Determining sequence-dependent DNA oligonucleotide hybridization and dehybridization mechanisms using coarse-grained molecular simulation, Markov state models, and infrared spectroscopy

Michael S. Jones,[†] Brennan Ashwood,[‡] Andrei Tokmakoff,[‡] and Andrew L. Ferguson[*,†]

†*Pritzker School of Molecular Engineering, The University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, United States*

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

E-mail: andrewferguson@uchicago.edu

Abstract

A robust understanding of the sequence-dependent thermodynamics of DNA hybridization has enabled rapid advances in DNA nanotechnology. A fundamental understanding of the sequence-dependent kinetics and mechanisms of hybridization and dehybridization remains comparatively underdeveloped. In this work, we establish new understanding of the sequence-dependent hybridization/dehybridization kinetics and mechanism within a family of self-complementary pairs of 10-mer DNA oligomers by integrating coarse-grained molecular simulation, machine learning of the slow dynamical modes, data-driven inference of long-time kinetic models, and experimental temperature-jump infrared spectroscopy. For a repetitive ATATATATAT sequence, we resolve a rugged dynamical landscape comprising multiple metsatable states, numerous competing hybridization/dehybridization pathways, and a spectrum of dynamical relaxations. Introduction of a G:C pair at the terminus (GATATATATC) or center (ATATGCATAT) of the sequence reduces the ruggedness of the dynamics landscape by eliminating a number of metastable states and reducing the number of competing dynamical pathways. Only by introducing a G:C pair midway between the terminus and the center to maximally disrupt the repetitive nature of the sequence (ATGATATCAT) do we recover a canonical "all-or-nothing" two-state model of hybridization/dehybridization with no intermediate metastable states. Our results establish new understanding of the dynamical richness of sequence-dependent kinetics and mechanisms of DNA hybridization/dehybridization, present a molecular basis with which to understand experimental temperature jump data, and furnish foundational design rules by which to rationally engineer the kinetics and pathways of DNA association and dissociation for DNA nanotechnology applications.

1 Introduction

Over the last couple of decades, DNA has proven to be much more than a vessel for genetic information. From sensing to computing to directed self-assembly, the programmable and predictable nature of DNA has unlocked numerous unforeseen nanotechnology applications.[1–4] Recently, single molecule localization techniques have exploited the rapid and transient binding of short DNA oligomers in order to achieve super-resolution microscopy and optical multiplexing.[5–7] Predictive understanding of the sequence-dependent thermodynamics of DNA hybridization/dehybridization – the assembly/disassembly of a DNA duplex from two single strands – has underpinned the rational design of DNA oligomer sequences for nanotechnology applications, where sequence-dependent nearest-neighbor models can accurately account for mismatched pairs, dangling ends, and other non-native bonding effects.[8,9] Secondary DNA structures such as hairpins and G-quadruplexes have also been studied in depth and leveraged for nanotechnology applications.[10–12] Predictive models of the dynamical, as opposed to purely thermodynamical, behaviors of DNA have become increasingly important in developing technologies such as DNA-PAINT (DNA Points Accumulation For Imaging In Nanoscale Topography), but these technologies have outpaced our fundamental understanding of the dynamics themselves.[13–15] Many experimental and computational studies have investigated DNA dynamical phenomena over picosecond to millisecond time scales.[16–20] Kinetic models have been developed for particular DNA processes such as toehold exchanges and optical barcoding[21,22] and supervised machine learning techniques have been combined with experimental measurements to predict the on/off rates as a function of sequence.[6,23,24]. A comprehensive understanding of the full dynamical landscape of hybridization/dehybridization accounting for the sequence-dependent metastable states and association/dissociation pathways remains lacking and fundamental questions remain unresolved. For example, it remains unclear the extent to which hybridization of short DNA oligomers largely proceeds in a conventionally assumed "all-or-nothing" fashion or if long-lived metastable states facilitate the transition.[17,19,25–28] Out-of-register "shifted" base paired

structures[17,18,25,29] and frayed structures[30–33] stand as candidates for metastable states with the potential to mediate substantial deviations from all-or-nothing behavior, but the degree to which these states are kinetically relevant is difficult to determine experimentally and is likely to be highly sequence-dependent. The development of predictive models and design rules with which to engineer DNA strands with tailored hybridization/dehybridization kinetics and pathways is vital to advancing rational design of DNA strands for nanotechnology applications and is also of importance in understanding fundamental biological processes such as transcription and gene regulation.

Our understanding of hybridization dynamics is built upon decades of experiments – such as temperature-jump, salt-jump, pH-jump, and other perturbative methods – that rapidly stimulate DNA and monitor relaxation to a new equilibrium.[28,34–40] More recently, single molecule diffusion and tethered multifluorophore assays have facilitated analyses under equilibrium conditions, but these results can be hampered by slow data collection rates and fluorescent tags effects on strand dynamics, particularly for shorter oligomers.[23,41–43] A number of computational modeling approaches have also been employed to provide molecular-level resolution of hybridization. Simplified lattice models can recapitulate the essential aspects of the hybridization pathways but lack the realism of continuous space representations[25,44]. The long time scales associated with hybridization/dehybridization events place them outside the reach of unbiased all-atom molecular dynamics simulations,[18] but they can be observed by employing enhanced sampling techniques[18,45–52] or by using elevated temperature or denaturing solvent concentrations to induce one-way dissociation events.[53,54] The effect of the applied bias upon the thermodynamics can be rigorously corrected for using standard reweighting techniques, but approaches to rigorously correct the kinetics, particularly under the conditions of high bias necessary for good sampling, are in their infancy.[55–60] A number of coarse-grained DNA force fields have been developed that enable direct observation of these events over microsecond time scales via unbiased coarse-grained molecular dynamics simulations,[29,47,52,61,62] which, up to a speedup factor associated with the smooth-

ing of the underlying free energy landscape inherent to the coarse-graining procedure, can preserve a faithful model of the unbiased dynamics and associated pathways. These models have previously been used to study biological phenomena such as nucleosome dynamics[63,64] and transcription factor binding[65,66] as well as nanotechnology applications such as strand displacement[67,68] and DNA origami[69,70].

In this work, we study a family of self-complementary pairs of 10-mer DNA oligomers using coarse-grained molecular simulation, machine learning of the slow dynamical modes, and data-driven inference of long-time kinetic models to establish new understanding of the influence of sequence upon hybridization/dehybridization kinetics and mechanisms. This family of oligomers – 5′-ATATATATAT-3′ (AT-all), 5′-GATATATATC-3′ (GC-end), 5′-ATATGCATAT-3′ (GC-core), and 5′-ATGATATCAT-3′ (GC-mix) – was designed to probe the influence of the placement of a G:C pair within an otherwise repetitive A:T sequence and has been the subject of our prior experimental investigations using temperature-jump infrared spectroscopy and simple lattice models.[19] We validate the new computational models of hybridization/dehybridization dynamics developed in this work against new experimental data and reinterpret our prior experimental observations in light of the new computational understanding. Consistent with previous studies,[18,25,52] we find the degree of repetitiveness in the sequence – and therefore the kinetic accessibility and thermodynamic stability of out-of-register shifted states – leads to richer dynamics populated by a diversity of long-lived metastable states. Our data-driven modeling and analysis rigorously quantifies these behaviors and furnishes accurate predictive models of the hybridization/dehybridization rates, dynamical pathways, and metastable states. Specifically, we demonstrate that disrupting repetitive stretches of A:T bases by placement of interrupting G:C base pairs enables us to tune the landscape from rich six-state to simple two-state "all-or-nothing" behavior, and the specific location of the interrupting pair can be used to modulate the stability of long-lived frayed states. Taken together, our analyses establish new molecular-level understanding of the sequence-dependent kinetics and pathways through quantitative predictive models for

the long-time system dynamics, resolution of the dynamical folding pathways and metastable states, and elementary design rules with which to sculpt and qualitatively alter the dynamical behaviors of the system. We anticipate that this new foundational understanding, and the extension of our approach to more extensive families of DNA sequences, can guide the rational design of DNA oligomers with tailored kinetic properties engineered for DNA nanotechnology applications such as DNA-PAINT.[6,7]

2 Methods

2.1 Computational Methods

2.1.1 Molecular dynamics simulations

We performed molecular dynamics simulations of four 10-base self-complementary double-stranded DNA sequences that we have previously studied by temperature-jump infrared spectroscopy[19]: 5′-ATATATATAT-3′ (AT-all), 5′-GATATATATC-3′ (GC-end), 5′-ATATGCATAT-3′ (GC-core), and 5′-ATGATATCAT-3′ (GC-mix). We modeled the DNA sequences using the coarse-grained 3-Site-Per-Nucleotide v2 (3SPN.2) model that uses three spherical beads to represent the phosphate, deoxyribose sugar, and nitrogenous base of each nucleotide and employs anisotropic interaction potentials to accurately treat intra-strand base-stacking, inter-strand cross-stacking, and base pairing.[47] The model was parameterized against experimental data on bond lengths, bend angles, torsional angles, base step energies, and base stacking free energies, and reliably reproduces the structure, melting temperatures, persistence lengths, and sequence, salt, concentration, and temperature effects on duplex formation.[47] The model enables access to millisecond time scales and has been has widely-adopted to study numerous phenomena including DNA packing in viral capsids, protein-DNA binding, and nucleosome unwrapping.[63,71,72] Although the 3SPN.2 model was not directly parameterized against dynamical experimental data, we will show below that the predicted

sequence-dependent kinetics and relaxation are, within a corrective scaling factor, in good agreement with observed experimental trends.

All calculations were performed using the LAMMPS simulation package (`http://lammps.sandia.gov`) in accordance with best practices for the 3SPN.2 model.[73] A single pair of self-complementary sequences were placed in a cubic periodic box with side length 7.8 nm corresponding to a single-strand concentration of 7 mol/L. Solvent effects were modeled implicitly by employing Langevin dynamics[74,75] with an experimentally motivated per-site friction coefficient of 9.94×10^{-11} m^2/s.[47,76] We specified a 240 mM implicit salt concentration and treated electrostatic interactions using the Debye-Hückel with a 5 nm cutoff radius.[77] Simulations were performed in the NVT ensemble employing a Langevin thermostat.[78]. Each sequence was simulated at its melting temperature – AT-all: 309 K, GC-end: 317 K, GC-core: 324 K, GC-mix: 324 K – in order to maximize the number of spontaneous transitions between dissociated and hybridized states. Melting temperatures for each sequence were determined empirically by comparing the ratio of hybridized to dissociated populations over a 10 K temperature ramp centered on the nearest neighbor (NN) model predicted melting temperature,[8,9] and selecting the temperature at which the hybridized and dissociated populations were approximately equal. The Langevin equations of motion were integrated using the scheme of Bussi and Parrinello[75] with a 20 fs integration time step. We performed 40 independent simulations for each of the four sequences with half of the runs initialized from the hybridized state and half from the dissociated state. The initial hybridized state was defined based on the crystal structure coordinates of Arnott et al.[79] The dissociated state was generated from the hybridized state by displacing one strand away from the other by 1 nm in each of the x, y, and z directions. The magnitude of this displacement did not affect the results so long as all native Watson-Crick (WC) bonds were completely broken. Initial bead velocities were assigned from a Maxwell-Boltzmann distribution at the temperature of interest. Each simulation was conducted for 26 μs and frames saved to disc every 100 ps. Each simulation required $\sim$24 CPU-hours on 28$\times$Intel E5-2680v4 CPU cores. The first 1 μs

of each run was discarded for equilibration providing us with 40×25 μs = 1 ms of simulation data for each sequence, during which time we observed 55-100 hybridization/dehybridization events.

2.1.2 Markov state model construction

Markov state models (MSMs) are a powerful approach to infer long-time kinetic models from short molecular simulation trajectories[80–84] that we employ in this work to construct high-resolution sequence-dependent kinetic models of DNA hybridization and dissociation. In brief, a MSM extracts from simulation trajectories an ensemble of long-lived metastable macrostates, their equilibrium occupancy populations, and the equilibrium probability fluxes between them. In this manner, they provide an interpretable and predictive model of the system thermodynamics and kinetics. MSMs have recently been implemented to study the hybridization mechanism of one particular 14-mer DNA oligomer, but determining the sequence-dependent kinetics and mechanisms was not the focus of this study.[85] An energy disconnectivity graph-based approach was used to interrogate the differences in hybridization rates and mechanisms between GGGGGG and GCGCGC hexamers to reveal strong deviations from "all-or-nothing" behaviors and the importance of zippering and slithering mechanisms.[17] Kinetic models were constructed not from the dynamical trajectories of the molecular model, but by estimating rate constants between local minima were estimated using a transition state theory approximation. A recent application of MSMs to the long-time dynamics of short RNA oligonucleotides revealed stacking time scales to be highly sequence dependent.[86] In this work, MSMs were constructed for each of the four DNA sequences from the 40×25 μs simulation trajectories following a four-step protocol detailed in Ref.[87]: (i) trajectory featurization, (ii) dimensionality reduction, (iii) microstate clustering and microstate transition matrix inference, (iv) macrostate clustering and macrostate transition matrix inference. Calculations were performed using the PyEMMA software package.[88]

Featurization. Trajectories comprising the Cartesian coordinates of the DNA strands

as a function of time were featurized using the MDtraj Python libraries[89] to represent the system in a manner that exposes the essential system dynamics but eliminates trivial translation and rotational invariances. We adopt intermolecular pairwise distances $d(i, j)$ between the centers of mass of the 10 bases as a natural rototranslationally invariant featurization that represents each system configuration as the 10×10=100-element vector of inter-strand pairwise distances. One additional symmetry arises from the self-complementary nature of these sequences – the sense and anti-sense strands in each pair are identical – such that the representation of the system under our featurization should remain unchanged upon inverting the arbitrary labeling of strand "1" and strand "2".[80] The 100-element pairwise distance vector is not invariant to this permutation, but can easily be made so via a simple symmetrization operation in which each of the $\binom{10}{2} = 45$ intermolecular pairwise distances are replaced by the mean of the two permutationally invariant distances. Specifically, $(d(i_1, j_2) = d(i_2, j_1)) \leftarrow 0.5\,(d(i_1, j_2) + d(i_2, j_1))$, where i_1 denotes the i^{th} base on strand 1 and j_2 the j^{th} base on strand 2.[80] Finally, we took the reciprocal of the permutationally-symmetrized pairwise distances to provide higher resolution and differentiation between proximate strand configurations in the near hybridized state compared to distantly separated dissociated strands. VAMP-2 scoring – calculation of the sum of the squared estimated eigenvalues of the transfer operator – of trajectories under a particular featurization provides a measure of the kinetic variance carried by that featurization.[90–93] Performing VAMP-2 scoring at a lag time of $\tau = 1.2$ ns and retaining the top five modes, reveals that the reciprocal permutationally-symmetrized pairwise distances can carry up to twice the kinetic variance as the non-reciprocal distances, suggesting that the higher resolution offered at close intermolecular distances can indeed boost the dynamical representational power of the model. Somewhat surprisingly, we observed that augmenting our set of intermolecular distances between bases with intramolecular distances between bases on the same strand led to no improvement of the VAMP-2 score. This indicates that the kinetically-relevant conformational state of the two strands are adequately represented via the intermolecular

distances and leading us to employ only intermolecular distances within our featurization.

