TBA

Mike Jones and Andrew L. Ferguson*

Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637

E-mail: andrewferguson@uchicago.edu

Abstract

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 applications[2–5]. Recently, structural DNA nanotechnology has enabled self-assembly on micro to milli scales, and dynamic DNA nanotechnology has been used to perform basic calculation and to probe single molecules via temporal DNA signatures[6–8]. [[Please correct reference "MhatreV.HoJi-AnnLee2012NIHAccess"]] 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[9,10]. 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[11,12]. 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 "slipped" base pairing in repetitive sequences, has been well documented in previous computational studies[1,10]). Furthermore, Sandstead et al suggested the existence of frayed metastable states during duplex dissociation, where the stability of these states was dictated by G:C bond placement in the 10-mer oligonucleotide sequence[13]. In this work, we study the same four sequences explored by Sandstead et al in an effort to uncover the sequence-dependent dynamics and their relation to metastable structures mentioned above.

Given the long timescales on which DNA hybridization and dissociation events occur, the study of these processes are generally not amenable to direct simulation techniques[1]. Instead,

many previous studies of DNA hybridization have employed accelerated sampling methods such as umbrella sampling[14] transition path sampling[15], and forward flux sampling[14–17]. Other computational studies use dramatically elevated temperature or denaturing solvent concentrations induce one-way dissociation events[18,19]. Experimental studies often employ temperature jump, concentration jumps, or other perturbative methods to drive DNA out of equilibrium and monitor relaxation processes in one direction[20,21]. Single molecule diffusion and tethered multifluorophore assays facilitate equilibrium analysis, however these present technical difficulties and are hampered by long timescales and data collection rates[11,22]. FoÌĹrster resonance energy transfer (FRET) analysis, particularly when coupled with methods mentioned above, provides additional resolution, but it unclear how fluorescent tags may interfere with the dynamics of short nucleotides[20].

In this work, we use the coarse-grain 3 site per nucleotide (3spn2) model to simulate hybridization and dissociation behavior at each sequence's melting temperature[23]. We leverage the properties of Markov State Models (MSMs) – namely that conditional probability depends only on the current state of the system[24] – 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[10,25], but the slowest sequence-dependent kinetics were not the focus of these studies. Pinamonti et al.[26] used MSMs to compare the slowest dynamics of short RNA nucleotides and found that stacking timescales are highly sequence dependent[26]. We take a similar approach to study 10-mer DNA oligonucleotides and introduce State Reversible Vampnets (SRVs) to directly learn the slowest sequence-dependent dynamical modes[27]. 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 the hybridization and dissociation process. Furthermore we can perform these analyses without biasing simulations or assuming that one processes is a strictly reversible version of the another. Because SRVs generate an optimized low dimensional basis, we show that we can access higher resolution MSMs (shorter 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. Taken together, our analysis reflects similar results to previous computational and experimental DNA work, while elucidating new insights into sequences dependent dynamics, meta-stable structures, and relative timescales.

2 Methods

2.1 3spn2 Model

Is some additional information warranted on the model itself before going into this specific setup?

2.2 Simulation set up

We initialized four sequences previously investigated by Sandstead 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[1,13]. We initialized explicit ions such that 240 mM NaCl and 18 mM MgCl2 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[28]. In order to maximize concentration without al-

lowing 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 A). 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 maintained a Debye-Huckel screening potential with an ionic strength of 240 mM – corresponding to NaCl concentration – to account for phosphate backbone interactions. These electrostatics preserve the persistence length and intrinsic curvature of DNA while taking the effects of ion-DNA interactions[28]. We run our simulations in the NVE ensemble and fix temperature via a Langevin thermostat in order to model implicit solvent interactions[29]. Simulation temperature was determined by sequence specific melting temperature such that strands were equally likely to spend time in the hybridized or dissociated state. Tmelt simulations were run for $3*10^8$ time steps, equating to 4.5μs simulation time, and frames were saved every 30 ps. A 15 fs timestep was used for all simulations runs. 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.

2.3 Featurization

All intermolecular pairwise distances for both oligonucleotides were calculated at each frame using the MDtraj software package[30]. Based on the self-complementary nature of each sequence, we averaged permutable distances (45 pairs in total) together. This permutation reduction follows a similar procedure used in TICAgg coordinates construction[31]. The VAMP-2 scoring method, which represents the sum of squared estimates for the transfer operator, was employed to evaluate the quality of this feature set[32]. We compared the VAMP-2

score for the permutation-free distances to the complete set of intermolecular distances and found only a small loss in kinetic variance when using the reduced data set. Furthermore, we found an increase in VAMP-2 score when using reciprocal pairwise distances and chose these reciprocal permutation-free coordinates as a consistent feature set. These features were normalized and passed into sequence-specific SRVs.

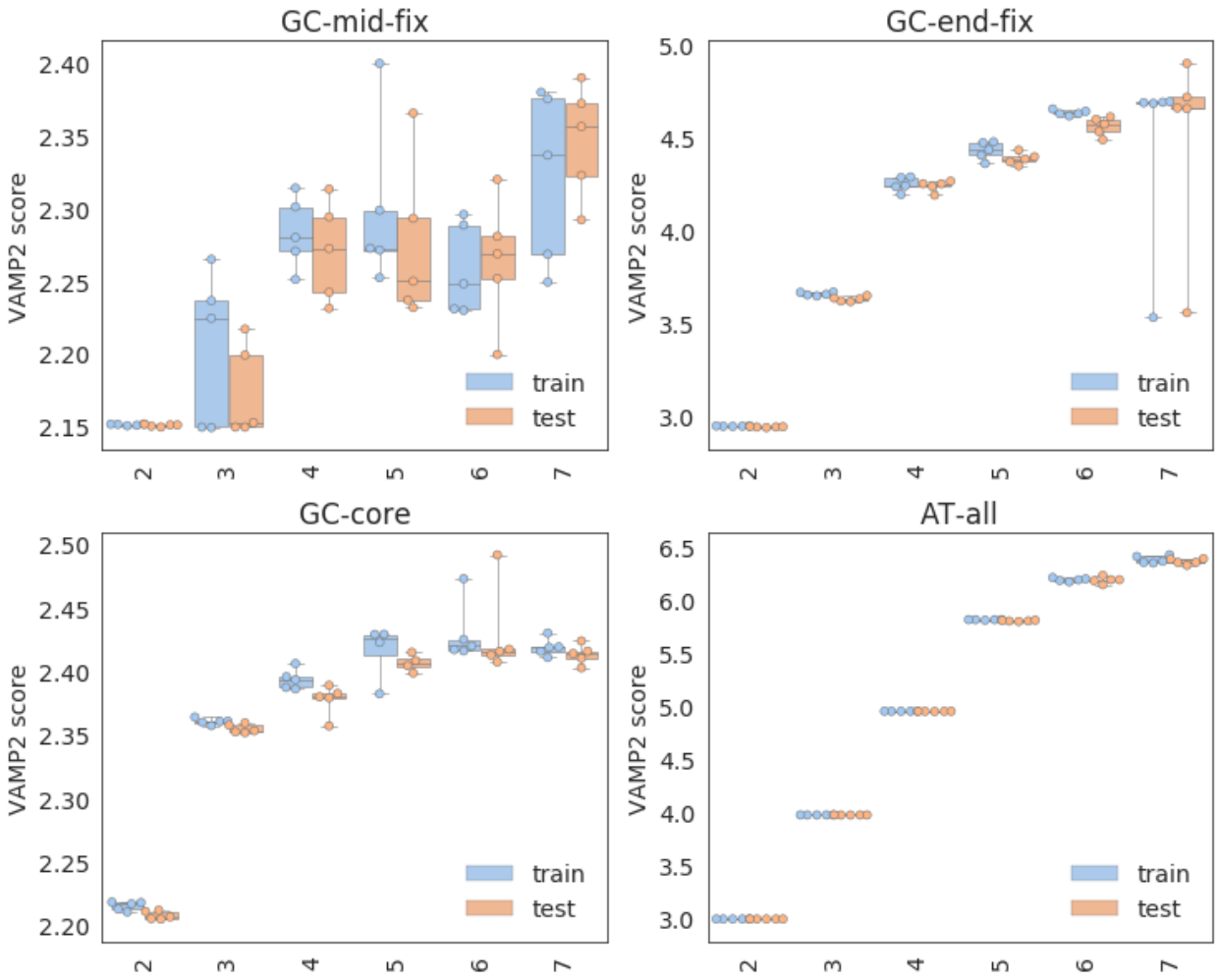


Figure 1: Cross val procedure to selection number of SRV coordinates (probably will be SI, but could include AT-all cross-val in the walkthrough section)

