Secondary–structure clustering of nucleic acid melting: Pseudo–random DNA

Swapnil Baral[1,2] and Michael Zwolak[1,*]

[1] *Biophysical and Biomedical Measurement Group, Microsystems and Nanotechnology Division, Physical Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA*

[2] *Department of Chemistry and Biochemistry, University of Maryland, College Park, MD, USA*

Biomolecular structural disorder can occur at all levels, from atomistic and secondary structures to tertiary formations and complexes. This disorder poses challenges for characterizing biomolecular behavior and function, as well as predictions, especially with all–atom models. Mapping atomistic or coarse–grained ensembles to secondary structures, though, removes entropic disorder due to flexible regions and atomic motion. What remains is a set of secondary structures with probabilities modified by the discarded atomistic configurational entropy. We further develop clustering based on this insight and apply it to the melting of pseudo–random, single–stranded DNA. Even without a well–defined fold, secondary–structure clustering, here using k–means, identifies order via common hybridization patterns. This includes a residual stem feature at high temperature with a probability following a Boltzmann factor. Moreover, we show how the evolution of clusters versus conditions helps refine the coarse graining of the structural space, supporting a view that clustering needs to capture the flow of information during a physical process. Overall, this method clarifies behavior during melting and advances the conceptual foundation of clustering. Its ability to perform despite disorder suggests it could be useful in other contexts, such as for intrinsically disordered proteins.

I. INTRODUCTION

Nucleic acid secondary–structure motifs and ensembles are essential for many biological functions, such as ribozyme activity and riboswitching [1–3]. They also play a key role in drug development targeting RNA [4–8]. As a result, substantial efforts have went into RNA secondary–structure prediction [9–15], including comparisons with experiments such as selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) [16, 17], dimethyl sulphate (DMS) [18, 19], and nuclear magnetic resonance (NMR) [20, 21]. Additionally, single–stranded DNA folding can alter assembly in bottom–up fabrication [22–26].

Predictions often directly use secondary–structure models—e.g., Ising–like models. These can produce a minimum free energy (MFE) structure as the expected fold [9, 11, 13] or directly target ensemble–level information, e.g., maximum expected accuracy [14, 15]. In the context here, Ding et al. developed statistical sampling and clustering to identify relevant folds within the ensemble, where the MFE may not always be the most representative [27–30]. Alternatively, deep learning counterparts both within the same thermodynamic framework [31] and outside of it [32, 33], are rapidly advancing.

Yet, there are many challenges. These include predicting structural response to cellular conditions in drug targeting [4–8], the role of dynamical factors in folding [34], and how competition with other nucleic acid species influences self–assembly [23–26]. All–atom molecular dynamics can address these challenges and has undergone extensive development for RNA dynamics and folding [35–37]. Some tools from secondary–structure prediction have not been incorporated into molecular dynamics. In particular, Ding et al. employ the base–pair distance for clustering to improve over MFE–based nearest–neighbor predictions [28] and to characterize ensembles [30]. The base–pair distance is also part of other secondary–structure packages to help analyze structures [11, 29].

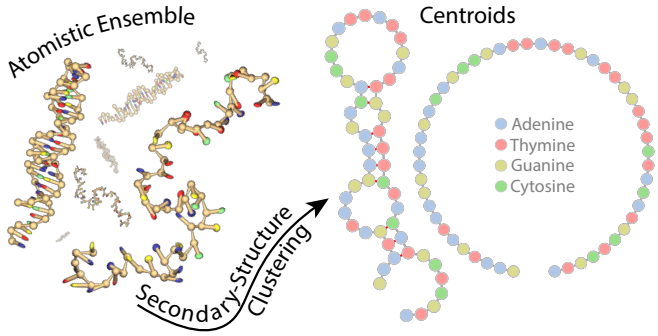


FIG. 1. **Clustering of melting.** Schematic of secondary–structure clustering of atomistic (here, coarse–grained atomistic) configurational ensembles. The centroids are from k–means with $k = 2$ near the melting temperature where the clusters have nearly equal probability. Clustering helps identify important folds/motifs, as well as partitions the ensemble.

We develop secondary–structure clustering to analyze molecular dynamics simulations. Specifically, we use the base–pair distance to quantify the dissimilarity between structures and drive clustering. This approach provides a hierarchical view of the ensemble: it captures atomistic details and response to molecular changes, while characterizing ensembles at a higher structural level that removes entropic disorder from flexibility. This provides an algorithmic approach to identify structural motifs and routes for assessing ensemble convergence and connecting simulations with experimental observables.

* mpz@nist.gov

II. MATERIALS AND METHODS

A. Secondary structure distance

We employ a base–pair distance variant,

$$d_{ij}^2 = n_i + n_j - 2n_{ij} \equiv \frac{1}{2}\mathrm{tr}\left|\Theta^i - \Theta^j\right|^2, \qquad (1)$$

where the squared distance is the number of bases that have to be broken and formed to convert between structures. We note that Ref. [28] employs a definition equivalent to $d_{ij} = n_i + n_j - 2n_{ij}$. Either definition furnishes a distance metric. We consider Eq. (1) so that k–means, which minimizes d_{ij}^2, minimizes the expression $n_i + n_j - 2n_{ij}$, although clustering techniques that minimize an L_1 distance could be an alternative, see Ref. [38] for an extended discussion of this and the clustering method here. This can be computed by taking the number of base pairs n_i in structure i and n_j in j, and subtracting twice the number of base pairs in common, n_{ij}. One can also employ the *secondary–structure matrix* $\mathbf{\Theta}^i$,

$$\mathbf{\Theta}^i_{pq} = \begin{cases} 1 & p \text{ paired with } q \text{ in } i \\ 0 & \text{otherwise,} \end{cases} \qquad (2)$$

and the Hilbert–Schmidt norm, $|\mathbf{A}|^2 = \mathrm{tr}\,\mathbf{A}^\dagger\mathbf{A}$, for operator $\mathbf{A}$. The matrix elements specify whether base p is paired with q, where we consider only singly–hybridized Watson–Crick pairs (thus, each row and column has at most one non–zero entry). The matrix representation, $\mathbf{\Theta}^i$, is more readily generalizable to other biomolecules.

B. Clustering

For clustering, we will employ k–means [39, 40] with a fixed k—a departure from normal practice—to a molecular ensemble. Specifically, we will use a variant of k–means that uses k–means++ [40] for initialization and employ a physical ensemble member for the centroid [38]. We do $\sharp\mathcal{S} = 500$ independent clusterings, taking the one that minimizes the objective function for the data set $\mathcal{D}$,

$$\mathcal{O} = \sum_{i\in\mathcal{D}} \min_{c\in\mathcal{C}} d_{ic}^2, \qquad (3)$$

via the set of centroids $\mathcal{C} = \{c_\kappa | \kappa = 1, 2, \ldots, k\}$.

We treat k as a fine–graining parameter, rather than a parameter that has to be heuristically optimized with an elbow or some other analysis. With $k = 1$, the cluster is just the whole ensemble, albeit having a single centroid as its representative. The $k = 1$ cluster characteristics will be the average ensemble properties, like $\langle n_b\rangle$ or average radius of gyration. These are informative and often directly connected to experimental measurements. Yet, higher k yield a more fine–grained view of the ensemble. The ability to successfully process and leverage larger k results determines their usefulness. This turns out to be an involved process, where a host of factors influence clustering quality. This includes usual issues with the number of trials $\sharp\mathcal{S}$ and dataset quality (molecular ensemble sampling, in our case). It also involves some issues specific to sweeps in variables, such as temperature, that revolve around consistency and the flow of information. We will focus on $k = 2$ here to address these issues.

