The Journal of Chemical Physics





# Investigation Into the Mechanism and Dynamics of DNA Association and Dissociation Utilizing Kinetic Monte Carlo Simulations

Ryan J. Menssen,[1] Gregory J. Kimmel,[2] and Andrei Tokmakoff[1, a)]

[1)] *Department of Chemistry, James Franck Institute, and Institute for Biophysical Dynamics, The University of Chicago, 929 East 57th Street, Chicago, Illinois 60637, United States*

[2)] *Moffitt Cancer Center, 12902 USF Magnolia Drive, Tampa, Florida 33612, United States*

(Dated: 18 December 2020)

In this work we present a kinetic Markov state Monte Carlo model designed to compliment temperature-jump (T-jump) infrared spectroscopy experiments probing the kinetics and dynamics of short DNA oligonucleotides. The model is designed to be accessible to experimental researchers in terms of both computational simplicity and expense while providing detailed insights beyond those provided by experimental methods. The model is an extension of a thermodynamic lattice model for DNA hybridization utilizing the formalism of the nucleation-zipper mechanism. Association and dissociation trajectories were generated utilizing the Gillespie algorithm and parameters determined via fitting the association and dissociation timescales to previously published experimental data. Terminal end fraying, experimentally observed following a rapid T-jump, in the sequence 5'-ATATGCATAT-3' was replicated by the model which also demonstrated that experimentally observed fast dynamics in the sequences 5'-C$(AT)_n$G-3', where $n$ = 2-6, were also due to terminal end fraying. The dominant association pathways, isolated by transition pathway theory, showed two primary motifs: initiating at or next to a G:C base pair, which is enthalpically favorable and related to the increased strength of G:C base pairs, and initiating in the center of the sequence, which is entropically favorable and related to minimizing the penalty associated with the decrease in configurational entropy due to hybridization.

[a)] Electronic mail: tokmakoff@uchicago.edu.

## I. INTRODUCTION

The kinetics and dynamics of nucleic acids have been shown to be a significant factor in both biological processes and a number of applications outside of biology. DNA duplexes are dynamic molecules that constantly undergo motions including configurational changes that can, but don't always, include a localized loss of base pairs. A large number of distortions to the "ideal" double helix have been identified that significantly impact the physical properties of DNA and play a role in biological processes.[1,2] DNA duplexes also dynamically break and form base pairs ranging from the fluctuations involving a single base pair, or base flipping,[3] to continuous stretches of broken base pairs, i.e. bubbles.[4] It has been proposed that the dynamics through which a base adopts an extrahelical configuration play a significant role in biological processes, for instance in the recognition of the target base by an enzyme.[3,5] DNA breathing modes, where base pairs dynamically open and close along a stretch of the double helix, have been thoroughly studied and shown to play a role in a large number of biological processes. Two examples among many others are the recognition of thymine dimers formed due to UV damage and the initiation of DNA transcription.[4,6–10]

While association mechanisms of short oligonucleotides have been discussed since the 1950s many open questions remain. One of the primary conceptual pictures used to discuss DNA association is the nucleation-zipper mechanism.[11–17] In the nucleation-zipper picture there are three distinct phases of the association process. In the first phase, two single DNA strands, known as monomers, diffuse together and form the first base pair, which is largely considered to be diffusion controlled.[13,17] Upon forming the first base pair the process enters a phase commonly known as the pre-equilibrium.[11,13,17] During this phase the partially formed duplex rapidly interconverts between a variety of configurations, all of which are thermodynamically metastable but include at least one intact base pair. The partially formed duplex remains in the pre-equilibrium until it either fully dissociates or the partially formed duplex forms the "critical nucleus." The critical nucleus contains the minimum number of intact base pairs such that the partially formed duplex is stable and has a significantly greater probability of rapidly zipping up the remaining bases relative to returning to the pre-equilibrium.[11,13] That is, it signifies that the transition state to hybridization has been crossed. Finally, downhill zipping describes the rapid sequential formation of the remaining base pairs from the critical nucleus to the fully formed duplex, a process that occurs orders

of magnitude faster than the formation of the critical nucleus.[11,12,15,17]

Recently, computational advances in this field have significantly outpaced experimental work. The hybridization, dehybridization, and dynamics of short DNA oligos have been studied utilizing both all-atom[18–20] and coarse-grained[21–24] molecular dynamics (MD) simulations. Computational studies have observed a number of different association mechanisms such as the "slithering" and "inch-worm" mechanisms that may play a role in the formation of DNA duplexes. The formation of non-native contacts occurs in both of these mechanisms suggesting that these contacts play a significant role in facilitating the hybridization of DNA. The significance of the "slithering" and "inch-worm" mechanisms relative to the "zippering" mechanism carries a sequence dependence.[21,22,25,26] However, the sophistication of modern computational methods, particularly all-atom or coarse-grained MD simulations, makes them not easily accessible to experimentally focused researchers due to their computational expense and the complexity of running and analyzing them. As a result, it would be useful to have a model that provides more detailed mechanistic insight than can be derived from experiment alone, while remaining accessible to experimentally focused researchers. Bridging this gap will allow experimentally focused researchers to harness the power of computational modeling to gain additional dynamical and mechanistic insights into DNA systems without making a significant resource investment.

Kinetic Monte Carlo methods, which simulate trajectories of a system evolving in a state space, provide computational and conceptual simplicity and are commonly utilized to study DNA and other biomolecules. Such models can be sophisticated and provide detailed structural information, including instances where their state space is based on Markov state models.[27–29] In the case of DNA, simpler approaches could use states generated by a lattice model,[30] or the thermodynamic nearest neighbor (NN) model.[31] In this paper, we describe a simplified kinetic Monte Carlo model for DNA hybridization that builds on states obtained from a new lattice model,[32] and implements Monte Carlo simulations in continuous-time using the Gillespie algorithm.[33,34] The Gillespie algorithm has been used to study DNA in a variety of contexts including breathing dynamics[7,35] and the hybridization of a variety of DNA motifs and structures.[31,36–38]

The Markov state kinetic Monte Carlo model presented here is applied to a number of DNA sequences of differing length and base pair composition to probe the different variables that impact the dynamics, mechanism, and energetic driving forces of DNA hybridiza-

tion and dehybridization. The main body of experimental results the model is applied to are recent temperature-jump (T-jump) dehybridization measurements on the sequences 5'-C(AT)$_n$G-3', where $n$ = 2-6, which will be referred to as the length series or CG-ends.[39] Additional T-jump experiments are analyzed for the sequence 5'-ATATATATAT-3', referred to as AT-all, and the sequence, 5'-ATATGCATAT-3', or GC-core.[40] The purpose of utilizing these varied sequences is twofold. First, to test how well the model matches experimental results when the number and location of G:C base pairs and the overall length of the sequence are altered. Second, these sequences provide a window into understanding how these variables effect the dynamics, mechanism, and energetic driving forces behind DNA association and dissociation.

## II. MODEL DESCRIPTION

The Markov state kinetic Monte Carlo model presented here utilizes a state space obtained from a lattice model that has been validated against short DNA oligonucleotides of differing length and composition.[32] In this lattice model, the thermodynamics of base pair association and dissociation are drawn from the NN parameters published by SantaLucia,[41] the configurational entropy of partially formed duplexes is determined from a self-avoiding random walk on a cubic lattice, and the translational entropy of the monomer and dimer states at fixed concentration comes from a separate cubic lattice that assigns configurations with two monomers residing in adjacent cells to a dimer. Drawing on the common nucleation-zipper mechanism,[11,13–16,42–44] where base pairs are added sequentially in hybridization, the kinetic model builds trajectories by stepping through configurations with each step adding or removing a single base pair. The thermodynamics from the lattice model are used in combination with detailed-balance relations in the calculation of rates for moving between states in the kinetic model.

