

SUPPORTING INFORMATION

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

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

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

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

E-mail: andrewferguson@uchicago.edu

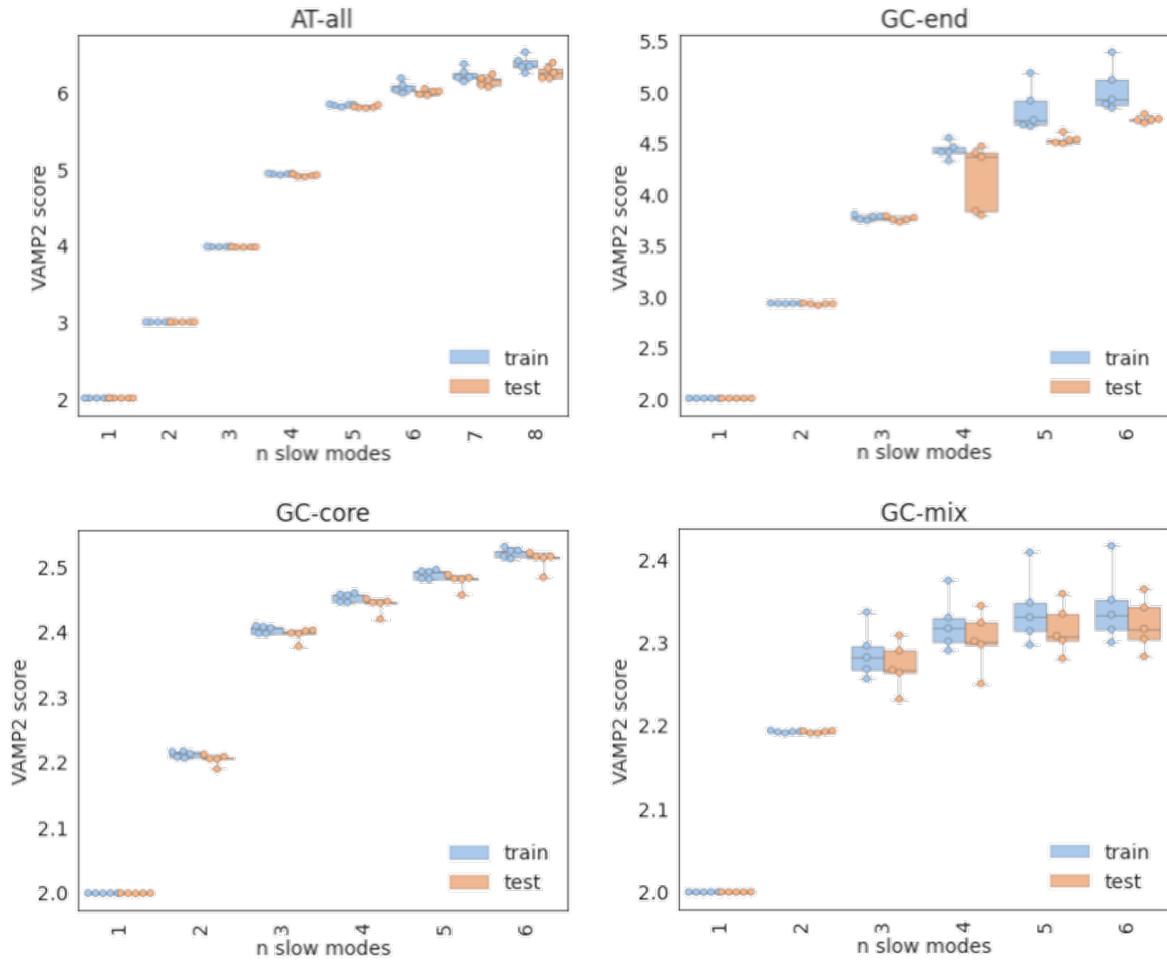


Figure S1: Five-fold cross-validation of the SRV VAMP-2 scores to select the optimal number of SRV coordinates for each sequence. A knee in the VAMP-2 plot was identified at the fifth, fourth, third, and second slow modes for AT-all, GC-end, GC-core, and GC-mix, respectively. An embedding of corresponding dimensionality was then used to cluster frames into discrete states. The absence of any significant separation in the training and testing VAMP-2 scores demonstrates that model is not overfitted.

