Supporting Material: Controlled Destabilization of Caged Circularized DNA Oligonucleotides Predicted by Replica Exchange Molecular Dynamics Simulations

Carsten Hamerla,[†] Padmabati Mondal,[‡] Rainer Hegger,[†] and Irene Burghardt[*,†]

[†] *Institute for Physical and Theoretical Chemistry, Goethe University Frankfurt, 60438 Frankfurt, Germany*

[‡] *Indian Institute of Science Education and Research (IISER) Tirupati, Panguru (G.P), Yerpedu Mandal, 517619 - Tirupati Dist., Andhra Pradesh, India*

E-mail: burghardt@chemie.uni-frankfurt.de

S1 Starting structures for REMD simulations

As explained in Sec. 2.2 of the main text, different protocols were used in order to construct the various N-M circularized structures. The relevant starting structures are shown in Figure S1.

For the native system shown in Figure S1a), standard color coding is used for the different bases, i.e., adenine (orange), cytosine (blue), thymine (pink), and guanine (grey). For the circularized structures including a linker, shown in Figure S1b)-f), the color coding was chosen as in Figure 2 of the main text where the linker region is emphasized. In the following, we refer to the circularized strand as the "primary" strand, while the complementary strand, which is unchanged as compared with the native system, is denoted "secondary" strand. The part of the primary strand that forms a double helix with the secondary strand is marked in orange, the dangling part of the primary strand is shown in gray, and the linker atoms are depicted in black. The secondary strand is shown in blue.

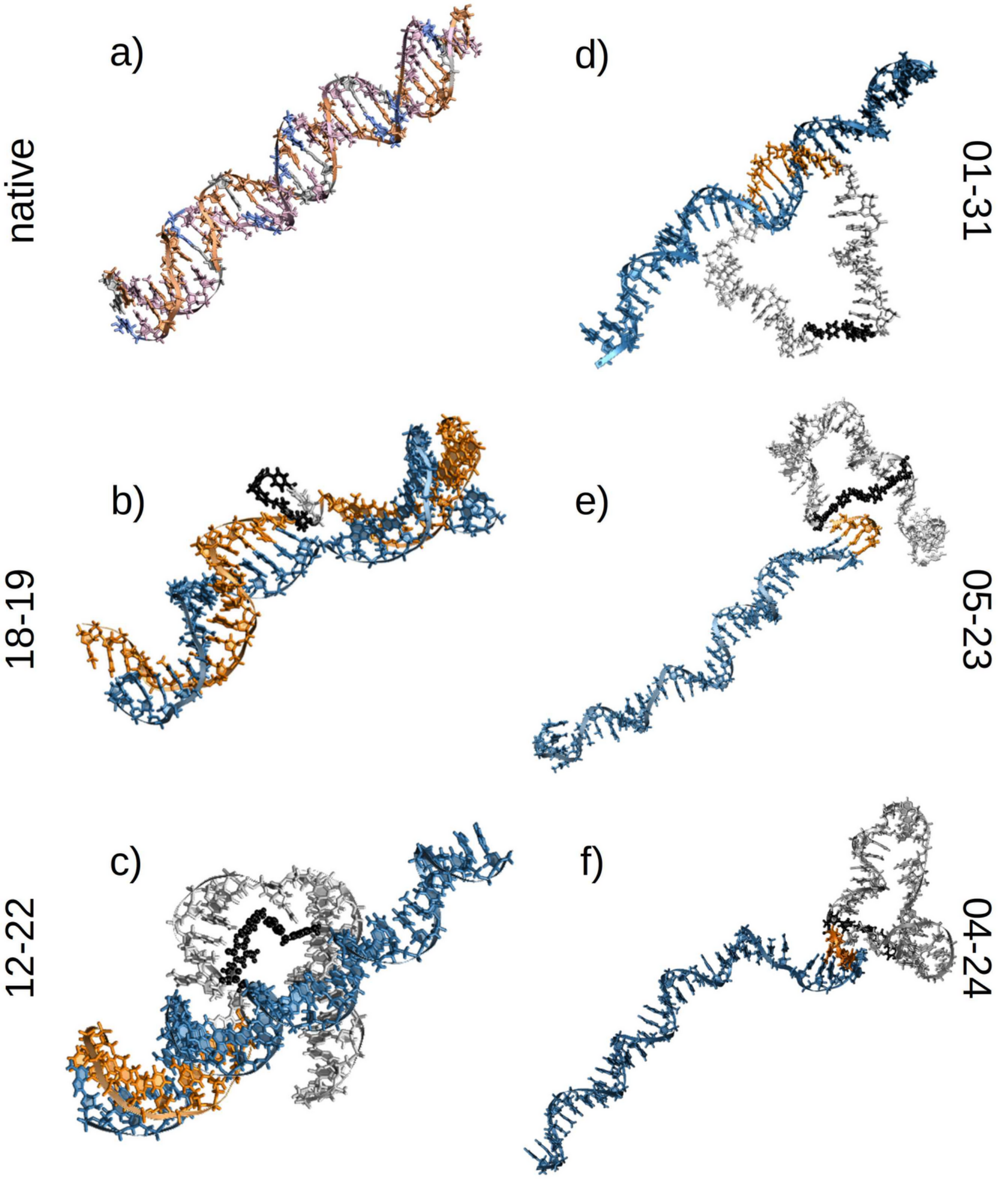

Figure S1: Starting structures employed in the REMD setup for the six systems under study. These structures are available as gro-files as part of the Supplementary Material. a) Native system, b) 18-19 system, c) 12-22 system, d) 01-31 system, e) 05-23 system, f) 04-24 system. The color coding is explained in the text.

