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

Add Melody?

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

Abstract

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

Content is good but a little delocalized and could be sharpened by tighter structure. Maybe the following paragraph structure:
1. DNA versatility & importance 3. Kinetics/mechanisms less understood.
2. Thermo well understood

1 Introduction

4. Prior kinetics mechanism work (Tokmakoff etc.)
5. What we do in this work - sophisticated analysis, diffusion, CG sims, experimental comparisons.
6. Main findings and conclusions. (Very brief.)

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 applications[1–4]. Recently, structural DNA nanotechnology has enabled self-assembly on micron to millimeter scales, and dynamic DNA nanotechnology has been used to perform basic calculation and to probe single molecules via temporal DNA signatures[5–7]. Both technologies rely on the hybridization reaction between complementary DNA strands and leverage the flexibility of shorter DNA oligomers to participate in these reactions. Although many experimental and computational studies have rigorously explored DNA dynamical phenomena such as hybridization, hairpin formation, and single base pair flipping, the sequence-dependent mechanisms of hybridization and dissociation dynamics are not fully understood[8–15]. Moreover, it is unclear the extent to which these processes evolve in an "all-or-nothing" fashion or if an ensemble meta-stable states facilitates the transition. Recent breakthrough studies have coupled experimental techniques with machine learning and MD simulations to investigate and predict sequence-dependent kinetics[16,17]. Where these studies focus on association and dissociation kinetics alone, we broaden our analysis into higher order dynamical processes and meta-stable intermediates. The stability of certain intermediates states, such as out-of-register or shifted base pairing in repetitive sequences, has been well documented in previous computational studies[9,18,19]). Furthermore, frayed structures and dynamics have been investigated in numerous computational and experimental studies[20–23]. Sanstead et al. highlighted the role of these dynamics during duplex dissociation, where the stability of frayed states was dictated by G:C base placement 10-mer oligonucleotide sequences[11]. In this work, we study the same four sequences explored by Sanstead et al in an effort to uncover the sequence-dependent dynamics and their relation to metastable structures mentioned above.

Our understanding of DNA dynamics has been built from decades of experiments – such as temperature-jump, salt-jump, pH-jump, or other perturbative methods – that drive DNA

out of equilibrium and monitor relaxation processes in one direction[13–15,24–28]. More recently, single molecule diffusion and tethered multifluorophore assays have facilitated equilibrium analysis, however these present technical difficulties and are hampered by data collection rates[16,29–31]. FoÌĹrster resonance energy transfer (FRET) analysis, particularly when coupled with methods mentioned above, provides additional resolution, but it is unclear how fluorescent tags may interfere with the dynamics of short nucleotides[24]. Given that dynamic insights from these experiments are limited, several high-level computational models have been employed to gain further detail[19,32–34]. Although these models provide experimentally verified speed-ups compared to all atom simulations, the long timescales on which DNA hybridization and dissociation events occur make these processes difficult to sample via direct simulation techniques[18]. Instead, many previous studies of DNA hybridization have employed accelerated sampling methods such as umbrella sampling, transition path sampling, and forward flux sampling[32,35–37].Other computational works use dramatically elevated temperature or denaturing solvent concentrations to induce one-way dissociation events[38,39]. Taken together, most experimental and computational work have studied certain aspect the overall dynamics process in one direction.

In this work, we use the coarse-grain 3 Sites Per Nucleotide (3spn.2) 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[40] – to combine many independent and unbiased trajectories and develop an understanding of sequence-specific kinetics and thermodynamics. MSMs have recently been implemented to study mechanisms and microstate distributions of DNA hybridization[9,41], but the slowest sequence-dependent kinetics were not the focus of these studies. Pinamonti et al.[42] used MSMs to compare the slowest dynamics of short RNA nucleotides and found that stacking timescales are highly sequence dependent[42]. We take a similar ap-

proach to study 10-mer DNA oligonucleotides and introduce State Free Reversible Vampnets (SRVs) to directly learn the slowest sequence-dependent dynamical modes[43]. Furthermore, we integrate SRVs into the MSM pipeline by generating an optimized low dimensional basis in which microstates 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 GC base pair placement in decamer oligonucleotides has a substantial effect on dynamical behavior. By evaluating equilibrium trajectories we can study the relevance of meta-stable states during both the hybridization and dissociation process. Because SRVs generate an optimized low dimensional basis, we show that we can access higher resolution MSMs (about a 50%reduction in required lag time) and generate more detailed models. Additionally, we can compare slow dynamical modes and meta-stable states between sequence-specific SRV-MSMs. Within these meta-stable 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 DNA dynamics. Taken together, our analysis reflects similar results to previous computational and experimental DNA work, while elucidating new insights into sequence-dependent dynamics, meta-stable structures, and relative timescales.

2 Methods

2.1 3spn.2 Model

No need for separate section. Abbreviate and integrate into section below.

The 3 site per nucleotide (3spn.2) coarse grain model was designed to accelerate DNA computation relative to all atom models while maintaining experimental melting temperatures, stacking energies, and persistence lengths[33]. Interactions 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[10,33]. 3spn.2 has been implemented to study DNA pack in viral capsids, protein-DNA binding, and nucleosome unwrapping, and recently served as a basis for the coarser 1spn model[44–47]. The model is amenable to an implicit or explicit ion environment – we chose the latter to include Mg^{2+} effects from experiment. We achieved better sampling of hybridization events with explicit ions, despite a more demanding computational cost.

2.2 Simulation set up

We initialized four sequences previously investigated by Sanstead et al. – 5'ATATATATAT3' (AT-all), 5'GATATATATC3' (GC-end), 5'ATATGCATAT3' (GC-core), and 5'ATGATATCAT3' (GC-mix) – along with their complementary strands according to 3SPN.2 documentation[11,18]. We initialized explicit ions such that 240 mM NaCl and 18 mM $MgCl_2$ were added to the box in addition to 18 Na counter ions to balance the charge from the 9 phosphate groups in each oligonucleotide backbone[48]. In order to maximize concentration without allowing strands to see each other through periodic boundaries, we set the box size just larger than the sum of the maximum end-to-end extension length of a single strand and the force cutoff (using Ewald summation method is set at 20 Å). This translated to a box size of 77.74 ÃĚ and an effective oligo concentration of 7 mM. We used an Ewald potential to calculate long range Coulombic interaction between DNA and ions. We used a Debye-Huckel screening potential to account for phosphate-backbone interactions and preserve the persistence length and intrinsic curvature of DNA while taking the effects of ion-DNA interactions[48]. We ran our simulations in the NVT ensemble and fix temperature via a Langevin thermostat in order to model implicit solvent interactions[49]. We set the simulation about 1 K higher

Simulations were conducted at the empirically determined

them empirical sequence-specific melting temperature in order to maximize transitions between dissociated and hybridized states. We use a 15 fs time step and $2 * 10^8$ steps in each simulation, equating to 4.5μs simulation time per run? Base pair coordinates were saved every 30 ps, and full frames including backbones and ions were saved every 3 ns for visualization purposes. For each sequences, 100 simulations were performed in parallel, consuming about 32 serial cpu-hours of computation time for each independent simulation. An approximate Tmelt Boltzman distribution was replicated by initializing half of runs from the hybridized state and half from a random dissociated state. In order to allow for further equilibration, the first third (1.5 μs) of each simulation was removed, resulting in 100 x 100000 frames and a total of 300μs simulation time per sequence.

10 times

CPU-h on a Intel XX processors

the rest initialized

Add something about no. of transition events seen per run? Demonstrate good sampling.

2.3 Featurization

More intro here. why?

All intermolecular pairwise distances for both oligonucleotides were calculated at each frame using the MDtraj software package[50]. Based on the self-complementary nature of each sequence – meaning that both the sense and anti-sense strands are identical to each other – we averaged permutable distances (45 pairs in total) together. This permutation reduction follows a similar procedure used in TICAgg coordinates construction[51]. The VAMP-2 scoring method was employed to evaluate the kinetic variance of the feature set and optimize hyperparameters[52,53]. The VAMP-2 score uses the covariances of a set of inputs estimates the transfer operator of a dynamical system (equation below), providing a robust and object means to evaluate various parameters and models. 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.

Structure of this section:
- Need for featurization to remove rigid translation + rotation.
- VAMP-2 comparison of different choices (inter + intra, inter, symmetrized inter, + reciprocal versions of these)
- Choose symmetrized due to self-complementarity and no VAMP-2 loss.

$$\mathscr{C}_{00} = \mathrm{E}\left[\chi(x_t)\chi(x_t)^\intercal\right]_t$$
$$\mathscr{C}_{01} = \mathrm{E}\left[\chi(x_t)\chi(x_{t+\tau})^\intercal\right]_t$$
$$\mathscr{C}_{11} = \mathrm{E}\left[\chi(x_{t+\tau})\chi(x_{t+\tau})^\intercal\right]_{t+\tau}$$
$$VAMP2[\chi] = \left\|\mathscr{C}_{00}^{-1/2}\mathscr{C}_{01}\mathscr{C}_{11}^{-1/2}\right\|_F^2 + 1$$

We compared scores for the permutation-free distances, intermolecular, and intra+intermolecular distances and found small differences between the three that varied by sequences (show all sequences in supplemental), but overall we did not observe a loss in kinetic information when using the reduced feature set (1). We found a substantial increase in VAMP-2 score when using reciprocal pairwise distances and chose reciprocal permutation-free coordinates to train the model. The smaller feature set enabled faster training times and better statistics over permutatable distances, without appearing to suffer a loss in generality or model resolution. These features were normalized and passed into sequence-specific SRVs.

2.3.1 SRVs

SRVs were first developed by Chen et al. as a means to directly learn slow eigenfunctions of the transfer operator[43]. The framework uses a twin-lobed artificial neural network, similar in structure to the network employed in VAMPnets[53], to learn a low dimensional representation of input features. This representation is optimized for the variational approach to conformational dynamics (VAC) from which the leading eigenfunctions of the transfer operator are then estimated. 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. Other methods, such

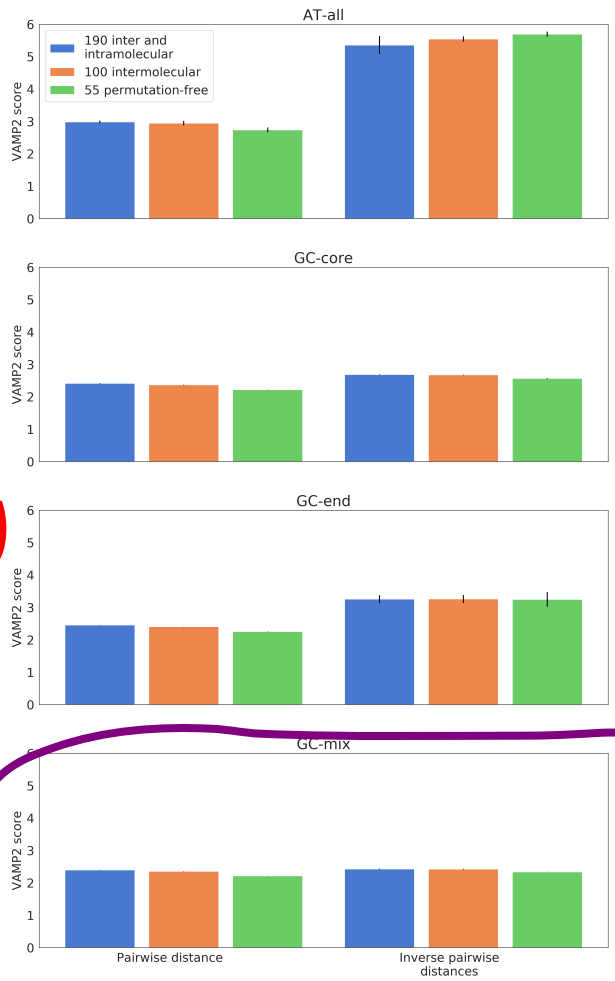


Figure 1: 5-fold cross validation to calculate the VAMP-2 score for each feature set. The inverse distances showed improvement across sequences, and the 55 reduced coordinate set performed about as well as larger features set.

