\,mol^{-1}\,K^{-1}}, \\ \sigma &= 0.766, \\ w &= 0.783. \end{aligned} \tag{5}$$

As σ and w are close to unity, our model shows only weak cooperative and end effects. The entropy and enthalpy parameters are similar to those found by Holbrook *et al.*,[75] who estimated $\Delta h^{st} = -5.7$ and $-5.3\,\mathrm{kcal\,mol^{-1}}$ and $\Delta s^{st} = -16.0$ and $-15.0\,\mathrm{cal\,mol^{-1}\,K^{-1}}$ for two different strands at $[\mathrm{Na^+}] = 120\,\mathrm{mM}$. Similar results at $[\mathrm{Na^+}] = 50\mathrm{mM}$ suggest weak salt dependence in this regime.[75]

Simulations performed in which the repulsive steric interactions were set to zero gave a slightly higher Δs^{st} and values of σ and w consistent with unity. Thus we conclude that the small cooperative effects in our model result from excluded volume. To understand the cause of the cooperativity, consider a chain of bases A, B, and C, and without loss of generality, consider B fixed whilst A and C move relative to it. Due to the requirement that base normals must point in the $3'$ to $5'$ direction to stack (see Appendix A), the regions of space in which A and C stack with B do not overlap. Therefore, if A and B are stacked, the excluded volume that A represents to C only prevents C adopting conformations in which it is unstacked. By contrast, if A and B are unstacked, the excluded volume of A can prevent C adopting both stacked and unstacked configurations. As a consequence, C has a slightly higher tendency to stack if A and B are stacked, and so there is a positive cooperativity. Similarly, end bases experience more freedom due to the reduction in excluded volume, and are therefore less likely to stack.

The statistical model is very successful. Fig. 3 compares its predictions to the results for a strand length (14 bases) and temperature (300 K) that are both well outside the ranges that were used in the fitting. Excellent agreement is found.

2. Duplex formation

Hydrogen bonding between bases can lead to the formation of bound pairs of DNA strands, which adopt the canonical 'B' double helix structure over a wide range of conditions due to stacking interactions. In contrast to

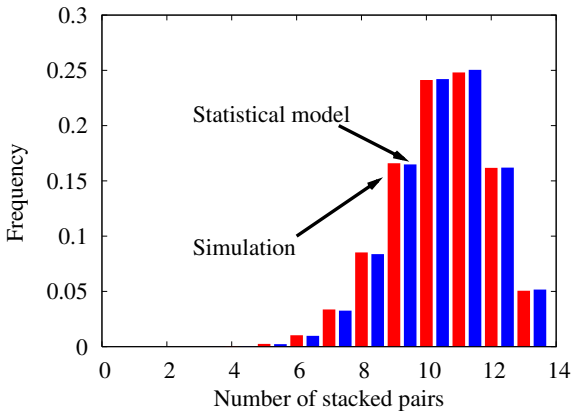

FIG. 3: Frequency of the total number of stacked bases in a 14-base single strand at 300 K from simulations of our model, and as predicted by the simpler statistical model with parameters as in Equation 5.

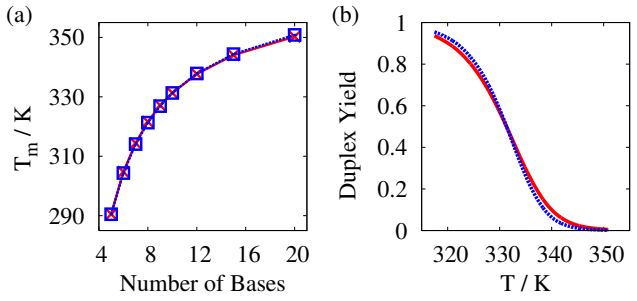

FIG. 4: (a) T_m as a function of strand length at an equal strand concentration of 3.36×10^{-4} M, as given by our model (crosses connected by a solid line) and averaged parameters from Ref. 42 (squares connected by a dashed line). (b) Fraction of 10-base strands bound in duplexes at a concentration of 3.36×10^{-4} M as a function of temperature, from our model (dashed line) and using the parameters of Ref. 42 (solid line).

the stacking transition, there is a reasonable consensus in the experimental literature on the melting temperature (T_m) of duplexes.

We fitted our model using the two-state model and parameters of Ref. 42, which is known to give a very good prediction of experimental T_m. Note that we do not reproduce two-state thermodynamics (see Appendix C), but rather treat Ref. 42 as a useful parameterization of experimental results for the melting temperatures of short duplexes. As our model contains no differentiation between A-T and G-C base pairs, we compare our results to strands consisting of 'average bases', the parameters for which, $\Delta h_{SL}^{step} = -8.2375\,\mathrm{kcal\,mol^{-1}}$ and $\Delta s_{SL}^{step} = -22.019\,\mathrm{cal\,mol^{-1}\,K^{-1}}$, were obtained from averaging over all possible complementary base-pair steps in Ref. 42. We also use the average helix initiation terms $\Delta h_{SL}^{init} = 1.1\,\mathrm{kcal\,mol^{-1}}$ and $\Delta s_{SL}^{init} = 3.45\,\mathrm{cal\,mol^{-1}\,K^{-1}}$, and an additional salt correction of $\Delta s_{SL}^{salt} = -0.12754\,\mathrm{cal\,mol^{-1}K^{-1}}$ per phosphate for $[\mathrm{Na^+}] = 500\,\mathrm{mM}$, again taken from Ref. 42.

We simulated pairs of complementary oligonucleotides in a periodic box for a range of strand lengths between 5 and 20 bases, and extrapolated to bulk statistics using the method discussed in Ref. 79.[117] Umbrella sampling, using the number of base pairs with a negative hydrogen-bonding energy as an order parameter Q, was used to ensure good sampling.

For the purposes of comparison with Ref. 42, we defined a state to be bound if any hydrogen-bonding interaction between strands had an energy below a cutoff of $-0.60\,\mathrm{kcal\,mol^{-1}}$, with typical hydrogen-bonding energies of a single base pair being larger by a factor of approximately 7. Doubling the cutoff had no significant effect on our results. T_m was taken as the temperature at which half of the strands would be bound in a bulk solution.

The variation in melting temperature with duplex length is shown in Fig. 4 (a), where it is compared to the predictions of the model of Ref. 42. The agreement in the dependence of T_m on length is extremely good: this dependence is essentially a measure of the cooperativity of the duplex forming transition, which is most strongly influenced by the relative contributions of hydrogen-bonding and stacking/cross-stacking to duplex stability.

The polynucleotide melting temperature (the melting temperature for infinitely long strands) at 500 mM $[\mathrm{Na^+}]$ for a strand of 50% C-G content, is predicted by the empirical relations given by Blake and Delcourt[92] and Frank-Kamenetskii[93] as 365.8 K and 363.2 K, respectively. An approximate value for our model can be estimated by simulating a pair of long, complementary strands in a partially bound state, and finding the temperature at which the free-energy change of adding an additional base pair to a partially formed duplex is zero. Simulations of partially formed 100–bp strands (with the duplex/single-stranded DNA interface at a variety of points) gave values of T in the range 364–366 K, in good agreement with the empirical relations.

Fig. 4 (b) compares the 10-bp duplex yield as a function of temperature for our model with the predictions of Ref. 42. The widths of the transitions are consistent to within a few degrees Kelvin, with our model consistently producing a marginally sharper transition for all duplex lengths. The width of the transition determines the response of the system to changes in concentration. Consider, for example, a simple two-state model of DNA hybridization, as used in Ref. 42 and expressed in Eqn. (1). Assuming equal total concentrations of each strand ($[A_0]$), the width of the transition scales approximately as:

$$\Delta T \sim \frac{k_\mathrm{B} T_m^2}{\Delta H}, \tag{6}$$

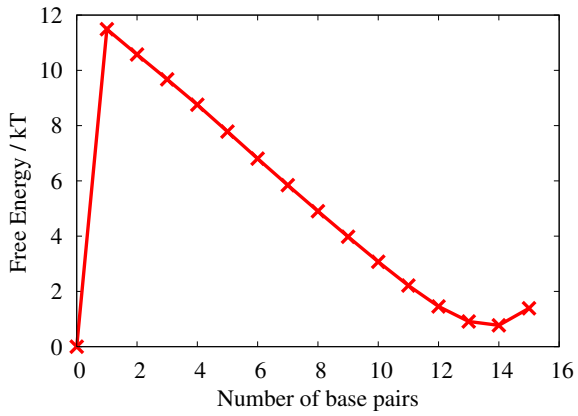

FIG. 5: Free-energy profile of bonding of a 15 bp duplex, as a function of the number of base pairs, at 343 K.

and the change in T_m with concentration is given by:

$$\frac{dT_m}{d[A_0]} = -\frac{k_B T_m^2}{[A_0]\Delta H} \sim \frac{\Delta T}{[A_0]}, \tag{7}$$

and hence agreement in both T_m and the transition width at a given concentration imply agreement in T_m over a range of concentrations.

3. *Free energy profile of duplex formation and fraying*

The free energy of duplex formation of a 15-bp duplex is plotted in Fig. 5 as a function of the number of base pairs (the order parameter for our umbrella sampling). To avoid complicating features in the free-energy profile due to hairpins and misbonds, which can conceal the underlying trends at low numbers of bonds, only base pairs that are present in the desired duplex had a non-zero strength of hydrogen bonding in this simulation. The general form of the free-energy profile is qualitatively similar to that found in Ref. 64 for another coarse-grained model of DNA, with an initial entropy penalty for the formation of the first base pair, followed by a downhill slope as the duplex 'zips up' in a cooperative fashion. As can be seen, the formation of the final base pair is actually free-energetically unfavourable, and the typical state consists of a duplex with 'frayed' ends. This fraying arises because bases at the end of the duplex lack the stabilizing influence of neighbouring base pairs on either side and entropy favours the open state.

Although fraying is a widely accepted phenomenon,[94] experimental data is rather sparse, though it is established that weaker AT ends fray more easily than CG capped helices.[95] Nonin *et al.*[95] inferred fraying probabilities of terminal AT bps of around 0.375 and 0.7 at 273 K and 298 K respectively, and found 0.015 and 0.12 for GC pairs at the same temperatures (at moderate salt concentrations), whereas Patel *et al.*[96] found much higher melting temperatures for terminal base AT pairs, concluding that they were around 50% frayed at 313 K at high salt concentration. Our model shows approximately 10% fraying at 273 K, increasing to around 21% at 300 K and reaching 50% at approximately 330 K, reasonable values for 'average' base pairs. We note that in many cases, particularly at low temperature, end bps in our model break but remain stacked, adopting conformations to maximize stacking at the expense of hydrogen bonding.

4. *Effect of stacking and fraying on thermodynamics of duplex formation*

We attempted to fit the duplex yield as a function of temperature, for each strand length l, using a two-state model of the form in Eqn. 1.

$$\frac{[A_l B_l]}{[A_l][B_l]} = v\frac{Z_{ll}}{Z_l^2} = \exp\Big(-\beta\big(\Delta H_l - T\Delta S_l\big)\Big), \tag{8}$$

where $[A_l]$ is the concentration of strand A of length l and $[B_l]$ and $[A_l B_l]$ are the concentrations of its complementary strand and the bound pair. v is the volume simulated, Z_{ll} and Z_l are the statistical weights (contributions to the partition function) of duplexes and single strands of length l in our simulations and ΔH_l and ΔS_l the (assumed T-independent) enthalpy and entropy of transition (we note that for our simulations in the canonical ensemble, ΔH corresponds to the change in internal energy of the system). It was found, however, to be an unsatisfying fit to the melting curves, and further attempts to fit ΔH_l and ΔS_l as a linear function in l (by analogy with the nearest-neighbour model), were unsuccessful. This failure should not come as a surprise, however, as several authors have indicated that the entropy and enthalpy of duplex formation show temperature dependence due to the single-stranded stacking transition.[75,82,88,90] A more sophisticated model which explicitly treats the stacking and fraying is detailed in Appendix C. We show that, for our model, temperature dependent effects can be incorporated into an extended nearest-neighbour description of the transition.