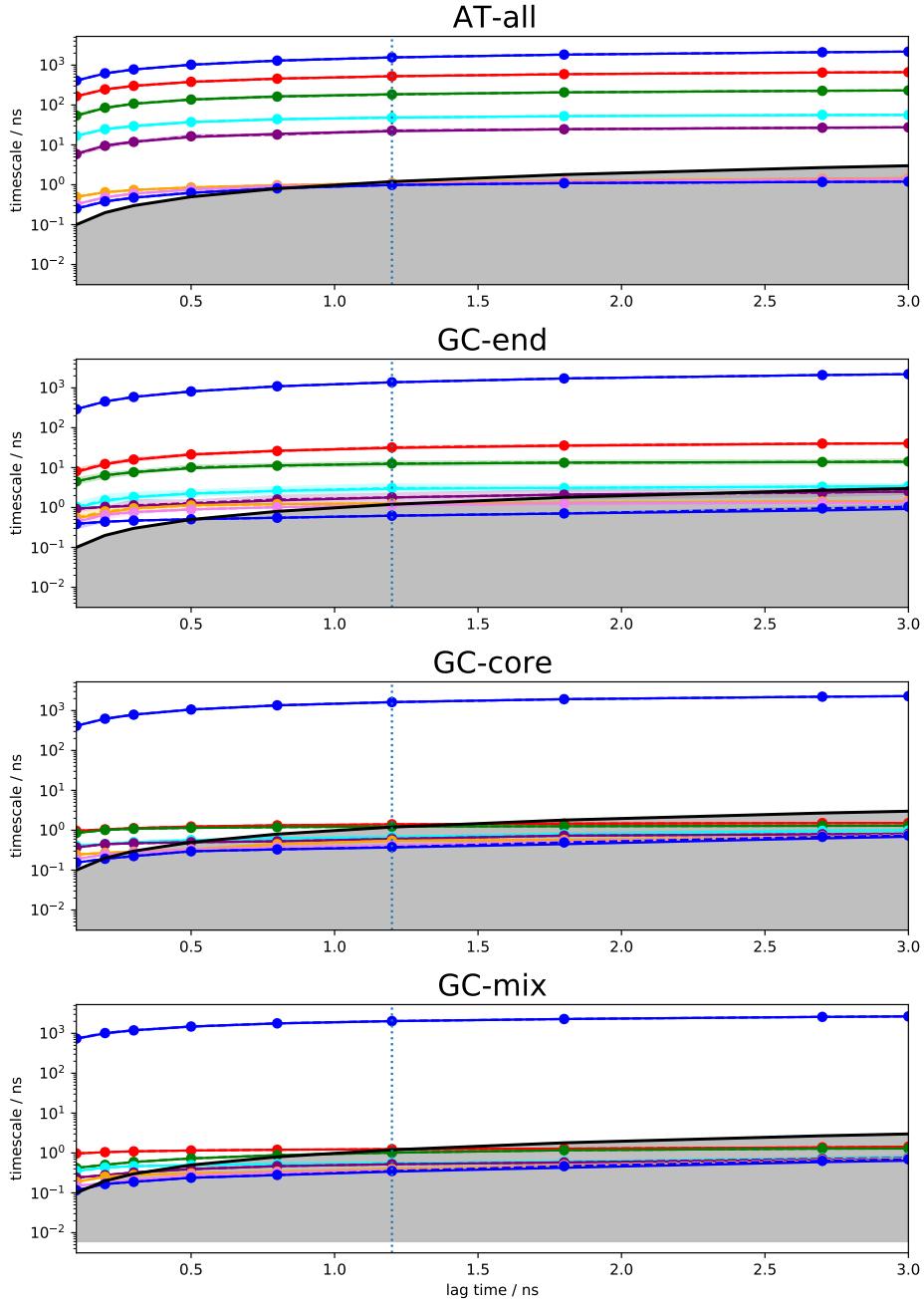
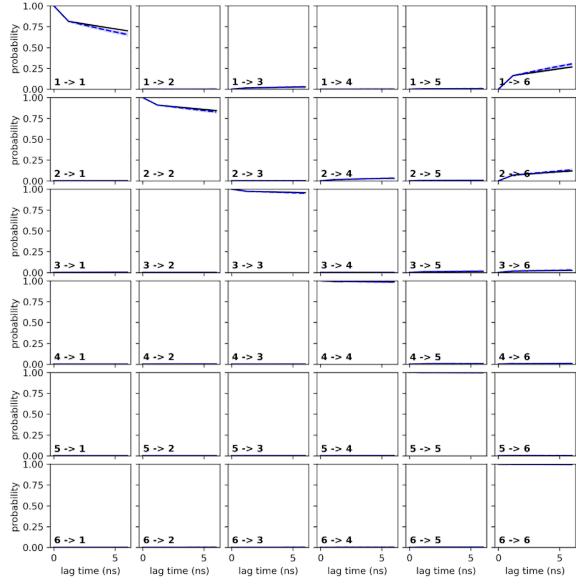
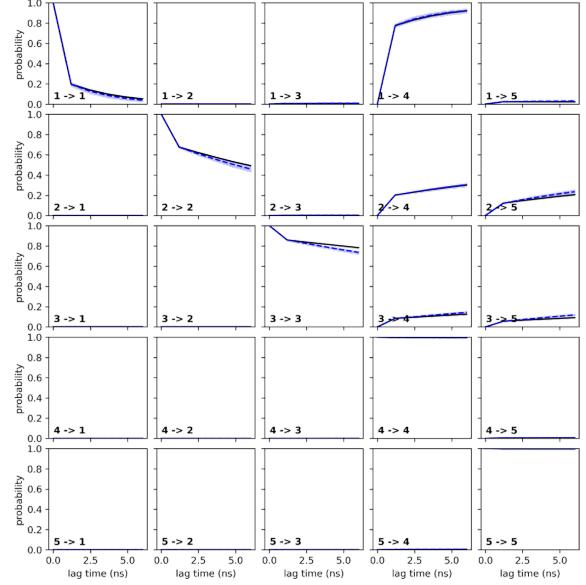