Dimensionality reduction. The featurized trajectories were then projected into a low-dimensional space in preparation for microstate clustering. The standard approach to doing so is to employ time-lagged independent components analysis (tICA) to learn a linear projection into a low-dimensional embedding that maximally preserves the kinetic variance in the data.[91,94,95] In this work, we instead employ state-free reversible VAMPnets (SRVs) that can be conceived of as a nonlinear version of tICA.[96] SRVs employ neural networks to learn flexible nonlinear functions of the trajectory featurization that better approximate the slow dynamical modes of the system and have been shown to produce substantially higher resolution MSMs than those developed using tICA.[87,96] SRV modes were learned independently for each system to best approximate the slow collective modes for that particular DNA sequence. SRVs were trained using the SRV package we previously developed (`https://github.com/hsidky/srv`) employing the default network architecture of two hidden layers each comprising 100 neurons and tanh activation functions, a learning rate of 0.001, and a batch size of 50,000. We adopted a lag time of $\tau = 1.2$ ns as appropriately short time scale to resolve the dynamical details of the hybridization/dehybridization dynamics.[97] As observed by Husic and Pande, the lag time cannot be treated as a hyperparameter to be optimized via the VAMP-2 score, but must be selected as a physically-motivated choice designed to expose the dynamical motions relevant at a particular time and length scale of interest.[98] As we shall show, this choice of lag time leads to high-resolution Markovian macrostate MSMs. We guarded against overfitting using five-fold cross-validation in which we divided the trajectory data into 50:50 training:validation splits. We observed plateau of the validation loss and no evidence of overfitting after 20 epochs of training requiring approximately 22 GPU-minutes on a single NVIDIA Tesla K80 GPU card. A VAMP-2 scoring of the cumulative kinetic variance explained as a function of number of SRV collective modes also showed no evidence of overfitting – as would be evinced by separation of the training and validation VAMP-2 scores[87] – and exhibited a knee for each of the four DNA sequences after

the fifth, fourth, third, and second slow modes for AT-all, GC-end, GC-core, and GC-mix, respectively (Fig. S1 in the Supporting Information), and motivating the construction of 5D, 4D, 3D, and 2D embeddings, respectively.

Microstate clustering. The SRV projections of the 10 million frames recorded over the course of the 1 ms molecular simulation trajectories collected for each DNA sequence were then clustered into microstates using k-means clustering. The VAMP-2 score of the microstate transition matrix constructed for each sequence at the selected $\tau = 1.2$ ns was insensitive to the choice of the number of microstates over the range 100-1000, motivating our selection of 200 microstate clusters for each system.

Macrostate clustering. The 200 microstates comprising each system were coarsened into our terminal macrostate MSM. The microstate transition matrix for each system was computed at a range of lag times τ and then diagonalized to recover the corresponding eigenvalues λ_i and associated implied time scales $t_i = -\tau / \ln |\lambda_i|$.[84] The implied time scale plots for the four DNA sequences are presented in Fig. S2. We observe rapid convergence of the implied time scales t_i with lag time τ for all systems, motivating the construction of high resolution macrostate MSMs at a lag time $\tau = 1.2$ ns. For this choice of lag time, we recover 5, 4, 2, and 1 implied time scales for the AT-all, GC-end, GC-core, and GC-mix systems, respectively. The identification of $(i-1)$ implied time scales implies the presence of $(i-1)$ slow modes and motivates the coarsening of the system into i macrostates. We estimate these i macrostates by applying PCCA+ spectral clustering to the leading $(i-1)$ eigenvectors of the microstate transition matrix.[99–101] We then estimate the corresponding 6, 5, 3, and 2 macrostate transition matrices $\mathbf{P}$ for the AT-all, GC-end, GC-core, and GC-mix systems, respectively, by projecting the molecular simulation trajectories into these discrete macrostates. These macrostate MSMs constitute our terminal kinetic models. We validate the Markovian nature of the four MSMs by subjecting them to the Chapman-Kolmogorov (CK) test[55,84,102]. This test asserts that the transition matrix for a Markovian (i.e., memoryless) MSM constructed at a lag time τ should satisfy the condition $\mathbf{P}(k\tau) =$

$\mathbf{P}^k(\tau)$, which states that k successive applications of the transition matrix constructed at a lag time τ should be equivalent to a single application of the transition matrix constructed at a lag time $k\tau$. We present in Fig. S3 the CK tests for each DNA sequence to demonstrate that the $\tau = 1.2$ ns models perform very well in predicting transition probabilities out to $k\tau = 7.2$ ns, validating the Markovian nature and kinetic validity of the four models.

2.2 Experimental Methods

2.2.1 Sample Preparation

Each DNA oligonucleotide was purchased from Integrated DNA Technologies (IDT) at desalt grade purity. Oligonucleotides were purified with 3 kD cutoff centrifugal filters (Amicon). All labile protons were exchanged in deuterium oxide (D_2O, Cambridge Isotopes, 99.9%). Oligonucleotides were prepared at a total strand concentration of 2 mM in 50 mM pD 7.2 sodium phosphate buffer with 240 mM NaCl and 18 mM $MgCl_2$. Prior to each measurement, DNA solutions were placed in a water bath at 90° C for three minutes and then cooled to room temperature under ambient conditions.

2.2.2 T-Jump IR Spectroscopy

The details of the technique and processing used to acquire temperature-jump infrared (T-jump IR) data have been described previously.[103–105] Briefly, heating was initiated through optical excitation of the O-D stretch overtone transition of D_2O. The 1.98 μm pulses (5 ns, 20 mJ, 20 Hz) used for heating were generated from the frequency-doubled output of a Nd:YAG laser sent through an optical parametric oscillator. Nonlinear IR spectra are collected from 5 ns to 50 ms delays after the T-jump with a synchronized 1 kHz spectrometer. T-jump heterodyne-detected vibrational echo (t-HDVE) IR spectra were acquired with Fourier transform spectral interferometry,[104] where the delay between the local oscillator (LO) and DVE signal was scanned in 5 fs steps between (-10) and 10 fs. t-HDVE spectra were acquired with a parallel pulse polarization scheme and presented as a dispersed pump-

probe (t-DPP) spectrum. t-DPP data are reported as the difference spectra relative to the initial temperature.

The sample was placed between two 1 mm thick CaF2 windows separated by a 50 μm Teflon spacer enclosed in a brass jacket. The initial temperature of the sample was set using a recirculating chiller connected to the brass sample jacket. The T-jump temperature change (ΔT) was set to 14-16 °C for all measurements and monitored using the change in transmission of the D_2O bend-libration combination band measured in the LO beam. The temperature change was quantified by comparing the change in transmission of the LO beam with a FTIR temperature series of D_2O.

2.2.3 Determination of Fast and Slow Dissociation Rates

To determine observed rates from the T-jump data, the time-domain t-HDVE data was inverse Laplace transformed into the rate domain using a maximum entropy approach (MEMiLT).[106] Observed rates λ^{fast} and λ^{slow} were computed from the amplitude-weighted mean rate across detected IR frequency, as previously described.[28] The fast response k_d^{fast} is defined as this amplitude-weighted mean rate, whereas the dissociation rate constant k_d^{slow} was extracted from the observed rate of the slow response λ^{slow} using a two-state model for self-complementary oligomers,[107]

$$\lambda^{\text{slow}} = k_d^{\text{slow}} + 4[S]_{T_f} k_a^{\text{slow}}, \tag{1}$$

where $[S]_{T_f}$ is the concentration of single-strand oligomer at the final temperature of the T-jump, and k_a^{slow} is the association rate constant. In practice, Eqn. 1 is recast in terms of the dissociation equilibrium constant K_d to solve for k_d^{slow} and k_a^{slow} as a function of K_d and $[S]_{T_f}$,

$$k_d^{\text{slow}} = \frac{\lambda^{\text{slow}} K_d(T_f)}{K_d(T_f) + 4[S]_{T_f}}, \tag{2}$$

$$k_a^{\text{slow}} = \frac{k_d^{\text{slow}}}{K_d(T_f)}. \tag{3}$$

FTIR temperature series were measured for each sequence as reported previously,[19] and the second SVD component along temperature was fit to a two-state all-or-nothing model to determine the fraction of intact duplex θ as a function of temperature.[108] K_d and $[S]_{T_f}$ were then determined from θ at the final temperature of each T-jump measurement,

$$[S]_{T_f} = [1 - \theta(T_f)]c_{tot}, \tag{4}$$

$$K_d(T_f) = \frac{2c_{tot}[1 - \theta(T_f)]^2}{\theta(T_f)}, \tag{5}$$

where c_{tot} is the total oligonucleotide concentration, which was 2 mM for all measurements.

3 Results and Discussion

3.1 Sequence-dependent coarse-grained kinetics recapitulate T-jump IR measurements

We first sought to demonstrate that the coarse-grained 3SPN.2 model accurately recapitulates the experimentally observed kinetics of DNA oligomer hybridization/dissociation by validating our computational predictions against temperature-jump infrared (T-jump IR) experiments. We conducted T-jump IR experiments for each DNA sequence as a function of temperature and extracted the "slow" k_d^{slow} (10-30 μs) and "fast" k_d^{fast} (70-100 ns) rates estimated from the T-jump response. The slow response has previously been attributed to duplex dissociation on microsecond time scales induced by the rapid heating of the initially hybridized duplex.[19,28] The fast response has been assigned to terminal base pair fraying,[19,28] with this process corresponding to a relatively complex dynamical process that can span time scales from picoseconds to microseconds.[31–33,109] All-atom simulations suggest that frayed ends can assume misaligned WC bonds, base-sugar hydrogen bonds, and terminal stacked

conformations.[30,110]

To computationally mimic the T-jump process in our 3SPN.2 simulations, we conducted 1 μs simulations of an initially hybridized DNA duplex over a range of temperatures and monitored its structural relaxation. We performed 120 independent simulations for each DNA sequence at each temperature, and from these extracted computational estimates of k_d^{slow} and fast k_d^{fast} (Fig. S4). First, we tracked the slow response corresponding to duplex dissociation in our simulations by compiling the distribution of times at which both of the central base pairs first separate to a distance of 1.3 nm starting from an initial fully hybridized duplex. This cutoff was selected as the distance beyond with the strands are effectively non-interacting and defines the dissociated state. We extracted our computational estimate of k_d^{slow} by fitting a decaying exponential to the fraction of hybridized sequences as a function of time $f_{\text{hybridized}}(t) = \exp(-k_d^{\text{slow}}t)$. Second, we tracked the fast response corresponding to terminal base pair fraying by compiling the distribution of times at which either of the terminal base pairs first separated to a distance of 1.3 nm, corresponding to a complete breakage of the WC interaction. We extracted our computational estimate of k_d^{fast} through a decaying exponential fit to the fraction of unfrayed sequences as a function of time $f_{\text{unfrayed}}(t) = \exp(-k_d^{\text{fast}}t)$.

We present in Fig. 1 a comparison of k_d^{slow} and k_d^{fast} estimated by computation and experiment. Although the 3SPN.2 model was not directly fitted against kinetic data,[47] its predictions of sequence dependent T-jump relaxation rates are, within a systematic scaling factor in time and systematic shift in temperature, in good agreement with observed experimental trends. It is well-known that the smoothing of the underlying free energy landscape induced by coarse-graining artificially accelerates the kinetics of coarse-grained molecular simulations and that different degrees of freedom may be accelerated by different factors.[111–113] We find that the simulated slow responses corresponding to center-of-mass translation of the strands during dissociation of the duplex is $\sim$10$\times$ accelerated relative to experiment, whereas the fast responses corresponding to fraying of the terminal bases is

$\sim$120$\times$ accelerated. We apply these sequence-independent scaling factors to our reported computational values. Although the 3SPN.2 model reproduces melting temperatures relatively well, we observed a systematic 4 K under-prediction relative to experiment and so we apply a universal (+4) K corrective calibration to our computational results. We note that these empirical calibration factors to the 3SPN.2 predictions are applied only for the purposes of making an experimental comparison, but note that there are uncertainties introduced by assuming that the equilibrium dynamics at fixed temperature can be compared directly to relaxation kinetics following a 15°C T-jump. The computational time scales and melting temperatures reported in the remainder of the paper are not corrected since we are only interested in the relative trends in the behaviors of the four sequences.