The actual temperature-dependent transition enthalpy can be deduced from:

$$\Delta H = -\frac{\mathrm{d}}{\mathrm{d}\beta}\ln K_{eq}, \tag{9}$$

where K_{eq} is the equilibrium constant of the reaction. The enthalpy changes at T_m for our model are slightly larger than predicted by Ref. 42, which is to be expected as the transitions are slightly narrower. The discrepancy rises from about 6% for 5-bp duplexes to around 22% for 20-bp double strands. The behaviour of ΔS is similar.

To investigate the details of the temperature dependence of enthalpy changes in duplex formation, we simulated 15 bp duplex formation over a wide range of temperatures (for clarity, we again only give "correct" pairs an attractive hydrogen-bonding interaction), with the data

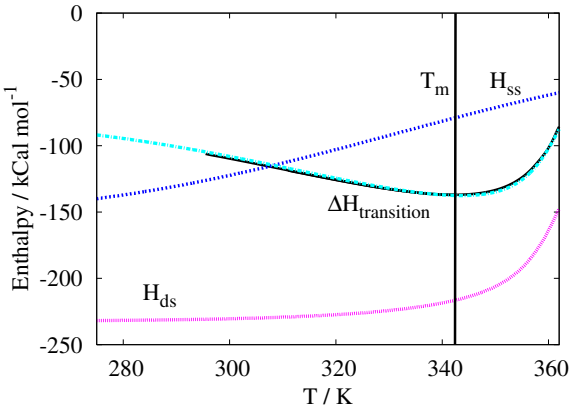

FIG. 6: Variation with T of enthalpies associated with the formation of a 15-bp duplex. Solid lines represent simulation results, dashed lines the predictions of the statistical model outlined in Appendix C. The lines labeled $\Delta H_{\text{transition}}$ give the enthalpy change upon duplex formation for the simulations and the statistical model. The lines labeled H_{ds} and H_{ss} are the enthalpies of the duplex and single strands respectively, relative to a completely unstacked state. The transition enthalpy in the statistical model is the difference between the latter two curves. The vertical line denotes the melting temperature $T_m = 342.6\,K$.

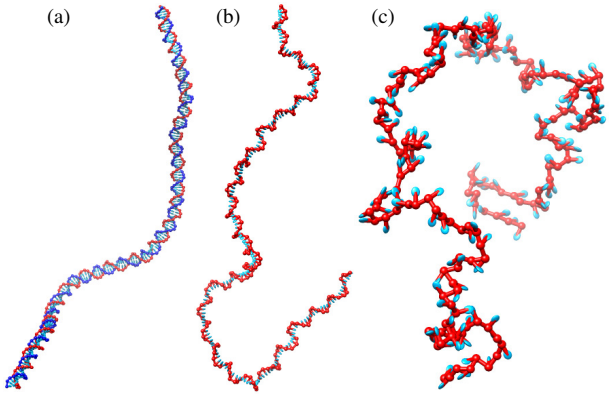

FIG. 7: Typical configurations indicating relative flexibility of double-stranded, stacked single-stranded and unstacked single-stranded DNA. (a) 202 bp double helix at 296.15 K. (b) Stacked single strand of 202 bases at 277.15 K. (c) Unstacked single strand of 160 bases at 296.15 K.

shown in Fig. 6. We find that at low temperatures (up to 342 K) ΔH becomes more negative with increasing temperature, with a gradient that reaches a maximum size of around $-0.055\,\text{kcal mol}^{-1}\,\text{K}^{-1}$ per base pair at approximately 328 K. At 342 K, however, ΔH reaches its most negative value, before increasing rapidly towards zero for higher temperatures.

The statistical model of Appendix C allows us to analyze this behaviour in terms of the enthalpy changes within the bound and unbound states. As shown in Fig. 6, the enthalpy of the bound state is approximately constant at lower temperatures, whereas the enthalpy of the single strands becomes less negative with increased temperature as they unstack, causing the observed tendency for ΔH of the transition to become more negative with increasing temperatures. As temperature continues to increase, however, the typical bound state changes from being a fully-formed duplex at low temperatures to a higher enthalpy partially-melted state at higher temperatures. Thus the enthalpy of the bound state becomes less negative as fraying becomes more significant, resulting in the observed increase in ΔH.

This change in enthalpy due to the stacking transition has been observed experimentally by several groups,[75,82,88,90,97] who deduced values for the typical enthalpy gradient of -0.050, -0.05 to -0.1, -0.062, -0.095 and -0.068 to $-0.87\,\text{kcal mol}^{-1}\,\text{K}^{-1}$ per base pair, respectively, in reasonable agreement with our model. These investigations were generally performed with either oligonucleotides with several CG pairs at the end[75,82,88,90] or polynucleotides,[97] both of which would massively reduce the impact of fraying. If we set the fraying contribution to zero, we obtain a typical value of -0.06 to $-0.07\,\text{kcal mol}^{-1}\,\text{K}^{-1}$, in even better agreement with experiment.

In addition, Jelesarov *et. al.*[90] also considered a duplex with AT bps at the end of the helix, for which ΔH becomes more negative with increasing T at low temperature, before flattening-off by around 310 K, in agreement with the predictions of our model for the consequences of fraying. Measurements were not performed at high enough T to check for an eventual reversal of the gradient of ΔH with temperature, but our model predicts the effect should be observable. In particular, duplexes with large AT end regions and stabilizing GC cores should demonstrate such an effect.

C. Mechanical properties

1. *Single-stranded persistence length*

Single strands, particularly when unstacked, are extremely flexible relative to dsDNA. This is crucial for nanotechnology, as it allows structures to contain highly bent ssDNA regions, such as at the vertices of polyhedra or the hinges of nanomachines.

Poly(dT) (long single stands of DNA in which all the bases are thymine) is generally assumed to be entirely unstacked at room temperature, and has little tendency to form secondary structure.[2,69] As a consequence, it can be used to test the inherent flexibility of unstacked single strands. Gapped helices have been used by Mills *et al.*,[69] who inferred a high salt persistence length of 20–30 Å from rotational decay rates, and Rivetti *et al.*,[98] who studied length distributions with atomic force microscopy, finding ~ 16 Å for short sections (< 5 bases), growing to around 28 Å for longer regions. Fluorescence resonance energy transfer between donors and acceptors

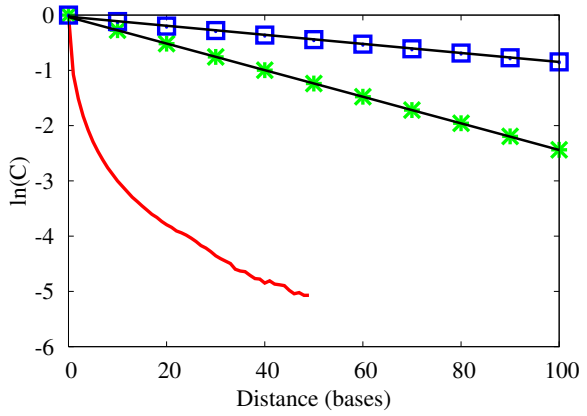

FIG. 8: Decay of the correlation (C) of helix axis plotted against base separation for a duplex at 296.15 K (squares) and a stacked single strand at 277.15 K (stars). The lines are fits to exponential decays. Also shown (solid line, no symbols) is the decay of the correlation of backbone vectors for an unstacked single strand at 296.15 K.

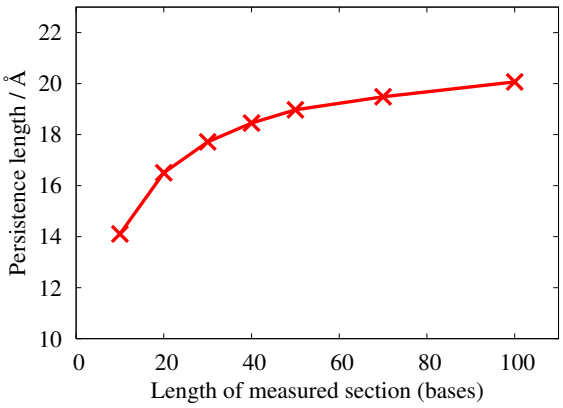

FIG. 9: L_{ps} plotted against the length of the single-stranded region of DNA analyzed at 296.15 K. For the purposes of comparison, the separation of successive backbone sites is approximately 6.4 Å.

attached to either end of poly(dT) has also been used to fit polymer models to chain end-to-end distributions, with Murphy *et al.* finding a persistence length of around 19.4 Å at 500 mM $[Na^+]$.[67] All of these results suggest persistence lengths on the scale of 2-5 bases.

To compare our model to experiment, we simulated single strands of one base type with stacking interactions set to zero, as shown in Fig. 7 (c), to mimic poly(dT). Poly(dT) is sometimes modeled as a worm-like chain,[67,98] in which a local stiffness opposes bending, resulting in an exponential decay of the correlations of backbone vectors with distance. In our model, however, unstacked ssDNA is essentially a freely-jointed chain with excluded volume, meaning that the conformation of backbone sites is restricted by steric clashes rather than local stiffness. As a result, the correlation of backbone-backbone vectors decays slower than exponentially (Fig. 8), due to steric interactions between non-neighbouring nucleotides. Such a decay implies that adjacent bases demonstrate larger kinking than would be expected from the picture of a worm-like chain with an equivalent overall stiffness of the strand. Consecutive backbone orientation is restricted only by steric clashes, hence large kinks are possible. More distant bases, however, still feel the excluded volume, and so the tendency is for directional correlation to decay slowly.

It was therefore difficult to obtain an unambiguous value for the persistence length to compare to experiment. We used the general definition from Ref. 99:

$$L_{ps} = \frac{\langle \mathbf{L}.\mathbf{l_0} \rangle}{\langle l_0 \rangle}, \tag{10}$$

with $\mathbf{L}$ being the end to end vector of the strand and $\mathbf{l_0}$ representing the first backbone-backbone vector. As the strand approaches infinite contour length, the value of L_{ps} should tend towards a constant, L_{ps}^{∞}. We estimated L_{ps}^{∞} by evaluating Eqn. 10 for single-stranded regions of lengths from 10 to 100 bases, embedded within strands of 70 to 160 bases. Four simulations were performed at 296.15 K for each length for at least 2.5×10^8 MC steps per particle, with the results plotted in fig. 9. The reason for embedding the measured length in a longer strand is that bases toward the end of single- or double-stranded DNA possess an increased relative flexibility. In order to obtain persistence length values that are valid for long strands where end effects are negligible bases near the end of strands were ignored.

This increased flexibility at the ends, which results from fewer restraining interactions, may be manifested in experimental systems. It is possible that interpretations that rely on the configuration of bases at the end of strands may be biased by such increased flexibility. Our results show that for strands of ~ 100 bases, the persistence length is similar to experimentally inferred values ($19 - 30$ Å). L_{ps} continues to grow noticeably for contour lengths much larger than L_{ps}^{∞} (~ 20 Å, or just over three bases), an effect that is consistent with the findings of Rivetti *et al.*,[98], and which indicates that non-nearest-neighbour interactions are important in providing the effective stiffness. This agreement with experimental results, together with the fact that our model provides a reasonably good representation of the effective excluded volume, suggests that our prediction that the freely-jointed chain gives a better representation of the conformational statistics of unstacked single-stranded DNA than the worm-like chain picture, should also hold for real DNA.