Figure S2: Convergence of the MSM implied time scales t_i as a function of lag time τ . Solid lines indicate maximum likelihood result while dashed lines show the Bayesian ensemble means superposed on shaded areas representing 95% confidence intervals. The high simulation data volume used in our MSM construction means that the maximum likelihood and Bayesian ensemble means are in good agreement and confidence intervals are narrow. The implied time scales for all sequences converge at a lag time of $\tau = 1.2$ ns (vertical line). The black solid curve marks equality of the implied time scale and lag time and delimits the shaded region wherein the implied time scales are shorter than the lag time and cannot be resolved.

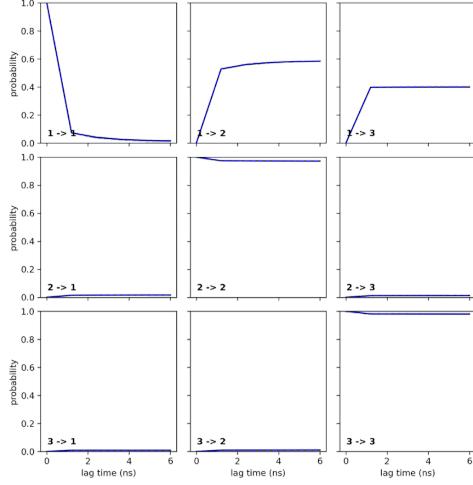
AT-all



GC-end



GC-core



GC-mix

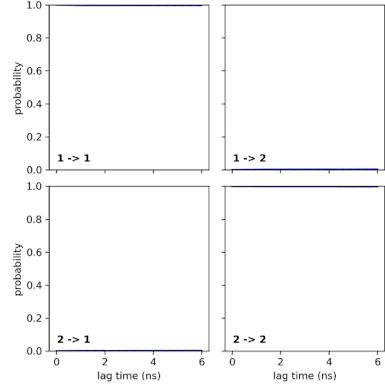


Figure S3: Chapman-Kolmogorov (CK) tests comparing the probabilities of remaining within each macrostates for each sequence as a function of lag time predicted by k applications of an MSM constructed at the $\tau = 1.2$ ns lag time $\mathbf{P}^k(\tau)$ (dashed blue line) versus those computed from an MSM constructed at that particular lag time $\mathbf{P}(k\tau)$ (solid black line). The blue shaded area indicates the estimated standard deviation in $\mathbf{P}^k(\tau)$. The good agreement between these two results provides numerical validation of the Markovian nature of the $\tau = 1.2$ ns lag time MSM.

