SUPPLEMENTAL 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, 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

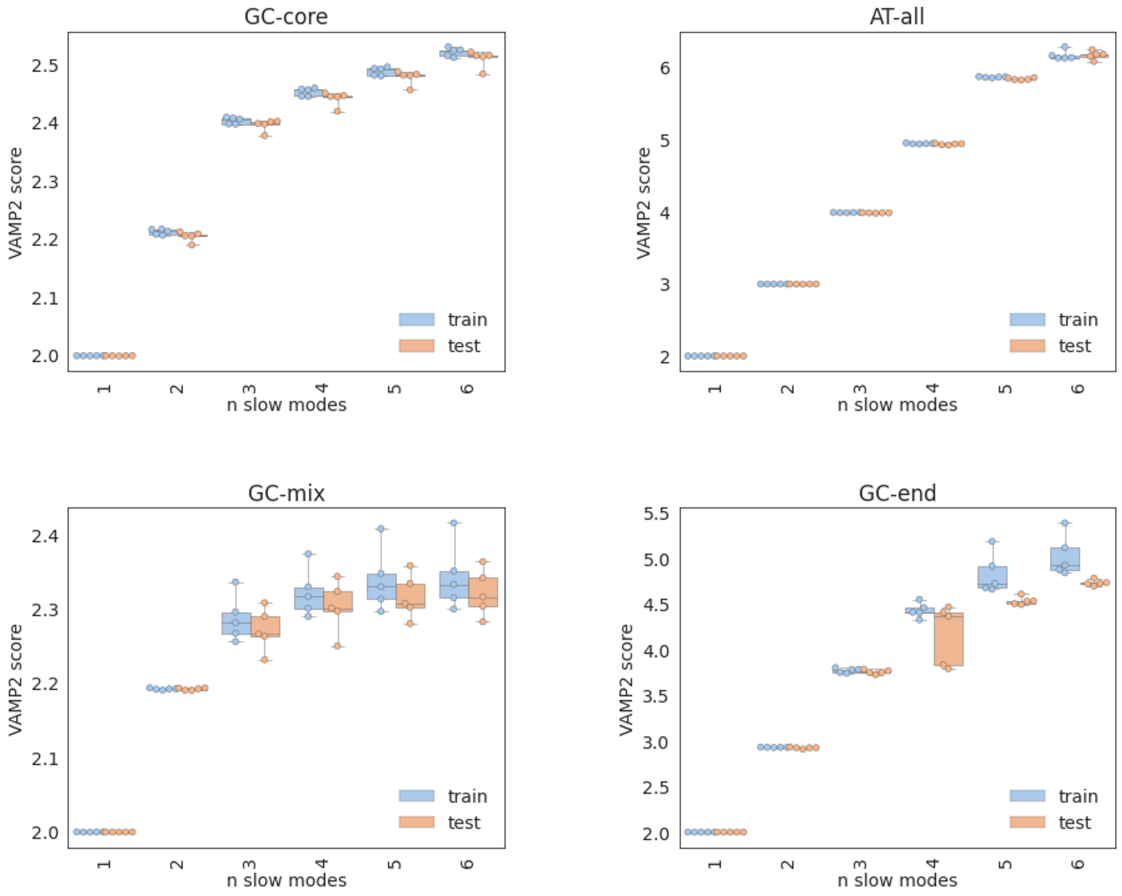


Figure S1: Five-fold cross-validation of the SRV VAMP-2 scores to select the optimal number of SRV coordinates for each sequence. A knee in the VAMP-2 plot at approximately the fifth slow mode for each of the four sequences motivates the choice of a 5D projection. The absence of any significant separation in the training and testing VAMP-2 scores demonstrates that the 5D model is not overfitted.