S2 Simulation details

As explained in Sec. 2.3 of the main text, simulations for each system were based on a single initial condition defined by the starting structures illustrated in Sec. S1. Each of the structures was placed in a cubic box whose diameter exceeded the size of the system by 1 nm in all directions. The box was filled with TIP3P water and as many Na ions as needed to neutralize the system. The number of water molecules depended on the actual size of the box and ranged from 50 $\times 10^3$ to about 75 $\times 10^3$. A standard equilibration protocol (Coulombic, NVT and NPT) was performed at 300 K. The resulting configuration of the full system was then used as the initial condition for all replicas. The Amber 99 force field with the Barcelona Supercomputing Center (bsc1) modifications[1] was employed. For all simulations, the GROMACS package,[2,3] version 2020.4, was used.

All runs were performed with the standard integration time step $\Delta t = 2$ fs. After each 1 ps interval, the condition for replica exchange in temperature space was checked. The configuration of the DNA system and the energy of the full system was written every 10 ps and the full system state was saved every 100 ps. Since we started with the same configuration for all temperatures, an initial equilibration phase occurs for the higher temperatures. The resulting equilibration time was identified as the time t_p where a plateau of the stability parameters sets in, as detailed in Sec. 3.1 of the main text. The plateau time differs significantly between the individual systems, see Table S1.

After equilibration, each system was propagated under replica exchange during a production run time t_prod. As in the case of the plateau time t_p, the temporal evolution of the stability parameters determined t_prod: If at a given time the stacking fraction (see Eq. (7) of the main text) reached zero for the highest simulation temperature, or at least was found to be close to zero, the simulation was stopped. This criterion did not apply to the native system: Here, we could not reach zero and it was clear from the progress of the stacking fraction that this was not achievable within a reasonable time.

Table S1: Onset of the plateau time t_p, production run time t_prod, and the sum $t_\mathrm{sim} = t_\mathrm{p} + t_\mathrm{prod}$ for all systems per replica. The last column shows the CPU times for single replicas.

system	t_p [ns]	t_prod [ns]	t_sim [ns]	CPU time [day]
native	30	32	62	88
18-19	12	24	36	52
12-22	8	10	18	26
01-31	8	14	22	31
05-23	10	13	23	33
04-24	7	28	35	50
Sum	75	121	196	280

Table S1 shows the values of t_p and t_prod as well as the overall simulation time $t_\mathrm{sim} = t_\mathrm{p} + t_\mathrm{prod}$ for each system. The last row gives the corresponding sums over all systems. Since for each system 110 replicas were simulated in parallel, the overall simulation time added up to 110 × 196 ns = 21.56 μs for all 660 replicas.

The simulations were performed on a cluster with Intel E5-2690 v4 CPUs with a Linux operating system. Each replica was simulated in a single thread. On average, we could propagate a replica by about 0.7 ns per CPU day, yielding a total CPU time of about 84

CPU years. The last column of Table S1 contains the CPU times for a single replica for each system. In our setup, CPU times are nearly identical to real (elapsed) times.

S3 Replica exchange acceptance ratio

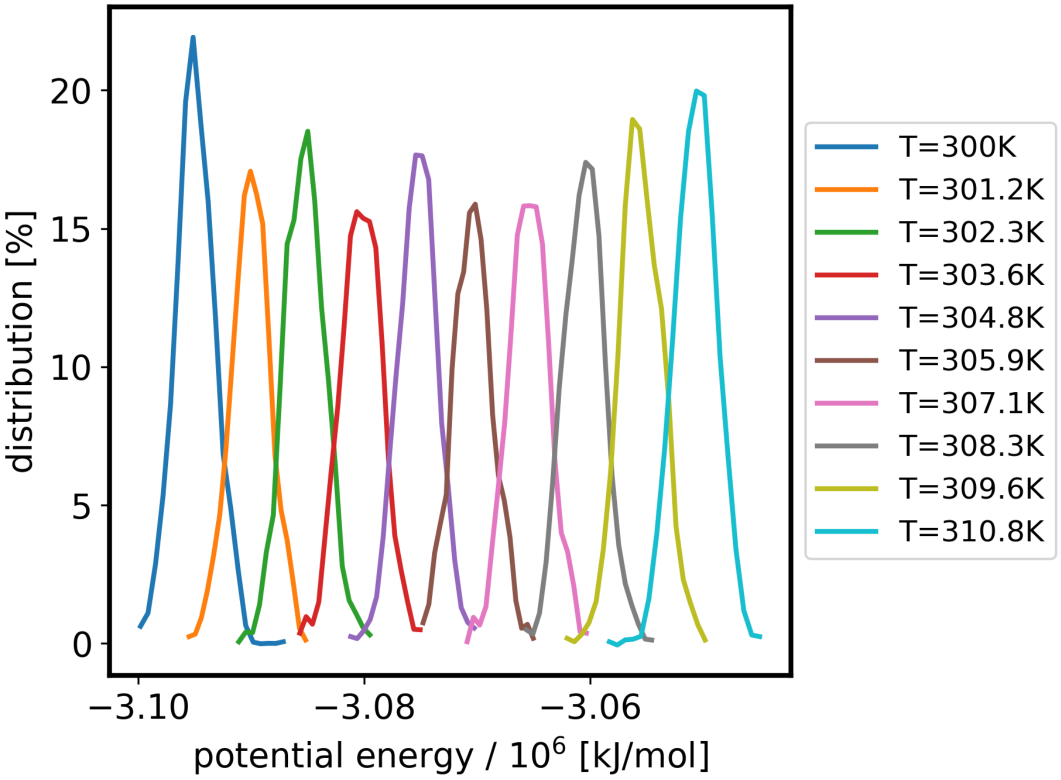

Figure S2: Distribution of the potential energies of the ten replicas with the lowest temperatures of the REMD simulation for the native system. The temperatures for these replicas, indicated in the figure, vary between 300.0 K and 310.8 K. The small temperature spacing of 1.2 K between neighboring replicas was necessary to obtain sufficient overlap of the potential energies for an acceptance ratio of about 4%.