2.3.1 SRVs

SRVs were first developed by Chen et al. as a means to directly learn slow eigenfunctions of the transfer operator[27]. The framework uses a twin-lobed artificial neural network, similar in structure to the network employed in VAMPnets[32], 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 can be estimated (equation here?). 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 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[33]. SRVs provide robust nonlinear approximation and computation timescales 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 as input data. 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[27,34]. 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 generated a hierarchical dynamic encoder (HDE) to transform 55 reciprocal pairwise distances into a low dimensional SRV feature 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 mode via cross-validation on the VAMP-2 score to ensure that the coordinate did not over fit on statistical noise[35]. In particular we looked for a diminishing increase in the vamp-2 score, inconsistency between cross validation scores – suggesting that the model maybe learning different modes, and a widening gap

between train and test scores – suggesting that the model might be overfitting on statistical noise in the training data. We used a batch size of 50000 and ran each model for 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[36,37].

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[24]. 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)[34]. Furthermore, SRV-MSM implied 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[38]. After passing in SRV coordinates, we perform k-means microstate clustering, Bayesian MSM construction, and PCCA+ 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

- lead with same kind of diagram for each sequence, add additional figures when necessary for a given sequence

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[39]. We found that most implied timescales (with the exception of the leading timescale) converge at a lag time of 600 ns. We kept a looser constraint on the convergence of the leading mode as it reflects the less Markovian dynamic of strand association and dissociation. 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[35]. It is also more difficult to perform kmeans clustering on a high dimensional space, especially when those higher dimensions are less kinetically relevant[24]. For these reasons, the number of slow SRV components should be carefully selected based on the specific system of interest. As shown in (Figure 1), we see a clear convergence of 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 dynamics of the overall hybridization and dissociation process. The next four SRV components all show a relatively high correlation along different offset diagonals. These diagonals correspond to the intermolecular distances between shifted base pairs and point to the existence of gradually faster shifting or "slithering" processes. Slow slithering mechanisms have similarly been reported in simulation studies to occur on orders of magnitude longer timescales than underlying fast dynamics[40],[10].

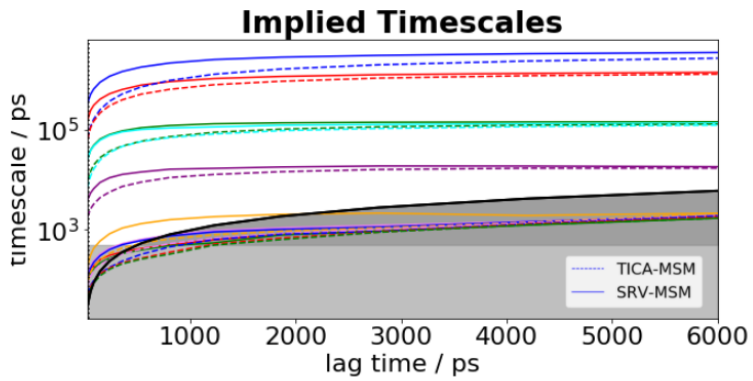
3.1.2 SRV-MSM construction and optimization

From this analysis we can determine that the slowest dynamics are fully characterized by the hybridization process and slithering 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 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 2 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. Using an MSM lag time of 1.2 ns, we then 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 represent a collection of metastable structures. Previous work 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)[41],[42]. 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 represent good high variance collective variables on which to visualize free energies and state assignments[34]. It is clear in Figure 3 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"[10],[1]. We found a substantial difference in thermodynamic stability and



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

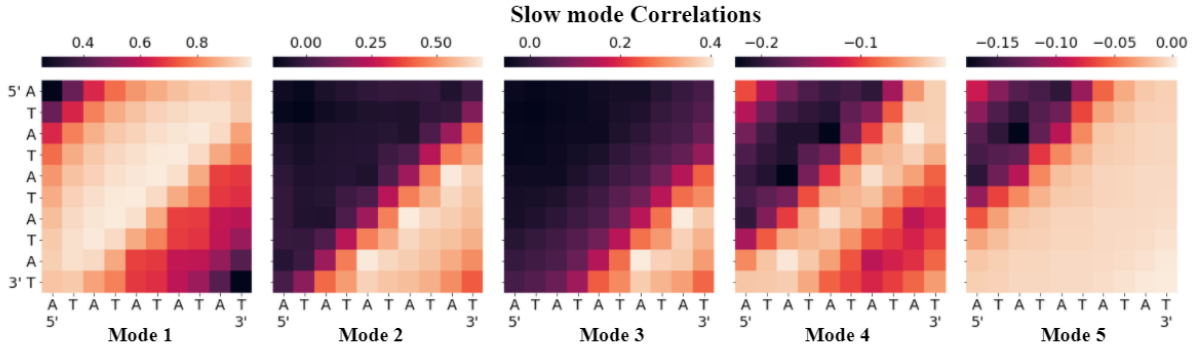
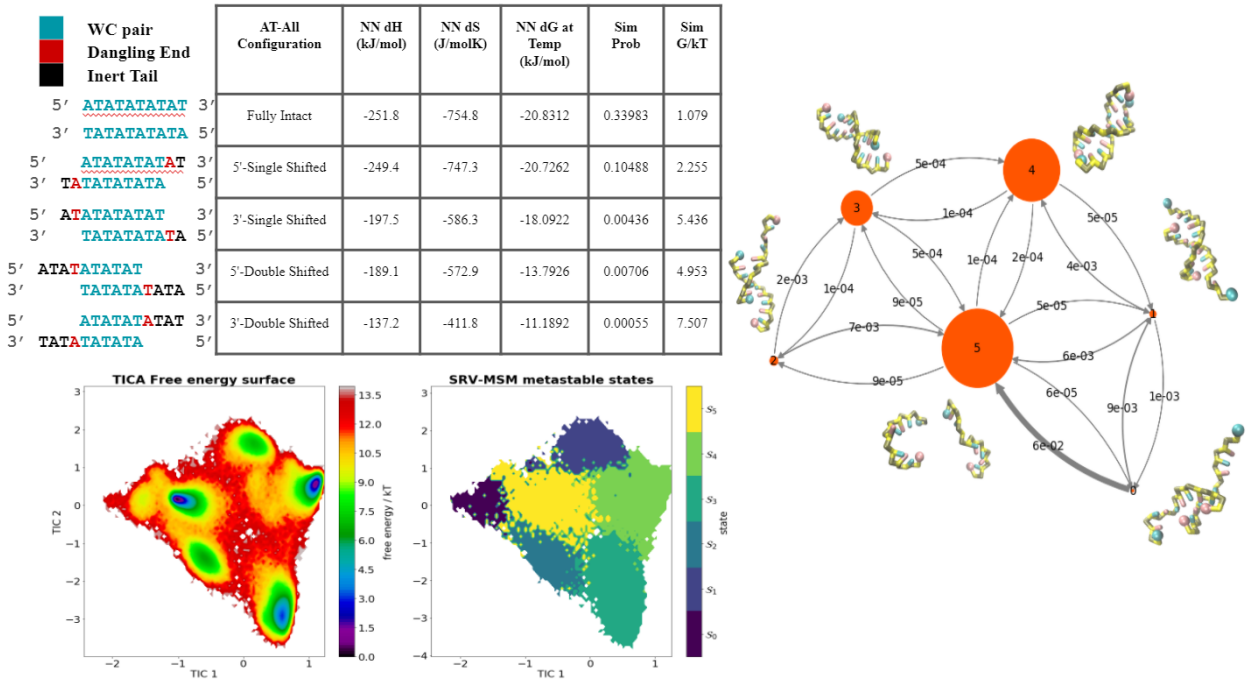