Mills *et al.*[69] also investigated the flexibility of gapped duplexes connected by poly(dA) at 4°C, when the bases are largely stacked into single helices. Although the interpretation depends on the probability of stacking, the intrinsic persistence length of the stacked regions was estimated to be in the region of 100 Å, corresponding to approximately 30 bases (stacked regions have a shorter

length per segment than unstacked sections due to twisting). This value is noticeably larger than that for unstacked strands, but smaller than for duplexes (approximately 150 bases at high salt concentration). For comparison we simulated single strands of 202 identical bases at 4°C for 8×10^9 MC steps (ignoring the data from the five bases at either end), requiring that all bases maintained a stacking interaction of $\geq -0.60\,\mathrm{kcal\,mol^{-1}}$ with their neighbours (doubling this value had no discernible effect). Unlike in the unstacked case, excluded volume does not play a large role as the length scale over which bending occurs is much larger than the size of one base (as can be seen in Fig 7 (b)). Hence, the relative alignment of vectors between stacking sites (which now act as the basic steps along the strand) was observed to decay exponentially, allowing a fit of the form:

$$\langle \mathbf{l_n}.\mathbf{l_0} \rangle = \exp(-n\langle l_0 \rangle / L_{ps}^{stack}), \tag{11}$$

from which we concluded that $L_{ps}^{stack}/\langle l_0 \rangle = 41.5$ bases (see Fig. 8) for our model. This value is higher than that reported by Mills *et al.*[69] by approximately 50%, but importantly it is much greater than the persistence length of unstacked ssDNA whilst also being much more flexible than dsDNA. Furthermore, the estimates in Ref. 69 assume unstacked bases behave as regions of persistence length 30 Å, which is at the upper end of estimates for poly(dT). As already noted, our results suggest that local kinking can be much larger than would be implied by the persistence length of unstacked bases. As such, the flexibility contribution from a single unstacked base may be larger than estimated, and consequently the flexibility of the stacked regions may be overestimated, possibly bringing our model into better agreement with the data.

2. *Double-stranded persistence length*

The persistence length of dsDNA is generally accepted to be approximately 450-500 nm at moderate to high [Na^+], corresponding to around 130–150 base pairs.[4,100] We performed three simulations of a duplex of length 202 bp at 296.15 K for 1.5×10^9 MC steps, ignoring the data from the ten base pairs at either end. Similar to our findings for stacked single helices, the correlation of the helix axis (defined as the distance between consecutive base-pair midpoints) at two points was observed to decay exponentially with distance, allowing an estimate of L_{ps}^{duplex} through Eqn. 11. Fig. 8 indicates a model persistence length of around 125 base pairs, in reasonable agreement with experiment. A typical configuration is shown in Fig. 7 (a).

3. *Double-stranded torsional and extensional stiffness*

Torsional rigidity (in the linear regime) is quantified by an elastic modulus C, which relates applied torque G to resultant twist $\Delta\theta$ of a duplex of length l: $C = Gl/\Delta\theta$. Estimates for C have been made using cyclization kinetics and topoisomer distributions for minicircles,[4,101,102] luminescence depolarization[103] and from twisting of DNA under tension,[104] giving values in the range 170–440 fJ fm. The effect of salt concentration on C is not entirely clear from the experimental literature.[103]

Calculating the response to torsion is non-trivial, as the curvature of the DNA axis makes the twist between two ends hard to define. In our previous work,[1] we attempted to infer an elastic modulus from the fluctuations in the angle between successive bases when projected onto the plane perpendicular to the vector joining their midpoints. Unfortunately, this method overestimates the torsional flexibility, presumably failing to decouple torsional variation from other fluctuations in a base-pair step. In this work, we instead obtain an approximate estimate of the torsional modulus by considering the twisting of the central 10 base pairs of a 20 bp duplex, and the central 20 base pairs of a 30 bp duplex at 296.15 K. Such short sections are extremely stiff, minimizing the natural bending fluctuations. To provide an unambiguous definition of torsion and twist, MC moves were chosen so that the base pairs at the end of the central section remained perpendicular to the vector between their midpoints, allowing the vector between the midpoints to define an axis about which torsion could be applied and twist measured.

Simulations were performed in which the torque applied to the end bases was varied between ±8 pN nm, and the resultant twist used to infer C. A separate estimate was also obtained using the equipartition result for the variance in twist at zero torque: $\langle \Delta\theta^2_{\mathrm{twist}} \rangle = kTl/C$. Further simulations used the equipartition result to estimate C under a tension of 9 pN, to ensure that stretching the duplexes had no effect. All estimates (for both 10- and 20-bp regions of interest) gave $C \sim 455 - 495\,\mathrm{fJ\,fm}$, suggesting that this is a reasonably robust estimate of the torsional stiffness of DNA duplexes in our model.

A long molecule of dsDNA under low tension responds as an extensible worm-like chain, with the behaviour initially dominated by the straightening of the chain, before stretching the base-pair rise itself becomes relevant as the chain extension approaches the contour length.[105,106] At higher forces, the duplex undergoes an overstretching transition and the B-DNA structure breaks down.[107] Experimental estimates for the extensional modulus K, obtained from fitting force-extension curves to extensible worm-like chain models, give K in the region of 1050–1250 pN at high salt.[105,106]

The extensional modulus K was estimated by applying tension to a 100-bp region within a 110-bp double helix, and fitting the resultant force-extension curve to the result of Odijk[108] for extensible worm-like chains:

$$x = L_0 \left(1 + \frac{FL_0}{K} - \frac{kT}{2F} \left[1 + y \coth y \right] \right), \tag{12}$$

where

$$y = \left(\frac{F L_0^2}{L_{ps} kT} \right)^{1/2}, \qquad (13)$$

in which x is the extension resulting from a force F applied to a duplex of contour length L_0 and persistence length L_{ps}. Performing an unconstrained three-parameter fit with the values of L_0, L_{ps} and K gave an excellent agreement with the data, as shown in Fig. 10, with $K = 2120\,\mathrm{pN}$, $L_0 = 339.4\,\text{Å}$ and $L_{ps} = 438\,\text{Å}$ (129 bp). The value of L_0 is similar to that expected from the rise of a short duplex (exactly 3.4 Å per base pair would give $L_0 = 336.6\,\text{Å}$), and L_{ps} is only slightly larger than the estimate from the decay of the correlation of the helix axis (415 Å). This agreement suggests that the extensible worm-like chain model provides a good description of the model's properties in this regime, and that the value of $K = 2120\,\mathrm{pN}$ is a reasonably robust one for our model.

Our model gives $C \approx 475\,\mathrm{fJ\,fm}$ (slightly larger than the top of the experimental range of 170–440 fJ fm) and $K \approx 2120\,\mathrm{pN}$, (about twice as large as typical experimental estimates). We do not believe the differences are crucial to the processes we are interested in investigating (although certain quantities, such as the critical twist density at which plectonomes are extruded, will be affected). It was found to be difficult to reparameterize the model to reduce these moduli without decreasing the persistence length, which is already slightly below experimental estimates. We feel that the current compromise, in which the persistence length is most faithfully reproduced, is a reasonable one as it is easier to imagine that nanostructures and nanodevices would be more sensitive to bending than torsional or extensional stiffness.

It is worth noting that recent investigations have suggested that DNA overwinds when stretched.[109] Our model does not reproduce this anti-intuitive behaviour, instead slightly untwisting as the stacking distance is extended. It is possible, therefore, that the model fails to capture the softness of a mode of deformation – perhaps the sloping of base pairs with respect to the axis[110] –that leads to this behaviour. If this is the case, it is perhaps unsurprising that the estimated moduli are larger than experimental observations

D. Structural motifs

1. *Hairpins*

DNA hairpins, which occur when a self-complementary strand binds to itself and forms a duplex stem and an unhybridized loop (Fig. 11), are a common structural motif. They have biological importance as a mechanism for release of superhelicity through cruciform formation.[31] Their relevance to nanotechnology includes metastable states (either occurring by accident[1] or through design).[28,68] In addition, they are an extremely common motif in biological RNA structures.[111] Aside from our earlier work using a previous parameterization of the current model,[1] we are unaware of any simultaneous application of a coarse-grained model to the formation of both hairpins and bimolecular duplexes. Our approach, in which the single strands have the potential to be extremely flexible, allows for hairpins and duplexes to have appropriate relative stabilities.

To demonstrate the ability of our model to represent hairpins, we simulated systems with stem sizes ranging from 6–12 bps, and loops of 6–18 bases. Four simulations for each hairpin were performed in the vicinity of T_m for 4×10^{10} MC steps (corresponding to at least 10^9 steps per nucleotide). Umbrella sampling as a function of hydrogen-bonded base pairs was used to ensure good statistics. In this case, we considered only states with at

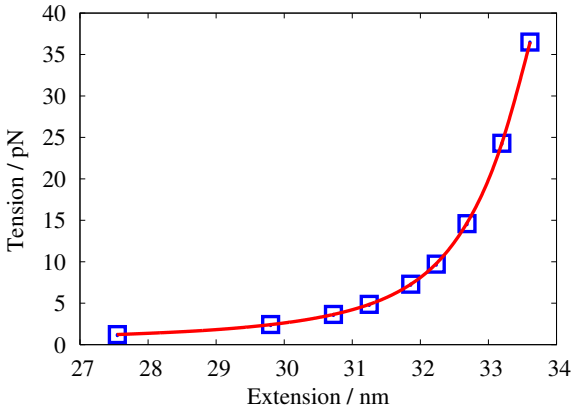

FIG. 10: Tension applied against extension for the central 100 bp of a 110-bp duplex at 296.15 K. The squares are simulation results, the solid line is a fit using Eqn. 12.

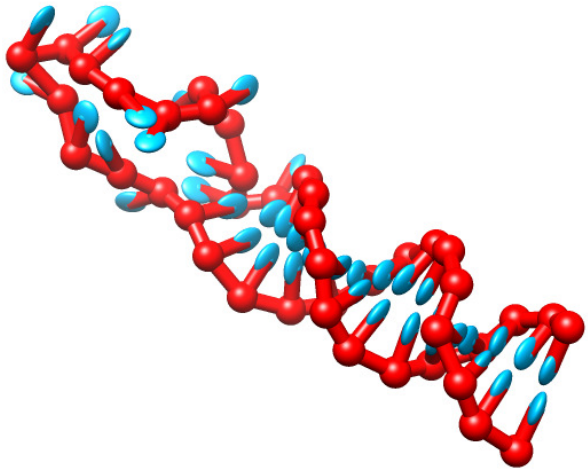

FIG. 11: A hairpin with a 12 bp stem and an 18-base loop at 343 K.

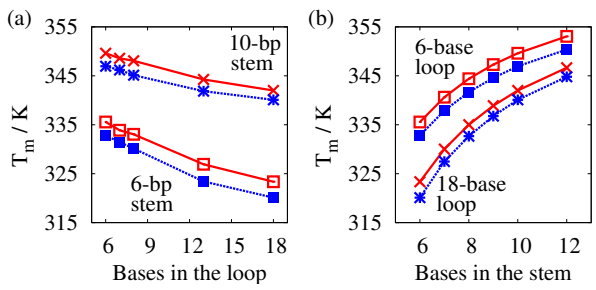

FIG. 12: Variation of hairpin melting temperature with (a) loop length and (b) stem length from our model (symbols connected by dashed lines) and from Ref. 42.

least one of the 'native' bps in the stem present as being a hairpin, as long loops have the potential to form transient base pairs with little relevance to the stability of the target structure. SantaLucia has presented parameters for estimating the melting temperature of hairpins,[42] which we again take as a good representation of experimental results. These parameters include sequence independent entropy penalties for loop formation and enthalpy/entropy terms for the stabilizing effect of the first mismatched bp in the loop (called a 'terminal mismatch': we compare to an average $\Delta h_{SL}^{term} = -2.91\,\mathrm{kcal\,mol^{-1}}$ and $\Delta s_{SL}^{term} = -7.33\,\mathrm{cal\,mol^{-1}\,K^{-1}}$). Our results for T_m are compared to the predictions of Ref. 42 in Figs. 12 (a) and (b). T_m is defined as the temperature at which a strand is in a hairpin state half of the time.