### A. Reaction Scheme

To help visualize the state space and how trajectories move through it, Figure 1 shows the reaction scheme for the sequence 5'-CATATG-3'. The main reaction coordinate trajectories progress along is the number of intact base pairs ($N_{\mathrm{BP}}$). This is demonstrated in Figure 1

FIG. 1. Kinetic model reaction scheme for the sequence 5'-CATATG-3'. The step forming the first base pair is referred to as the nucleation step while all other steps are referred to as propagation steps. The different possible configurations are shown below each state, with each row representing a different possible configuration of intact (black) and broken (white) base pairs. The lines between states denote the allowed moves for transitioning between configurations.

by the states denoted either as 2M, designating the monomer state, two dissociated single strands of DNA, and $D_n$ which designates a state with $n$ intact base pairs. Within each possible $N_{BP}$, the model can adapt any configuration where all intact base pairs are in-register, meaning they are aligned properly for the formation of the fully formed dimer, and the intact base pairs form a continuous stretch. All other configurations are excluded. The configurations of unpaired bases, either frayed ends or monomers, are not explicitly considered in the kinetic model, though they are included indirectly when calculating the free energy of each configuration. In doing so, the model in essence assumes that free terminal chains sample all possible configurations very quickly relative to the making and breaking of base pairs. The movement of the two monomers as they diffuse and encounter one another is also not explicitly resolved. When moving between states only one base pair can form or break during a single step and the value of $N_{BP}$ must change. As a result a base pair may only form or break at the end of the continuous stretch of intact base pairs, with the exception of moving between the 2M and $D_1$ states. The first base pair can form anywhere along the sequence with each position sharing the same probability.

### B. Model Parameters and Rate Determination

Moves that form a base pair are broken down into two categories: nucleation, the formation of the first base pair, and propagation, the formation of a base pair next to an already formed base pair. The forward rates utilize three parameters, $k_f$, $\sigma_i$, and $\beta$ that are conceptually similar to the parameters utilized in the model of Craig, Crothers, and Doty[11], but with somewhat different definitions. $k_f$, which has units of $s^{-1}$, is the zipping rate constant, and can be thought of as the "speed limit" for forming a base pair next to an already formed base pair.[11–13,45] $\sigma_i$ and $\beta$ are unitless scaling factors that describe the attenuation of the base pair formation rate relative to $k_f$.[11].

The attenuation factor for the nucleation step, $\beta$, describes the rate of forming the first base pair in a nucleation step, $k_N$, relative to the zipping speed limit, i.e. $k_N = \beta k_f$. Calculating $\beta$ assumes nucleation can be broken down into two independent and sequential steps, the single strands diffusing into the proper orientation followed by the formation of the base pair. With this assumption, the timescale for forming the first base pair can be written as

$$\tau_N = \tau_D + \tau_f \tag{1}$$

where $\tau_{\mathrm{D}}$ is the timescale for the two monomers diffusing into proper orientation and $\tau_{\mathrm{f}}$ is the timescale for the formation of the first base pair. Assuming that the two monomers are aligned after the first step, the rate of base pair formation can be presumed to be $k_{\mathrm{f}}$, thus $\tau_{\mathrm{f}} = \frac{1}{k_{\mathrm{f}}}$. Utilizing this and Equation 1, $\beta$ can be expressed as

$$\beta = \frac{\tau_{\mathrm{f}}}{\tau_{\mathrm{D}} + \tau_{\mathrm{f}}} \tag{2}$$

To determine $\tau_{\mathrm{D}}$, we make use of the Stokes-Einstein expression for the diffusion-limited association rate for two identical spheres for a concentration of monomers [M][46]

$$\tau_{\mathrm{D}} = \frac{3\eta}{8RT[\mathrm{M}]} \tag{3}$$

giving

$$\beta = \frac{8RT[\mathrm{M}]}{3\eta k_{\mathrm{f}} + 8RT[\mathrm{M}]} \tag{4}$$

Here $R$ is the ideal gas constant, $\eta$ is the viscosity, and [M] is the initial monomer concentration of the system, which in the context of this work is the equilibrium monomer concentration at the initial temperature prior to the arrival of the T-jump pulse. Determining $\beta$ in this way also incorporates the expected concentration dependence for the association of self-complementary DNA single strands.

The attenuation parameter for propagation steps is denoted $\sigma_i$, the values of which are contained in the interval $(0, 1]$.[11] The subscript $i$ denotes $N_{\mathrm{BP}}$ for the initial state, with respect to the forward direction, for the formation of a base pair along an already formed stretch. The value of $\sigma_i$ increases with $N_{\mathrm{BP}}$ in agreement with the conceptual understanding of $k_{\mathrm{f}}$ being the rate of formation for a base pair at the long end of a series of intact base pairs. While there are a number of different factors that contribute to the increasing rate of formation for each sequential base pair, it has been considered to be primarily due to the additional stability that is associated with the formation of the helical structure that occurs when multiple consecutive intact base pairs exist.[11]

The definition of $\sigma_i$ requires that the values fall between zero and one and that it starts small, monotonically increases, and asymptotically approaches a value of one[11]

$$\lim_{i \to \infty} \sigma_i = 1 \tag{5}$$

The hyperbolic tangent function fits these requirements and the intuitive understanding of $\sigma_i$ resulting in the definition

$$\sigma_i = \tanh \frac{\alpha x_i}{1 - x_i} \tag{6}$$

where $\alpha$ is a fit parameter in the model that dictates how quickly $\sigma_i$ approaches a value of one and $x_i$ is the fraction of intact base pairs relative to the total number of base pairs $\left(\frac{N_{\mathrm{BP}}}{N}\right)$.

The only remaining term in Figure 1 is the free energy difference between the initial state ($n$) and the final state ($m$), $\Delta G_{nm}$. The value of $\Delta G_{nm}$ is the difference between the free energy of the configurations in the final and initial states for a particular move, with the free energy of both configurations calculated using the thermodynamic lattice model previously published by our group,[32] and is used in the calculation of the reverse rate starting with the detailed-balance relation

$$s = \frac{k_{n\to m}}{k_{m\to n}} = e^{\frac{-\Delta G_{nm}}{RT}} \tag{7}$$

where $s$ denotes an equilibrium constant for a single base pairing step following the notation used by many in the literature,[11,17,43] $k_{n\to m}$ is the forward rate, and $k_{m\to n}$ is the reverse rate. Since the free energy of each individual configuration is used to calculate $\Delta G_{nm}$ the reverse steps in the model carry all of the sequence specificity, an assumption widely used in kinetic Monte Carlo models utilized to study DNA.[7,31,35–37,47]

Two more assumptions are made when calculating the rates, both involving $k_{\mathrm{f}}$. The first assumes $k_{\mathrm{f}}$ is independent of base pair composition, an assumption commonly made for similar models in the literature[7,11,12,17,35,45,48] since A:T and G:C base pairs are sterically similar and $k_{\mathrm{f}}$ should not significantly depend on stacking interactions.[7]

It is also assumed that $k_{\mathrm{f}}$ is independent of temperature, which is more contentious in the literature. Models exist that do not include a temperature dependence,[45] while others do by incorporating an activation energy or directly fitting each individual temperature; however, among these models the results are inconclusive. It has been proposed that the activation barrier is small and positive, generally in the range of 1-5 kcal mol$^{-1}$.[12,17,47,48] This leads to the proposal that the elementary formation of a single base pair adjacent to an intact base pair is diffusion-controlled.[12,17,48] However, caution should be exercised due to studies in the literature examining significantly longer sequences,[17] or fitting as few as two temperatures and acknowledging that under certain experimental conditions the correct rate as a function of temperature was obtained using an activation energy of zero.[12] Other experimental results, examining sequences with lengths of 8-14 base pairs, demonstrate that $k_{\mathrm{f}}$ varies insignificantly and inconsistently with temperature for a given chain length.[11] It is also worth noting that the formation of a base pair being either diffusion-controlled or

barrierless provides further support for both $k_{\mathrm{f}}$ and the overall forward rate not carrying a sequence dependence, as one would not be expected to exist in either case.