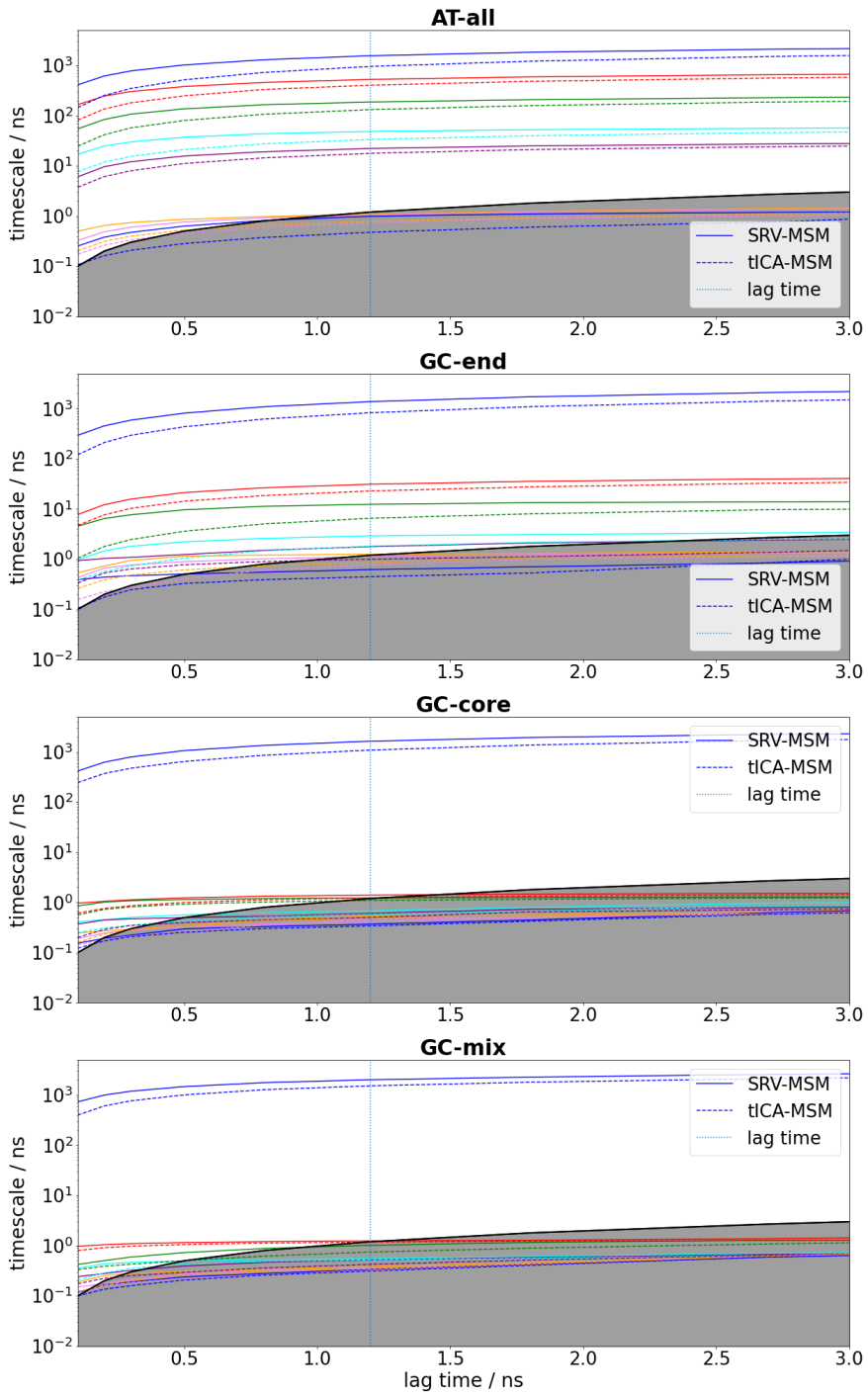


Figure S2: Convergence of the MSM implied time scales t_i as a function of lag time τ. Solid lines indicate maximum likelihood result while dashed lines show the Bayesian ensemble means. The implied time scales for all sequences converge at a lag time of $\tau = 1.2$ ns (vertical line). The black solid 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. [[(i) Eliminate tICA data. (ii) Eliminate legend. (iii) Add Bayesian ensemble uncertainties as colored shading behind the solid line. PyEMMA will do this calculation for you.]]

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 an MSM constructed at the $\tau = 1.2$ ns lag time (dashed blue line) versus those computed from an MSM constructed at the particular lag time (solid black line). The good agreement between these two results provides numerical validation of the Markovian nature of the $\tau = 1.2$ ns lag time MSM.