~~as time-lagged independent component analysis (TICA) and kernel TICA (kTICA), have been employed in place of SRVs but suffer limitations from the linear transformation in the case of the former and high computational cost in the case of the latter~~[54,55]. SRVs provide robust nonlinear approximation and computation time that scales linearly with the amount of input data. 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[43,56]. For these latter

system, SRV-MSMs were contructed in order to find the stability of metastables states as well as transition probabilities between those states.

Using optimized hyperparameters and featurized trajectory data, we transform 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[57]. In particular, we looked for convergence in the vamp-2 score and inconsistency between cross validation scores – suggesting that the model may be overfitting on artifacts in the training data. We used a batch size of 50000 and ran each model for a total of 30 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[58,59].

2.3.2 SRV-MSMs

MSMs are a powerful tool for interpreting large amounts of simulation data in a statistically robust and experimentally comparable way. The technique relies on the discretitization of kinetically similar conformations into microstates and finds the conditional probability between states 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[40]. This becomes crucial in our system given the large difference in timescales between leading modes. Because SRV eigenfunctions translate simulation features into their slowest kinetic representations, they are optimally suited as an MSM basis. To prove this, SRV-MSM VAMP-2 scores were shown to be consistently higher than MSMs constructed from TICA coorinates (TICA-MSMs)[56]. Furthermore, SRV-MSM im-

plied timescales converge faster than TICA-MSM timescales, enabling a shorter lag time and therefore a higher resolution model. To build our SRV-MSM framework, we employed the PyEmma MSM pipeline and generated independent models for each sequence[60]. After passing in SRV coordinates, we perform k-means microstate clustering, Bayesian MSM construction, and PCCA+ hierarchical macrostate assignments. 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.

3 Results

3.1 AT-all

3.1.1 SRV optimization and analysis

In our analysis, we found that the AT-all sequence, given its repetitive structure and lack of GC-content, produced the cleanest dynamics and displayed a canonical spectral gap between modes. For this reason, we lead our discussion with this sequence and use it as 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[61]. We found that most implied timescales converge at an SRV lag time of 1.2 ns. We kept a looser constraint on the convergence of the leading mode – corresponding to the overall hybridization/dissociation process – as it tends to have a longer transit period and becomes more difficult to fulfill Markovian conditions. It should be noted that our lag time selection was informed by the other sequences as well to maintain consistency. 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[57]. It is also more difficult to perform kmeans clustering on a high dimensional space, especially when those higher dimensions are less kinetically relevant[40]. For these reasons, the number of slow SRV components should be carefully selected based on the specific system of interest. As shown in (Figure 2), we see diminishing returns in the VAMP-2 score after five slow modes and select these modes as our optimized SRV basis. 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 affected by (or effective on) each process. For example, the first slow mode shows a significant positive correlation to each distance and the strongest correlation with matching base pair distances (shown along the main diagonal). Given these relationships 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[9,19,32,34].

3.1.2 SRV-MSM construction and optimization

From this analysis we can determine that the slowest dynamics are fully characterized by the association/dissociation process and shifting behavior of out of register base pairs. Although this is qualitatively informative, 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

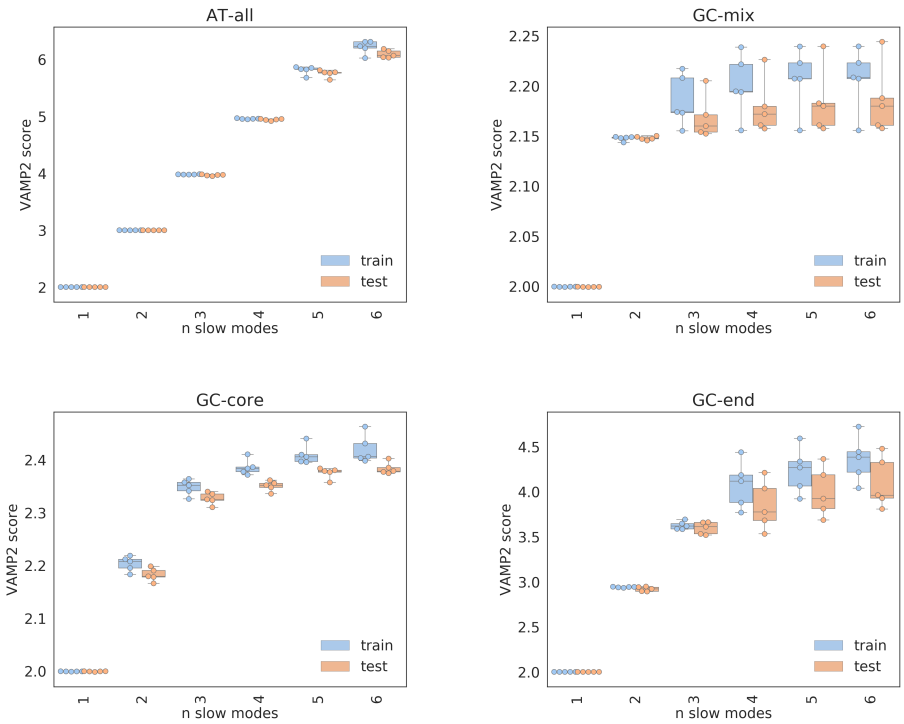


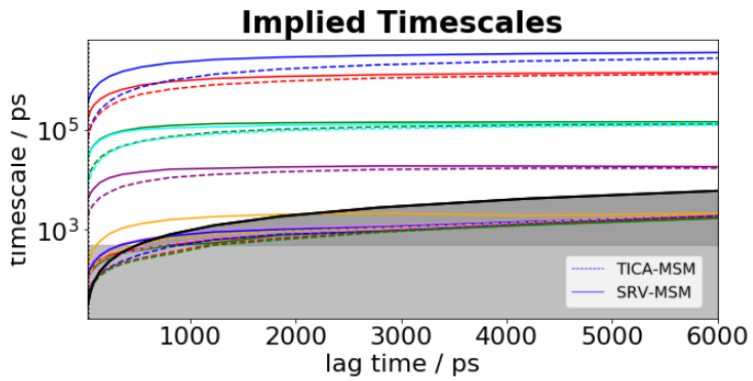
Figure 2: 5-fold cross validation procedure to select number of SRV coordinates. We look for the converge of the VAMP-2 scores, inconsistent scores between folds, and deviation between training and test data as indicators that the model has begun fitting on statistical noise.

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. In Figure 3 we compare the convergence of SRV-MSM and TICA-MSM timescales, where SRV-MSM timescales converge consistently faster and to higher values of all leading modes. 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)[62,63]. 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. For AT-all, we kept to the convention of clustering into n+1 macrostates, where n is the number of slow components captured by the MSMs. To visualize these six 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[56]. It is clear in Figure 4 that the PCCA state assignments are capturing free energy minima in TICA space as independent states. After assigning these macrostates we can then calculate their relative probabilities and free energies, visualize representative molecular renderings, and estimates transition probabilities between states. In this final step, we use a minimum flux cutoff of 2e-6 in order to mitigate erroneous quick transition or skipping between states.

3.1.3 Implied timescales for TICA-MSMs for

In our coarse-grained SRV-MSM, we observe an "aligned hybridized" state, dissociated state, and four "shifted" states characterized by different combinations of out-of-register base pairings. These shifted states consist of 2 or 4 base pair shifting (single or doubled shifted) in either the 5' or 3' direction with varied stability and transition fluxes between kinetically neighboring states. This state decomposition is expected given the repetitive nature of the AT-all sequence and previous computational results that find shifted conformations form "deep kinetic traps"[9,18]. We found a substantial difference in thermodynamic stability and kinetic behavior between the 5'A shifted states and 3'T shifted states. This difference can be accounted for by examining experimental studies on the thermodynamics of "dangling ends" – unpaired bases adjacent to the paired duplex – and "inert tails" – free bases that



Slow mode (Lag = 1.2 ns)	TICA-MSM (ns)	SRV-MSM (ns)
1	1242.6	2476.3
2	771.7	1051.4
3	86.9	130.8
4	80.4	111.1
5	12.7	16.9

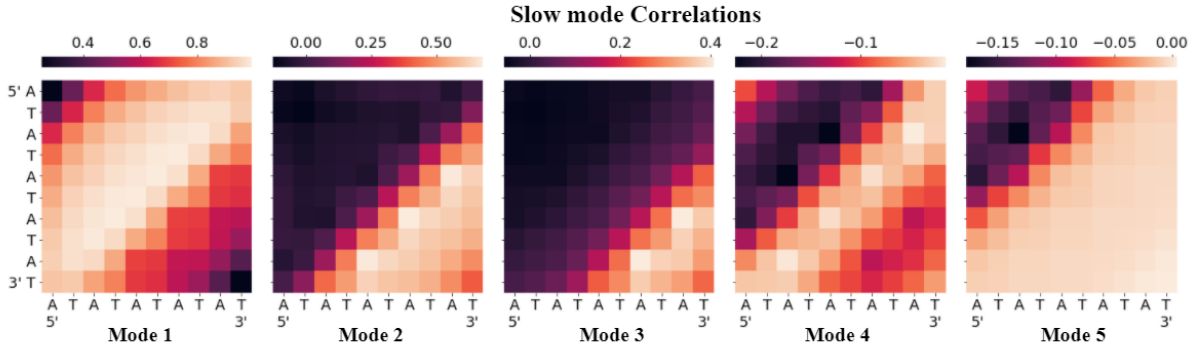
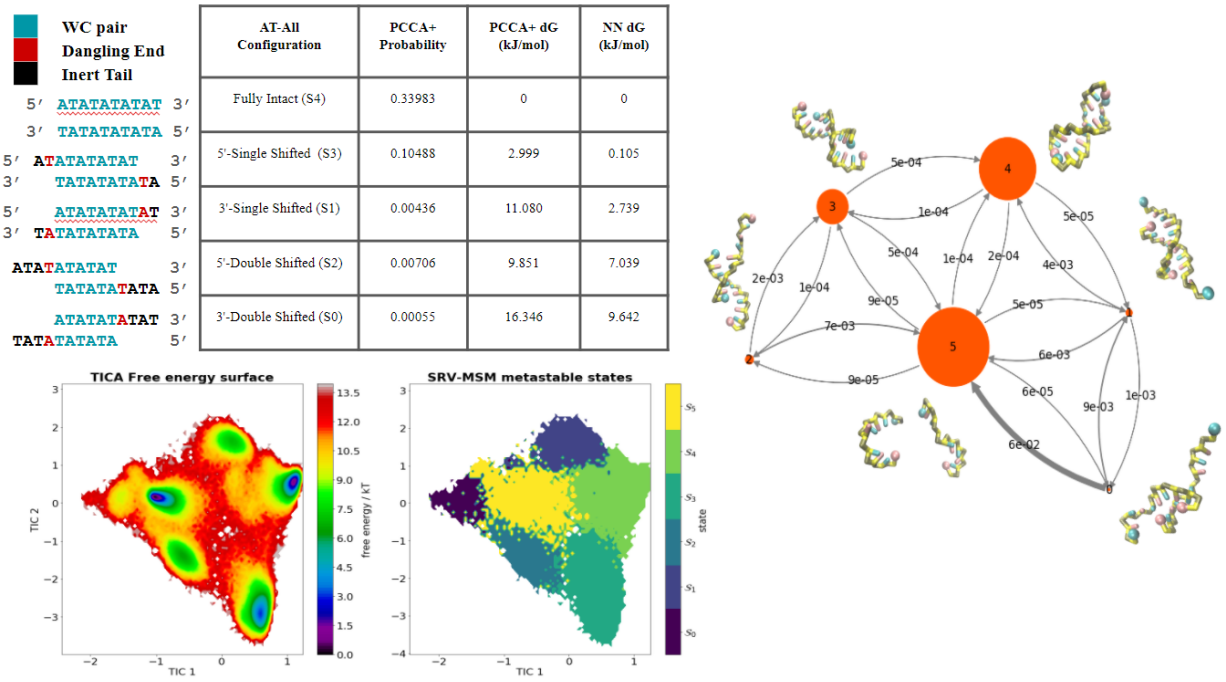