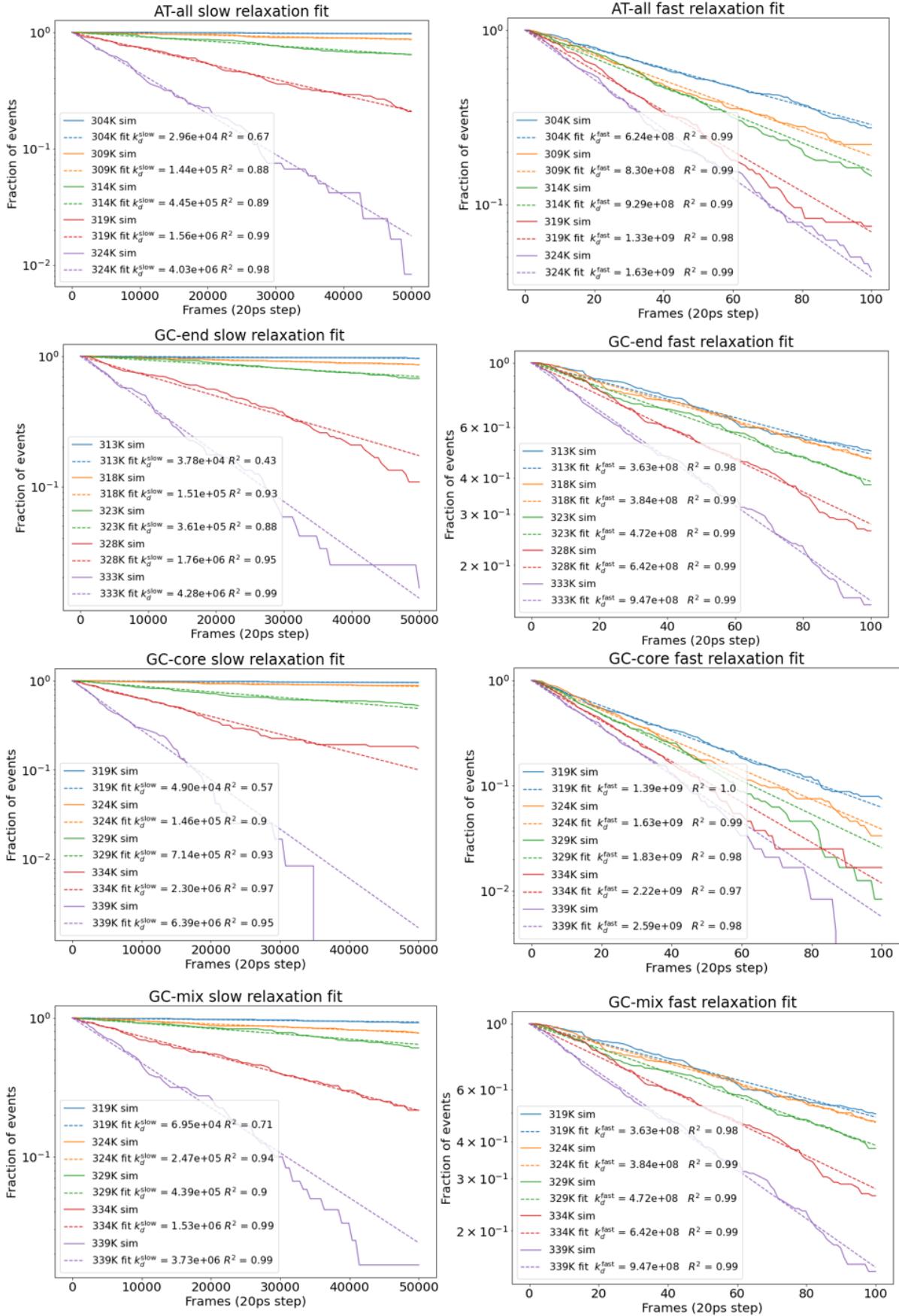
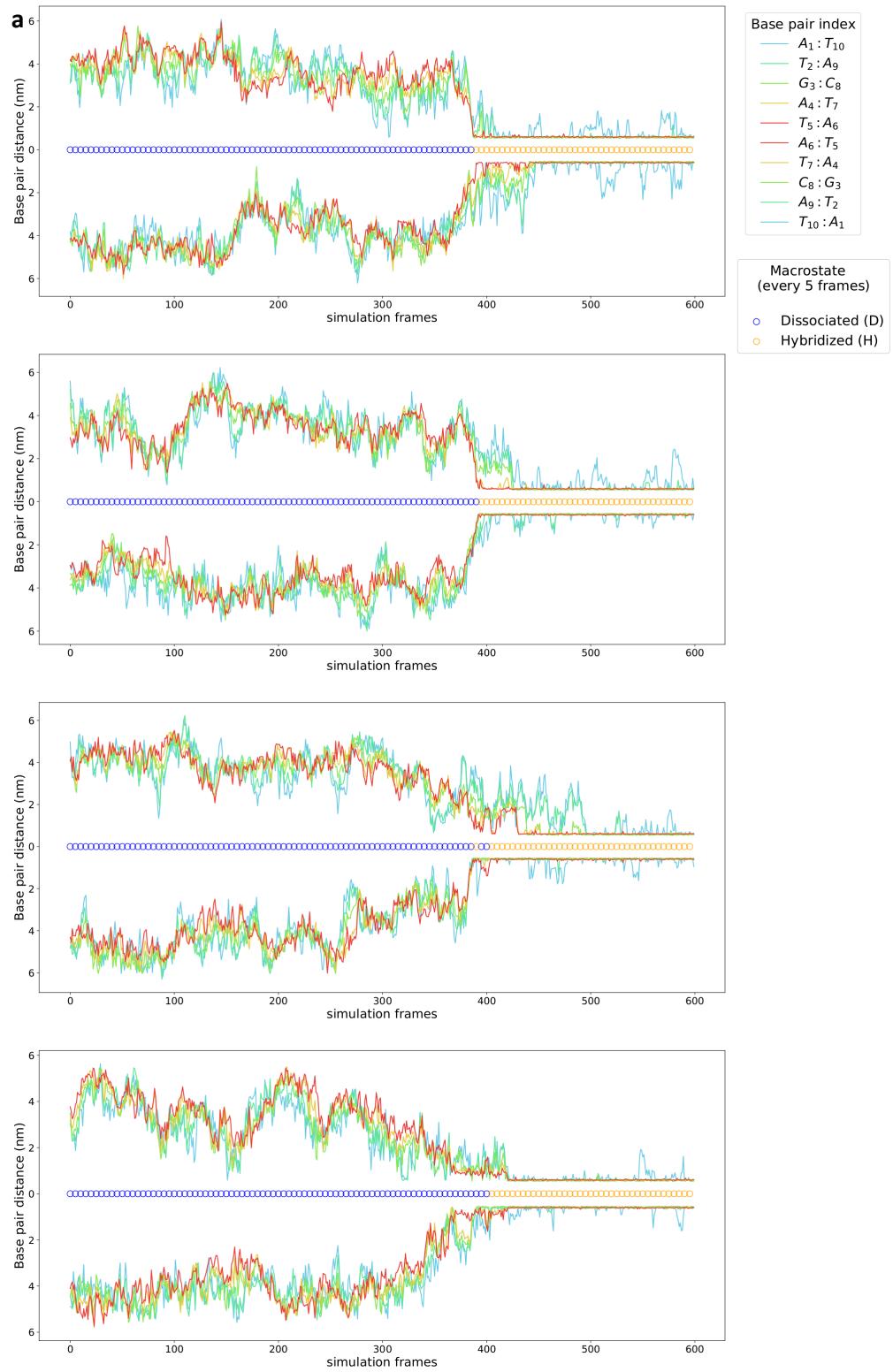


