Applying State-free Reversible VAMPNets and Markov State Models to Learn Dynamics of DNA Oligonucleotides

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

†*Pritzker School of Molecular Engineering, The University of Chicago, 929 East 57th Street, 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

NOTES

- All figs as pdf (makes compilation faster) or tiff

- For quotation marks in tex use “XXX” rather than "XXX"

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 pathways, 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 two-state model of hybridization/dehybridization over a smooth landscape 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 in burgeoning 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 nanotechnological 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 developments in dynamic DNA nanotechnologies have outpaced our understanding of the dynamics themselves.[13–15] Many experimental and computational studies have investigated DNA dynamical phenomena from the picosecond to millisecond range.[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.[23,24] [[Is this red clause correct? Are there simulation components to the work? Was it on/off rates that were predicted?]] 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. [[Accurate? More here? Does this adequately convey wrt to both citations why the two cited investigations are not the final

word. i.e., why is there still a need for the present paper – why is the problem not solved by these works, what did they leave unanswered?]] 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.[25,26] [[Should these references also be included here?[19,27,28]]] Out-of-register "shifted" base paired structures[17,25,29,30] and frayed structures[31–34] stand as candidates for metastable states with the potential to mediate substantial deviations from all-or-nothing behavior, but the extent 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 has been built from 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.[28,35–41] More recently, single molecule diffusion and tethered multifluorophore assays have facilitated equilibrium analysis, but these results can be hampered by slow data collection rates and fluorescent tags effects on strand dynamics, particularly for shorter oligomers.[23,42–44] 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[25,45]. The long time scales associated with hybridization/dehybridization events place them outside the reach of unbiased all-atom molecular dynamics simulations,[29] [[I moved this citation here – is that OK?]] but they can be observed by employing enhanced sampling techniques such as metadynamics, umbrella sampling, transition path sampling, and forward flux sampling,[18,46–51] or by using elevated temperature or denaturing solvent

concentrations to induce one-way dissociation events.[52,53] 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. {ADD REF: Girsanov (Keller ? 10.1063/1.4989474, 10.1063/1.5027728 ,10.1137/17M1124772), TRAM (Noe, Hummer ? 10.1073/pnas.1525092113), and MBAR for paths (Chodera, Shirts, Noe ? 10.1063/1.3592152, 10.1063/1.3592153)} A number of coarse-grained DNA force fields have been developed that enable direct observation of these events over microsecond time scales [[micro is OK? milli?]] via unbiased coarse-grained molecular dynamics (CGMD) simulations,[30,46,51,54,55] which, up to a speedup factor associated with the smoothing 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 do X, Y, and Z. {ADD REF: }

In this work, we study a family of self-complementary pairs of 10-mer DNA oligomers using coarse-grained molecular simulation, machine learning, 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′ATATATATAT3′ (AT-all), 5′GATATATATC3′ (GC-end), 5′ATATGCATAT3′ (GC-core), and 5′ATGATATCAT3′ (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 and existing experimental data and reinterpret our prior experimental observations in light of the new computational understanding. Consistent with previous studies,[25,29,51] we find the degree of repetitiveness in the sequence – and therefore the kinetic accessibility and thermodynamic stability of non-native 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 that facilitate the hybridization/dehybridization process. Analysis of the configurational ensemble within each metastable state reveals the conformational heterogeneity and dynamical interconversions proceeding on time scales that lie below those resolvable within the long-time kinetic model. 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 and DNA barcoding. {ADD REF: } [[Is barcoding appropriate here, or just PAINT is sufficient?]]

2 Methods

2.1 Molecular dynamics simulations

We performed molecular dynamics simulations of four 10-base self-complementary double-stranded DNA sequences that we have previously studied by ultra-fast temperature-jump infrared spectroscopy[19]: 5′ATATATATAT3′ (AT-all), 5′GATATATATC3′ (GC-end), 5′ATATGCATAT3′ (GC-core), and 5′ATGATATCAT3′ (GC-mix). We modeled the DNA sequences using the coarse-grained 3-Site-Per-Nucleotide v2 (3SPN.2) that uses three spherical beads to repre-

sent 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.[46] 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.[46] The model enables access to millisecond time scales and has been has widely-adopted to study numerous phenomenon including DNA packing in viral capsids, protein-DNA binding, and nucleosome unwrapping.[56–58] 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 excellent agreement with observed experimental trends.

All calculations were performed using the LAMMPS simulation package in accordance with best practices for the 3SPN.2 model. {ADD REF: LAMMPS} 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 XX mol/L. Solvent effects were modeled implicitly by employing Langevin dynamics[59] [[Is this the right reference for Langevin dynamics?]] with an experimentally motivated per-site friction coefficient of $\xi = 9.94 \times 10^{-11}$ m^2/s.[46] {ADD REF: A. E. Nkodo, G. M. Garnier, B. Tinland, C. Desruisseaux, L. C. Mc-Cormick, G. Drouin, and G. W. Slater,Electrophoresis22, 2424 (2001)} We specified a 240 mM implicit salt concentration and treated electrostatic interactions using the Debye-Hückel with a 5 nm cutoff radius. {ADD REF: Debye-Huckel} Simulations were performed in the NVT ensemble employing a Langevin thermostat. {ADD REF: Langevin thermostat} Each sequence was simulated at its melting temperature – AT-all: XX K, GC-end: XX K, GC-core: XX K, GC-mix: XX K – in order to maximize the number of spontaneous transitions between dissociated and hybridized states. Melting temperatures for each sequence were determined empirically by XXX. The Langevin equations of motion were integrated using

the scheme of Bussi and Parrinello {ADD REF: G. Bussi and M. Parrinello,Phys.Rev.E75, 056707 (2007).} 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. Initial system configurations were prepared How? How specify initial configurations and apply any energy minimization? Initial bead velocities assigned from a Maxwell-Boltzmann distribution at the system temperature. {ADD REF: Maxwell-Boltzmann} Each simulation was conducted for 26 μs and frame saved to disc every 100 ps, requiring $\sim$24 CPU-hours on 28$\times$Intel E5-2680v4 CPU cores. The first 1 μs of each run was discarded for equilibration providing us with 40$\times$25 μs = 1 ms of simulation data for each sequence, during which time we observed 55-100 hybridization/dehybridization events.

2.2 Markov state model construction

Markov state models (MSMs) are a powerful approach to infer long-time kinetic models from short molecular simulation trajectories[60] {ADD REF: (1) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. J. Am. Chem. Soc. 2018, 140, 2386?2396. (2) 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. (3) https://doi.org/10.33011/livecoms.1.1.5965 (4) 10.1016/j.sbi.2014.04.002} that we employ to construct high-resolution sequence-dependent kinetic models of DNA hybridization and dissociation. MSMs have recently been implemented to study mechanisms of DNA hybridization, but the slowest sequence-dependent kinetics were not the focus of these studies.[17,61] A recent application of MSMs to the long-time dynamics of short RNA oligonucleotides revealed stacking timescales to be highly sequence dependent. Pinamonti et al.[62] MSMs were constructed for each of the four DNA sequences from the 40$\times$25 μs simulation trajectories following the six-step protocol detailed in Ref.[63]: (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. {ADD REF: }

Featurization. Trajectories comprising the Cartesian coordinates of the DNA strands as a function of time were featurized using the MDtraj Python libraries[64] to represent the system in a manner that exposes the essential system dynamics but eliminates trivial translation and rotational invariances. We adopt pairwise distances $d(i, j)$ between the centers of mass of the 10 bases / nucleotides / what? as a natural rototranslationally invariant featurization that represents each system configuration as the $\binom{20}{2} = 190$-element vector of pairwise distances between all distinguishable bases within the complementary sequence pair. This featurization naturally accounts for intramolecular configurations via the pairwise distances between bases within the same strand and intermolecular configurations via the pairwise distances between bases on different strands. 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".[60] The 190-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, $(d(i_1, j_2) = d(i_2, j_1)) \leftarrow 0.5\,(d(i_1, j_2) + d(i_2, j_1))$ [[Right?]], where i_1 denotes the i^{th} bases / nucleotides / what? on strand 1 and j_2 the j^{th} bases / nucleotides / what? on strand 2.[60] Finally, we elected to work with 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.[65] {ADD REF: https://doi.org/10.1021/acs.jctc.5b00553. (38) Scherer,M.K.;Husic,B.E.;Hoffmann,M.;Paul,F.;Wu,H.; Noe? F. Variational Selection of Fea-

tures for Molecular Kinetics. J. Chem. Phys. 2019, 150, No. 194108. (39) Wu, H.; Noe? F. Variational Approach for Learning Markov Processes from Time Series Data. 2017} Performing VAMP-2 scoring at a lag time of $\tau =$ XX ps and retaining the top XX modes, reveals that the reciprocal permutationally-symmetrized pairwise distances can carry up to XX% more kinetic variance than the non-reciprocal distances, suggesting that the higher resolution offered at close intermolecular distances can indeed boost the dynamical representational power of the model.