Figure 2: AT-all SRV-MSM timescales convergence and implied timescales using a lag time of 1.2 ns. Leading slow mode correlations to all 100 intermolecular distances between oligos.

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 extend beyond the dangling end[43]. Dangling 5' ends have a consistently stabilizing effect, and inert tails decrease stability as they increase in length. It has been shown that two base pair 5' dangling ends (or dangling ends with one inert tail) have higher melting temperature and are enthalpically favorable compared to 3' ends[44,45]. This is likely due to 5' ends preferentially stacking on the core duplex compared to 3' ends. Furthermore, Santalucia thermodynamic calculations (Figure 3) predict that the specific nearest neighbors bonds in the 5' shifted state are more energetically favorable than those in the 3' shifted state[46] (Note that these calculations are performed at higher salt concentration, and that salt seems to have a significant effect on the impact of inert tails on stability). Taken together, these experimental

results substantiate why we observe the single-shifted 5' state to be only slightly less stable relative to the aligned hybridized state. Furthermore, we expect implied timescales of the two leading modes to be on the same order of magnitude as these reflect the average lifetimes of these states. The stability of the single-shifted 3' state is notably lower than that of the single-shifted 5' state. In fact, the free energy is about equal to that of the double-shifted 5' state, despite it having more less intact base pairs. This is partially explained by nearest neighbor calculations, however these do not account for the differential effects of the 5' and 3' inert tails or kinetic considerations which we will discuss more below. The final 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 destabilitization – longer tails tend to have a diminishing effect on overall strand stability [43]. Accordingly, the longer inert tails in the double-shifted states have less of an effect on relative 5' vs. 3' stability compared to the shorter tails in the single-shifted states.



AT-All Configuration	NN dH (kJ/mol)	NN dS (J/molK)	NN dG at Temp (kJ/mol)	Sim Prob	Sim G/kT
Fully Intact	-251.8	-754.8	-20.8312	0.33983	1.079
5'-Single Shifted	-249.4	-747.3	-20.7262	0.10488	2.255
3'-Single Shifted	-197.5	-586.3	-18.0922	0.00436	5.436
5'-Double Shifted	-189.1	-572.9	-13.7926	0.00706	4.953
3'-Double Shifted	-137.2	-411.8	-11.1892	0.00055	7.507

Figure 3: a) Thermo value calculated from Santalucia data and from MSMs b) TICA projections of freen energies and state clustering c) and transitions probabilities between states

3.1.4 Transition probabilities between states

Beyond state probabilities and free energy approximations, the macrostate MSM also yields discrete kinetic information in the form of transition probabilities between states. Figure 3 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 slithering-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 may have some role in the substantially lower state populations we see above. 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.

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. Although these might not represent fully optimized hyperparameters for each sequence, they preserve Markovian properties of the system and provide adequate resolution to evaluate macrostates. The number of SRV coordinates and MSM macrostates are varied based on the kinetics and thermodynamics behavior of specific sequences.

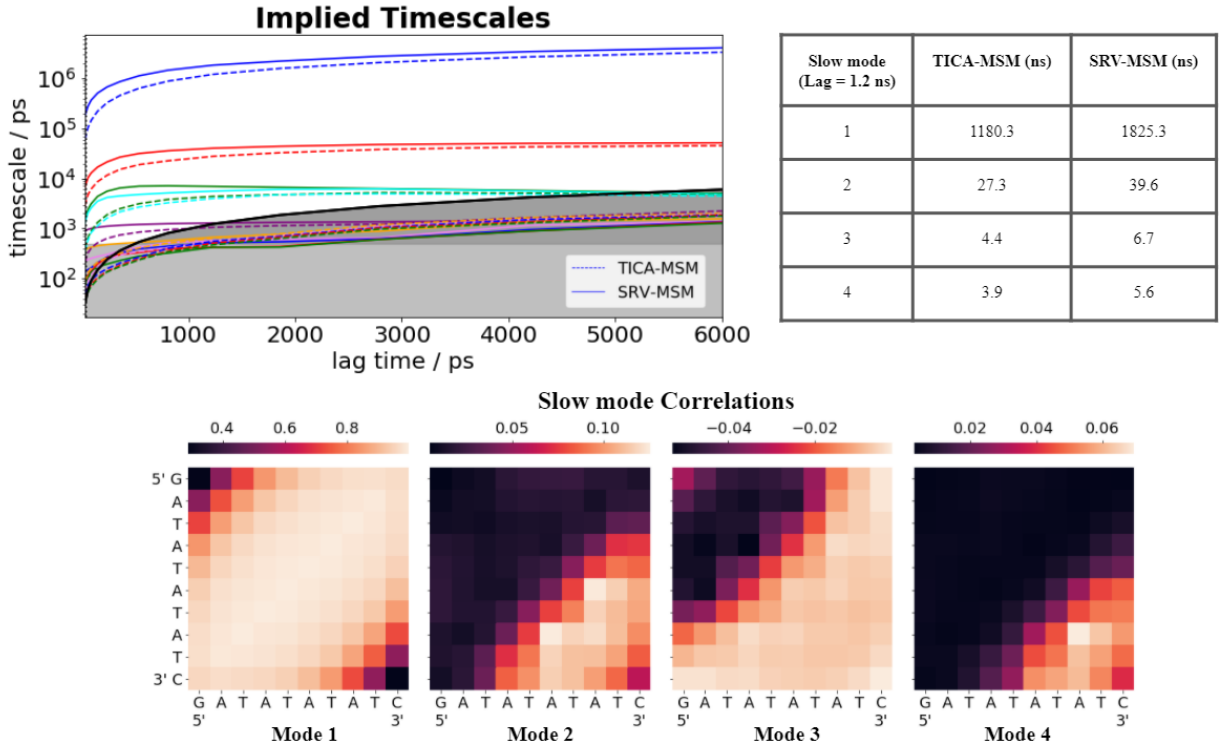
3.2 GC-end

3.2.1 SRV-MSM construction and states

Next we examine the GC-end sequence, which add GC 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 slithering 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 4) 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 1). 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 decreases the number of out-of-register WC pairs by two relative to corresponding AT-all states. This substantially diminishes the stability of the GC-end shifted states, resulting in lower state populations and shorter timescales.