To ensure robustness and consistency, we identify two factors that are important: (1) The distribution of equidistant structures and (2) centroid stability. In cases where a structure is equidistant from both centroids, we assign it to the cluster with lower occupancy to partition structures consistently across temperature. Otherwise, the cluster probability can be noisy rather than have smooth behavior as conditions change. For the second factor, we use the centroid pairs (for $k = 2$) across the whole temperature sweep as reference set. After an initial clustering across all temperatures, we check each temperature to see if the centroid pairs in this set lower the cost function, Eq. (3). The rationale is as follows: Clustering with k–means and some other methods is a random process that will typically fall short of finding the global optimum. Thus, there is some randomness built in to the determination of centroids. Having $\sharp\mathcal{S}$ independent clustering attempts to regularize this, but centroids

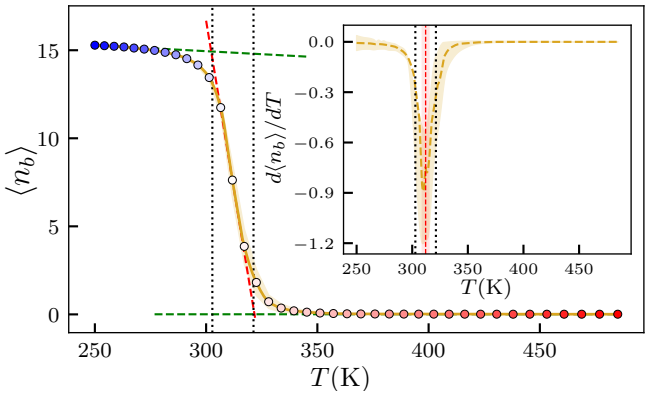


FIG. 2. **Melting.** Ensemble average base pairing, $\langle n_b\rangle$, versus temperature. The barely visible shaded region represents the block standard error (BSE) [41]. The solid curve is an interpolation with piecewise cubic Hermite polynomials. The inset shows the first derivative of the interpolation (red, dashed line). We use a bootstrap analysis to obtain the melting temperature, $T_m = (312 \pm 3)$ K (vertical, dashed red line with shaded red uncertainty), from the peak in the derivative of the interpolation. The transition width of $\Delta T_m = (19 \pm 2)$ K (black, dotted lines with uncertainty not shown) quantifies the sharpness of the melting. We define this width via the intersection of a tangent at T_m with linear fits to the low– and high–temperature baselines (i.e., the first eight data points for low temperature and $\langle n_b\rangle = 0$ for high temperature). The peak in $d\langle n_b\rangle/dT$ for the interpolation and the bootstrapped T_m have a small offset. The uncertainties in T_m and ΔT_m are plus and minus one standard deviation within the bootstrapping realizations.

can still change from one temperature to the next due to insufficient $\sharp\mathcal{S}$. This will create noise in both the centroids themselves, their probability, and other characteristics. Clustering at two nearby temperatures or other conditions, however, effectively doubles the number of independent clusterings $\sharp\mathcal{S}$, and even more when clustering many temperatures. Thus, there is more information contained in a parameter sweep than just at the isolated points. This is an expectation of continuity with the exception of phase transitions and sharp crossovers. Thus, the algorithm should leverage this information. Overall, this self–consistency improves cluster interpretability.

We will leverage that secondary–structure clustering removes, by construction, atomistic disorder from structural flexibility, as well as fluctuations and vibrations around bonds [38]. This is a feature that holds regardless of the clustering technique or molecular example under study. In the context of k–means, we observe two useful features of secondary–structure clustering. One is that the clusters have a connection to energetics. This feature is not expected to hold with all clustering techniques. For instance, density–based methods will likely break connections to pathways and energies. Two, the clustering provides an algorithmic, human–independent, way to capture the melting and other processes. We expect this to generally hold true.

C. Molecular Dynamics

We use clustering to study a pseudo–random DNA sequence of 50 bases that is a fragment of the M13 bacteriophage [42, 43]. Its sequence is 5′–GAATGATAAGGAAAGACAGCCGATTATTGATTGGTTTCTACATGCTCGTA–3′. To generate the ensemble, we use a coarse–grained DNA model, oxDNA [44, 45] implemented as CGDNA [46] in LAMMPS [47], and perform extensive replica exchange molecular dynamics (REMD) simulations across a range of temperatures at very fine temperature increments, chosen to ensure accurate sampling of melting transitions in nucleic acids. As in typical REMD, we employ an inhomogeneously (exponentially) distributed grid in order to get homogeneous exchange frequencies across neighboring temperatures,

$$T_r = T_{\min} \cdot e^{\lambda\cdot(r-1)} \quad , \quad r = 1, 2, \ldots, N_r, \tag{4}$$

where T_r is the temperature for replica r, $T_{\min} = 250$ K, $T_{\max} = 485$ K, and $N_r = 40$ is the total number of replicas. The constant $\lambda = [1/(N_r - 1)] \ln(T_{\max}/T_{\min})$, ensures the last replica temperature is exactly $T_{\max}$.

We note that, since this model has a continuum solvent, we examine its behavior below and above physically relevant temperatures. This enables obtaining a more complete understanding of behavior and clustering performance. The REMD simulation is run for 64×10^7 steps with an exchange attempt frequency (EAF) of 0.02 per step, i.e., one every 50 steps, which corresponds to one exchange attempt per ≈ 85.3 ps. We set the continuum dielectric to approximate 0.5 mol/L NaCl. We randomly sample 100 k structures for each r to form the clustering data sets, $\mathcal{D}_r$, giving a total of 4 M structures across 40 replicas. We employ the base pairing criteria of Ref. [38].

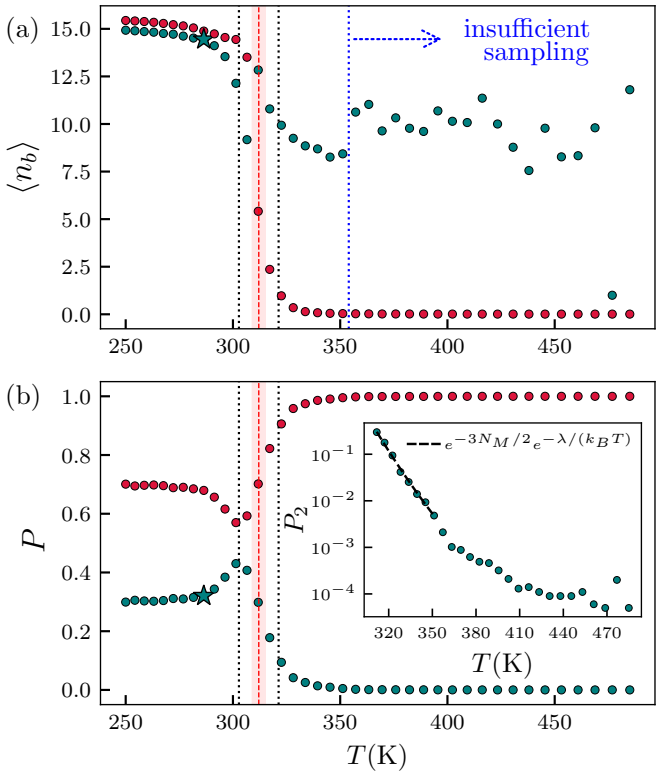


FIG. 3. **Secondary–structure clustering.** (a) Average number of base pairs, $\langle n_b\rangle_\kappa$, in cluster $\kappa = 1$, 2 (red, teal) versus temperature. (b) The cluster probabilities, P_κ, versus temperature. The vertical lines indicate the melting temperature (dashed red with the shaded red uncertainty), transition width (black dotted), and the region with insufficient sampling to cluster properly (blue dotted). The teal star marks the end of a stable clustering regime, see Fig. 4, where the lower cluster now has a frayed end to absorb ensemble diversity. A cluster with the same centroid reappears post–melting with about 8 base pairs on average. This cluster's probability follows Eq. (10) (inset). A fit gives $\lambda = (-1.00 \pm 0.02)$ eV as the free energy change from the unfolded state as a reference and $N_M = 25.5 \pm 0.6$ as the number of bases with restricted entropy in the stem region. The uncertainties for λ and N_m are plus and minus one standard error of the fit. This λ is reasonable value for structures with about six base pairs and a couple additional pairs that are mismatches, as is N_M given that 16 bases are locked in pairs on average and more are restricted in the stem.