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. {ADD REF: [27] Pérez-Hernández G, Paul F, Giorgino T, Fabritiis GD, Noé F. Identification of slow molecular order parameters for Mar- kov model construction. J Chem Phys. 2013; 139(1):015102. https://doi.org/10.1063/1.4811489. [28] SchwantesCR,PandeVS.ImprovementsinMarkovStateModel Construction Reveal Many Non-Native Interactions in the Fold- ing of NTL9. J Chem Theory Comput. 2013; 9(4):2000?2009. https://doi.org/10.1021/ct300878a. [29] Noé F, Clementi C. Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation. J Chem Theory Comput. 2015; 11(10):5002?5011. https://doi.org/10.1021/acs.jctc.5b00553.} In this work, we instead employ state-free reversible VAMPnets (SRVs) that can be conceived of as a nonlinear version of tICA.[66] 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 in applications to WW-domain and Trp-cage mini-proteins.[63,66] Importantly, the 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 0.01, and a batch size of 50,000. We adopted a lag time of $\tau = 1.2$ ns as appropriate short to resolve the dynamical details of the hybridization/dehybridization dynamics.[67] As observed by Husic and Pande, it is conceptually incorrect to treat the lag time as a model hyperparameter to be self-consistently optimized via the VAMP-2 score, but rather as a physically-motivated choice designed to expose the dynamical motions relevant at a particular time and length scale of interest. {ADD REF: https://doi.org/10.1063/1.5002086} As we shall show, this choice of lag time leads to converged and Markovian macrostate MSMs. We guarded against overfitting using cross-validation in which we divided the trajectory data into XX contiguous segments of XX ns that were then partitioned into an 80:20 training:validation split. We observed plateau of the validation loss and no evidence of overfitting after 20 epochs of training for each system that required approximately 22 GPU-minutes on a single NVIDIA XXX GPU card. A VAMP-2 scoring of the cumulative kinetic variance explained as a function of number of SRV collective modes retained also showed no evidence of overfitting – as would be evinced by separation of the training and validation VAMP-2 scores[63] – and exhibited a knee for each of the four DNA sequences after the fifth slow mode (Fig. S1). This motivated the construction of low-dimensional embeddings of the trajectories for each of the four DNA sequences into a 5D space of leading SRV modes.

Microstate clustering. The 5D SRV projections of the XX frames recorded over the course of the 1 ms MD 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 highly insensitive to the choice of the number of microstates over the range XX-YY, motivating our selection of 200 microstate clusters for each system. [[Accurate statement?]]

Macrostate clustering. The 200 microstates comprising each system were finally 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 cor-

responding eigenvalues λ_i and associated implied time scales $t_i = -\tau / \ln |\lambda i|$. {ADD REF: https://doi.org/10.33011/livecoms.1.1.5965} The implied timescale plots for the four DNA sequences are presented in Fig. S2. We observe convergence of the implied time scales very rapidly with lag time τ for all systems, motivating the construction of high resolution macrostate MSMs at a lag time $\tau = 1.2$ ns. At 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-mid 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. {ADD REF: (43) 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. (44) Deuflhard, P.; Weber, M. Robust Perron Cluster Analysis in Conformation Dynamics. Linear Algebra Appl. 2005, 398, 161?184. (45) Kube, S.; Weber, M. A Coarse Graining Method for the Identification of Transition Rates Between Molecular Conformations. J. Chem. Phys. 2007, 126, No. 024103.} 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-mid systems, respectively, by projecting the MD 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 {ADD REF: https://doi.org/http://dx.doi.org/10.1063/1.3565032., https://doi.org/10.1073/pnas.0905466106., https://doi.org/10.33011/livecoms.1.1.5965}. 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 τshould be equivalent to a single application of the transition matrix constructed at a lag time $k\tau$. We present in Fig. S3-6 the CK tests for each DNA sequence to demonstrate that the $\tau = 1.2$ ns models performs excellently in predicting transition probabilities out to $k\tau =$ XX ns, validating the

Markovain nature and kinetic validity of the four kinetic models.

2.4 Experimental T-jump IR spectroscopy

TBA

3 Results

3.1 Sequence-dependent coarse-grained kinetics agree with T-jump IR data

As an initial validation of 3SPN2 kinetics, we compared simulated relaxation measurements with temperature-jump (T-jump) IR experiments. Coarse-grained timescales are inherently difficult to match against experiments due to smoothing of the free energy landscape and variable acceleration across different degrees of freedom. As such, we evaluated multiple temperatures for each sequence and inferred acceleration factors from temperature-dependent trends. Experimental temperature-dependent relaxation timescales were measured for each sequence using T-jump IR spectroscopy as described previously.[28] We recorded "fast" and "slow" amplitude-weighted responses – previously attributed to terminal base pair fraying and duplex dissociation, respectively – for each temperature and sequence. To compare results with 3SPN2, we ran 120 shorter (1 us) simulations initialized in the hybridized state at final T-jump temperatures for each sequence. We calculated relaxation times for a slow dissociation response and fast fraying response at each temperature and compared these with experimental temperature-dependent relaxation fits.

We measured the slow dissociation response by fitting the distribution of times at which the core base pairs separate beyond a 1.3 nm cutoff. We found the inverse of these relaxation times – the effective dissociation rate – to increase exponentially with temperature, which is expected given the large enthalpic barrier of dissociation.[20,37,39] We noted an aver-

age temperature shift of about 4°C between experiment and simulation fits. While 3SPN2 captures melting temperature relatively well, some deviation is expected given varied ion conditions and other coarse-grained effects. We also noted an acceleration of about one order of magnitude compared to experiment, although this factor is sensitive to the exact definition of melting temperature. After accounting for the melting temperature shift and acceleration factor, we saw strong agreement between the experimental data and simulated relaxation fits (Figure 1)). At lower temperatures, T-jump slow responses likely contain a mixture of dissociation and hybridization dynamics, therefore rates do not drop off to the same extent as our analysis, which considers dissociation alone. GC-core showed the largest deviation from experiment, indicating that increased temperature triggers dissociation at a disproportionately higher rate in simulation compared to experiment.