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

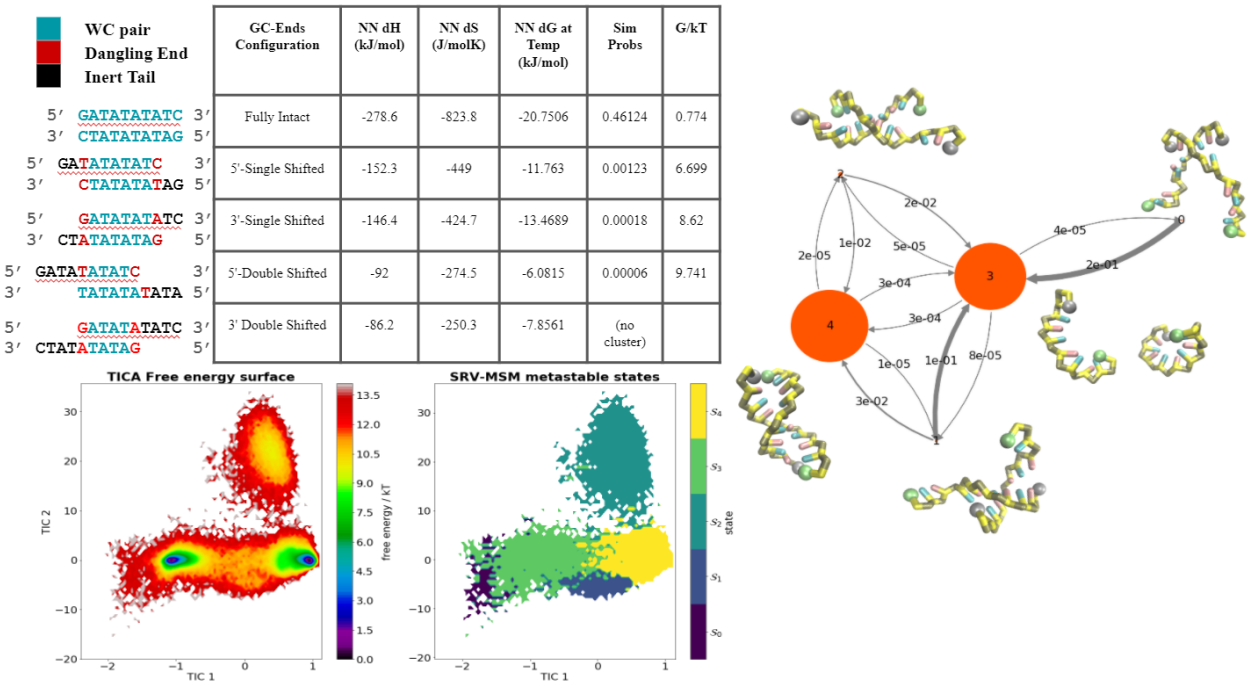
For our analysis, we consider C:T and G:A base pairs in the GC-ends shifted states as non-interacting dangling ends such that each shifted conformation has four total dangling ends and two inert tails. Although internal base pair mismatches can cause substantial conformational distortions such as kinking, terminal mismatches have been shown to be slightly stabilizing and have a minimal effect on helical character[47],[48]. In the context of the 3spn2 model, these ends are accounted for via intra-strand base stacking and inter-strand cross-stacking interactions[23]. The only direct interaction between non-WC basepairs



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 4: (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.

is parameterized by isotropic excluded volume potential, which is likely much more simplistic than the true mismatch interaction. Nevertheless, we find these bonds to be stabilizing enough to account for the relatively small population of conformation we observed in these shifted states (Figure 5). Contrary to AT-all thermodynamics, nearest neighbor calculations predict the 3' shifted states to be slightly more stable than 5' shifted states, again without taking into account entropic effects from inert tails. 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.



GC-Ends Configuration	NN dH (kJ/mol)	NN dS (J/molK)	NN dG at Temp (kJ/mol)	Sim Probs	G/kT
Fully Intact	-278.6	-823.8	-20.7506	0.46124	0.774
5'-Single Shifted	-152.3	-449	-11.763	0.00123	6.699
3'-Single Shifted	-146.4	-424.7	-13.4689	0.00018	8.62
5'-Double Shifted	-92	-274.5	-6.0815	0.00006	9.741
3' Double Shifted	-86.2	-250.3	-7.8561	(no cluster)	

Figure 5: a) Thermo value calculated from Santalucia data and from MSMs b) TICA projections of freen energies and state clustering c) and transitions probabilities between states

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 5' 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[49]. 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 6 we present the first two non-trivial diffusion map eigenfunctions and show representations of the degenerate third coordinate in the SI (Figure 7. 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 GC-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. 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, we find the second diffusion map coordinate to be indicative of inert tails that fold back onto these terminal out-of-register pairings.

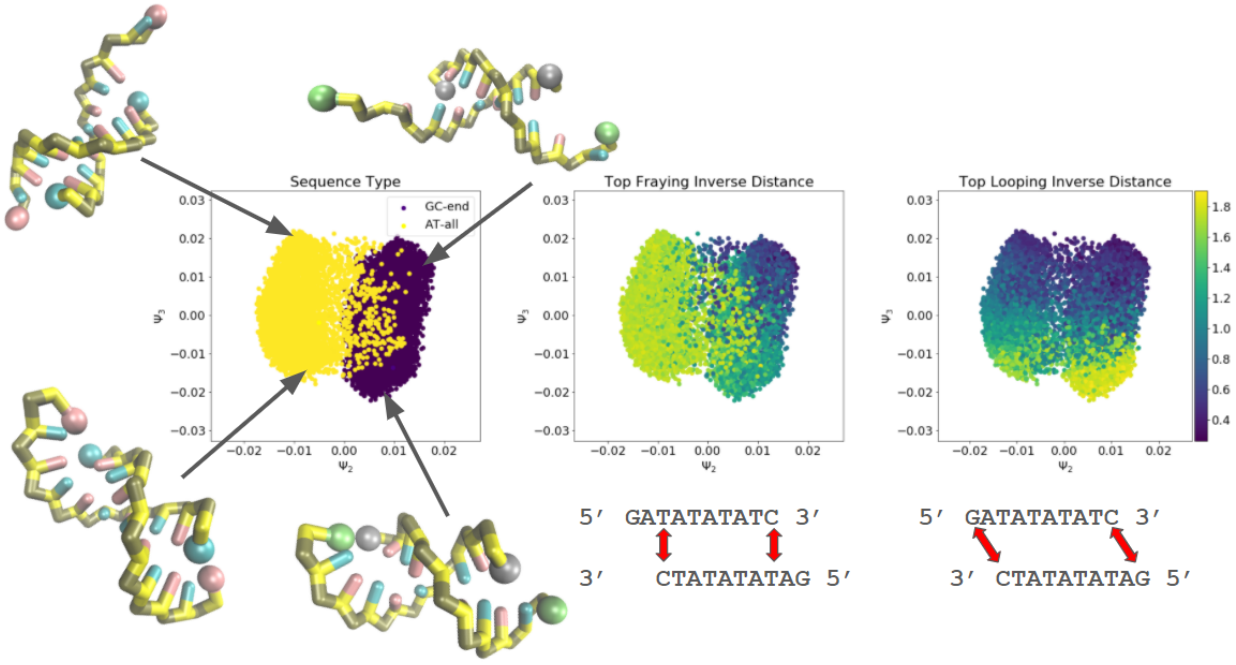


Figure 6: First two diffusion map coordinates built from 10000 single-shifted 5' states, equally sampled from AT-all and GC-end. Color maps show inverse distances between out-of-register ends and complementary ends.