III. RESULTS AND DISCUSSION

Figure 1 shows the folded and unfolded ensemble centroids. Figure 2 quantifies the transition by showing the average number of base pairs as a function of the replica temperature, yielding the full melting profile. As temper-

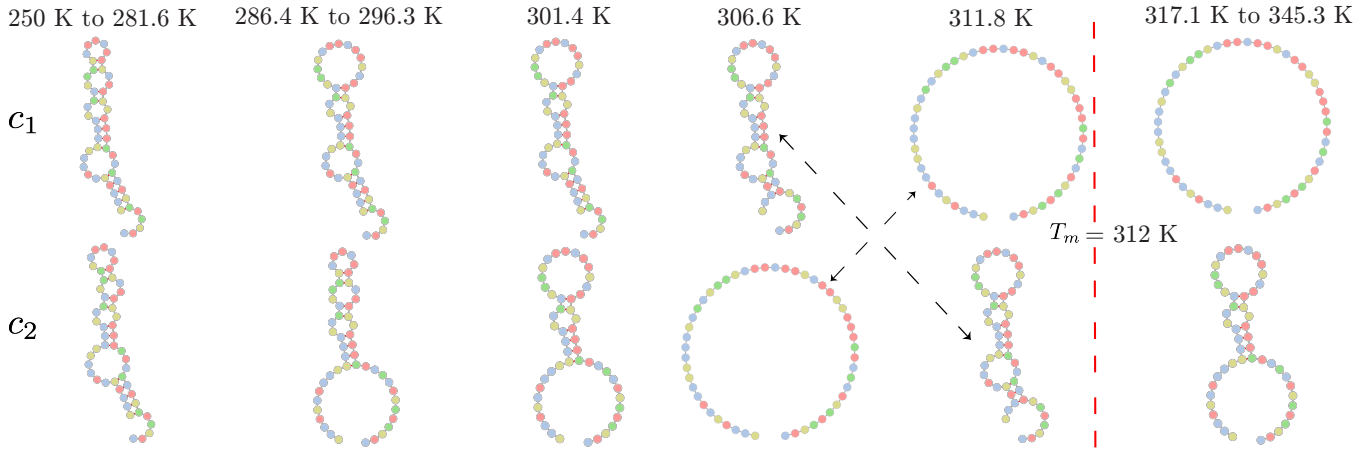


FIG. 4. **Secondary–structure centroids.** From left to right, the centroids versus increasing temperature, as labeled above each centroid pair c_κ with $\kappa = 1, 2$. The centroids remain constant across many ranges of temperatures, displaying a continuity in the flow of information during the melting process. From 250 K to 281.6 K, the centroids are both mostly folded. As temperature increases, though, the ensemble diversity increases. Eventually, the clustering changes (marked by a teal star in Fig. 3), with the lower probability cluster absorbing the diversity. This gives a stark change of its centroid to a frayed structure. While conforming to typical melting paradigms, what is happening is more intricate: k–means identifies the frayed structure because it has the most common motif to what is otherwise still a well–folded cluster, as seen by the average number of base pairs in Fig. 3a. This centroid evolves to the unfolded state before reappearing as a persistent stem feature at high temperature. The centroids for replicas above about 350 K are shown in the SI, as this region is insufficiently sampled to properly cluster.

ature increases, there is a moderately sharp decrease in base–pairing at the melting temperature, T_m. The minimum in the first derivative yields $T_m = (312 \pm 3)$ K as displayed in inset. This value and its uncertainty are from bootstrapping over 10 k noisy realizations of the melting curve. We determine the transition width, $\Delta T_m = (19 \pm 2)$ K, from the bootstrapping calculation via the intersection of the tangent at T_m with the low– and high–temperature baselines. All quoted uncertainties for T_m and ΔT_m correspond to one standard deviation from the bootstrap realizations of the melting curves.

We now employ k–means to the melting transition. Figure 3 shows the evolution of the average number of base pairs and the probability of each of the two $k = 2$ secondary–structure clusters as the pseudo–random DNA melts. Cluster 1 (red, more occupied) and cluster 2 (teal, less occupied) are two folds around a similar stem but with different hairpin and internal loop motifs, see the centroids in Fig. 4. They have a d_{ij}^2 of 5. At low temperature (250 K to 281.6 K), these centroids stay constant. While most of this regime is below physically relevant temperatures, the continuity of the centroids reflects a stable structural regime. Moreover, algorithmically, employing the reference centroid set (i.e., cross checking centroids from different temperatures to see if the cost function decreases) and consistently assigning equidistant structures were both necessary to identify this stable regime. Otherwise, there will be noise, e.g., in the cluster centroid and probability, making it more difficult to identify a meaningful structural partitioning of the ensemble, see the Supplemental Information (SI). While a larger $\sharp\mathcal{S}$ can alleviate the need for cross–checking, it doesn't make maximal use the information already extracted from the parameter sweep, nor does one know *a priori* how large $\sharp\mathcal{S}$ needs to be. Equidistant structures need to be consistently assigned, regardless.

At 286.4 K (the starred data point for the teal cluster in Fig. 3), the probabilities of the two clusters start to approach 1/2, and cluster 2 adopts a frayed–end centroid. At first glance, this seems to be fully consistent with the typical paradigm for melting, where the ends first fray. Yet, this cluster still has a large average number of base pairs, larger than its centroid would indicate. In this situation, k–means is identifying a common—a "base"—stem, around which the molecular ends assume many different folds. The nearest secondary structure to all those different folds is one with the frayed end.

This is essentially an argument of structural diversity: There are many folds but there is a common base stem. To make this more quantitative and to also help understand the impact on the clustering process, we compute an entropic measure of structural diversity,

$$\mathcal{S} = -\frac{1}{L}\sum_{i=1}^{L}\sum_{j=0}^{L} p_{ij}\log p_{ij}, \tag{5}$$

where p_{ij} is the probability that base i pairs with base j, the $j = 0$ term, $p_{i0} = 1 - \sum_{j=1}^{L} p_{ij}$, accounts for the unpaired state, and L is the length of DNA sequence. A related binary measure, $H_{\rm D} = -\frac{1}{L}\sum_{i=1}^{L}[p_i \log p_i + (1 - p_i)\log(1 - p_i)]$, where $p_i = \sum_{j=1}^{L} p_{ij}$, satisfies $H_{\rm D} < \mathcal{S}$ since $p_i, 1 - p_i$ majorizes $\{p_{ij}\}$.