In Fig. 1a we observe an exponential increase of k_d^{slow} with temperature, as expected from the large enthalpic barrier to duplex dissociation.[20,36,38] Under the time and temperature calibration corrections, we see generally very good agreement between the computational and experimental curves. The computational rates do tend to systematically under-predict the experimental values at low temperatures, an effect that we ascribe to the fact that the experimental T-jump responses likely contain a mixture of dissociation and hybridization events – the latter of which persist to lower temperatures than the former – whereas the computational analysis considers dissociation alone. Of the four sequences, GC-core shows the largest discrepancy between computation and experiment, although the general exponential trend is preserved. This may be a result of its high propensity to fray (cf. Section 3.5). In Fig. 1b we expose a largely linear dependence of k_d^{fast} upon temperature for AT-all, GC-end, and GC-mix compared to an exponential dependence for GC-end. These trends can be understood in light of the comparatively larger enthalpic barrier for dissociation of the terminal G:C base pair in GC-end compared to that for the A:T terminal pair in the other three sequences. Again we see good agreement between the scaled computational predictions and the experimental T-jump measurements. The favorable comparison of computation and experiment provides support for the capacity of the 3SPN.2 model to reliably reproduce

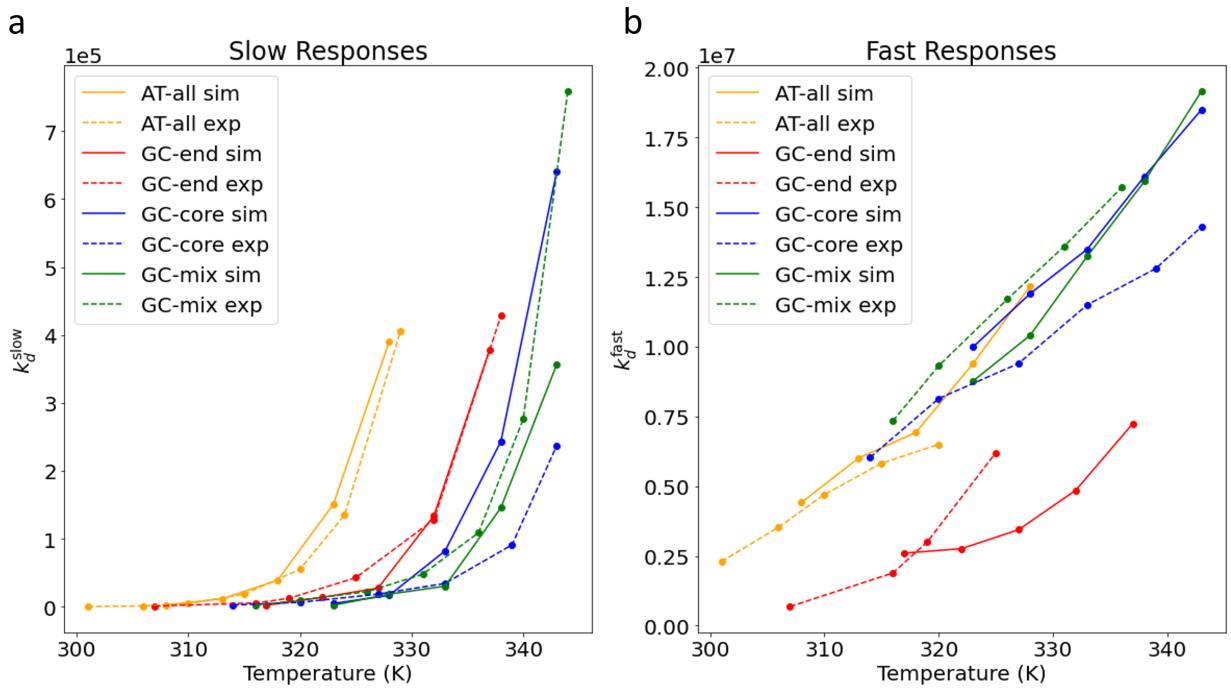


Figure 1: Experimental measurements and computational predictions of slow and fast at T-jump IR responses. Results are reported in terms of the final T-jump temperature. (a) The experimental and simulated slow rate constants k_d^{slow} corresponding to duplex dissociation over long time scales. (b) The experimental and simulated fast rate constant k_d^{fast} corresponding to terminal base-pair fraying on short time scales. The simulation results are corrected by a sequence-independent scaling factor that corrects for a 10× acceleration of the slow dissociation dynamics and 120× acceleration of the fast fraying dynamics. The simulated temperature in all cases is subjected to a (+4) K corrective calibration to account for an observed systematic under-prediction of the melting temperature by the 3SPN.2 model

sequence-dependent trends in the slow and fast kinetics of the four DNA oligomers.

3.2 SRV-MSMs for each sequence

We now proceeded to construct Markov state models (MSMs) from 1 ms of aggregated simulation trajectories for each of the four sequences at their respective melting temperatures to generate sequence-dependent kinetic models. MSMs define the long-lived metastable macrostates of the system, their equilibrium occupancies, and the equilibrium transition probabilities between them. As such, they are extremely valuable in providing both a quantitative predictive model and a physically comprehensible mechanistic understanding of the

long-time dynamical evolution of the system between an ensemble of metastable macrostates. We present in Fig. 2 the inferred MSMs for each of the four 10-base DNA sequences. Across all four sequences we identify a totality of seven metastable macrostates corresponding to the fully hybridized state (H) in which all native base pairings are intact, four shifted states in which the strands are translated out-of-register by two or four bases in the $5'$ (5S2, 5S4) or $3'$ (3S2, 3S4) direction, a frayed state (F4) – unique to GC-core – in which four terminal A:T base pairs are unbound, and the fully dissociated state (D). We present these seven macrostates in Fig. 2a along with schematic and cartoon renderings of representative microstates contained within each of these macrostates. In Fig. 2b we present the occupancy probabilities of each state at thermodynamic equilibrium. By virtue of the fact that each sequence is simulated at its corresponding melting temperature (T_m), the probability of the dissociated state (D) is, by construction, approximately equal to the sum of the probabilities over the remaining six states (H, 5S2, 3S2, 5S4, 3S4, F). In Fig. 2c we present a visualization of the macrostate MSMs for each of the four sequences showing the connectivity between the identified macrostates. The macrostates are represented as orange circles in proportion to their equilibrium probabilities and the grey arrows indicate the probability of hopping from one macrostate to another under one time step of the kinetic model. The fluxes between the macrostates provide a wealth of high-level, interpretable information on the sequence-dependent metastable states and hydbrization/dehybridization pathways. Immediately, we identify that the AT-all sequence possesses a rich and complex dynamical landscape comprising six metastable states whereas at the other end of the spectrum GC-mix exhibits far simpler two-state "all-or-nothing" behavior. In Fig. 2d we present the so-called implied time scales of each MSM. These time scales correspond to the relaxation times of the DNA dimer among its constituent metastable macrostates. The leading implied time scale for each system corresponds to the characteristic time scale for hybridization/dehybridization. Since each system is simulated at the same concentration and at its respective melting temperature, it is not surprising that the leading time scale is approximately equal for all four

systems and corresponds to the characteristic time scale for hybridization/dehybridization. The spectrum of higher order time scales corresponds to increasingly quicker relaxations between the metastable macrostates within the kinetic model and resolve the interesting sequence-dependent differences in the hybridization/dehybridization kinetic pathways. The total number of implied time scales is typically one fewer than the number of metastable macrostates and the existence of large implied time scales is indicative of slowly relaxing kinetic processes. The dense spectrum of slow implied time scales for AT-all is indicative of its relatively complex kinetic landscape whereas that for the two-state GC-mix comprises only a single time scale corresponding to hybridization/dehybridization. We now proceed to analyze in detail the sequence-dependent thermodynamics, kinetic, and mechanisms exposed by the four MSMs.

3.3 Comparison of MSM thermodynamic predictions with NN model

We first compare the thermodynamic predictions for the equilibrium macrostate probabilities to those of predictive sequence-dependent models for the thermodynamics of DNA association. The nearest neighbor (NN) model is a popular empirical model of DNA hybridization thermodynamics that predicts the free energy of duplex formation as a sum over helix initiation terms and the hybridization free energies of nearest neighbor pairs of bases that account for both the specific WC pairings and the modulating effects of the local (i.e., nearest neighbor) environment.[8,9] The parameters of the NN model were estimated by regressing over 108 experimental measurements to furnish a predictive model for the free energy of association as a function of DNA sequence and explicitly account for stacking contributions of native pairs, internal mismatches, and dangling ends. We apply the NN model to predict the free energy F^{NN} of each of the macrostates occupied by each of the four sequences. A full accounting of our application of the NN model is provided in the Supporting Information. The free energy of each macrostate is related to its equilibrium occupancy probability P via the statistical mechanical relationship $F = -k_B T \ln P + C$, where T is temperature, k_B is Boltzmann's

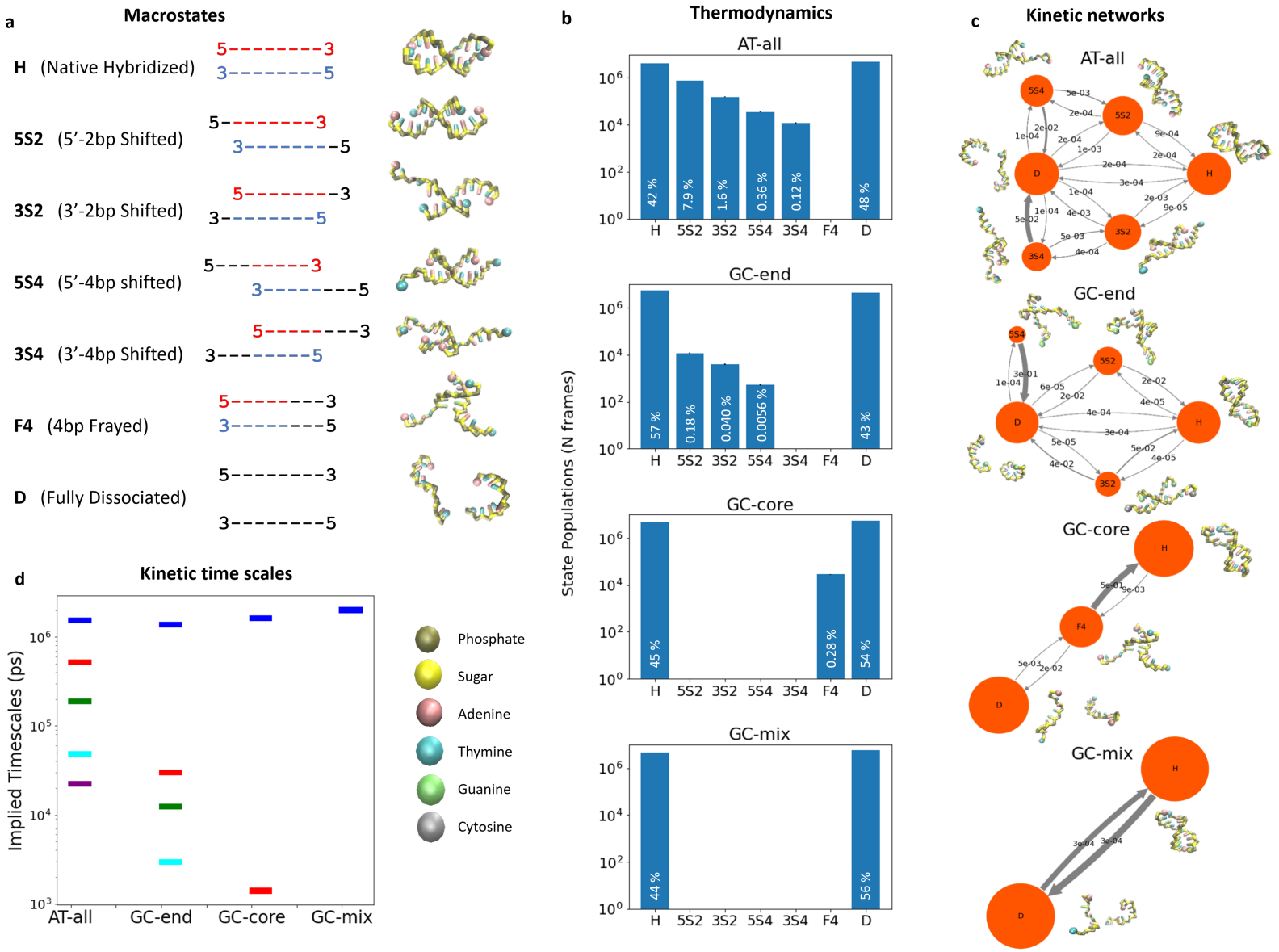

a
Macrostates
H (Native Hybridized)
5--------3
3--------5
5S2 (5'-2bp Shifted)
5-------3
3--------5
3S2 (3'-2bp Shifted)
5--------3
3-------5
5S4 (5'-4bp shifted)
5---------3
3-------5
3S4 (3'-4bp Shifted)
5---------3
3---------5
F4 (4bp Frayed)
5---------3
3---------5
D (Fully Dissociated)
5--------3
3--------5
b
Thermodynamics
State Populations (N frames)
AT-all
42 %
7.9 %
1.6 %
0.36 %
0.12 %
48 %
GC-end
57 %
0.18 %
0.040 %
0.0056 %
43 %
GC-core
45 %
0.28 %
54 %
GC-mix
44 %
56 %
H 5S2 3S2 5S4 3S4 F4 D
c
Kinetic networks
AT-all
GC-end
GC-core
GC-mix
d
Kinetic time scales
Implied Timescales (ps)
AT-all
GC-end
GC-core
GC-mix
Phosphate
Sugar
Adenine
Thymine
Guanine
Cytosine