Figure S2 shows the distribution of potential energies for the first ten replicas out of 110 replicas that were used in the calculations. The overlap between these distributions reflects the probability of exchange between adjacent configurations, and, hence, the acceptance ratio in the REMD simulation. The acceptance ratio estimated from the simulations is given as the ratio between the successful exchange attempts and the number of all exchange attempts. In our set-up, an acceptance ratio of 4% is obtained.

The acceptance ratio of 4% is smaller than typical values of about 10%-30% which are considered optimal.[4–7] The reason for the small overlap of neighboring distributions lies in the size of the systems. Neglecting correlations, one can qualitatively argue as follows: On the one hand, the standard deviation of energy fluctuations is proportional to $\sqrt{N}$, where N is the number of particles in the system (noting that the standard deviation scales as $1/\sqrt{N}$ which is multiplied by N to obtain the standard deviation of the N-particle energy). On the other hand, the distance between the peaks of neighboring distributions is approximately proportional to $Nk_B\Delta T$. Due to this difference in scaling, the temperature spacing ΔT needs to decrease as the system size increases, such as to obtain a reasonable overlap of the distributions. Due to limitations imposed by the computational effort, the trade–off between small ΔT and computational effort results in a rather small acceptance ratio in our calculations.

In order to assess whether the small acceptance ratio of about 4% influences the efficiency of the simulations on the time scales we can reach, we recorded the diffusion of configurations through temperature space, see Figure S3a). To achieve a mixing effect, a configuration that started at a given temperature T should repeatedly visit all other temperatures in the course of the simulation time. Figure S3 shows this process for four different starting temperatures. Even though the configurations are not confined to a limited temperature range around the starting temperature, it is clear that the diffusion of a trajectory through temperature space is not efficient. The diffusion process is not converged in terms of the statistics of a random process, such that much longer trajectories – likely of the order of several hundred ns – would be needed to see a proper mixing.

As an additional test, we increased the number of replicas to 224, leading to a larger overlap of the energy distributions and an acceptance ratio of about 20%. For the – very short – simulation time of 18 ns where a comparison could be made, the stability parameters follow a very similar profile as in the case of 110 replicas, see Figure S4. This suggests that the small acceptance ratio of 4% in the case of 110 replicas likely does not introduce a bias on our propagation time scale. However, the results shown in Figure S3b) for the diffusion through temperature space also suggest that increasing the number of replicas does not significantly enhance the diffusion. Indeed, an increase in the number of replicas reduces the efficiency of sweeping the temperature space, due to the increased number of swaps of REMD configurations. As a result, convergence becomes highly challenging in large systems.[8,9] This is likely one major reason for the overestimation of T_m values in our simulations.

Finally, in Figure S5 we show a comparison between single temperature MD simulations (black traces), and REMD trajectories (orange traces) at the same temperature. The left-hand-side plot shows results for T = 430 K, while the right-hand-side plot shows results for the highest temperature T = 450 K. Much larger fluctuations are found to occur in the REMD trajectories as compared with the MD trajectories, which reflects the exchange with other configurations in the REMD case. For the left-hand-side plot, illustrating trajectories at T = 430 K, the MD trajectory always exhibits minimal fluctuations as compared with the REMD trajectory, and always stays at the lower end of the fluctuation spectrum of the REMD trajectory. This implies that the melting process will be accelerated in the REMD simulation, as expected. Conversely, though, the MD trajectory at T = 450 K is found at