Figure 5a shows the structural diversity versus temperature for the full ensemble and for each cluster, and

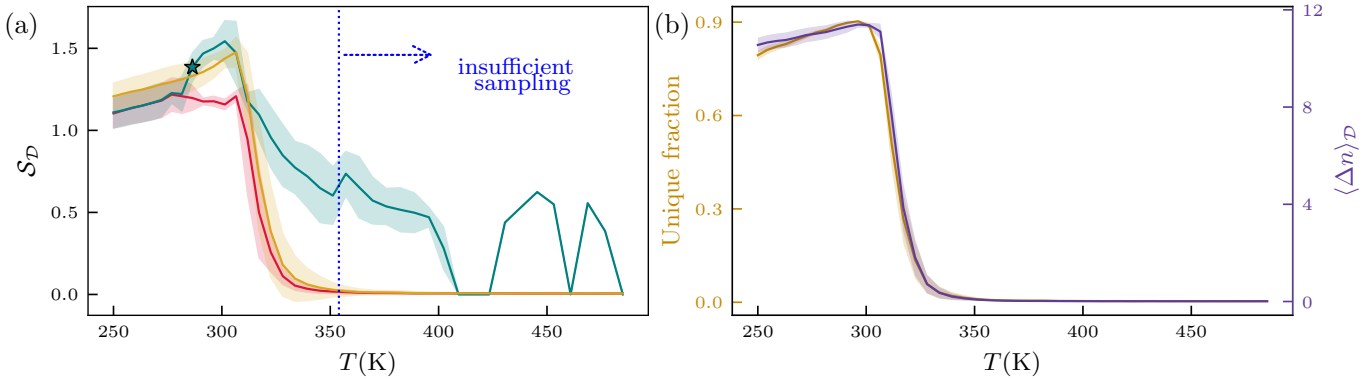


FIG. 5. **Structural diversity.** (a) The structural diversity, Eq. (5), versus temperature for the ensemble (orange) and individual clusters (red, teal). The shaded regions are the BSE. This measure of diversity is essentially the average base–to–base pairing entropy. Prior to melting, the diversity is steadily increasing. At 286.40 K—the starred data point in Fig. 3—the teal cluster starts to absorb this diversity, reflected by a discontinuity in its first derivative. This cluster assumes a centroid with a frayed end, yet most of its structures are more folded, just with different folds. Above melting, the teal cluster has a slowly decreasing diversity (prior to sampling artifacts after about 350 K). (b) Other metrics for diversity, such as the unique fraction of secondary structures and the spread in base pairing, confirm this picture of ensemble diversity increasing pre–melting and the lower cluster absorbing this diversity. The connecting lines are guides to the eye only.

Fig. 5b shows other measures. At low temperature, the ensemble diversity takes on a relatively high value due to the nature of the pseudo–random DNA, which does not have a particular target fold. This value steadily increases as the temperature increases towards melting. The diversity of the clusters initially tracks the ensemble value, albeit lower because the clustering process partitions the ensemble in to more well–defined—yet still quite diverse here—structural distributions. Near the starred data point in Fig. 3b, there is an abrupt uptick in the diversity of the teal cluster. This is a response to the steadily increasing ensemble diversity, requiring that one or both the clusters become more heterogeneous. For this molecule, the second cluster absorbs this extra diversity by taking on the "lowest common denominator" centroid, a frayed structure.

We designate the appearance of this frayed centroid as the start of a pre–melting regime. The phrase is not because of the frayed–end structure, per se, but rather the large diversity of folds—essentially, a hybridization entropy—that have to be captured by one of the centroids. However, the boundaries of this regime are not well defined and are in part a consequence of the clustering technique (and the use of $k = 2$) itself. As seen in Fig. 4, the folded and frayed centroid pair continues for about 10 K (three replicas). The lower probability centroid then melts further (the hairpin loop opens more), becoming fully unfolded just before the melting transition. The folded and unfolded centroids then swap places at the melting temperature (at the closest temperature replica, which is at 311.8 K).

Already at this stage, secondary–structure clustering and, specifically, with k–means at a fixed k, is useful: It is partitioning the ensemble into important structures and common motifs and is doing so without human intervention. Yet, this clustering has done something more. The identified common motif in the pre–melting regime is actually a persistent motif post–melting, even as most of the ensemble becomes unstructured. The lower probability centroid post–melting, from 317.1 K to 345.3 K, is the same as the pre–melting one at 301.4 K. At higher temperatures still (> 350 K), the second cluster starts to have very few members. Just after 350 K, it has about 400 members out of 100 k total for the ensemble at a fixed temperature. This goes down to 5 members at the last temperature. Thus, this region is insufficiently sampled to form two clusters, resulting in noisy clustering. Larger ensembles, and associated more costly REMD simulations, would enable tracking its evolution to higher temperatures. We expect that the same motif state will continue to have an exponential decrease in probability, although it is possible that clustering would identify, e.g., the motif with fewer base pairs as temperature increases. In principle, such a configuration could be an artifact of REMD, but its quantitative properties are in line with thermodynamic expectations, see below.

At the qualitative level, the above observations indicate that the secondary structure clustering with k–means is identifying important features of the ensemble. The common stem motif that persists to high temperature but that also captures heterogeneity at low temperature is particularly surprising. This residual fold is not visible in projections using principal component analysis (PCA) due to dominance of flexible–end motion, but is clearly revealed by secondary–structure clustering.

At the quantitative level, the probability of the lower probability cluster—the one that contains this motif—follows a Boltzmann factor, as shown in the inset to Fig. 3b. Above melting, the dominant state is the unfolded state—cluster 1— and the partition function will

be approximately just its contribution,

$$\mathcal{Z} = \alpha_1 e^{-E_1/k_BT} + \alpha_2 e^{-E_2/k_BT} \approx \alpha_1 e^{-E_1/k_BT}, \quad (6)$$

where k_B is Boltzmann's constant, α_i the number of configurations of state i, and E_i their energy. Thus, the probability for the motif state is

$$P_2 = \alpha_2 e^{-E_2/k_BT}/\mathcal{Z} \approx (\alpha_2/\alpha_1) e^{-\lambda/k_BT}, \quad (7)$$

with $\lambda = E_2 - E_1$. Moreover, we also can estimate the prefactor, which is given by the change in configurational entropy when going from the unfolded to motif state,

$$\alpha_2/\alpha_1 = e^{(S_2 - S_1)/k_B}. \quad (8)$$

For dangling, single–strand regions, the entropy is related to a freely–jointed chain model (FJCM) with N repeat units of length of l. For completeness, the average radius of gyration of the unfolded state is about 3.4 nm. Using $\langle R_g \rangle = \sqrt{N} l/\sqrt{6}$ yields $l \approx 1.2$ nm. In our approximation below, l will cancel out so we do not need its value. For the unfolded state, $N = 50$. The motif state has about 8 base pairs on average, although 28 bases are within the stem motif. From the FJCM, the change in entropy if N_M nucleotides have to be fully extended is

$$S_2 - S_1 = -\frac{3}{2} k_B \frac{R^2}{N_M l^2} = -\frac{3}{2} k_B N_M, \quad (9)$$

where we took $R = N_M \cdot l$ as the extension. The remaining nucleotides, $N - N_M$, are all still "FJCM" and thus they don't contribute to the entropy change. For the approximation here, we take fully extended to be equivalent to fully structured since the oligonucleotide is not actually extended but fixed into a stem motif.