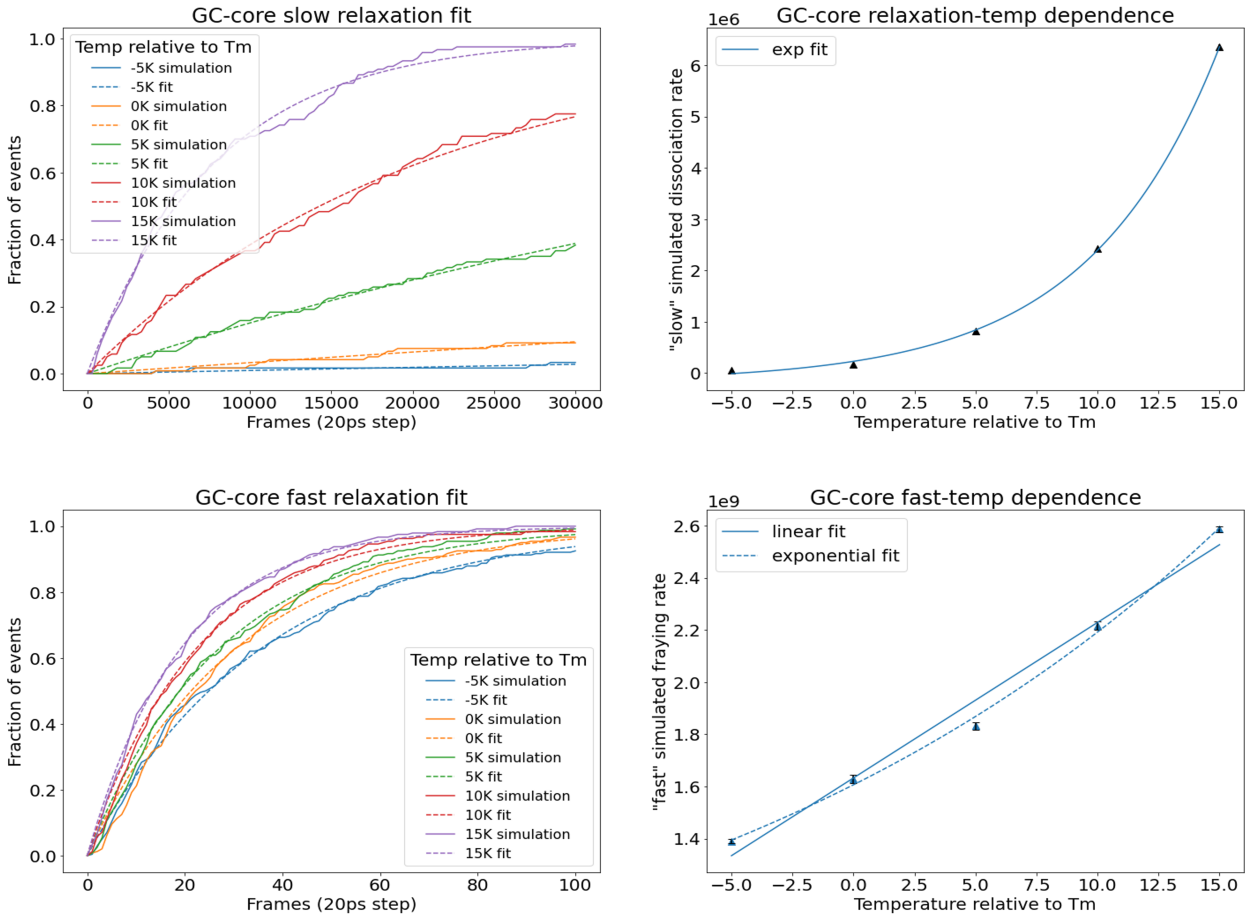
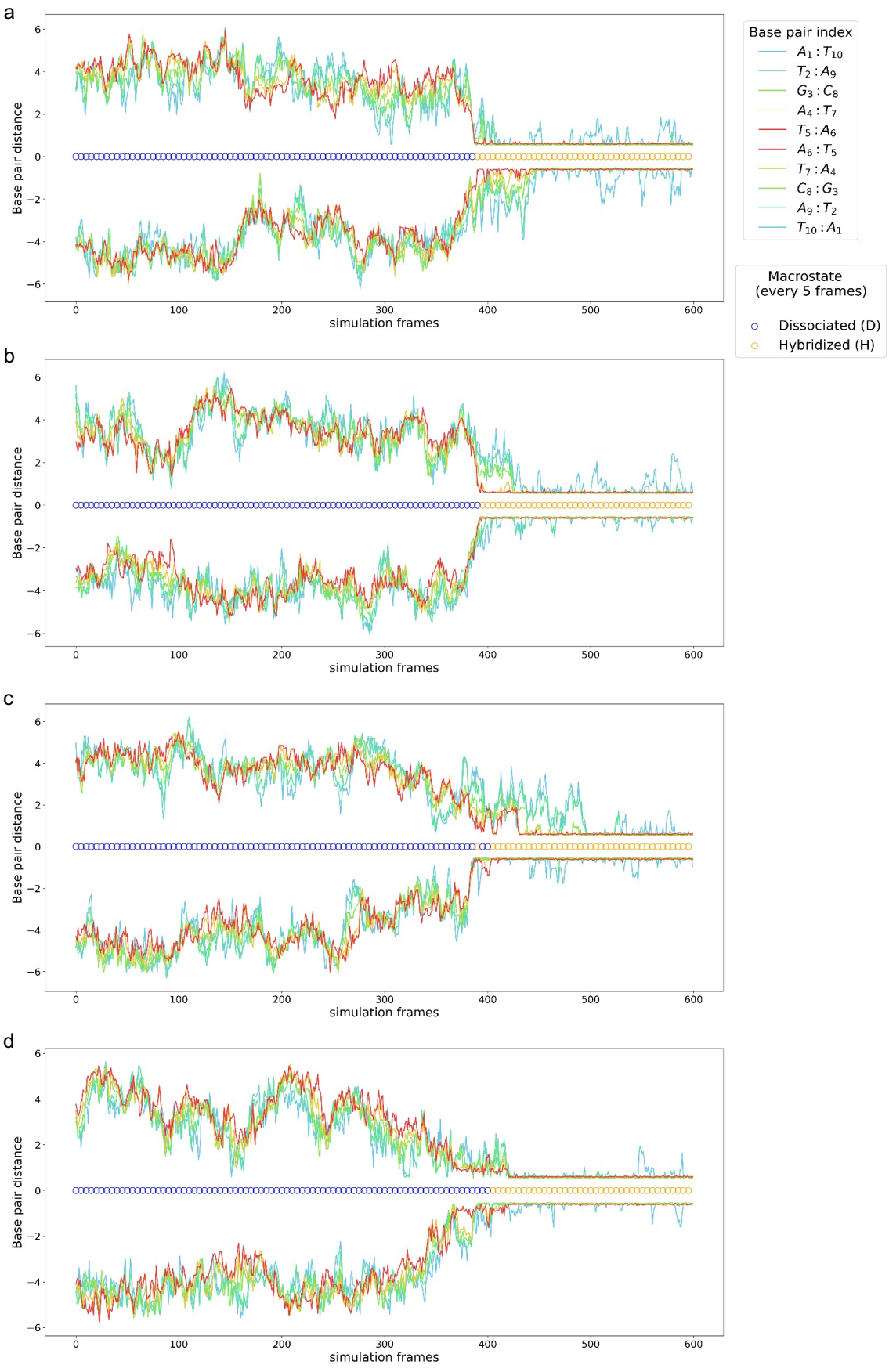
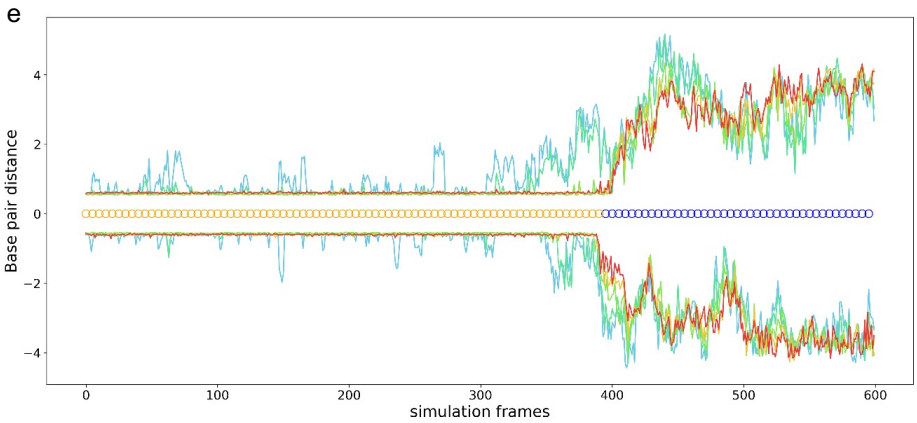