The results indicate that our model slightly underestimates T_m for hairpins relative to the predictions of Ref. 42 (and by extension, experiment) by approximately 3 K, which is slightly less than 1% of the absolute melting temperature (at the T_m predicted by Ref. 42, our hairpins constitute approximately 25% of the ensemble rather than 50%). Encouragingly, the trends with loop length and stem size are well reflected by our model (this is particularly pleasing, as the dependence on loop length was not used in parameterization), an indication that the majority of the physics of hairpin formation is well represented by our model. We note that our model is less successful for the smallest loops (3-5 bases), possibly because it does not incorporate specific interactions within a tightly packed loop that may provide extra stability.[112] As found with duplex formation, transition widths for our model are slightly smaller than predicted by Ref. 42 (the difference is very similar to that observed in Fig. 4(b)).

2. *Mismatches, bulges and internal bubbles*

A variety of other DNA motifs exist, such as duplexes involving mismatches between non-complementary base pairs or with one strand carrying extra, unpaired bases. SantaLucia [42] has provided parameters for the influence of these motifs on T_m. In many cases, they are highly sequence dependent and it is less clear than in the simple double helix case (where the variations in parameters are relatively smaller) that averaging over ΔS and ΔH contributions for all sequences is a reasonable approach to find an average effect. It should, however, give a rough estimate of the typical change in melting temperature due to a motif.

We compared the effect of several motifs on model duplex T_m to the predictions of Ref. 42, again averaged over all possible sequences (Table I). The simplest possible case is that of a single unpaired base at the end of a strand, generally referred to as a 'dangling end'. Typically, dangling ends are observed to provide a stabilizing influence, assumed to result from cross-stacking with the final base pair of the duplex, although the degree of stabilization is highly sequence dependent.[42,76] The cross-stacking interaction included in our model provides such a stabilizing effect, and the degree of stabilization is in good agreement with the predictions of Ref. 42.

In contrast to dangling ends, extra, unpaired bases on one strand within the helix are highly destabilizing, as they disrupt the helix structure. In the terminology of SantaLucia, these are known as bulges. In general, our model slightly underestimates the destabilization of helices due to bulges compared to the predictions of Ref. 42, although the observed melting temperatures remain within 2% of the predictions.

If a non-complementary pair of bases is added to an otherwise complementary duplex to form a mismatch, the effect is generally stabilizing at the end of a duplex (this is a "terminal mismatch") and destabilizing in the interior. Our model reproduces this tendency as shown in Table I, and also captures the increase in destabilization if the mismatch region is extended (to form an internal "bubble"). Once again, the destabilizing effect of motifs internal to the duplex tend to be slightly underestimated relative to the predictions of Ref. 42, and the observed melting temperatures again remain within around 2% of the predictions.

The motifs provide a good test of the model, as many were not considered in parameterization (although the dangling ends and terminal mismatches were used to constrain the strength of cross-stacking). In addition, misbonded structures involving these motifs may have a role in the kinetics of nanostructure assembly, and hence it is important that the model provides a reasonable representation of them. Although in some cases the quantitative agreement with Ref. 42 is not perfect, the model represents these motifs in a physically sensible way and the trends in stability at least qualitatively reflect the average properties of DNA. Furthermore, the typical magnitudes of ΔT_m are reasonable, with the T_m remaining within 2% of the average predictions of Ref. 42. It is possible that an underestimate of the disruptive effect of extra bases on the helical structure,[31] perhaps because the excluded volume of bases is smaller than in reality, causes the underestimate of ΔT_m due to internal motifs. This effect,

Motif	Complementary bp	Motif size	ΔT_m / K	
			Our Model	Ref. 42
Dangling end	5	1 base	+3.95	+4.24
	8	1 base	+1.20	+1.44
	15	1 base	+0.74	+0.61
Bulge	8	1 base	−18.58	−23.40
		2 bases	−24.64	−27.23
	15	1 base	−8.86	−12.58
		2 bases	−11.51	−11.67
		5 bases	−16.91	−13.78
Terminal mismatch	5	1 base / strand	+6.85	+6.95
	8	1 base / strand	+2.73	+2.55
	15	1 base / strand	+0.74	+0.63
Internal mismatch / bubble	8	1 base / strand	−8.77	−14.09
		2 bases / strand	−15.77	−21.86
		5 bases / strand	−25.83	−28.81
	15	1 base / strand	−5.35	−4.97
		2 bases / strand	−9.53	−11.60
		5 bases / strand	−15.62	−15.74

TABLE I: Effect on the melting temperature of a complementary duplex due to the addition of a motif. In this table, ΔT_m is the difference between the T_m of a structure with the motif and a fully complementary duplex consisting of the same number of complementary bps as the motif structure. For internal mismatches, bulges and bubbles, the motif was placed at the centre of the duplex.

however, would be expected to be larger for bulges than for mismatched pairs or symmetric bubbles.

Given the good agreement between the model and Ref. 42 for a single mismatch added to a 15-bp duplex, we investigated how the position of the mismatch affected stability. T_m is plotted against the position of the mismatch in Fig. 13. As can be seen, there are two distinct regimes, with the melting temperature initially decreasing as the mismatch is moved from the end of the strand (where it is stabilizing) towards the centre. Eventually, however, it reaches a plateau at around five bases from the end of the strand.

The cause of this plateau can be identified from examining the free-energy profiles for duplexes with mismatches located two and six bp from the end (Fig. 14). The first point to note is that the stability of duplexes with the maximum number of base pairs (15) is nearly identical, despite the difference in mismatch position. This suggests that provided a mismatch is surrounded by base pairs on either side, changing its location has little effect on the total free energy. The difference in T_m arises instead from a difference in the lowest free-energy state.

When the mismatch is near to the strand end (in the regime where T_m depends on mismatch position), the most stable state consists of the larger section of duplex formed with the bases beyond the mismatch unpaired. In this regime, the total free-energy gain from pairing the bases beyond the mismatch does not compensate for the free-energy cost of enclosing a mismatch in a helix. As the mismatch is moved towards the centre, the larger section loses bases and so becomes less stable, with the consequence that T_m drops. At some point, however, it becomes favorable for the bases in the shorter region to also bond. From this point onwards, the most stable state consists of the two duplex regions surrounding the mismatch. The net effect of moving the mismatch further towards the centre only marginally affects the overall stability of the duplex. As a result a plateau in T_m should occur.

As the temperature is lowered, the free-energy gain from base pair formation increases. As a consequence, the number of bases required before the region beyond the mismatch is stable as a duplex decreases. For example, we find that for a mismatch two bases from the end of a 15 bp duplex, the enclosed mismatch state becomes the most stable just below 320 K.

It is claimed in Ref. 42 that the stability of a mismatch is independent of its position, except for terminal mismatches and mismatches occurring one base from the end, which may cause the final base pair to be unstable. Our simulations suggest, however, that the distance of the mismatch from the duplex end at which T_m plateaus should increase with strand length (as longer strands melt at higher temperature). Furthermore, a similar temperature-dependent influence of motif location should hold for all destabilizing internal bubbles and bulges, as the beginning of the plateau simply indicates

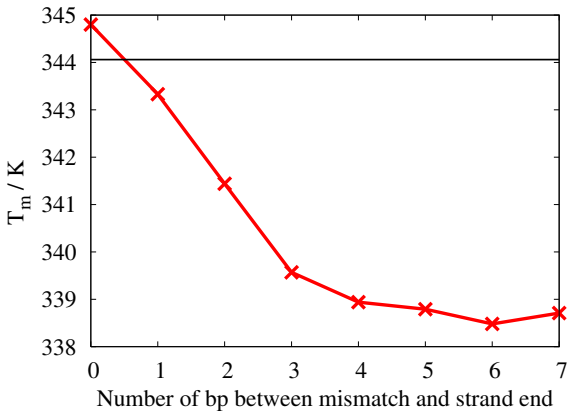

FIG. 13: Melting temperature of 15-bp complementary helix with an additional mismatch added against the distance of that mismatch from the end of the strand. The melting temperature in the absence of a mismatch is indicated via the horizontal line.

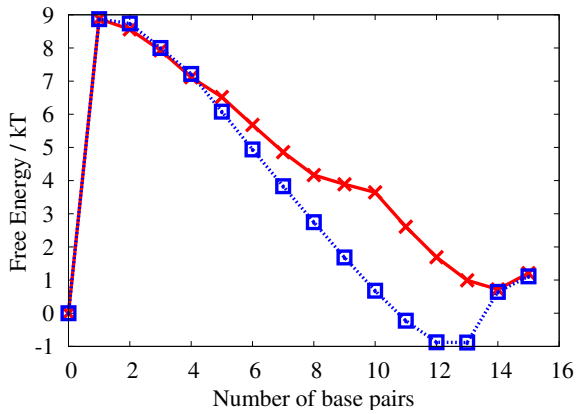

FIG. 14: Free energy profile at 339 K for a 15 base pair duplex with one additional mismatch placed 2 bases from the end (squares) and 6 bases from the end (crosses).

the point at which it is free-energetically favourable to enclose the disruption. This result should be qualitatively robust to the approximations in the model. In particular, sequence dependence will likely cause fluctuations but not destroy the general trend.

IV. DISCUSSION

We have examined in detail the structural, mechanical and thermodynamic properties of a coarse-grained model of DNA based on that presented in Ref. 1 (and used there to simulate a full cycle of DNA tweezers, an iconic nanodevice). Several small alterations to the model were made in order to improve the description of DNA flexibility and allow for the calculation of forces and torques. The aim of the model is to embed the known thermodynamics of B-DNA into a dynamical, coarse-grained

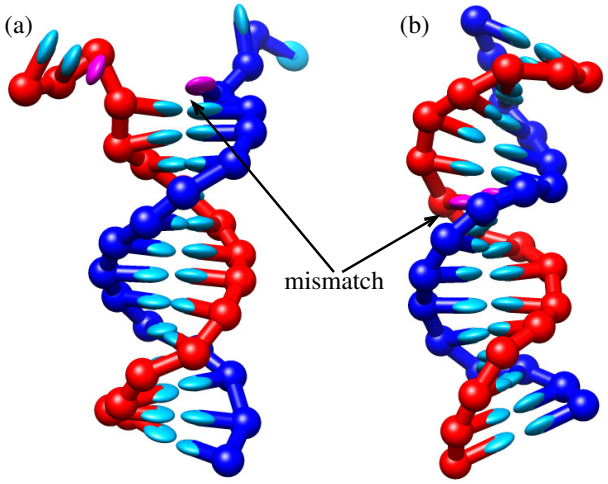

FIG. 15: Typical configurations of a duplex with 15 complementary bp and 1 internal mismatch at 335 K. a) Mismatch two bp from the end of the strands, with unpaired bases after the mismatch. b) Mismatch six bp from the end of the strand, enclosed by two intact helices.

representation of DNA while simultaneously providing a reasonably accurate description of the structural and mechanical properties of B-DNA and ssDNA.

The model provides a good quantitative representation of the three key thermodynamic processes that affect self-assembly: single-stranded stacking, duplex hybridization and hairpin formation. To our knowledge, this is the first coarse-grained model for which all three processes have been considered simultaneously.