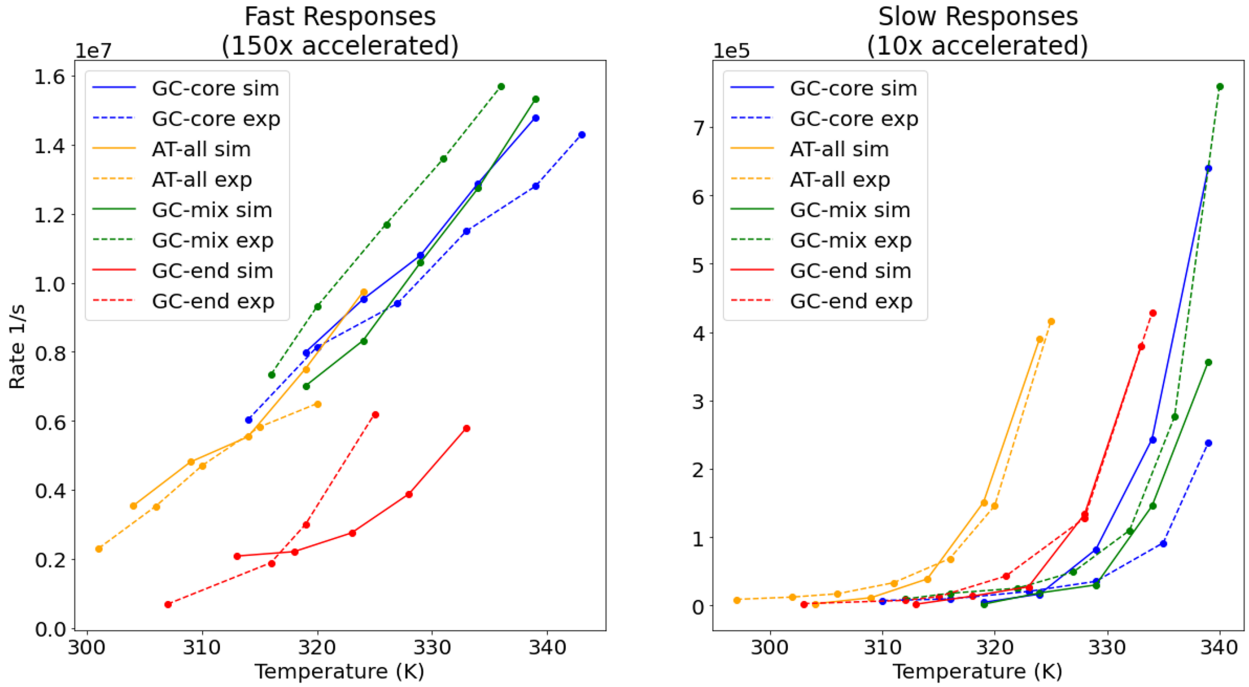


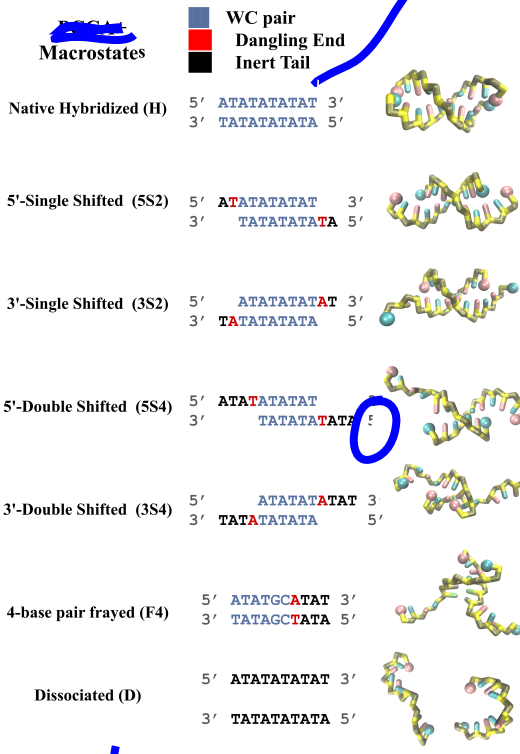
Figure 1: Fitting temperature-dependent trends to the "slow" dissociation mode and "fast" fraying mode detected in experiments. The effective simulation temperature is shifted by 4 degrees celcius to account for systematic differences in simulation and experimental melting temperature. Different scaling factors are applied to fast and response to account for variable coarse-grained acceleration along different degrees of freedom.

The fast response – which Sanstead et. al attributed to terminal base pair fraying signatures – was more challenging to compare against simulation observables. Numerous experimental and computational studies have shown that DNA and RNA fraying is a complex dynamical process with timescales that span 5 ps to several microseconds.[32–34,68] All-atom simulations suggest that frayed ends can assume misaligned WC bonds, base-sugar hydrogen bonds, and terminal stacked conformations.[31,69] Given that there is only one interaction site parameterized on each 3SPN2 base, we would not expect to resolve this diverse collection of states and dynamics. Instead, we measured the fast fraying response by counting frames until duplex terminal ends splits beyond a cutoff. This approach assumes that fraying on the permutable top and bottom of the duplex are independent from each other, and that a base pair distance is a reliable approximation for the ensemble spectroscopy signal. This is a reasonable assumption given that the amplitude-weighted timescales should consist largely of terminal fraying events. Again we fit relaxation curves to the ensemble of fraying timescales in order to extract rate approximations for each sequence at a series of temperatures.

For A:T terminal sequences, the simulated fast response appears linear with temperature, indicating a barrierless and diffusion-driven process. In contrast, the GC-end responses are distinctly slower and increase exponentially with temperature, likely due to a greater enthalpic barrier associated with G:C fraying. We observe similar trends in the experimental data, although comparisons at the highest temperatures were limited due to mixing between the fast and slow responses (Figure 1). We found the optimal acceleration factor to be dependent on the choice of the cutoff parameter, but we can approximate that 3SPN2 fraying dynamics are accelerated by about 100-200x relative to experiment. It is not surprising that we see different rates of acceleration for the dissociation and fraying processes given that coarse-grained effects can vary across different degrees of freedom. In particular, the simplified treatment of the fraying process may smoothen the free energy landscape and speed up dynamics relative to a more global processes like dissociation. Although one should rely on higher resolution models to study in depth mechanisms of fraying, our results indicate that

terminal fraying is a reasonable assignment for the spectroscopic fast response in Sanstead et al. and that 3SPN2 can recapitulate some sequence-dependent fraying effects.

3.2 SRV-MSMs for each sequence



Stationary Probabilities

	AT-all	GC-end	GC-core	GC-mix
Native Hybridized (H)	$(4.152 \pm 0.003) \times 10^{-1}$	$(5.720 \pm 0.006) \times 10^{-1}$	$(4.535 \pm 0.004) \times 10^{-1}$	$(4.572 \pm 0.005) \times 10^{-1}$
5'-Single Shifted (5S2)	$(7.88 \pm 0.02) \times 10^{-2}$	$(1.18 \pm 0.06) \times 10^{-3}$	--	--
3'-Single Shifted (3S2)	$(1.57 \pm 0.03) \times 10^{-2}$	$(4.0 \pm 0.3) \times 10^{-4}$	--	--
5'-Double Shifted (5S4)	$(3.62 \pm 0.07) \times 10^{-3}$	$(5.4 \pm 0.5) \times 10^{-5}$	--	--
3'-Double Shifted (3S4)	$(1.20 \pm 0.07) \times 10^{-3}$	--	--	--
4-base pair frayed (F4)	--	--	$(2.77 \pm 0.02) \times 10^{-3}$	--
Dissociated (D)	$(4.855 \pm 0.008) \times 10^{-1}$	$(4.208 \pm 0.007) \times 10^{-1}$	$(5.445 \pm 0.006) \times 10^{-1}$	$(5.428 \pm 0.008) \times 10^{-1}$

Figure 2: Nearest neighbor representations, molecular renderings, and sequence-dependent probabilities for PCCA+ macrostates. Shifted state abbreviations indicate the direction of the shifted overhang (5′ vs. 3′) and the number of shifted base pairs relative the native state (2 vs. 4).

We followed the SRV-MSM analysis pipeline described in the Methods section to generate sequence-dependent kinetic models for each oligomer. We identified seven kinetically relevant states that were captured within the resolution of the model. Figure 2 shows cartoon and molecular renderings for each of these states as well as their sequence-specific stationary probabilities. By row, these include the fully hybridized state (H) in which all native base