### C. Optimizing Fit Parameters and Obtaining Trajectories

Association and dissociation trajectories were generated with the Gillespie algorithm[33,34] using the "direct method"[33] to determine both the time interval and the final state for each step. For association (dissociation) trajectories, the system starts in the monomer (fully formed dimer) state and terminates upon reaching the fully formed dimer (monomer) state. To determine the fit parameters $\alpha$ and $k_{\mathrm{f}}$, the model was parameterized against the experimental T-jump results for sequences of varying base pair composition and length that have been published previously.[39,40] The parameters for each sequence were fit independently to the observed rate constants, $k_{\mathrm{obs}}$, with five or six temperatures included for each sequence. To compare the simulations to the experimental results, a set of association and dissociation trajectories were run for a given set of parameters to determine the mean first passage time for both directions. The first passage time is the time it takes for a trajectory to proceed from the initial state to the final state which in the case of association (dissociation) are the monomer (fully formed dimer) and the fully formed dimer (monomer) respectively. The mean first passage time is then determined by averaging over all trajectories. To model the conditions of the T-jump experiments the model is run at the final temperature of the sample after the arrival of the T-jump pulse.

The observed rate constant, which is the standard relaxation rate from perturbative chemical kinetics for the two state D $\rightleftharpoons$ 2M process, is calculated according to a standard two-state kinetic analysis making the assumption that these rates are in response to a weak perturbation, which the temperature jump is assumed to be for the experiments the model is fit to in this work.[39,40,49] Under this assumption, the observed rate constant is given by[50]

$$k_{\mathrm{obs}} = k_{\mathrm{d}} + 4[\mathrm{M}]k_{\mathrm{a}} \tag{8}$$

where [M] is the monomer concentration at the initial temperature prior to the T-jump pulse, $k_{\mathrm{a}}$ is the association rate constant, and $k_{\mathrm{d}}$ is the dissociation rate constant. The association rate $k_{\mathrm{a}}$ is calculated from[38]

$$k_{\mathrm{a}} = \frac{1}{[\mathrm{M}]\langle\tau_{\mathrm{a}}\rangle} \tag{9}$$

where $\langle \tau_a \rangle$ is the association mean first passage time from the model. The dissociation rate, $k_d$, is calculated from[38]

$$k_d = \frac{1}{\langle \tau_d \rangle} \tag{10}$$

where $\langle \tau_d \rangle$ is the dissociation mean first passage time from the model. The parameters were optimized utilizing a pattern search algorithm that minimized the sum of the squared residuals at each temperature. It is worth noting that these equations are correct for the self-complimentary sequences analyzed here and would need to be altered for the case of non-self-complimentary sequences.

The number of trajectory sets run during each iteration of the fitting algorithm is twice the number of fit parameters, so in the case of fitting $k_f$ and $\alpha$, four trajectory sets must be run each iteration and thousands of iterations are required. To reduce computational expense, the trajectory sets run during the course of the fitting are relatively small, on the order of hundreds of trajectories. A number of optimization routines were run for each sequence with randomized initial parameters until a more concise range in which the parameters were converging was determined. The best parameters from these initial fits were selected and used as the initial parameters for additional optimization routines, using the same method, to determine the final parameters. Once these parameters were determined, a large trajectory set was run with these parameters to ensure that the values compared to experiment during the fitting routine were representative of the results of the large trajectory set. For the 5'-ATATGCATAT-3' (GC-core) sequence, 5,000 association and dissociation trajectories were run while for all other sequences, 100,000 trajectories were run. This ensured there was no error due to the small trajectory sets used during the optimization routines. These large trajectory sets, and the transition rate matrices used to generate them, are the results analyzed in this paper.

The parameters returned by the fitting algorithm are given in Table I. To get a better sense for the values of $\sigma_i$, Figure 2 shows the $\sigma_i$ values at each $N_{BP}$ value for each sequence in the length series. This helps to both demonstrate the functional form of $\sigma_i$ given in Equation 6 and provide a clearer conceptual understanding of how the rate of formation for a single base pair increases with an increasing number of previously intact base pairs. One note on the inclusion of $\sigma_i$ in this form is that preliminary results were obtained from an earlier version of the model that did not incorporate $\sigma_i$ values of any kind, making $\beta$ the only attenuation parameter. This early version generated fits to the experimental data

TABLE I. Fit parameters returned by the kinetic model for each sequence.

| sequence | length | $k_f$ ($s^{-1}$) | $\alpha$ |
|---|---|---|---|
| CG-ends | 6 | 5.4344 x $10^{11}$ | 0.6101 |
| | 8 | 2.0969 x $10^{11}$ | 1.0790 |
| | 10 | 5.9513 x $10^{10}$ | 1.5424 |
| | 12 | 4.0526 x $10^{10}$ | 1.5665 |
| | 14 | 7.4036 x $10^{9}$ | 2.2705 |
| GC-core | 10 | 3.5993 x $10^{10}$ | 3.7285 |
| AT-all | 10 | 3.2769 x $10^{10}$ | 2.8186 |

that were relatively comparable to those from the model presented here. The use of a single attenuation parameter would be comparable to the use of an apparent $\beta$, or $\beta_{app}$, which has been done by others.[11,16] However, even though the fit quality was not substantially improved the $\sigma_i$ parameter was incorporated in an effort to provide a clearer physical interpretation of the model's parameters since it is expected that the rate of formation for multiple base pairs should be attenuated.

It is interesting to note that for all sequences in the length series, due to the value of $\alpha$ increasing with increasing length, the values of $\sigma_i$ consistently approach a value of one within the formation of 4-5 base pairs, in good agreement with predictions from the literature.[11] This suggests that the values of $\sigma_i$ approach a value of one within approximately a half turn of the DNA double helix, supporting the idea that beginning to obtain the helical structure, and the structural stability associated with it, likely plays a significant role in the rate at which a single base pair forms. The consistency across the different sequences also suggests that defining $\sigma_i$ based on the number of previously intact base pairs, rather than the number of unpaired base pairs in the frayed end, results in a more accurate representation of the degree to which the rate of forming an individual base pair is attenuated. This supports the definition of $\sigma_i$ utilized in this model while also providing further evidence for the physical

FIG. 2. The $\sigma$ values as a function of normalized $N_{\mathrm{BP}}$ with (•) marking the position of each base pair for a given sequence with an associated $\sigma_i$ value less than 0.9 for the 5'-C$(\mathrm{AT})_n$G-3' sequences with $n$ = 2-6.

interpretation of this parameter.

The observed rate constants, calculated from the mean first passage time for association and dissociation using the two-state analysis, are compared to those determined by experiment in Figure 3 for all sequences except AT-all. AT-all was excluded since only two temperatures are available making it a poor metric of fit quality relative to the other sequences. The model is generally in good agreement with the experimental data, particularly at higher temperatures.

The main discrepancy between the model and experimental results is the ability of the model to replicate the degree to which the different sequences and lengths demonstrate nonlinear trends in the Arrhenius plots. The observed rate constant is closely related to the dissociation rate constant, and a linear Arrhenius plot is indicative of two-state kinetics dictated by a single temperature independent activation barrier. The Arrhenius plots for the dissociation rates are also shown in Figure 4. The values of $k_{\mathrm{obs}}$ from the model consistently demonstrate a small degree of nonlinearity such that for shorter CG-ends sequences, where the experimental trends are linear, the model does not fully replicate the linear trend resulting in some deviation at low temperature. The degree of nonlinearity demonstrated

FIG. 3. Arrhenius plots of the observed rate constants from the kinetic model (red) and experiment[39,40] (black) for (a) 5'-CATATG-3', (b) 5'-CATATATG-3', (c) 5'-CATATATATG-3', (d) 5'-CATATATATATG-3', (e) 5'-CATATATATATATG-3', and (f) 5'-ATATGCATAT-3'.

by the model appears to be relatively unaffected by sequence length and composition which can be seen by the fact that the model is unable to fully replicate the degree of nonlinearity observed in the GC-core sequence, again deviating at low temperature. Outside of capturing the change in the degree of linearity as a function of length and sequence, the results from

FIG. 4. Arrhenius plots of the dissociation rate constants from the kinetic model (red) and experiment[39,40] (black) for (a) 5'-CATATG-3', (b) 5'-CATATATG-3', (c) 5'-CATATATATG-3', (d) 5'-CATATATATATG-3', (e) 5'-CATATATATATATG-3', and (f) 5'-ATATGCATAT-3'.

the model are shown to be in excellent agreement with the experimental results.