The mechanical properties of DNA are also reasonably well represented by the model, with the singled-stranded persistence length (for stacked and unstacked bases) and double-stranded persistence length, stretch modulus and torsional modulus all of similar size to typical experimental estimates. Importantly, the inclusion of the stacking transition allows single strands to be unstacked and flexible, which facilitates the formation of hairpins as well as other DNA nanostructures for which single-stranded regions are important.

The model contains several simplifications, the most important of which are the lack of sequence dependence beyond the specificity of A-T and G-C bonds and the absence of explicit electrostatic interactions. Thus the model cannot predict screening effects without a new parameterization at each salt concentration. Furthermore, in its current state, the model may incorrectly represent structures that involve the close proximity of strands that are not bound to each other, where it would fail to capture the cumulative repulsion resulting from adjacent phosphate sites. The model is also only capable of representing structures involving B-DNA and ssDNA, and the equal groove size in the model may also mask subtle effects related to major and minor grooving.

Ignoring sequence heterogeneity dramatically lowers the number of parameters needed for the coarse-grained

model. It also simplifies the analysis of the physical processes, naturally generating results for an "average strand". This picture may be particularly advantageous when sequence effects obscure an important general trend. Of course there are also many processes where sequence heterogeneity is critical, for example, preferred sites for bubble nucleation. Such effects are not resolved by our model. Nevertheless, for many applications in DNA nanotechnology, sequence dependent effects beyond complementarity are not that critical to design or functionality. For example, in Ref. 1 we show that the entropy cost of bringing an anti-fuel strand in close proximity to the tweezer complex slows down the displacement-mediated detachment of the first arm of the fuel strand. Such predictions should be fairly robust and independent of sequence heterogeneity effects. We also show how metastable hairpin formation in the anti-fuel strand can further affect the free-energy profile and the related kinetics of the displacement process. Again, this general prediction should be fairly robust, but how it plays out for a particular set of tweezers will depend on how easily the anti-fuel strand sequence forms hairpins. For example, if the metastable hairpin formation is undesirable, then our predictions could be supplemented by methods such as the nearest-neighbour model in order to design strands that minimize hairpin formation.

Similarly, in the current paper we make a series of predictions that should be relevant to experiment. For example, we predict that a maximum in the magnitude of the enthalpy change of duplex formation, ΔH, should occur as the temperature nears the polynucleotide melting temperature and fraying begins to reduce the number of base pairs in the bound state. The exact location and magnitude of the maximum will depend on sequence-dependent effects such as the exact melting temperature, whether the end bases form weaker AT or stronger CG bonds that promote or repress fraying, respectively, as well as the thermodynamics of single-stranded stacking. But our prediction of a maximum in the absolute value of ΔH should be fairly robust.

We also predict that the T_m of a duplex containing a destabilizing motif should depend on the location of the motif in a temperature-dependent fashion. As the destabilizing motif is moved towards the centre of a duplex, the melting temperature should decrease before reaching a plateau. The distance from the end at which the plateau is observed will increase with T_m and the destabilizing effect of the motif. Both of these effects result from sufficiently generic properties that we expect them to be resilient to the approximations of the model.

When compared to the nearest-neighbor model, our model tends to slightly underpredict the effect of dangling ends, bulges, terminal mismatches and internal mismatches on the duplex melting temperature. Again, it should be kept in mind that for real DNA the effect of each of these motifs will depend very much on the exact sequence, whereas our predictions are for an average over all possible sequence permutations. Nevertheless, in a system where multiple kinetic traps are relevant, extra care should be taken when interpreting the simulations because the relative stabilities of different states could be somewhat misrepresented.

We also make some predictions for the conformational statistics of dsDNA, suggesting that single strands behave much more like freely-jointed chains with excluded volume than like worm-like chains. A particular consequence of this difference is that freely-jointed chains typically undergo much larger local kinking than worm-like chains with an equivalent effective persistence length.

Finally, we have demonstrated that a nearest-neighbour two-state model of duplex formation can be extended to incorporate stacking and fraying. This extension suggests a way to reconcile the appealing simplicity of nearest-neighbour models with temperature variation of both single- and double-stranded states. To develop such a model, however, would require a much greater consensus in the properties of single-stranded stacking and fraying than currently exists.

As it stands, we believe the model has the potential (both in terms of accuracy and computational efficiency) to open up a range of previously inaccessible general problems involving the interplay between single- and double-stranded DNA, including many aspects of DNA nanotechnology. For example, we are investigating the operation of a DNA walker,[27] the force-induced melting of DNA,[113] the assembly of a DNA tetrahedron[12] and binding of hairpins in the presence of a DNA catalyst.[68] The model may also be applied to biologically relevant processes such as the extrusion of cruciforms in supercoiled DNA containing inverted repeats.[31] Future work will aim to incorporate sequence-dependent interaction strengths (we note that much of the sequence dependence should arise from stacking), major and minor grooving and an implicit model for electrostatics, as well as comparing to atomistic simulations to improve the description of fluctuations on the base pair level.

[1] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, Phys. Rev. Lett. **104**, 178101 (2010).

[2] W. Saenger, *Principles of Nucleic Acid Structure* (Springer-Verlag, New York, 1984).

[3] J. D. Watson and F. H. C. Crick, Nature **171**, 737 (1953).

[4] P. J. Hagerman, Annu. Rev. Biophys. Biophys. Chem. **17**, 265 (1988).

[5] N. R. Kallenbach, R.-I. Ma, and N. C. Seeman, Nature **305**, 829 (1983).

[6] T. J. Fu and N. C. Seeman, Biochemistry **32**, 3211 (1993).

[7] H. Yan, S. H. Park, G. Finkelstein, J. H. Reif, and T. H. LaBean, Science **301**, 1882 (2003).

[8] E. Winfree, F. R. Liu, L. A. Wenzler, and N. C. Seeman, Nature **394**, 539 (1998).

[9] J. Malo, J. C. Mitchell, C. Venien-Bryan, J. R. Harris, H. Wille, D. J. Sherrat, and A. J. Turberfield, Angew. Chem. Int. Ed. **44**, 3057 (2005).
[10] J. Chen and N. C. Seeman, Nature **350**, 631 (1991).
[11] Y. Zhang and N. C. Seeman, J. Am. Chem. Soc. **116**, 1661 (1994).
[12] R. P. Goodman, I. A. T. Sharp, C. F. Tardin, C. M. Erben, R. M. Berry, C. F. Schmidt, and A. J. Turberfield, Science **310**, 1661 (2005).
[13] C. M. Erben, R. P. Goodman, and A. J. Turberfield, J. Am. Chem. Soc. **129**, 6992 (2007).
[14] W. M. Shih, J. D. Quispe, and G. F. Joyce, Nature **427**, 618 (2004).
[15] F. F. Andersen, B. Knudsen, C. L. P. Oliveira, R. F. Frohlich, D. Kruger, J. Bungert, M. Agbandje-McKenna, R. McKenna, S. Juul, C. Veigaard, et al., Nucl. Acids Res. **36**, 1113 (2008).
[16] Y. He, T. Ye, M. Su, C. Zhang, A. Ribbe, W. Jiang, and C. Mao, Nature **452**, 198 (2008).
[17] P. W. K. Rothemund, Nature **440**, 297 (2006).
[18] E. S. Andersen, M. Dong, M. M. Nielsen, K. Jahn, R. Subramani, W. Mamdouh, M. M. Golas, B. Sander, H. Stark, C. L. P. Oliveira, et al., Nature **459**, 73 (2009).
[19] S. M. Douglas, H. Dietz, T. Liedl, B. Högberg, F. Graf, and W. M. Shih, Nature **459**, 414 (2009).
[20] Z. Li, B. Wei, J. Nangreave, C. Lin, Y. Liu, Y. Mi, and H. Yan, J. Am. Chem. Soc. **131**, 13093 (2009).
[21] F. A. Aldaye and H. F. Sleiman, J. Am. Chem. Soc. **129**, 13376 (2007).
[22] J. Zimmermann, M. P. Cebulla, S. Monninghoff, and G. von Kiedrowski, Angew. Chem. Int. Ed. **47**, 3626 (2008).
[23] J. Bath and A. J. Turberfield, Nat. Nanotechnol. **2**, 275 (2007).
[24] B. Yurke, A. J. Turberfield, A. P. Mills, F. C. Simmel, and J. Neumann, Nature **406**, 605 (2000).
[25] W. B. Sherman and N. C. Seeman, Nano Lett. **4**, 1203 (2004).
[26] J.-S. Shin and N. A. Pierce, J. Am. Chem. Soc. **126**, 10834 (2004).
[27] J. Bath, S. J. Green, K. E. Allan, and A. J. Turberfield, Small **5**, 1513 (2009).
[28] S. J. Green, J. Bath, and A. J. Turberfield, Phys. Rev. Lett. **101**, 238101 (2008).
[29] C. Zhang, J. Yang, and J. Xu, Langmuir **26**, 1416 (2010).
[30] T. Liedl and F. C. Simmel, Nano Lett. **5**, 1894 (2005).
[31] R. R. Sinden, *DNA structure and function* (Academic Press Inc., London, 1994).
[32] M. Orozco, A. Pérez, A. Noy, and F. J. Luque, Chem. Soc. Rev. **32**, 350 (2003).
[33] R. Lavery, K. Zakrzewska, D. Beveridge, T. C. Bishop, D. A. Case, I. Cheatham, Thomas, S. Dixit, B. Jayaram, F. Lankas, C. Laughton, et al., Nucl. Acids Res. **38**, 299 (20010).
[34] A. Pérez, F. J. Luque, and M. Orozco, J. Am. Chem. Soc. **129**, 14739 (2007).
[35] C. Mura and A. J. McCammon, Nucl. Acids Res. **36**, 4941 (2008).
[36] S. Kannan and M. Zacharias, Phys. Chem. Chem. Phys. **11**, 10589 (2009).
[37] E. J. Sorin, Y. M. Rhee, B. J. Nakatani, and V. S. Pande, Biophys. J. **85**, 790 (2003).
[38] S. Kannan and M. Zacharias, Biophys. J. **93**, 3218 (2007).
[39] J. Marko, *Multiple aspects of DNA and RNA: From biophysics to bioinformatics* (Elsevier, Amsterdam, 2005), chap. 7, pp. 211–27, Les Houches Session LXXXII.
[40] D. Poland and H. A. Scheraga, J. Chem. Phys. **45**, 1464 (1966).
[41] J. SantaLucia, Jr., Proc. Natl. Acad. Sci. U.S.A **17**, 1460 (1998).
[42] J. SantaLucia, Jr. and D. Hicks, Annu. Rev. Biophys. Biomol. Struct. **33**, 415 (2004).
[43] R. Everaers, S. Kumar, and C. Simm, Phys. Rev. E **75**, 041918 (2007).
[44] T. Dauxois, M. Peyrard, and A. R. Bishop, Phys. Rev. E **47**, 684 (1993).
[45] N. B. Becker and R. Everaers, J. Chem. Phys. **130**, 135102 (2009).
[46] F. Lankaš, O. Gonzalez, L. M. Heffler, G. Stoll, M. Moakher, and J. H. Maddocks, Phys. Chem. Chem. Phys. **11**, 10565 (2009).
[47] M. Paliy, R. Melnik, and B. A. Shapiro, Physical Biology **7**, 036001 (2010).
[48] F. Trovato and V. Tozzini, J. Phys. Chem. B **112**, 13197 (2008).
[49] M. Sayar, B. Avşaroğlu, and A. Kabakçıoğlu, Phys. Rev. E **81**, 041916 (2010).
[50] P. D. Dans, A. Zeida, M. R. Machado, and S. Pantano, J. Chem. Theory Comput. **6**, 1711 (2010).
[51] K. Voltz, J. Trylska, V. Tozzini, V. Kurkal-Siebert, J. Langowski, and J. Smith, J. Comput. Chem. **29**, 1429 (2008).
[52] A. Morriss-Andrews, J. Rottler, and S. S. Plotkin, J. Chem. Phys. **132**, 035105 (2010).
[53] A. Louis, J. Phys.: Condens. Matter **14**, 9187 (2002).
[54] M. E. Johnson, T. Head-Gordon, and A. A. Louis, J. Chem. Phys. **126**, 144509 (2007).
[55] K. Drukker, G. Wu, and G. C. Schatz, J. Chem. Phys. **114**, 579 (2001).
[56] M. Sales-Pardo, R. Guimera, A. A. Moreira, J. Widom, and L. Amaral, Phys. Rev. E **71**, 051902 (2005).
[57] M. Kenward and K. D. Dorfman, J. Chem. Phys. **130**, 095101 (2009).
[58] F. Ding, S. Sharma, P. Chalasani, V. V. Demidov, N. E. Broude, and N. V. Dokholyan, RNA **14**, 1164 (2008).
[59] S. Pasquali and P. Derreumaux, J. Phys. Chem. B **114**, 11957 (2010).
[60] C. Hyeon and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **102**, 6789 (2005).
[61] C. Hyeon and D. Thirumulai, Biophys. J. **92**, 731 (2007).
[62] T. E. Ouldridge, I. G. Johnston, A. A. Louis, and J. P. K. Doye, J. Chem. Phys. **130**, 065101 (2009).
[63] E. J. Sambriski, V. Ortiz, and J. J. de Pablo, J.Phys.: Condens. Matter **21** (2009).
[64] E. J. Sambriski, D. C. Schwartz, and J. J. de Pablo, Biophys. J. **96**, 1675 (2009).
[65] S. Niewieczerzał and M. Cieplak, J. Phys.: Condens. Matter **21**, 474221 (2009).
[66] S. Pitchiaya and Y. Krishnan, Chem. Soc. Rev. **35**, 1111 (2006).
[67] M. C. Murphy, I. Rasnik, W. Chang, T. M. Lohman, and T. Ha, Biophys. J. **86**, 2530 (2004).
[68] J. Bois, S. Venkataraman, H. M. T. Choi, A. J. Spakowitz, Z. Wang, and N. A. Pierce, Nucl. Acids Res. **33**, 4090 (2005).
[69] J. B. Mills, E. Vacano, and P. J. Hagerman, J. Mol. Biol. **285**, 245 (1999).
[70] S. A. Harris, C. A. Laughton, and T. B. Liverpool, Nucl.