Figure 2: Thermodynamic and kinetic predictions of the sequence-specific MSMs for AT-all, GC-end, GC-core, and GC-mix. (a) **Macrostates.** Schematic representation of the seven metastable macrostates occupied by one or more of the four sequences: fully hybridized state (H), two-base out-of-register shifted states in the $5'$ (5S2) or $3'$ (3S2) direction, four-base out-of-register shifted states in the $5'$ (5S4) or $3'$ (3S4) direction, and the fully dissociated state (D). The line drawings represent the 10-base self-complementary sequences, where red-to-blue contacts indicates (possible) WC base pairing and black indicates an unbound bases. Adjacent to each line drawing we provide representative molecular structures corresponding to that macrostate. (b) **Thermodynamics.** Histograms reporting the number of the 10^7 total frames within the 1 ms of simulation trajectories observed to occupy each of the seven macrostates, corresponding to our numerical estimates of the equilibrium occupancy probabilities. Uncertainties are calculated across 100 MSMs using a Bayesian MSM estimation are reported for each bar and are very small compared to the total counts. (c) **Kinetic networks.** MSMs illustrating the kinetic network for each sequence. The orange circles correspond to the macrostates occupied by each sequence and are labeled by the macrostate codes reported in panel a. The radii of the circles is proportional to the logarithm of the equilibrium occupancy probabilities reported in panel b. Molecular renderings of an illustrative snapshot from the coarse-grained molecular simulations are provided next to each macrostate. The grey arrows between macrostates indicate the presence of a probability flux between this pair of states at equilibrium and the arrow thickness is proportional to the flux. (To avoid congesting the diagram, arrows are not reported for probability fluxes lower than 3×10^{-6}.) The numerical value overlaid on each arrow reports the conditional probability that a system occupying the macrostate at the start of the arrow at time t will transition to the macrostate at the end of the arrow by time $(t + \tau)$, where τ=1.2 ns is the lag time corresponding to a single time step of the MSM. Large orange circles correspond to thermodynamically favorable states and large grey arrows correspond to kinetically favorable transitions. (d) **Kinetic time scales.** Distribution of MSM implied time scales for each sequence. The leading implied time scale corresponds to the characteristic time scale for hybridization/dehybridization and is approximately equal for all systems since simulations were conducted at the same concentration and at the respective melting temperatures. The higher order implied time scales correspond to a spectrum of kinetic relaxations between the constituent macrostates in the MSM corresponding to shifted and/or frayed states.

constant, and C is an additive constant reflecting our ignorance of the absolute scale of free energies. We use this relationship to convert the equilibrium occupancy probabilities predicted by our MSM and reported in Fig. 2b into free energies F^{MSM}. The unknown additive constants preclude comparisons of absolute free energies between the MSM and NN model, but it is legitimate to compare relative free energies between pairs of macrostates since the additive constant cancels in taking differences. As such, we arbitrarily set the additive constant C in both the MSM and NN model such that the hybridized state H defines a zero free

energy reference state and we report the stability of all macrostates relative to the hybridized state as $\Delta F^{\mathrm{NN}} = F^{\mathrm{NN}} - F_H^{\mathrm{NN}}$ for the NN model and $\Delta F^{\mathrm{MSM}} = F^{\mathrm{MSM}} - F_H^{\mathrm{MSM}}$ for the MSM.

As illustrated in Fig. 3, we see that the MSM tends to predict higher relative free energies for the out-of-register shifted macrostates 5S2, 3S2, 5S4, and 3S4 compared to the NN model, such that the MSM predicts lower equilibrium occupancies of these states. The NN and MSM predictions for the frayed state F4 are in good agreement. Although the trends are in qualitative agreement, what is the root of the quantitative discrepancy of the MSM and NN models in the predicted relative stability of the shifted states? First, the MSM is constructed bottom-up from molecularly detailed 3SPN.2 simulations whereas the NN model is fitted top-down by regression against experimental data. There are approximations inherent in the 3SPN.2 model, not least of which is the coarse-grained representation that integrates over atomic degrees of freedom, and in the NN model that was fitted to limited experimental data assuming a low-order expansion in terms of nearest neighbor additive contributions. Second, although 3SPN.2 is expected to capture some dangling end stabilization effects through base stacking and cross-stacking interactions, the model was not parameterized to fully capture the enthalpic contributions of the interaction of unbound bases with terminal base pairs, whereas this term is explicitly included within the NN model. Third, a well known deficiency of the NN model is the absence of any treatment of inert tails – free bases that extend beyond the dangling end tend to destabilize the duplex.[114]

We can further explore the role of inert tails upon macrostate stability by analyzing the AT-all and GC-end sequences that occupy out-of-register shifted macrostates 5S2 and 3S2 comprising a one-base inert tail and 5S4 and 3S4 comprising a three-base inert tail (cf. Fig. 2a). For both sequences, our simulations show that 5S2 is both the most stable of shifted states relative to H (Fig. 3a) and has the smallest discrepancy ($\sim$4 kJ/mol AT-all, $\sim$10 kJ/mol GC-end) compared to NN predictions (Fig. 3b). The 3S2 and and 5S4 states have nearly the same deviation from NN predictions ($\sim$7 kJ/mol AT-all, $\sim$15 kJ/mol GC-end), indicating that longer tails in the 5′ direction are as destabilizing as shorter tails

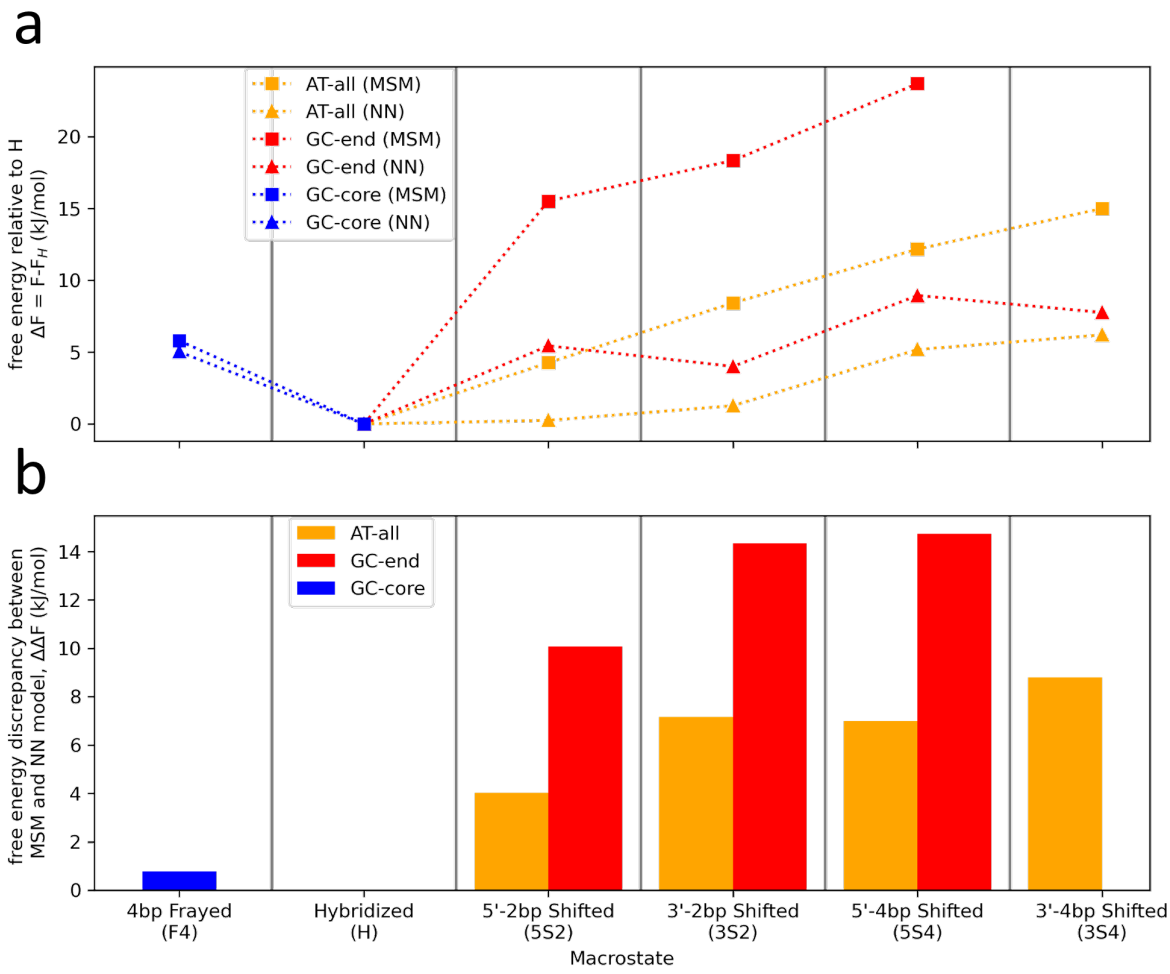


Figure 3: Comparison of the macrostate free energy predictions of the MSMs and nearest neighbor (NN) thermodynamic model.[8,9] (a) Free energies of each macrostate relative to the hybridized state $\Delta F = F - F_H$. We define the hybridized state H to possess a free energy of zero and take care to only compare relative free energies (i.e., ΔF) between the MSM and NN model. (b) Discrepancy between the macrostate relative free energy predictions $\Delta\Delta F = \Delta F^{\mathrm{MSM}} - \Delta F^{\mathrm{NN}}$ of the MSM relative to the NN model. The MSM tends to predict higher relative free energies (i.e., lower occupancy probabilities) for the out-of-register shifted states 5S2, 3S2, 5S4, 3S4 relative to the hybridized state H compared to the NN model.

in the 3′ direction. It is known from differential scanning calorimetry studies[115] that 3′ inert tails are more destabilizing than 5′ tails, with the differential behavior attributed to a combination of 5′ tails preferentially stacking on the core duplex and 3′ tails perturbing the duplex structure.[114–116]. Both the NN and MSM predictions for AT-all are consistent

with this trend (i.e., $F_{3S2} > F_{5S2}$ and $F_{3S4} > F_{5S4}$). For GC-end, the MSM and NN models both predict the out-of-register shifted states to be less stable relative to the hybridized state than the corresponding predictions for AT-all. This is in line with expectations since the terminal G:C pairs in GC-end decrease by two the number of available WC pairings in out-of-register shifted states compared to AT-all. The GC-end NN predictions run contrary to the expectation that the $3'$ inert tails should be more destabilizing than the $5'$ tails, whereas the MSM predictions are consistent with this trend. Indeed, the MSM model for GC-end does not identify the 3S4 macrostate as a stable metastable conformation for the duplex.

In sum, the qualitative trends in the macrostate thermodynamic stabilities are in good agreement between the MSM and NN models, but show quantitative discrepancies for macrostates possessing inert tails. In these instances the MSM predicts these macrostates to be less stable relative to the hybridized state compared to the NN model predictions by 4.0-14.5 kJ/mol. The MSM predictions are also consistent with the experimental expectation that $3'$ inert tails should be more destabilizing than the $5'$ tails, whereas the NN predictions can be in conflict with this trend.

3.4 Out-of-register states facilitate hybridization and dissociation dynamics (AT-all, GC-end)

In addition to thermodynamic stabilities, the macrostate MSM also furnishes quantitative and interpretable predictions of hybridization and dehybridization pathways and mechanisms. We now proceed to analyze these predictions for each of the four sequences and illuminate the relationship between sequence and dynamics. Two of our sequences, AT-all and GC-end, support out-of-register metastable states, and we commence our analysis with the role of these shifted states.

AT-all possesses the richest and most complex MSM of the four sequences by virtue of its repetitive nature, comprising a hybridized state (H), dissociated state (D), and four out-of-register shifted states (5S2, 3S2, 5S4, 3S4) (Fig. 2c). Analysis of the MSM transition

probabilities reveal a critical role of the shifted states in mediating hybridization and dehybridization. Commencing from the dissociated state D, we observe approximately equal probabilities for transitions to each of the other five states, such that a transition to one of the out-of-register shifted states 5S2, 3S2, 5S4, or 3S4 is approximately 2.2 times more likely than a direct transition to the hybridized state H. Commencing from the hybridized state H, however, a direct transition to the dissociated state is approximately 1.2 times more likely than a transition to one of the two-base shifted states 5S2 or 3S2. Once in one of the four shifted states, the 5′ vs. 3′ overhang and degree of shifting play an important role in determining whether the duplex will transition to more shifted states, more aligned states, or completely dissociate. Transitions from more shifted states towards more aligned states (i.e., 5S4 $\rightarrow$ 5S2, 5S2 $\rightarrow$ H, 3S4 $\rightarrow$ 3S2, 3S2 $\rightarrow$ H) are approximately an order of magnitude more probable than the reverse transitions from more aligned states to more shifted states. The largest single transition probability from the four shifted states 5S2, 3S2, 5S4, and 3S4 is, however, back to the dissociated state D. Consistent with the higher destabilizing effect of 3′ inert tails relative to 5′ tails,[114–116] the 3S4 $\rightarrow$ D transition probability is twice as large as the 5S4 $\rightarrow$ D, and the 3S2 $\rightarrow$ D is four times larger than the 5S2 $\rightarrow$ D. The transition probability from the 5S2 and 3S2 states back to the dissociated state D is equal to or greater than the transition probability to the hybridized state. A transition path theory analysis of the MSM reveals that 33% of productive hybridization trajectories D $\dashrightarrow$ H (where the dashed arrow indicates the combination of both direct and indirect pathways) proceed through one or more out-of-register shifted states. Among these out-of-register pathways, the D $\dashrightarrow$ 5S2 $\dashrightarrow$ H transition is predicted to occur 57% of the time. A mean first passage time (MFPT) analysis returns a MFPT for D $\dashrightarrow$ H of 3.0 μs and for H $\dashrightarrow$ D of 2.5 μs. As expected by the fact that the calculations are performed at the approximate melting temperature of the sequence, the MFPTs are approximately equal.