Using Eq. (9), the probability, Eq. (7), becomes

$$P_2 \approx e^{-3N_M/2} e^{-\lambda/k_BT}. \quad (10)$$

The fit yields $\lambda = (-1.00 \pm 0.02)$ eV and $N_M = 25.5 \pm 0.6$, where the quoted uncertainties are plus and minus one standard error from the fit of P_2. This level of stabilization (λ) is consistent with having six Watson–Crick base pairs and a couple additional mismatches: Using standard nearest–neighbor thermodynamics for a six base pair stem with one bulge and AT–rich termini gives $\lambda \approx -1.15$ eV (≈ -26.6 kcal/mol) [48]. The two mismatched pairs that appear in many of the cluster structures will give a stabilizing correction to this, whereas the bulge will give some destabilizing correction, keeping the overall estimate roughly in line with this value.

The value of N_M is in reasonable agreement with the number of bases fixed in the stem region. A better approximation would examine the change in entropy to a rigid rod and also account for the loops in the stem region. In either case, one expects $N_M \lesssim 28$, since the hairpin loop retains significant entropy. That Eq. (10) captures the minority cluster supports that secondary–structure clustering with k–means is identifying a characteristic feature of the ensemble.

IV. CONCLUSION

We further developed the base–pair distance and secondary–structure clustering to compress and understand the conformational ensemble from atomistic (here, coarse–grained) simulations. By construction, this removes disorder due to atomistic motion (flexibility and vibrations) that can obscure underlying order. We considered the evolution of clusters versus temperature and with k fixed as a fine graining parameter. To examine such a variable sweep, we identified and solved two methodological problems: consistently assigning equidistant clusters and employing a concept of reference sets. These reflect the natural flow of information versus some physical parameter (here, temperature). Other issues were standard and include requiring reasonable $\sharp\mathcal{S}$ and sufficient REMD to yield large, high fidelity ensembles.

Secondary–structure clustering with k–means identifies sub–populations of structures and an important motif. Qualitatively, we find that it provides a simple, digestible view of melting. Quantitatively, the motif had clear thermodynamic signatures in terms of both its configurational entropy change and its energetics. A strength of the secondary–structure clustering is the identification of this particular motif (with six base pairs in the centroid and eight on average) as the leading order thermodynamic contribution—i.e., the "first excitation"—above the unfolded state. While it will be challenging due to the small probability, this feature may be identifiable experimentally in, e.g., single–molecule measurements with nanopores or other techniques.

We expect that for other molecules and clustering algorithms, secondary–structure clustering will be able to identify features of ensembles that will otherwise be obscured by disorder and be concealed within, e.g., PCA. Yet, there are two observations in this work that likely are particular to certain molecule types and clustering approaches. While k–means clustering is effective for compressing the ensemble, the centroids tend to have fewer base pairs than the cluster average, indicating that they under represent the structural content. The hybridization disorder of the pseudo–random DNA seems to be in part responsible for this. There are many folds that have to be captured by a limited set of centroids. Moreover, this same observation indicates that k–means is partitioning the ensemble rather than "clustering" it. Despite this, the clustering of the heterogeneous ensemble is supplying information about common structural motifs. Moreover, it acts, in a way, as a histogram of the ensemble, which will help to assess convergence and to quantify how well two state models of melting work. The observation of a quantitative thermodynamic connection is especially promising. We expect the approach will be helpful in an array of applications, such as DNA nanotechnology and disordered RNA drug targets.

V. ACKNOWLEDGEMENTS

We thank J. W. Robertson for helpful comments.

VI. SUPPLEMENTARY DATA

The supporting information file is provided.

VII. CONFLICT OF INTEREST

None declared.

VIII. DATA AVAILABILITY

The data underlying this article are available in the article and in its online supplementary material.

[1] L. R. Ganser, M. L. Kelly, D. Herschlag, and H. M. Al-Hashimi, Nat. Rev. Mol. Cell Biol. **20**, 474 (2019).
[2] C. M. Schmidt and C. D. Smolke, Cold Spring Harb. Perspect. Biol. **11** (2019).
[3] L. Cheng, E. N. White, N. L. Brandt, A. M. Yu, A. A. Chen, and J. B. Lucks, Nucleic Acids Res. **50**, 12001 (2022).
[4] K. D. Warner, C. E. Hajdin, and K. M. Weeks, Nat. Rev. Drug Discov. **17**, 547 (2018).
[5] J. P. Falese, A. Donlic, and A. E. Hargrove, Chem. Soc. Rev. **50**, 2224 (2021).
[6] J. L. Childs-Disney, X. Yang, Q. M. R. Gibaut, Y. Tong, R. T. Batey, and M. D. Disney, Nat. Rev. Drug Discov. **21**, 736 (2022).
[7] C. L. Haga and D. G. Phinney, Expert Opin. Drug Discov. **18**, 135 (2023).
[8] S. Kovachka, M. Panosetti, B. Grimaldi, S. Azoulay, A. Di Giorgio, and M. Duca, Nat. Rev. Chem. **8**, 120 (2024).
[9] M. Zuker and P. Stiegler, Nucleic Acids Res. **9**, 133 (1981).
[10] J. S. McCaskill, Biopolymers **29**, 1105 (1990).
[11] M. Zuker, Nucleic Acids Res. **31**, 3406 (2003).
[12] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner, Proc. Natl. Acad. Sci. U. S. A. **101**, 7287 (2004).
[13] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, Monatsh. Chem. **125**, 167 (1994).
[14] Z. J. Lu, J. W. Gloor, and D. H. Mathews, RNA **15**, 1805 (2009).
[15] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, Algorithms Mol. Biol. **6** (2011).
[16] J. M. Taliaferro, N. J. Lambert, P. H. Sudmant, D. Dominguez, J. J. Merkin, M. S. Alexis, C. Bazile, and C. B. Burge, Mol. Cell **64**, 294 (2016).
[17] A. Spasic, S. M. Assmann, P. C. Bevilacqua, and D. H. Mathews, Nucleic Acids Res. **46**, 314 (2018).
[18] S. Rouskin, M. Zubradt, S. Washietl, M. Kellis, and J. S. Weissman, Nature **505**, 701 (2014).
[19] M. Zubradt, P. Gupta, S. Persad, A. M. Lambowitz, J. S. Weissman, and S. Rouskin, Nature Methods **14**, 75 (2017).
[20] S. D. Kennedy, in *RNA Structure Determination: Methods and Protocols*, edited by D. H. Turner and D. H. Mathews (Springer, New York, NY, 2016) pp. 253–264.
[21] M. Marušič, M. Toplishek, and J. Plavec, Curr. Opin. Struct. Biol. **79**, 102532 (2023).
[22] P. W. K. Rothemund, Nature **440** (2006).
[23] J. M. Majikes, P. N. Patrone, D. Schiffels, M. Zwolak, A. J. Kearsley, S. P. Forry, and J. A. Liddle, Nucleic Acids Res. **48**, 5268 (2020).
[24] J. M. Majikes, P. N. Patrone, A. J. Kearsley, M. Zwolak, and J. A. Liddle, ACS Nano **15**, 3284 (2021).
[25] J. M. Majikes, M. Zwolak, and J. A. Liddle, Biophys. J. **121**, 1986 (2022).
[26] J. Majikes, A. Hasni, S. Haridas, J. Robertson, A. Pintar, M. Zwolak, and J. A. Liddle, bioRxiv (2025).
[27] Y. Ding and C. E. Lawrence, Nucleic Acids Res. **31**, 7280 (2003).
[28] Y. Ding, C. Y. Chan, and C. E. Lawrence, RNA **11**, 1157 (2005).
[29] C. Chan, C. Lawrence, and Y. Ding, Bioinformatics **21**, 3926 (2005).
[30] Y. Ding, C. Y. Chan, and C. E. Lawrence, J. Mol. Biol. **359**, 554 (2006).
[31] K. Sato, M. Akiyama, and Y. Sakakibara, Nat. Commun. **12**, 941 (2021).
[32] X. Chen, Y. Li, R. Umarov, X. Gao, and L. Song, arXiv (2020).
[33] T. Shen, Z. Hu, Z. Peng, J. Chen, P. Xiong, L. Hong, L. Zheng, Y. Wang, I. King, S. Wang, S. Sun, and Y. Li, arXiv (2022).
[34] S. W. Schaffter and E. A. Strychalski, Sci. Adv. **8** (2022).
[35] K. M. Weeks, Curr. Opin. Struct. Biol. Nucleic acids / Sequences and topology, **20**, 295 (2010).
[36] J. Šponer, G. Bussi, M. Krepl, P. Banáš, S. Bottaro, R. A. Cunha, A. Gil-Ley, G. Pinamonti, S. Poblete, P. Jurečka, N. G. Walter, and M. Otyepka, Chem. Rev. **118**, 4177 (2018).
[37] O. Languin-Cattoën and G. Bussi, arXiv (2025).
[38] S. Baral and M. Zwolak, to appear, Biophys. J. (2025).
[39] S. Lloyd, IEEE Trans. Inf. Theory **28**, 129 (1982).
[40] D. Arthur and S. Vassilvitskii, in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (SIAP, 2007) pp. 1027–1035.
[41] A. Grossfield and D. Zuckerman, in *Annual Reports in Computational Chemistry*, Vol. 5 (2009) pp. 23–48.
[42] J. Messing, in *Methods in Enzymology*, Recombinant DNA Part C, Vol. 101 (Academic Press, 1983) pp. 20–78.
[43] P. M. van Wezenbeek, T. J. Hulsebos, and J. G. Schoenmakers, Gene **11**, 129 (1980).
[44] B. E. K. Snodin, J. S. Schreck, F. Romano, A. A. Louis, and J. P. K. Doye, Nucleic Acids Res. **47**, 1585 (2019).
[45] B. E. K. Snodin, F. Randisi, M. Mosayebi, P. Šulc, J. S. Schreck, F. Romano, T. E. Ouldridge, R. Tsukanov, E. Nir, A. A. Louis, and J. P. K. Doye, J. Chem. Phys.