Acids Res. **36**, 21 (2008).
[71] T. Schlick, *Molecular Modeling and Simulation* (Springer-Verlag, New York, 2002).
[72] S. Whitelam, E. H. Feng, M. F. Hagan, and P. L. Geissler, Soft Matter **5**, 1521 (2009).
[73] M. Swart, T. van der Wijst, C. F. Guerra, and F. M. Bickelhaupt, J. Mol. Model. **13**, 1245 (2007).
[74] J. Sponer, P. Jurečka, I. Marchan, F. J. Luque, M. Orozco, and P. Hobza, Chem. Eur. J. **12**, 2854 (2006).
[75] J. Holbrook, M. Capp, R. Saecker, and M. Record, Biochemistry **38**, 8409 (1999).
[76] K. M. Guckan, B. A. Schweitzer, R. X.-F. Ren, C. J. Sheils, D. C. Tahmassebi, and E. T. Kool, J. Am. Chem. Soc. **122**, 2213 (2000).
[77] G. Torrie and J. P. Valleau, J. Comp. Phys. **23**, 187 (1977).
[78] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, J. Comput. Chem. **13**, 1011 (1992).
[79] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, J. Phys.: Condens. Matter **22**, 104102 (2010).
[80] P. Chen and C. M. Li, Small **3**, 1204 (2007).
[81] C. R. Calladine, H. R. Drew, B. F. Luisi, and A. A. Travers, *Understanding DNA* (Elsevier Academic Press, London, 2004).
[82] G. Vesnaver and K. J. Breslauer, Proc. Natl. Acad. Sci. U.S.A **88**, 3569 (1991).
[83] M. Leng and G. Felsenfeld, J. Mol. Biol. **15**, 455 (1966).
[84] R. M. Epand and H. A. Scheraga, J. Am. Chem. Soc. **89**, 3888 (1967).
[85] D. Pörschke, Biochemistry **15**, 1495 (1976).
[86] S. M. Freier, K. O. Hill, T. G. Dewey, L. A. Marky, and K. J. Breslauer, Biochemistry **20**, 1419 (1981).
[87] J. Zhou, S. Gregurick, S. Krueger, and F. Schwarz, Biophys. J. **90**, 544 (2006).
[88] P. J. Mikulecky and A. L. Feig, Biopolymers **82**, 38 (2006).
[89] J. Applequist and V. Damle, J. Am. Chem. Soc. **88**, 3895 (1966).
[90] I. Jelesarov, C. Crane-Robinson, and P. L. Privalov, J. Mol. Biol. **294**, 981 (1999).
[91] D. Poland and H. A. Scheraga, *Theory of Helix-Coil Transitions in Biopolymers: Statistical Mechanical Theory of Order-disorder Transitions in Biological Macromolecules* (Academic Press, New York, 1970).
[92] R. D. Blake and S. G. Delcourt, Nucl. Acids Res. **26**, 3323 (1998).
[93] M. D. Frank-Kamenetskii, Biopolymers **10**, 2623 (1971).
[94] D. Andreatta, S. Sen, J. L. Pérez Lustres, S. A. Kovalenko, N. P. Ernsting, C. J. Murphy, R. S. Coleman, and M. A. Berg, J. Am. Chem. Soc. **128**, 6885 (2006).
[95] S. Nonin, J.-L. Leroy, and M. Gueron, Biochemistry **34**, 10652 (1995).
[96] D. J. Patel and C. W. Hilbers, Biochemistry **14**, 2651 (1975).
[97] A. Tikhomirova, N. Taulier, and T. V. Chalikian, J. Am. Chem. Soc. **126**, 16387 (2004).
[98] C. Rivetti, C. Walker, and C. Bustamante, J. Mol. Biol. **280**, 41 (1998).
[99] P. Cifra, Polymer **45**, 5995 (2004).
[100] C. G. Baumann, S. B. Smith, V. A. Bloomfield, and C. Bustamante, Proc. Natl. Acad. Sci. USA **94**, 6185 (1997).
[101] D. M. Crothers, J. Drak, J. D. Kahn, and S. D. Levene, Methods Enzymol. **212**, 3 (1992).
[102] M. Vologodskaia and A. Vologodskii, J. Mol. Biol. **317**, 205 (2002).
[103] B. S. Fujimoto, G. P. Brewood, and J. M. Schurr, Biophys. J. **91**, 4166 (2006).
[104] Z. Bryant, M. D. Stone, J. Gore, S. B. Smith, N. R. Cozzarelli, and C. Bustamante, Nature **424**, 338 (2003).
[105] M. Wang, H. Yin, R. Landick, J. Gelles, and S. Block, Biophys. J. **72**, 1335 (1997).
[106] J. R. Wenner, M. C. Williams, I. Rouzina, and V. A. Bloomfield, Biophys. J. **82**, 3160 (2002).
[107] S. B. Smith, Y. Cui, and C. Bustamante, Science **271**, 795 (1996).
[108] T. Odijk, Macromolecules **28**, 7016 (1995).
[109] J. Gore, Z. Bryant, M. Nöllman, M. U. Le, N. R. Cozzarelli, and C. Bustamante, Nature **442**, 836 (2006).
[110] T. Lionnet, S. Joubaud, R. Lavery, D. Bensimon, and V. Croquette, Phys. Rev. Lett. **96**, 178102 (2006).
[111] D. K. Hendrix, S. E. Brenner, and S. R. Holbrook, Q. Rev. Biophys. **38**, 221 (2005).
[112] S. Kuznetsov, Y. Shen, A. S. Benight, and A. Ansari, Biophys. J. **81**, 2864 (2001).
[113] J. van Mameren, P. Gross, G. Farge, P. Hooijman, M. Modesti, M. Falkenberg, G. J. L. Wuite, and E. J. G. Peterman, Proc. Natl. Acad. Sci. U.S.A. **106**, 18231 (2009).
[114] Bases were counted as stacked if their interaction was stronger than $-0.60\,\mathrm{kcal\,mol^{-1}}$ (relative to a typical stacked interaction of $-6\,\mathrm{kcal\,mol^{-1}}$). Adjusting the cutoff to $-1.2\,\mathrm{kcal\,mol^{-1}}$ had a negligible effect.
[115] Our simulations are performed in the canonical ensemble, and hence should be described in terms of energy and entropy changes. We assume that, as dilute DNA strands contribute a very small partial pressure, discrepancies between constant volume and constant pressure results are small: we therefore use the term "enthalpy" to describe what are in fact energies in our model, for consistency with experimental literature.
[116] Throughout this article, lower case symbols represent enthalpy and entropy changes per pair of interacting bases, whereas capitals correspond to enthalpy and entropy changes per pair of interacting strands.
[117] Unless otherwise stated, all melting temperature calculations in this article used four simulations of 4×10^{10} MC steps, and were performed at a reference concentration of 3.36×10^{-4} M. Simulations of duplexes with more than 12 bp necessitated using a larger periodic cell, and hence a lower concentration. The fraction of bound duplexes was scaled to the higher concentration assuming the separate species are approximately ideal, as justified in Ref. 79.

Appendix A: Model details and parameterization

The current model is based on that introduced in Ref. 1, with some changes introduced to give duplexes more flexibility (having performed a wider range of structural tests, the stiffness was found to be overestimated in the old version). Truncated interactions have also been quadratically smoothed (making the potential continuous and differentiable, allowing simulation with methods like Langevin dynamics). Although this introduces further parameters, the thermodynamic and structural proper-

ties are largely unaffected by the details of smoothing.

The functional forms used in the interactions are given below:

- FENE spring (used to connect backbones):

$$V_{\rm fene}(r) = -\frac{k}{2}\ln\left(1-\frac{(r-r_0)^2}{\Delta^2}\right). \tag{A1}$$

- Morse potential (used for stacking and H-bonding):

$$V_{\rm Morse}(r,\epsilon,r_0,a) = \epsilon\big(1-\exp\left(-(r-r_0)a\right)\big)^2. \tag{A2}$$

- Harmonic potential (used for cross-stacking):

$$V_{\rm harm}(r,\epsilon,r_0) = \frac{\epsilon}{2}\left(r-r_0\right)^2. \tag{A3}$$

- Lennard - Jones potential (used for soft repulsion);

$$V_{\rm LJ}(r,\epsilon,\sigma) = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12}-\left(\frac{\sigma}{r}\right)^{6}\right). \tag{A4}$$

- Quadratic terms (used for modulation)

$$V_{\rm mod}(\theta,a,\theta_0) = 1-a(\theta-\theta_0)^2 \tag{A5}$$

- Quadratic smoothing terms:

$$V_{\rm smooth}(x,b,x_c) = b(x_c-x)^2 \tag{A6}$$

These functional forms are combined to give the following smooth and differentiable functions:

- The radial part of the stacking and hydrogen-bonding potentials:

$$f_1(r) = \begin{cases} V_{\rm Morse}(r,\epsilon,r_0,a) - V_{\rm Morse}(r_c,\epsilon,r_0,a)) & \text{if } r^{low} < r < r^{high}, \\ \epsilon V_{\rm smooth}(r,b^{low},r_c^{low}) & \text{if } r_c^{low} < r < r^{low}, \\ \epsilon V_{\rm smooth}(r,b^{high},r_c^{high}) & \text{if } r^{high} < r < r_c^{high}, \\ 0 & \text{otherwise.} \end{cases} \tag{A7}$$

