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

1. Better section titles
2. Tighten up hypotheses & results
3. Main text figures
   - pdf format
   - in main LaTeX directory
   - named Fig1.pdf, Fig2.pdf, ..

Paper Structure

Intro ✓

Methods ✓

Results
- MSM presentation
- series of hypothesis driven analyses
- experimental comparison

Conclusions
- what do we know now that we didn't before?

Abstract

Despite rapid advances in DNA nanotechnology and a robust understanding of the associated thermodynamics, the sequence-dependent mechanisms of DNA hybridization are not fully understood. In this work, we investigate these dynamics by performing equilibrium coarse-grained simulations of oligonucleotide sequences with varied G:C placement. We employ State-Free Reversible VAMPnets to directly learn the slowest dynamical modes of each sequence and optimize Markov State Model (MSM) construction. Furthermore, we perform elevated temperature simulations to recapitulate temperature-jump IR and FTIR data collected on the oligonucleotides. For repetitive sequences, we find a spectrum of slow dynamics associated with out-of-register base pairing and kinetically relevant transitions between these states. In contrast, G:C pairs near the center of the duplex induce more pronounced fraying dynamics through which hybridization and dissociation are facilitated. In both cases, these mechanisms deviate from an âĂIJall-or-nothingâĂİ hybridization model. Our computational predictions show agreement with experiments, and provide new fundamental understanding of the sequence-dependent kinetics and mechanisms of DNA hybridization.

https://www.overleaf.com/project/5e9e5110c524b8000192c548

1 Introduction

Over the last couple decades, DNA has proved 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]. In particular, DNA dynamical behavior has become increasingly important in developing these technologies and understanding fundamental biological processes such as transcription and gene regulation[5–7]. A thorough understanding of the hybridization process – the formation of a DNA duplex from two single strands – and its associated dynamics is integral to further advancements in these fields. The effect of sequence on hybridization thermodynamics has been rigorously explored, and nearest-neighbor calculations can 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]. Although many experimental and computational studies have investigated DNA dynamical phenomena from the picosecond to millisecond range, the sequence-dependent mechanisms of hybridization and dissociation dynamics are not fully understood[13–20]. Moreover, it is unclear the extent to which these processes evolve in an "all-or-nothing" fashion or if an ensemble metastable states facilitates the transition[21,22].

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 drive DNA out of equilibrium and monitor relaxation processes in one direction[18–20,23–27]. More recently, single molecule diffusion and tethered multifluorophore assays have facilitated equilibrium analysis, however these results can be hampered by slow data collection rates and fluorescent tags effects on strand dynamics, particularly for shorter oligomers[23,28–31]. Given that dynamic insights from these experiments are limited, several coarse-grained molecular dynamics (MD) models have been employed to gain further detail[32–35]. Although these models provide experimentally verified speed-ups compared to all-atom simulations, the long timescales on which DNA hybridization and dissociation events occur can still make these

processes difficult to sample via direct simulation techniques[36]. Instead, many previous studies of DNA hybridization have employed accelerated sampling methods such as metadynamics, umbrella sampling, transition path sampling, and forward flux sampling[32,37–39]. Simplified lattice models can expedite computation time and yield useful thermodynamics and transition state information, however some configurations may not be accessible and kinetics are less informative[21,40]. Other computational works use dramatically elevated temperature or denaturing solvent concentrations to induce one-way dissociation events[41,42]. Taken together, most experimental and computational work have studied certain aspects the overall dynamics process in one direction or in a biased fashion.

Recent studies have coupled experimental techniques with machine learning (ML) and MD simulations to investigate and predict sequence-dependent kinetics[29,43]. Where these studies focus on association and dissociation kinetics alone, we seek to broaden our analysis into higher order dynamical processes and metastable intermediates. For example, stable out-of-register or "shifted" base pairing in repetitive sequences have been documented in previous computational studies[14,21,34,36]. The extent to which these states are kinetically relevant is difficult to determine experimentally, and appears highly sequence-dependent. Frayed structures and dynamics have also been investigated in numerous computational and experimental studies[44–47]. Sanstead et al. highlighted the role of these dynamics during duplex dissociation, where the stability and relaxation timescale of frayed states was dictated by G:C base placement 10-mer oligonucleotides[16]. Together, these dynamics represent substantial deviations from all-or-nothing behavior that can be variably expressed via subtle sequence differences in short oligomers.

In this work, we study the same four sequences explored by Sanstead et al. in an effort to uncover sequence-dependent dynamics and their relation to metastable structures mentioned above[16]. We use the coarse-grained 3 Sites per Nucleotide (3spn2) model to simulate hybridization and dissociation behavior near each sequence's melting temperature[33]. We perform these analyses without biasing simulations or assuming that one processes is a

strictly reversible version of the another. Furthermore, we leverage the properties of Markov State Models (MSMs) – namely that conditional probability depends only on the current state of the system – to combine many independent and unbiased trajectories and develop an understanding of sequence-specific kinetics and thermodynamics[48]. MSMs have recently been implemented to study mechanisms and microstate distributions of DNA hybridization, but the slowest sequence-dependent kinetics were not the focus of these studies[14,49]. Pinamonti et al.[50] used MSMs to compare the slowest dynamics of short RNA nucleotides and found that stacking timescales are highly sequence dependent. We take a similar approach to study 10-mer DNA oligonucleotides and introduce State Free Reversible Vampnets (SRVs) to directly learn the slowest sequence-dependent dynamical modes[51]. Furthermore, we integrate SRVs into the MSM pipeline by generating an optimized low dimensional basis in which microstate clustering can be performed. We show that SRV coordinates can be useful for both directly interpreting dynamical trends and for improving overall SRV-MSM quality when compared to more conventional methods such as time structure independent components analysis (tICA).

We find that G:C base pair placement in decamer oligonucleotides has a substantial effect on dynamical behavior. By evaluating equilibrium trajectories we can study the relevance of metastable states during both the hybridization and dissociation process. Because SRVs generate an optimized low dimensional basis, we show that higher resolution MSMs (measured by a reduction in the required lag time) are accessible. Additionally, we compare slow dynamical modes and metastable states between sequence-specific SRV-MSMs. Within these metastable dynamical states, we leverage diffusion maps to analyze the diversity of structures whose inter-conversion rate are too fast to produce unique slow modes. Finally, we run higher temperatures simulations to investigate the temperature dependent nature of some experimentally relevant dynamical responses. Taken together, our analysis reflects similar results to previous computational and experimental DNA work, while elucidating new insights into sequence-dependent dynamics, metastable structures, and relative timescales.

2 Methods

2.1 Simulation set up

The 3 site per nucleotide (3spn2) coarse grained model was designed to accelerate DNA computation relative to all-atom models while maintaining experimental melting temperatures, stacking energies, and persistence lengths[33]. Interaction sites are located at phosphate, sugar, and base centers of mass; anisotropic potentials are designed to model non-bonded interactions such as intra-strand base-stacking, inter-strand cross-stacking, and base pairing. The model has been validated against experimentally determined structural properties and hybridization rates, although no dynamic information was used to parameterize the model other than Langevin friction coefficients[15,33]. 3spn2 has been widely-adopted to study numerous phenomenon including DNA packing in viral capsids, protein-DNA binding, and nucleosome unwrapping.[52–54].

We initialized four sequences previously the subject of ultra-fast T-jump experimentation – 5'ATATATATAT3' (AT-all), 5'GATATATATC3' (GC-end), 5'ATATGCATAT3' (GC-core), and 5'ATGATATCAT3' (GC-mix) – along with their complementary strands in a periodic box of dimensions 7.774 nm[16,36]. We specified a 240 mM implicit salt concentration and used a Debye-Huckel approximation for electrostatic interactions within a 5 nm cutoff. We ran our simulations in the NVT ensemble and fixed temperature via a Langevin thermostat to account for implicit solvent interactions[55]. Simulations were conducted at empirically determined melting temperatures in order to maximize transitions between dissociated and hybridized states. We used a 20 fs time step and 1.3×10^9 steps in each simulation, saving frames every 100 ps. For each sequence, 40 simulations were performed in parallel, consuming about 24 serial CPU-hours on 28xIntel E5-2680v4 processors per simulation. Half of runs were initialized from the native hybridized state and half from a dissociated state. The first 10000 frames (1 μs of simulation time) of each simulation were removed, resulting in 40 x 250000 frames and a total of 1 ms simulation time per sequence. Depending on sequence,

We observed between 55-100 hybridizing and dissociating transitions across all runs.

2.2 Featurization

All intermolecular pairwise distances for both oligonucleotides were calculated at each frame using the MDTraj software package[56]. Intermolecular distances removed translational and rotational symmetries, but additional symmetries arise from the self-complementary nature of these sequences – both the sense and anti-sense strands are identical to each other. To account for this, we averaged permutable distances (45 pairs in total) together following a similar procedure used in tICAgg coordinates construction[57]. The VAMP-2 scoring method was employed to evaluate the kinetic variance of the feature set and optimize hyperparameters[58,59]. The VAMP-2 score uses the covariances of a set of inputs to estimate the transfer operator of a dynamical system, providing a robust and object means to evaluate various parameters and models. (Can include equations or additional info in SI?) We compared scores for the intermolecular, intra+intermolecular, and symmetrized intermolecular distances, in addition to the reciprocal value for each of these coordiantes (figure 1). We found that reciprocal coordinates performed better across the board, but we saw no substantial difference between the three featurization methods. This indicated that minimal kinetic information was lost in the symmetrization procedure, therefore we chose reciprocal symmetrized coordinates to train the model. The smaller feature set enabled faster training times and better statistics over permutable distances, without appearing to suffer a loss in generality or model resolution. These features were normalized and passed into sequence-specific SRVs.

