Employing artificial neural networks to find reaction coordinates and pathways for self-assembly

Jörn H. Appeldorn, Simon Lemcke, Thomas Speck,* and Arash Nikoubashman*

Institute of Physics, Johannes Gutenberg-University Mainz, Staudingerweg 7-9, 55128 Mainz, Germany

E-mail: thomas.speck@uni-mainz.de; anikouba@uni-mainz.de

Abstract

Capturing the autonomous self-assembly of molecular building blocks in computer simulations is a persistent challenge, requiring to model complex interactions and to access long time scales. Advanced sampling methods allow to bridge these time scales but typically require to construct accurate low-dimensional representations of the transition pathways. In this work, we demonstrate for the self-assembly of two single-stranded DNA fragments into a ring-like structure how autoencoder architectures based on unsupervised neural networks can be employed to reliably expose transition pathways and to provide a suitable low-dimensional representation. The assembly occurs as a two-step process through two distinct half-bound states, which are correctly identified by the neural net. We exploit this latent space representation to construct a Markov state model for predicting the four molecular conformations and transition rates. Our work opens up new avenues for the computational modeling of multi-step and hierarchical self-assembly, which has proven challenging so far.

Keywords

DNA, self-assembly, molecular dynamics, Markov state model, machine learning, neural networks

As an alternative to the top-down manipulation of a material's nano(colloidal) constituents, directing its autonomous organization through the judicious design of building blocks has proven to be a viable strategy.[1–4] Considering the myriad number of ways these (macro)molecular building blocks can be combined, sufficiently specific yet reversible interactions are required to achieve the desired target structures.[5] This challenge has been mastered by nature, where, *e.g.*, amphiphilic phospholipids self-assemble into bilayer cell membranes, and the amino acid sequence of a protein dictates its folding and subsequent biological function. Following this design principle, enormous progress has been made in the past decades to engineer and leverage the self-folding of elongated (macro)molecules (typically DNA) into predetermined complex shapes, stabilized by native contacts programmed into their base-pair sequences ("DNA origami").[4,6,7] The specificity offered by designing base-pair sequences has also been exploited for the self-assembly of "patchy" colloidal particles coated with complementary DNA strands.[8–15] Such patchy colloids can also be viewed as a highly coarse-grained representation of folded proteins;[16] although this simplified description might seem too reductionist at first glance, previous simulations have demonstrated excellent agreement between the phase diagrams of such patchy particles and experimental results for γ-crystallin and lysozyme.[17]

Accurate yet efficient numerical methods are indispensable to, *inter alia*, elucidate microscopic pathways and scan (often vast) parameter spaces to guide experiments. However, the inherent multiscale aspect of self-assembly (from the molecular building blocks to extended mesoscale structures) poses a formidable challenge. In particular, the large time-scale separation between the atomistic motion of the building blocks and the time on which the mesoscopic target structure is assembled needs to be bridged. An arsenal of advanced sampling techniques has been devised for this purpose, including forward-flux sampling,[18–20]

umbrella sampling,[21] transition path sampling,[18,22,23] and the Wang-Landau algorithm.[24] However, these methods typically rely on the *a priori* knowledge of reaction coordinates or at least a suitable low-dimensional representation of the system (called order parameters or collective variables).[25] Dimensionality reduction through constructing physically and chemically informed collective variables requires detailed insights and extensive verification. Moreover, multiple barriers and competing pathways pose severe challenges for the majority of existing methods.

While data-driven modeling and machine-learning techniques play an increasing role also in computational soft matter,[26–30] their systematic use in self-assembly is still largely unexplored. A promising route to construct a suitable low-dimensional space in an unsupervised manner are autoencoder architectures.[31,32] These neural networks consist of an encoder section, which compresses the input data to a so-called *latent space*, and a decoder section, which reconstructs the reduced data to its original size. Both parts are trained simultaneously to minimize the deviations between input and reconstructed data. To make this approach viable, as an intermediate step raw particle configurations are mapped onto *structural descriptors* that ensure simple symmetries (*e.g*, invariance with respect to translation, rotation, and permutation).[29] These structural descriptors are the input data to the autoencoder so that each point in the latent space parameterizes a manifold of structurally indistinguishable configurations.

Here, we study the hybridization of two single-stranded DNA fragments, which are designed such that they can self-assemble into a ring as shown in Fig. 1. This system exhibits two competing pathways from an unbound to a fully bound state that are insufficiently resolved by standard collective variables. Identifying long-lived molecular conformations and transition pathways requires a suitable distance metric for the latent space representation. One successful candidate for this task is sketch-map,[33] which orders points in a low-dimensional space so that their relative distance reflects the relative distance of points in the input space of structural descriptors. EncoderMap, the combination of an autoen-