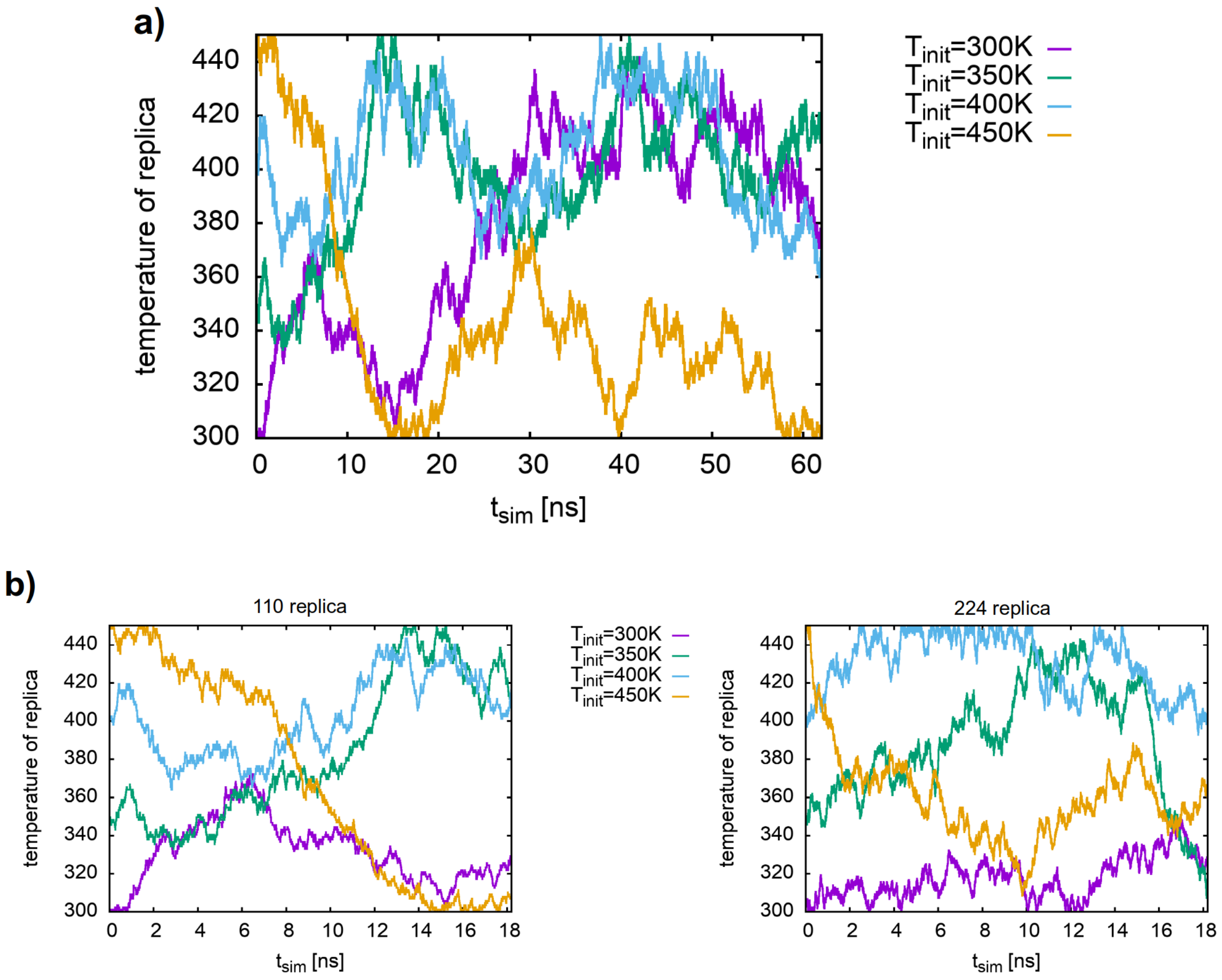

Figure S3: (a) Diffusion of four REMD configurations through temperature space (taken from the simulations comprising 110 replicas). The starting temperatures were 300 K, 350 K, 400 K, and 450 K, respectively. This analysis suggests that the REMD trajectories do not exhibit confinement to a narrow temperature range, but the diffusion process is inefficient on the propagation time scale. (b) Comparison between the diffusion processes in temperature space for the system comprising 110 replicas and an additional REMD set-up comprising 224 replicas, on a short time scale of 18 ns (see also Figure S5). It is seen that the larger number of replicas does not enhance the diffusion process on the observation time scale.

the upper end of the fluctuation spectrum of the REMD trajectory, such that the latter exhibits, on average, a higher stability than the MD trajectory. This can be explained by the mixing of the REMD trajectory with lower-temperature – more stable – replicas, in the absence of higher-temperature replicas. From these observations, we expect that the REMD simulations should overall have a “smoothing” effect on the melting profiles, while a corresponding MD analysis would likely exhibit a sharper profile.

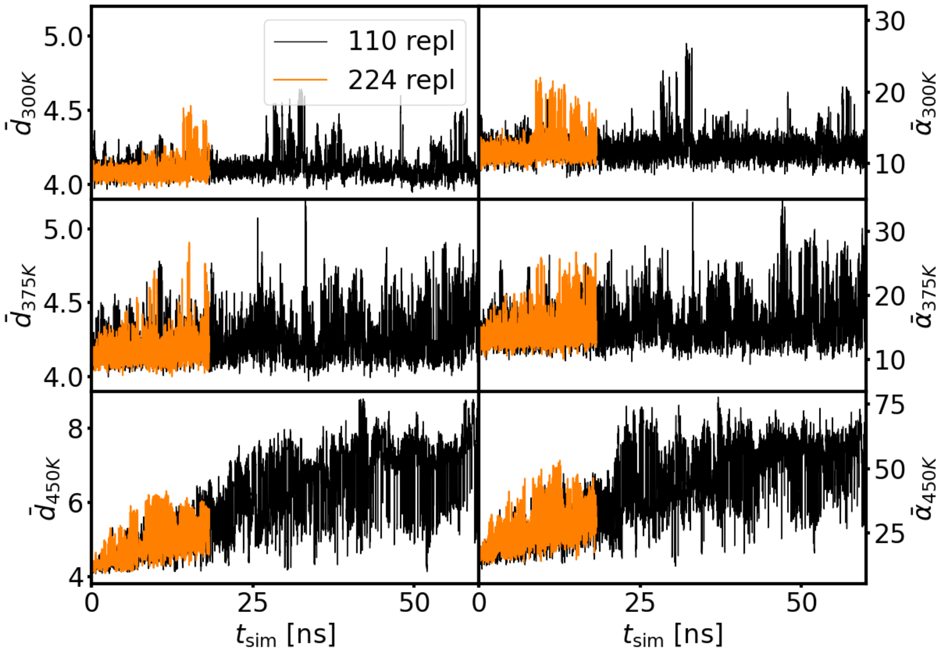

Figure S4: Comparison of the stability parameters $(\bar{d}, \bar{\alpha})$ for the native system, for two different REMD setups, including 110 replicas (black curves) and 224 replicas (orange curves), respectively. For the latter set-up, the acceptance ratio is increased from 4% to about 20%; however, the simulation time was restricted to 18 ns. Three different REMD realizations, with starting temperatures $T = 300$ K, $T = 375$ K, and $T = 450$ K, are shown.