2.3 SRVs

SRVs were first developed by Chen et al. as a means to directly learn slow eigenfunctions of the transfer operator[51]. The method is a descendant of VAMPnets[59], deep canonical correlation analysis[60], and extended dynamic mode decomposition with dictionary learning (EDMD-DL)[61]. The framework uses a twin-lobed artificial neural network to learn an opti-

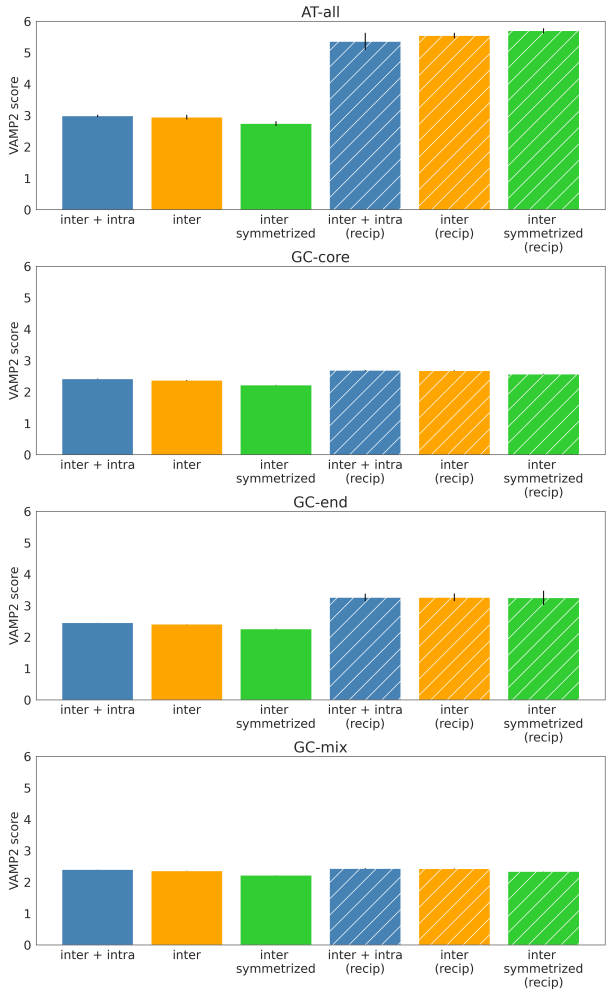


Figure 1: 5-fold cross validation to calculate the VAMP-2 score for each feature set. The inverse distances showed improvement across sequences, and the 55-dimension symmetrized coordinate set performed about as well as larger features set.

mal basis set for the variational approach to conformational dynamics (VAC) from which the leading eigenfunctions of the transfer operator are then estimated[62]. The resulting orthogonal modes are associated with the slowest dynamical processes in a system, and can be used to interpret kinetic information directly (such as physical correlations and timescales) and to construct MSMs[48,63]. SRVs provide robust nonlinear approximation and computation time that scales linearly with the amount of input data and are more powerful and efficient that tICA and ktICA approaches. This is a key attribute to our system as 10 million frames with 55 features in each frame are used for each sequence. The SRV framework has been tested on

toy systems where the true eigenfunctions of the transfer operator are known and on small protein simulation data such as the WW-domain and Trp-cage mini-protein[51,63]. For these latter systems, SRV-MSMs were contructed in order to find the stability of metastables states as well as transition probabilities between these states.

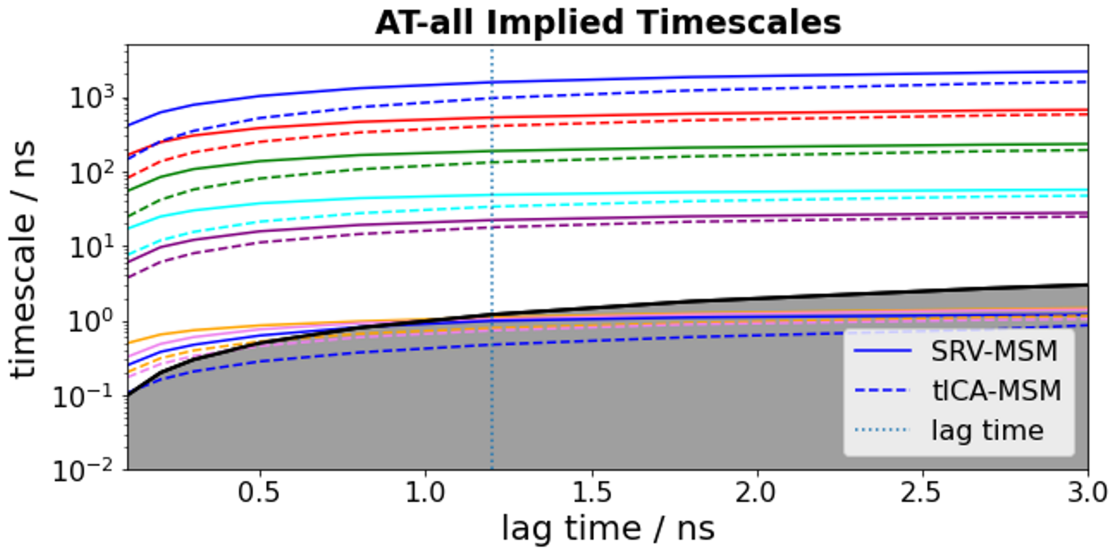
2.4 SRV-MSMs

MSMs are a powerful tool for interpreting large amounts of simulation data in a statistically robust and experimentally comparable way[64,65]. Kinetically similar conformations are first discretized into microstates and the conditional probability of transition between state is calculated within some lag time. The reliance on conditional probabilities allows for many independent simulations (longer than the lag time) to be collectively interpreted. To take full advantage of the MSM frameworks, however, the input basis should be as kinetically meaningful as possible[48]. Because SRV eigenfunctions translate simulation features into their slowest kinetic representations, they are optimally suited as an MSM basis. To build our SRV-MSM framework, we employed the PyEmma MSM pipeline and generated independent models for each sequence[66]. In a similar approach to Sidky et al. we performed k-means microstate clustering, Bayesian MSM construction, and PCCA+ hierarchical macrostate assignments[63]. The number of microstates were determined by VAMP-2 score, and the SRV-MSM lag time was selected based on implied timescales convergence. The number of PCCA+ macrostates was determined based on the characteristic of each system and will be discussed more in depth in the results and supplemental information.

3 Results

3.1 SRV-MSMs provide better resolution than tICA-MSMs

As a preliminary check on SRV-MSM performance, we compared implied timescales with a more conventional tICA-MSM approach. Previous work showed that SRV-MSM implied



Timescale at lag = 1.2 ns	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5
SRV-MSM (ns)	1561	527	186	48	22
tICA-MSM (ns)	956	404	131	37	18

Figure 2: SRV-MSM implied timescales converged faster than tICA-MSM implied across the five leading AT-all modes. Timescales are directly compared at the chose lag time of 1.2 ns

timescales converged faster than tICA-MSM timescales, enabling a shorter lag time and therefore a higher resolution model[63]. Here we observed the same trend across all sequences, and we highlight these results for AT-all in figure 2, emphasizing that faster convergence was observed across all leading timescales. Similar plots for the other sequences are included in the SI. We also note that the infrequency of transitions relative to the individual trajectory length leads to slower convergence of the leading mode. Because this mode corresponds the overall hybridization and dissociation process, we found that it displayed similar behavior across all sequences and that higher order modes were more informative for lag time selection.

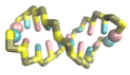
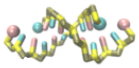
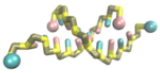
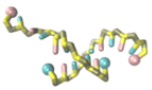
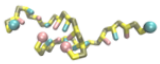
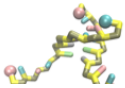
PCCA+ Macrostates	WC pair / Dangling End / Inert Tail		PCCA+ Stationary Probabilities			
			AT-all	GC-end	GC-core	GC-mix
Native Hybridized (H)	5′ ATATATATAT 3′ 3′ TATATATATA 5′		(4.152 ± 0.003) E-1	(5.720 ± 0.006) E-1	(4.535 ± 0.004) E-1	(4.572 ± 0.005) E-1
5'-Single Shifted (5S1)	5′ ATATATATAT 3′ 3′ TATATATATA 5′		(7.88 ± 0.02) E-2	(1.18± 0.06) E-3	--	--
3'-Single Shifted (3S1)	5′ ATATATATAT 3′ 3′ TATATATATA 5′		(1.57 ± 0.03) E-2	(4.0 ± 0.3) E-4	--	--
5'-Double Shifted (5S2)	5′ ATATATATAT 3′ 3′ TATATATATA 5′		(3.62 ± 0.07) E-3	(5.4 ± 0.5) E-5	--	--
3'-Double Shifted (3S2)	5′ ATATATATAT 3′ 3′ TATATATATA 5′		(1.20 ± 0.07) E-3	--	--	--
4-base pair frayed (F4)	5′ ATATGCATAT 3′ 3′ TATAGCTATA 5′		--	--	(2.77 ± 0.02)E-3	--

Figure 3: Nearest neighbor representations, molecular renderings, and sequence-dependent probabilities for PCCA+ macrostates. The dissociated state is not shown but represents the remainder of the stationary probability for each sequence.

3.2 Interpreting PCCA+ Macrostates

Following the analysis pipeline described above, we generate SRV-MSMs for each sequence. We identified seven kinetically relevant states that are captured within the resolution of the model. These include the fully hybridized state (H) in which all native base pairings are intact, the fully dissociated state (D), four "shifted" states (5S1, 5S2, 3S1, 3S2) where complementary base pairs form out-of-register, and a frayed state (F4), unique to GC-core, in which the four terminal A:T base pairs are unbound. Shifted states are only observed for AT-all and GC-end; state abbreviations indicate the direction of the shifted overhang (5' vs. 3') and the number of shifted motifs relative the native state (1 vs. 2). GC-mix displays two-state behavior within the resolution of the model, however we will show that the hybridization and dissociation are still characterized by an ensemble of shorter-lived states. Stationary probabilities in each of these states are shown in figure 3. We also show that PCCA+ clustering assigns states to relative free energy minima in the tICA space.

(3) Finish with sentence like: "We now proceed to analyze the sequence-dependent thermodynamic and kinetic features of each MSM."

3.3 ~~Shifted state stability~~ ← Better heading, eg "MSM stabilities in good accord with nearest-neighbor thermodynamic models."