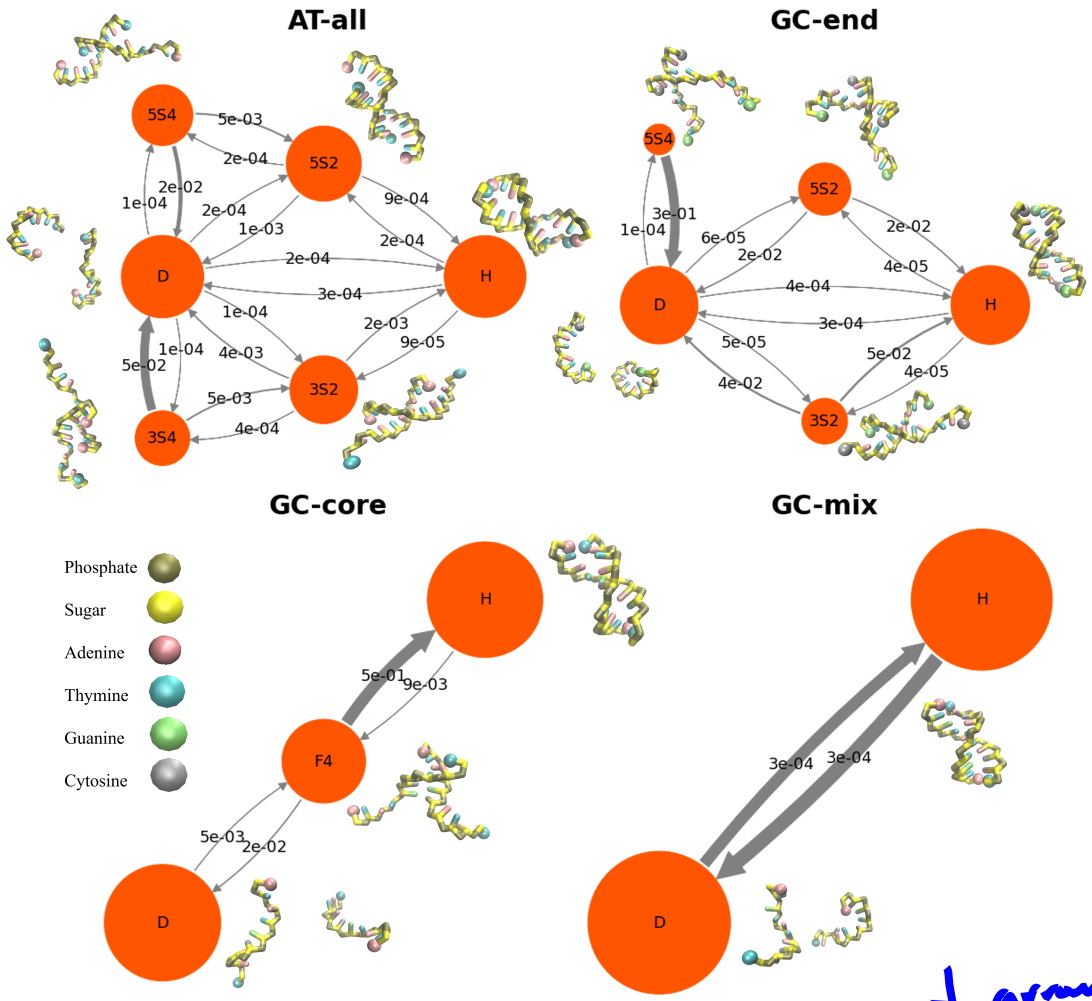


Figure 3: Flux diagram between PCCA+ states are shown for each sequences, accompanied by representative structures are for each macrostate. Arrows indicate the probability of transitioning between states within the lag time. Circle areas are proportional to the log of stationary probabilities to enhance visualization of less stable states

pairings are intact, four "shifted" states (5S2, 3S2, 5S4, 3S4) where complementary base pairs form out-of-register, a frayed state (F4) – unique to GC-core – in which the four terminal A:T base pairs are unbound, and the fully dissociated state (D). Figure 3 shows how these states fit into sequence-specific low-dimensional representations and kinetic models. Arrows between states show the probability of a transition to another state within the MSM lag time. We now proceed to analyze the sequence-dependent thermodynamics and kinetic features of

each MSM.

3.3 Inert tails produce deviations from nearest neighbor thermodynamic models

Macrostates	AT-all		GC-end	
	~~PCCA+ dG~~ (kJ/mol)	~~NN dG~~ (kJ/mol)	PCCA+ dG (kJ/mol)	NN dG (kJ/mol)
5S1	4.26	0.08	15.5	7.48
3S1	8.42	2.23	18.3	5.67
5S2	12.2	6.49	23.7	12.5
3S2	15.0	8.61	--	10.6

Figure 4: Comparisons between free energies based on simulation macrostate populations and nearest neighbor calculations. All free energies are normalized such that that the native hybridized state is set to zero. Calculations included dangling end contributions but do not take inert tails effects.

Nearest neighbor (NN) models have been experimentally optimized to account for stacking contributions of native pairs, internal mismatches, and dangling ends.[9] The effects of inert tails – free bases that extend beyond the dangling end – have been shown to destabilize the duplex but are not included in NN models.[70] It is informative to compare our MSM predictions against NN calculations to understand how inert tail and kinetic effects may cause deviation from these predictions. In particular, we examine the out-of-register AT-all and GC-end states which are amenable to NN comparisons (Figure 4). For AT-all, NN calculations predict that conformations in the 5′ shifted states (5S2 and 5S4) are more energetically favorable than those in the 3′ shifted state (3S2 and 3S4). We find qualitative agreement to our macrostate free energies, where inert tail effects likely contribute a 4-6.5 kJ/mol increase in free energy compared to dangling end predictions alone. For GC-end, we consider C:T and G:A mismatches as non-interacting dangling ends such that each shifted conformation has four total dangling ends. Based on this treatment, NN calculations yield

higher overall free energies due to fewer native base pair contacts. Contrary to our simulation results, however, the NN model predicts the 3S2 state to be more stable than 5S2 state. Moreover, we do not observe a stable 3S4 cluster for GC-end. This indicates that 5′ vs. 3′ inert tail differences may outweigh NN stacking effects alone. These inert tail effects show qualitative agreement with previous experimental studies, which attribute the differential behavior to some combination of 5′ tails preferentially stacking on the core duplex and 3′ tails perturbing the duplex structure.[70–72].

3.4 Out-of-register dynamics facilitate hybridization and dissociation

In addition to state probabilities and free energy approximations, the coarse-grained MSM yields valuable discrete kinetic information in the form of transition probabilities between states. For AT-all, we observe approximately equal probability of transitioning from D to any other state. Once a transition has been made, however, the 5′ vs. 3′ overhang and degree of shifting play an important role in determining whether the duplex will continue to shift out-of-register or re-dissociate. Transition probabilities are higher when moving towards a more aligned state than towards a more shifted state – e.g. 5S2 → H is more likely than 5S2 → 5S4 – suggesting that more shifted states play a greater role in facilitating the hybridization process than dissociation. Furthermore, we see equal or higher transition probability from shifted states to the dissociated state (5S2 → D) than to more aligned states (5S2 → H), indicating that the shifting-hybridization process is frequently disrupted by complete dissociation. In particular, we observe that the transition probability 3S4 → D is 10x higher than the 3S4 → 3S2 path towards native hybridization.

For GC-end we observe no significant flux between the 5S4 state and the structurally similar 5S2 state, indicating that the former acts more as a kinetic trap than a pathway to H. The 5S2 and 3S2 states still readily convert with H, however there is an order of magnitude lower probability of reaching these states from D when compared to AT-all. This

suggests that inert tails inhibit out-of-register binding – accounting in part for higher free energies discussed above – but may not substantially disrupt out-of-register hybridization mechanisms along 5S2 → H and 3S2 → H once shifted states have formed. This phenomenon will be explored more in depth below.

Taken together, our results show that AT-all hybridization and dissociation kinetics are substantially modulated by out-of-register pathways. A majority of initial hybridization contacts occur out-of-register, and, based on transition probabilities, 35% of D → H pathways and 37% of H → D pathways pass through some out-of-register state. The 3S4 and 5S4 states are more prone to re-dissociation and thus contribute to about 5% of these pathways. While GC-end strands can access out-of-register states, only 11% of native hybridization events pass through any out-of-register states. These observations show that small disruptions to repetitive tracts shift hybridization pathways towards more two-state behavior. Indeed, there remain some repetitive tracts in GC-mix and GC-core, however we observe no out-of-register pathways for these sequences. Furthermore, our results indicate that internal displacement mechanisms reported in previous simulation studies are capable of disrupting non-native pairing and correcting base pair mismatches without fully dissociating strands.[30,51,54] These processes can occur even when terminal mismatches are present in the non-native structure.

Experimentally, out-of-register states are suspected to occur but are difficult to capture due to short lifetimes and similarity in response to other processes such as fraying. To explicitly minimize out-of-register base pairing, similar AT repeat motif sequences have been padded by GCG clamps during experimental analysis.[27] Recent all-atom results identified analogous out-of-register states as "deep kinetic traps" along the hybridization pathway for repetitive dGCGCGC hexamers.[17] Contrary to our kinetic results, however, "slithering" mechanisms along out-of-register tracts were not observed to be a dominant pathway, especially when compared to the high rates of slithering exhibited by the homogenous dGGGGGG strand. It is unclear if these differences were a consequence of varied simulation conditions or if GC repeat motifs are less susceptible to direct out-of-register transitions compared to AT

motifs. This is possible given that stronger hydrogen binding in GC motifs[73,74] may prevent fluctuation-driven rearrangement, however further computational and experimental studies are required to verify these differences.