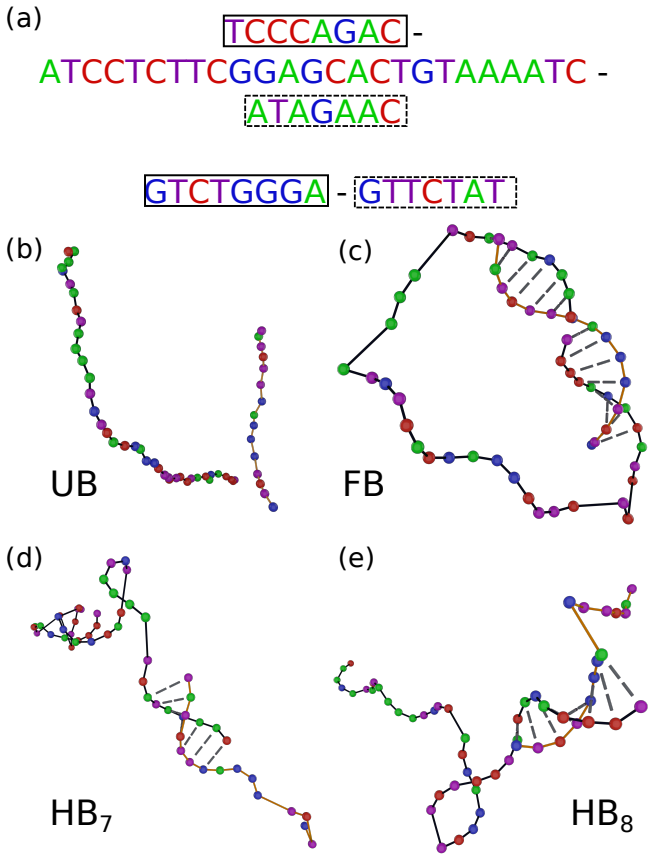

Figure 1: (a) Nucleobase sequence of the long (top) and short (bottom) ssDNA molecule. The major and minor sides are indicated by solid and dashed boxes, respectively. (b-e) Representative snapshots of the (b) unbound (UB), (c) fully bound (FB), (d) 7-half-bound (HB_7), and (e) 8-half-bound states (HB_8).

coder with the objective function of sketch-map has recently been applied to identify distinct (albeit rarely occurring) clusters of conformations in folding simulations of the Trp-cage protein.[34] In this work, we demonstrate that this approach can be leveraged to numerically predict the self-assembly of multiple (bio)macromolecules into a desired superstructure.

Results and Discussion

DNA self-assembly

We study one single-stranded DNA (ssDNA) molecule with 40 nucleobases and one molecule with 15 nucleobases through molecular dynamics (MD) simulations of a coarse-grained rep-

resentation,[35] see *Methods*. By visually inspecting the simulations, we can already classify the system into four well-defined macro-states: All simulations start in the *unbound state* (UB), where the two molecules do not form hydrogen bonds (H-bonds) and thus can move independently. There are two different hybridized half-bound states due to the asymmetry of the nucleobase sequences (Fig. 1): The *8-half-bound state* (HB_8), where the major sides of the short and the long molecule hybridize, and the *7-half-bound state* (HB_7), where their minor sides hybridize. The fourth state is the *fully bound state* (FB), where both the major and minor sides of the two ssDNA molecules hybridize into a ring-like structure.

In a first attempt to quantify the hybridization states, we introduce two order parameters: The end-to-end distance d_{cl} of the long ssDNA molecule (a small value indicates that it has bent into a circle), and the average base pair distance

$$\langle d_{\mathrm{bp}} \rangle = \frac{1}{N_{\mathrm{bp}}} \sum_{i=1}^{N_{\mathrm{bp}}} d_{\mathrm{bp},i}, \quad d_{\mathrm{bp},i} = |\mathbf{r}_i - \mathbf{R}_i| \tag{1}$$

between the $N_{\mathrm{bp}} = 15$ binding pairs on the two DNA strands. Here, $\mathbf{R}_i$ and $\mathbf{r}_i$ are the positions of the i-th nucleobase on the long and small molecule, respectively (counting only bases in the binding regions of the DNA strands). To determine whether two nucleobases have bound, we first compute the distribution of base pair distances in the FB state. We find a rather sharp maximum at $\langle d_{\mathrm{bp}} \rangle_{\mathrm{FB}} \approx 10.4\,\text{Å}$ with only negligible variations for the different base pairs. We then go through all configurations and count the number of base pairs ($n_{\mathrm{bp}} \leq N_{\mathrm{bp}}$) with $d_{\mathrm{bp},i} \leq 10.8\,\text{Å}$. Figure 2 shows the number of bound base pairs, n_{bp}, in the plane of the two variables d_{cl} and $\langle d_{\mathrm{bp}} \rangle$. These two order parameters work reasonably well to separate the UB state from the other configurations, but the remaining three states (HB_7, HB_8, and FB) are poorly separated from each other.