The repetitive AT motifs in the AT-all and GC-end sequences produces a collection of out-of-register states similar to those shown in previous coarse-grained and all-atom studies [14,21,32,36]. The thermodynamic stability of these states can be evaluated against experimental prediction by defining each structure in terms of "dangling ends" – unpaired bases adjacent to the paired duplex – and "inert tails" – free bases that extend beyond the dangling end [67]. Dangling ends tend to have small stabilizing effects, and inert tails decrease stability as they increase in length – 3' inert tails have stronger effects than 5' tails. Although sequence-dependent

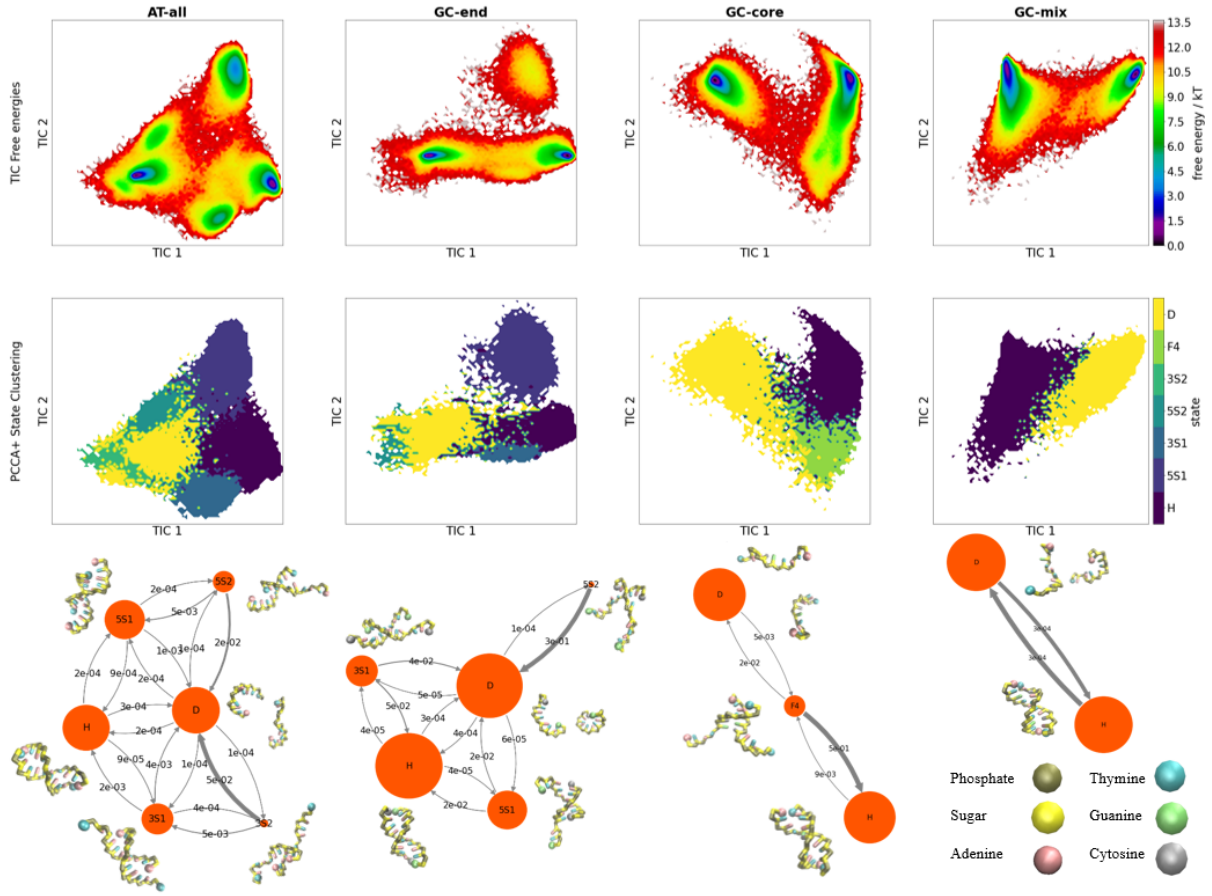


Figure 4: Free energies maps and PCCA+ states clusterings are shown in the tICA space in order to enhance visibility in dimensions. Flux diagram between PCCA+ states are shown for each sequences, accompanied by representative structures are for each state. Arrows indicate the probability of transitioning out of the state within the lag time. Circle areas are proportional to state stability.

effects of inert tails are less well known, dangling end effects can be factored into nearest neighbor (NN) calculations[9]. 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. For AT-all, NN calculations (figure 5) predict that conformations in the 5' shifted states (5S1 and 5S2) are more energetically favorable than those in the 3' shifted state (3S1 and 3S2). We find qualitative agreement to our PCCA+ 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 in the GC-ends shifted states 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 PCCA+ results, however, 3' out-of-register states to be more stable than 5' states. This leads us to believe that 5' vs. 3' differences – attributed to some combination of 5' tails preferentially stacking on the core duplex and 3' tails perturbing the duplex structure – may out-weight NN stacking effects alone[68].

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 5: 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.

3.4 Shifting dynamics

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. Figure 4 shows the probability of moving from one state to another within the MSM lag time. For AT-all, we observe approximately equal probability of transitioning from D to any other state, indicating that this is primarily a diffusion-driven process. 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 – 5S1 → H is more favorable than 5S1 → 5S2 – suggesting that these metastable shifted states play a more significant role in facilitating the hybridization process than dissociation. Furthermore, we see equal or higher transition probability from shifted states to the dissociated state (5S1 → D) than to more aligned states (5S1 → H), indicating that the shifting-hybridization process is frequently disrupted by complete dissociation. In particular, we observe that the transition probability 3S2 → D is 10x higher than the 3S2 → 3S1 path to native hybridization. Indeed, for GC-end we observe no significant flux between the 5S2 state and the structural similar 5S1 state, indicating that the former acts more as a kinetic trap than a pathway to H. The 5S1 and 3S1 states still readily convert with H, however there is an order of magnitude lower probability of reaching these states from D. This indicates 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 5S1 → H and 3S1 → H once shifted states have formed.

Taken together, our results indicate that AT-all hybridization and dissociation kinetics are substantially modulated by out-of-register pathways. A majority of hybridization events occur out-of-register, and, based on transition probabilities, 35% of D → H pathways pass through some out-of-register state. While GC-end strands can access out-of-register states, only 11% of native hybridization events pass through these states. These observations show 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[32,34,35]. Experimentally, out-of-register states are suspected to oc-

cur, but are difficult to capture due to short lifetimes and subtle kinetic traces. To explicitly minimize out-of-register base pairing, similar AT repeat motif sequences have been padded by GCG clamps during experimental analysis.[69] Recent all-atom results identified similar out-of-register states as "deep kinetic traps" along the hybridization pathway for repetitive dGCGCGC hexamers[14]. 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 whether these differences were a consequence of varied simulation conditions or whether GC repeat motifs are less susceptible to out-of-register transitions compared to AT motifs. This is possible given that stronger hydrogen binding in GC motifs may prevent fluctuation-driven rearrangement, however further computational and experimental studies are required to verify these differences.

What is M.S section doing? Delete.

3.5 Shifting experimental comparisons

In examining spectroscopic T-jump signatures for these four sequences, we observed a significantly stretched fast response for AT-all, and a more subtle GC-end stretch relative to GC-core and GC-mix fast responses. For each sequence, the fast response is attributed to terminal base fraying, but stretching indicates a broader distribution of dynamics at this timescale. We expect that fraying would still occur in out-of-register conformations but at a different rate than intact fraying. Furthermore, internal displacement mechanisms facilitating direct out-of-register transitions may produced similar responses contributing to the ensemble. Given that it is experimentally challenging to distinguish between these contributions, we cannot confirm whether these responses originate from some population of out-of-register configurations prior the temperature jump, shifting mechanisms during the dissociation process, or some combination of the two. In the context of our SRV-MSMs, we see qualitative agreement between more complex transition network and a more stretched response. We interrogate these differences by examining high temperature simulations later

in this manuscript. (Can discuss more details about fig 14 here and potentially include this fig in the main text)?

3.6 GC-core metastability

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 suggests that once key contacts are made, the zippering mechanism ensures that the helix will quickly form outward[13,32]. 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 $\leftrightarrow$ F4 transition has occurred, a F4 $\leftrightarrow$ H transition will likely proceed it. On the other hand, H $\leftrightarrow$ F4 events are more frequent but unlikely to initiate complete dissociation.

Lattice model studies have shown that frayed intermediates make substantial contributions to the GC-core conformational ensemble[21,40]. Araque et al. defines 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[21]. When examining all four sequences using T-jump IR and 2D IR spectroscopy, Sanstead et al. found that the GC-core had the highest deviation from two-state behavior during dissociation[16]. As their lattice model did not consider previously mentioned shifted states, 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, lattice model predictions showed that GC-core had substantially more frayed base pairs[40]. 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 inter-conversion between all accessible frayed states[27]. We see the same rapid fraying in simulation data – which is too fast to be attributed to a converged SRV mode – however we stipulate that this inter-conversion first relies on the slower formation of the the A:T bond nearest to the GC center. Although this process occurs much slower than single A:T base bonding and breaking, it may be difficult to experimentally discern from the overall hybridization process which contains both G:C and A:T character and occurs on a similar timescale.

3.7 GC-mix displays more canonical hybridization and dissociation

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 process, we designated this transitions as effectively two-state within the resolution of our model. We did, however, we do 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[16]. While AT-termini fraying is surely a prerequisite to dissociation, we find these states to be so common and fleeting 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.

Given the lack of a repetitive AT interior (as in AT-all and GC-end) or consecutive AT exterior (as in GC-core), we expect more canonical dynamics from GC-mix. For this analysis, we looked at qualitative trends in our trajectory data, paying close attention to the distances

between matching WC-pairs (6). During hybridization, we observed the formations of 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[13,17,24]. 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. There's evidence of a short-lived state composed of 2-4 base pair contacts immediately before full dissociation occurs. In contrast to the F4 state we observe in GC-core, these conformations do not form a distinct free-energy minima in SRV or tICA space, nor do they tend to reform intact duplexes. As a whole, these dynamics are similar to previously reported "fraying-peeling" mechanism[41,42,44]. 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.