Figure 3: AT-all SRV-MSM timescales convergence and implied timescales using a lag time of 1.2 ns. Leading slow mode correlations to all 100 intermolecular distances between oligos.

extend beyond the dangling end[64]. Dangling 5' ends have a consistently stabilizing effect, and inert tails tend to decrease stability as they increase in length, particularly at lower ionic strengths. It has been shown that 5' dangling ends with one inert tail have higher melting temperatures and are energetically favorable compared to 3' ends[65,66]. These effects might be attributed to some combination of 5' tails preferentially stacking on the core duplex and 3' tails perturb the duplex structure and solvation environment[67]. Furthermore, DNA nearest neighbor (NN) calculations that include dangling end contributions (Figure 4) predict that conformations in the 5' shifted state are more energetically favorable than those in the 3' shifted state. Although NN calculations were accurate in predicting the intact thermodynamics behavior of these sequences, they assume all-or-nothing hybridization and do not take inert tails into account[68,69]. Therefore it makes sense that our PCCA+ free energies are all higher than thermodynamics prediction, due to some destabilizing effect of the inert tails.

Notably, the 3' dangling states have a 6-8 kJ/mol higher free energy compared to theory, whereas the 5' dangling states deviate by less than 3 kJ/mol. In fact, the free energy of the 5'-double shifted state is slightly lower than that of the 3'-single shifted state, despite the former having two fewer intact base pairs. The double-shifted 3' state is more stable than one might expect given the difference in stability between the single and double 5' shifted states. This might be attributed to the asymptotic effects of inert tail destabilization – longer tails tend to have a diminishing effect on overall strand stability[64].



AT-All Configuration	PCCA+ Probability	PCCA+ dG (kJ/mol)	NN dG (kJ/mol)
Fully Intact (S4)	0.33983	0	0
5'-Single Shifted (S3)	0.10488	2.999	0.105
3'-Single Shifted (S1)	0.00436	11.080	2.739
5'-Double Shifted (S2)	0.00706	9.851	7.039
3'-Double Shifted (S0)	0.00055	16.346	9.642

Figure 4: a) AT-all free energies calculated from nearest neighbor data and from MSM PCCA+ probabilities b) TICA projections of free energies and PCCA+ state clusterings c) Transitions probabilities between states within the 1.2 ns lag time

3.1.4 Transition probabilities between states

Beyond state probabilities and free energy approximations, the macrostate 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 (1.2 ns). All transition probabilities are higher when moving towards a more aligned state than

towards a more shifted state, 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 rates from shifted states to the dissociated state than to more aligned states, indicating that the shifting-hybridization process is frequently disrupted by complete dissociation. In particular, we observe that the transition probability from double-shifted 3' state to the dissociated state is 6x higher than to the neighboring single-shifted 3' state. Although there is an approximately equal probability of transitioning from the dissociated state to the hybridized and 5' shifted states, there is a 2x lower probability of moving to the 3' shifted states. These asymmetric kinetic effects, in addition to NN and inert tail contributions described above, may further lower the 3' shifted states populations relative to 5'. From these observations we conclude that shifted states are important in facilitating the hybridization process, but that increasingly shifted states (in particular when shifted in the 3' direction) are more likely to dissociate than to evolve into a fully aligned hybridized state. We recognize that these effects could be sensitive to the specific ion environment – especially considering the effect of ionic strength on inert tails – and warrant further computational and experimental investigation beyond this work.

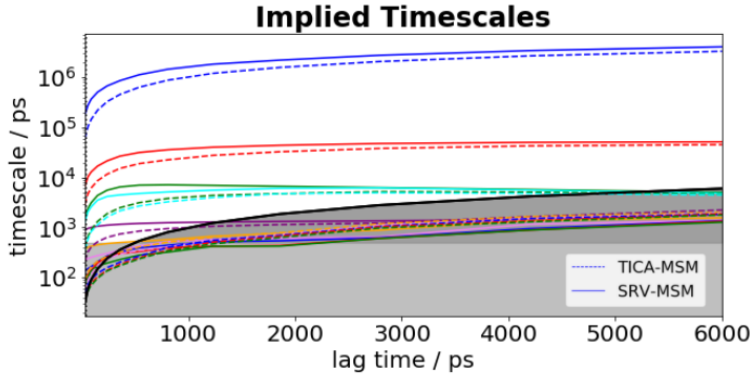
For all proceeding sequences, we use the same SRV and MSM lag times, number of microstates, and cross-validation procedures as in the AT-all case. For each sequence, these preserve Markovian properties of the system and provide some consistency to compare across models. The number of SRV coordinates and MSM macrostates are determined based on cross-validation and the kinetic behavior of specific sequences.

3.2 GC-end

3.2.1 SRV-MSM construction and states

Next we examine the GC-end sequence, which add G:C caps to the same repetitive AT motifs. Based on SRV cross-validation (SI) we use the leading four slow modes to define our MSM basis. These coordinates are qualitatively similar to those we studied for AT-

all, where the three faster modes each represent some shifting dynamics. When we build the SRV-MSM for the GC-end sequence we observe a larger separation between the leading SRV-MSM timescale and the proceeding modes, which is shown in SI Figure 5) by a distinct spectral gap. We observe a second spectral gap after the fourth slow modes along with a drop in the model vamp-2 score (SI Figure 13). The clustered macrostates show a similar distribution where shifted populations are notably lower. The 3' double-shifted state is no longer identified as a metastable state, and the total number of states is reduced to five (again one more than the total number of slow modes). Although these states appear similar to the AT-all shifted states, the presence of the C:T or G:A mismatches replace two intact WC bonds in the corresponding AT-all states. This substantially diminishes the stability of the GC-end shifted states, resulting in lower state populations and shorter timescales.



Slow mode (Lag = 1.2 ns)	TICA-MSM (ns)	SRV-MSM (ns)
1	1180.3	1825.3
2	27.3	39.6
3	4.4	6.7
4	3.9	5.6

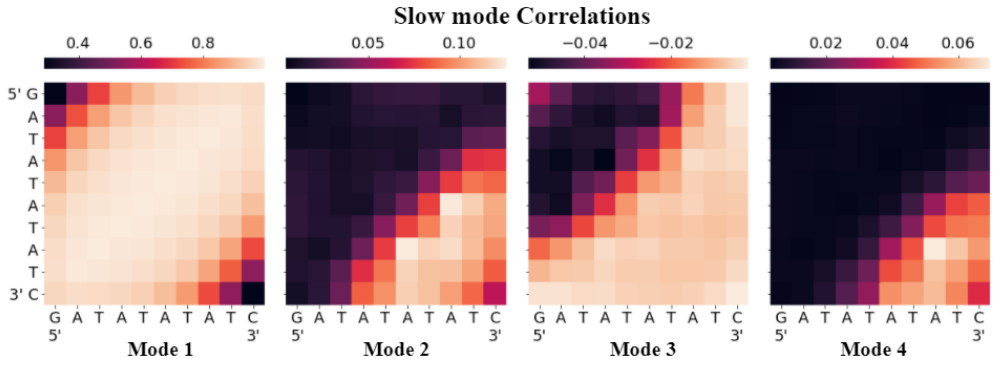


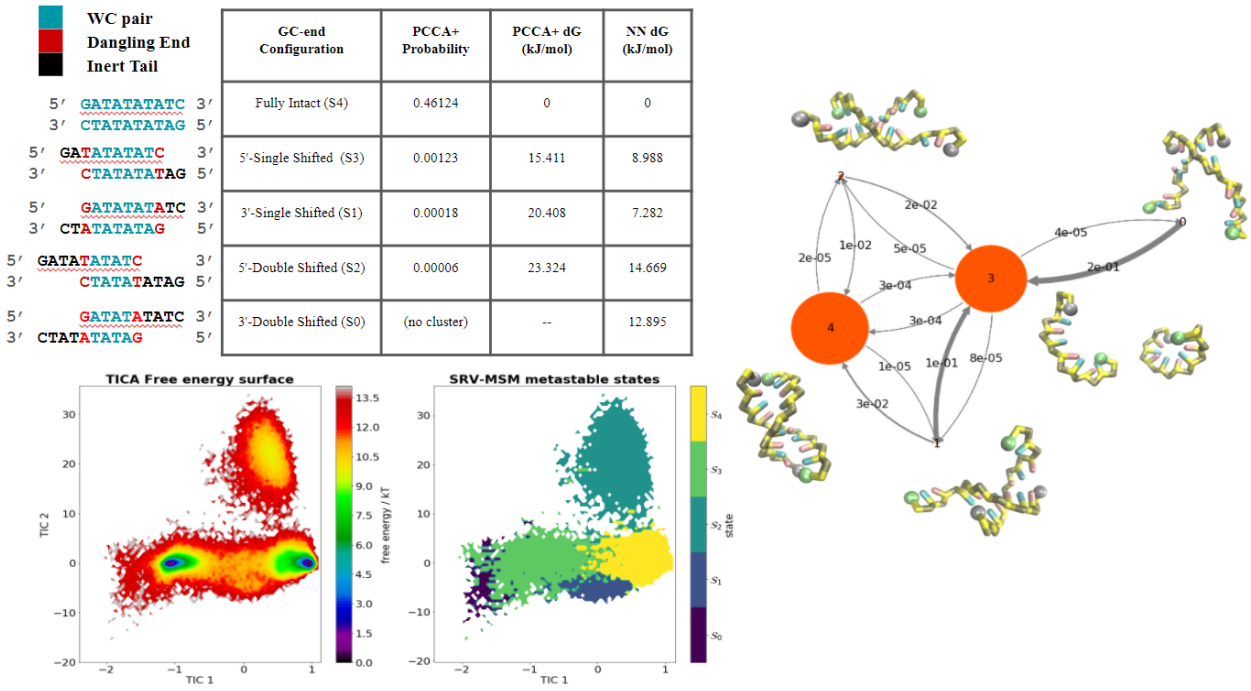
Figure 5: (SI) GC-end SRV-MSM timescales convergence and implied timescales using a lag time of 1.2 ns. Leading slow mode correlations to all 100 intermolecular distances between oligos.