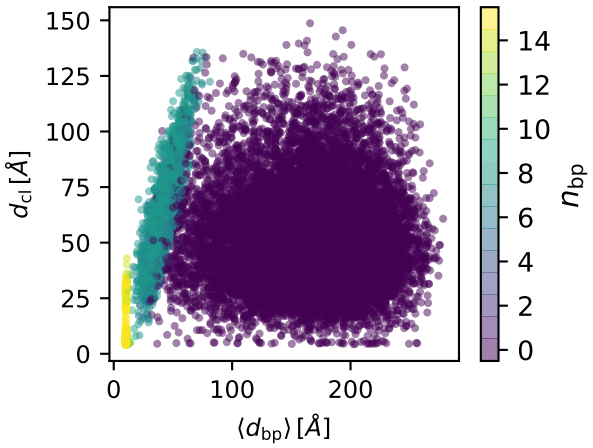

Figure 2: End-to-end distance d_{cl} of the long ssDNA molecule *vs.* average base pair distance $\langle d_{\mathrm{bp}} \rangle$ [Eq. (1)]. Each dot represents a different configuration with the color indicating the number of bound base pairs n_{bp} (see scale on the right).

Latent space as order parameter

To optimize the classification of the different states, we employ the EncoderMap[34] dimensionality reduction algorithm, which is based on a neural network autoencoder in combination with a non-linear distance metric. This machine-learning approach reduces the input data to a low-dimensional latent space, from which it is expanded again to its original dimensionality while minimizing the loss between the input and output data. This bottleneck structure is a key attribute of the network design, which constrains the amount of information that can traverse the full network, forcing a learned compression of the input data.[36,37] As input data, we use the $N_{\mathrm{bp}} = 15$ nucleobase pair distances $d_{\mathrm{bp},i}$ serving as a high-dimensional structural descriptor.

The first step is to transform the N_{bp} distances through

$$\zeta_i = \pi \left[1 + \exp\left(-x \frac{d_{\mathrm{bp},i}}{\langle d_{\mathrm{bp}} \rangle_{\mathrm{FB}}} + y \right) \right]^{-1} \tag{2}$$

with $\langle d_{\mathrm{bp}} \rangle_{\mathrm{FB}} \simeq 10.4\,\text{Å}$ being the average base-pair distance in the FB state. We set $x = 3$ and $y = 7$ so that nucleobase pair distances larger than 3 times $\langle d_{\mathrm{bp}} \rangle_{\mathrm{FB}}$ will result in a value close to π. We thus map the configurations of the UB state with large (and for the

transitions irrelevant) distances to similar values $\zeta_i \simeq \pi$, while at the same time amplifying differences for smaller distances through the non-linear function.

In contrast to linear techniques such as principal component analysis, EncoderMap can capture non-linear features in the input data through the sketch-map cost function[33]

$$C_{\mathrm{sm}} \propto \sum_{j<k} |f(R_{jk}) - f(r_{jk})|^2 \tag{3}$$

with Euclidean distances r_{jk} in latent space and distances

$$R_{jk} = \sqrt{\sum_{i=1}^{N_{\mathrm{bp}}} \left|\zeta_i^{(j)} - \zeta_i^{(k)}\right|^2} \tag{4}$$

in the input parameter space, where $\zeta_i^{(j)}$ is the i-th input parameter of the j-th configuration. The longest possible distance between two configurations is $\sqrt{N_{\mathrm{bp}}\pi^2} \simeq 12.17$, which is reached between the FB and UB state. The histogram of the calculated distances R_{jk} is

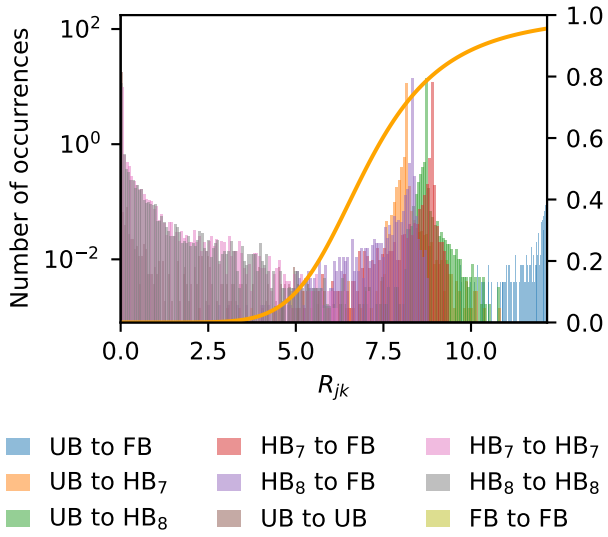

Figure 3: Histogram of the Euclidean distances R_{jk} [Eq. (4)] between different configurations (left axis). The different colors indicate the pair of states to which the configurations correspond. The orange line is the sigmoid function $f(R)$ for the weight of the input parameters (right axis).

shown in Fig. 3 for all relevant configuration pairs.

The basic premise of sketch-map is that not all pairwise distances are equally important. The statistics of both very small and very large distances contains little information, and sketch-map selects intermediate distances through the sigmoid function $f(x) = 1 - [1 + (2^{a/b} - 1)(x/\sigma)^a]^{-b/a}$, where σ defines the inflection point, a determines the left slope and b the right slope. Specifically, for R_{jk} the parameters are set to $\sigma = 7$, $a = 7$, and $b = 5$ so that the inflection point of the sigmoid function lies close to the transitions between the HB states and the UB and FB states, hence prioritizing these regions when learning the latent space representation [Fig. 3]. Distances in latent space are mapped using $\sigma = 1$, $a = 2$, and $b = 5$.