3.5 Out-of-register dynamics may stretch T-jump IR responses

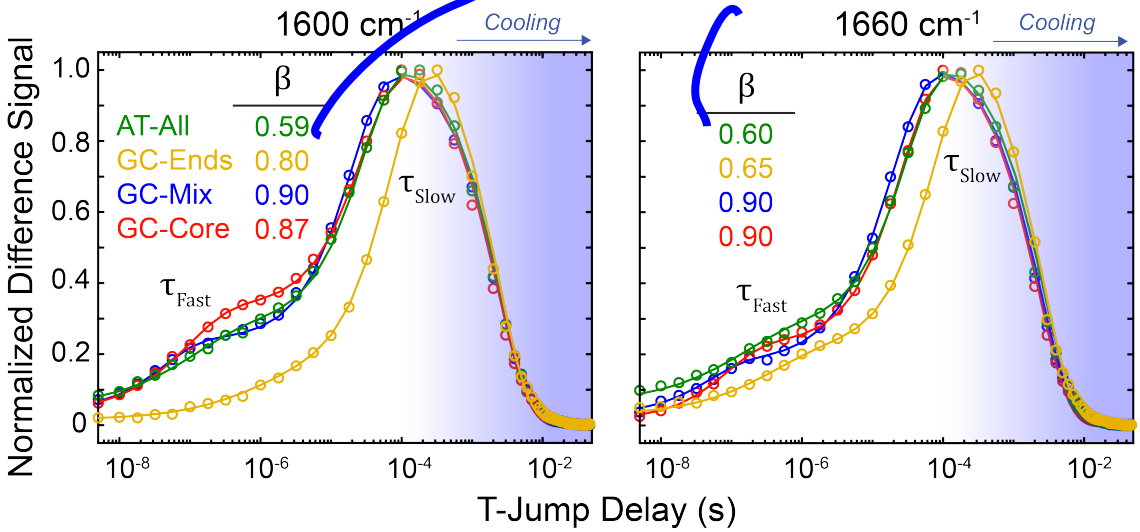


Figure 5: Normalized time traces from T-jump HDVE data probing the change in in-plane nucleobase vibrations. The signal at 1600 cm-1 corresponds to A and T nucleobases while the signal at 1660 cm-1 contains G, C, and T contributions. Each T-jump was performed with a temperature change of 15 ÂřC to a final temperature near the Tm of each respective sequence. Each time trace was fit to the sum of a stretched exponential and two exponentials (solid lines). The stretched exponential describes the component from 10 ns to 1 Îijs (tau_{Fast}), and the two exponentials describe the signal rise from 1 - 300 Îijs (tau_{Slow}) and thermal relaxation of the sample back to the initial temperature. The stretch factor (Îš) from each fit is listed.

In addition to identifying relaxation timescales, the T-jump IR response profile can inform the distribution of dynamics contributing to the response. In particular, the fast response of AT-all and GC-end exhibits more stretched exponential character than for GC-core and GC-mix (Figure 5). Furthermore, the GC-end response contains both G:C and A:T vibrational contributions, suggesting that native terminal base fraying may not be the unique source of the response. On the other hand, GC-mix and GC-core are characterized by a

distinct fast response comprised only of A:T features that points to A:T fraying as the predominant dynamic source. We predict that the mid-IR responses of fraying and shifting are not clearly distinguishable for GC-ends and AT-all, therefore we cannot determine the stretched responses originate from fraying dynamics in out-of-register configurations, shifting mechanisms during the dissociation process, or some combination of the two. Although reproducing these stretched signals is not within the scope of this analysis, we see qualitative agreement between more complex transition network (AT-all > GC-end > GC-core, GC-mix) and increased stretching in the fast response.

3.6 Centralizing GC placement induces long-lived, kinetically relevant fraying

The GC-core sequence represents a departure from the dominant shifting dynamics observed for AT-all and GC-end. Instead, the dynamical analysis describes a hybridization/dissociation pathway facilitated by a unique, highly frayed state (F4). Previous studies suggest that once key native contacts are made, the zippering mechanism ensures that the helix will quickly form outward.[16,51] Our results indicate, however, that the relative instability of A:T bonds compared the GC core can interrupt this process and form a longer lived metastable state. This occurs during the dissociation process as well, where one half of the A:T base contacts are entirely broken for a substantial period of time before the full dissociation event occurs. We observe these events to occur with equal probability on either permutable end of the helix.

The F4 state is more accessible from an already bound helix, and once oligos are in F4, they are over 20x more likely to return to H than to D. Thus once a D $\rightarrow$ F4 transition has occurred, a F4 $\rightarrow$ H transition will likely proceed. On the other hand, H $\rightarrow$ F4 events are more frequent but unlikely to initiate complete dissociation. We previously noted that frayed dynamics are substantially accelerated in the 3SPN2 model, which may lead to the F4 state being sampled more frequently than in experiment. Because GC-core stability is

most sensitive to fraying, this effect could cause GC-core to dissociate more frequently and lead to the deviation in slow response we note in Figure 1.

Lattice model studies have shown that frayed intermediates make substantial contributions to the GC-core conformational ensemble.[25,45,75] Araque et al. defined a similar 8-mer sequence (dATGCGCAT) as non-two-state, where a stable, symmetrically A:T frayed state is a crucial part of the duplex transition path.[25] When examining all four sequences using T-jump IR and 2D IR spectroscopy, Sanstead et al. found that GC-core had the highest deviation from two-state behavior during dissociation (when excluding out-of-register contributions).[19] As T-jump IR data only showed a loss of A:T contacts during the fast response, this intermediate state was defined by a high degree of fraying about the central core. While 1-2 base pair fraying was commonly observed for GC-mix and AT-all as well, their lattice model predicted GC-core to have substantially more frayed base pairs.[45] Variable 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 interconversion between all accessible frayed states.[28] We see the same rapid fraying in simulation data – which is too fast to be attributed to a converged SRV mode – however our SRV-MSM results show this inter-conversion first relies on the slower formation of the the A:T bond nearest to the GC center.

3.7 Heterogeneous base pair distribution promotes nucleation-zippering hybridization and dissociation

Given the lack of a repetitive AT interior (as in AT-all and GC-end) or exterior (as in GC-core), we expect more canonical dynamics from GC-mix. Although GC-mix dynamics are most similar to those of GC-core, we did not observe a converged slow mode corresponding to multi-base fraying behavior for GC-mix. Instead, we observed two modes converge, corresponding to the association/dissociation dynamic and diffusive behavior while strands are dissociated, respectively. Given that this second mode did not inform the hybridization

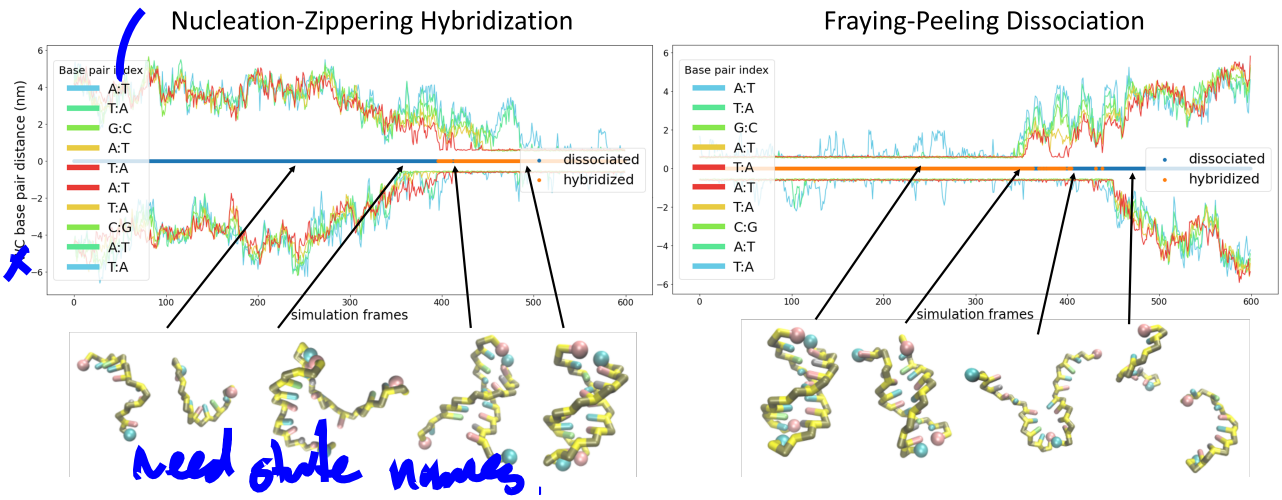


Figure 6: WC base pair distance and molecular renderings along two representative GC-mix hybridization and dissociation events. Permutable distances are reflected across the x-axis and show that fraying in the hybridized states tends to occur independently on either end of the duplex until full dissociation occurs. The SRV-MSM state is shown along the x-axis, indicating the point at which a transition has been determined. Fluctuation between states is common during the transition.

process, we designated this transitions as effectively two-state within the resolution of our model. We did, however, observe substantial fraying of the two AT termini in the simulation data. Although these frayed states may be too short-lived to resolve a distinct slow mode, this behavior shows qualitative agreement with experimental analysis of this sequence which attributed fraying prior to dissociation as a deviation from all-or-nothing behavior.[19] While AT-termini fraying is surely a prerequisite to dissociation, we find these states to be so common and transient that very few progress to full dissociation compared to the F4 state. Furthermore, one or two-base-pair fraying does not fundamentally disrupt the helix in such a way that its re-formation is kinetically inhibited by the intermediate structures we present for GC-core.