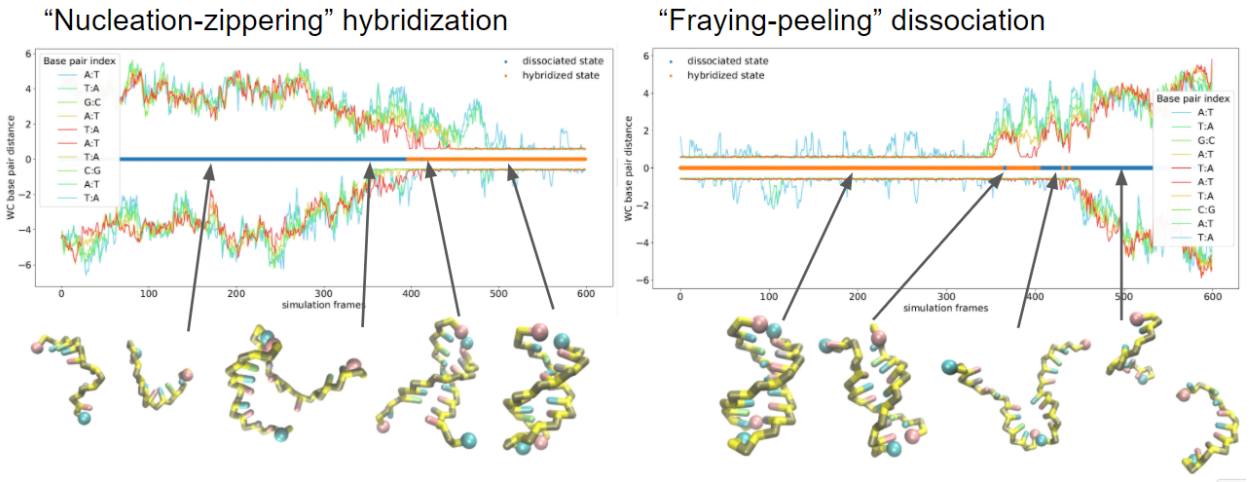


Figure 6: WC base pair distance and molecular renderings along two representative GC-mix hybridization and dissociation events. Permutable distances are reflected across the x-axis and show that fraying in the hybridized states tends to occur independently on either end of the duplex until fully dissociation 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.

Better title. Problem- not method-centric. what do dMaps help you find out? MSM

3.8 Diffusion maps show structural diversity within ~~PCCA+~~ macrostates

Although our SRV-MSMs provide useful insight into slow processes, we were also interested in the ensemble of configurations within PCCA+ macrostates. These configurations interchange significantly faster than the SRV and MSM lag times, but we can use a diffusion map embedding to glean a structural understanding of the macrostate population. Diffusion maps generate a low dimensional embedding of the data into high variance structural modes and are well-suited to find subtle differences in temporally disconnected data[70,71]. We were particularly interested in using this method to explore the 5s1 and F4 macrostates as these represent the most popluated and kinetically relevant macrostates for AT-all, GC-end, and GC-core.

Reads like exploratory analysis. Make hypothesis and use dMaps to support.

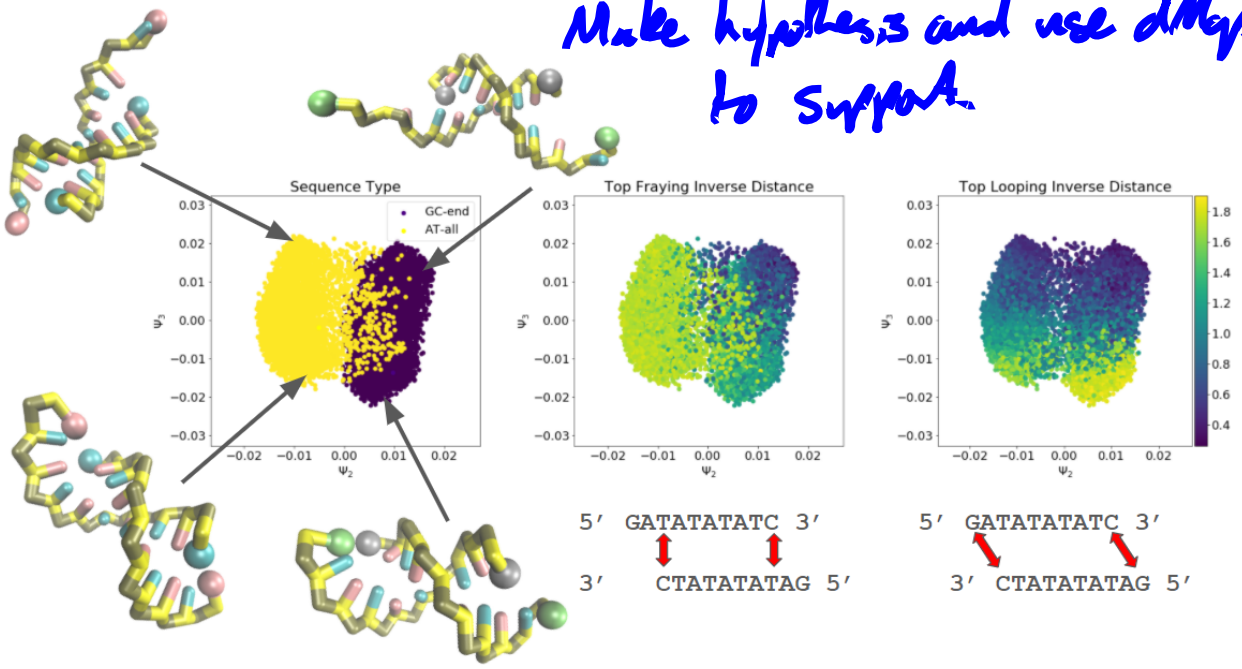


Figure 7: First two diffusion map coordinates built from 10000 5S1 states, equally sampled from AT-all and GC-end. Color maps show inverse distances between out-of-register ends and complementary ends.

3.8 5s1 state analysis

While examining molecular renderings, we noticed that a significant proportion of GC-end 5s1 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 5S1 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 (Figure 15). Diffusion maps built from samples of the 3S1 states are also shown in the SI (Figure 16).

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. This reveals that the mismatched C:T pairs are never bound – a consequence of the 3spn2 excluded volume interaction – whereas the AT-all pairs are mostly bound with occasional fraying indicated by small AT-all overlap in the GC-end region. This effect may be exaggerated given that C and T base pairs are assigned slightly higher excluded volume radii in 3spn2[33]. The second diffusion mode, which correlates highly with the average distance between 3' and 5' ends, has higher values for GC-end than AT-all. Because the GC-end termini do not bind out of register, we find that they are readily able to form stabilizing contacts despite the shifted conformation of the duplex as a whole. These "shifted-loop" bonds are shown to be uniquely stable for GC-end conformations in the 5' shifted state, and their existence in the simulations is confirmed by molecular renderings sampled from

these regions. Although AT-all shifted ends tend to stay bound out-of-register, the second diffusion coordinate shows some population of inert tails that fold back onto the helix.

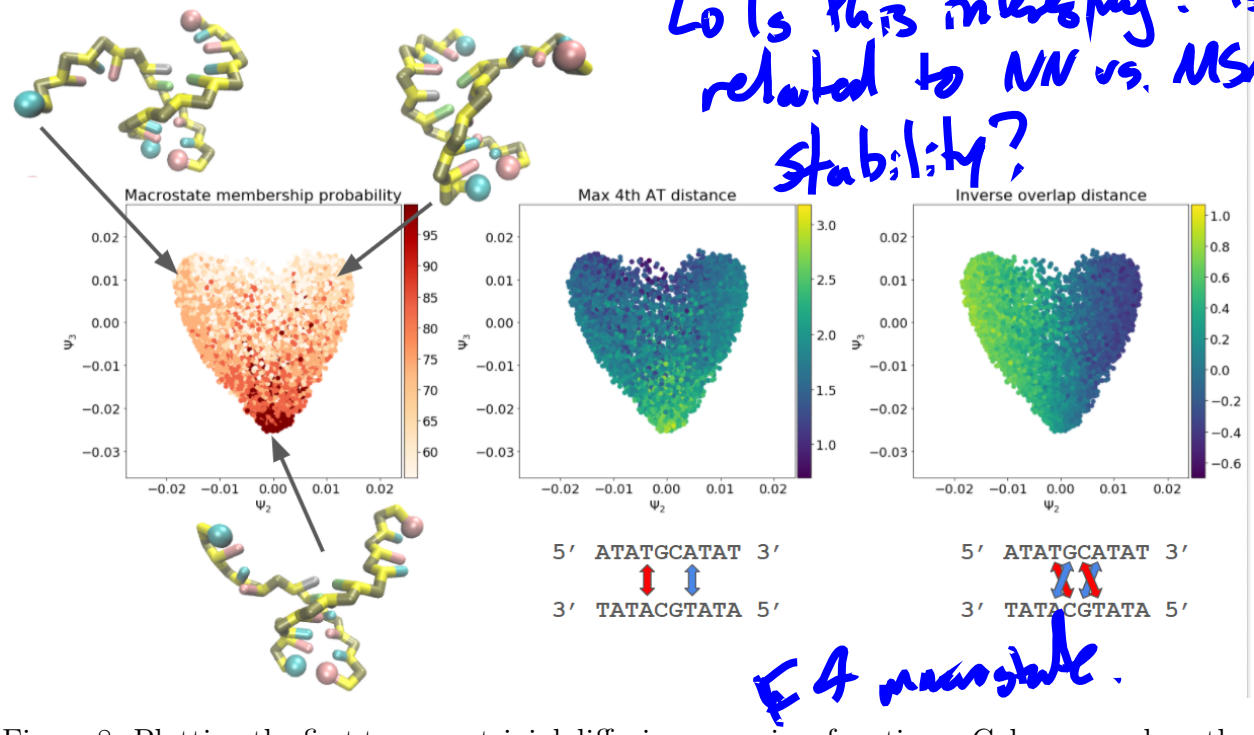


Figure 8: Plotting the first two non-trivial diffusion map eigenfunctions. Color maps show the probability that conformations are clustered into the frayed macrostate, maximum distance between 4th AT basepairs, and the overlap between adjacent A:T basepairs and the GC core.