The hidden layer of the neural network consists of two fully connected layers with 128 neurons each, a bottleneck layer with 2 neurons, followed again by two layers with 128 neurons each, which are connected to an output layer with $N_{\mathrm{bp}} = 15$ neurons. The network is trained with data from 32 different MD runs each with 180,000 time steps, where the two DNA strands either hybridize *via* the HB_7 state, *via* the HB_8 state, or remain unbound. The two-dimensional latent space is spanned by two real numbers Φ_1 and Φ_2, where each point (Φ_1, Φ_2) represents a manifold of indistinguishable molecular configurations. Note that

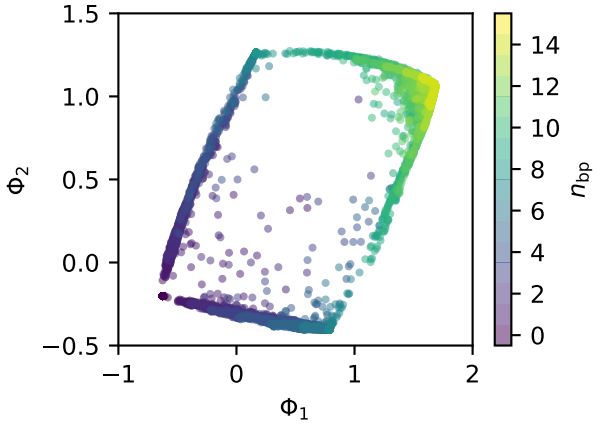

Figure 4: Latent space representation of the configurations showing a subset of the training data. Each dot represents one configuration and the color visualizes the number of bound nucleobases n_{bp}.

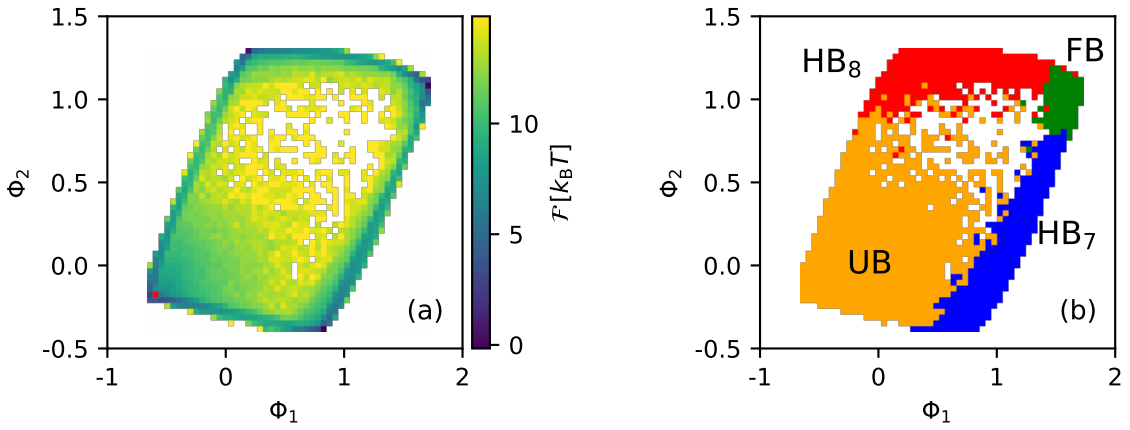

Figure 5: (a) Latent space free energy $\mathcal{F}(\Phi_1, \Phi_2)$. The magnitude of $\mathcal{F}$ is indicated by the color bar, with dark blue and yellow representing the lowest and highest free energy, respectively. The red bin contains all configurations with $\langle d_{\mathrm{bp}} \rangle > 3\langle d_{\mathrm{bp}} \rangle_{\mathrm{FB}}$ where the two fragments are fully unbound. The free energy of this bin strongly depends on the molecular concentration. (b) Metastable basins in latent space. Each color indicates one of the four basins identified by the PCCA+ clustering and is labeled according to Fig. 1.

the actual numbers do not carry a physical meaning but the latent space preserves relative distances, *i.e.*, pairs of configurations that are close in input space are also close in latent space.

The latent space representation of the training data is shown in Fig. 4. We now observe a clear separation of the four hybridization states, roughly corresponding to the four corners of a rhombus: Points in the bottom left corner of this graph ($\Phi_1 \simeq -0.5$, $\Phi_2 \simeq -0.25$) belong to UB states, while points in the top right corner ($\Phi_1 \simeq 1.5$, $\Phi_2 \simeq 1.0$) represent FB states. These two regions are connected by two pathways corresponding to the binding processes *via* the HB$_8$ and HB$_7$ states, respectively: The left branch belongs to transitions where first the two major sides and then the minor sides hybridize, while the right branch undergoes the transitions with the opposite binding order. Hence, the two order parameters Φ_1 and Φ_2 uncovered by the autoencoder architecture clearly separate the different molecular conformations and pathways in contrast to our previous *ad hoc* representation in terms of the end-to-end distance d_{cl} and average base pair distance $\langle d_{\mathrm{bp}} \rangle$, *cf.* Fig. 2.