142, 234901 (2015).
[46] O. Henrich, Y. A. Gutiérrez Fosado, T. Curk, and T. E. Ouldridge, Eur. Phys. J. E **41**, 57 (2018).
[47] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton, Comput. Phys. Commun. **271**, 108171 (2022).
[48] J. John SantaLucia and D. Hicks, Annu. Rev. Biophys. Biomol. Struct. **33**, 415 (2004).

Supplemental Information: Secondary–structure clustering of nucleic acid melting: Pseudo–random DNA

Swapnil Baral[1,2] and Michael Zwolak[1]

[1]*Biophysical and Biomedical Measurement Group, Microsystems and Nanotechnology Division, Physical Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA*

[2]*Department of Chemistry and Biochemistry, University of Maryland, College Park, MD, USA*

I. CLUSTERING ARTIFACTS

Figure S1 compares three alternative strategies within k–means clustering. In Fig. S1a, the basic procedure takes a typical k–means assignment approach, randomly assigning cluster numbers to structures that are equidistant to multiple centroids. This introduces a temperature–to–temperature variability in cluster assignment, resulting in fluctuations in cluster probabilities and other characteristics. In Fig. S1b, we assign equidistant structures deterministically, putting them into the lower probability cluster. This stabilizes assignments and removes some of the variance in the evolution of the cluster characteristics. Assigning equidistant structures to the highest probability cluster instead does not change our conclusions, but does modify the probability curves. We note that other consistent assignment strategies are possible, including randomly assigning equidistant structures at the temperature at which they first appear and then maintaining consistency thereafter. Figure S1c shows the results with a cross–temperature global check: centroid sets produced at all temperatures are tested for lowering the cost function at all other temperatures. This additional step selects the centroids with the lowest cost, mitigating local–optimum artifacts that arise when there are not enough independent trials of the clustering.

II. CLUSTER AND ENSEMBLE CENTROIDS

Below, we show the ensemble centroid and cluster centroids corresponding to all the temperature replicas in Fig. 3 of the main text. For temperatures with a common cluster