GC-end comprises the next most complex MSM. The introduction of the G:C pairs at the termini of the strands maximally preserves the repetitive tract of A:T base pairings such

that the GC-end MSM possesses all of the same macrostates in its dynamical landscape with the exception of the 3S4 state (Fig. 2c). As discussed in Section 3.3, the 3S4 state is rendered unstable within the lag time of our MSM due to the presence of the destabilizing $3'$ inert tail and only four WC base pairings compared to six in the case of AT-all. Analysis of the transition probabilities reveal significant differences compared to those in the AT-all kinetic network. Commencing from the dissociated state D, we observe a similar transition probability to the 5S4 state as for AT-all, but once in the 5S4 state there are no significant transition probabilities to any other state except back to D. As such, the 5S4 state acts as a kinetic trap rather than as an intermediate to hybridization. The D $\rightarrow$ H and H $\rightarrow$ D transition probabilities are commensurate with those for AT-all. However, the D $\rightarrow$ 5S2 and D $\rightarrow$ 3S2 transition probabilities are half or less of those in AT-all, and the reverse transitions are an order of magnitude larger. This may be attributed to the reduced thermodynamic stability of the 5S2 and 3S2 states in GC-end that comprise only six WC pairs compared to eight in AT-all (cf. Fig. 3). The 5S2 $\rightarrow$ H and 3S2 $\rightarrow$ H transition probabilities are more than an order of magnitude larger than in AT-all, which may again be attributed to the lower thermodynamic stability of the two shifted states relative to the hybridized state H. Again, the transition probabilities out of the 3S2 state to D or H are comparatively higher than those out of the 5S2 state, consistent with the increased destabilizing effect of $3'$ inert tails.[114–116] Commencing from the hybridized state H, a direct transition to the dissociated state is approximately seven times more likely than a transition to one of the two-base shifted states 5S2 or 3S2. A transition path theory analysis of the MSM reveals that only 7% of hybridization events D $\dashrightarrow$ H and dehybridization events H $\dashrightarrow$ D proceed through one or more out-of-register shifted states. The significantly reduced role for out-of-register shifted states in mediating the hybridization and dissociation pathways for GC-end relative to AT-all is consistent with the reduced thermodynamic stability of these states due to the elimination of possible out-of-register WC base pairing for the terminal G:C pairs and therefore a reduced accessibility of these states in the GC-end kinetic network. We compute a MFPT for D $\dashrightarrow$

H of 1.6 μs and for H $\dashrightarrow$ D of 2.1 μs, which are again approximately equal.

The out-of-register kinetic landscape that defines AT-all and GC-end hybridization have been explored by a number of previous computational studies. Simulations have identified internal displacement mechanisms capable of correcting base pair alignment in 3SPN.2[18] as well as in the coarse-grained oxDNA[52] and BioModi[61] models. In all cases, these mechanisms were shown to be crucial components of the hybridization pathway for homogeneous and repetitive sequences. Xiao et al. performed an all-atom energy landscape-based analysis of 5′-GGGGGG-3′ and 5′-GCGCGC-3′ hexamers.[17] Out-of-register states for 5′-GCGCGC-3′ hexamers were identified as deep kinetic traps along the hybridization pathway and "slithering" through these states did not provide a significant hybridization pathway compared to an alternative "zippering" mechanism. In contrast, slithering through out-of-register shifted states and zippering served as two parallel pathways for hybridization of 5′-GGGGGG-3′. This stands in contrast to our results for our AT-all (5′-ATATATATAT-3′) sequence, in which out-of-register states participated in 33% of productive hybridization events. It is conceivable that the stronger hydrogen bonding in G:C WC pairs relative to A:T pairs may render out-of-register shifted states less favorable to hybridization by suppressing fluctuation-driven rearrangements,[117,118] but additional studies would be required to reconcile these observations.

3.5 Central GC placement induces long-lived frayed states (GC-core)

The GC-core MSM represents a departure from the relatively rich and complex kinetic networks dominated by out-of-register shifted states to a much simpler one dominated by fraying (Fig. 2c). The MSM contains only three states – hybridized H, dehybridized D, and frayed F4. The F4 state is unique to GC-core and contains up to six WC pairs – the two central G:C core pairs and as many as four A:T pairs on one side or other of the core, while the other run of four A:T pairs remains free. (As expected by symmetry, the particular

AT run that is free occurs with equal probability on either side of the core.) Although partially frayed states containing less than four free A:T bases on either end of the duplex are common, these tend to inter-convert faster than the lag time and are not registered as metastable within our MSM. Our model reveals the absence of any direct hybridization or dehybridization transitions between the H and D states, with all pathways passing through the frayed state F4. Previous studies would suggest that hybridization of this sequence should proceed via a zippering mechanism, wherein upon formation of the strong central G:C WC base pairings the duplex helix should rapidly assemble in a middle-out fashion.[16,52] Our results are partially consistent with this expectation, but reveal the frayed state F4 to be unexpectedly metastable, serving as a long-lived state with an mean life time of 1.8 ns. The stability of the state is attributable to the enthalpic stabilization offered by the up to six WC pairs and the entropic stabilization associated with the configurational entropy of the two free AT-tails.

Analysis of the transition probabilities show that commencing from the F4 state, progression to the hybridized state F4 $\rightarrow$ H is 25 times more likely than dissociation F4 $\rightarrow$ D. Thus once a D $\rightarrow$ F4 transition has occurred, a F4 $\rightarrow$ H transition will likely proceed; concomitantly, H $\rightarrow$ F4 events tend to fall back to the H state and are unlikely to proceed to complete dissociation. The transition probabilities H $\rightarrow$ F4 and D $\rightarrow$ F4 are, respectively, one half and one quarter as likely as the F4 $\rightarrow$ H transition probability. We noted in Section 3.1 that fraying dynamics in the 3SPN.2 model appear to be significantly accelerated relative to center-of-mass translation, and it is conceivable that this may lead to elevated sampling of the F4 state within the computational model relative to experiment and the induction of more frequent dissociation. Moreover, since GC-core is the sequence most prone to fraying, this effect could be the root of the relatively poorer agreement of the k_d^{slow} response for GC-core compared to the other sequences due to an artificially elevated computational prediction of this rate (Fig. 1a). Our model predicts a MFPT for D $\rightarrow$ F4 $\rightarrow$ H of 3.4 μs and for H $\rightarrow$ F4 $\rightarrow$ D of 2.9 μs, which are again approximately equal.

Lattice models have previously identified frayed states as putative intermediates in DNA hybridization/dehybridization.[19,25,44] Araque et al. studied a 5′-ATGCGCAT-3′ octomer using a lattice model and identified a symmetrically A:T frayed state as a crucial part of the duplex transition path.[25] We previously studied the four sequences that are the subject of the present work using T-jump IR and 2D IR spectroscopy and identified GC-core as possessing the highest deviation from two-state behavior during dissociation when neglecting out-of-register contributions.[19] This result was interpreted to arise from loss of A:T contacts and fraying around the central G:C core, and this hypothesis was supported by lattice model calculations that predicted the GC-core conformational ensemble to possess substantially more frayed configurations than the other three sequences.[44] Follow-up T-jump measurements and Smoluchowski simulations on model 1D free energy landscapes showed that AT termini fraying was an effectively barrierless process characterized by rapid inter-conversion between all accessible frayed states.[28] These prior results are consistent with the present findings that expose the GC-core sequence to be the only sequence that occupies the F4 frayed state and therefore the only one possessing a metastable frayed state on time scales exceeding the τ=1.2 ns lag time of our MSMs.

3.6 Disruption of repetitive AT tracts promotes two-state "all-or-nothing" kinetics (GC-mix)

The GC-mix sequence is the only one of the four sequences studied that exhibits simple two-state "all-or-nothing" behavior.[17,19,25,26] The GC-mix MSM comprises just two states, the hybridized H and dehybridized D (Fig. 2c), indicating that association and disassociation of the strands proceeds directly without passing through any metastable intermediate states resolvable under the τ=1.2 ns lag time of our MSM. The two-state behavior appears to arise as a consequence of the placement of the G:C pair that maximally disrupts the repetitive AT tract within the decamer and destabilizing either out-of-register shifted states or frayed states. We note we do observe transient fraying of the terminal two-base AT tails within our

dynamical simulations, but these frayed states are not sufficiently thermodynamically stable to produce a metastable macrostate within the resulting MSM. This stands in contrast to the metastable F4 state populated by GC-core. Our MSM predicts a MFPT for D $\rightarrow$ H of 2.9 μs and for H $\rightarrow$ D of 2.3 μs, which are again nearly equal.

Given the very simple two-state "all-or-nothing" behavior of GC-mix and the absence of any intermediate metastable states, we sought to interrogate our simulation trajectory data to elucidate the hybridization and dehybridization mechanisms. To do so, we followed all 10 intermolecular distances between native WC base pairs and tracked their evolution through a number of hybridization and dehybridization events. We present one representative example of each event in Fig. 4 and four more in Fig. S5. During the hybridization process, we observe a global decrease in all 10 distances as the strands approach one another and the formation of key native WC contacts immediately prior to duplex formation: specifically, one of the G:C WC pairs and at least one neighboring A:T pair or 2-3 central A:T pairs. This behavior is consistent with a "nucleation-zippering" mechanism as has been reported in previous studies.[16,20,35,50] In dehybridization, we observe fraying on the two-base AT-tails on one or both sides of the duplex followed by rapid dissociation of the central WC base pairs. Qualitatively, we observed some short-lived states composed of two to four native WC base pair contacts immediately before full dissociation occurs, but, in contrast to the F4 state we observe in GC-core, these conformations do not constitute a metastable state within our MSM nor do they tend to reform intact duplexes. These dissociation dynamics are consistent with a "fraying-peeling" dehybridization mechanism.[30,53,54]

3.7 Long-lived metastable shifted states predicted by the MSM are resolved by T-jump IR

Finally, we sought to validate the predictions of our sequence-dependent MSMs against experimental T-jump IR spectroscopy. T-jump IR measurements commence from a low temperatures, apply a step jump in temperature, and track the relaxation of the system

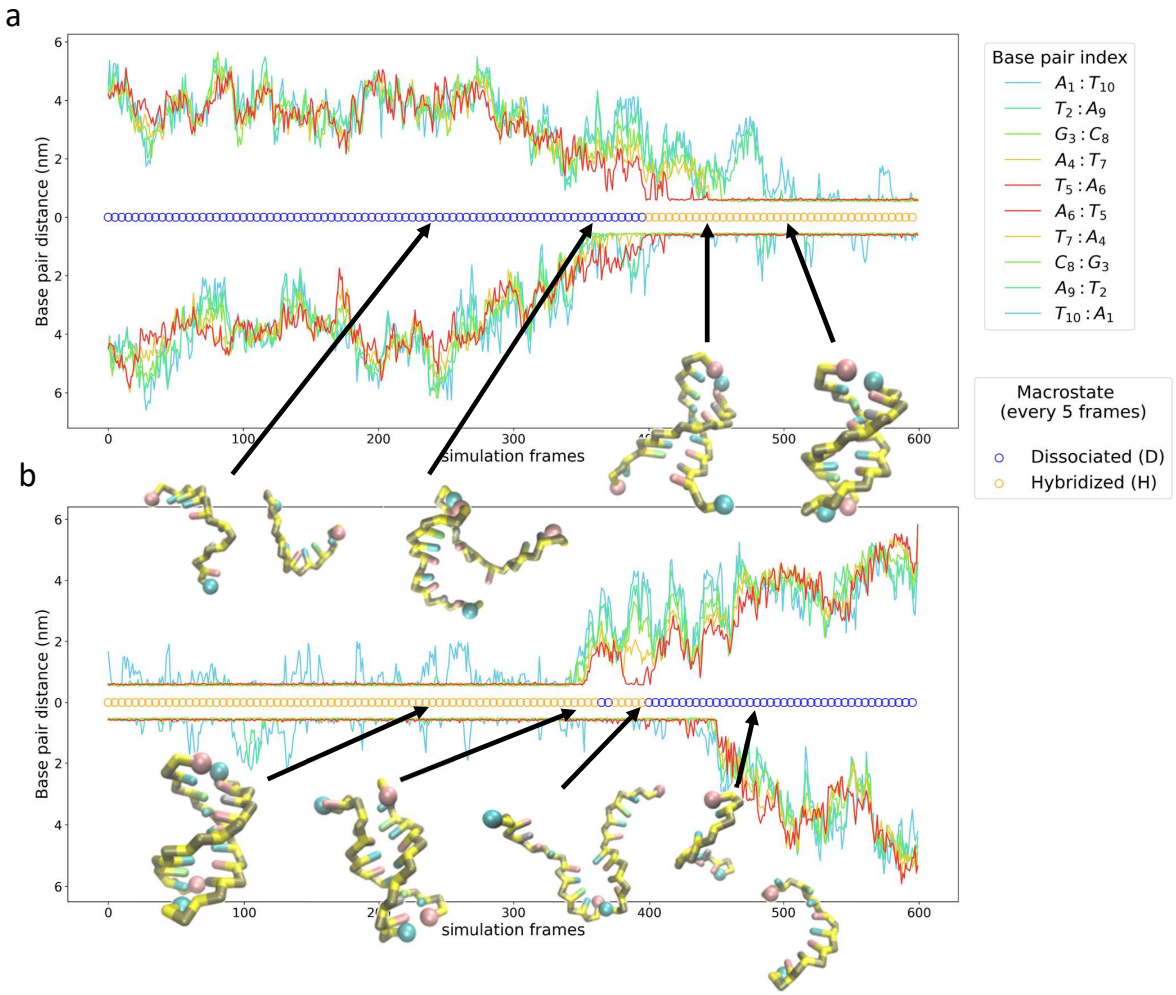


Figure 4: GC-mix hybridizes by nucleation-zippering and dehybridizes by fraying-peeling. Tracking of the 10 intermolecular distances between native WC base pairs over the course of an (a) hybridization event and (b) dehybridization event. Symmetrically permutable distances (e.g., A_1:T_{10} and T_{10}:A_1) are reflected across the x-axis to avoid congestion in the plot. Circles superposed on the x-axis indicate the instantaneous MSM state assignment as dissociated D (blue) or hybridized H (orange). Hybridization tends to occur by a nucleation-zippering mechanism, wherein a native G:C pair and adjacent A:T pair or 2-3 central A:T pairs first form prior to rapid formation of the duplex. Dehybridization tends to occur by a fraying-peeling mechanism wherein fraying of the two-base AT-tails on one or both sides of the duplex precedes dissociation of the central native base pairs and complete dissolution of the duplex. Four additional hybridization events and four additional dehybridization events are presented in Fig. S5.