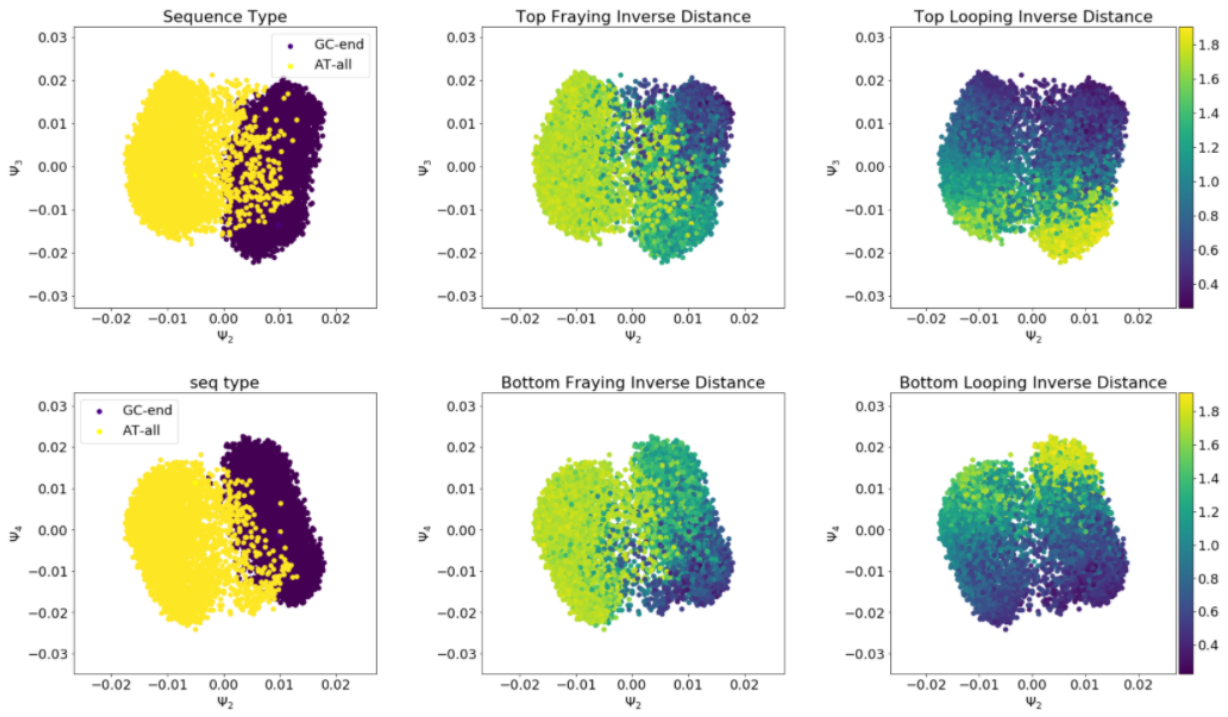


Figure 7: SI fig

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 5). 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 experiments, the GC-end sequences was observed to have less deviation from the two-state dissociation model compared to oligonucleotides with GC pairs closer to the core[13]. Futhermore, 2D spectroscopy analysis

showed a distinct kinetic response – particularly at lower temperature – that did not clearly match the physical processes under investigation.[50]. Given that this signal contained some A:T and G:C character and was slower than single base pair fraying, it is possible that it represented some oligomer population shifting out-of-register as we observe in our MSMs. This cannot be verified given that the signal was recorded in a congested spectroscopic range, however this dynamic should be taken into consideration when analyzing or utilizing oligos of a similar sequence.

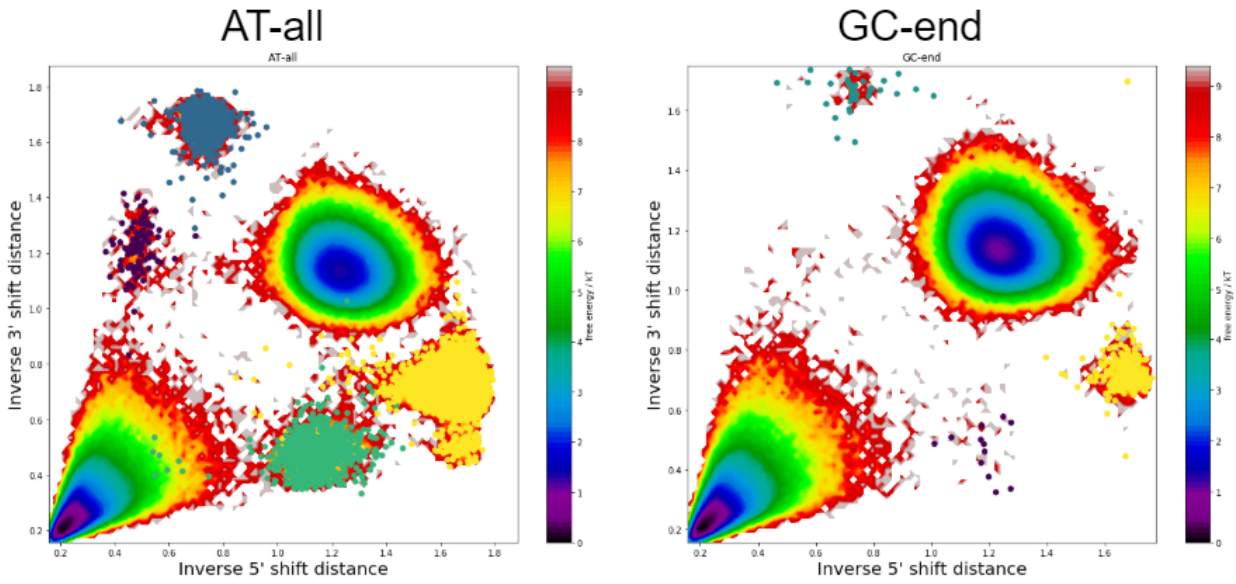


Figure 8: Shows similarities between shifting modes and state memberships for AT-all and GC-ends (probs better to grey out underlying free energy landscape and move this to the SI.