Figure S4: Exponential fits for both slow (dissociation) and fast (fraying) response for all four sequences during “computational T-jump” experiments. From 120 independent $1 \mu\text{s}$ simulations, we compiled the slow response data by recording the fraction of sequences with both central Watson-Crick base pairs intact as a function of time, and the fast response data as the fraction of sequences with both terminal Watson-Crick base pairs intact as a function of time. We define a Watson-Crick base pair to be intact if the centers-of-mass of the two complementary bases lie within a linear distance of 1.3 nm. We extracted our computational estimate of k_d^{fast} by fitting a decaying exponential to the fraction of bound A:T termini as a function of time $f_{\text{unfrayed}}(t) = \exp(-k_d^{\text{fast}} t)$. Similarly, we extracted our computational estimate of k_d^{slow} by fitting a decaying exponential to the fraction of hybridized sequences as a function of time $f_{\text{hybridized}}(t) = \exp(-k_d^{\text{slow}} t)$. We report within the figure legend to each panel the coefficient of determination R^2 for a least squares linear fit of the model to the data in log space (i.e., $\log(f) = -k_d t$) and data are plotted on log-linear axes to facilitate visual comparison of the fits. In all cases we observe excellent fits of the models to the data with all R^2 values equal to or better than 0.88 (except for the slow response at the lowest temperature $T_m-5\text{K}$ where dissociation events are sparse).