to the dehybridized state. We hypothesized that the influence of the out-of-register shifted states present in the AT-all and GC-end sequences upon the system relaxation kinetics should

be manifest in the slow and/or fast responses measured by T-jump IR. As discussed in Section 3.1, the slow IR response is largely attributed to dissociation events and the fast to terminal base fraying. With regards to the slow response, our MFPT analyses of our MSMs predict out-of-register shifting events (i.e., H $\rightarrow$ 3S2, 5S2, 3S4, 5S4) to proceed on microsecond time scales, which are commensurate with the 1.4 - 2.9 μs time scales for dehybridization (i.e., H $\rightarrow$ D) for each of the four sequences. As such, we anticipate that the dynamical relaxations associated with out-of-register shifted states proceed on similar time scales to, and may not be distinguishable from, the relaxation to the dehybridized state. Nevertheless, the presence of these out-of-register shifted states in the low-temperature equilibrium ensemble prior to the T-jump step may be observable via their influence on the fast T-jump IR response attributable to fraying. Specifically, we hypothesize that the dangling ends and inert tails present in the out-of-register shifted states should promote an elevated fraying response over the course of the relaxation, and should therefore be manifest in the observation of a more stretched relaxation over experimental time scales of 70-100 ns. Analysis of the MSM equilibrium distributions (Fig. 2b) reveals 24% of the equilibrium ensemble to reside in out-of-register shifted states 3S2, 5S2, 3S4, and 5S4 for AT-all, compared to just 0.3% for GC-end, and 0% for GC-core and GC-mix. It is our conjecture that the substantial presence of out-of-register shifted states in the pre-T-jump AT-all ensemble should be distinguishable from the GC-end, GC-core, and GC-mix as an elongation of the fast relaxation response associated with terminal base fraying.

We present in Fig. 5 our T-jump IR t-HDVE difference spectra and corresponding normalized time traces at 1600 cm^{-1} and 1660 cm^{-1}. The signal at 1600 cm^{-1} corresponds to changes in A and T ring vibrations while the signal at 1660 cm^{-1} contains contributions from G and T carbonyl vibrations. Each time trace was fitted to the sum of a stretched exponential and two exponentials $S(t) = A\exp\left(-(t/\tau_{\text{fast}})^{\beta_{\text{fast}}}\right) + B\exp\left(-t/\tau_{\text{slow}}\right) + C\exp\left(-t/\tau_{\text{cool}}\right)$. The stretched exponential describes the relaxation process from 5 ns to 1 μs, and the two exponentials describe the signal increase from 1-320 μs and signal decay from re-hybridization

induced by thermal relaxation back to the initial temperature. The fitting parameters A, B, and C correspond to the relative amplitudes of the three kinetic responses and the stretch factor β_{fast} to the heterogeneity of fraying dynamics at the fast timescale. A testable prediction of our hypothesis is that the long-lived out-of-register shifted states in AT-all should result in a significantly smaller value for the fitted β_{fast} parameter (i.e., a more stretched response) relative to those for GC-end, GC-core, and GC-mix. This hypothesis was supported by the experimental time series at both 1600 cm^{-1}, where $\beta_{\text{fast}}^{\text{AT-all}} = 0.3$ compared to $\beta_{\text{fast}}^{\text{GC-end}} = 0.6$, $\beta_{\text{fast}}^{\text{GC-core}} = 0.7$, and $\beta_{\text{fast}}^{\text{GC-mix}} = 0.6$, and 1660 cm^{-1}, where $\beta_{\text{fast}}^{\text{AT-all}} = 0.4$ compared to $\beta_{\text{fast}}^{\text{GC-end}} = 0.7$, $\beta_{\text{fast}}^{\text{GC-core}} = 0.6$, and $\beta_{\text{fast}}^{\text{GC-mix}} = 0.6$. This result supports our hypothesis and validates a testable prediction of our sequence-dependent MSMs.

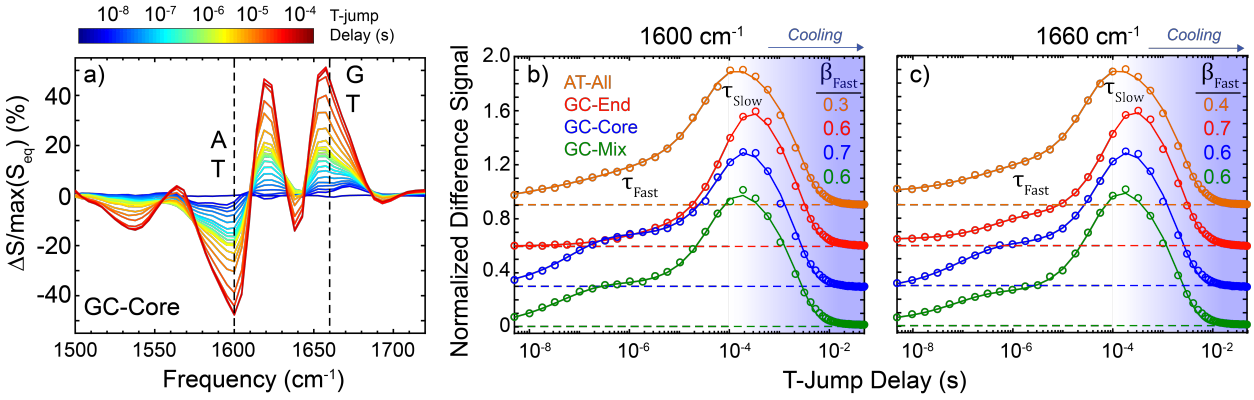


Figure 5: T-Jump IR responses reflect sequence-dependent conformational heterogeneity. (a) MSM predictions of equilibrium macrostate life times at the sequence melting temperature. (b) Mid-IR t-HDVE difference spectra for GC-core at time delays from 5 ns to 560 μs. Normalized time traces for each sequence are shown at (c) 1600 cm^{-1} and (d) 1660 cm^{-1}. The signal at 1600 cm^{-1} corresponds to changes in A and T ring vibrations while the signal at 1660 cm^{-1} contains contributions from G and T carbonyl vibrations. Each time trace is fit to the sum of stretched exponential with two exponentials (solid lines): $S(t) = A\exp\left(-(t/\tau_{\text{fast}})^{\beta_{\text{fast}}}\right) + B\exp\left(-t/\tau_{\text{slow}}\right) + C\exp\left(-t/\tau_{\text{cool}}\right)$. The stretched exponential describes the process from 5 ns to 1 μs, and the two exponentials describe the signal increase from 1-320 μs and signal decay from re-hybridization induced by thermal relaxation back to the initial temperature. The stretch factor β_{fast} for the fits at 1600 cm^{-1} and 1660 cm^{-1} are reported directly on the plots in panels c and d.

4 Conclusions

We have conducted an integrated computational and experimental study of the sequence-dependent kinetic mechanisms for the hybridization and dehybridization dynamics of a family of four self-complementary 10-mer DNA oligomers: ATATATATAT (AT-all), GATATATATC (GC-end), ATATGCATAT (GC-core), and ATGATATCAT (GC-mix). We conducted 1 ms of unbiased coarse-grained molecular dynamics simulations and employed deep learning techniques to construct high-resolution Markov state models as predictive and interpretable models of the sequence dependent dynamics. T-jump IR spectroscopy was used to calibrate the kinetic time scales of the coarse-grained molecular model and validate the kinetic prediction of the Markov state models that the AT-all sequence should possess long-lived out-of-register shifted states that are detectable within T-jump IR t-HDVE time traces. Our results reveal that the specific placement of interrupting G:C pairs within an otherwise repetitive AT sequence can have a profound impact on the kinetic pathways and mechanisms for association and dissociation of the DNA duplex. In particular, we found AT-all to possess the richest and most complex kinetic landscape of the four sequences that is dominated by out-of-register shifted states that participate in 33% of hybridization and dehybridization events. Introduction of the G:C pairs at the end of the strand maintains an eight-base pair repetitive AT tract and the GC-end kinetic landscape possess all but one of the same out-of-register shifted states as AT-all. Destabilization of the GC-end shifted states relative to AT-all, however, results in a far more limited participation of these states with only 7% of GC-end hybridization and dehybridization events passing through one or more shifted states. Placing the G:C pairs in the center of the strand maintains two four-base AT tracts either side of the core and results in qualitatively different kinetic behaviors for GC-core. In this case, no metastable out-of-register shifted states are registered by our model with the hybridization and dehybridization pathways all passing through a strongly metastable frayed state in which one or other of the four-base AT-tracts is unbound to produce two free AT-tails. Finally, placing the G:C bases between the center and end of the strand to maximally disrupt the

repetitive AT tracts results in no metastable out-of-register or frayed states for GC-mix and results in simple two-state "all-or-nothing" hybridization/dehybridization behavior. Analysis of the specific pathways reveals hybridization to largely proceed by a nucleation-zippering mechanism and dehybridization to proceed by a fraying-peeling mechanism.

The ordering of the computationally predicted kinetic landscapes from most to least complex – AT-all > GC-end > GC-core > GC-mix – is largely dictated by sequence repetitiveness, specifically the number of consecutive AT motifs. We note that this ordering differs from our previously reported ordering in terms of deviation from two-state behavior of GC-core > GC-mix > AT-all > GC-end.[19,28] We can understand these two apparently discrepant orderings by understanding that the latter was deduced based on experimental analyses and lattice models that did not account for out-of-register states and focussed largely on fraying behaviors. Indeed, under the assumption that fraying is the dominant kinetic process relative to out-of-register shifting, we can harmonize the predictions of the present work with our prior work by eliminating all out-of-register shifted states in our fitted MSMs (Fig. 2c), in which case we find GC-core to contain an F4 frayed intermediate and the remaining sequences to all have simple two-state dynamics such that the predicted ordering is GC-core > GC-mix $\approx$ AT-all $\approx$ GC-end.

In sum, our results demonstrate the profound effect of sequence upon the kinetic landscapes, metastable states, and hybridization/dehybridization mechanisms of short DNA oligomers. Our analysis of this small family of sequences expose preliminary design principles for the (meta)stability of out-of-register and frayed states but we anticipate much greater richness in the landscapes will emerge with studies of longer and more diverse sequences. Going forward, we will extend this work to discern more general trends in sequence-dependent hybridization/dehybridization for a wider range of oligomer sequences and motivate strategies for experimental comparisons. We anticipate that these insights may provide foundational design rules by which to improve understanding of *in vivo* hybridization processes and rationally engineer optimized sequences for DNA nanotechnology applications such as

DNA-PAINT[7] and DNA barcoding.[6]

Supporting Information

Details of nearest neighbor (NN) model calculations, supplementary figures illustrating selection of optimal number of SRV slow modes for each sequence, convergence of MSM implied time scales, MSM Chapman-Kolmogorov tests, "computational T-jump" calculations, and GC-mix hybridization and dehybridization trajectories.

Conflict of Interest Statement

A.L.F. is a co-founder and consultant of Evozyne, Inc. and a co-author of US Provisional Patents 62/853,919 and 62/900,420 and International Patent Applications PCT/US2020/035206 and PCT/US20/50466.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. CHE-1841805. A.T. thanks the National Institute of General Medical Sciences of the National Institutes of Health (Award No. R01-GM118774) for support of this research. B.A. acknowledges support from the NSF Graduate Research Fellowship Program. This work was completed in part with resources provided by the University of Chicago Research Computing Center. We gratefully acknowledge computing time on the University of Chicago high-performance GPU-based cyberinfrastructure supported by the National Science Foundation under Grant No. DMR-1828629.

References

(1) Seeman, N. C.; Sleiman, H. F. DNA Nanotechnology. *Nat. Rev. Mater.* **2017**, *3*, 1–23.

(2) Adleman, L. Molecular Computation of Solutions to Combinatorial Problems. *Science* **1994**, *266*, 1021–1024.

(3) Rothemund, P. W. Folding DNA to Create Nanoscale Shapes and Patterns. *Nature* **2006**, *440*, 297–302.

(4) Gu, H.; Chao, J.; Xiao, S. J.; Seeman, N. C. A Proximity-based Programmable DNA Nanoscale Assembly Line. *Nature* **2010**, *465*, 202–205.

(5) Schnitzbauer, J.; Strauss, M. T.; Schlichthaerle, T.; Schueder, F.; Jungmann, R. Super-resolution Microscopy with DNA-PAINT. *Nat. Protoc.* **2017**, *12*, 1198–1228.

(6) Shah, S.; Dubey, A. K.; Reif, J. Improved Optical Multiplexing with Temporal DNA Barcodes. *ACS Synth. Biol.* **2019**, *8*, 1100–1111.

(7) Strauss, S.; Jungmann, R. Up to 100-fold Speed-up and Multiplexing in Optimized DNA-PAINT. *Nat. Methods* **2020**, *17*, 789–791.

(8) SantaLucia, J. A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 1460–1465.

(9) Santalucia, J.; Hicks, D. The Thermodynamics of DNA Structural Motifs. *Annu. Rev. Biophys. Biomol. Struct* **2004**, *33*, 415–440.

(10) Tsukanov, R.; Tomov, T. E.; Masoud, R.; Drory, H.; Plavner, N.; Liber, M.; Nir, E. Detailed Study of DNA Hairpin Dynamics using Single-Molecule Fluorescence Assisted by DNA Origami. *J. Phys. Chem. B* **2013**, *117*, 11932–11942.