To supplement GC-mix analysis, we looked at qualitative trends in our trajectory data near MSM-identified hybridization and dissociation events. We present two representative events in Figure 6 along with all native base pair distances over time. During hybridization, we observed the some key base pair contacts before the full duplex formed. Specifically, first

contacts tended to involve one of the G:C bonds and at least one neighboring A:T. This behavior is indicative of a nucleation-zippering mechanism as has been reported in previous studies.[16,20,36,49] For dissociation events, we noted two base pair fraying on one or both sides of the duplex followed by more rapid dissociation of the central base pairs. Qualitatively, we observed some short-lived states composed of 2-4 base pair contacts immediately before full dissociation occurs. However, in contrast to the F4 state we observe in GC-core, these conformations do not form a distinct free-energy minima in SRV space, nor do they tend to reform intact duplexes. As a whole, these dynamics are similar to previously reported "fraying-peeling" mechanism.[31,52,53] We observed similar fast dynamics and transition states in the other three sequences, however they are more difficult to discern as they occur in concert with the longer lasting metastable states discussed above.

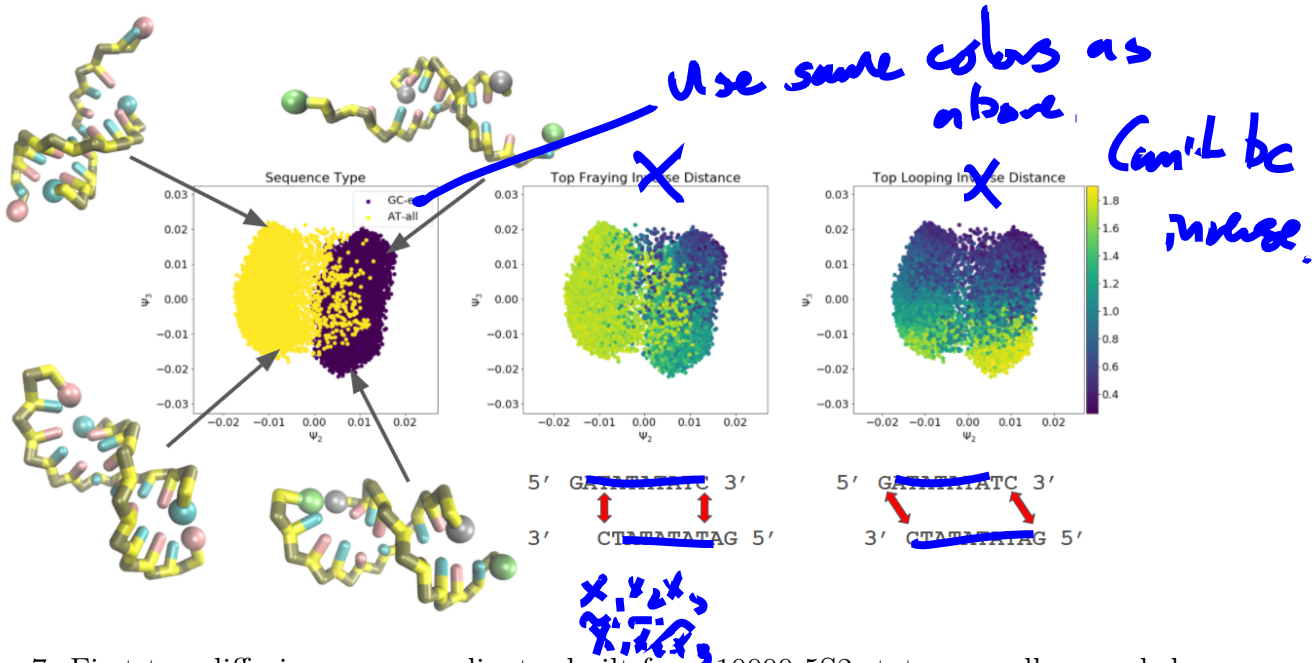


Figure 7: First two diffusion map coordinates built from 10000 5S2 states, equally sampled from AT-all and GC-end. Color maps show inverse distances (nm^{-1}) between out-of-register ends and complementary ends. The lower right region of the diffusion map space shows a high concentration of "shifted loop" base pairs that may facilitate the out-of-register GC-end transition.

3.8 5S2 state distributions reveal differences between AT-all and GC-end kinetics

Although our SRV-MSMs provide useful insight into slow processes, we were also interested in the ensemble of configurations within PCCA+ macrostates and during transition periods between states. These configurations interchange significantly faster than the SRV-MSM lag time, but we can use a diffusion map approach to glean a structural details of the configuration in these regions.[76,77] While examining molecular renderings, we noticed that a significant proportion of GC-end 5S2 configurations retained one native G:C bond, even when all available A:T bonds were formed out-of-register. We would not necessarily expect duplexes to sacrifice helical conformational entropy in order to facilitate termini bonding. To compare how these state populations differ between GC-end and AT-all, we employed diffusion maps built on an equal sampling of 5000 conformations from the 5S2 state of both sequences. We used all 100 intermolecular distances (as opposed to the 55 permutation free coordinates used to construct SRV-MSM) as our distance metric, making it easier to discern structures that form on either permutable end of the shifted conformation. This created degenerate 2nd and 3rd diffusion modes, with nearly equal eigenvalues, differentiating effects at the identical "top" and "bottom" of the strands. In Figure 7 we present the first two non-trivial diffusion map eigenfunctions and show representations of the degenerate third coordinate in the SI. Diffusion maps built from samples of the 3S2 states are also shown in the SI.

The first diffusion mode clearly delineates between the GC-end and AT-all shifted conformations and correlates highly with the average distance between the $3'$ end and its shifted complementary pair. As expected, this shows that mismatched C:T pairs are never bound whereas the AT-all pairs are mostly bound with occasional fraying indicated by small AT-all overlap in the GC-end region. The second diffusion mode, which correlates highly with the average distance between $3'$ and $5'$ ends, highlights structural variances within the two sequences distributions. We notice that about 40% of GC-end configurations fall below

$\Psi_3 = -0.01$ compared to only 10% of AT-all configuration. We find that these GC-end populations share stabilizing termini contacts despite all A:T contacts being shifting out-of-register. These "shifted-loop" bonds appear uniquely stable for GC-end conformations, and suggest a transition state through which hybridization is facilitated. This may account for the GC-end 5S2 $\rightarrow$ H transition being more favorable than 5S2 $\rightarrow$ D despite the state free energy being substantially higher than the equivalent AT-all pathways.

3.9 Sequence-dependent fraying determines direct hybridization pathways

To compare fast hybridization dynamics across all sequences, we trimmed simulations centered on direct D$\rightarrow$H (or D$\rightarrow$F4$\rightarrow$H for GC-core) events and excluded transitions that passed through long-lived out-of-register states. To achieve higher resolutions, we reran selected simulation with a 10x higher save rate and analyzed a 1000 frame region (10 ns simulation time) for each event. We built diffusion maps on a set of 48 such "trimmed" events evenly sampled across sequences. We found a distinct spectral gap after the first non-trivial diffusion map eigenvalue, indicating that the leading mode Ψ_2 contained most of the structural variance. Indeed, Ψ_2 had a 0.89 Pearson correlation with simulation time and a -0.94 correlation with inter-strand core distance with minimal variation by sequence. This shows that Ψ_2 is strongly correlated with hybridization progress for all sequences and represents a consistent collective variable for comparison between sequences.