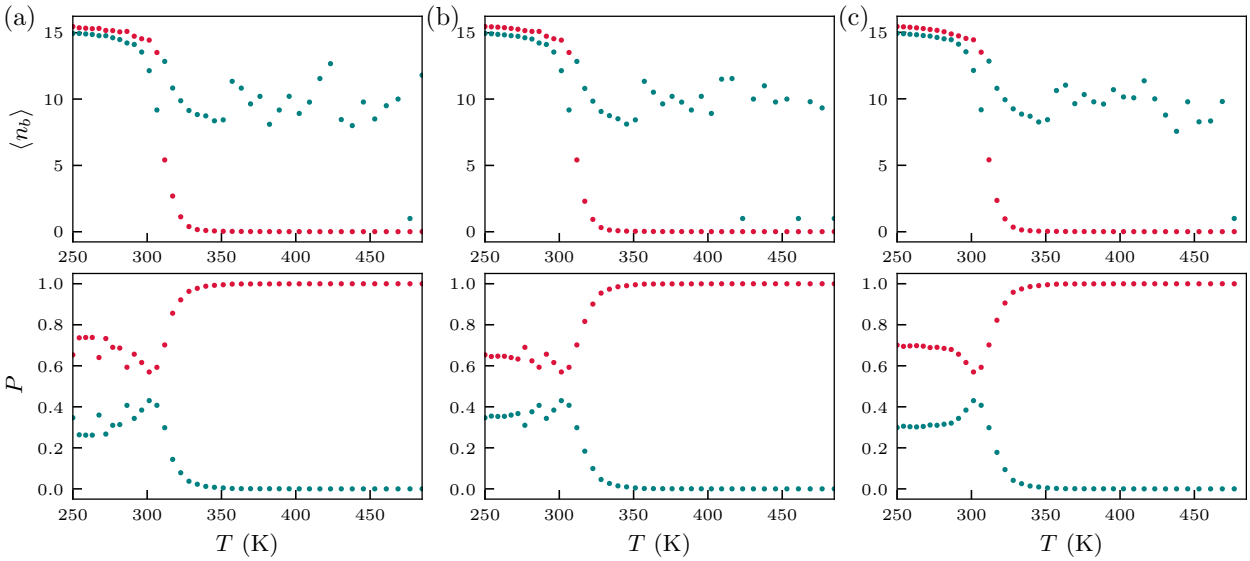
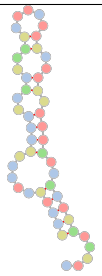
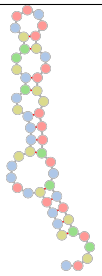
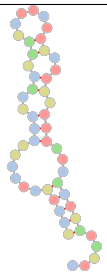
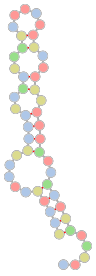
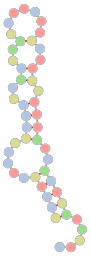
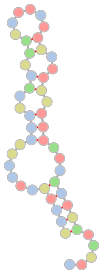
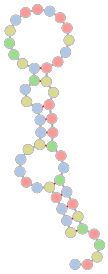
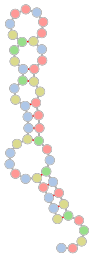
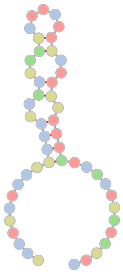
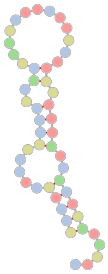
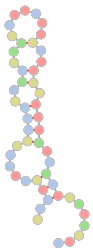
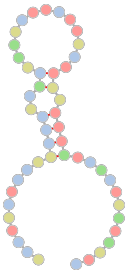
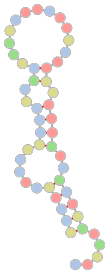
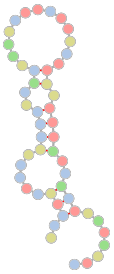
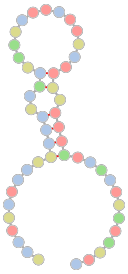
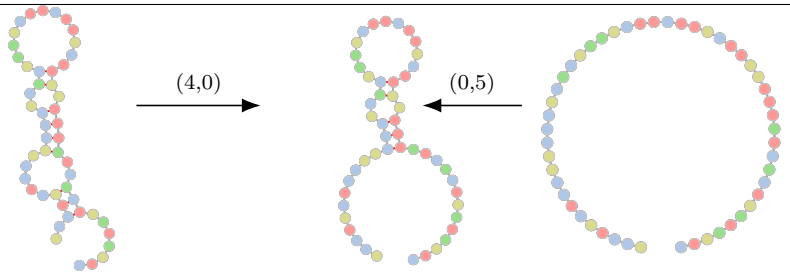
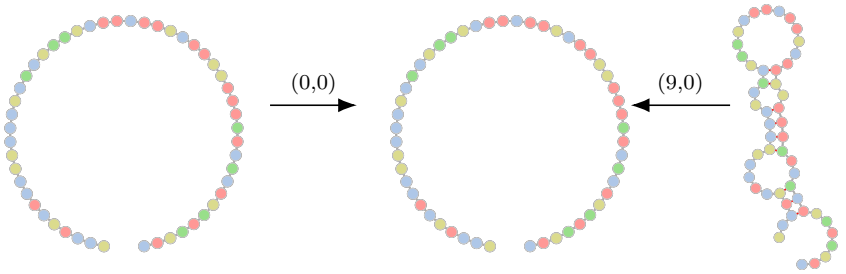
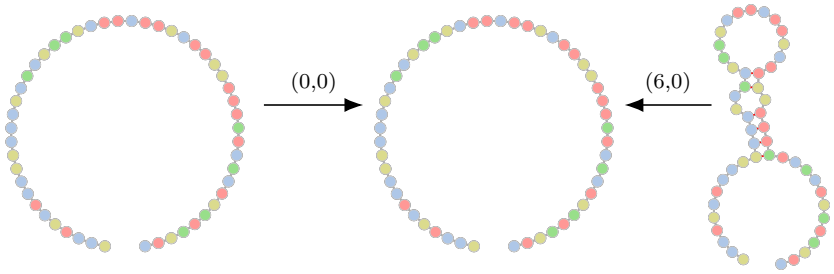
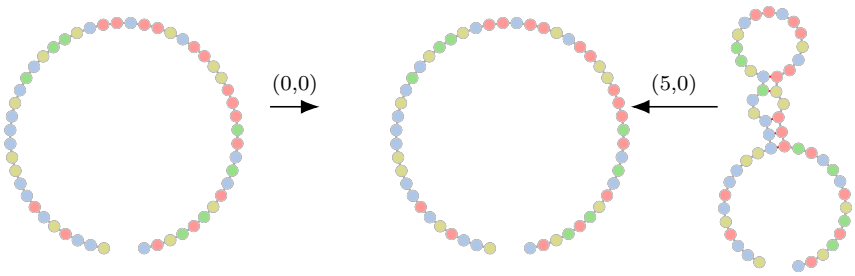
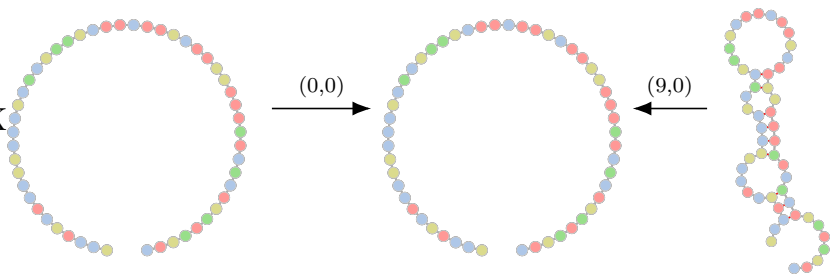
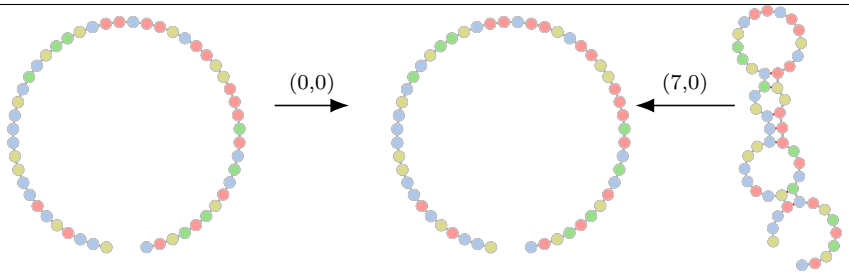
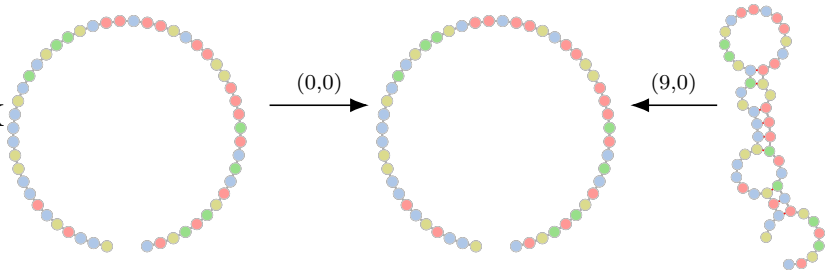
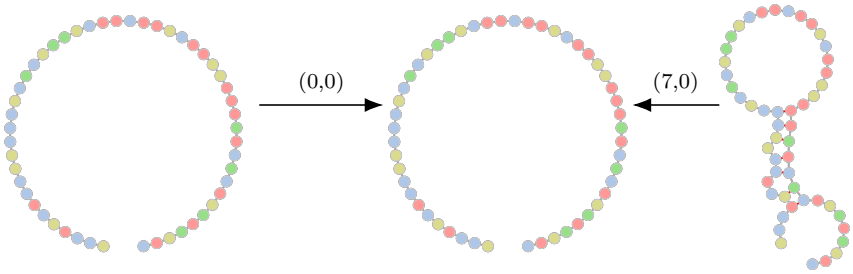
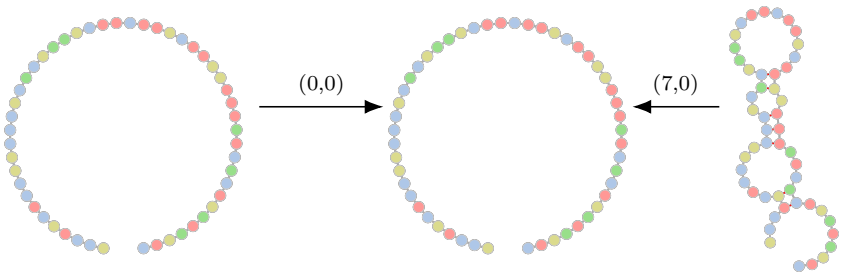
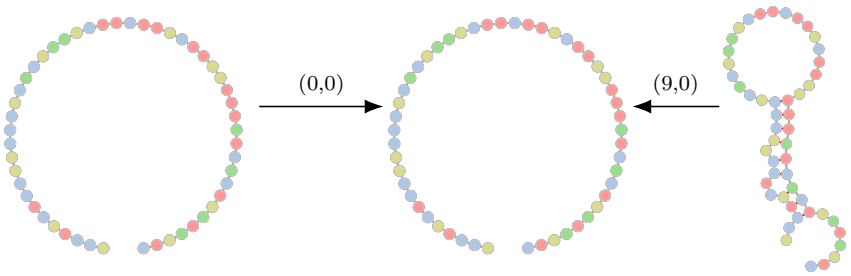