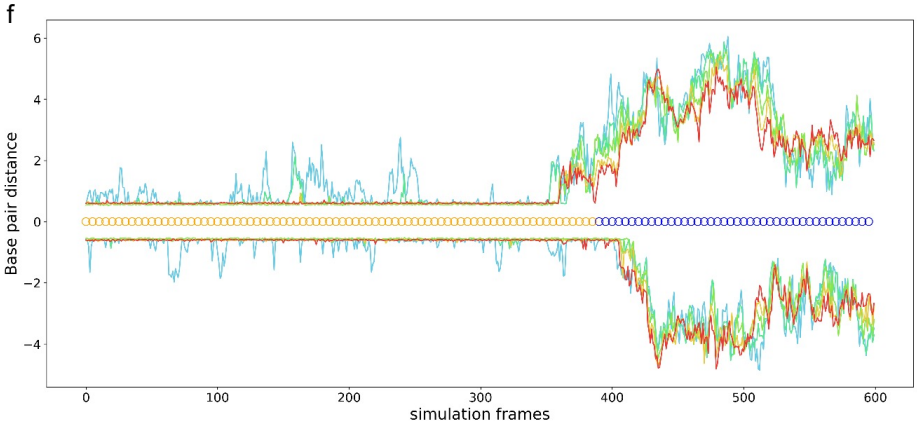
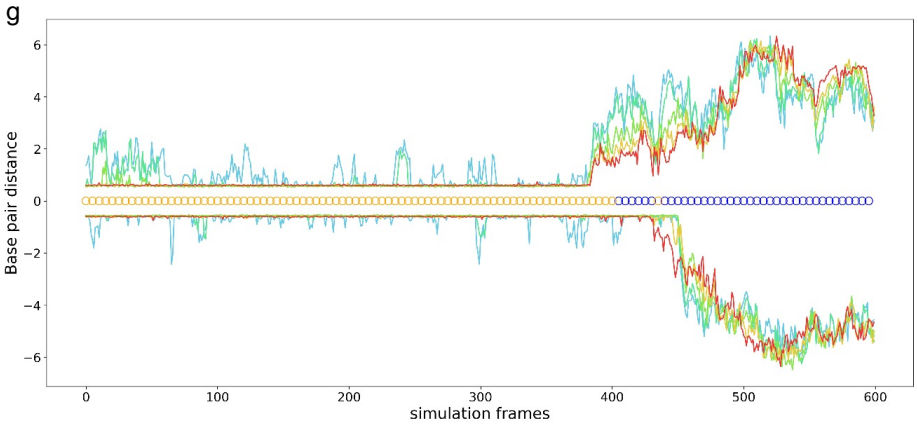