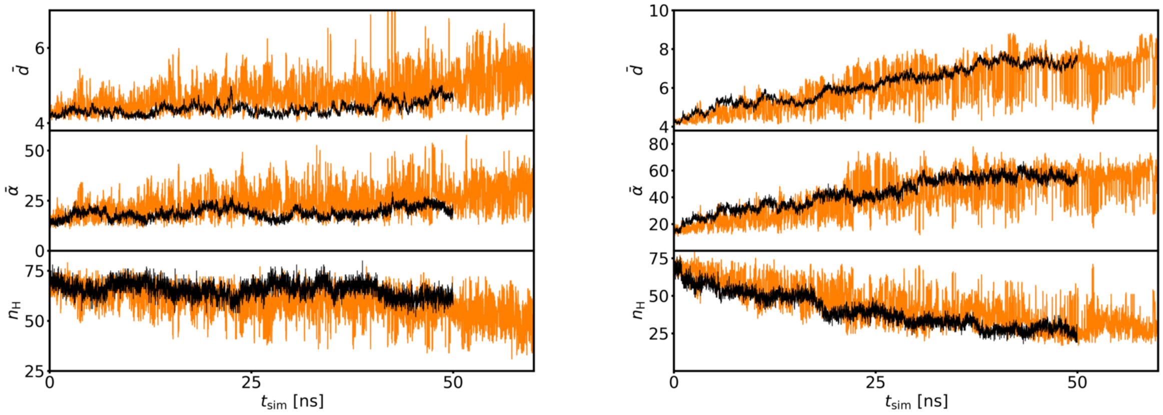

Figure S5: Comparison of the stability parameters $(\bar{d}, \bar{\alpha}, n_{\mathrm{H}})$ for single temperature MD simulations (black traces) and corresponding REMD "trajectories" (orange traces) for the native system. The l.h.s. plot corresponds to $T = 430$ K while the r.h.s. plot corresponds to $T = 450$ K. The initial configuration was the same in all cases.

S4 Analysis of hydrogen bonding fractions

In Sec. 2.4 of the main text, we introduced stability parameters relating to both stacking and hydrogen bonding interactions. In the case of the native system, it was shown that these stability parameters yield similar results (see Sec. 3.2 and Figure 6 of the main text). However, we mentioned that in the case of the circularized systems, the hydrogen bonding parameter does not yield meaningful results, such that the comparative analysis of Sec. 3.2 was conducted exclusively for the stacking parameters. Here, we illustrate why the melting temperature analysis based on the hydrogen bonding parameter is not reliable in the context of the circularized systems.

To this end, Figure S6 shows again the melting curves shown in Figure 7 of the main text, while adding the melting curves that would be obtained based upon the hydrogen bonding parameter n_{H} (see Sec. 2.4). It is seen that among the circularized systems, only the 18-19 system – with the minimal loop size – exhibits a profile of the hydrogen bonding fraction P_{H} that resembles the stacking fractions P_d and P_α. For the 18-19 system, this is obviously the expected result.

For the other circularized systems, two trends are observed: (i) For the 12-22 and 01-31 systems, the melting profile for the hydrogen bonding fraction P_{H} appears shifted to much higher T_{m} values than those obtained from the stacking fractions, (ii) for the 05-23 system and the 04-24 system, the stacking fractions do not yield a typical melting curve at all. In the following, we summarize the reasons for these observations.

In the 12-22 and 01-31 systems, the initial number of hydrogen bonds is smaller than in the native system (or the 18-19 system), but yields an acceptable statistics. Nevertheless, a drastic overestimation of the T_{m} values results from the P_{H} analysis: The apparent melting temperatures based on P_{H} are augmented by 40-50 K as compard with the result obtained from the stacking fractions, as can be inferred from Figure S6. The reason lies in the primary structure of these systems, which exhibit a comparatively high number of C–G pairs in the double helix region, as compared with the native system. In contrast to A–T Watson–Crick base pairs which possess two hydrogen bonds, C–G base pairs feature three hydrogen bonds and tend to be more stable. As can be inferred from Figure 5 of the main text, C–G base pairs do not fully open even at T_{max}. Therefore these systems appear more stable than the native system based on the P_{H} analysis. As a result, artificially high T_{m} values are obtained from the hydrogen bonding fraction P_{H}.

In the 04-24 system and the 05-23 system, the situation is different since these systems exhibit a very short double helix region, with a small number of initial hydrogen bonds. This leads to two effects. Firstly, the statistical fluctuations are much stronger as is clearly visible in Figure S6. Secondly, since the two strands do not separate completely within the simulation time, random hydrogen bond formation occurs. Due to the small initial number of hydrogen bonds, this effect gives rise to a large bias in the estimated temperature.