3.8 F4 state analysis

We applied a similar approach to investigate the structural composition of the F4 macrostate. We build diffusion maps using 10000 frames sampled from the macrostate (figure 8). This time, we set our distance metric to the same permutation-free coordinates we used to build the SRV-MSMs. Again, we were able to identify a combination of physical coordinates that closely correlated to the first two non-trivial diffusion modes. We found that a larger distance between internal A:T pairs increased the PCCA+ probability of inclusion into macrostate. This distance also correlated closely with the second diffusion map mode. Interestingly, we observed that the first non-trivial mode – the feature that describes the most structural diversity in the system – corresponds to difference in "overlap" distance between adjacent

A or T bases and the GC core. We defined overlap as the distance between a G/C and the A/T adjacent to its native complement. In these conformations, one of the strands maintains some helical character while the other twists out of place, resulting in WC bonds being obstructed by the oligo backbone. These states represent another potential way in which the hybridization process (or helix reformation) can be kinetically frustrated. We observed this mode to be mostly symmetric, however there is slight tendency for the 3'T end to fray farther out of place relative to its 5'A counterpart. This might be another consequence of differential excluded volume radii in the 3spn2 force field.

3.9 SRV correlations to physical coordinates

Is this section useful?

Having constructed and interpreted SRV-MSMs, we revisited our original SRV basis to investigate how slow modes are constructed from input features. We found these GC-core modes to be of particular interest as they reveal the hierarchical and nonlinear nature of the dynamical encoding. In particular, we examined a collection of "trimmed" trajectories centered on both hybridization or dissociation events. For each trajectory, we compared the first three SRV coordinates with a corresponding collective variables with which they shared behavior. Two representative trajectories are shown in figure (3.9). Complementary G:C pairs are the best indicators for a hybridization/dissociation event, and we see a sharp change in the first SRV mode (SRV1) as these bonds form or break. The second slow mode (SRV2) is most active when G:C pairs are bound but the adjacent AT pairs are not. There is a small signal for fraying at the outer base pairs, but the mode overwhelmingly learns about these neighboring A:T/G:C bonds. Moreover, this behavior reflects movement in and out of the F4 state in the corresponding SRV-MSM discussed above. SRV3 is most active during dissociation, and seems to track closely with the average distance between all complementary base pairs. We attribute this to the SRV learning about the diffusive motions of the two body system and evaluating the likelihood of an imminent hybridization event. The third mode also peaks when the oligos are close together but configured in such a way that is

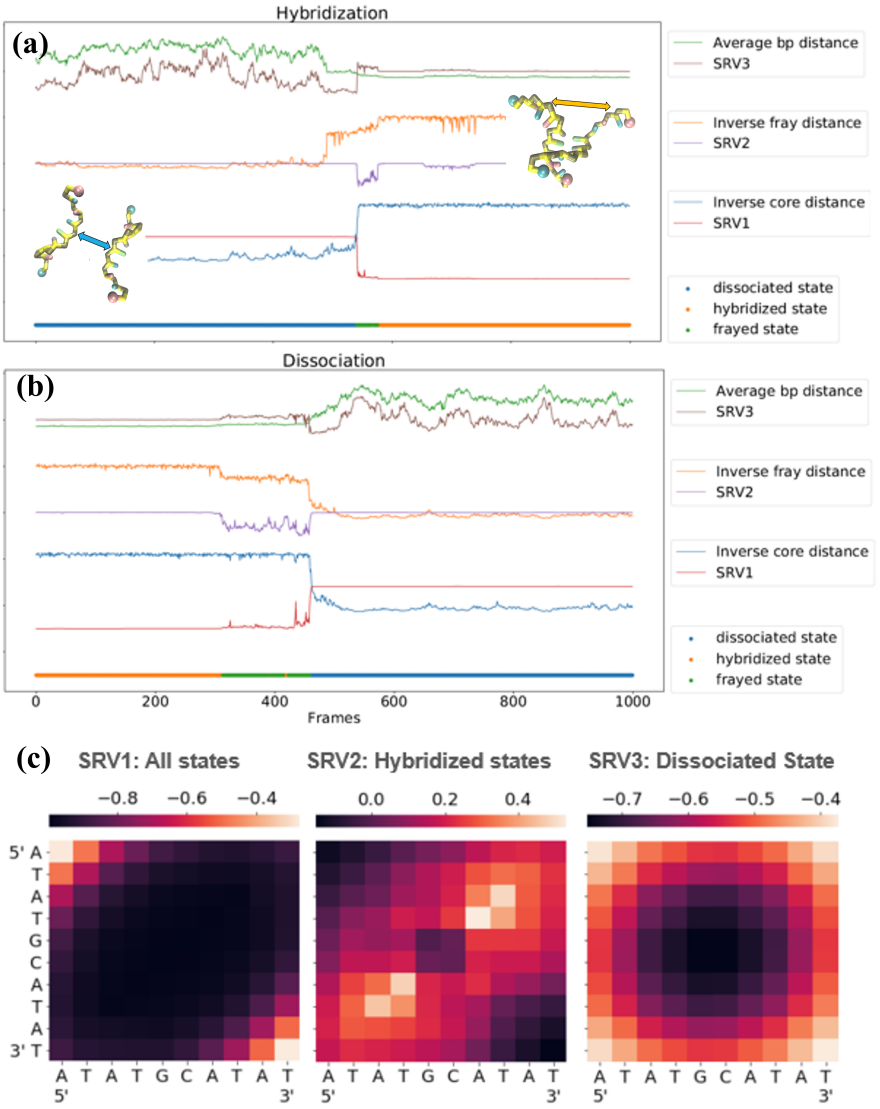


Figure 9: Leading GC-core SRV coordinates correlated with identifiable physical coordinates during sample hybridization (a) and dissociation (b) events. (c) Pearson correlations between all 100 intermolecular base pair distances and each the three leading slow modes. SRV2 correlations were calculated in the H and F4 states; SRV correlations were calculated in the D state.

not amenable hybridization. These misaligned conformations include inverse contacts where 5'/5' and 3'/3' ends meet and looped conformations where one strand is folded in on itself and preventing satisfactory WC contacts.

Despite the qualitative trends we observe between physical coordinates and leading SRV modes, we found it difficult to find correlations across full trajectories. With the exception of SRV1, Pearson correlations between physical coordinates and SRV modes were near zero and showed erroneous trends. However, when we performed this analysis along smaller sections of trajectories we noticed that the sign of SRV2 and SRV3 correlations switched depending on whether the oligos were in the hybridized or dissociated state. This shows that these modes are inherently nonlinear and provide support on top of the first mode – which serves as an indicator function for hybridized vs. dissociated state. With respect to our SRV-MSM macroscates, we found that SRV2 was "turned on" in the H and F4 states – corresponding to the intact helix and frayed state, respectively – and SRV3 was turned on in the dissociated state. Accordingly, we calculated Pearson correlations between each SRV mode and all distances in states where the modes are active. Figure shows the highest correlation between SRV2 and inner A:T pairs, weak correlation with outer A:T pairs, and an inverse correlation with 5'/5' and 3'/3' pairs which tend to approach each when the duplex is in the F4 state. We also observed a highly symmetric correlation between SRV3 and central base pairs distances, which indicative of overall diffusive behavior. Taken together, these analyses reveal how the SRV learns and represents the dynamical space in an hierarchical manner, providing enhanced resolution to a linear method like tICA and producing the MSM results we observe above.

Better title.

Hypothesis & methods

3.10 Temperature-dependent timescale comparisons

Given differences in temperature and ensemble distributions between Tmelt simulations and T-jump experiments, we found it difficult to make direct timescales comparisons based on our equilibrium models alone. To supplement our analysis, we ran short simulations initialized

This section can be much lighter.

in the hybridized state at a series of elevated temperatures for each sequence. We derived relaxation times for a "slow" dissociation response and "fast" fraying response at each temperature and compared these with experimental temperature-dependent relaxation fits for GC-end and GC-core. Additionally, we repeated the protocol in Sanstead and Tokmakoff[27] to generate temperature series data for AT-all and GC-mix sequences.

We measured the slow dissociation response by fitting the distribution of times at which the core base pairs separate beyond a 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. Furthermore, we see an acceleration of about one order of magnitude compared to experiment, although this factor is sensitive to the exact definition of melting temperature which can vary between simulation experiments. After accounting for the acceleration factor, we saw agreement between the experimental data and simulated relaxation fits. GC-core showed the largest deviation experiment, with dissociation rate increasing much more quickly with temperature in simulations.

The fast response – which Sanstead et. al attributed to base pair fraying signatures – was more difficult 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[45–47,72]. All-atom simulations suggest that frayed ends can assume misaligned WC bonds, base-sugar hydrogen bonds, and terminal stacked conformations[44,73]. 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 either duplex terminal end to split 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

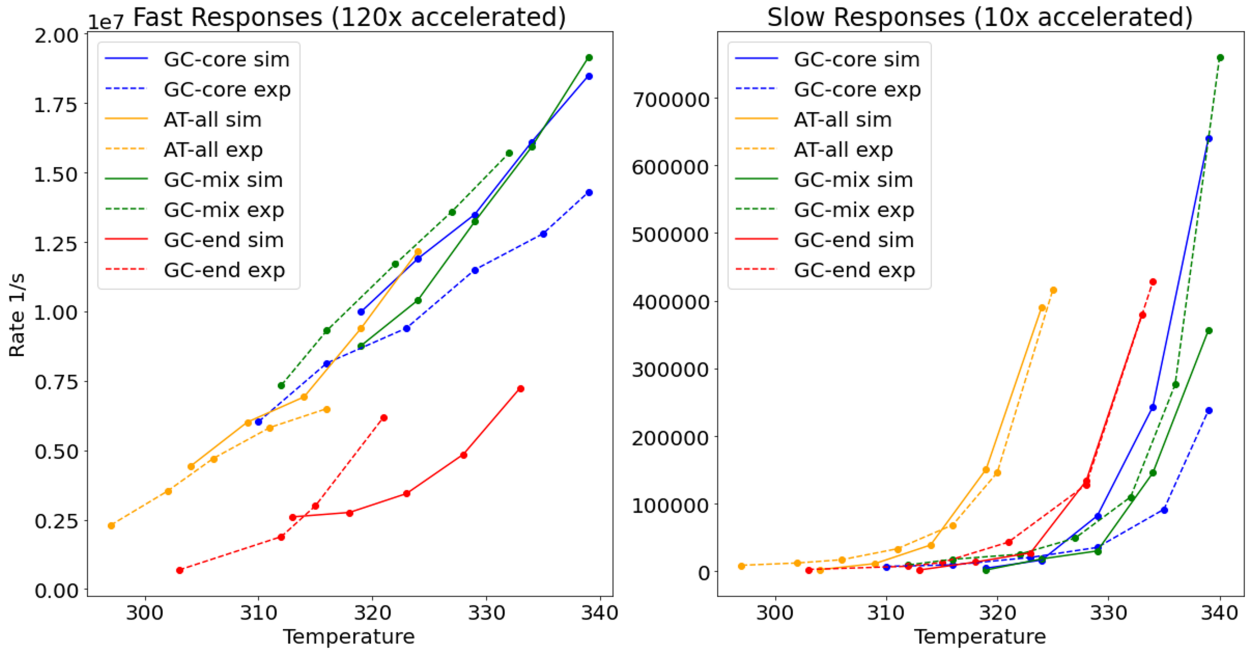


Figure 10: Fitting temperature-dependent trends to the "slow" dissociation mode and "fast" fraying mode detected in experiments. The effective simulation temperature is shifted by 4 degrees K to account for systematic differences in simulation and experimental melting temperature. Different scaling factors are applied to fast and response to account for variable coarse-grained acceleration along different degrees of freedom.