Markov state modeling

Employing the latent space of the autoencoder as a suitable low-dimensional representation of the assembly process, we now construct a Markov state model (MSM) to identify metastable states and estimate the kinetic rates.[38,39] To construct the MSM, we first compute the two-dimensional latent space representation for all 256 trajectories and discretize the resulting order parameter space into 49×49 bins (note that some of the bins are empty since we use a regular grid for the discretization). Including all non-empty bins, we then compute the transition probability matrix for this discretized representation using PyEMMA.[40] The lag time is set to $10000\Delta t = 151.5\,\mathrm{ps}$.

From the stationary probability $\mathcal{P}(\Phi_1, \Phi_2)$, we calculate the free energy $\mathcal{F} = -k_\mathrm{B}T \ln \mathcal{P}$ for each bin. The resulting two-dimensional free energy landscape is shown in Fig. 5(a), where we have shifted the energy scale so that $\mathcal{F}_*(\mathrm{FB}) = 0$ in the FB state. The fully unbound state is indicated by a red bin, which represents all configurations with an average nucleobase pair distance $\langle d_\mathrm{bp} \rangle > 3\langle d_\mathrm{bp} \rangle_\mathrm{FB}$. The minimal free energy of the UB state $\mathcal{F}_*(\mathrm{UB}) \simeq 0.17\,k_\mathrm{B}T$ is very close to that of the FB state. To rationalize this finding, let us consider the difference in free energy going from the UB to the FB state (at the same temperature T), given by $\Delta\mathcal{F} = \Delta U - T\Delta S$ in the canonical ensemble. In the FB state, the system gains $\Delta U \approx 16\,k_\mathrm{B}T$ in potential energy due to the hybridization of complementary nucleobases ($\simeq 1.077\,k_\mathrm{B}T$ per hybridized pair[41]). In the UB state, the two ssDNA fragments can move independently from each other, whereas they move as a unit in the FB state. Hence, the difference in translational entropy between the UB and FB configuration is $\Delta S_\mathrm{t} = k_\mathrm{B} \ln(1/\phi) \approx 7\,k_\mathrm{B}$, with molecular volume fraction $\phi \approx 10^{-3}$ in our simulations. Further, the conformational entropy of the system is smaller in the FB state since large portions of the two strands are locked into position [see Fig. 1(e)]. Of course, this calculation is just a rough estimation, which neglects, *e.g.*, the change in bending energy as the two ssDNA fragments form a ring. Nevertheless, it serves as a useful check for the consistency of our MSM.

Out of the UB state, the free energy increases as the base pairs start to approach due to

the reduction of translational and configurational intramolecular entropy. However, as more base pairs hybridize, the gain in potential energy starts to outweigh the loss in entropy so that the free energy again decreases after crossing the barrier, reaching the minima $\mathcal{F}_*(\mathrm{HB}_7) \simeq -0.16\,k_\mathrm{B}T$ and $\mathcal{F}_*(\mathrm{HB}_8) \simeq 0.59\,k_\mathrm{B}T$. We observe the same qualitative behavior as the remaining open halves of the two DNA strands bind.

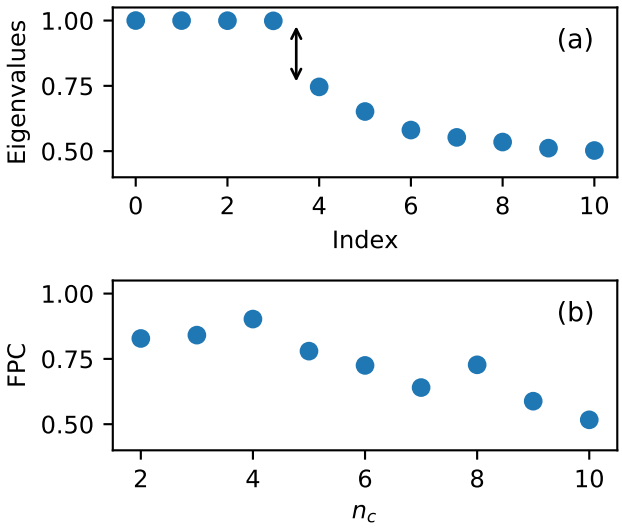

Figure 6: (a) The 11 largest eigenvalues of the transition matrix obtained from the discretized latent space. The eigenvalue with index zero is exactly unity and corresponds to the stationary state. Note the gap between fourth and fifth eigenvalue (arrow). (b) Fuzzy partition coefficient [Eq. (5)] as a function of the number n_c of metastable basins.