- The radial part of the cross-stacking potential:

$$f_2(r) = \begin{cases} V_{\rm harm}(r,\epsilon,r_0) - V_{\rm harm}(r_c,\epsilon,r_0) & \text{if } r^{low} < r < r^{high}, \\ \epsilon V_{\rm smooth}(r,b^{low},r_c^{low}) & \text{if } r_c^{low} < r < r^{low}, \\ \epsilon V_{\rm smooth}(r,b^{high},r_c^{high}) & \text{if } r^{high} < r < r_c^{high}, \\ 0 & \text{otherwise.} \end{cases} \tag{A8}$$

- The radial part of the excluded volume potential:

$$f_3(r) = \begin{cases} V_{\rm LJ}(r,\epsilon,\sigma) & \text{if } r < r^\star, \\ \epsilon V_{\rm smooth}(r,b,r_c) & \text{if } r^\star < r < r_c, \\ 0 & \text{otherwise.} \end{cases} \tag{A9}$$

- The angular modulation factor used in stacking, hydrogen bonding and cross-stacking:

$$f_4(\theta) = \begin{cases} V_{\rm mod}(\theta,a,\theta_0) & \text{if } \theta_0-\Delta\theta^\star < \theta < \theta_0+\Delta\theta^\star, \\ V_{\rm smooth}(\theta,b,\theta_0-\Delta\theta_c) & \text{if } \theta_0-\Delta\theta_c < \theta < \theta_0-\Delta\theta^\star, \\ V_{\rm smooth}(\theta,b,\theta_0+\Delta\theta_c) & \text{if } \theta_0+\Delta\theta^\star < \theta < \theta_0+\Delta\theta_c, \\ 0 & \text{otherwise.} \end{cases} \tag{A10}$$

- Another modulating term which is used to impose right handedness (effectively a one-sided modulation):

$$f_5(\phi) = \begin{cases} 1 & \text{if } x > 0, \\ V_{\rm mod}(x,a,0) & \text{if } x^\star < x < 0, \\ V_{\rm smooth}(x,b,x_c) & \text{if } x_c < x < x^\star, \\ 0 & \text{otherwise.} \end{cases} \tag{A11}$$

The potentials and parameters used to describe each interaction are listed in the Table. II. When more than

Interaction	Functional form	Parameters			
backbone spring $V_{backbone}$	$V_{\rm fene}(r_{\rm backbone})$	$k = 2$	$\Delta = 0.25$	$r_0 = 0.7525$	
hydrogen bond V_{HB}	$f_1(r_{\rm bond})$	$\epsilon = 1.077$	$a = 8$	$r_0 = 0.4$	$r^{low} = 0.34$
				$r_c = 0.75$	$r^{high} = 0.70$
	$f_4(\theta_1)$	$a = 1.50$	$\theta_0 = 0$	$\Delta\theta^\star = 0.70$	
	$f_4(\theta_2)$	$a = 1.50$	$\theta_0 = 0$	$\Delta\theta^\star = 0.70$	
	$f_4(\theta_3)$	$a = 1.50$	$\theta_0 = 0$	$\Delta\theta^\star = 0.70$	
	$f_4(\theta_4)$	$a = 0.46$	$\theta_0 = \pi$	$\Delta\theta^\star = 0.70$	
	$f_4(\theta_7)$	$a = 4.00$	$\theta_0 = \pi/2$	$\Delta\theta^\star = 0.45$	
	$f_4(\theta_8)$	$a = 4.00$	$\theta_0 = \pi/2$	$\Delta\theta^\star = 0.45$	
stacking V_{stack}	$f_1(r_{\rm stack})$	$\epsilon = 1.2145$	$a = 6$	$r_0 = 0.4$	$r^{low} = 0.32$
		$+2.6568\,kT$		$r_c = 0.9$	$r^{high} = 0.75$
	$f_4(\theta_4)$	$a = 1.30$	$\theta_0 = 0$	$\Delta\theta^\star = 0.8$	
	$f_4(\theta_5)$	$a = 0.90$	$\theta_0 = 0$	$\Delta\theta^\star = 0.95$	
	$f_4(\theta_6)$	$a = 0.90$	$\theta_0 = 0$	$\Delta\theta^\star = 0.95$	
	$f_5(\cos(\phi_1))$	$a = 2.00$	$x^\star = -0.65$		
	$f_5(\cos(\phi_2))$	$a = 2.00$	$x^\star = -0.65$		
cross-stacking V_{c_stack}	$f_2(r_{\rm cstack})$	$\epsilon = 47.5$	$r_0 = 0.575$	$r_c = 0.675$	$r^{low} = 0.495$
					$r^{high} = 0.655$
	$f_4(\theta_1)$	$a = 2.25$	$\theta_0 = 2.35$	$\Delta\theta^\star = 0.58$	
	$f_4(\theta_2)$	$a = 1.70$	$\theta_0 = 1.00$	$\Delta\theta^\star = 0.68$	
	$f_4(\theta_3)$	$a = 1.70$	$\theta_0 = 1.00$	$\Delta\theta^\star = 0.68$	
	$f_4(\theta_4) + f_4(\pi - \theta_4)$	$a = 1.50$	$\theta_0 = 0$	$\Delta\theta^\star = 0.65$	
	$f_4(\theta_7) + f_4(\pi - \theta_7)$	$a = 1.70$	$\theta_0 = 0.875$	$\Delta\theta^\star = 0.68$	
	$f_4(\theta_8) + f_4(\pi - \theta_8)$	$a = 1.70$	$\theta_0 = 0.875$	$\Delta\theta^\star = 0.68$	
excluded volume V_{exc}	$f_3(r_{\rm ex1}) + f_3(r_{\rm ex2})$ $+f_3(r_{\rm ex3}) + f_3(r_{\rm ex4})$	$\epsilon = 2.00$	$\sigma_1 = 0.70$	$r_1^\star = 0.675$	
			$\sigma_2 = 0.33$	$r_2^\star = 0.32$	
			$\sigma_3 = 0.515$	$r_3^\star = 0.50$	
			$\sigma_4 = 0.515$	$r_4^\star = 0.50$	

TABLE II: Parameter values in the model. All lengths are defined with respect to a reduced lengthscale (1 unit = 8.518Å), all angles are given in radians and all energies are defined with respect to a reduced temperature ($kT = 0.1$ corresponding to 300 K). The variables in the potential are defined in Fig. 16

one function is listed for an interaction, the total interaction is a product of all the terms. Given the parameters of the main part of the interaction (for example, ϵ, r_0, a and r_c for the $V_{\rm Morse}$ part of $f_1(r)$), the parameters of the smoothed cutoff regions are uniquely determined by ensuring continuity and differentiability at the boundaries (r^{low} and r^{high} for $f_1(r)$). The nucleotide geometry and definition of the angles and vectors used in the potential are shown in Fig. 16.

The potential of the system is given by:

$$\begin{aligned} V \quad &= \sum \big(V_{backbone} + V_{stack} + V'_{exc}\big) \\ &+ \sum_{\text{other pairs}}^{\text{nn}} \big(V_{HB} + V_{c\_stack} + V_{exc}\big), \end{aligned} \tag{A12}$$

where the sum over nn runs over consecutive bases within strands, and V'_{exc} is equal to V_{exc} except that it does not include an $f_4(r_{\rm ex1})$ term. Note the directional dependence in the stacking interaction: the angles are defined between normal vectors of bases and a vector joining bases in the 3' to 5' direction. Only complementary base pairs posses non-zero hydrogen-bond energies.

To ensure right-handed helices, the modulation of stacking interactions is somewhat subtle, involving chiral terms. Consider two consecutive bases in a strand, i and j, with $i \to j$ corresponding to the $3' \to 5'$ direction. The angles θ_5 and θ_6 are defined as the angles between the normals of i and j and $\mathbf{r}^{\mathbf{ij}}_{\rm stack}$ (with $\mathbf{r}^{\mathbf{ij}}_{\rm stack}$ being defined as the vector from the stacking site of i to that of j). Thus, stacked bases have normals pointing in the $3' \to 5'$ direction, allowing the definition of a local axis. The angles ϕ_1 and ϕ_2, defined in terms of this axis, provide helicity. For each base we define a normalized vector $\hat{\mathbf{v}}^\alpha_{\rm helicity} = \hat{\mathbf{r}}^{ij} \times \hat{\mathbf{r}}^\alpha_{\rm back-base}$, where $\alpha = i, j$ and $\hat{\mathbf{r}}^\alpha_{\rm back-base}$

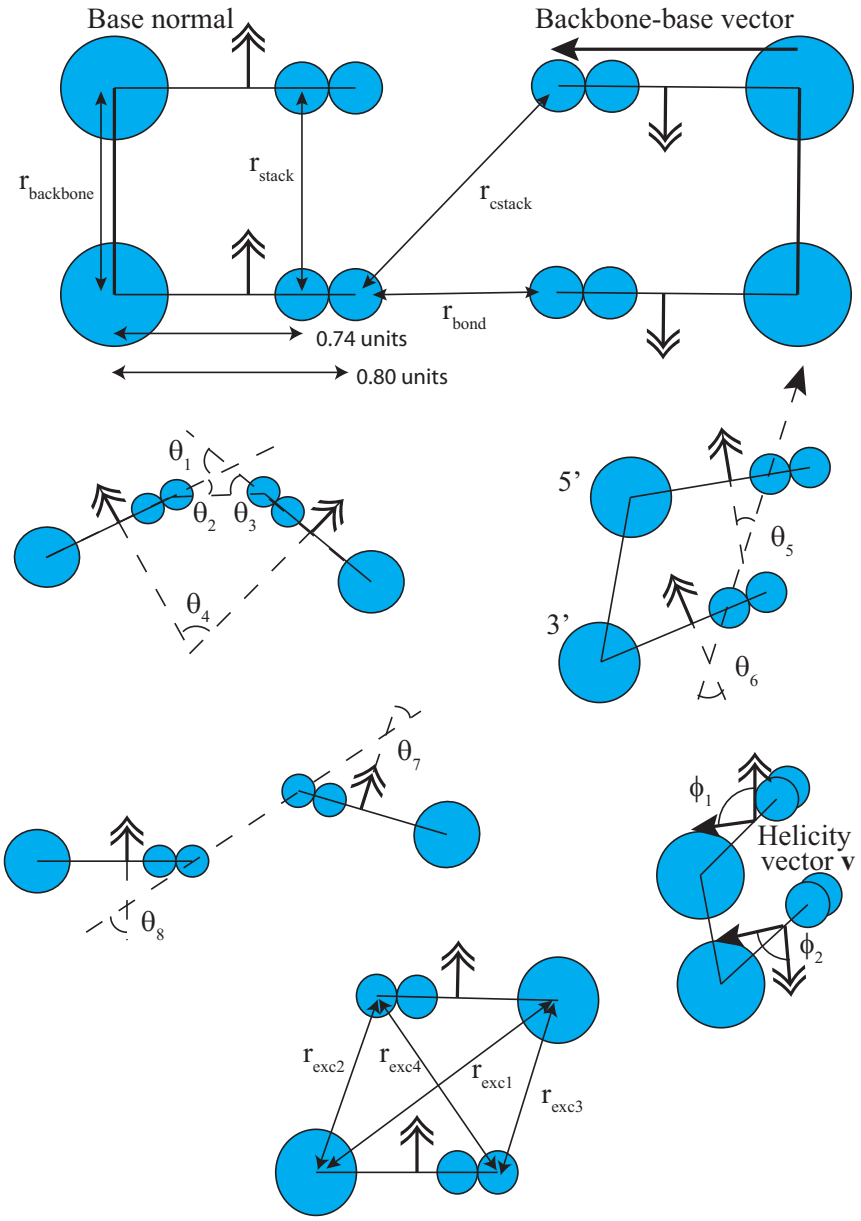

FIG. 16: Illustration of variables used in the potential of the DNA model.

is the normalized backbone site to stack site vector of base α. ϕ_1 and ϕ_2 are the angles between $\hat{\mathbf{v}}^\alpha$ and the base normals: for a right handed helix, these are $< \pi/2$ and the stacking interaction is modulated to disfavour greater angles.