One final note on the agreement between the model and the experimental data concerns the primary insights gained from the model. The primary analysis presented here focuses on the mechanistic insights gained from the association and dissociation trajectories and how

they compare to the experimental results. While trends in the model's parameters exist that are worth discussing, caution should be taken before drawing strong conclusions from them since, while rooted in physical processes, they are a combination of a number of factors that are difficult to disentangle. Further studies that expand the number of different sequences examined by the model should help to alleviate this by providing a larger data set of model parameters and a more robust analysis, particularly with regards to trends with sequence, length, and other variables, that will allow conclusions to be drawn with more confidence. With respect to future studies, it is worth noting that in practice we have found that the lattice model is currently useful for sequences of up to 20 base pairs in length. The Markov state Monte Carlo model has not been used for sequences longer than 14 base pairs due to a lack of experimental data to fit to. However, its current form is capable of simulating sequences of up to 20 base pairs, the limit of the lattice model, given sufficient computing power.

## III. RESULTS

The analysis here will have two different foci. The first aspect analyzed in detail is the barrier crossing event in the trajectories. The barrier crossing event is defined here for association (dissociation) as the portion of the trajectory between the last time the trajectory is in the monomer (fully formed dimer) state until it reaches the fully formed dimer (monomer) state. The barrier crossing events generally occur on a timescale of hundreds of picoseconds for the dissociation and nanoseconds for the association. These barrier crossing events can involve both forward and backward moves meaning that there is no set number of steps that make up the barrier crossing. However, the minimum number of steps is equal to the number of base pairs in the sequence. For certain sequences it is not uncommon for the barrier crossing event to include two to three times more steps than the minimum.

It is also informative to analyze the entire trajectory as a whole. The main focus here is examining the early time dynamics, that take place on an approximately nanosecond timescale prior to the dissociation of the full duplex, experimentally observed in some sequences. These early time dynamics will in some cases be referred to as the "fast response" as they experimentally appear as an increase in signal that occurs on a faster time scale than the full dissociation. This analysis examines how the trajectory moves through the state

space prior to the barrier crossing event and how long the trajectory spends in states with each $N_{\mathrm{BP}}$. Doing so allows for quantitative comparison between the dissociation trajectories for each sequence, and the resulting trends with length and sequence composition, to changes observed in the experimental results to further clarify the dynamics that our occuring.

### A. Barrier Crossing Pathways

Utilizing transition pathway theory (TPT)[51–54] individual pathways for barrier crossing events can be isolated and ranked according to the frequency at which they occur. It is important to distinguish individual pathways from the overall mechanism. Individual pathways are one possible way the system can move through different configurations between the initial and final states in the association or dissociation barrier crossing events. In this context the overall mechanism incorporates the entire distribution of individual pathways and is a more general view of how the system progresses through a barrier crossing event.

Figure 5 shows the top six association pathways isolated by TPT for three different sequences and the probability of a successful association event occurring along each of the pathways, shown in the bar graphs on the left. The probability is calculated from the percentage of overall flux between the monomer state and fully formed dimer state for each individual pathway.[27] Black boxes represent intact base pairs and time proceeds up the chart with the first base pair formed in the second row from the bottom and the reaction proceeding upwards to reach the fully formed dimer state in the top row. Since these sequences are all self-complimentary each pathway has a symmetric partner that is identical and carries the same probability. For example, the first 5'-CATATATATG-3' pathway in Figure 5 initiates at one end and zips up sequentially across the sequence. The symmetric partner of this pathway is identical except that it starts at the other terminus. For all self-complimentary sequences each pair will be referred to as a unit, for example referring to the two most dominant pathways refers to the two most dominant sets of pathways which are the top four individual pathways. The probabilities shown in the bar graphs in Figure 5 represent the probability of an association event following each individual pathway. For example, the probability of an association event following the top 5'-CATATATATG-3' pathway shown or its symmetric pair is 16%, since the probability for each pathway is 8%.

After examining the different pathways shown in Figure 5 two prominent motifs emerge.

FIG. 5. Top six association pathways for 5'-ATATATATAT-3' at 308 K, 5'-ATATGCATAT-3' at 333 K, and 5'-CATATATATG-3' at 334 K. For each sequence these pathways are ordered from most to least probable from left to right with their ranking denoted by the number above each column and the bar graph showing the probability for each pathway. For each sequence these six pathways, and their symmetric partner, make up 64.7%, 74.2%, and 69.0% respectively of the total flux between the monomer state and the fully formed dimer state across all pathways isolated by TPT at the temperatures shown.

The first pathway motif initiates in the center of the sequence and the two sides symmetrically form base pairs, keeping the two frayed ends of similar or identical length, until the duplex is formed. This pathway motif can be observed in Figure 5 as the top pathways for both 5'-ATATATATAT-3' and 5'-ATATGCATAT-3' in addition to the third pathway for 5'-CATATATATG-3'. The second motif initiates at or near a G:C base pair and in the case where the pathway does not initiate at the G:C base pair it forms as early as possible. This is best seen in the top two pathways for 5'-CATATATATG-3', where this pathway is distinct from the center-initiated pathway, and the top pathways for 5'-ATATGCATAT-3', where the two motifs overlap due to the location of the G:C base pairs. The probability of each pathway occurring is correlated to how closely the pathway follows each of these motifs. Initiating closer to the center increases the probability of the pathway as does initiating closer to a G:C base pair. Looking at 5'-ATATGCATAT-3', where the two pathway motifs overlap, initiating further from the center and increasing the difference in length between the two frayed ends results in the probability of these pathways rapidly decreasing. While deviating from either motif results in decreasing probability, the G:C base pair driven motif demonstrates a more rapid drop off. This can be seen by comparing the probability of the first two pathways for 5'-CATATATATG-3' and 5'-ATATATATAT-3'. Comparing these two sequences also shows that initiating at or near a G:C base pair carries a more dominant effect, as seen by examining the ranking of the top six pathways for 5'-CATATATATG-3'.

### B. Overall Mechanistic Insights

While the individual pathways are informative on a microscopic scale, it is important to more generally consider the mechanism for monomer-dimer transitions in terms of the overall two-state reaction. However, our focus at this point remains on the barrier crossing event itself. One interesting aspect of the overall mechanism is the probability of initiating a barrier crossing at different positions. This can be determined by summing over all TPT pathways, the result of which is shown in Figure 6.

One of the more interesting observations of Figures 5 and 6 is that while the dominant individual pathway for 5-CATATATATG-3' initiates at a termini those positions are the least likely to initiate a successful association barrier crossing event. The most probable position is actually either next to the G:C termini, the position at which the second most probable

individual pathway initiates, or in the center of the sequence depending on temperature. It is also interesting to note that the third position has a lower probability than either of its neighbors, which is related to the previously discussed pathway motifs. The center-initiated preference results in the dome shape and explains why the fourth position from the end is more probable than the third position. The second position is more probable than the third position because it receives a significant benefit from forming next to a terminal G:C base pair. This does not however explain why the most dominant pathway for 5'-CATATATATG-3' initiates at the least likely position to initiate a successful association barrier crossing. The reason for this has to do with Figure 6 incorporating all possible pathways. Combinatorics dictates that more pathways initiate at positions near the center relative to positions near the end. The number of potential pathways at each position can be found by looking along the row of Pascal's triangle whose length is equal to the number of base pairs in the sequence. Even though these pathways become less probable, the cumulative effect makes a significant difference. This explains why the termini are the least likely position for initiating a successful association event, while the most probable pathway initiates at that position, it is the only pathway. The results in Figure 6 for 5'-CATATATATG-3' are in stark contrast to those for the GC-core sequence that demonstrate a very strong preference for initiating in the middle of the sequence, which makes sense when considering the pathways in Figure 5 and the overlap between the two main pathway motifs.