(11) Mosayebi, M.; Romano, F.; Ouldridge, T. E.; Louis, A. A.; Doye, J. P. The Role of Loop Stacking in the Dynamics of DNA Hairpin Formation. *J. Phys. Chem. B* **2014**, *118*, 14326–14335.

(12) Mergny, J. L.; Sen, D. DNA Quadruple Helices in Nanotechnology. *Chem. Rev.* **2019**, *119*, 6290–6325.

(13) Deluca, M.; Shi, Z.; Castro, C. E.; Arya, G. Dynamic DNA Nanotechnology: Toward Functional Nanoscale Devices. *Nanoscale Horiz.* **2020**, *5*, 182–201.

(14) Cordes, T.; Santoso, Y.; Tomescu, A. I.; Gryte, K.; Hwang, L. C.; Camará, B.; Wigneshweraraj, S.; Kapanidis, A. N. Sensing DNA Opening in Transcription Using Quenchable Förster Resonance Energy Transfer. *Biochemistry* **2010**, *49*, 9171–9180.

(15) Naimark, O. B.; V, B. Y.; A, B. Y.; Gagarskich, O. N.; Grishko, V. V.; Nikitiuk, A. S.; Voronina, A. O. DNA Transformation, Cell Epigenetic Landscape and Open Complex Dynamics in Cancer Development. *Math. Biol. Bioinforma.* **2020**, *15*, 251–267.

(16) Yin, Y.; Zhao, X. S. Kinetics and Dynamics of DNA Hybridization. *Acc. Chem. Res.* **2011**, *44*, 1172–1181.

(17) Xiao, S.; Sharpe, D. J.; Chakraborty, D.; Wales, D. J. Energy Landscapes and Hybridization Pathways for DNA Hexamer Duplexes. *J. Phys. Chem. Lett.* **2019**, *10*, 6771–6779.

(18) Hinckley, D. M.; Lequieu, J. P.; De Pablo, J. J. Coarse-grained modeling of DNA oligomer hybridization: Length, sequence, and salt effects. *J. Chem. Phys.* **2014**, *141*.

(19) Sanstead, P. J.; Stevenson, P.; Tokmako, A. Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved through Infrared Spectroscopy. *J. Am. Chem. Soc.* **2016**, *138*.

(20) Pörschke, D.; Eigen, Cooperative Nonenzymic Base Recognition III. Kinetics of the Helix-Coil Transition. *J. Mol. Biol.* **1971**, *62*, 361–381.

(21) Zhang, D. Y.; Winfree, E. Control of DNA Strand Displacement Kinetics Using Toehold Exchange. *J. Am. Chem. Soc.* **2009**, *131*, 17303–17314.

(22) Shah, S.; Dubey, A. K.; Reif, J. Programming Temporal DNA Barcodes for Single-Molecule Fingerprinting. *Nano Lett.* **2019**, *19*, 2668–2673.

(23) Schickinger, M.; Zacharias, M.; Dietz, H.; Schickinger, M.; Zacharias, M.; Dietz, H. Tethered Multifluorophore Motion Reveals Equilibrium Transition Kinetics of Single DNA Double Helices. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*.

(24) Zhang, J. X.; Fang, J. Z.; Duan, W.; Wu, L. R.; Zhang, A. W.; Dalchau, N.; Yordanov, B.; Petersen, R.; Phillips, A.; Zhang, D. Y. Predicting DNA Hybridization Kinetics from Sequence. *Nat. Chem.* **2018**, *10*, 91–98.

(25) Araque, J. C.; Robert, M. A. Lattice Model of Oligonucleotide Hybridization in Solution. II. Specificity and Cooperativity. *J. Chem. Phys.* **2016**, *144*.

(26) Sikora, J. R.; Rauzan, B.; Stegemann, R.; Deckert, A. Modeling Stopped-Flow Data for Nucleic Acid Duplex Formation Reactions: The Importance of Off-Path Intermediates. *J. Phys. Chem. B* **2013**, *117*, 8966–8976.

(27) Wyer, J. A.; Kristensen, M. B.; Jones, N. C.; Hoffmann, S. V.; Nielsen, S. B. Kinetics of DNA Duplex Formation: A-Tracts versus AT-Tracts. *Phys. Chem. Chem. Phys.* **2014**, *16*, 18827–18839.

(28) Sanstead, P. J.; Tokmakoff, A. Direct Observation of Activated Kinetics and Downhill Dynamics in DNA Dehybridization. *J. Phys. Chem. B* **2018**, *122*, 3088–3100.

(29) Maciejczyk, M.; Spasic, A.; Liwo, A.; Scheraga, H. A. DNA Duplex Formation with a Coarse-Grained Model. *J. Chem. Theory Comput.* **2014**, *10*, 5020–5035.

(30) Zgarbová, M.; Otyepka, M.; Šponer, J.; Lankaš, F.; Jurečka, P. Base Pair Fraying in Molecular Dynamics Simulations of DNA and RNA. *J. Chem. Theory Comput.* **2014**, *10*, 3177–3189.

(31) Nonin, S.; Leroy, J. L.; Guéron, M. Terminal Base Pairs of Oligodeoxynucleotides: Imino Proton Exchange and Fraying. *Biochemistry* **1995**, *34*, 10652–10659.

(32) Nikolova, E. N.; Bascom, G. D.; Andricioaei, I.; Al-Hashimi, H. M. Probing Sequence-Specific DNA Flexibility in AâĂŚTracts and Pyrimidine-Purine Steps by Nuclear Magnetic Resonance 13C Relaxation and Molecular Dynamics Simulations. *Biochemistry* **2012**, *51*, 8654–8664.

(33) Andreatta, D.; Sen, S.; Pérez Lustres, J. L.; Kovalenko, S. A.; Ernsting, N. P.; Murphy, C. J.; Coleman, R. S.; Berg, M. A. Ultrafast Dynamics in DNA: âĂIJFrayingâĂİ at the End of the Helix. *J. Am. Chem. Soc.* **2006**, *128*, 6885–6892.

(34) Morrison, L. E.; Stols, L. M. Sensitive Fluorescence-Based Thermodynamic and Kinetic Measurements of DNA Hybridization in Solution. *Biochemistry* **1993**, *32*, 3095–3104.

(35) Wetmur, J. G.; Davidson, N. Kinetics of Renaturation of DNA. *J. Mol. Biol.* **1968**, *31*, 349–370.

(36) Craig, M. E.; Crothers, D. M.; Doty, P. Relaxation Kinetics of Dimer Formation by Self Complementary Oligonucleotides. *J. Mol. Biol.* **1971**, *62*, 383–401.

(37) Pörschke, D.; Uhlenbeck, O. C.; Martin, F. H. Thermodynamics and Kinetics of the Helix-Coil Transition of Oligomers Containing GC Base Pairs. *Biopolymers* **1973**, *12*, 1313–1335.

(38) Williams, A. P.; Longfellow, C. E.; Freier, S. M.; Kierzek, R.; Turner, D. H. Laser

Temperature-Jump, Spectroscopic, and Thermodynamic Study of Salt Effects on Duplex Formation by dGCATGC. *Biochemistry* **1989**, *28*, 4283–4291.

(39) Narayanan, R.; Zhu, L.; Velmurugu, Y.; Roca, J.; Kuznetsov, S. V.; Prehna, G.; Lapidus, L. J.; Ansari, A. Exploring the Energy Landscape of Nucleic Acid Hairpins Using Laser Temperature-Jump and Microfluidic Mixing. *J. Am. Chem. Soc.* **2012**, *134*, 18952–18963.

(40) Chen, C.; Wang, W.; Wang, Z.; Wei, F.; Zhao, X. S. Influence of Secondary Structure on Kinetics and Reaction Mechanism of DNA Hybridization. *Nucleic Acids Res.* **2007**, *35*, 2875–2884.

(41) Liu, C.; Obliosca, J. M.; Liu, Y. L.; Chen, Y. A.; Jiang, N.; Yeh, H. C. 3D Single-Molecule Tracking Enables Direct Hybridization Kinetics Measurement in Solution. *Nanoscale* **2017**, *9*, 5664–5670.

(42) Chen, X.; Zhou, Y.; Qu, P.; Xin, S. Z. Base-by-Base Dynamics in DNA Hybridization Probed by Fluorescence Correlation Spectroscopy. *J. Am. Chem. Soc.* **2008**, *130*, 16947–16952.

(43) Dupuis, N. F.; Holmstrom, E. D.; Nesbitt, D. J. Single-Molecule Kinetics Reveal Cation-Promoted DNA Duplex Formation Through Ordering of Single-Stranded Helices. *Biophys. J.* **2013**, *105*, 756–766.

(44) Sanstead, P. J.; Tokmakoff, A. A Lattice Model for the Interpretation of Oligonucleotide Hybridization Experiments. *J. Chem. Phys.* **2019**, *150*, 1–13.

(45) Piana, S. Atomistic Simulation of the DNA Helix-Coil Transition. *J. Phys. Chem. A* **2007**, *111*, 12349–12354.

(46) Zerze, G. H.; Stillinger, F. H.; Debenedetti, P. G. Thermodynamics of DNA Hybridization from Atomistic Simulations. *J. Phys. Chem. B* **2021**,

(47) Hinckley, D. M.; Freeman, G. S.; Whitmer, J. K.; De Pablo, J. J. An Experimentally-Informed Coarse-Grained 3-Site-Per-Nucleotide Model of DNA: Structure, Thermodynamics, and Dynamics of Hybridization. *J. Chem. Phys.* **2013**, *139*, 1–17.

(48) Schmitt, T. J.; Rogers, J. B.; Knotts IV, T. A. Exploring the Mechanisms of DNA Hybridization on a Surface. *J. Chem. Phys.* **2013**, *138*.

(49) Sambriski, E. J.; Schwartz, D. C.; De Pablo, J. J. Uncovering Pathways in DNA Oligonucleotide Hybridization via Transition State Analysis. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 18125âĂŞ18130.

(50) Sambriski, E. J.; Ortiz, V.; De Pablo, J. J. Sequence Effects in the Melting and Renaturation of Short DNA Oligonucleotides: Structure and Mechanistic Pathways. *J. Phys. Condens. Matter* **2009**, *21*.

(51) Hoefert, M. J.; Sambriski, E. J.; José De Pablo, J. Molecular Pathways in DNA-DNA Hybridization of Surface-Bound Oligonucleotides. *Soft Matter* **2011**, *7*, 560–566.

(52) Romano, F.; Doye, J. P. K.; Ouldridge, T. E.; Petr, S.; Louis, A. A. DNA Hybridization Kinetics : Zippering , Internal Displacement and Sequence Dependence. *Nucleic Acids Res.* **2013**, *41*, 8886–8895.

(53) Wong, K. Y.; Pettitt, B. M. The Pathway of Oligomeric DNA Melting Investigated by Molecular Dynamics Simulations. *Biophys. J.* **2008**, *95*, 5618–5626.

(54) Perez, A.; Orozco, M. Real-Time Atomistic Description of DNA Unfolding. *Angew. Chemie - Int. Ed.* **2010**, *49*, 4805–4808.

(55) Prinz, J. H.; Chodera, J. D.; Pande, V. S.; Swope, W. C.; Smith, J. C.; Noé, F. Optimal Sse of Data in Parallel Tempering Simulations for the Construction of Discrete-State Markov models of Biomolecular Dynamics. *J. Chem. Phys.* **2011**, *134*.

(56) Chodera, J. D.; Swope, W. C.; Noé, F.; Prinz, J. H.; Shirts, M. R.; Pande, V. S. Dynamical reweighting: Improved Estimates of Dynamical Properties from Simulations at Multiple Temperatures. *J. Chem. Phys.* **2011**, *134*.

(57) Stelzl, L. S.; Kells, A.; Rosta, E.; Hummer, G. Dynamic Histogram Analysis To Determine Free Energies and Rates from Biased Simulations. *J. Chem. Theory Comput.* **2017**, *13*, 6328–6342.

(58) Donati, L.; Hartmann, C.; Keller, B. G. Girsanov Reweighting for Path Ensembles and Markov State Models. *arXiv* **2017**, *244112*, 1–15.

(59) Donati, L.; Keller, B. G. Girsanov Reweighting for Metadynamics Simulations. *J. Chem. Phys.* **2018**, *149*.

(60) Quer, J.; Donati, L.; Keller, B. G.; Weber, M. An Automatic Adaptive Importance Sampling Algorithm for Molecular Dynamics in Reaction Coordinates. *SIAM J. SCI. Comput* **2018**, *40*, 653–670.

(61) Markegard, C. B.; Fu, I. W.; Reddy, K. A.; Nguyen, H. D. Coarse-Grained Simulation Study of Sequence Effects on DNA Hybridization in a Concentrated Environment. *J. Phys. Chem. A* **2015**, *119*, 1823–1834.

(62) Dans, P. D.; Walther, J.; Gómez, H.; Orozco, M. Multiscale Simulation of DNA. *Curr. Opin. Struct. Biol.* **2016**, *37*, 29–45.

(63) Lequieu, J.; Córdoba, A.; Schwartz, D. C.; De Pablo, J. J. Tension-Dependent Free Energies of Nucleosome Unwrapping. *ACS Cent. Sci.* **2016**, *2*, 660–666.

(64) Lequieu, J.; Schwartz, D. C.; De Pablo, J. J. In Silico Evidence for Sequence-Dependent Nucleosome Sliding. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E9197–E9205.

(65) Terakawa, T.; Takada, S. P53 Dynamics Upon Response Element Recognition Explored By Molecular Simulations. *Sci. Rep.* **2015**, *5*, 1–10.

(66) Tan, C.; Takada, S. Dynamic and Structural Modeling of the Specificity in Protein-DNA Interactions Guided by Binding Assay and Structure Data. *J. Chem. Theory Comput.* **2018**, *14*, 3877–3889.