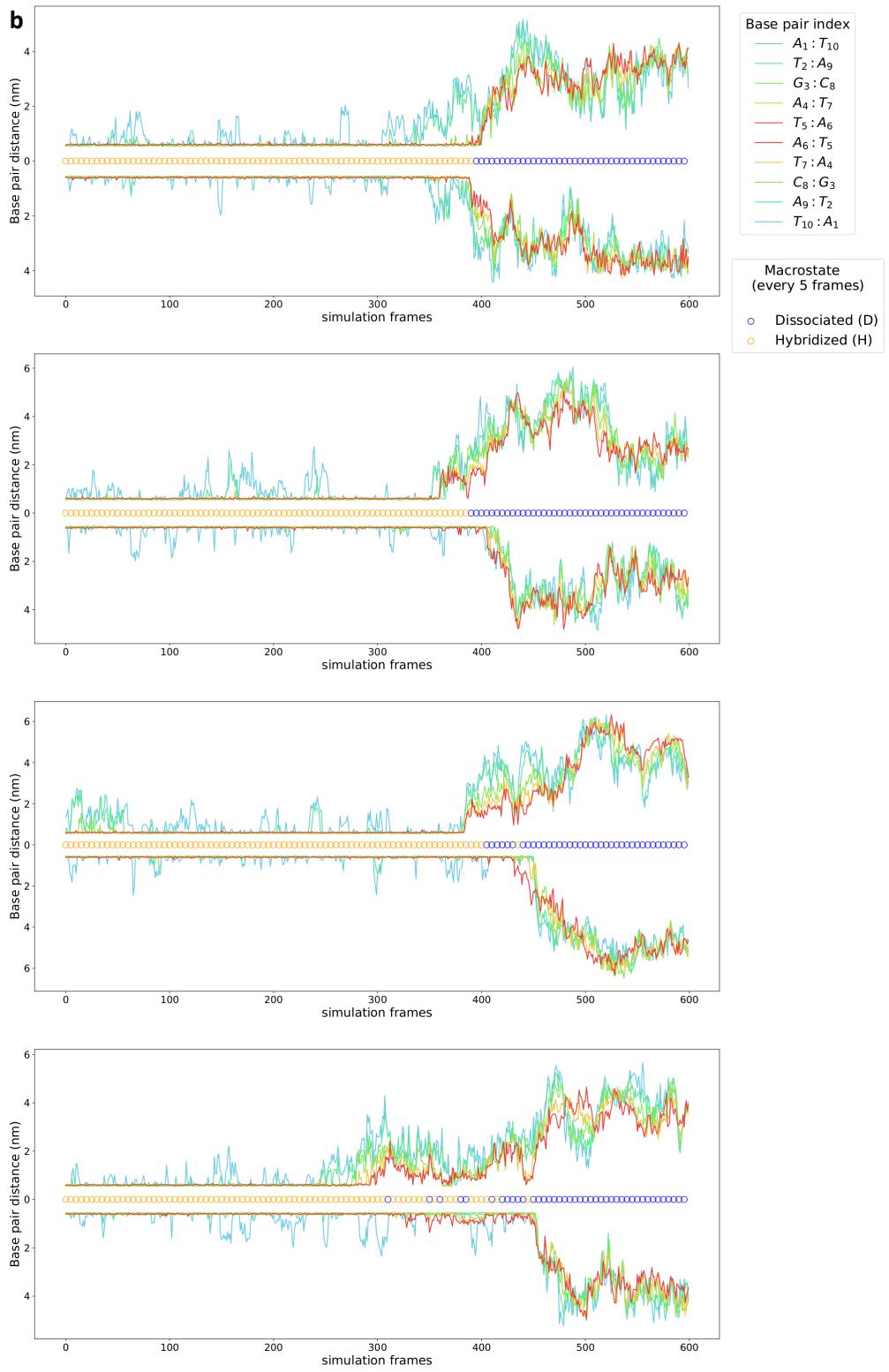


Figure S5: GC-mix hybridizes by nucleation-zippering and dehybridizes by fraying-peeling. Tracking of the 10 intermolecular distances between native WC base pairs over the course of four additional (a) hybridization and (b) dehybridization events. Symmetrically permutable distances (e.g., $A_1:T_{10}$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

Nearest neighbor model of duplex thermodynamics

DNA duplex hybridization thermodynamics are commonly predicted using nearest-neighbor (NN) models wherein the stability of a given base pair is assumed to solely depend on the identity and orientation of its adjacent base pairs. We apply the popular NN model reported by SantaLucia⁷ to estimate the stability of fully intact, shifted, and frayed duplex configurations observed in our MSMs.

The NN model assumes hybridization proceeds in a two-state all-or-nothing manner. The Gibbs free energy difference between the hybridized and dissociated states can be determined from calculation of NN enthalpy and entropy contributions,

$$\Delta G_{NN}^\circ(T) = \Delta H_{NN}^\circ - T\Delta S_{NN}^\circ. \quad (1)$$

ΔH_{NN}° and ΔS_{NN}° are assumed to be independent of temperature T , and are computed from the sum over all NN terms in a given configuration,

$$\Delta H_{NN}^\circ = \sum_i^{n_{NN}} \Delta H_{NN,i}^\circ + \sum_j^{n_{DE}} \Delta H_{DE,j}^\circ + \Delta H_{init}^\circ, \quad (2)$$

$$\Delta S_{NN}^\circ = \sum_i^{n_{NN}} \Delta S_{NN,i}^\circ + \sum_j^{n_{DE}} \Delta S_{DE,j}^\circ + \Delta S_{init}^\circ + \Delta S_{sym}^\circ. \quad (3)$$

$\Delta H_{NN,i}^\circ$ is the NN enthalpy for a given dinucleotide step within duplex DNA, $\Delta H_{DE,j}^\circ$ corresponds to the enthalpic contribution from a given dangling end (DE) base next to the duplex⁷, and ΔH_{init}° is a duplex initiation term that accounts for terminal effects on duplex enthalpy. A schematic illustration of the NN and DE contributions for a fully hybridized duplex and an out-of-register shifted state is presented in Fig. S6. ΔS_{NN}° is constructed in an analogous manner from NN, dangling end, and initiation terms, plus a symmetry term, ΔS_{sym}° , that applies an entropic penalty to self-complementary sequences for maintaining C2 symmetry.

The SantaLucia NN parameters are defined for a solution condition at neutral pH with a sodium concentration of 1M. We use the empirical salt correction developed by Owczarzy⁷ to determine the NN parameters at the sodium concentration used for simulations in this work. The Owczarzy model assumes the effect of counter-ions on DNA hybridization thermodynamics to be purely entropic,

$$\Delta S_{NN}^\circ(c_{Na^+}) = S_{NN}^\circ(1M) + \Delta H_{NN}^\circ[(4.29f_{GC} - 3.95) \times 10^{-5} \ln(c_{Na^+}) + 9.4 \times 10^{-6}(\ln(c_{Na^+}))^2], \quad (4)$$

where f_{GC} is the fractional G:C content of the duplex and c_{Na^+} is the concentration of sodium counter-ions measured in M.