3.2.2 Comparison between GC-end and AT-all shifted states

For our analysis, we consider C:T and G:A base pairs in the GC-ends shifted states as non-interacting dangling ends such that each shifted conformation has four total dangling ends and two inert tails. Although internal base pair mismatches can cause substantial conformational distortions such as kinking, terminal mismatches have been shown to be slightly stabilizing and have a minimal effect on helical character[68,70]. In the context of the 3spn.2 model, these ends are accounted for via intra-strand base stacking and inter-strand cross-stacking interactions[33]. The only direct interaction between non-WC basepairs is parameterized by isotropic excluded volume potential, which is likely more simplistic than the true mismatch interaction. Despite fewer intact base pairs, we find these structures stabilizing enough to account for the relatively small population of conformations we observed in these shifted states (Figure 6). Contrary to AT-all thermodynamics, NN calculations predict the 3' shifted states to be slightly more stable than 5' shifted states, again without taking into account entropic or enthalpic effects from inert tails. However, state populations reveal that the 5' single-shifted state is again significantly more stable than the 3', indicating that dangling end and inert tail effects outweigh nearest neighbor effects alone. As in the AT-all case, we observe a smaller deviation between 5' shifted theoretical and simulated free energies (6.5-8.7 kJ/mol) compared to the 3' case (13 kJ/mole). Interestingly, we found that 5' single-shifted stability might be elevated by a substantial portion of conformations retaining one intact GC base pair. Visualizations reveal that the shifted oligos – particularly in the 5' shifted state – have a tendency to sacrifice some helical conformational entropy in a way that facilitates G:C termini bonding even when all available A:T bonds are formed out-of-register.

3.2.3 Diffusion maps show sequence differences in shifted states

To compare these single-shifted 5' prime states with the corresponding AT-all state, we employ diffusion maps built on an equal sampling of 5000 conformations from the 5' single-



GC-end Configuration	PCCA+ Probability	PCCA+ dG (kJ/mol)	NN dG (kJ/mol)
Fully Intact (S4)	0.46124	0	0
5'-Single Shifted (S3)	0.00123	15.411	8.988
3'-Single Shifted (S1)	0.00018	20.408	7.282
5'-Double Shifted (S2)	0.00006	23.324	14.669
3'-Double Shifted (S0)	(no cluster)	--	12.895

Figure 6: GC-end free energies calculated from nearest neighbor data and from MSM PCCA+ probabilities b) TICA projections of free energies and PCCA+ state clusterings c) Transitions probabilities between states within the 1.2 ns lag time

shifted state of both sequences. Diffusion maps generate a low dimensional embedding based on some metric for diffusive distance and are well-suited to find subtle structural differences in temporally disconnected data[71,72]. 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 a degenerate 2nd and 3rd diffusion modes, with nearly equal eigenvalues, differentiating between looping 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 14). Diffusion maps built from samples of the 3' shifted states are also shown in the SI (Figure 15). 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 mismatch C:T-pairs are never bound – a consequence of the 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 confirmed by molecular renderings of 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.

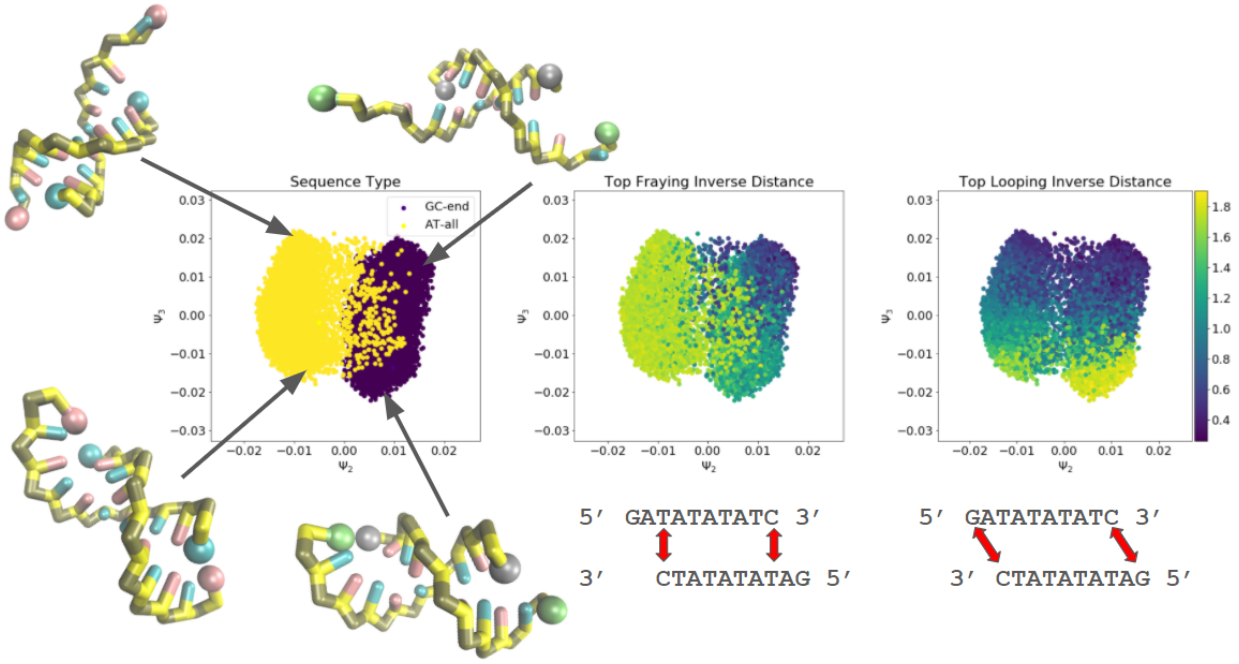


Figure 7: First two diffusion map coordinates built from 10000 single-shifted 5' states, equally sampled from AT-all and GC-end. Color maps show inverse distances between out-of-register ends and complementary ends.

3.2.4 GC-end dynamics are largely two-state

The kinetic behavior between GC-end macrostates is also distinct compared to AT-all. We do not observe any significant flux between the double-shifted to single-shifted 5' states. While this could be due to insufficient sampling, it is likely that the double-shifted states has too few contacts to directly transition into a single-shifted state without first dissociating. This is further demonstrated by the kinetic distance between this state and the aligned hybridized state (S0 and S4, respectively) shown in TICA projections (Figure 6). We observe some flux between the single-shifted states and aligned hybridized state, but these events are more rare than their AT-all equivalents. This suggests that although these transitions are possible routes for hybridization/dissociation, it is substantially more common for the transition to proceed in a two-state manner. In T-jump, FTIR, and 2D IR experiments, the GC-end sequence was observed to have less deviation from the two-state dissociation model compared to oligonucleotides with GC pairs closer to the core[11]. Furthermore, GC-ends revealed a distinct T-jump response containing evidence for the loss of G:C and A:T base pairing prior to full dissociation of the duplex[28]. Given that this response was slower than the A:T fraying response observed in other sequences, it is possible that it corresponds to shifting or related looping dynamics. This cannot be verified given that the signal was recorded in a congested spectroscopic range and could be hard to differentiate from aligned G:C fraying. We will discuss this more in our investigation of higher temperatures.

3.3 GC-core

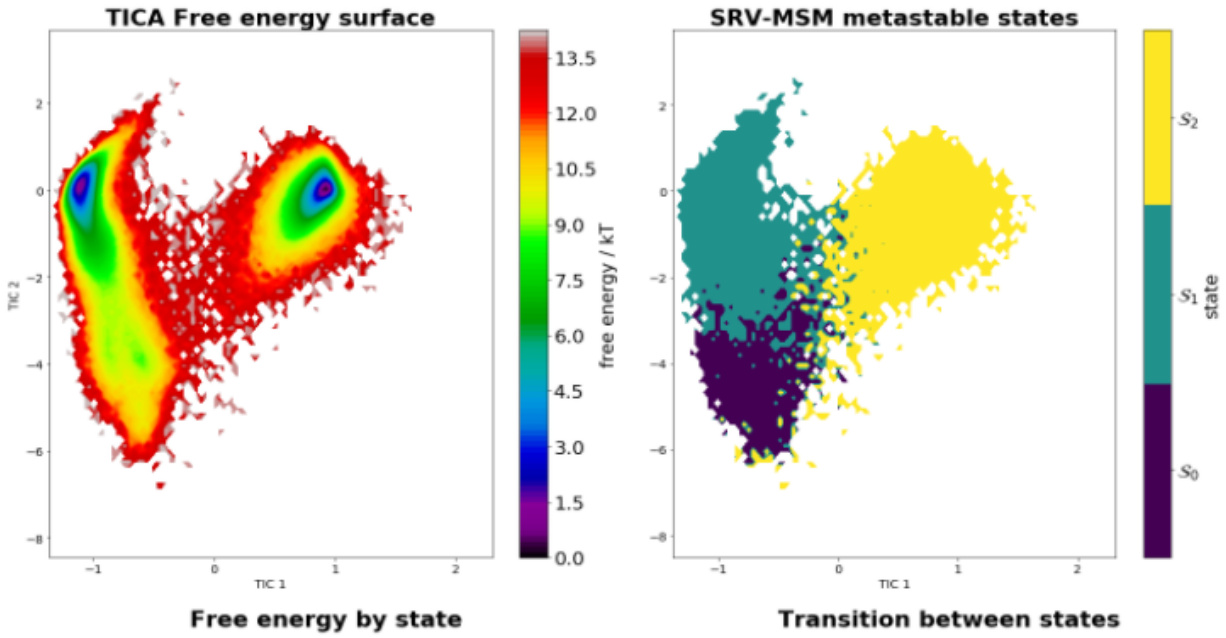
3.3.1 Building SRV-MSM

The GC-core sequence represents a departure from the dominant shifting dynamics observed for AT-all and GC-end. Based on SRV cross-validation (SI Figure 13), we selected the first three components for our analysis. Next, 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 3rd mode is mostly providing information about dissociation dynamics – so we performed PCCA+ clustering into three macrostates representing the hybridized, dissociated, and "4bp-frayed" states. Again, we found transition probabilities between states and visualized representative molecular renderings. We observe that the 4bp-frayed meta-stable state is solely composed of trajectories where both G:C core base pairs are bound and one of the adjacent A:T bonds are not. Previous work suggests that once key contacts are made, the zippering mechanism ensures that the helix will quickly form outward[8,32]. Our results indicate, however, that the relative instability of AT bonds compared the the GC-core can interrupt this process and form a longer lived frayed metastable states. 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 side of the helix. The MSM transition probabilities suggest that this is key intermediate state for both the hybridization and dissociation processes, however the pathways differ between directions. We find that the hybridized state has a 5x higher probability of transitioning into the frayed state within the lag time compared to transitions from the dissociated state. This is expected considering that this state is more accessible from an already bound helix. Moreover, once oligos are in this state, they are over 10x more likely to return to the hybridized state than to fully dissociate. Thus once a dissociated to 4bp-frayed transition has occurred, it is a likely to proceed into a fully hybridized conformation. On the other hand, transitions from the hybridized to 4bp-frayed state are much more frequent and are unlikely to proceed to a fully dissociated state.

3.3.2 Diffusion maps show ensemble of frayed states

To further examine this 4bp-frayed state, we build diffusion maps using 10000 frames sampled from the macrostate 9. 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 combi-



GC-core Configuration	Sim Prob	Sim G/kT
Dissociated (S2)	0.5805	0.544
Hybridized (S1)	0.4176	0.873
Frayed (S0)	0.0019	6.284

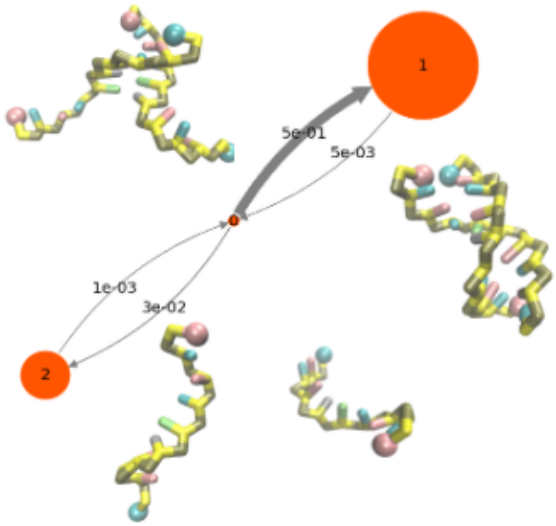


Figure 8: Full MSM output for GC-core

nation of physical coordinates that closely correlated to the first two non-trivial diffusion modes. We found that a larger distance between the third and fourth 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. In these conformations, one of the strands maintains some helical character while the other twist 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.