In Fig. 6(a), we plot the sorted eigenvalues of the transition matrix. As ensured by the Perron-Frobenius theorem, the largest eigenvalue is unity corresponding to the stationary state plotted in Fig. 5(a). The next three eigenvalues remain close to unity and correspond to the slowest modes, whereas the following eigenvalues are substantially smaller. The dynamics projected onto the latent space thus exhibits a separation between slow collective degrees of freedom and fast relaxation. Each of the slow eigenvectors describes a hierarchical division of the latent space into basins that we identify as long-lived metastable states.[39] For three slow eigenvectors we thus expect four basins. This is corroborated by the fuzzy partition

coefficient (FPC)[42]

$$\mathrm{FPC}(n_\mathrm{c}) = \frac{1}{m}\sum_{\alpha=1}^{n_\mathrm{c}}\sum_{j=1}^{m} u_{i\alpha}^2, \tag{5}$$

where $u_{i\alpha}$ is the probability to find bin i in the basin α, m is the number of non-empty bins, and n_c is the number of metastable basins. The value of FPC thus varies from $1/n_\mathrm{c}$ for equal membership (completely fuzzy) to unity (for each bin the value of one probability is unity and the rest are zero). To calculate $u_{i\alpha}$, we employ the robust Perron cluster analysis (PCCA+) algorithm. In Fig. 6(b), the FPC is plotted as a function of n_c. It exhibits a non-trivial maximum close to unity at $n_\mathrm{c} = 4$, which confirms that the configuration space is optimally partitioned into four metastable basins. Figure 5(b) shows the resulting identification of these metastable basins in the latent space (membership is determined by the maximal probability $u_{i\alpha}$). We see that each basin represents a connected manifold of bins and thus configurations that are structurally similar. Importantly, each basin includes one local minimum of the free energy and basins are separated by free-energy barriers. The configurations contributing to the four minima exhibit H-bonds according to the UB, HB_7, HB_8, and FB states (Fig. 1), and we identify the four basins with these conformations.

Transition pathways

With this low-dimensional representation, we now discuss the two competing pathways from the UB to the FB state *via* either the HB_7 or the HB_8 states. We construct the corresponding two minimum free energy paths (MEPs) as follows: Starting from the fully unbound state [the red bin in Fig. 5(a)], we choose the neighboring bin with the smallest free energy as the next position and repeat this procedure until the FB state is reached.

The resulting MEPs $\mathcal{F}_*(s)$ are shown in Fig. 7 as functions of the length s of the path. Both paths are qualitatively similar with two barriers on the order of $\Delta\mathcal{F} \approx 7 - 11\, k_\mathrm{B}T$ that have to be overcome to bind the two ssDNA fragments. The specific value of s does not have a direct meaning but it conveys how similar two configurations are along the MEP. All four

barriers are asymmetric along s, with a slow approach toward the barriers and a sudden drop into the following minimum. Moreover, we find that the maxima of the MEPs separate the metastable basins, indicating that the latent space indeed is a connected representation of the slowest degrees of freedom. Figure 8 shows configurations representing the four transition-state ensembles at the free energy barriers. Looking more closely at the values of s when base pairs bind, we find a qualitative difference between the first and second barrier: Coming from the UB state, base pairs fully close (establish H-bonds) only after crossing the free energy barrier [cf. Fig. 8(a,c)], whereas the closing events are spread along the flatter approach to the FB barrier. At the barrier, all base pairs have bound [cf. Fig. 8(b,d)], quickly falling into the FB state.

The final MSM comprising the four metastable basins is shown in Fig. 9, for which the transition probability matrix is computed again using PyEMMA.[40] Staying in the current basin has the highest probabilities ($\simeq 0.99$), implying that transitions are indeed rare, which is in agreement with the timescale separation. Also, direct (fast) transitions between the UB and FB states, as well as between HB_7 and HB_8 states, are suppressed (the probabilities are at least one order of magnitude smaller than the smallest probability for UB$\rightarrow$HB$_8$). The probabilities to enter or leave the FB state are of the same order ($\simeq 10^{-4}$), with a larger

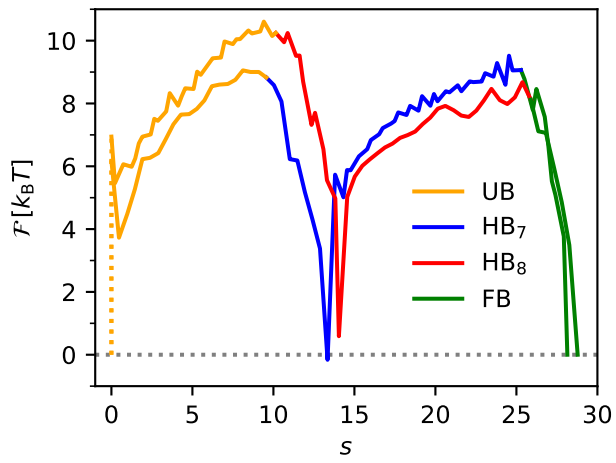

Figure 7: Minimum free energy paths $\mathcal{F}_*(s)$ of the competing pathways UB$\rightarrow$HB$_7\rightarrow$FB and UB$\rightarrow$HB$_8\rightarrow$FB as functions of the Euclidean path length s in latent space units.