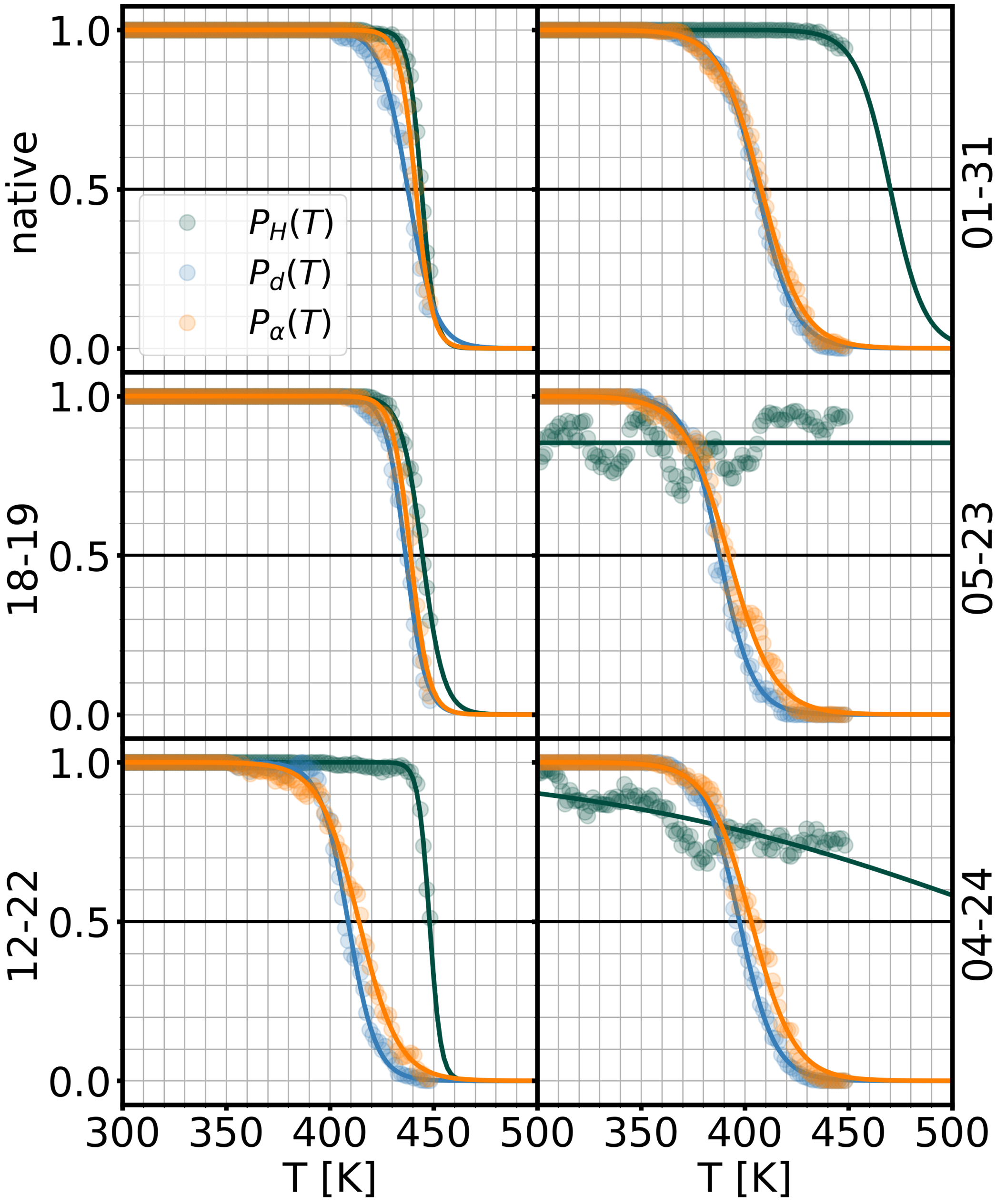

Figure S6: Analogously to Figure 7 of the main text, melting curves of the native system and all circularized systems, are shown. In addition to the distance fraction P_d and the tilt fraction P_α, the hydrogen bond fraction P_{H} is also shown (green traces).

S5 Alternative computation of stacking fractions

As detailed in Sec. 2.4 of the main text, the computation of the stacking fractions shown in Figure 6 and Figure 7 of the main text involves several averages. First, an average is taken over all nucleotides (see Eq. (5) of the main text), followed by a time average that leads to the fractions of Eq. (6) of the main text. In further detail, the procedure is as follows, e.g., for the stacking parameters: For a given configuration of the system, the stacking parameters $\alpha_i(t,T)$ and $d_i(t,T)$ are initially averaged over all pairs $i = 1, \ldots, 62$ such as to yield the averages $\bar{d}(t,T)$ and $\bar{\alpha}(t,T)$. Next, time averages are computed and the quantities $\langle n_d \rangle(T)$ and $\langle n_\alpha \rangle(T)$ are obtained, which count the number of instances where the native stacking interactions are preserved. To this end, threshold values are introduced as detailed in the main text. The fractions $P_d(T)$, $P_\alpha(T)$ are then obtained as the ratio between these time averages $\langle n_d \rangle(T)$ and $\langle n_\alpha \rangle(T)$ and the reference value obtained for the native interactions in the lowest-temperature simulation. An analogous procedure is used to obtain the hydrogen bond fractions.

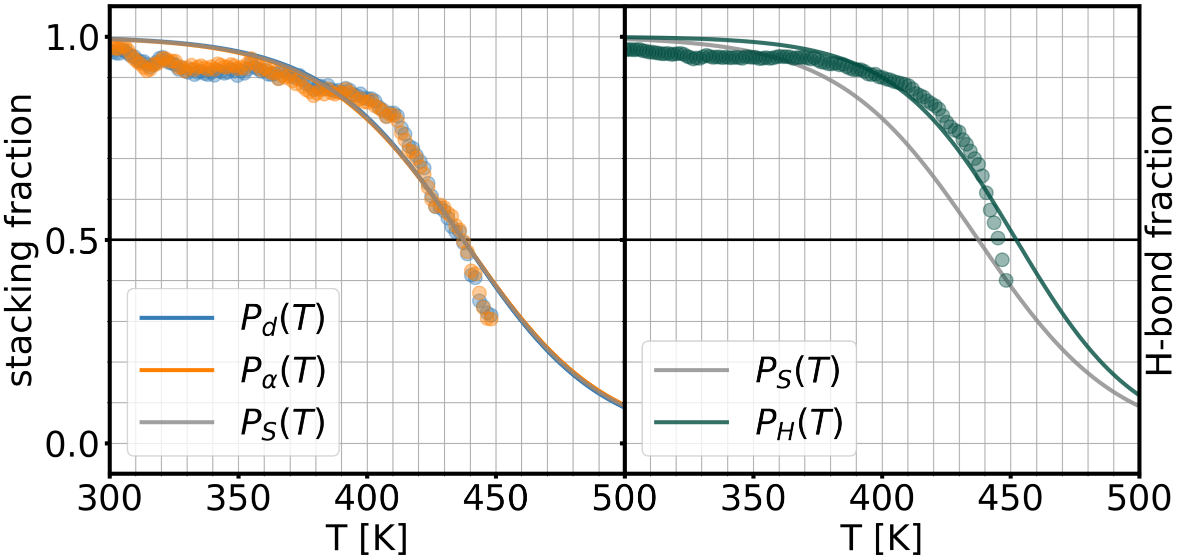

Figure S7: Melting curves for the native system, computed with the alternative protocol described in the text. Stacking fractions and hydrogen bond fractions for the native system are shown, as in Figure 6 of the main text. In the alternative scheme described in the text, time averages are carried out for each nucleotide pair.