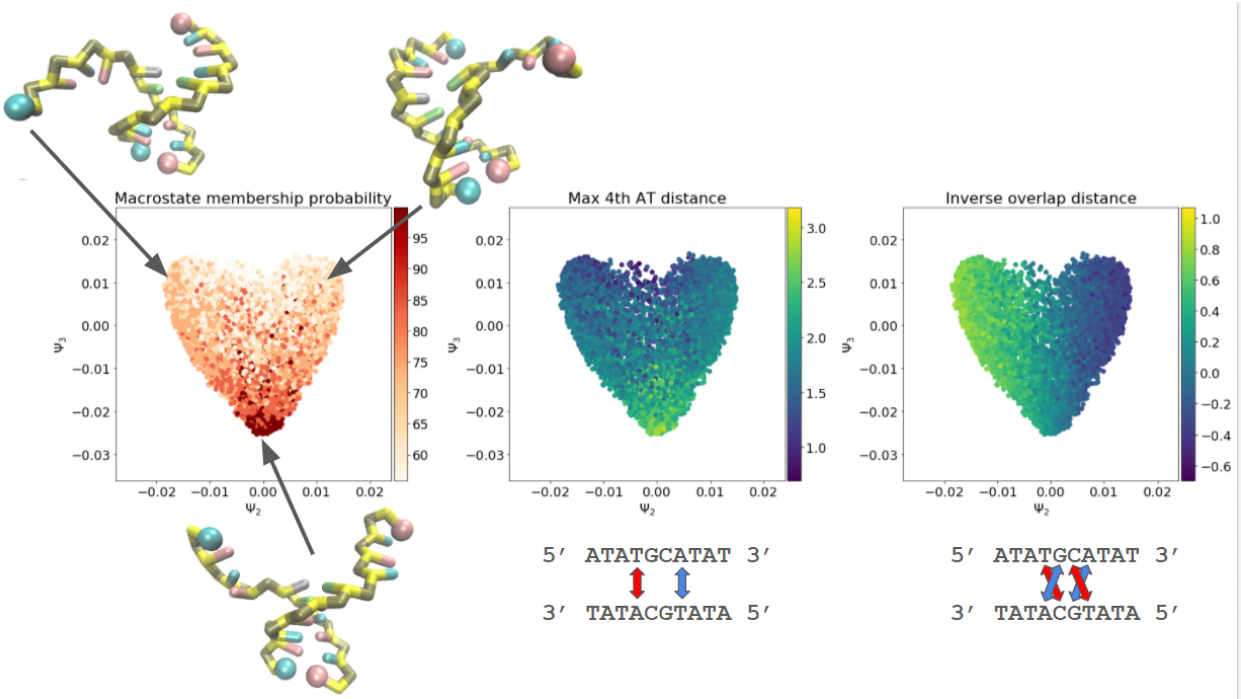


Figure 9: Plotting the first two non-trivial diffusion map eigenfunctions. Color maps show the probability that conformations are clustered into the frayed macrostate, maximum distance between 4th AT basepairs, and the the overlap between adjacent AT basepairs and the GC core.

3.3.3 SRV correlations to physical coordinates?

Having constructed and interpreted our SRV-MSM, we revisited our original SRV basis to analyze what physical correlations could be interpreted from the model. We found these

GC-core modes to be of particular interest as they reveal the hierarchical nature of the dynamical encoding. We examined a collection of 1000-frame "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 showed strong correlation. Two representative trajectories are shown in figure: (10). 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. This behavior reflects the above-mentioned kinetic trap between core binding and the fully hybridized state. 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. In addition to picking up on dissociation behavior, the third mode peaks when the oligos are close together but configured in such a way that is 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 strong qualitative trends we observe between physical coordinates and leading SRV modes, we only find high Pearson correlations for SRV1. For the next two SRV modes, the sign of their correlation switches depending on whether the oligos are in the hybridized or dissociated state. This shows that these modes are providing support on top of the first mode – which determines hybridization vs. dissociation – and thus can display very different behaviors in either state. With respect to our SRV-MSM macroscates, we found that SRV2 was "turned on" in states 0 and 1 – corresponding to the intact helix and frayed state, respectively – and SRV3 was turned on in the dissociated state (S2). Accordingly, we recalculated 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 other during 4bp-fraying. We also observed a highly symmetric correlation between SRV3 and central base pairs distances, indicative of overall diffusive behavior. Taken together, these analyses reveal how the SRV learns and represents the dynamical space, leading to the MSM results we observe above.

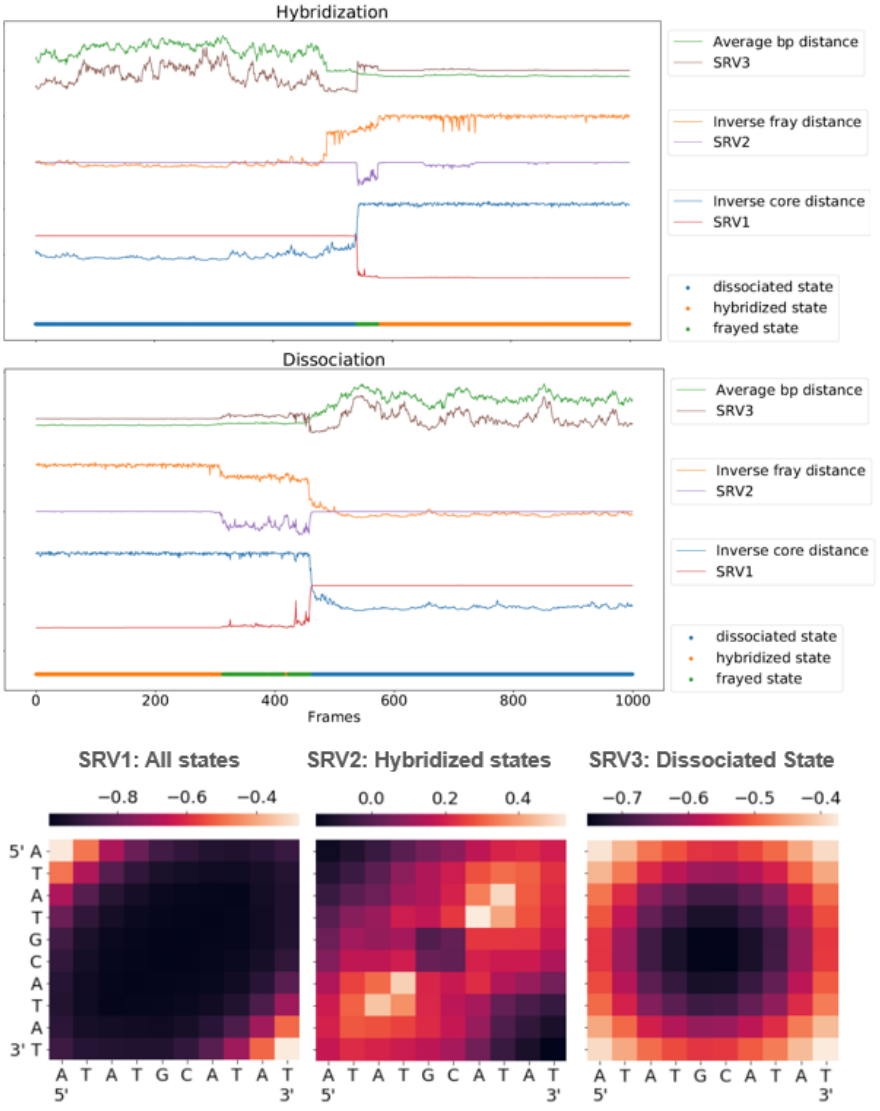


Figure 10: Show how SRV coordinates correspond to physical features over time and during sample dissociation and hybridization events. Should probably go to SI if we also include GC-core diffusion maps.

3.3.4 Comparison to Experiment

When examining these same 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[11]. 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[73]. 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[28]. 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 formation of the the A:T bond nearest to the GC center. Our diffusion map analysis further shows how the ensemble of 4-bp frayed configurations impede helix formation. 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.

3.4 GC-mix

The GC-mix sequence shows a similar implied timescale distribution to GC-core, however we no longer see 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, however we do observe substantial fraying of the two AT termini in the simulation data. Although

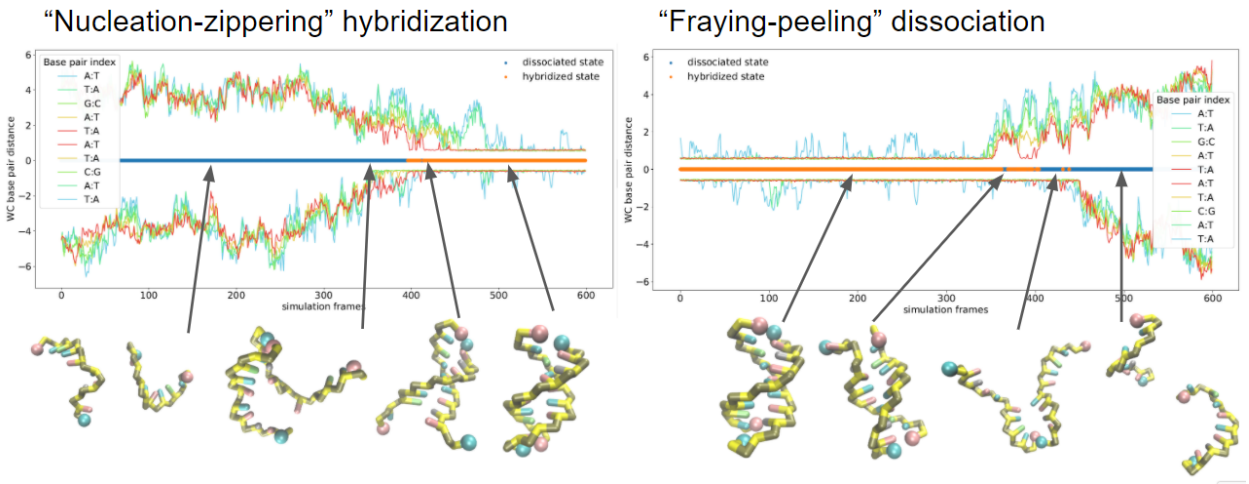


Figure 11: WC base pair distance and molecular renderings along two representative GC-mix hybridization and dissociation events.

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[11]. 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. On the other hand, the 4bp-frayed state in GC-core had a substantial probability (10%) of fully dissociated rather reforming an intact helix. 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 (11). During hybridization, we observe the formations of some key base pair contacts before the full duplex forms. We observe that first contacts tend to involve one of the G:C bonds, which is not surprising that these are more evenly spaced out than in GC-core and GC-end. This behavior is indicative of a nucleation-zippering mechanism as has been reported in previous studies[8,12]. We obverse dissociation events proceeded

by two base pair fraying on one side of the duplex followed by more rapid dissociation of the central base pairs. There tends to be a short-lived state composed of 2-4 base pair contacts shortly before fully dissociation occurs. In contrast to the 4-bps frayed 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[20,38,39]. 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.