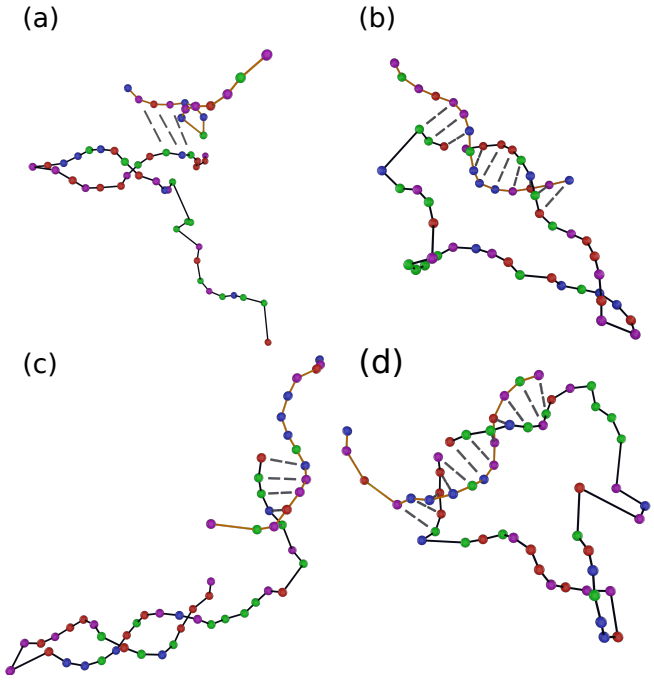

Figure 8: Transition configurations. First row (a) from UB to HB_8 and (b) from HB_8 to FB. Second row (c) from UB to HB_7 and (d) from HB_7 to FB. The configurations shown have the smallest residual mean-square displacement to the average (centroid) of all configurations contributing to the corresponding bin of the barrier.

probability to occur *via* the HB_8 state. The probabilities that the half-bound states dissolve to the unbound state are of the same order as reaching the FB state. Figure 9 shows that the kinetic bottleneck is reaching one of the half-bound states from the fully unbound state with probabilities that are one (UB→HB_7) or two (UB→HB_8) orders of magnitude smaller than reaching/leaving the FB state. This finding agrees qualitatively with the MEPs shown in Fig. 7, where the barriers separating the HB_7 state are first smaller and then higher (between the half-bound and fully bound states) compared to the HB_8 barriers.

Conclusions

We have demonstrated how unsupervised machine learning can be employed to gain molecular insight into the self-assembly of two single-stranded DNA (ssDNA) fragments. Somewhat similar to the folding of biomolecules, we can think about the self-assembly process in terms

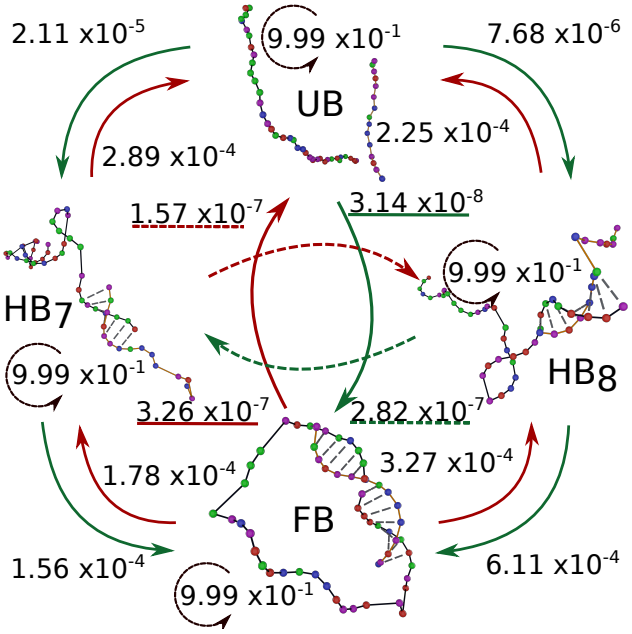

Figure 9: Final Markov state model for the four clustered states of Fig. 5(b). The values inside the configurations are the transition probabilities to stay in the same state. The outer (green) arrows indicate the transition probabilities toward a more bound state, while the outer (red) arrows symbolize the transition from a more bound state to a less bound state. In the center, the transition from the UB to the FB state is indicated by the right (green) arrow and the left (red) indicates the opposite direction. The dashed arrows represent the transition probabilities between the two half-bound states HB_7 and HB_8.