Once we determined a promising hybridization coordinate, we compared probability distributions for each sequence along Ψ_2 (Figure 8). The region where $\Psi_2 < 0$ is noisy and mainly reflects the distance between approaching strands, but we noted statistically meaningful differences between sequence populations once first Watson-Crick contacts are made. In the $0.001 < \Psi_2 < 0.0035$ region, heatmaps show that contacts tend to form near the center of the duplex, and frayed structures progressively zipper closed as Ψ_2 increases. As expected, GC-core shows a distinct peak in the frayed region, corresponding to the F4 macrostate iden-

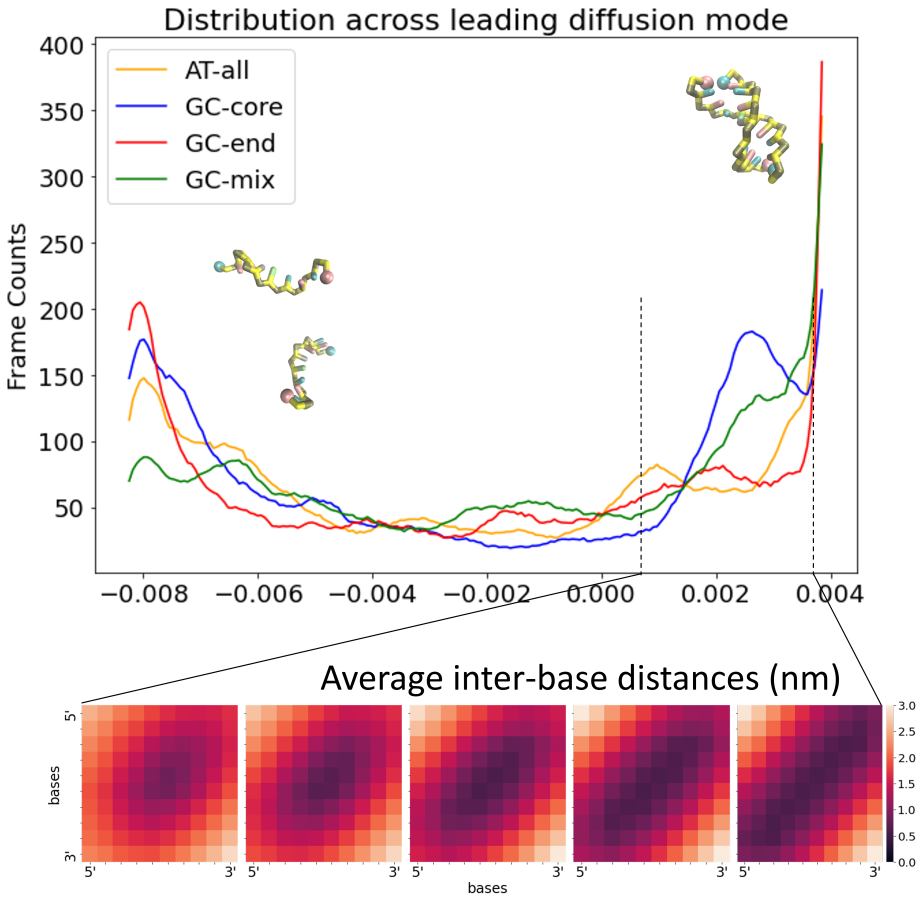


Figure 8: Sequence-dependent population distributions for trimmed hybridization events. Ψ_2 represents the maximum structural variance across all sequence configurations and correlates strongly with hybridization progression. Populations are discretized in 200 bins and smoothed by averaging with the nearest 5 bins. Heat maps show the average intermolecular base distances in nm at increasing values of Ψ_2, where the main diagonal shows native base pair distances.

tified by SRV-MSMs. GC-mix has a notably higher population than AT-all and GC-end but displays a more gradual progression toward the native hybridized state. This is consistent with the downhill nucleation-zippering mechanism described above. Although differences between AT-all and GC-end are more subtle, overall deviations from two-state behavior follow predictions by Sanstead et. al.[19] where GC-core > GC-mix > AT-all > GC-end. This qualitative agreement shows that MSM information can be adapted to compare sequences along common pathways and elucidate dynamics occurring faster than the lag time. Furthermore, we are able to distinguish out-of-register deviations from fraying deviations and

describe relevant experimental comparisons for each.

3.10 Limitations

Although we were able to obtain improved resolutions on several relevant dynamics by using a shared lag time across sequences, we found it difficult to converge faster processes such as duplex nucleation and zippering. These processes are crucial to duplex formation, however they do not appear to be kinetically metastable or slow relative to other timescales of interest. Furthermore, they can initiate at various points along the strand, which, under out present featurization method, may appear as a collection of modes instead of as one distinct process. We also found the need to strike a balance between adequate sampling of hybridization events and frame save rate in order to maintain tractable SRV-MSM calculations. Indeed, we collected over 10 GB of equilibrium trajectory data for each sequence and were working near memory limits when training SRVs and building SRV-MSMs.

In any high-level model there are inevitable simulation artefacts produced by coarse-grained approximations. For example, the treatment of non-interacting base pairs as an excluded volume potential alone may not be representative of dynamics produced from mismatched dangling ends. In general, coarse-grained models produce a smoother free-energy surface which can results in much faster motions between states. This is illustrated by substantial accelerations in temperature-dependent responses when compared to experiment. Furthermore, it may be easier to cross between states – e.g. a 5S2 $\rightarrow$ H transition – when the usually rough free energy path becomes more easily traversed.

4 Conclusions

We have demonstrated how G:C placement in an otherwise repetitive AT sequence has a profound impact on equilibrium hybridization dynamics. By supplementing coarse-grained MD with data-driven time-lagged analysis, we constructed high resolution SRV-MSMs ca-

pable of distinguishing sequence-dependent kinetic behavior. In particular, we found that AT-all and GC-end sequences both participate in some degree of out-of-register base pair shifting, although these dynamics are more relevant for AT-all. On the other hand, GC-core hybridization transits through, or is perhaps facilitated by, a frayed intermediate in which one half of A:T bonds are broken and the duplex is significantly disrupted. Our computational approach and results show strong comparisons to T-jump experiments at ns and us timescales. Going forward, we expect that this work will be extended to predict more general trends in sequence-dependent hybridization and motivate strategies for experimental comparisons. These insights should be leveraged for enhanced understanding of *in vivo* hybridization processes, optimized sequence design in application such as DNA-PAINT and barcoding, and novel developments in the growing field of dynamic DNA nanotechnology.

References

(1) Seeman, N. C.; Sleiman, H. F. DNA nanotechnology. *Nature Reviews Materials* **2017**, *3*.

(2) Adleman, L. Molecular Computation of Solutions to Combinatorial Problems. 1994.

(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. *Nature Protocols* **2017**, *12*, 1198–1228.

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

(7) Strauss, S.; Jungmann, R. Up to 100-fold speed-up and multiplexing in optimized DNA-PAINT. *Nature Methods* **2020**, *17*, 789–791.

(8) SantaLucia, J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences of the United States of America* **1998**, *95*, 1460–1465.

(9) Santalucia, J.; Hicks, D. T t dna s m. **2004**,

(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. *Journal of Physical Chemistry 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. *Journal of Physical Chemistry B* **2014**, *118*, 14326–14335.

(12) Mergny, J. L.; Sen, D. DNA quadruple helices in nanotechnology. *Chemical Reviews* **2019**, *119*, 6290–6325.

(13) Deluca, M.; Shi, Z.; Castro, C. E.; Arya, G. Dynamic DNA nanotechnology: Toward functional nanoscale devices. *Nanoscale Horizons* **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. **2020**, 251–267.

(16) Yin, Y.; Zhao, X. S. Kinetics and dynamics of DNA hybridization. *Accounts of Chemical Research* **2011**, *44*, 1172–1181.

(17) Xiao, S.; Sharpe, D. J.; Chakraborty, D.; Wales, D. J. Energy Landscapes and Hybridization Pathways for DNA Hexamer Duplexes. *Journal of Physical Chemistry Letters* **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. *Journal of Chemical Physics* **2014**, *141*.

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

(20) Pörschke, D.; Eigen, Cooperative nonenzymic base recognition III.âĂă Kinetics of the Helix-Coil Transition. **1971**, 361–381.

(21) Zhang, D. Y.; Winfree, E. Control of DNA strand displacement kinetics using toehold exchange. *Journal of the American Chemical Society* **2009**, *131*, 17303–17314.