3.4.1 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 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[28]. Although there was no experimental data for AT-all and GC-mix, we used our relaxation fits to predict how these might compare with the other sequences. We measure the slow dissociation response by fitting the distribution of times at which the core base pairs separate beyond a 2 nm cutoff. We find 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.

The fast response – which Sanstead et. al attributed to base pair fraying signatures – was more difficult to compare given the limits our coarse-grained model. Numerous experimental

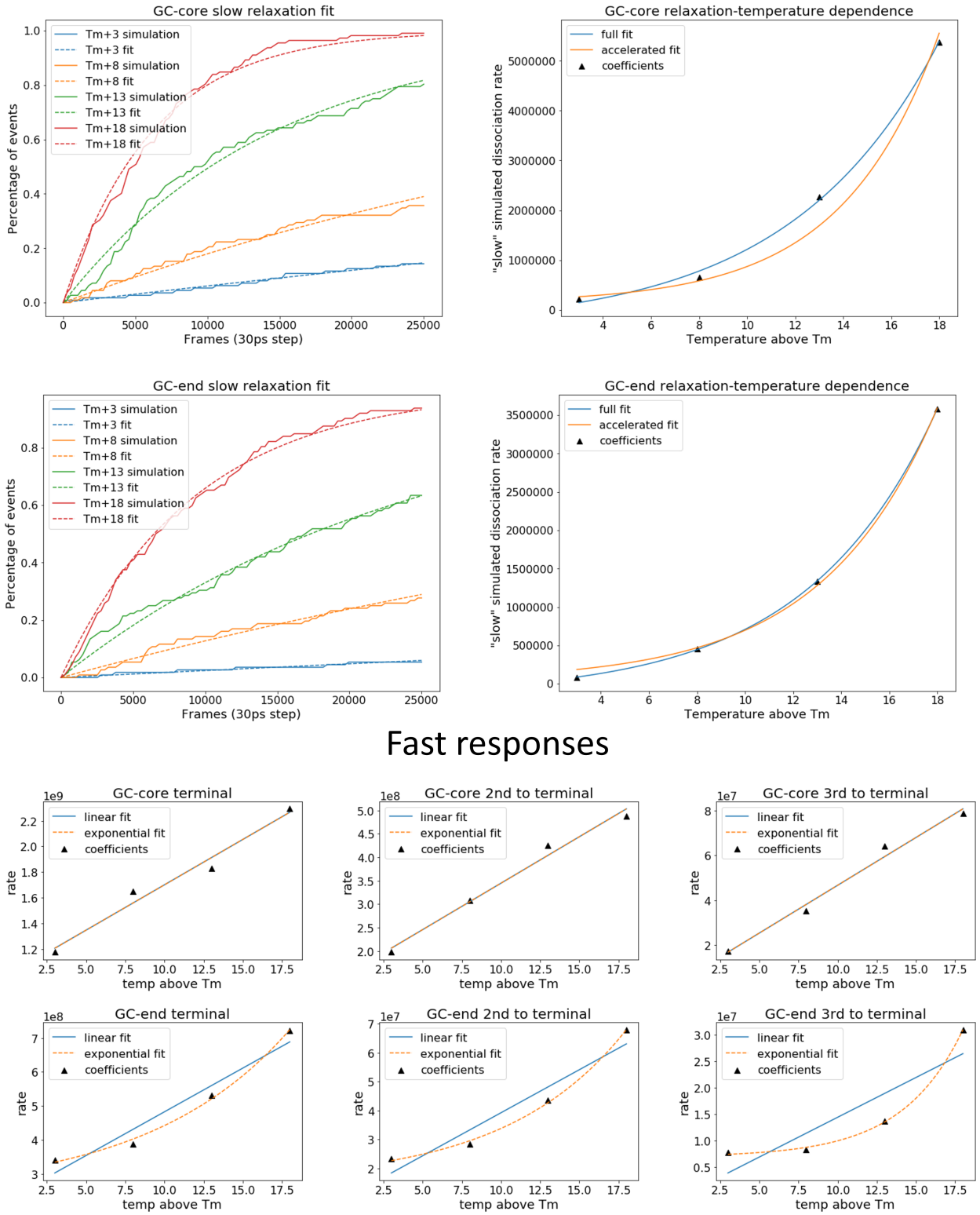


Figure 12: Fitting temperature-dependent trends to the "slow" dissociation mode and "fast" fraying mode detected in experiments. Both show an exponential relationship between dissociation and temperature, however fraying appears more linear for GC-core than for GC-end.

and computational studies have shown that DNA and RNA fraying is a complex dynamical process with timescales that span 5 ps to several microseconds[21–23,74]. Furthermore, all-atom simulations suggest that frayed ends can assume misaligned WC bonds, base-sugar hydrogen binding, and terminal stacked conformations[20,63]. 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 measure the fast fraying response by observing how long it takes for either duplex terminal end to split beyond a given 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. For further comparison, we include fits over a range of cutoffs and repeat these calculations on the second and third interior base pairs.

For GC-core, each base pair and cutoff combination appears linear with temperature, indicating a barrierless and diffusion-driven process. As expected, the acceleration factor is also highly dependent on the choice of parameters, but, given that terminal fraying likely dominates the signal, we approximate that fraying dynamics are accelerated by about two order of magnitude relative to experiments. In contrast, GC-end is distinctly exponential, especially at base pairs closer to the core of the duplex. This behavior makes sense given the thermodynamic barrier to G:C fraying, but it also deviates the GC-end experimental linear fit. In addition to fraying, we notice that the rate of shifting increases at higher temperature. Moreover, the thermodynamic stability of these shifted conformations increases relative to that of intact duplex. For T-jumps at higher initial and final temperatures, it is possible that some portion of the starting ensemble is in a shifted state (cite Ryan's work) and that the fast dynamics are increasingly influenced by alternative "looping" G:C bonding that we observed in our diffusion map analysis above. It is less likely that shifting mechanisms would have their own distinct signals, given that these occur at far lower frequency than fraying or full-dissociation, but we believe that further experimental work is necessary to investigate

this phenomenon.

The temperature-dependent analysis can also inform out interpretation of the 3spn2 model. Previous work has shows that kinetic association rates were accelerated by about one to two orders of magnitude relative to experiment[33]. It is not surprising that different dynamics might be accelerated at different rates, especially considering that 3spn2 was not parameterized based on any dynamical properties[?] . Furthermore, more data is needed for explicit ion model comparisons.

4 Conclusion

4.1 Will look a lot like the last paragraph of the intro/abstract

4.1.1 Limitations

1. Does not explicitly pick up on nucleation/zippering events. Because these are all unique sites (and zippering dynamics could vary at each site) it would be difficult to properly sample and interpret the unique corresponding modes. Furthermore these zippering dynamics occur significantly faster than shifiting or GC-core fraying and cannot be captured via our save rate.

2. Coarse grain-models makes it difficult to compare timescales. Different degrees of freedom might have different accelerations.

3. Limited by memory (number of frames) required to capture equilibrium trajectories and rarity of hybridization/dissociation events

4. A smoother free energy surface minimizes the time spent in these intermediate states

5. Sensitive to 3spn2 artefacts. Shifting may be overexpressed (and is in fact suppressed in[32], but similar results have been shown in all-atom studies as well[9]. Base-pair mismatch interactions may be somewhat unphysical and alter state populations. Although

we present shifting as a possible explanation for GC-end signal, there has been no conclusive evidence of these dynamics outside out all-atom and coarse-grained simulation studies.

4.2 SI figs

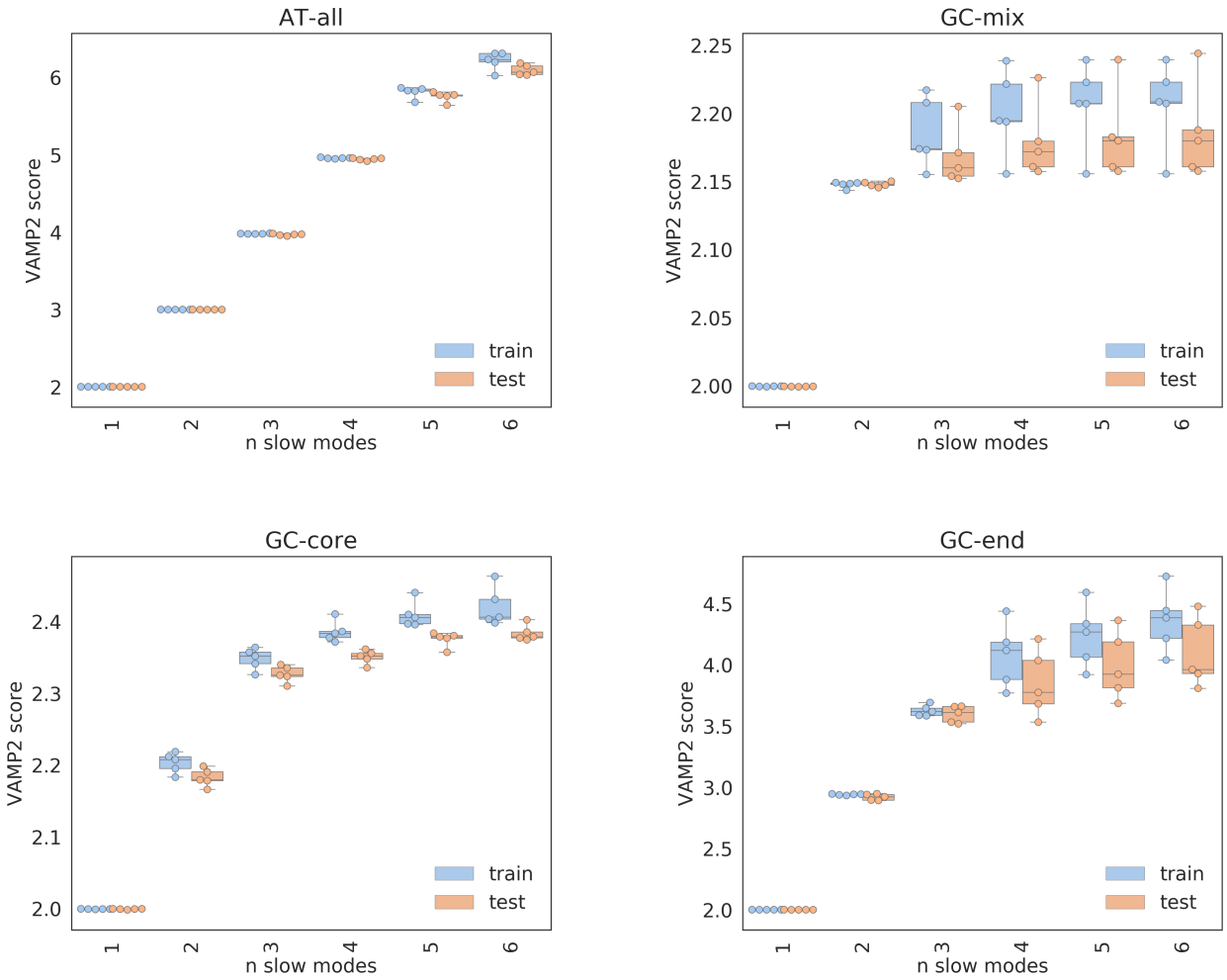


Figure 13: 5-fold cross validation procedure to select number of SRV coordinates. We look for both the converge of the VAMP-2 and consistent scores between folds to ensure that the model is not fitting on statistical noise. (probably will be SI, but could include AT-all cross-val in the walkthrough section)