Figure 6 can also be analyzed to gain insights into the likely locations of the critical nucleus for a given sequence. The critical nucleus is an important structure in the mechanism of DNA hybridization and dehybridization since its identity impacts the energetics and resulting reaction rates for the processes.[11,13,21,55] While the size of the critical nucleus is debated in the literature, and likely depends on factors including sequence composition[14] and temperature,[21] it is generally considered to be small and on the order of two to four base pairs.[11,13,21,55] Considering the relatively small size of the critical nucleus, the probability of initiating at different positions within the sequence can be used as a proxy for the location of the critical nucleus. As such we can consider what the preference for initiating centrally or at or near a G:C base pair means with respect to the critical nucleus. The fact that association events that initiate near a G:C base pair will form the G:C base pair as early as possible, as shown in Figure 5, shows that the preference for initiating association at or near a G:C base pair implies that there is a preference for including the G:C base pair in the critical

FIG. 6. Percentage of all association barrier crossing events that initiate at each position for 5'-ATATGCATAT-3', and 5'-CATATATATG-3' at the lowest and highest temperatures studied.

nucleus, which makes sense considering the added stability provided by G:C base pairs. The preference for initiating near the center of the sequence implies the critical nucleus also has a preference for forming centrally. This means that the critical nucleus preferentially forms either in the center of a sequence or, in the case where there are no G:C base pairs in the center of the sequence, incorporates one or more G:C base pairs. In instances where there are one or more G:C base pairs near the center of a sequence, such as the GC-core sequence analyzed in this work, the critical nucleus will very likely incorporate them.

It is worth briefly comparing these results to the literature, particularly coarse-grained MD simulations. Coarse-grained MD simulations have found that contacts in the center of the sequence are critical for hybridization, particularly in the case of more randomized sequences.[23,24] For both randomized and repetitive sequences nucleation is biased towards the center[23] and one study found that middle to middle nucleation events represent more than 80% of all those possible for all oligos examined.[22] All of which is consistent with the findings from the model presented here which also demonstrates a strong preference for initiating association events centrally.

It is also interesting to examine the influence of G:C base pairs on the association initiation

position. It has been proposed that sequences that contain them are expected to initiate at their position.[14,56] While a preference for forming at or near G:C base pairs exists, it is still location dependent and not overwhelming. While the findings for GC-core do show that a large number of initiations will occur at the G:C base pairs it is still less than 50% of all initiations for all temperatures studied here. For CG-ends this number is even lower with initiation at the terminal G:C base pairs making up less than 28% of all initiations for the shortest sequence and less than 12% for the longest sequence. This further demonstrates that, for CG-ends, while initiating at a G:C base pair does appear to result in a dominant individual pathway, when considering the mechanism as a whole the relative significance of that pathway diminishes, particularly for longer lengths.

### C. Energetic Driving Forces

Our attention now turns to the driving forces behind the trends observed in the individual pathways and overall barrier crossing mechanisms. Two general motifs were observed, initiating association in the center and initiating at or near a G:C base pair that forms early on. We will now discuss why the center-initiated motif is entropically driven while the G:C base pair initiated motif is enthalpically driven. These motifs may overlap, resulting in the enthalpic and entropic components driving the same, or competing, pathways.

To demonstrate how entropy favors the center-initiated motif, consider the increased preference for the pathways that follow it, best observed in the top pathway for AT-all in Figure 5. There are two contributions to this entropically driven preference for initiating centrally. When treating the configurational entropy of frayed ends as a self-avoiding random walk, as our thermodynamic lattice model does[32], the highest entropy state for configurations with a given $N_{\mathrm{BP}}$ has two frayed ends of equal length. The next highest entropy states are those that have two frayed ends of nearly equal lengths and the entropy decreases as the difference between the length of the two ends grows. Finally, the lowest entropy configuration has a single frayed end.[32] As a result, the preferential configurational entropy drives the preference for initiating association in the center and forming the remaining base pairs by building out symmetrically.

The second factor that drives preference for initiating association events in the center of the sequence, as seen in Figure 6 is also entropic in nature. Initiating in the center of a

sequence results in an increased multiplicity of pathways to the final fully hybridized state. While this additional driving force contributes to the preference for initiating in the center, unlike the contribution from the configurational entropy it does not result in any preference for building symmetrically out from the center.

Since the pathways that initiate at or next to the terminal G:C base pairs in the CG-ends sequences are expected to have a higher entropic cost, the relative preference for these pathways must be enthalpically driven. This is not particularly surprising as G:C base pairs are known to be more stable than A:T base pairs and it has been previously proposed that they play a role in the early stages of the association process for this reason.[14]

### D. Full Trajectory Analysis

Now we will step back from looking at the barrier crossing event and examine the entire trajectory, with a particular eye towards the experimentally observed fast response. The primary purpose of this is twofold. To understand how fraying appears in the model by looking at GC-core, where fraying has been experimentally observed, and using this knowledge to gain insight into the fast response that grows in with length in the CG-ends sequences. Analyzing the entire trajectory is difficult since there are thousands of steps in each trajectory. However, due to the construction of the model requiring that all dissociation initiate at the ends any early time response must be fraying. As a result the individual configurations are insignificant and the $N_{\mathrm{BP}}$ at each step can simply be tracked.

It is worth briefly discussing the previously published experimental results that the model is compared to. The GC-core sequence demonstrated a significant early time rise in signal that was not observed for the sequence 5'-GATATATATC-3'. This early time response was attributed to the fast fraying dynamics of the terminal A:T base pairs.[40] Since the fraying of the terminal A:T base pairs, with no loss of G:C base pairs, was so clearly observed in the experiment this will be used as a point of comparison for the model to investigate if the model correctly captures this behavior. The study examining the CG-ends sequences found that the sequences showed a small increase in early time signal, significantly smaller than GC-core, that grew with both increasing length and temperature for a given length. Interestingly, rather than adopting the biexponential signal trace observed for the GC-core sequence[40] the fast response for the CG-ends series adopted a stretched exponential form

and was nearly indistinguishable between A:T and G:C base pairs.[39] Because the CG-ends results did not demonstrate a clear qualitative mechanistic interpretation like the GC-core sequence the origin of the fast response in the case of the CG-ends series could not be determined through experiment alone.

Figure 7 contains plots showing the percentage of time spent in states as a function of $N_{\mathrm{BP}}$ averaged over all trajectories for each length, except for the shortest sequence, of the CG-ends series and the GC-core sequence at either 333 or 334 K. Looking at the partially intact states there is a clear distinction between GC-core and the CG-ends sequences. The increased amount of time GC-core spends in partially intact states is consistent with the experimentally observed fraying.[40] This demonstrates that the kinetic model replicates the fraying behavior experimentally observed in GC-core while indicating less early time dissociation in the ten base pair CG-ends sequence, also in agreement with experiment.[39,40]

Not only does the kinetic model agree with the experimental results but also with the thermodynamic lattice model.[32] Figure 7 compares the percentage of time spent in states during a trajectory with the probability, presented as a percentage, of occupying the same states according to the thermodynamic lattice model, both as a function of $N_{\mathrm{BP}}$. Note that the probability of occupying a state is proportional to the free energy of the system and the free energy surface along $N_{\mathrm{BP}}$, which is given by $-k_{\mathrm{B}}T\ln(P)$. This demonstrates that over a sufficient number of trajectories the amount of time spent in different states is primarily dictated by the thermodynamic free energy of the system rather than any significant kinetic factors.

The agreement with experiment is particularly good since the model also demonstrates that for GC-core primarily A:T base pairs are dissociating at early time. Any configuration with six or more intact base pairs must have both G:C base pairs intact. While configurations with four or five base pairs are not accessed to any significant degree. The lattice model shows that even when states with four or five intact base pairs are accessed the probability of still having both G:C base pairs intact is over 99%.[32] This demonstrates that the kinetic model is accurately depicting the fast fraying dynamics of the GC-core sequence.

Now that it has been established that the kinetic model demonstrates the fraying behavior expected for GC-core; the same analysis can be applied to the CG-ends sequences to understand what is behind the experimentally observed fast response that grows in with increasing length. Figure 7 shows that with increasing length, and temperature held constant,

FIG. 7. Average percentage of time during spent in states with each $N_{\mathrm{BP}}$ for a trajectory starting in the fully formed dimer state (red) and the probability of occupying a non-monomer state with a given $N_{\mathrm{BP}}$ determined by the thermodynamic lattice model[32] (black) for (a) 5'-CATATATG-3' at 333 K, (b) 5'-CATATATATG-3', (c) 5'-CATATATATATG-3', and (d) 5'-CATATATATATATG-3' at 334 K and (e) 5'-ATATGCATAT-3' at 333 K.

the amount of time spent in partially intact states increases. This provides clear evidence that even with the stabilizing G:C base pairs on the termini these sequences become more susceptible to fraying with increasing length. This is in agreement with studies from the literature that have also observed fraying in sequences with G:C end caps.[57,58] While the trends with length are not nearly of the magnitude observed for GC-core, this is reasonable since the trends observed in the experimental results for the CG-ends sequences are also of smaller magnitude. This suggests that fraying is the likely source of the increasing stretching factor observed in the CG-ends sequences with increasing length. It is worth noting briefly that in conjunction with CG-ends the sequence 5'-GATATATATC-3' was briefly examined to identify if switching the bases at the 5' and 3' ends affected the behavior observed by the model. Preliminary work examined the probability of occupying non-monomer states with a given $N_{\mathrm{BP}}$, similar to Figure 7. This showed that the 5'-GATATATATC-3' sequence would undergo similar fraying behavior to CG-ends with the only difference, resolved by the model, being a slight increase in terms of the average size of the frayed ends.