(67) Srinivas, N.; Ouldridge, T. E.; Šulc, P.; Schaeffer, J. M.; Yurke, B.; Louis, A. A.; Doye, J. P.; Winfree, E. On the Biophysics and Kinetics of Toehold-Mediated DNA Strand Displacement. *Nucleic Acids Res.* **2013**, *41*, 10641–10658.

(68) Haley, N. E.; Ouldridge, T. E.; Mullor Ruiz, I.; Geraldini, A.; Louis, A. A.; Bath, J.; Turberfield, A. J. Design of hidden Thermodynamic Driving for Non-Equilibrium Systems via Mismatch Elimination During DNA Strand Displacement. *Nat. Commun.* **2020**, *11*.

(69) Snodin, B. E.; Schreck, J. S.; Romano, F.; Louis, A. A.; Doye, J. P. Coarse-Grained Modelling of the Structural Properties of DNA Origami. *Nucleic Acids Res.* **2019**, *47*, 1585–1597.

(70) Doye, J. P.; Fowler, H.; Prešern, D.; Bohlin, J.; Rovigatti, L.; Romano, F.; Šulc, P.; Wong, C. K.; Louis, A. A.; Schreck, J. S. et al. The oxDNA Coarse-Grained Model as a Tool to Simulate DNA Origami. *arXiv* **2020**,

(71) Córdoba, A.; Hinckley, D. M.; Lequieu, J.; de Pablo, J. J. A Molecular View of the Dynamics of dsDNA Packing Inside Viral Capsids in the Presence of Ions. *Biophys. J.* **2017**, *112*, 1302–1315.

(72) Lu, W.; Bueno, C.; Schafer, N. P.; Moller, J.; Jin, S.; Chen, X.; Chen, M.; Gu, X.; de Pablo, J. J.; Wolynes, P. G. OpenAWSEM with Open3SPN2: A Fast, Flexible, and Accessible Framework for Large-Scale Coarse-Grained Biomolecular Simulations. *bioRxiv* **2020**, 1–21.

(73) Plimpton, S. Fast Parallel Algorithms for ShortâĂŞRange Molecular Dynamics. *Journal of Computational Physics* **1997**, *117*, 1–42.

(74) Dunweg, W., B.; Paul Brownian Dynamics Simulations Without Gaussian Random Numbers. *Int. J. Mod. Phys. C* **1991**, *2*, 817–27.

(75) Bussi, G.; Parrinello, M. Accurate Sampling Using Langevin Dynamics. *Phys. Rev. E* **2007**, *75*.

(76) Nkodo, A. E.; Garnier, J. M.; Tinland, B.; Ren, H.; Desruisseaux, C.; McCormick, L. C.; Drouin, G.; Slater, G. W. Diffusion Coefficient of DNA Molecules During Free Solution Electrophoresis. *Electrophoresis* **2001**, *22*, 2424–2432.

(77) Debye, P.; E., H. Zur Theorie der Elektrolyte. *Phys. Zeitschrift* **1923**, 185–206.

(78) Schneider, T.; Stoll, E. Molecular-dynamics Study of a Three-Dimensional One-Component Model for Distortive Phase Transitions. *Phys. Rev. B* **1978**, *17*, 1302–1322.

(79) Arnott, S.; Smith, P. J. C.; Chandrasekaran, *CRC Handbook of Biochemistry and Molecular Biology*; 1976; pp 411–422.

(80) Sengupta, U.; Carballo-pacheco, M.; Strodel, B. Automated Markov State Models for Molecular Dynamics Simulations of Aggregation and Self-Assembly. *J. Chem. Phys.* **2019**, *115101*, 2–5.

(81) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know About Markov State Models But Were Afraid to Ask. *Methods* **2010**, *52*, 99–105.

(82) Chodera, J. D.; Noé, F. Markov State Models of Biomolecular Conformational Dynamics. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135–144.

(83) Husic, B. E.; Pande, V. S. Markov State Models : From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396.

(84) Wehmeyer, C.; Scherer, M. K.; Hempel, T.; Husic, B. E.; Olsson, S.; Noé, F. Introduction to Markov State Modeling with the PyEMMA Software. *Living J. Comput. Mol. Sci.* **2019**, *1*, 1–12.

(85) Jin, R.; Maibaum, L. Mechanisms of DNA Hybridization: Transition Path Analysis of a Simulation-Informed Markov Model. *J. Chem. Phys.* **2019**, *150*.

(86) Pinamonti, G.; Zhao, J.; Condon, D. E.; Paul, F.; Noe, F.; Turner, D. H.; Bussi, G. Predicting the Kinetics of RNA Oligonucleotides Using Markov State Models. *J. Chem. Theory Comput.* **2017**, *13*, 926–934.

(87) Sidky, H.; Chen, W.; Ferguson, A. L. High-Resolution Markov State Models for the Dynamics of Trp-Cage Miniprotein Constructed over Slow Folding Modes Identified by State-Free Reversible VAMPnets. *J. Phys. Chem. B* **2019**, *123*, 7999–8009.

(88) Scherer, M. K.; Trendelkamp-schroer, B.; Paul, F.; Pe, G.; Ho, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-h.; Noe, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525–5542.

(89) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L. P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532.

(90) Noé, F.; Nüske, F. A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems. *Multiscale Model. Simul.* **2013**, *11*, 635–655.

(91) Noé, F.; Clementi, C. Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2015**, *11*, 5002–5011.

(92) Scherer, M. K.; Husic, B. E.; Hoffmann, M.; Paul, F.; Wu, H.; Noé, F. Variational Selection of Features for Molecular Kinetics. *J. Chem. Phys.* **2019**, *150*.

(93) Wu, H.; Noé, F. Variational Approach for Learning Markov Processes from Time Series Data. *J. Nonlinear Sci.* **2020**, *30*, 23–66.

(94) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov model Construction. *J. Chem. Phys.* **2013**, *139*.

(95) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000–2009.

(96) Chen, W.; Sidky, H.; Ferguson, A. L. Nonlinear Discovery of Slow Molecular Modes using State-Free Reversible VAMPnets. *J. Chem. Phys.* **2019**, *150*.

(97) Phys, J. C.; Prinz, J.-h.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. et al. Markov Models of Molecular Kinetics : Generation and Validation. *J. Chem. Phys.* **2011**, *134*.

(98) Husic, B. E.; Pande, V. S. Note: MSM Lag Time Cannot Be Used for Variational Model Selection. *J. Chem. Phys.* **2017**, *147*, 2015–2017.

(99) Röblitz, S.; Weber, M. Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification. *Adv. Data Anal. Classif.* **2013**, *7*, 147–179.

(100) Weber, M. Implications of PCCA+ in Molecular Simulation. *Computation* **2018**, *6*, 1–16.

(101) Kube, S.; Weber, M. A Coarse Graining Method for the Identification of Transition Rates Between Bolecular Conformations. *J. Chem. Phys.* **2007**, *126*.

(102) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the Equilibrium Ensemble of Folding Pathways From Short Off-Equilibrium Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011–19016.

(103) Chung, H. S.; Khalil, M.; Smith, A. W.; Tokmakoff, A. Transient Two-Dimensional IR Spectrometer for Probing Nanosecond Temperature-Jump Kinetics. *Rev. Sci. Instrum.* **2007**, *78*.

(104) Jones, K. C.; Ganim, Z.; Tokmakoff, A. Heterodyne-Detected Dispersed Vibrational Echo Spectroscopy. *J. Phys. Chem. A* **2009**, *113*, 14060–14066.

(105) Jones, K. C.; Ganim, Z.; Peng, C. S.; Tokmakoff, A. Transient Two-Dimensional Spectroscopy with Linear Absorption Corrections Applied to Temperature-Jump Two-Dimensional Infrared. *J. Opt. Soc. Am. B* **2012**, *29*, 118.

(106) Kumar, A. T.; Zhu, L.; Christian, J. F.; Demidov, A. A.; Champion, P. M. On the Rate Distribution Analysis of Kinetic Data Using the Maximum Entropy Method: Applications to Myoglobin Relaxation on the Nanosecond and Femtosecond Timescales. *J. Phys. Chem. B* **2001**, *105*, 7847–7856.

(107) Bernasconi, C. *Relaxation Kinetics*; Elsevier, 2012.

(108) Marky, L. A.; Breslauer, K. J. Calculating Thermodynamic Data for Transitions of any Molecularity from Equilibrium Melting Curves. *Biopolymers* **1987**, *26*, 1601–1620.

(109) Galindo-Murillo, R.; Roe, D. R.; Cheatham, T. E. Convergence and Reproducibility in Molecular Dynamics Simulations of the DNA Duplex d(GCACGAACGAACGAACGC). *Biochim. Biophys. Acta* **2015**, *1850*, 1041–1058.

(110) Pinamonti, G.; Paul, F.; Rodriguez, A.; Bussi, G. The Mechanism of RNA Base Fraying: Molecular Dynamics Simulations Analyzed with Core-Set Markov State Models. *J. Chem. Phys.* **2019**, *150*.

(111) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; De Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

(112) Fritz, D.; Koschke, K.; Harmandaris, V. A.; Van Der Vegt, N. F.; Kremer, K. Multiscale Modeling of Soft Matter: Scaling of Dynamics. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10412–10420.

(113) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini Model. *Chem. Soc. Rev.* **2013**, *42*, 6801–6822.

(114) Di Michele, L.; Mognetti, B. M.; Yanagishima, T.; Varilly, P.; Ru, Z.; Frenkel, D.; Eiser, E. Effect of Inert Tails on the Thermodynamics of DNA Hybridization. *J. Am. Chem. Soc.* **2014**, *136*, 6538–6541.

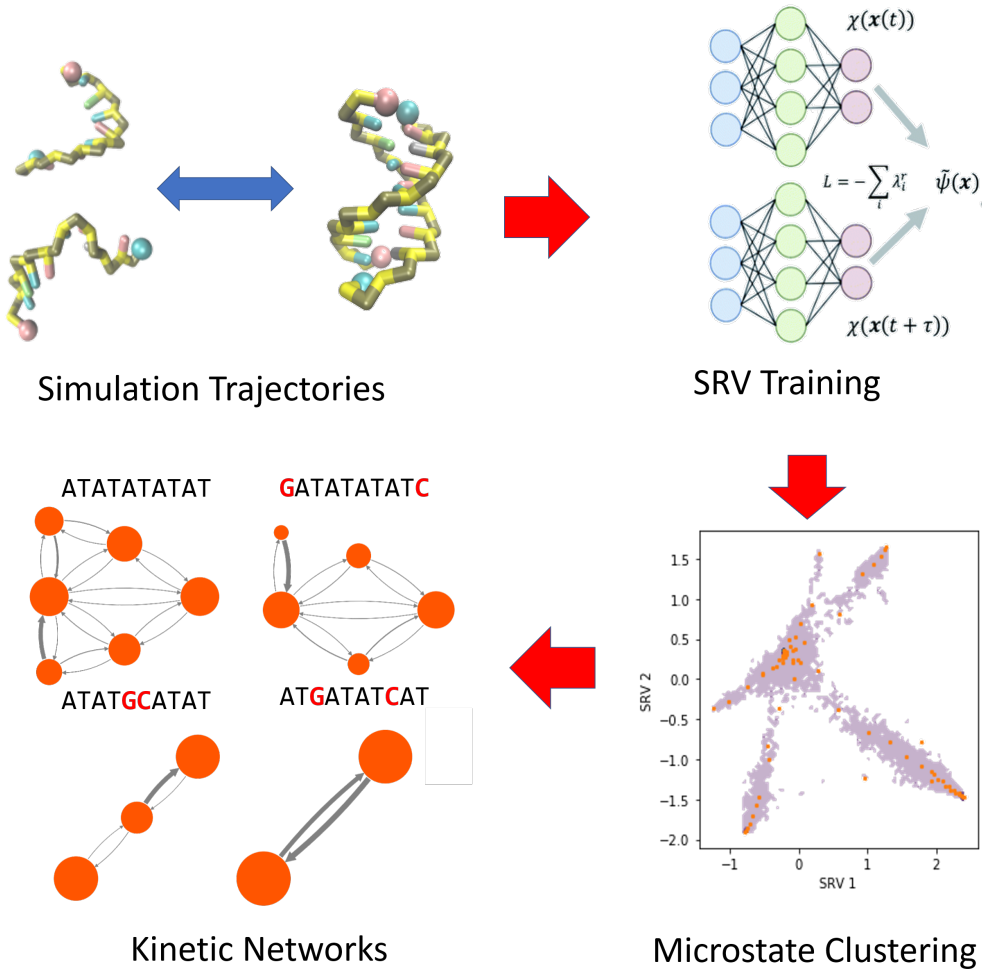
(115) Dickman, R.; Manyanga, F.; Brewood, G. P.; Fish, D. J.; Fish, C. A.; Summers, C.; Horne, M. T.; Benight, A. S. Thermodynamic Contributions of 5âĂŹ- and 3âĂŹ-Single Strand Dangling-Ends to the Stability of Short Duplex DNAs. *J. Biophys. Chem.* **2012**, *3*, 1–15.

(116) Doktycz, M. J.; Paner, T. M.; Amaratunga, M.; Benight, A. S. Thermodynamic Stability of the 5 Dangling-Ended DNA Hairpins Formed from Sequences 5- (XY) 2GGATAC (T) ,GTATCC-3, Where X, Y = A, T, G, C. *Biopolymers* **1990**, *30*, 829–845.

(117) Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M. D. Base-Stacking and Base-Pairing Contributions into Thermal Stability of the DNA Double Helix. *Nucleic Acids Res.* **2006**, *34*, 564–574.

(118) Zacharias, M. Base-Pairing and Base-Stacking Contributions to Double-Stranded DNA Formation. *The J. Phys. Chem. B* **2020**, *124*, 10345–10352.

TOC Image



For Table of Contents Only