3.3 GC-core

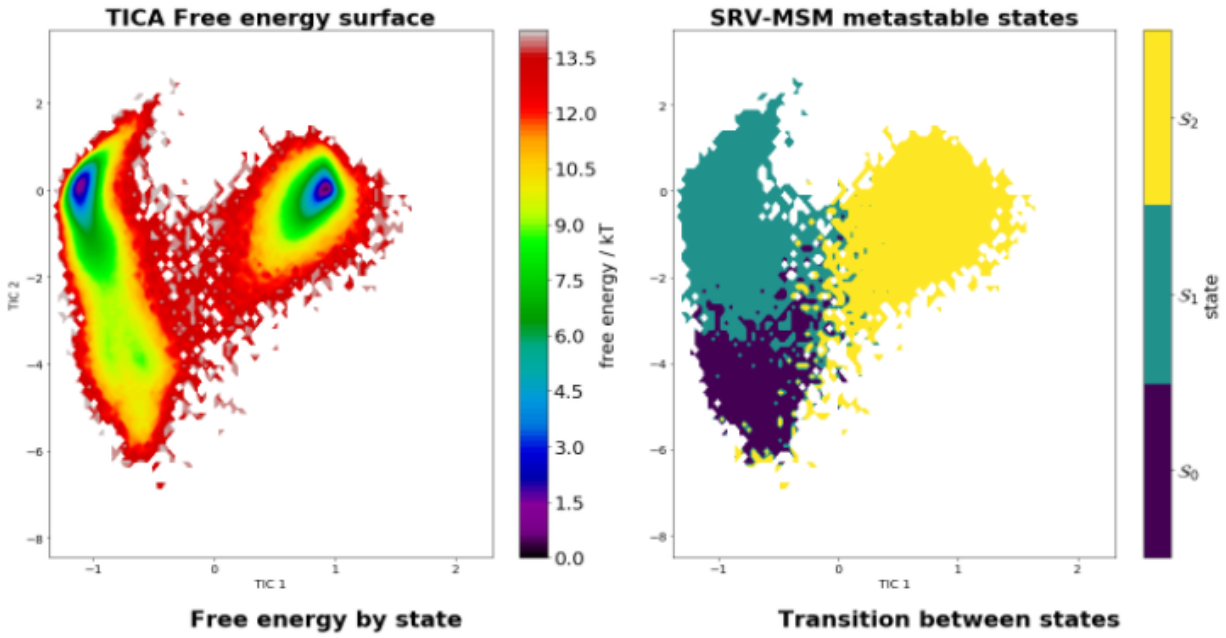
3.3.1 Building SRV-MSM

The GC-core sequence represents a departure from the dominant slithering dynamics observed for AT-all and GC-end. Based on SRV cross-validation (SI Figure 1), 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. As we observed from the 2nd SRV mode analysis above, 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[17],[9]. 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. The MSM transition probabilities suggests 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. This time, we set out distance metric to the same permutation-free coordinates we used to build the SRV-MSMs. Again, we were able to identify a combination



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

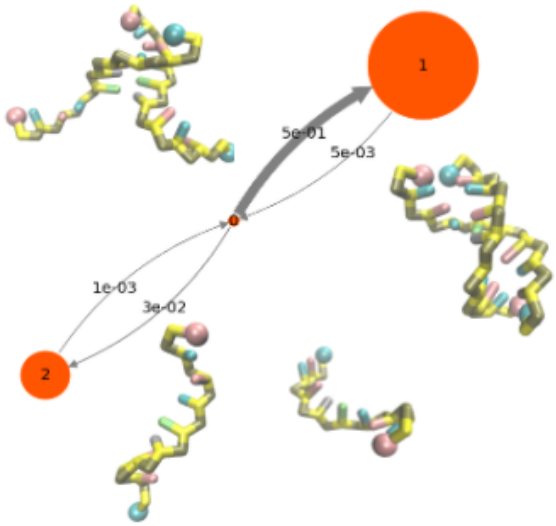


Figure 9: Full MSM output for GC-core

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 likelihood that a configuration was clustered into the meta-stable 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 tendency for the 3'T end to fray farther out of place relative to its 5'A counterpart.

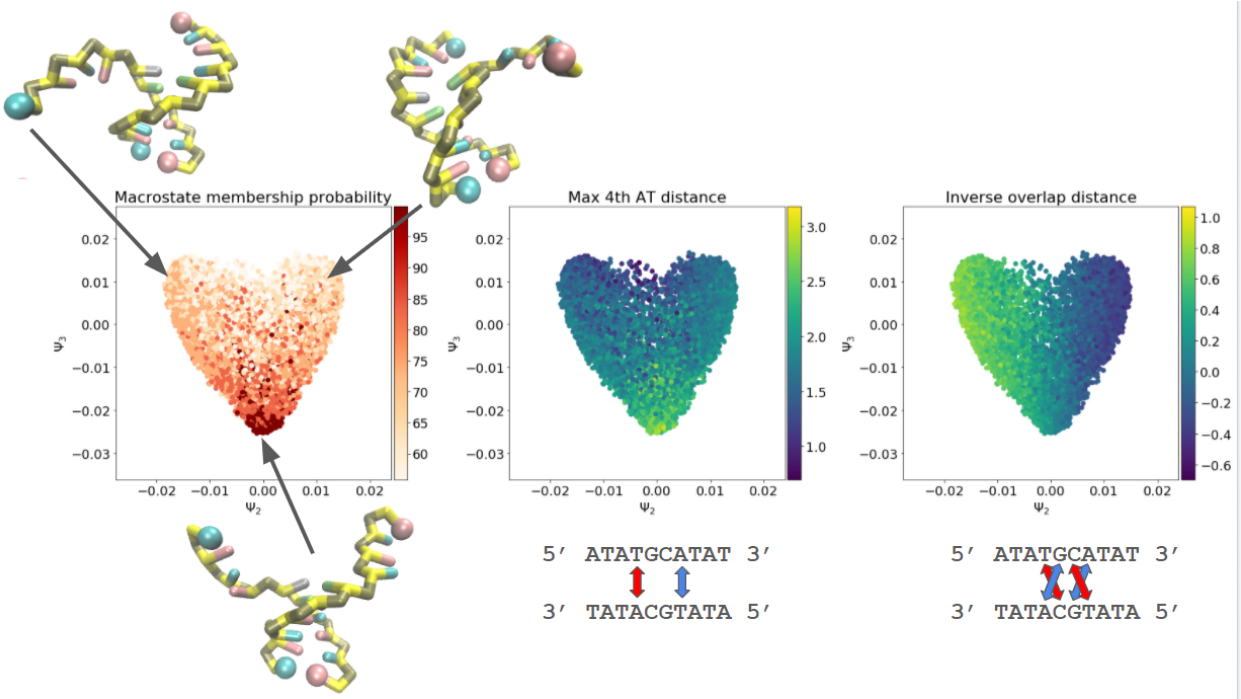


Figure 10: Plotting the first two non-trivial diffusion map eigenfunctions. Color maps show the probability that conformations are clustered into the frayed macrostate, maximum distance between 4th AT basepairs, and the the overlap between adjacent AT basepairs and the GC core.

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 dy-

namical 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: (). 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 2. 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.