## IV. CONCLUSION

The two parameter Markov state Monte Carlo model presented here is able to reasonably reproduce the experimental results and is in agreement with existing coarse-grained MD simulations of more significant complexity. With regards to fast dynamics prior to the full dissociation of the duplex, the model recreates the fast fraying dynamics experimentally observed for GC-core and provides further evidence that these dynamics are primarily driven by thermodynamic factors and the reshaping of the free energy surface rather than kinetic factors. Additionally, the model suggests that the origins of the increasing fast response observed with increasing length in the CG-ends series is similar to those observed in the GC-core sequence suggesting that fraying plays a significant role in these dynamics.

The model also shows that the initiation position for a successful association barrier crossing, which corresponds to the location of the critical nucleus and transition state, is driven by two factors. Entropic favorability, due to both a contribution from minimizing the penalty due to configurational entropy and from an increased number of available association pathways, preferentially drives initiating at the center of the sequence. The second factor is an enthalpic favorability that preferentially drives initiating near G:C base pairs,

if present in the sequence. The effects of these energetic forces, particularly the enthalpic benefit, become far less significant after the formation of the first few base pairs which is in agreement with the canonical nucleation-zipper mechanism and the corresponding critical nucleus. This demonstrates that the simple, accessible, two parameter Markov state kinetic Monte Carlo model presented here provides additional insights that compliment existing experimental methods to aid in understanding the mechanisms and energetics of DNA hybridization and dehybridization. Additionally, the insights gained from this model build the foundation for continuing research on this topic moving forward. Utilizing it hand-in-hand with future sequence specific kinetic results from different sequence and length motifs will further validate the model and drive a more complete understanding of DNA association and dissociation.

## ACKNOWLEDGMENTS

The authors would like to thank Brennan Ashwood for his thoughtful comments and suggestions and Paul Sanstead for sharing his sequence-dependent dehybridization data. The authors would also like to thank the University of Chicago Research Computing Center for their support of this work. This work was supported by a grant from the National Science Foundation (CHE-1561888).

## DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## REFERENCES

[1]R. Lavery, K. Zakrzewska, D. Beveridge, T. C. Bishop, D. A. Case, I. Cheatham, Thomas, S. Dixit, B. Jayaram, F. Lankas, C. Laughton, J. H. Maddocks, A. Michon, R. Osman, M. Orozco, A. Perez, T. Singh, N. Spackova, and J. Sponer, "A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA," Nucleic Acids Res. **38**, 299–313 (2010).

[2]W. K. Olson, M. Bansal, S. K. Burley, R. E. Dickerson, M. Gerstein, S. C. Harvey, U. Heinemann, X.-J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C.-S. Tung, E. Westhof, C. Wolberger, and H. M. Berman, "A standard reference frame for the description of nucleic acid base-pair geometry," J. Mol. Biol. **313**, 229–237 (2001).

[3]R. J. Roberts and X. Cheng, "Base flipping," Annu. Rev. Biochem. **67**, 181–198 (1998).

[4]G. Altan-Bonnet, A. Libchaber, and O. Krichevsky, "Bubble dynamics in double-stranded DNA," Phys. Rev. Lett. **90**, 138101 (2003).

[5]Q. Dai, P. J. Sanstead, C. S. Peng, D. Han, C. He, and A. Tokmakoff, "Weakened N3 hydrogen bonding by 5-formylcytosine and 5-carboxylcytosine reduces their base-pairing stability," ACS Chem. Biol. **11**, 470–477 (2016).

[6]P. H. von Hippel, N. P. Johnson, and A. H. Marcus, "Fifty years of DNA "breathing": Reflections on old and new approaches," Biopolymers **99**, 923–954 (2013).

[7]T. Ambjörnsson, S. K. Banik, O. Krichevsky, and R. Metzler, "Breathing dynamics in heteropolymer DNA," Biophys. J. **92**, 2674–2684 (2007).

[8]K. Blagoev, B. Alexandrov, E. Goodwin, and A. Bishop, "Ultra-violet light induced changes in DNA dynamics may enhance tt-dimer recognition," DNA Repair **5**, 863–867 (2006).

[9]G. Kalosakas, K. Ø. Rasmussen, A. R. Bishop, C. H. Choi, and A. Usheva, "Sequence-specific thermal fluctuations identify start sites for DNA transcription," Europhys. Lett. **68**, 127–133 (2004).

[10]B. S. Alexandrov, V. Gelev, S. W. Yoo, L. B. Alexandrov, Y. Fukuyo, A. R. Bishop, K. Ø. Rasmussen, and A. Usheva, "DNA dynamics play a role as a basal transcription factor in the positioning and regulation of gene transcription initiation," Nucleic Acids Res. **38**, 1790–1795 (2010).

[11]M. E. Craig, D. M. Crothers, and P. Doty, "Relaxation kinetics of dimer formation by self complementary oligonucleotides," J. Mol. Biol. **62**, 383–401 (1971).

[12]D. Pörschke, "A direct measurement of the unzippering rate of a nucleic acid double helix," Biophys. Chem. **2**, 97–101 (1974).

[13]D. Pörschke and M. Eigen, "Co-operative non-enzymatic base recognition III. kinetics of the helix-coil transition of the oligoribouridylic • oligoriboadenylic acid system and of oligoriboadenylic acid alone at acidic pH," J. Mol. Biol. **62**, 361–381 (1971).

[14]D. Pörschke, O. C. Uhlenbeck, and F. H. Martin, "Thermodynamics and kinetics of the

helix-coil transition of oligomers containing GC base pairs," Biopolymers **12**, 1313–1335 (1973).

[15]J. Applequist and V. Damle, "Theory of the effects of concentration and chain length on helix—coil equilibria in two-stranded nucleic acids," J. Chem. Phys. **39**, 2719–2721 (1963).

[16]J. Applequist and V. Damle, "Thermodynamics of the helix-coil equilibrium in oligoadenylic acid from hypochromicity studies," J. Am. Chem. Soc. **87**, 1450–1458 (1965).

[17]J. G. Wetmur and N. Davidson, "Kinetics of renaturation of DNA," J. Mol. Biol. **31**, 349–370 (1968).

[18]A. Pérez, F. J. Luque, and M. Orozco, "Dynamics of B-DNA on the microsecond time scale," J. Am. Chem. Soc. **129**, 14739–14745 (2007).

[19]A. Pérez and M. Orozco, "Real-time atomistic description of DNA unfolding," Angew. Chem. Int. Ed. **49**, 4805–4808 (2010).

[20]A. Pérez, F. J. Luque, and M. Orozco, "Frontiers in molecular dynamics simulations of DNA," Acc. Chem. Res. **45**, 196–205 (2012).

[21]T. E. Ouldridge, P. Šulc, F. Romano, J. P. K. Doye, and A. A. Louis, "DNA hybridization kinetics: Zippering, internal displacement and sequence dependence," Nucleic Acids Res. **41**, 8886–8895 (2013).

[22]D. M. Hinckley, J. P. Lequieu, and J. J. de Pablo, "Coarse-grained modeling of DNA oligomer hybridization: Length, sequence, and salt effects," J. Chem. Phys. **141**, 035102 (2014).

[23]E. J. Sambriski, D. C. Schwartz, and J. J. de Pablo, "Uncovering pathways in DNA oligonucleotide hybridization via transition state analysis," Proc. Natl. Acad. Sci. U. S. A. **106**, 18125–18130 (2009).