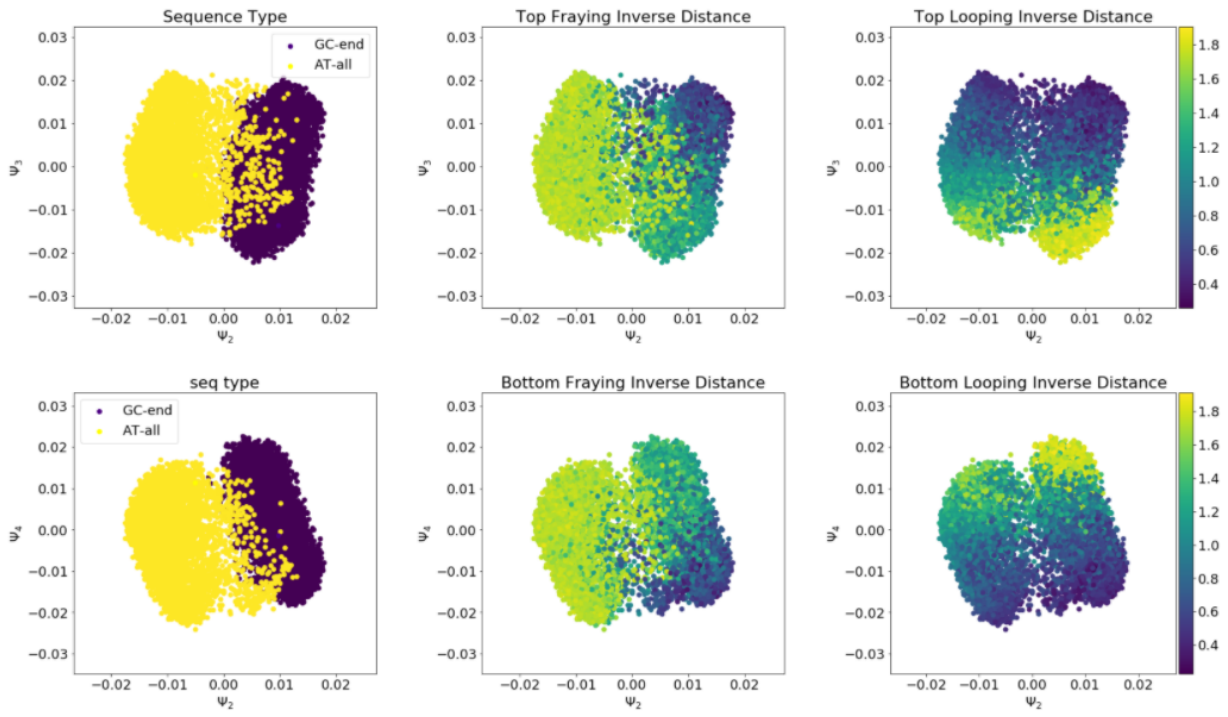


Figure 14: Full 5' diffusion maps including degenerate third diffusion map mode

4.2.1 Code to provide on GitHub

1. Cross val data
2. lammps run script and each sequence input files
3. Featurization with mdtraj (reindexing and sparse saving)
4. Dmaps for GC-core and GC-end
5. timescales comparison
6. SRV cross valss
7. MSM construction

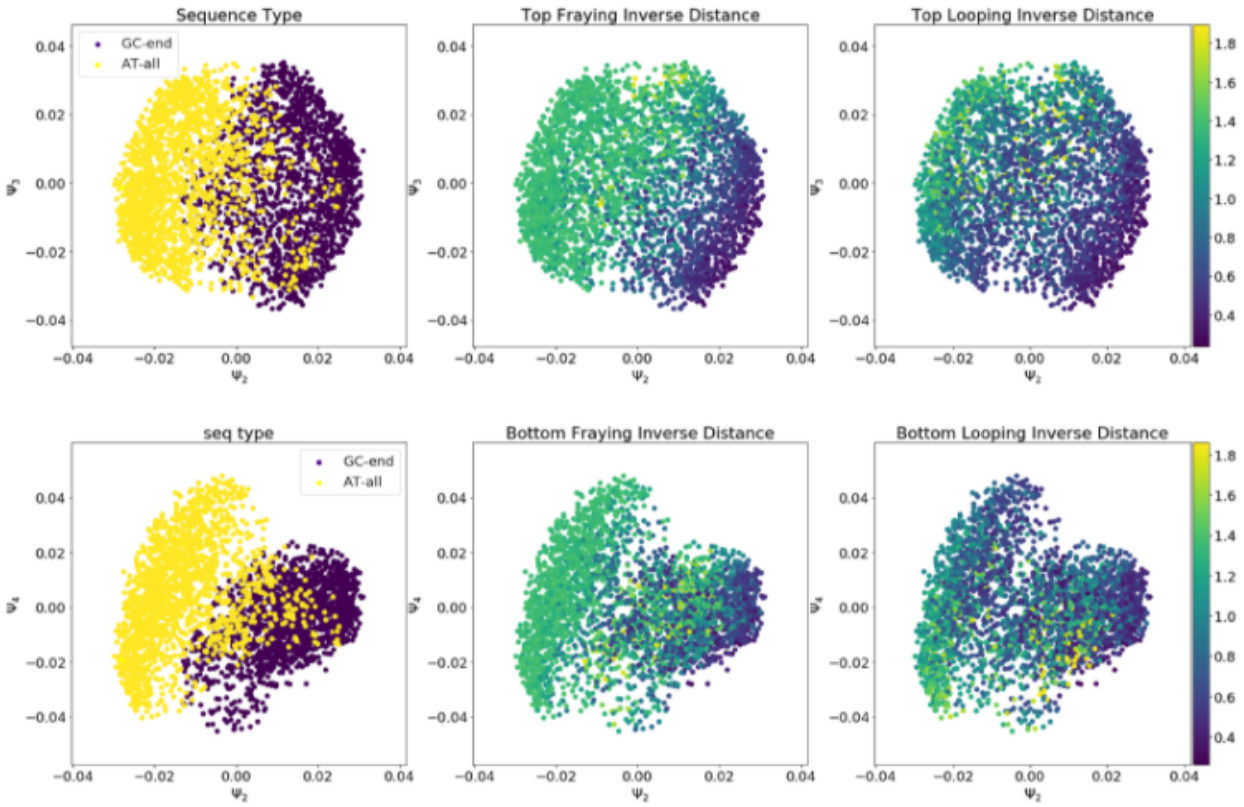


Figure 15: Full 3' diffusion maps, shows greater overlap between the AT-all and GC-end populations, as well as a less distinct "looping" region for intact GC bonds

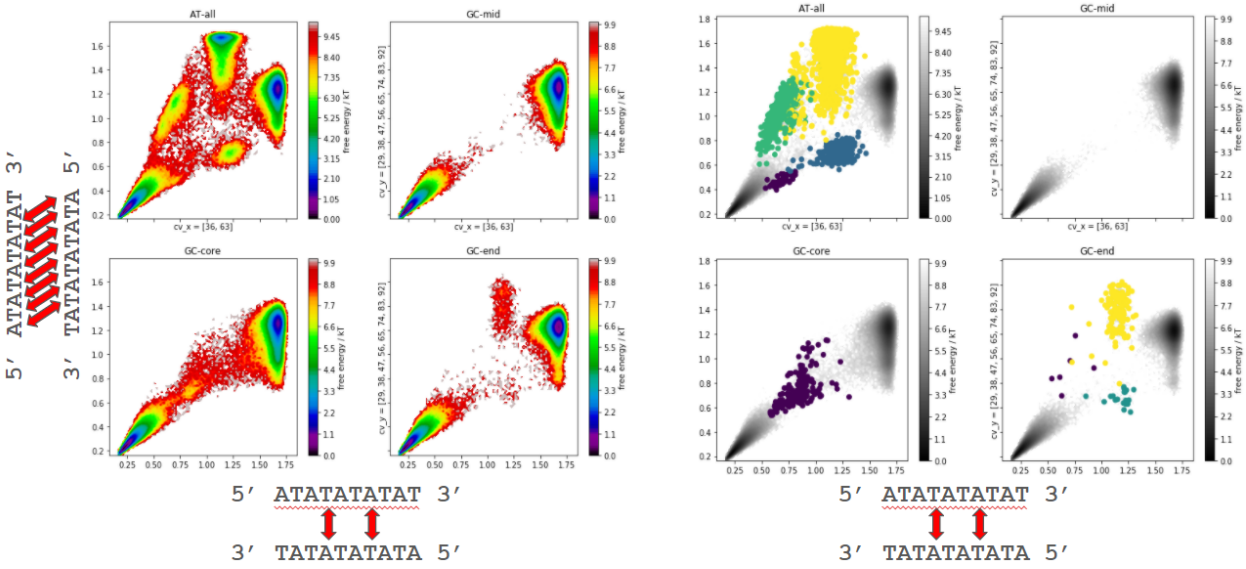


Figure 16: All sequences and metastable macrostates plotted on frayed/shifted collective variable axes.

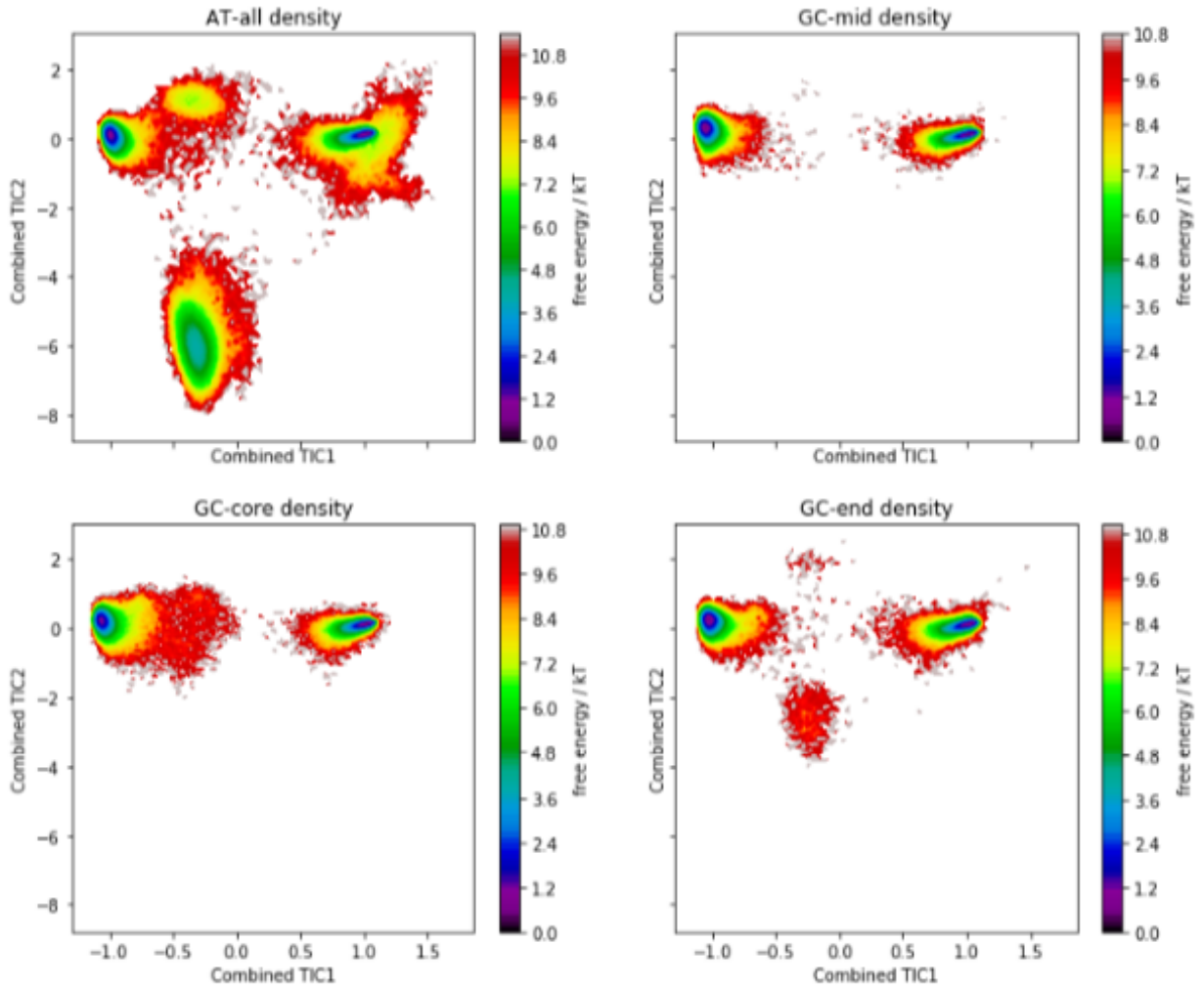


Figure 17: All sequences plotted on combined TICA axes. Might include in SI or not at all.