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 thermodynamic barrier associated with G:C fraying. We observe similar trends in the experimental data, although comparisons at high temperatures were limited due to mixing between the fast and slow responses (figure 10). 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 two order of magnitude relative to experiments. 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 we 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 capture some sequence-dependent fraying effects.

4 Conclusion

We have demonstrated how coarse-grained MD can be supplemented by data-driven time-lagged analysis to efficiently learn the slowest processes of DNA hybridization. We constructed high resolution SRV-MSMs capable of distinguishing sequence-dependent dynamics, and showed the relevance of these dynamics to experimental results. We found that AT-all and GC-end sequences both participate in some degree of out-of-register base pair shifting, although these states have higher kinetic relevance 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. This approach allowed us to aggregate an ensemble of trajectories in order to properly sample hybridization and dissociation without bearing a large computational expense. Furthermore, we replicated T-jump analysis across a series of temperatures to verify previous spectroscopic assignments and sequence-dependent effects at different timescales.

4.1 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 relevant 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 must 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 5S1 $\rightarrow$ H transition – when the usually rough free energy path becomes more easily traversed.

4.1.1 Not sure whether to discuss explicit comparisons

We also acknowledge that the implicit model may not fully capture the role of ions in facilitating hybridization and dissociation. In addition to implicit simulations, we followed the same pipeline to generate simulations and SRV-MSMs given the ion conditions specified in the Sanstead et al.[16] using the 3spn2 explicit implementation[74]. These results look very similar to the implicit case, although with slightly slower dynamics and higher uncertainty in stationary probabilities. Comparisons were limited by the higher computational demand required to simulate 5x the beads in the box, a shorter integration timestep, and reliance on Ewald summation to account for electrostatics. We caution over-interpretation of these results as the explicit model has not been as widely adopted or thoroughly validated compared to the implicit model.

Going forward, we believe that this work can be extended to predict more general trends in sequence-dependent hybridization and strategies for experimental comparisons. These insights should be leveraged for sequence design and incorporated into the growing field of dynamic DNA nanotechnology.

5 SI Methods notes

5.1 VAMP-2 Scoring

Optimization of hyperparameters including feature choice, number of microstates, and number of slow modes was evaluated with via the VAMP-2 score. In the equations below, we show how covariances are obtained from some featurization χ of a time series x_t and its time-lagged pairs $x_{t+\tau}$. The VAMP-2 score can then be found for χ by applying the VAMP principle with cross-validation.

$$\mathscr{C}_{00} = \mathrm{E}\left[\chi(x_t)\chi(x_t)^\intercal\right]_t$$
$$\mathscr{C}_{01} = \mathrm{E}\left[\chi(x_t)\chi(x_{t+\tau})^\intercal\right]_t$$
$$\mathscr{C}_{11} = \mathrm{E}\left[\chi(x_{t+\tau})\chi(x_{t+\tau})^\intercal\right]_{t+\tau}$$

$$VAMP-2[\chi] = \left\|\mathscr{C}_{00}^{-1/2}\mathscr{C}_{01}\mathscr{C}_{11}^{-1/2}\right\|_F^2 + 1$$

5.2 SRV training

Using optimized hyperparameters and featurized trajectory data, we transformed 55 reciprocal pairwise distances into a low dimensional SRV basis set. In order to maintain consistency between sequences, we kept all SRV training hyperparameters the same with the exception of the number of outputed slow modes. We determined the number of slow modes via cross-validation on the VAMP-2 score to ensure that the coordinate did not over fit on statistical noise[75]. In particular, we looked for convergence in the VAMP-2 score and inconsistency between cross validation scores – suggesting that the model may be over-fitting on artifacts in the training data. We used a batch size of 50000 and ran each model for a total of 20 training epochs. We used two hidden layers and set the size of each layer to 100. For cross-validation and comparison between different hyper-parameters, we used a 80/20 validation split training. SRV training required about 22 GPU-minutes across 1 GPU and 10 CPUs. SRV training was implemented using Keras and Tensorflow[76,77].

5.3 SRV-MSM walkthrough and analysis

In our analysis, we found that the AT-all sequence, given its repetitive structure and lack of GC-content, produced the simplest dynamics and displayed a clear spectral gap between

modes. For this reason, we use this sequence a case study to work through our SRV-MSM pipeline step-by-step. Our first task was to identify the SRV lag time that was longer than the intrinsic Markov timescales of the system, yet short enough to resolve the dynamics of interest[78]. We found that most implied timescales converge at an SRV lag time of 1.2 ns. Next we selected an optimal number of SRV components to include in our analysis. After a certain point, higher order dynamical modes provide diminishing contributions the overall kinetic variance as measured by the Vamp-2 score, and the model can begin fitting on statistical noise in the trajectory data instead of the true dynamics[75]. It is also more difficult to perform kmeans clustering on a high dimensional space, especially when those higher dimensions are less kinetically relevant[48]. For these reasons, the number of slow SRV components should be carefully selected based on the specific system of interest. As shown in figure 12), we see diminishing returns in the VAMP-2 score after five slow modes and select these modes as our optimized SRV basis.

Although SRV coordinates alone provide some information, we can access a more holistic picture of sequence kinetics and thermodynamics by using these SRV coordinates as a basis on which to construct an MSM. Because these coordinates are already capturing a majority of the system's kinetic variance, they serve as an ideal basis on which to group frames into microstates. We performed k-means clustering, and optimized the number of microstates at 200 by monitoring VAMP-2 score. Next, we selected an MSM lag time in a similar fashion to our SRV lag time selection process. This enables us to select a shorter lag time and build a higher resolution model than we could from an analogous tICA basis. Setting the MSM lag time to the same 1.2 ns we used for our SRVs, we built a Bayesian MSM to calculate transition probability matrix between each microstate. Finally, PCCA+ spectral clustering was implemented to group these microstates into macrostates that each represented a collection of metastable structures. Previous works have used a common set of microstates and/or performed manual clustering of microstates based on physical read outs from simulation data (stacking score, energies, etc)[73,79]. Although these techniques are useful for performing

comparisons between sequences, we saw better results when optimizing MSMs to capture the most detail of sequence individually and thus developed an independent set of microstates and macrostates for each sequence.

Next, we seek to interpret the physical relevance of these leading modes by plotting the Pearson correlation of each mode with the 100 intermolecular distances between strands. The quantitative meaning of these coordinates can be difficult to interpret given their nonlinear relationship to the SRV collective variables, but the relative different between these correlations shows which coordinates are most relevant to each process. For example, the first slow mode shows a positive correlation to each distance and the strongest correlation with native base pair distances (shown along the main diagonal). Given these correlations and the substantially longer timescale of this process, we can deduce that this leading mode corresponds to the overall hybridization and dissociation process. The next four SRV components all show a relatively high correlation along offset diagonals. These diagonals correspond to the intermolecular distances between complementary but out-of-register base pairs and point to the existence higher order "shifting" processes between sets of such base pairings. Previous "inchworm" and "pseudoknot" mechanisms have similarly been reported in simulation studies to correct base pair mismatches and occur on orders of magnitude longer timescales than underlying fast dynamics such as fraying[32,34,35].

We found GC-end had a similar implied timescales distribution to AT-all, with a distinct spectral gap after the fourth mode. For AT-all and GC-end, we kept to the convention of clustering into n+1 macrostates, where n is the number of slow components captured by the MSMs. For GC-core we built an SRV-MSM using these first three SRV modes as a basis and proceeded along the pipeline as described above. We found that four macrostate clustering was unstable – likely because the third mode is mostly providing information about dissociation dynamics – so we performed PCCA+ clustering into three macrostates representing the hybridized (H), dissociated (D), and 4 4 base pair frayed (F4) states. The GC-mix sequence showed a similar implied timescale distribution to GC-core, however we

no longer saw a converged slow mode corresponding to multi-base fraying behavior. Instead, we observed two modes converge, corresponding to the association/dissociation dynamic and diffusive behavior while strands are dissociated. These correlate closely with the first and third GC-core SRV modes. Although we built our SRV-MSM using these two coordinates, we again were unable to form a stable third state along the second coordinate. As such, we designated this transitions as effectively two-state within the resolution of our model.

To visualize macrostates, we project the data into the two leading tICA coordinates. Although SRVs outperform these coordinates for the purpose of MSM construction, tICA coordinates represent good high variance collective variables on which to visualize free energies and state assignments[63]. Furthermore, we found that all macrostate were discernable on the first two tICA coordinates, whereas multiple SRV dimensions would be necessary to visualize states. After assigning these macrostates we calculated their stationary and transition probabilities averaged across 100 Gibbs sampled MSM and PCCA+ assignment. Means and standard deviations were calculated from this ensemble.