[24]M. J. Hoefert, E. J. Sambriski, and J. J. de Pablo, "Molecular pathways in DNA-DNA hybridization of surface-bound oligonucleotides," Soft Matter **7**, 560–566 (2011).

[25]S. Xiao, D. J. Sharpe, D. Chakraborty, and D. J. Wales, "Energy landscapes and hybridization pathways for DNA hexamer duplexes," J. Phys. Chem. Lett. **10**, 6771–6779 (2019).

[26]M. Maciejczyk, A. Spasic, A. Liwo, and H. A. Scheraga, "DNA duplex formation with a coarse-grained model," J. Chem. Theory Comput. **10**, 5020–5035 (2014).

[27]F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations," Proc.

Natl. Acad. Sci. U. S. A. **106**, 19011–19016 (2009).

[28]V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, "Molecular simulation of *ab Initio* protein folding for a millisecond folder ntl9(1-39)," J. Am. Chem. Soc. **132**, 1526–1528 (2010).

[29]F. Noé and S. Fischer, "Transition networks for modeling the kinetics of conformational change in macromolecules," Curr. Opin. Struc. Biol. **18**, 154–162 (2008).

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

[31]F. Dannenberg, K. E. Dunn, J. Bath, M. Kwiatkowska, A. J. Turberfield, and T. E. Ouldridge, "Modelling DNA origami self-assembly at the domain level," J. Chem. Phys. **143**, 165102 (2015).

[32]P. J. Sanstead and A. Tokmakoff, "A lattice model for the interpretation of oligonucleotide hybridization experiments," J. Chem. Phys. **150**, 185104 (2019).

[33]D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," J. Comput. Phys. **22**, 403–434 (1976).

[34]D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," J. Phys. Chem. **81**, 2340–2361 (1977).

[35]S. K. Banik, T. Ambjörnsson, and R. Metzler, "Stochastic approach to DNA breathing dynamics," Europhys. Lett. **71**, 852–858 (2005).

[36]K. E. Dunn, F. Dannenberg, T. E. Ouldridge, M. Kwiatkowska, A. J. Turberfield, and J. Bath, "Guiding the folding pathway of DNA origami," Nature **525**, 82–86 (2015).

[37]J. M. Schaeffer, C. Thachuk, and E. Winfree, "Stochastic simulation of the kinetics of multiple interacting nucleic acid strands," in *DNA Computing and Molecular Programming*, Boston and Cambridge, MA, August 17-21, 2015, edited by A. Phillips and P. Yin (Springer: Cham, 2015) pp. 194–211.

[38]S. Zolaktaf, F. Dannenberg, E. Winfree, A. Bouchard-Côté, M. Schmidt, and A. Condon, "Efficient parameter estimation for DNA kinetics modeled as continuous-time markov chains," in *DNA Computing and Molecular Programming*, Seattle, WA, August 5-9, 2019, edited by C. Thachuk and Y. Liu (Springer: Cham, 2019) pp. 80–99.

[39]R. J. Menssen and A. Tokmakoff, "Length-dependent melting kinetics of short DNA oligonucleotides using temperature-jump IR spectroscopy," J. Phys. Chem. B **123**, 756–767 (2019).

[40]P. J. Sanstead, P. Stevenson, and A. Tokmakoff, "Sequence-dependent mechanism of DNA oligonucleotide dehybridization resolved through infrared spectroscopy," J. Am. Chem. Soc. **138**, 11792–11801 (2016).

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

[42]J. H. Gibbs and E. A. DiMarzio, "Statistical mechanics of helix-coil transitions in biological macromolecules," J. Chem. Phys. **30**, 271–282 (1959).

[43]B. H. Zimm, "Theory of "melting" of the helical form in double chains of the DNA type," J. Chem. Phys. **33**, 1349–1356 (1960).

[44]M. Eigen and D. Pörschke, "Co-operative non-enzymic base recognition: I. thermodynamics of the helix-coil transition of oligoriboadenylic acids at acidic pH," J. Mol. Biol. **53**, 123–141 (1970).

[45]K. Marimuthu and R. Chakrabarti, "Sequence-dependent theory of oligonucleotide hybridization kinetics," J. Chem. Phys. **140**, 175104 (2014).

[46]D. F. Calef and J. Deutch, "Diffusion-controlled reactions," Ann. Rev. Phys. Chem. **34**, 493–524 (1983).

[47]X. Chen, Y. Zhou, P. Qu, and X. S. Zhao, "Base-by-base dynamics in DNA hybridization probed by fluorescence correlation spectroscopy," J. Am. Chem. Soc. **130**, 16947–16952 (2008).

[48]D. Pörschke, "Model calculations on the kinetics of oligonucleotide double helix coil transitions. evidence for a fast chain sliding reaction," Biophys. Chem. **2**, 83–96 (1974).

[49]P. J. Sanstead and A. Tokmakoff, "Direct observation of activated kinetics and downhill dynamics in DNA dehybridization," J. Phys. Chem. B **122**, 3088–3100 (2018).

[50]B. Nölting, *Protein Folding Kinetics: Biophysical Methods* (Springer: Berlin, 2006).

[51]E. Weinan and E. Vanden-Eijnden, "Towards a theory of transition paths," J. Stat. Phys. **123**, 503 (2006).

[52]P. Metzner, C. Schütte, and E. Vanden-Eijnden, "Illustration of transition path theory on a collection of simple examples," J. Chem. Phys. **125**, 084110 (2006).

[53]P. Metzner, C. Schütte, and E. Vanden-Eijnden, "Transition path theory for markov jump processes," Multiscale Model. Simul. **7**, 1192–1219 (2009).

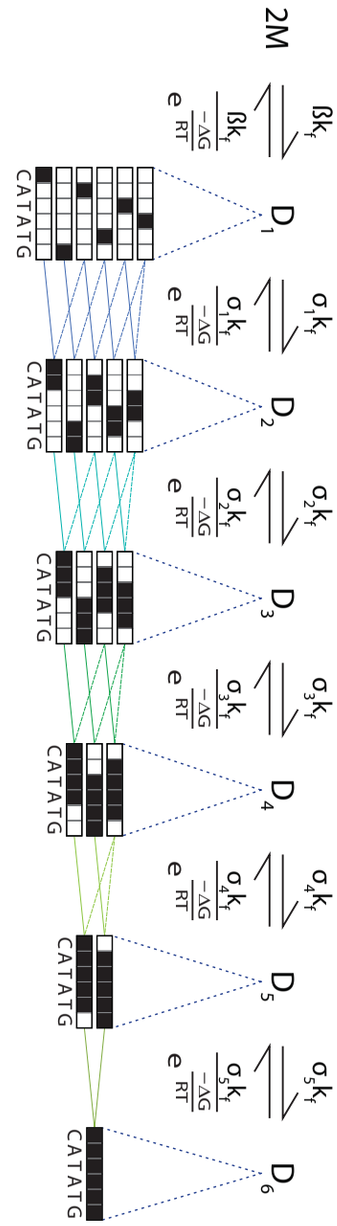
[54]E. Weinan and E. Vanden-Eijnden, "Transition-path theory and path-finding algorithms for the study of rare events," Ann. Rev. Phys. Chem. **61**, 391–420 (2010).

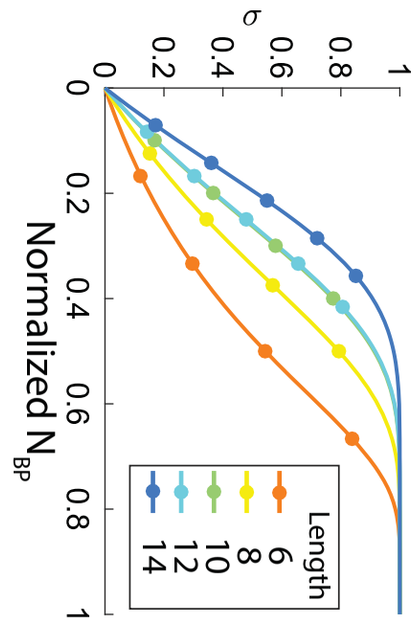
[55]T. Ohmichi, H. Nakamuta, K. Yasuda, and N. Sugimoto, "Kinetic property of bulged helix formation: Analysis of kinetic behavior using nearest-neighbor parameters," J. Am. Chem. Soc. **122**, 11286–11294 (2000).