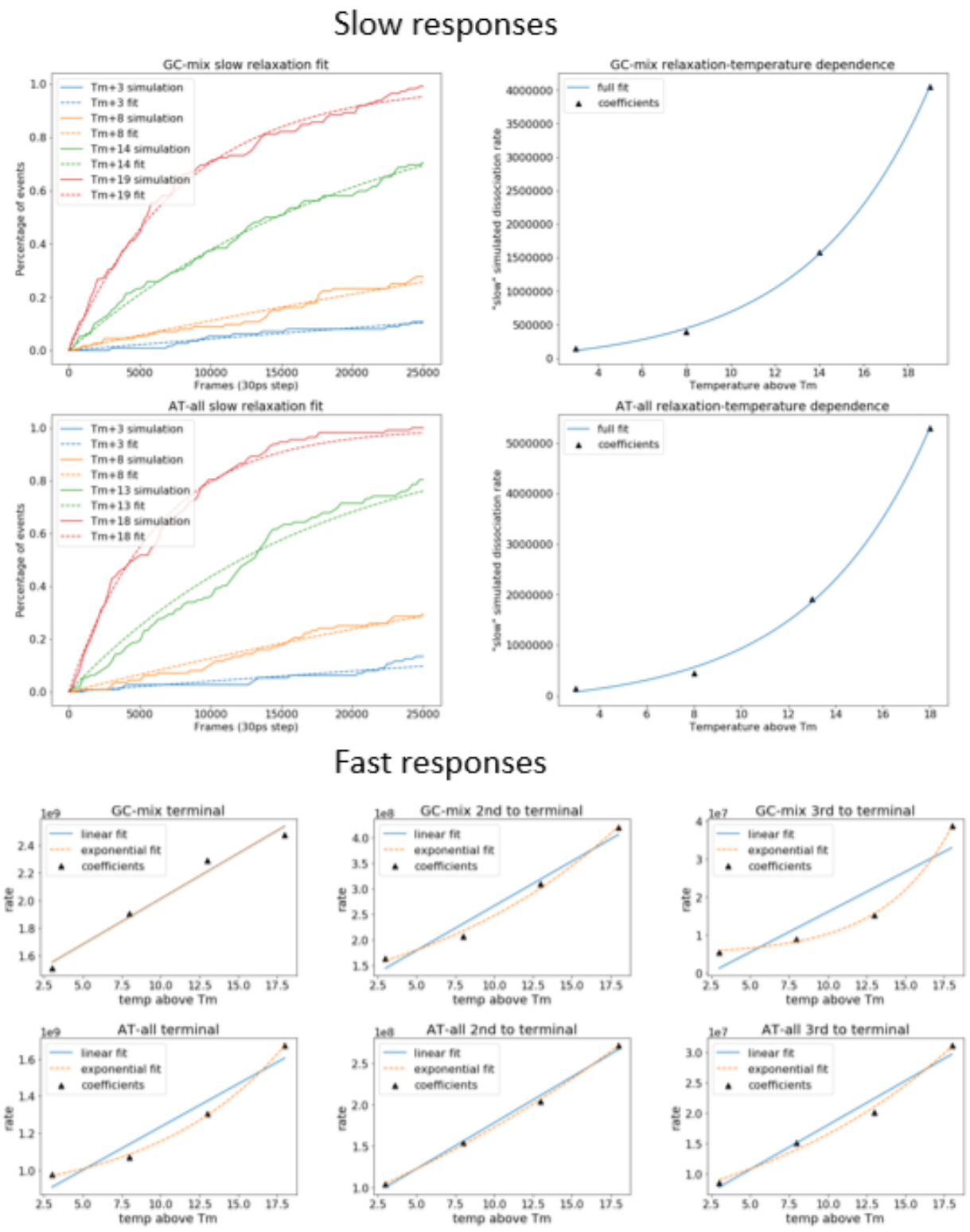


Figure 18: Slow and Fast fits for GC-mix and AT-all, no acceleration values were calculated as there was not the same temperature-dependent data available for these sequences. GC-mix again shows barrierles fraying at the terminus, with a noticeable barrier for G:C bonds breaking at the second base pair in. AT-all shows a more exponential relationship for terminal fraying, likely because the simulations are run at lower relative temperatures.

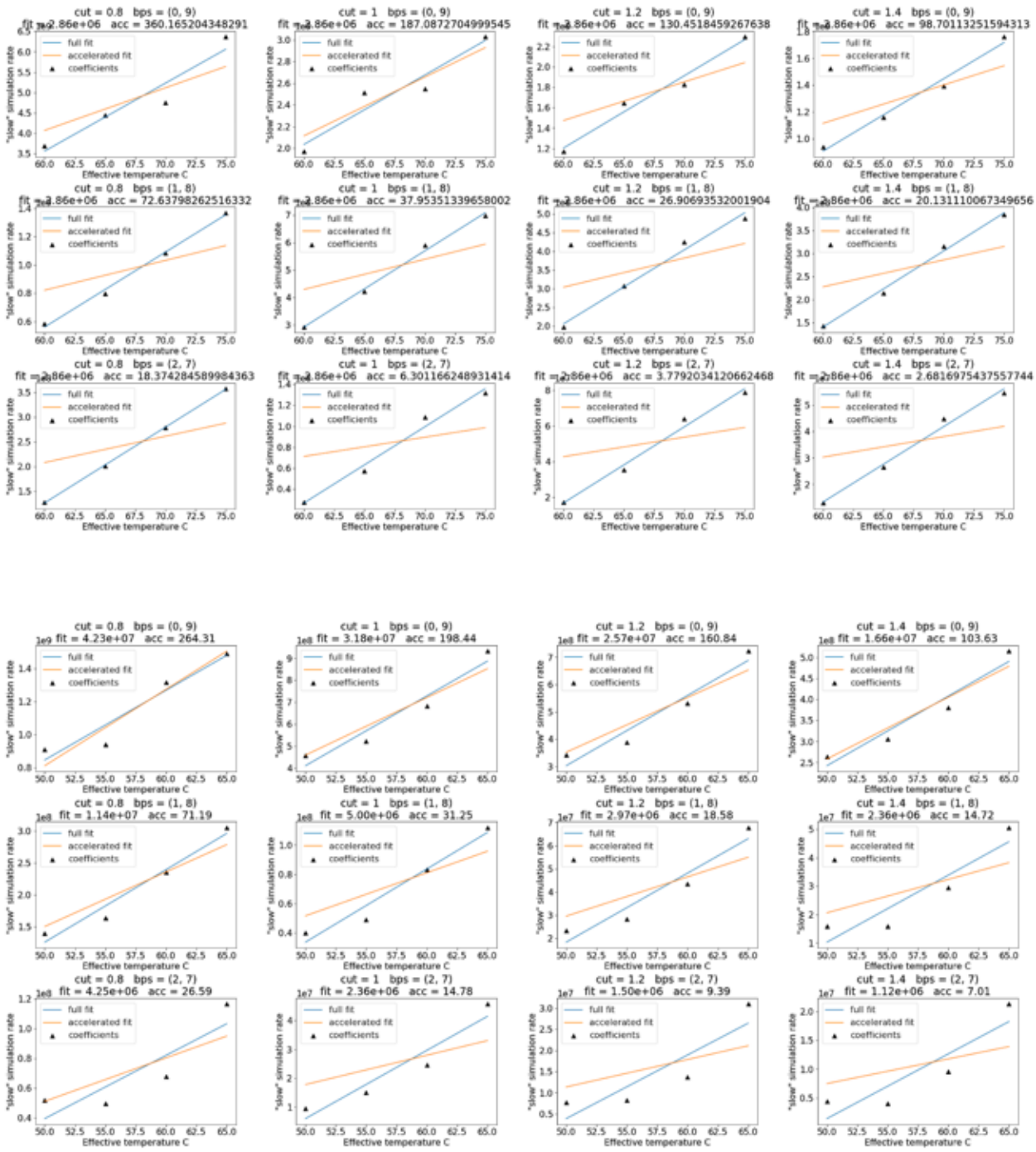


Figure 19: Fits across multiple cutoffs and basepairs for GC-core and GC-end fast response. (This is a placeholder until I reformat these to match the main text).

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) Mhatre V. Ho, J.-A. L.; Martin, K. C. NIH Public Access. *Bone* **2012**, *23*, 1–7.

(6) Bui, H.; Shah, S.; Mokhtar, R.; Song, T.; Garg, S.; Reif, J. Localized DNA Hybridization Chain Reactions on DNA Origami. *ACS Nano* **2018**, *12*, 1146–1155.

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

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

(9) 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.

(10) 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*.

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

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

(13) 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.

(14) 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.

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

(16) 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*.

(17) 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.

(18) 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*.

(19) 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.

(20) 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.

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

(22) 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.

(23) 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.

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

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

(26) 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.

(27) 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.

(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) 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.

(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) 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.

(35) 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*.

(36) 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.

(37) 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.

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

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

(40) 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.

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

(42) 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**,

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

(44) 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.

(45) 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.

(46) 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.

(47) Lequieu, J.; Córdoba, A.; Moller, J.; De Pablo, J. J. 1CPN: A coarse-grained multi-scale model of chromatin. *Journal of Chemical Physics* **2019**, *150*.

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

(49) 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.

(50) 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.

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

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

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

(54) Molgedey, L.; Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters* **1994**, *72*, 3634–3637.

(55) Harrigan, M. P.; Pande, V. S. Landmark Kernel tICA for Conformational Dynamics. **2017**,

(56) 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.

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

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

(59) 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**,

(60) 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**,

(61) 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*.

(62) 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.

(63) 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*.

(64) 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.

(65) Senior, M.; Jones, R. A.; Breslauer, K. J. Influence of Dangling Thymidine Residues on the Stability and Structure of Two DNA Duplexes. *Biochemistry* **1988**, *27*, 3879–3885.

(66) Dickman, R.; Manyanga, F.; Brewood, G. P.; Fish, D. J.; Fish, C. A.; Summers, C.; Horne, M. T.; Benight, A. S. Thermodynamic contributions of 5âĂŹ- and 3âĂŹ-single

strand dangling-ends to the stability of short duplex DNAs. *Journal of Biophysical Chemistry* **2012**, *03*, 1–15.

(67) 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.

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

(69) 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.

(70) Di Michele, L.; Mognetti, B. M.; Yanagishima, T.; Varilly, P.; Ruff, Z.; Frenkel, D.; Eiser, E. Effect of inert tails on the thermodynamics of DNA hybridization. *Journal of the American Chemical Society* **2014**, *136*, 6538–6541.

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

(72) 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.

(73) 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*.

(74) 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.