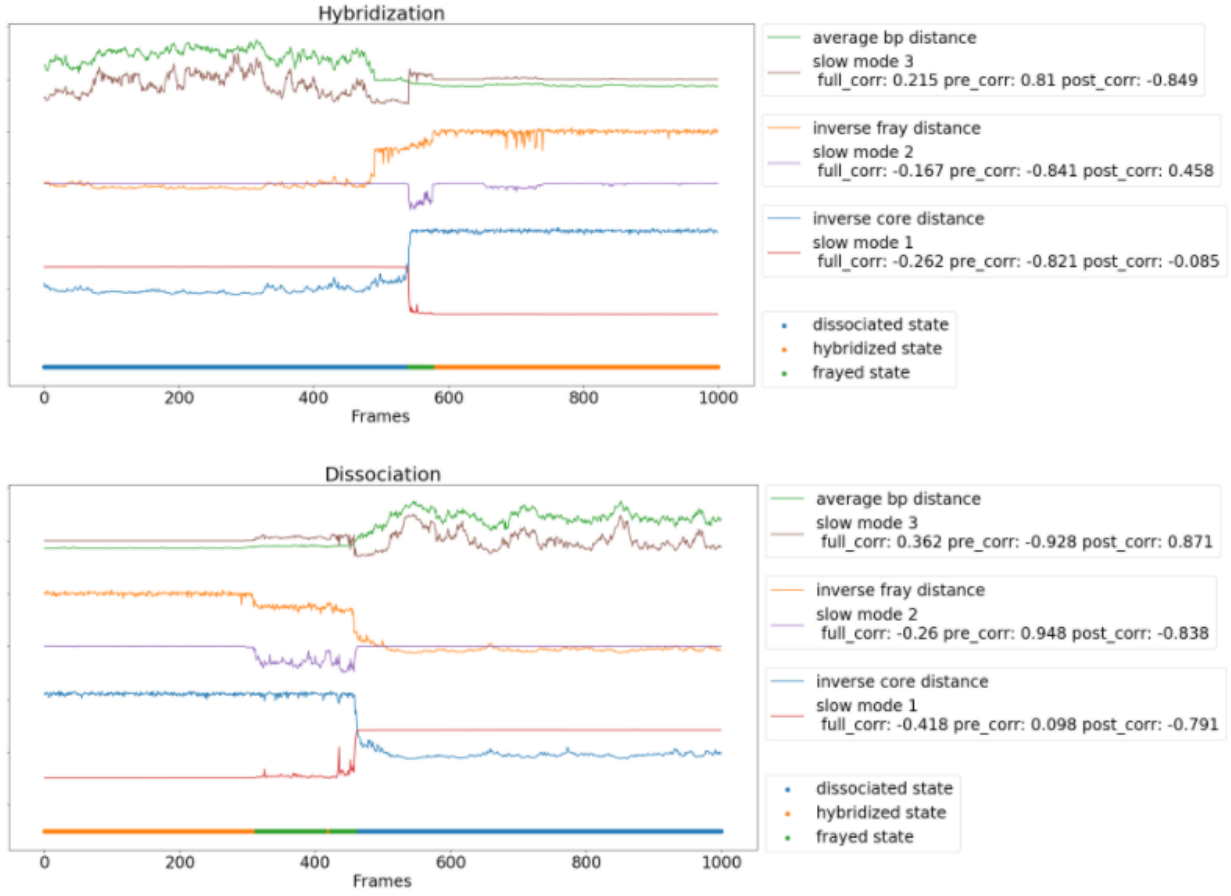


Figure 11: Show how SRV coordinates correspond to physical features over time and during sample dissociation and hybridization events. Should probably go to SI if we also include GC-core diffusion maps

3.3.4 Comparison to Experiment?

When examining these same four sequences using T-jump and 2D spectroscopy, Sandstead et al. found that the GC-core had the highest deviation from two-state behavior during

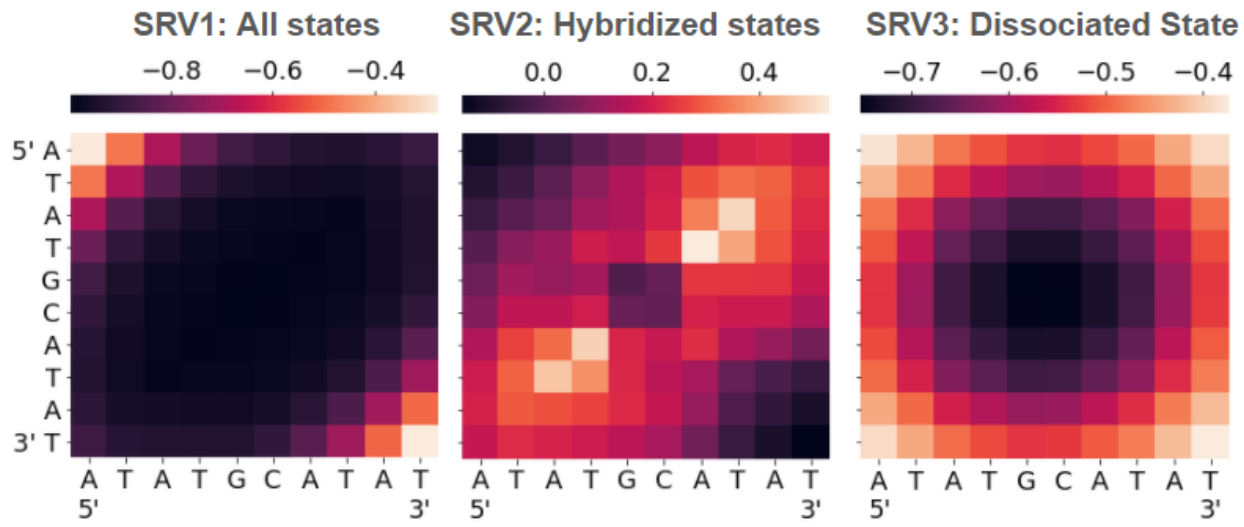


Figure 12: Shows Pearson correlations between leading SRV modes and all 100 intermolecular distances (needs labels for slow modes)

dissociation[13]. As their analysis did not consider the 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. Further analysis showed that AT termini fraying was an effectively barrierless process characterized by rapid inter-conversion between all accessible frayed states[50]. 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 "macrofrayed trajectories" 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 frayed behavior. Although we observe substantial fraying of the two AT termini in the simulation data, these dynamics are too fast to converge as distinct frayed modes. Instead, we observed two modes converge, corresponding to the association/dissociation dynamic and diffusive behavior the strands while 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 an effectively two-state process with no long-lived or kinetically relevant metastable states. Previous work has attributed AT-all termini fraying prior to dissociation as a deviation from two-state behavior[13]. Here we consider these AT-frayed states to be part of the overall hybridization state. While AT-termini fraying is surely a prerequisite for dissociation, we find these states to be so common and fleeting that very few evolve initiate dissociation. Furthermore, they do 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.

3.4.1 Definition of two-state and nucleation

In the simulation data, we do observe sustained regions of partial base pairing during hybridization. These states reflect initial contact that form between base pairs prior to a fast zippering mechanism that binds the oligos into a fully hybridized state. Qualitatively, these observations match well with the three-state assumption for hybridization that proceeds through a "nucleation" state of around 3 base pairs bound[9] (cite earlier porske 1971). Our featurization method, however, was designed to elucidate sequence-dependent – thus location-dependent – dynamics, and does not naturally group these kinds of state. This limitation also does not allow us to pick up on proceeding zippering mechanisms, as these would each appear as separate dynamics per the independent nucleation site and need to be

sampled accordingly.

Other ways to investigate GC-mix:

1. Can try building simplified MSM here: (use number of base pairs bound as features)
2. Timescale analysis shows that GC-mix is reversible two-state process. We note that it takes about 40 percent longer for a fully hybridized oligomer to fully dissociate (once $>$ 2 bp are borken) for GC-core compared to GC-mix.
3. Given that GC-mix and GC-core are run at very similar temperature, we would expect the extra basepairs broken during GC-mix dissociation process to slow dissociation, however in the opposite effect was observed in the slow mode data
4. Evidence of fraying correlations in higher order modes, but these are not consistent and their inclusion does not generate a metastable third mode
5. Show 2nd slow mode correlations in SI plus some configuration snapshots
6. Discuss misaligned configurations here instead of in the GC-core section

3.5 Project results onto shared basis

Here I've constructed a shared basis from physical coordinates informed by the AT-all SRV2 and GC-core SRV2 (Figure 13) and TICA coordinates learned on a combination of all four sequence (Figure 14). I'm leaning towards presenting the first in the main text and including the other in the SI, but I'm not sure how clear either of these are.

3.5.1 Limitations

1. Does not explicitly pick up on nucleation/zippering events
2. Coarse grain-models makes it difficult to compare timescales. Furthermore our implied timescales function more as an average frequency of a process rather than a value that can be directly compared to experiment.