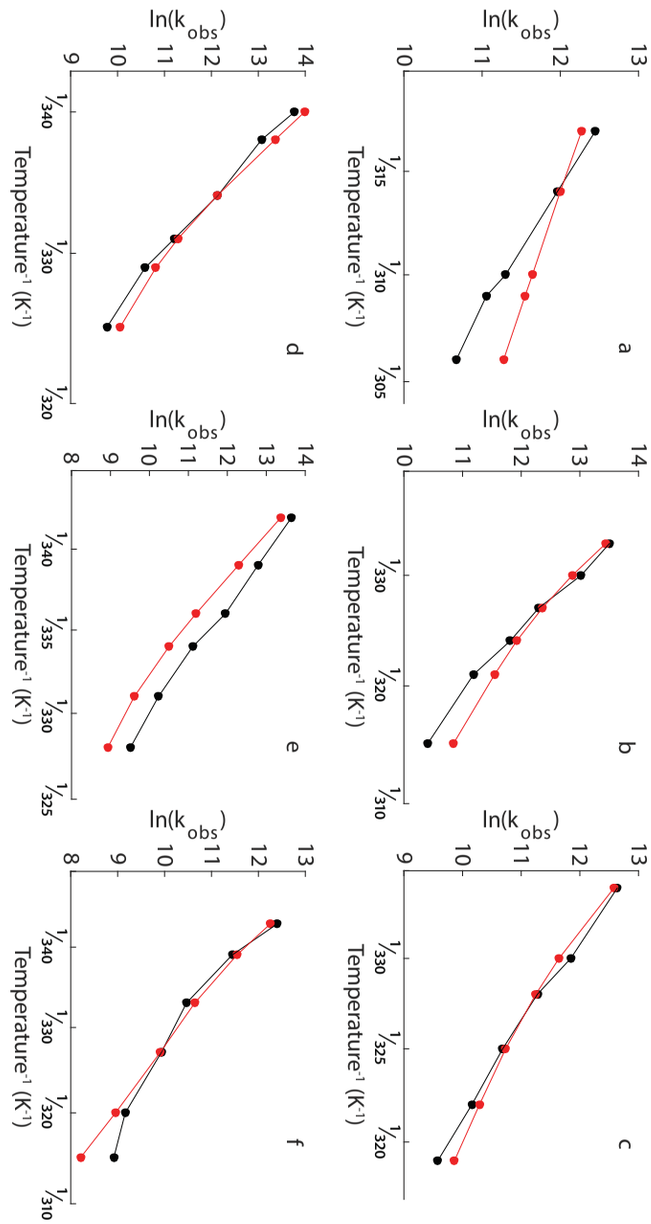
[56]C. Chen, W. Wang, Z. Wang, F. Wei, and X. S. Zhao, "Influence of secondary structure on kinetics and reaction mechanism of DNA hybridization," Nucleic Acids Res. **35**, 2875–2884 (2007).

[57]E. N. Nikolova, G. D. Bascom, I. Andricioaei, and H. M. Al-Hashimi, "Probing sequence-specific DNA flexibility in A-tracts and pyrimidine-purine steps by nuclear magnetic resonance $^{13}$C relaxation and molecular dynamics simulations," Biochemistry **51**, 8654–8664 (2012).

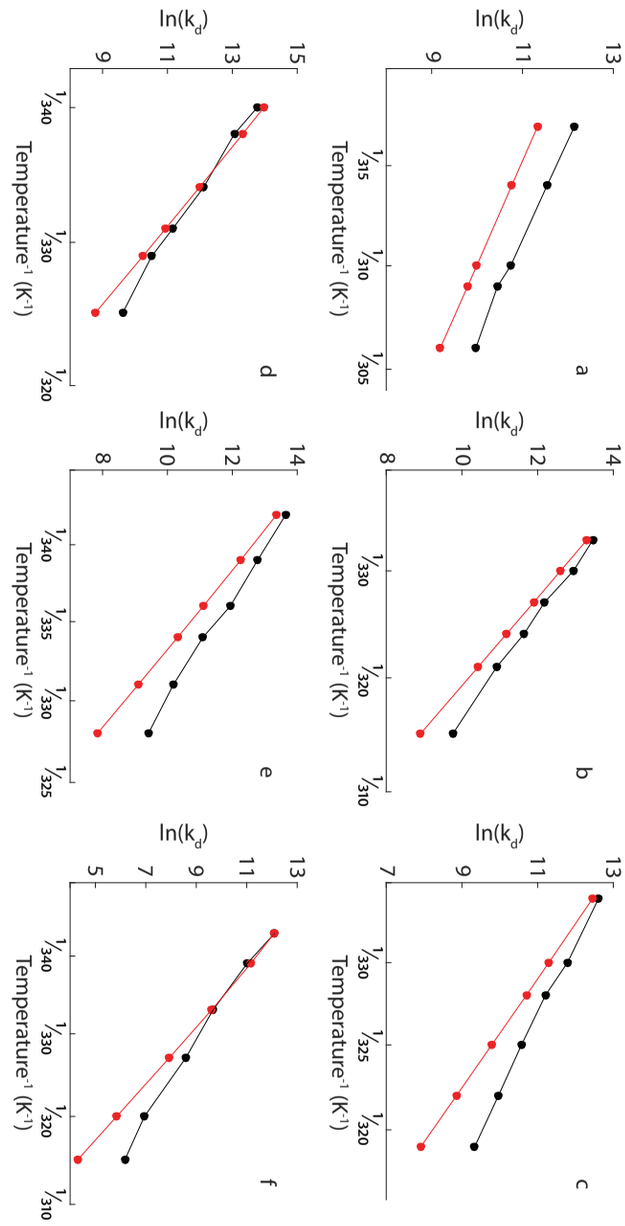
[58]M. Zgarbová, M. Otyepka, J. Šponer, F. Lankaš, and P. Jurečka, "Base pair fraying in molecular dynamics simulations of DNA and RNA," J. Chem. Theory Comput. **10**, 3177–3189 (2014).



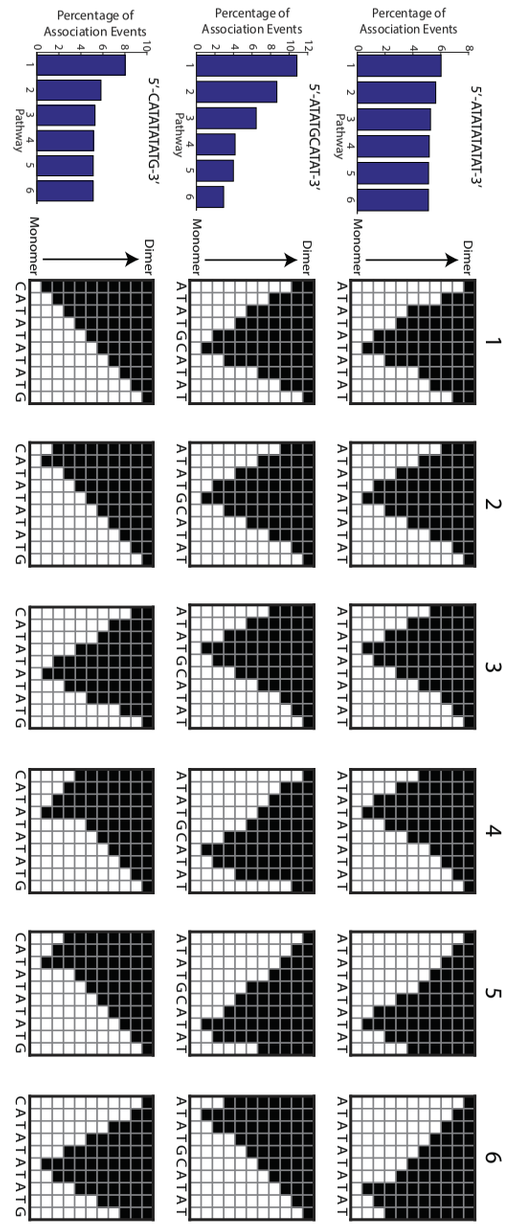




PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0035187



PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0035187



PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0035187