5.4 SI figs

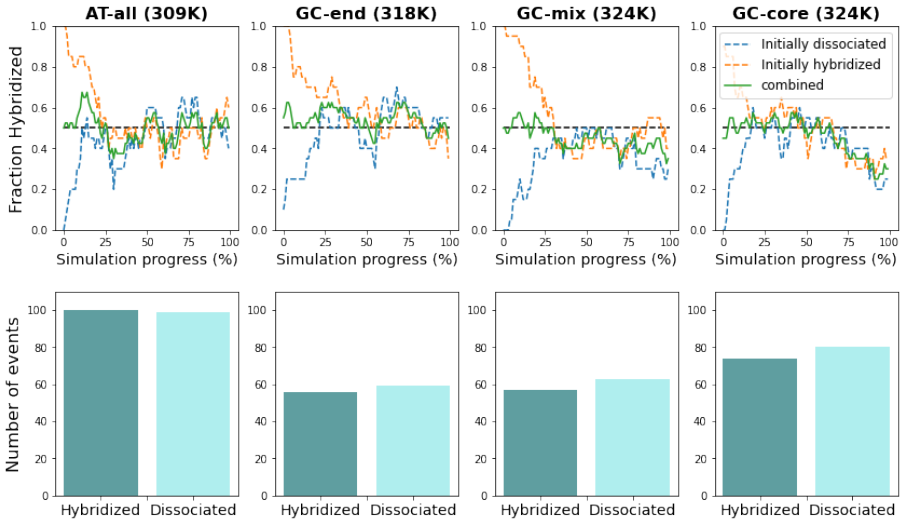


Figure 11: All sequences consist of 20 trajectories initialized in the hybridized states and 20 in dissociated state. The fraction of hybridized duplexes averaged across these sets is shown over time. For all sequences these curves converge near 0.5, but stochasticity of rare-event makes a definitive Tmelt difficult to identify. The number of hybridization/dissocation combined across trajectories is shown. Not every trajectory contained a transition event, but on average more than one full event occurred (H→D→H) per trajectory, providing adequate sampling of dynamics. AT-all undergoes substantially more transition, but many of these do not reach a native (in-register) hybridized state.

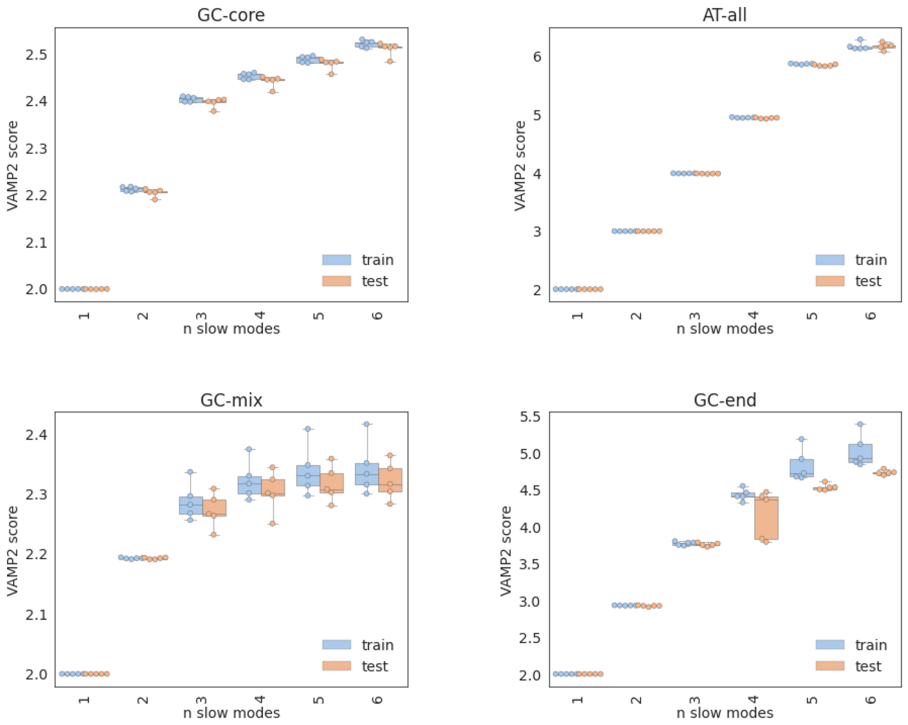


Figure 12: 5-fold cross validation procedure to select number of SRV coordinates. We look for the converge of the VAMP-2 scores, inconsistent scores between folds, and deviation between training and test data as indicators that the model has begun fitting on statistical noise.

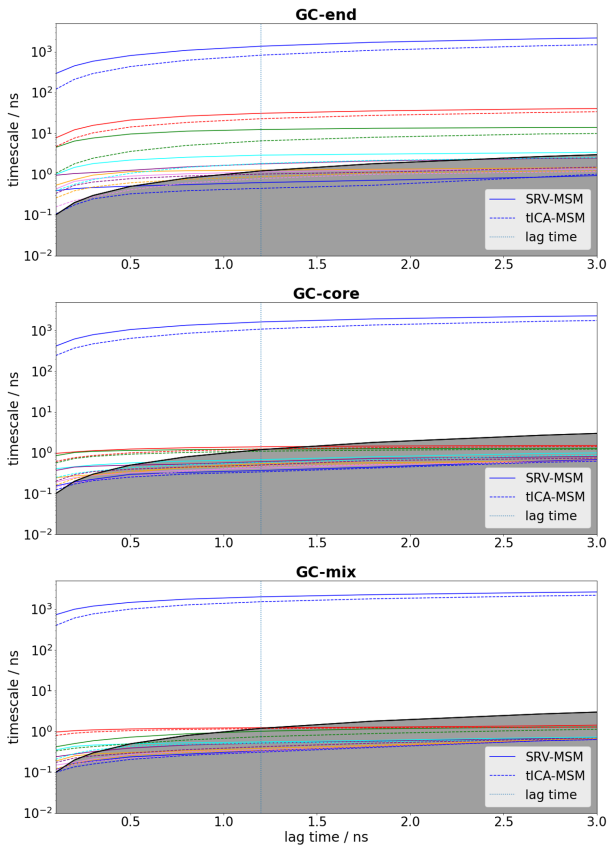


Figure 13: SRV-MSM vs. tICA-MSM implied timescales convergence for GC-end, GC-core, and GC-mix. Progressively larger spectral gaps are observed between the first mode and higher order modes. We observe convegenve of resolvable higher order modes at a shared lag time of 1.2 ns. It is difficult to converge the leading mode in this regime due to the infrequency of hybridization/dissociation events.

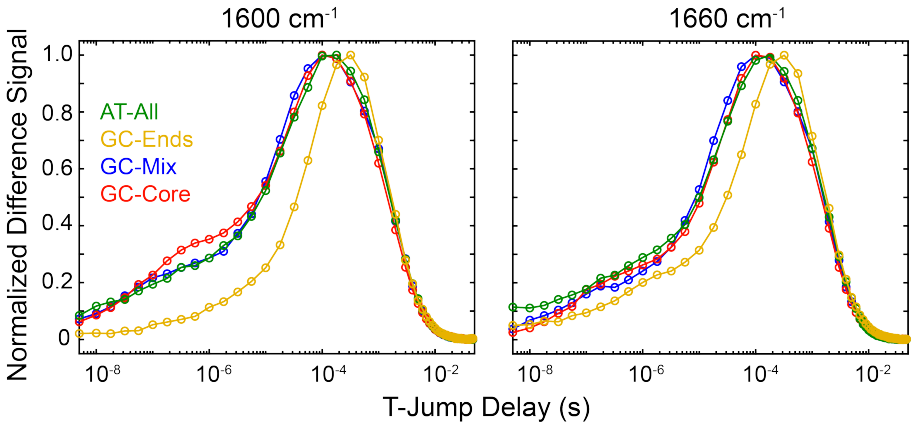


Figure 14: Kinetic traces of the fast spectroscopic response show varied degrees of shifting based on sequences. More info from Brennan on subtle differences here, discuss stretching coefficients? (could include this in main text as well)

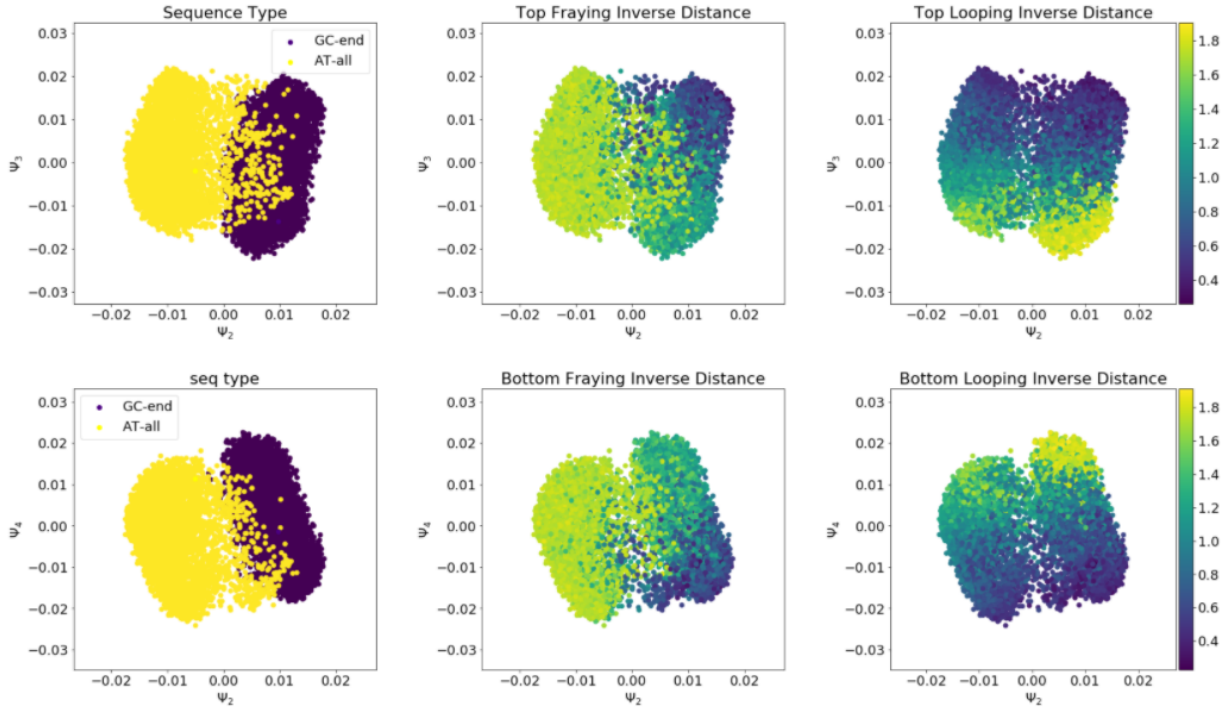


Figure 15: Full 5' diffusion maps including degenerate third diffusion map mode.