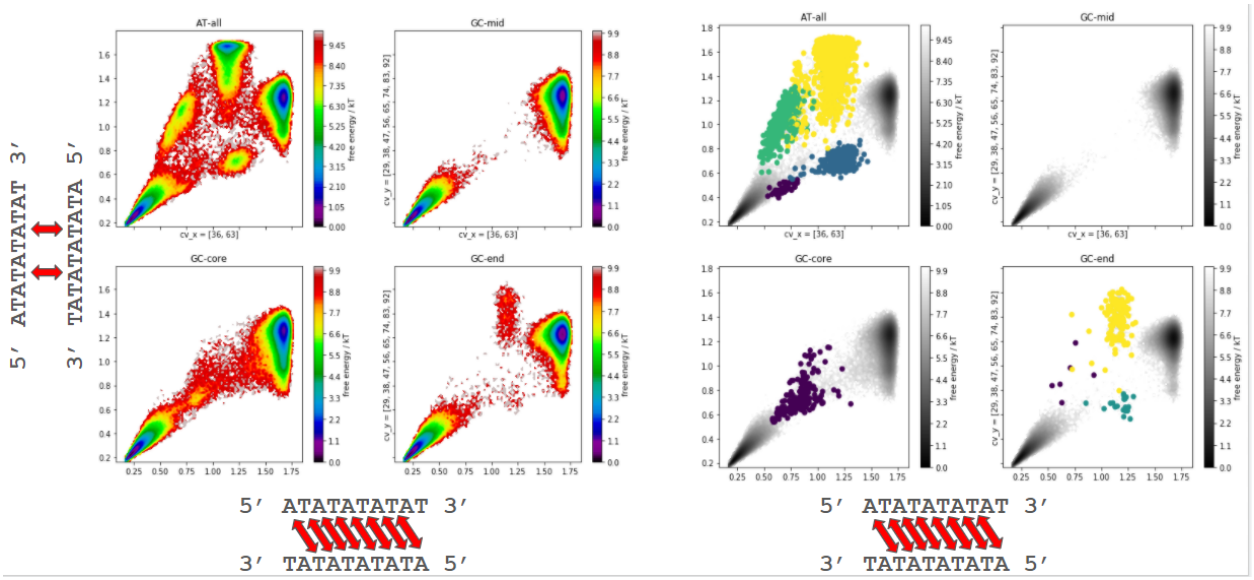


Figure 13: All sequences and metastable macrostates plotted on frayed/shifted collective variable axes, Not sure how helpful these axes drawings are – also need to switch x and y here.

3. Limited by memory (number of frames) required to capture equilibrium trajectories at a save rate that enabled adequate resolution for the dynamics of interest.

4 Conclusion

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

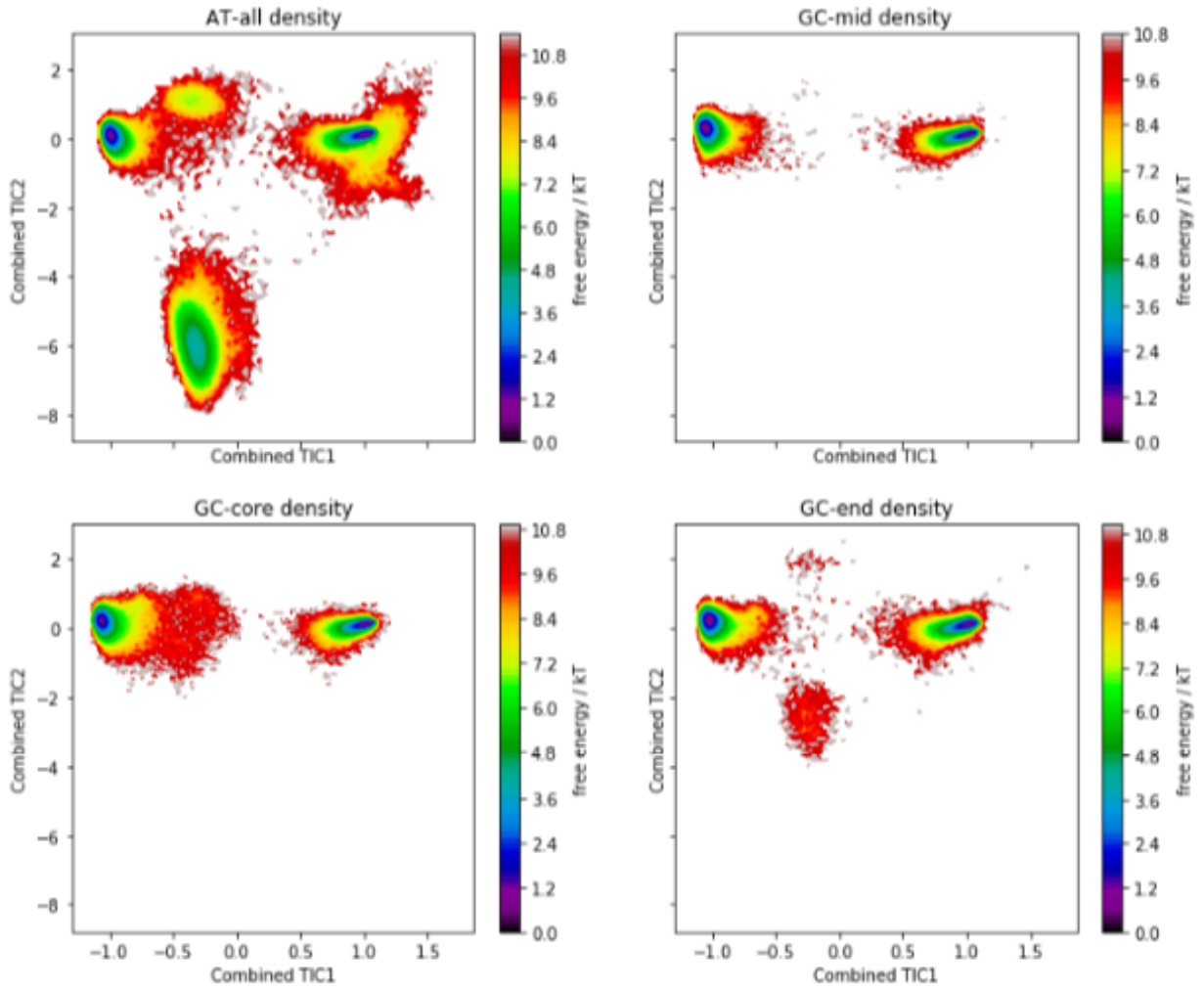


Figure 14: All sequences plotted on combined TICA axes. Can add similar macrostate distribution as shown in the above figure.

References

(1) 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*.

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

(3) Adleman, L. Molecular solution to computational problems.pdf. 1994.

(4) Rothemund, P. W. Folding DNA to create nanoscale shapes and patterns. *Nature* **2006**, *440*, 297–302.

(5) Gu, H.; Chao, J.; Xiao, S. J.; Seeman, N. C. A proximity-based programmable DNA nanoscale assembly line. *Nature* **2010**, *465*, 202–205.

(6) Mhatre V. Ho, J.-A. L.; Martin, K. C. å§žåŻăçŽĎæŤźåŔŸNIH Public Access. *Bone* **2012**, *23*, 1–7.

(7) 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.

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

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

(10) 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.

(11) 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*.

(12) 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.

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

(14) 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*.

(15) 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.

(16) 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.

(17) 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.

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

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

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

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

(22) 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.

(23) 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*.

(24) 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.

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

(26) 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**,

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

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

(29) 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.

(30) 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.

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

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

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

(34) 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.

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

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

(37) 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**,

(38) 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**,

(39) 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*.

(40) 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.

(41) 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.

(42) Pinamonti, G.; Paul, F.; Rodriguez, A.; Bussi, G. analyzed with core-set Markov state models. *43*.

(43) 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.

(44) 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.

(45) 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.

(46) Allawi, H. T.; SantaLucia, J. Nearest neighbor thermodynamic parameters for internal G·A mismatches in DNA. *Biochemistry* **1998**, *37*, 2170–2179.

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

(48) 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.

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

(50) 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.