(22) Shah, S.; Dubey, A. K.; Reif, J. Programming Temporal DNA Barcodes for Single-Molecule Fingerprinting. *Nano Letters* **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. **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. *Nature Chemistry* **2018**, *10*, 91–98.

(25) Araque, J. C.; Robert, M. A. Lattice model of oligonucleotide hybridization in solution. II. Specificity and cooperativity. *Journal of Chemical Physics* **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. *Journal of Physical Chemistry 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. *Physical Chemistry Chemical Physics* **2014**, *16*, 18827–18839.

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

(29) Phys, J. C.; Hinckley, D. M.; Lequieu, J. P.; Pablo, J. J. D. Coarse-grained modeling of DNA oligomer hybridization : Length , sequence , and salt effects. **2014**, *035102*.

(30) Maciejczyk, M.; Spasic, A.; Liwo, A.; Scheraga, H. A. DNA duplex formation with a coarse-grained model. *Journal of Chemical Theory and Computation* **2014**, *10*, 5020–5035.

(31) Zgarbová, M.; Otyepka, M.; Šponer, J.; Lankaš, F.; Jurečka, P. Base pair fraying in molecular dynamics simulations of DNA and RNA. *Journal of Chemical Theory and Computation* **2014**, *10*, 3177–3189.

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

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

(34) 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: "Fraying" at the end of the helix. *Journal of the American Chemical Society* **2006**, *128*, 6885–6892.

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

(36) Wetmur, J. G.; Davidson, N. Kinetics of renaturation of DNA. *Journal of Molecular Biology* **1968**, *31*, 349–370.

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

(38) Pörschke, D.; Uhlenbeck, O. C.; Martin, F. H. Thermodynamics and kinetics of the helixâĂŘcoil transition of oligomers containing GC base pairs. *Biopolymers* **1973**, *12*, 1313–1335.

(39) 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.

(40) 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. *Journal of the American Chemical Society* **2012**, *134*, 18952–18963.

(41) 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 Research* **2007**, *35*, 2875–2884.

(42) 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.

(43) Chen, X.; Zhou, Y.; Qu, P.; Xin, S. Z. Base-by-base dynamics in DNA hybridization probed by fluorescence correlation spectroscopy. *Journal of the American Chemical Society* **2008**, *130*, 16947–16952.

(44) Dupuis, N. F.; Holmstrom, E. D.; Nesbitt, D. J. Single-molecule kinetics reveal cation-promoted DNA duplex formation through ordering of single-stranded helices. *Biophysical Journal* **2013**, *105*, 756–766.

(45) Phys, J. C.; Sanstead, P. J.; Tokmakoff, A. A lattice model for the interpretation of oligonucleotide hybridization experiments A lattice model for the interpretation of oligonucleotide hybridization experiments. **2019**, *185104*.

(46) 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. *Journal of Chemical Physics* **2013**, *139*.

(47) Schmitt, T. J.; Rogers, J. B.; Knotts IV, T. A. Exploring the mechanisms of DNA hybridization on a surface. *Journal of Chemical Physics* **2013**, *138*.

(48) Sambriski, E. J.; Schwartz, D. C.; De Pablo, J. J. Uncovering pathways in DNA oligonucleotide hybridization via transition state analysis (Proceedings of the National Academy of Sciences of the United States of America (2009) 106, (18125-18130) DOI: 10.1073/pnas.0904721106). *Proceedings of the National Academy of Sciences of the United States of America* **2009**, *106*, 21007.

(49) Sambriski, E. J.; Ortiz, V.; De Pablo, J. J. Sequence effects in the melting and renaturation of short DNA oligonucleotides: Structure and mechanistic pathways. *Journal of Physics Condensed Matter* **2009**, *21*.

(50) 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.

(51) Romano, F.; Doye, J. P. K.; Ouldridge, T. E.; Petr, S.; Louis, A. A. DNA hybridization kinetics : zippering , internal displacement and sequence dependence. **2013**, *41*, 8886–8895.

(52) Wong, K. Y.; Pettitt, B. M. The pathway of oligomeric DNA melting investigated by molecular dynamics simulations. *Biophysical Journal* **2008**, *95*, 5618–5626.

(53) Perez, A.; Orozco, M. Real-time atomistic description of DNA unfolding. *Angewandte Chemie - International Edition* **2010**, *49*, 4805–4808.

(54) 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. *Journal of Physical Chemistry A* **2015**, *119*, 1823–1834.

(55) Dans, P. D.; Walther, J.; Gómez, H.; Orozco, M. Multiscale simulation of DNA. *Current Opinion in Structural Biology* **2016**, *37*, 29–45.

(56) 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. *Biophysical Journal* **2017**, *112*, 1302–1315.

(57) Lu, W.; Bueno, C.; Schafer, N. P.; Moller, J.; Jin, S.; Chen, X.; Chen, M.; Gu, X.; Pablo, J. J. D.; Peter, G. OpenAWSEM with Open3SPN2 : a fast , flexible , and accessible framework for large-scale coarse-grained biomolecular simulations Author summary. **2020**, 1–21.

(58) Lequieu, J.; Córdoba, A.; Schwartz, D. C.; De Pablo, J. J. Tension-dependent free energies of nucleosome unwrapping. *ACS Central Science* **2016**, *2*, 660–666.

(59) Schneider, T.; Stoll, E. Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Physical Review B* **1978**, *17*, 1302–1322.

(60) Sengupta, U.; Carballo-pacheco, M.; Strodel, B. Automated Markov state models for molecular dynamics simulations of aggregation and self-assembly. **2019**, *115101*, 2–5.

(61) Jin, R.; Maibaum, L. Mechanisms of DNA hybridization: Transition path analysis of a simulation-informed Markov model. *Journal of Chemical Physics* **2019**, *150*.

(62) 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. **2017**,

(63) 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. 1–13.

(64) 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. *Biophysical Journal* **2015**, *109*, 1528–1532.

(65) Noé, F.; Nüske, F. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Modeling and Simulation* **2013**, *11*, 635–655.

(66) Chen, W.; Sidky, H.; Ferguson, A. L. Nonlinear Discovery of Slow Molecular Modes using State-Free Reversible VAMPnets. 1–19.

(67) 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 Markov models of molecular kinetics : Generation and validation. **2018**, *174105*.

(68) Galindo-Murillo, R.; Roe, D. R.; Cheatham, T. E. Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochimica et Biophysica Acta - General Subjects* **2015**, *1850*, 1041–1058.

(69) Pinamonti, G.; Paul, F.; Rodriguez, A.; Bussi, G. The mechanism of RNA base fraying: molecular dynamics simulations analyzed with core-set Markov state models. *43*.

(70) Michele, L. D.; Mognetti, B. M.; Yanagishima, T.; Varilly, P.; Ru, Z.; Frenkel, D.; Eiser, E. E ff ect of Inert Tails on the Thermodynamics of DNA Hybridization. **2014**, 0–3.

(71) Doktycz, M. J.; Paner, T. M.; Amaratunga, M.; Benight, A. S. Thermodynamic stability of the 5âĂš danglingâĂŘended DNA hairpins formed from sequences

5âĂšâĂŘ(XY)2GGATAC(T)4GTATCCâĂŘ3âĂš, where X, Y = A,T,G,C. *Biopolymers* **1990**, *30*, 829–845.

(72) Dickman, R.; Manyanga, F.; Brewood, G. P.; Fish, D. J.; Fish, C. A.; Summers, C.; Horne, M. T.; Benight, A. S. Thermodynamic contributions of 5'- and 3'-single strand dangling-ends to the stability of short duplex DNAs. *Journal of Biophysical Chemistry* **2012**, *03*, 1–15.

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

(74) Zacharias, M. Base-Pairing and Base-Stacking Contributions to Double-Stranded DNA Formation. *The Journal of Physical Chemistry B* **2020**,

(75) Sanstead, P. J.; Ashwood, B.; Dai, Q.; He, C.; Tokmakoff, A. Oxidized Derivatives of 5-Methylcytosine Alter the Stability and Dehybridization Dynamics of Duplex DNA. *Journal of Physical Chemistry B* **2020**, *124*, 1160–1174.

(76) Coifman, R. R.; Lafon, S. Diffusion maps. *Applied and Computational Harmonic Analysis* **2006**, *21*, 5–30.

(77) Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America* **2010**, *107*, 13597–13602.