Figure S4: [[Can this be (i) simplified – just show the left panels, the right panels are in the main text, (ii) report the kd values in the legends for the fits and provide the equation in the caption, and (iii) show data for all four seqs. Rewrite the caption and axes labels to reflect the terminology used in the main text (e.g., k_d^{slow}, report Tm by writing it as text within white space of each figure for each sequence).]]


a
b
c
d
Base pair distance
simulation frames
Base pair index
$A_1 : T_{10}$
$T_2 : A_9$
$G_3 : C_8$
$A_4 : T_7$
$T_5 : A_6$
$A_6 : T_5$
$T_7 : A_4$
$C_8 : G_3$
$A_9 : T_2$
$T_{10} : A_1$
Macrostate
(every 5 frames)
Dissociated (D)
Hybridized (H)


e
Base pair distance
simulation frames


f
Base pair distance
simulation frames


g
Base pair distance
simulation frames

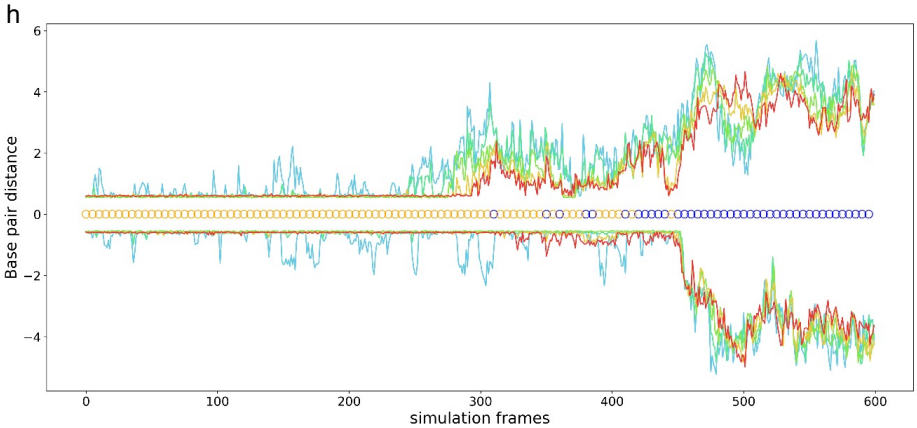

h
Base pair distance
simulation frames

Figure S5: [[Add caption mirroring that from the main text.]]

Nearest neighbor model of duplex thermodynamics

TBA – full accounting complete with mathematics and references for model and its parameters of the application of NN model to predict free energy of each macrostate for each sequence[1]

References

(1) 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.