However, the sequence of the two averages that are taken – i.e., an average over nucleotides and a time average – could introduce a bias. Therefore, we compare here with an alternative approach where the time average is taken for each nucleotide pair, followed by an average over nucleotide pairs. For the stacking parameters, this alternative sequence implies that for each base pair, a time average is taken and a local stacking fraction is computed. Finally, an average is taken over these local stacking fractions. Figure S7 shows the corresponding fractions. Clearly the results differ from those shown in Figure 6 of the main text for the native system. Conspicuously, the plateau for lower temperatures has disappeared. Apparently, stacking fractions tend to be underestimated at low temperatures, but overestimated at high temperatures. This is likely due to the fact that fluctuations are felt more strongly at the level of the individual base pairs, and would tend to make a local base pair appear less stable at low temperatures – but at high temperatures, fluctuations would rather tend to suggest higher local stability. As shown in Figure S7 for the native system, the melting point is obtained in good agreement with the procedure presented in the main text. Since the latter averaging procedure is far better compatible with a sigmoidal fit, this is the scheme which has been employed throughout in our analysis.

S6 T_{m} analysis based on fixed simulation times

The duration of the simulations was not the same for all systems, as can be seen from Table S1. The reason lies in the melting profiles of the different systems, i.e., the simulations were stopped when the fractions P_α and P_d approached or reached zero. As a result, the statistics for the melting temperature analysis differs between the six systems. To assess the effect of the different simulation times, we made an estimation of the melting temperatures for data sets with the same duration. To this end, we restricted the analysis of the data to the first 10 ns starting from the onset of the plateau $[t_p, t_p + 10\,\mathrm{ns}]$. The resulting T_{m} values are shown in Table S2 and compared to the data of Table 1 of the main text. The differences can be explained by statistical fluctuations.

Table S2: T_{m} values based on the first 10 ns after the onset of the plateau (t_p), in comparison with the data of Table 1 of the main text. The numbers indicated in brackets are deviations from the temperatures as obtained from the full data sets.

system	T_{m}^d	T_{m}^α	T_{m}^S
native	443 (4)	445 (4)	444 (5)
18-19	443 (7)	446 (7)	445 (7)
12-22	409 (0)	414 (0)	411 (0)
01-31	407 (1)	411 (3)	409 (2)
05-23	388 (1)	395 (3)	391 (1)
04-24	400 (3)	408 (5)	404 (4)

S7 Melting temperatures for individual strands

The T_{m} values reported in the manuscript and in the preceding sections of the Supplementary Material are based on an analysis of both strands. Due to the very different character of the circularized (primary) strand and the native-like secondary strand, one could conjecture that the stacking fractions yield different melting profiles for the two strands taken separately.

Therefore, we here report calculation results of T_m values based on stacking fractions for the individual strands of the various systems. The analysis for the individual strands as compared with the analysis for both strands, as carried out in the main text, is summarized in Table S3. It is seen that the differences between the values calculated from the primary strand and the ones from the secondary strand are system dependent, and can become quite pronounced. However, there is no systematic trend – i.e., melting temperatures obtained from the secondary strand can be both higher or lower as compared with the average.

Table S3: Melting temperatures obtained from the distance and tilt parameters reproduced from the main text (columns 2 and 3), as compared with analogous results computed for the secondary (sec) or primary (prim) strand taken separately. The secondary strand is unmodified and can be taken as identical for all systems. Differences with respect to the calculation for the full system are indicated in brackets.

system	T_m^d	T_m^α	$(T_\mathrm{m}^d)^\mathrm{sec}$	$(T_\mathrm{m}^\alpha)^\mathrm{sec}$	$(T_\mathrm{m}^d)^\mathrm{prim}$	$(T_\mathrm{m}^\alpha)^\mathrm{prim}$
native	437	441	438 (1)	441 (0)	437 (0)	442 (1)
18-19	436	439	436 (0)	439 (0)	436 (0)	438 (-1)
12-22	409	414	406 (-3)	408 (-6)	416 (7)	426 (12)
01-31	406	408	409 (3)	406 (-2)	401 (-5)	410 (2)
05-23	387	392	395 (8)	396 (4)	382 (-5)	389 (3)
04-24	397	403	394 (-3)	393 (-10)	398 (1)	410 (7)

S8 Free energy landscapes

In Figure 8 of the main text, free energy profiles are shown as a function of the distance parameter $\bar{d}$ and the tilt parameter $\bar{\alpha}$, for all relevant systems at different temperatures (T = 300 K, T = 350 K, T = 400 K and T = 450 K). These free energy landscapes include the initial nonequilibrium phase preceding the plateau time t_p (see Sec. 3.1 of the main text and Table S1), and hence, illustrate a pseudo-dynamics of the melting process. Here, we separate the initial transients ($t < t_\mathrm{p}$) from the equilibrated phase ($t > t_\mathrm{p}$) and show the corresponding free energy landscapes in Figure S8 (see Table S1 for the times $t \geq t_\mathrm{p}$ for the individual systems). It is seen that the ($\bar{d}$, $\bar{\alpha}$) regions reached beyond $t = t_\mathrm{p}$ are clearly distinct from the initial transients, reflecting that the melting process takes the various systems to regions of configuration space which feature large $(\bar{d}, \bar{\alpha})$ values. The trends are in line with the analysis of the melting curves for the different systems.