Appendix B: Statistical model of stacking

It is instructive to characterize the thermodynamics of the model using a simpler, statistical model, as it highlights the causes of certain behaviour. We model the stacking transition using a statistical description based on that of Poland and Scheraga.[91] In this model, a given

pair of neighbours can be either stacked or unstacked, and the list of stacked pairs specifies the system configuration.

If each stacking pair were independent, the contribution to the partition function from a configuration (its relative probability of occurring) would be given by:

$$Z_{\text{config}} = z_0 u^{N_i} v^{N_j} \tag{B1}$$

where u and v represent the contributions to the partition function ("statistical weight") of a stacked and an unstacked pair respectively, N_i and N_j are the number of stacked an unstacked pairs and z_0 denotes the trivial contribution from translation and orientation of the whole strand. As discussed in Section III B 1, the excluded volume of nucleotides means that pairs of neighbours are not independent. To deal with this, we introduce two new parameters. The statistical weight of a continuous section of n stacked pairs is now given by:

$$u(n) = \sigma u^n w^x, \tag{B2}$$

with x being equal to the number of bases in the run of stacked pairs that lie at the end of the strand. n unstacked pairs contribute the same statistical weight as before:

$$v(n) = v^n. \tag{B3}$$

If σ and w are unity, each neighbour pair is independent, and we return to Eqn. B1. σ takes the role of a cooperativity parameter: for $0 < \sigma < 1$, stacking is cooperative, in that configurations with multiple separate regions of stacking are disfavoured, and for $\sigma > 1$ stacking is anticooperative. w accounts for end effects: for $0 < w < 1$, end bases are less likely to stack, and for $w > 1$ the opposite is true.

Using these definitions, the total partition function for a strand of length l becomes:

$$Z_l = \sum_{\{n_i, m_j; l\}} z_0 w^x \prod_i \sigma u^{n_i} \prod_j v^{m_j}. \tag{B4}$$

Here, $\{n_i, r_j; l\}$ specifies a configuration, n_i being the number of stacked pairs in the i^{th} contiguous sequence of stacked neighbours, m_j being the number of unstacked pairs in the m^{th} sequence of unstacked bases and $x = \sum_i x_i$ is the total number of bases at the end of the strand involved in stacking.

Defining $t = u/v$, $n = \sum_i n_i$ and letting p be the total number of stacked regions, we obtain:

$$Z_l = Z_l^u \sum_{\{n_i, m_j; l\}} w^x \sigma^p t^n. \tag{B5}$$

with $Z_l^u = z_0 v^{l-1}$ being the partition function of a completely unstacked strand. To compare directly with simulations, we require the ratio of the probability of observing r stacked pairs to the probability of observing a completely unstacked strand:

$$\frac{Z_l(r)}{Z_l^u} = \sum_{\{n=r; l\}} w^x \sigma^p t^n = t^r \sum_{x=0}^{2} w^x \sum_p \sigma^p \Omega_{\{x,r,p;l\}}, \tag{B6}$$

with $\Omega_{\{x,r,p;l\}}$ defined as the number of distinct configurations of length l with r stacked pairs, of which x are at the end of the strand, divided between p contiguous regions of stacking. The advantage of this representation is that finding $\Omega_{\{x,r,p;l\}}$ is simply a matter of combinatorics. It can be shown that:

$$\Omega_{\{x,r,p;l\}} = \frac{(1+\delta_1^x)(r-1)!(l-r-2)!}{(r-p)!(p-1)!(l-r-2-p+x)!(p-x)!}, \tag{B7}$$

for all possible values of x, r and p for a strand of length l, with the exception that $\Omega_{\{0,0,0;l\}} = \Omega_{\{2,l-1,1;l\}} = 1$.

We assume that the temperature dependence of stacking is manifested in the parameter t, which is defined as $t = \exp(-\Delta h^{st}/RT + \Delta s^{st}/R)$, with Δh^{st} and Δs^{st} representing the (assumed constant) enthalpy and entropy changes associated with stack formation. As w and σ arise from excluded volume effects, they are assumed to be entropic and hence temperature independent. We fitted this 4-parameter model to data obtained in simulations, the results are shown in Section III B 1.

Appendix C: Statistical model for duplex formation

Eqn. 8 assumes a constant entropy and enthalpy difference between bound and unbound states. It is well known, however, that ΔS_l and ΔH_l should both become more negative with temperature, as the unbound strands become increasingly disordered due to unstacking.[75,82,88,90] Using the formalism of Appendix B, we can factor out this effect:

$$\frac{[A_l B_l]}{[A_l][B_l]} = v\frac{Z_{ll}}{Z_l^2} = \frac{\exp\Big(-\beta\big(\Delta H_l' - T\Delta S_l'\big)\Big)(Z_l^u)^2}{Z_l^2}, \tag{C1}$$

where in this case $\Delta H_l'$ and $\Delta S_l'$ are the enthalpy and entropy difference between the duplex and unstacked single-stranded macrostates.

Although fitting to Eqn. C1 with constant $\Delta H_l'$ and $\Delta S_l'$ was more successful than assuming constant ΔH_l and ΔS_l, it overcorrected for the variations in ΔS_l and ΔH_l with temperature. The failure resulted from neglecting the changes in the bound state with temperature, which were dominated by two effects:

- As temperature increases, increased fraying leads to smaller entropy and enthalpy differences between typical bound states and completely unstacked single strands, as bound states become more disordered.

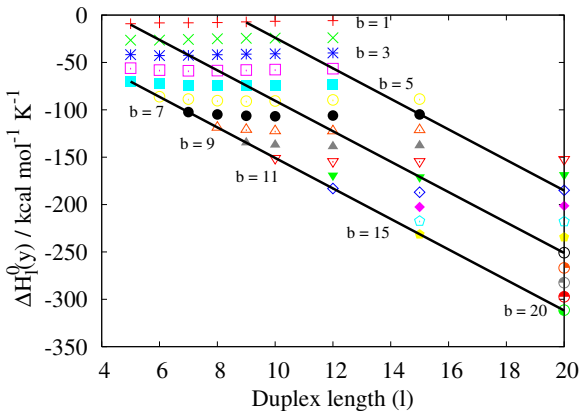

FIG. 17: $\Delta H_l^0(y)$ against duplex length l. Points are styled according to the total number of bonds formed, $b = l - y$. The solid lines are linear fits for the dependence of $\Delta H_l^0(y)$ on l for fixed y.

- Frayed ends themselves undergo a stacking transition, once more resulting in the entropy and enthalpy of bound states relative to unstacked strands becoming less negative with temperature.

To incorporate these effects within a statistical model, we separately consider the entropy and enthalpy differences between unstacked single strands and macrostates with y out of l possible base pairs formed. We then approximately adjust for stacking of the frayed ends by treating the $2(l-y)$ unpaired bases as undergoing stacking with the same Δh^{st} and Δs^{st} given in Section III B 1. Cooperativity and end effects are ignored as it would be difficult to include them consistently when stacking is initiated adjacent to a duplex region.

We thus define $Z_{ll}(y)$ as:

$$Z_{ll} = \sum_y Z_{ll}(y), \tag{C2}$$

and $Z_{ll}^u(y)$ as the contribution to $Z_{ll}(y)$ in which none of the unpaired bases are stacked. $Z_{ll}^u(y)$ is approximated by:

$$Z_{ll}(y) = Z_{ll}^u(y)\Big(1 + \exp\big(-\beta(\Delta h^{st} - T\Delta s^{st})\big)\Big)^{2(l-y)}, \tag{C3}$$

Our hypothesis is that the enthalpy and entropy differences between unstacked single strands and the states contributing to $Z_{ll}^u(y)$ should be approximately constant for given l and y, as the temperature variation due to breaking stacks and fraying has been factored out. The values of $Z_{ll}(y)/Z_l^2$ were extracted from the fraying data, and $Z_{ll}^u(y)/(Z_l^u)^2$ inferred using Eqns. B6 and C3. Fitting to

$$v\frac{Z_{ll}^u(y)}{(Z_l^u)^2} = \exp\Big(-\beta\big(\Delta H_l^0(y) - T\Delta S_l^0(y)\big)\Big) \tag{C4}$$

with constant $\Delta H_l^0(y)$ and $\Delta S_l^0(y)$ (which represent the enthalpy and entropy differences between unstacked single strands and the states contributing to $Z_{ll}^u(y)$) was very successful.

Furthermore, as shown in Fig. 17, $\Delta H_l^0(y)$ (and $\Delta S_l^0(y)$, which is not shown) are to an excellent approximation linear in l for fixed y. Thus, having factored out sources of variation with temperature in the initial and final states, we arrive at a statement similar to the initial hypothesis of the nearest-neighbour model: adding an extra bp to a helix (i.e., increasing the length of the strands by one base, and forming one extra base pair, so that the number of unpaired bases is constant) contributes a constant enthalpy and entropy change relative to unstructured single strands.

This finding suggests an extension of the nearest-neighbour model to non-two-state behaviour to incorporate fraying and stacking, and thus predict the values of $\Delta S(T)$ and $\Delta H(T)$ for oligonucleotides. To achieve this description, fraying and stacking transitions must be sufficiently well characterized, and the assumption that helix stability is predominantly due to nearest-neighbour effects must hold, as it does in our model. It should be noted that at low temperatures, certain oligomers may also have significant contributions to the single-stranded state from hairpins, which are not incorporated into this model.

We are finally in a position to characterize the hybridization transition with completely temperature independent parameters. Combining Eqns. 8, B6, C2, C3 and C4, we find:

$$K_{eq} = \exp\Big(-\beta\big(\Delta H_l - T\Delta S_l\big)\Big) = v\frac{Z_{ll}}{Z_l^2} = \frac{\sum_y \exp\Big(-\beta\big(\Delta H_l^0(y) - T\Delta S_l^0(y)\big)\Big)\Big(1 + \exp\big(-\beta(\Delta h^{st} - T\Delta s^{st})\big)\Big)^{2(l-y)}}{\sum_r \exp\big(-\beta(\Delta h^{st} - T\Delta s^{st})\big)^r \sum_x^2 w^x \sum_p \sigma^p \Omega_{\{x,r,p;l\}}} \tag{C5}$$

$$\Delta H_l = -\frac{\mathrm{d}}{\mathrm{d}\beta}\ln K_{eq} = \frac{\sum_y \Big(\Delta H_l^0(y) + 2(l-y)\Delta h^{st}\frac{\exp(-\beta(\Delta h^{st} - T\Delta s^{st}))}{1+\exp(-\beta(\Delta h^{st} - T\Delta s^{st}))}\Big) Z_{ll}(y)}{Z_{ll}} - 2\frac{\sum_r (r\Delta h_{st} Z_l(r))}{Z_l}. \tag{C6}$$

Eqn. C6 is used in Section III B 4 to produce the fit of ΔH_{15} to simulations. The first term gives the enthalpy of duplexes with respect to unstacked single strands and the second term the enthalpy of two single strands with respect to their unstacked state. As can be seen, the agreement is good over a wide range of temperatures. Had hairpins been possible in the simulations of Section III B 4, they may have distorted the enthalpy at temperatures far below T_m. Hairpins were excluded to make the interpretation of results clearer, and their presence would have made the single-stranded state's enthalpy more negative. This change would have led to a smaller transition enthalpy between single strands and duplexes at these temperatures.