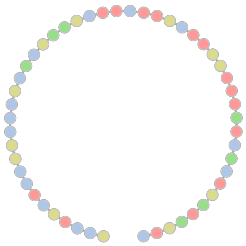
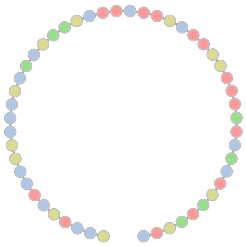
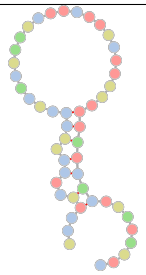
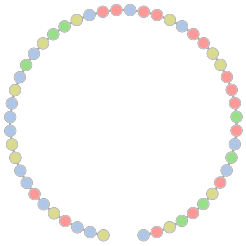
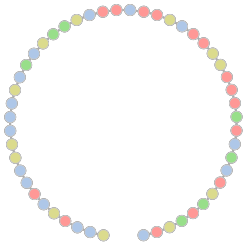
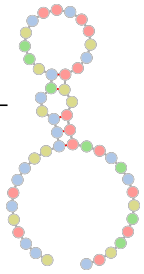
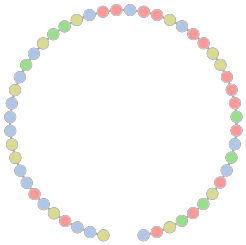
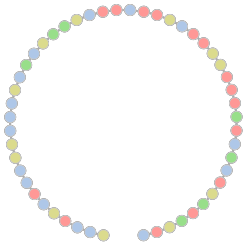
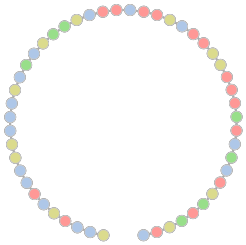
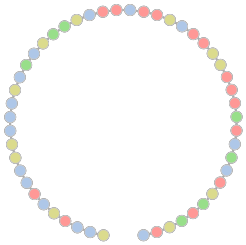
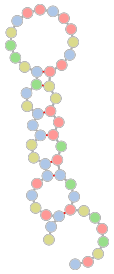
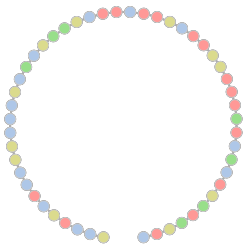
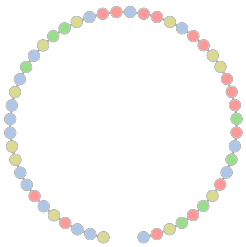
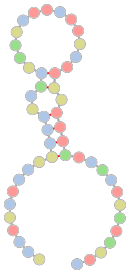
FIG. S1. ***k*–means strategies.** (a) Typical k–means: structures equidistant to multiple centroids are assigned at random. This results in large fluctuations of the probability, as well as smaller fluctuations in the average number of base pairs. (b) Equidistant structures are consistently allocated to the lower probability cluster. This regularization some of the fluctuations. (c) The same as in (b) plus a cross–temperature global check: centroid sets from all temperatures are used to compute cost at all other temperatures, and the set that yields the lowest cost is selected. This yields consistent, smooth characteristics, except above about 350 K here, where sampling would need to be exponentially larger to capture the minority cluster.

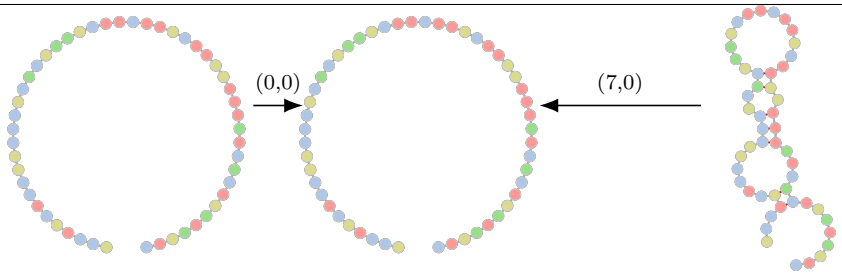
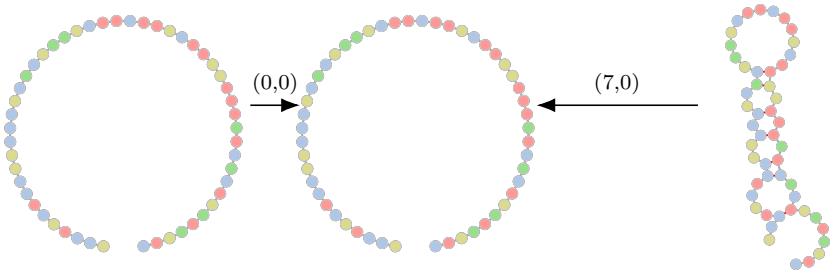
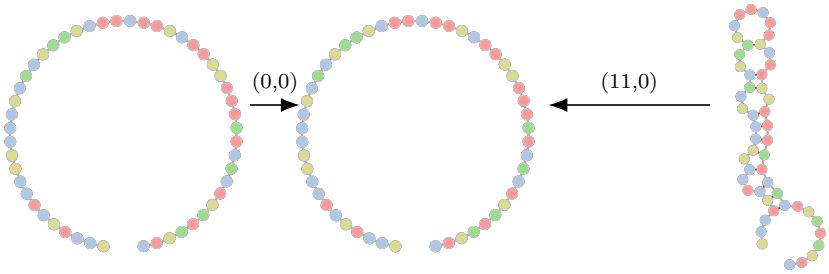
and ensemble centroids, we show that set just one time along with the relevant temperature range on the left. The centroids highlight the main structural motifs and their temperature dependence across the melting transition. We also show above the arrows the number of base pairs that must be broken (formed) to transform the cluster centroids into the ensemble.

Temperature	Cluster 1		Ensemble		Cluster 2
250 K to 267.5 K		(0,0) →		← (2,2)	
272.2 K to 281.6 K		(1,0) →		← (1,1)	
286.4 K to 291.0 K		(0,1) →		← (1,5)	
296.3 K		(2,1) →		← (0,4)	
301.4 K		(2,0) →		← (3,0)	

Temperature	Cluster 1	Ensemble	Cluster 2
306.8 K			
311.8 K			
317.1 K to 345.3 K			
351.2 K			
357.2 K to 363.3 K			

Temperature	Cluster 1	Ensemble	Cluster 2
369.6 K			
375.9 K to 402.3 K			
409.3 K			
416.2 K			
423.4 K			

Temperature	Cluster 1		Ensemble		Cluster 2
430.6 K		(0,0)		(7,0)	
438 K		(0,0)		(5,0)	
445.5 K		(0,0)		(9,0)	
453.1 K		(0,0)		(7,0)	
460.9 K		(0,0)		(6,0)	

Temperature	Cluster 1	Ensemble	Cluster 2
468.8 K		(0,0)	(7,0)
476.8 K		(0,0)	(7,0)
485 K		(0,0)	(11,0)