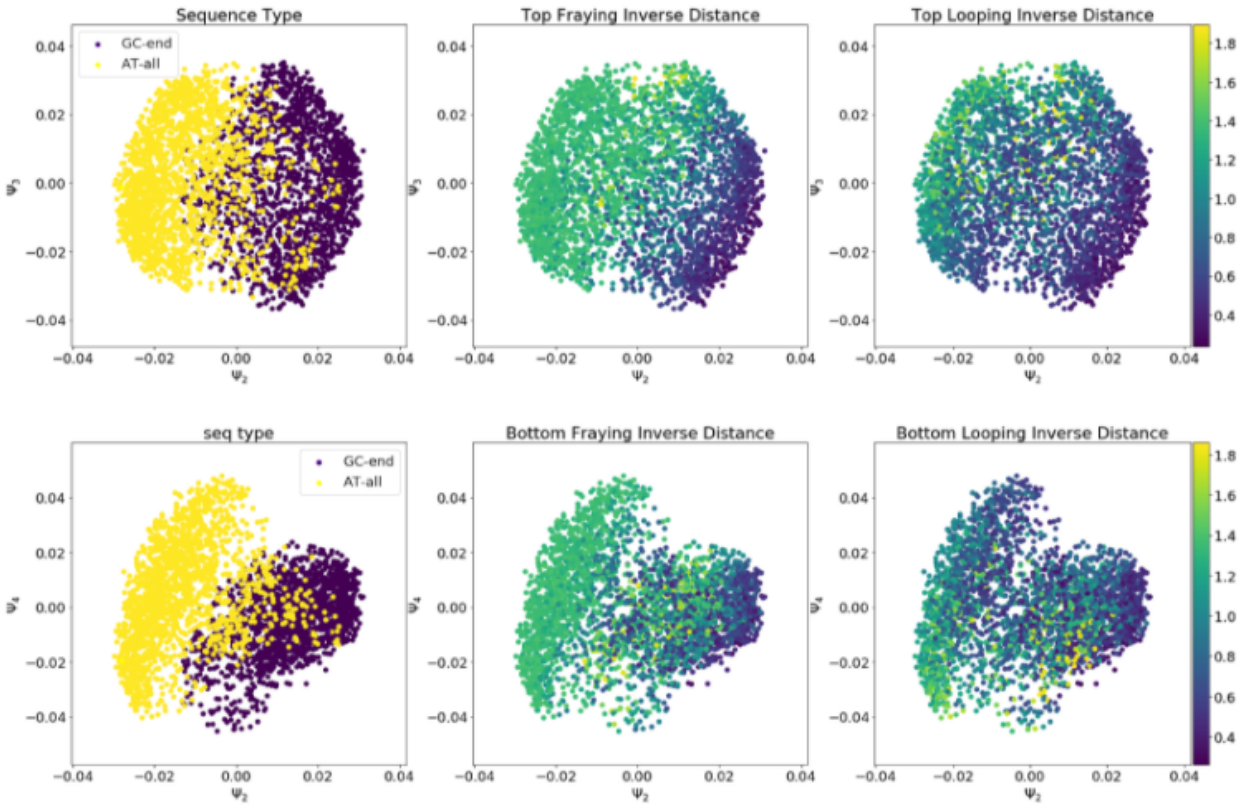


Figure 16: Full 3' diffusion maps, shows greater overlap between the AT-all and GC-end populations, as well as a less distinct "looping" region for intact GC bonds

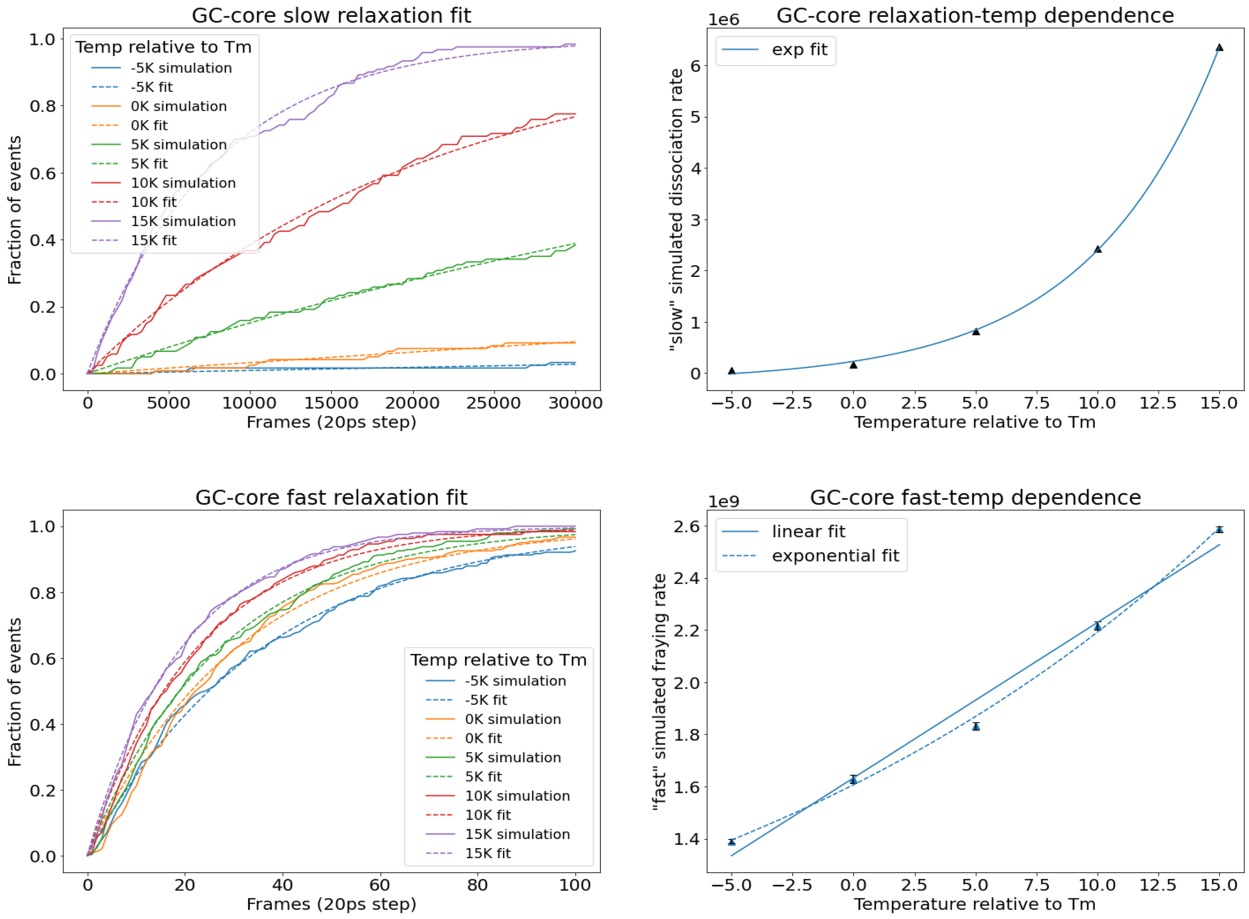


Figure 17: Slow and Fast modes were determined by fitting over a distribution of 120 independent simulations. For each temperature in the series, a relaxation curve was fit to the distribution and the associated exponential coefficient was used to the calculate rate. Temperatures series between -5 - +15K relative to empirically determined sequence melting temperature were explored. This process was repeated for each sequence (only GC-core is shown above).

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) Deluca, M.; Shi, Z.; Castro, C. E.; Arya, G. Dynamic DNA nanotechnology: Toward functional nanoscale devices. *Nanoscale Horizons* **2020**, *5*, 182–201.

(6) 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.

(7) 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.

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

(14) 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.

(15) 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*.

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

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

(18) 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.

(19) 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.

(20) Craig, M. E.; Crothers, D. M.; Doty, P. elaxation Kinetics of Dimer elf Complementary Oligon. *J. Mol. Biol.* **1971**, *62*, 383–401.

(21) Araque, J. C.; Robert, M. A. Lattice model of oligonucleotide hybridization in solution. II. Specificity and cooperativity. *Journal of Chemical Physics* **2016**, *144*.

(22) 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.

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

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

(25) 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.

(26) 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.

(27) 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.

(28) 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.

(29) 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*.

(30) 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.

(31) 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.

(32) 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.

(33) 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*.

(34) 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.

(35) 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.

(36) 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*.

(37) 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*.

(38) 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.

(39) 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.

(40) 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*.

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

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

(43) 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.

(44) 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.

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

(46) 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.

(47) 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.

(48) 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.

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

(50) 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**,

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

(52) 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.

(53) 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.

(54) 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.

(55) 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.

(56) 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.

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

(58) Wu, H. Variational approach for learning Markov processes from time series data. 1–30.

(59) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nature Communications* **2018**, *9*, 1–11.

(60) Andrew, G.; Bilmes, J.; Livescu, K. Deep Canonical Correlation Analysis. **2013**, *28*.

(61) Li, Q.; Dietrich, F.; Bollt, E. M.; Kevrekidis, I. G. Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the koopman operator. *Chaos* **2017**, *27*, 1–25.

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

(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) 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 : Gener-

ation and validation Markov models of molecular kinetics : Generation and validation. **2011**, *174105*.

(65) Husic, B. E.; Pande, V. S. Markov State Models : From an Art to a Science. **2018**,

(66) 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. **2015**,

(67) 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.

(68) 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.

(69) 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.

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

(71) 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.

(72) 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.

(73) 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*.

(74) Hinckley, D. M.; Pablo, J. J. D. Coarse-Grained Ions for Nucleic Acid Modeling. **2015**,

(75) McGibbon, R. T.; Pande, V. S. Variational cross-validation of slow dynamical modes in molecular kinetics. *Journal of Chemical Physics* **2015**, *142*.

(76) Keras @ Github.Com. `https://github.com/fchollet/keras`.

(77) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. **2016**,

(78) 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*.

(79) Pinamonti, G.; Zhao, J.; Condon, D. E.; Paul, F.; Noè, F.; Turner, D. H.; Bussi, G. Predicting the Kinetics of RNA Oligonucleotides Using Markov State Models. *Journal of Chemical Theory and Computation* **2017**, *13*, 926–934.