In applying the NN models to each macrostate, we made the simplifying assumption that the ensemble of microstates constituting each macrostate could be represented by a single pattern of Watson-Crick base pairing that are schematically illustrated in Fig. 2a. Indeed, each macrostate is composed of an ensemble of partially frayed states with higher free energy contributions than the intact pattern. However, when averaging across the frayed contributions of all microstates in each macrostate, we have found that incorporating additional NN configurations changes $\Delta\Delta F$ values by < 1 kJ/mol when compared to using a single microstate. Given the extent to which a single NN representation captures each macrostate (which is further demonstrated in Fig. S7), we have elected to maintain a single microstate free energy estimate for each macrostate

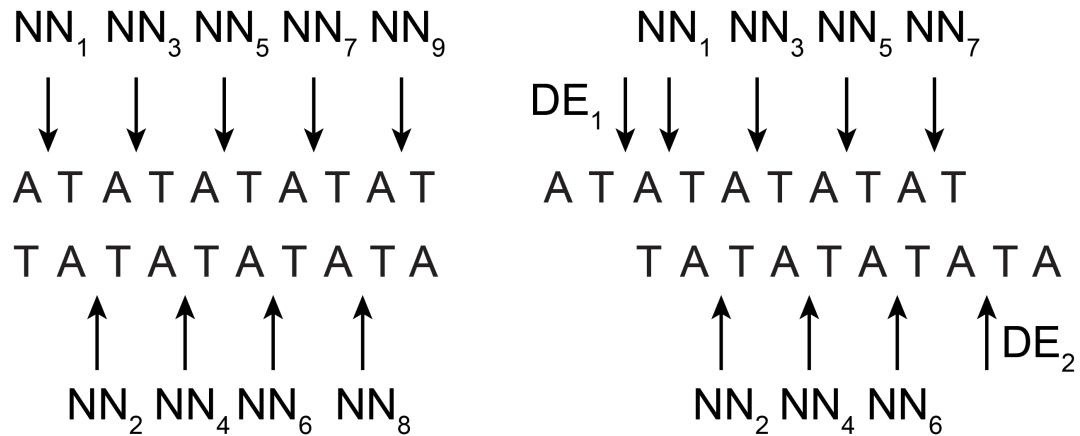


Figure S6: Schematic of nearest neighbor (NN) contributions for a fully hybridized AT-all duplex and an out-of-register duplex with dangling ends (DE).