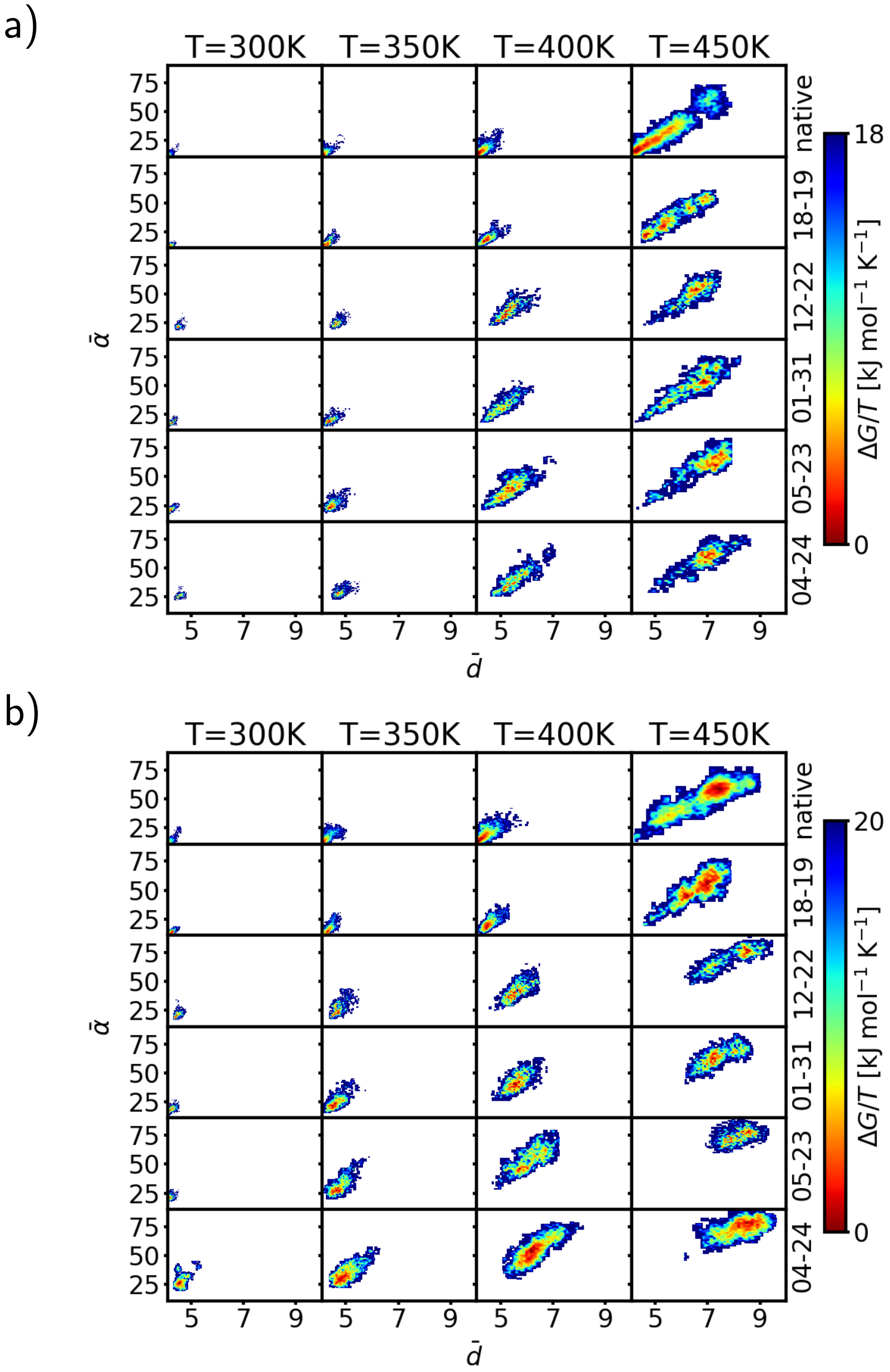

Figure S8: Analogously to Figure 8 of the main text, free energy landscapes are shown: (a) Free energy landscape computed for the transient period $t < t_\mathrm{p}$. (b) Free energy landscape computed for times $t > t_\mathrm{p}$, i.e., during t_prod. See Table S1 for the definition of t_p and t_prod for the individual systems.

References

(1) Ivani, I. et al. Parmbsc1: a refined force field for DNA simulations. *Nat Methods* **2016**, *13*, 55–58, Number: 1 Publisher: Nature Publishing Group.

(2) Bekker, H.; Berendsen, H.; Dijkstra, E.; Achterop, S.; Vondrumen, R.; Vanderspoel, D.; Sijbers, A.; Keegstra, H.; Renardus, M. GROMACS - A PARALLEL COMPUTER FOR MOLECULAR-DYNAMICS SIMULATIONS: 4th International Conference on Computational Physics (PC 92). *PHYSICS COMPUTING '92* **1993**, 252–256, Place: SINGAPORE Publisher: World Scientific Publishing.

(3) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **1995**, *91*, 43–56.

(4) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **1999**, *314*, 141–151.

(5) Zhang, W.; Wu, C.; Duan, Y. Convergence of replica exchange molecular dynamics. *J. Chem. Phys.* **2005**, *123*, 154105, Publisher: American Institute of Physics.

(6) Sindhikara, D.; Meng, Y.; Roitberg, A. E. Exchange frequency in replica exchange molecular dynamics. *J. Chem. Phys.* **2008**, *128*, 024103, Publisher: American Institute of Physics.

(7) Qi, R.; Wei, G.; Ma, B.; Nussinov, R. Replica Exchange Molecular Dynamics: A Practical Application Protocol with Solutions to Common Problems and a Peptide Aggregation and Self-Assembly Example. *Methods Mol Biol* **2018**, *1777*, 101–119.

(8) Kim, J.; Keyes, T.; Straub, J. E. Replica Exchange Statistical Temperature Monte Carlo. *Journal of Chemical Physics* **2009**, *130*, 124112.

(9) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *Journal of Chemical Physics* **2002**, *116*, 9058.