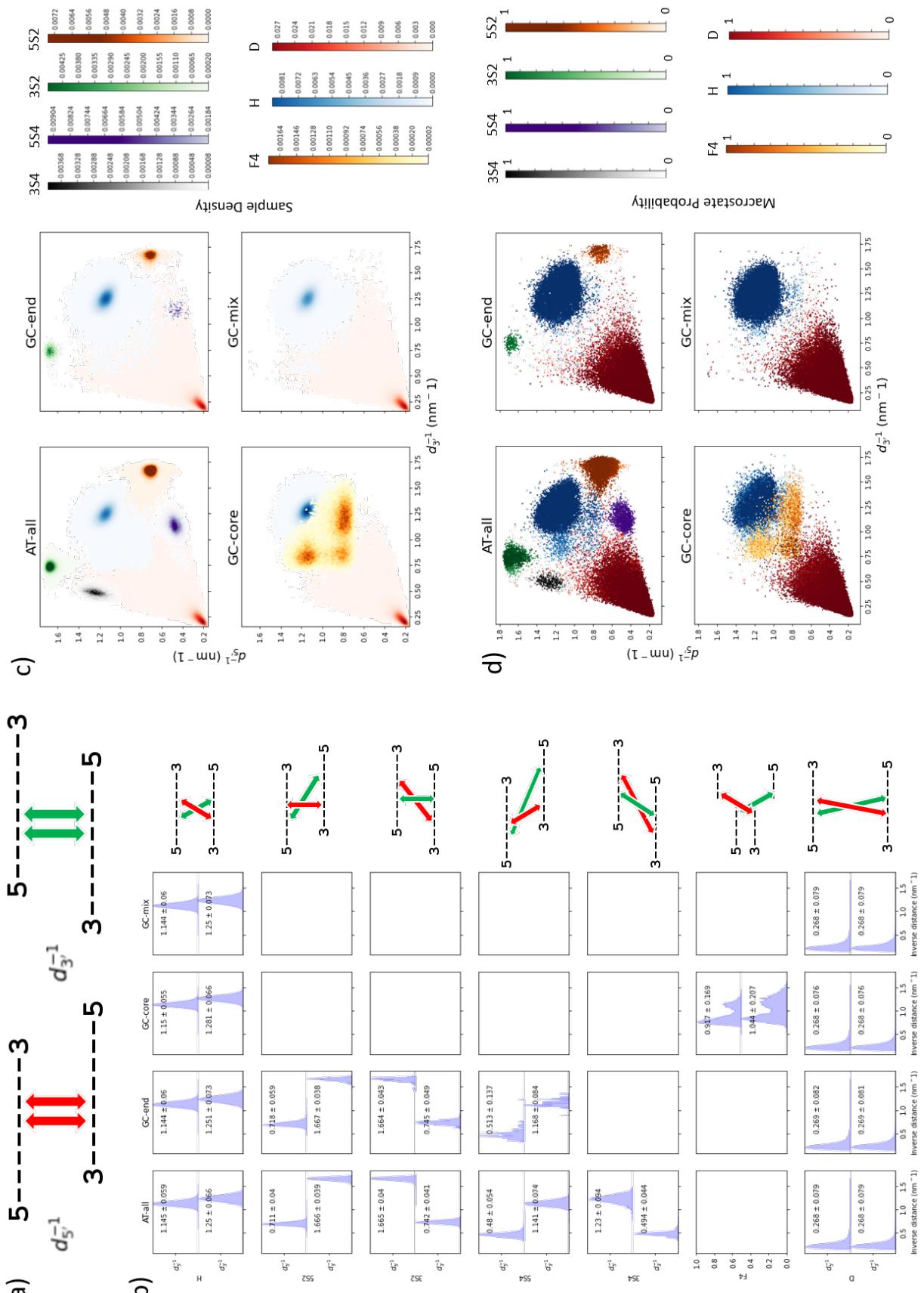


Figure S7. Macrostate interpretation and structural analysis. The Markov state model macrostates are learned from the simulation trajectories using the featurization and deep learning procedure described in Section 2.1.2 of the main text, but it is informative to perform a *post hoc* projection of the simulation data into a low-dimensional space spanned by physically interpretable order parameters in order to determine the structural characteristics of each macrostate and the structural heterogeneity of the microstate ensemble. (a) Definition of two physically-interpretable order parameters useful in characterizing the structure and base pairing of each microstate under both shifting and fraying. We define the 3' shift for a particular base on a strand as the linear distance between that base and the base on the complementary strand offset by two base pairs in the 3' direction. This distance reaches a minimum when the two strands are mutually shifted two base pairs in the 3' direction. We define $d_{3'}$ as the mean 3' shift averaged over the two central bases, represented by red arrows, in the complementary strands. It is more convenient to work the reciprocal $d_{3'}^{-1}$, since it has a finite range whereas $d_{3'}$ diverges for dehybridized strands. We define $d_{5'}^{-1}$ analogously as the reciprocal mean 5' shift. We have found this pair of variables extremely informative in revealing the microstate and macrostate structure and base pairing patterns. (b) We report the mean and standard deviation of $d_{3'}$ and $d_{5'}$ taken over all microstates comprising each macrostate for each sequence. The small standard deviations that are consistent with a narrow distribution of the bulk of the microstates around the free energy minimum of each macrostate. (c) Superposition of sample density for each macrostate for each sequence AT-all, GC-end, GC-core, and GC-mix projected into $d_{3'}^{-1}$ - $d_{5'}^{-1}$. This panel conveys two important pieces of information. First, each macrostate is anchored around a single free energy minimum containing the preponderance of the probability mass of the microstate ensemble. These projections indicate a relatively narrowly peaked distribution of microstates around the free energy minimum with a broad tail. Second, the location of each macrostate free energy minimum in the $d_{3'}^{-1}$ - $d_{5'}^{-1}$ projection readily illuminates the base pair pattern within the duplex. Within the fully hybridized H state (blue) $d_{3'}^{-1} \approx d_{5'}^{-1} \approx 1.2 \text{ nm}^{-1}$. A 2 b.p. shift to 3' is manifest in a decrease in $d_{3'}^{-1}$ (increase in $d_{3'}$) and an increase in $d_{5'}^{-1}$ (decrease in $d_{5'}$) producing the stable 3S2 state (green). A further 2 b.p. shift to 3' results in a further decrease in $d_{3'}^{-1}$ and also a decrease in $d_{5'}^{-1}$ (back down to the same value as in the H state) to produce the 3S4 state (black). Analogous arguments apply to the 2 b.p. 5' shifted state 5S2 (orange) and 4 b.p. 5' shifted state 5S4 (purple). The separated dehybridized state D exists not at the origin, but, due to the finite size of our simulation box, at $d_{3'}^{-1} \approx d_{5'}^{-1} \approx 0.2 \text{ nm}^{-1}$ (red). The GC-core frayed state F4 (orange) exists as a heart-shaped region between the H and D states, with each symmetric lobe of the heart corresponding to fraying of one or other of the loose ends. (d) Coloring the microstates by their membership probabilities to their assigned macrostate under a fuzzy clustering procedure shows that microstates lying in the vicinity of the free energy minima have high probabilities whereas those at the macrostate cluster boundaries have less definitive assignments. The former comprise the preponderance of the microstate ensemble while the latter represent intermediate configurations that are transiently occupied as the system transitions from one macrostate to another.