of long-lived molecular conformations and rare transitions. To identify these conformations, we employ an autoencoder neural network to compress a set of structural descriptors into a two-dimensional latent space representation. Our setup differs from other approaches to learn Markov State models (MSM) through machine-learning techniques. First, we do not "learn" the dynamics as, *e.g.*, time-lagged autoencoders[43] but apply a two-step process in which we first determine the mapping of structural descriptors to a latent space and then construct the MSM in this low-dimensional space. This separation allows us to inspect the suitability of the latent space coordinates as order parameters. Second, we endow the latent space with a metric that forces configurations that are close in the original space to remain close in latent space, which is an important ingredient to preserve the kinetic connectivity of the high-dimensional space.[44] For the ssDNA fragments, we find that the metastable basins

are compact sets in the latent space well-separated by free energy barriers. Although the latent space has been determined from structural descriptors alone, the agreement between free energy barriers and the boundaries of metastable basins identified through a MSM strongly indicates that the latent space is a close representation of the slowest degrees of freedom. Extracting one-dimensional minimum free energy paths allows to study transition pathways and in principle to apply tools from transition state theory.[45]

The techniques described here will also be useful to study self-assembly into extended d-dimensional periodic structures,[46] the theory of which is closely related to nucleation.[47] Even for simple systems, it is challenging to uncover reaction coordinates for such highly cooperative transitions as occurring in nucleation and self-assembly. While optimal linear combinations of a set of candidate order parameters have been determined through maximum-likelihood methods,[48–50] autoencoder architectures have the potential to uncover strongly non-linear optimal mappings to reaction coordinates. Our results show that a comprehensive computational understanding for multi-step and hierarchical self-assembly is in reach through a combination of unsupervised dimensionality reduction of a set of structural descriptors with Markov state modeling.

Methods

We perform molecular dynamics (MD) simulations in the NVT ensemble using the oxDNA2 coarse-grained model,[35] which treats ssDNA as a string of rigid nucleotides moving in an implicit solvent (the specific base pair sequences are taken from an example shipped with the code). The temperature is fixed to $T = 303\,\mathrm{K}$ using an Andersen-like thermostat.[51] The edge length of the cubic simulation box is $340.72\,\text{Å}$, with periodic boundary conditions applied to all three Cartesian directions. The salt concentration of the aqueous solution is $0.5\,M$, resulting in strongly screened electrostatic interactions with a Debye length of roughly $4\,\text{Å}$. The equations of motion are integrated using a time step of $\Delta t = 15.15\,\mathrm{fs}$, and we let

the systems evolve for approximately 30 μs to 45 μs in simulation time. To collect data, we perform 256 independent simulations, each starting from a configuration in which the two DNA strands are well separated from each other.

Acknowledgments

We acknowledge financial support by the German Research Foundation (DFG) through the collaborative research center TRR 146 (Grant No. 404840447), and A.N. further acknowledges funding by the DFG through project NI 1487/2-2. S.L. acknowledges support from the Emergent AI Center funded by the Carl-Zeiss-Stiftung.

References

1. Whitesides, G. M. Self-Assembly at All Scales. *Science* **2002**, *295*, 2418–2421.

2. Glotzer, S. C.; Solomon, M. J. Anisotropy of building blocks and their assembly into complex structures. *Nat. Mater.* **2007**, *6*, 557–562.

3. Sacanna, S.; Pine, D. J.; Yi, G.-R. Engineering shape: The novel geometries of colloidal self-assembly. *Soft Matter* **2013**, *9*, 8096–8106.

4. Seeman, N. C.; Sleiman, H. F. DNA nanotechnology. *Nat. Rev. Mater.* **2017**, *3*, 17068.

5. Whitelam, S.; Jack, R. L. The Statistical Mechanics of Dynamic Pathways to Self-Assembly. *Annu. Rev. Phys. Chem.* **2015**, *66*, 143–163.

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

7. Kuzyk, A.; Jungmann, R.; Acuna, G. P.; Liu, N. DNA Origami Route for Nanophotonics. *ACS Photonics* **2018**, *5*, 1151–1163.

8. Alivisatos, A. P.; Johnsson, K. P.; Peng, X.; Wilson, T. E.; Loweth, C. J.; Bruchez, M. P.; Schultz, P. G. Organization of 'nanocrystal molecules' using DNA. *Nature* **1996**, *382*, 609–611.

9. Angioletti-Uberti, S.; Mognetti, B. M.; Frenkel, D. Re-entrant melting as a design principle for DNA-coated colloids. *Nat. Mater.* **2012**, *11*, 518–522.

10. Yi, G.-R.; Pine, D. J.; Sacanna, S. Recent progress on patchy colloids and their self-assembly. *J. Phys. Condens. Matter* **2013**, *25*, 193101.

11. Maye, M. M.; Thilak Kumara, M.; Nykypanchuk, D.; Sherman, W. B.; Gang, O. Switching binary states of nanoparticle superlattices and dimer clusters by DNA strands. *Nat. Nanotechnol.* **2010**, *5*, 116–1209.

12. Zhang, Y.; Lu, F.; Yager, K. G.; van der Lelie, D.; Gang, O. A general strategy for the DNA-mediated self-assembly of functional nanoparticles into heterogeneous systems. *Nat. Nanotechnol.* **2013**, *8*, 865–872.

13. Sciortino, F.; Zhang, Y.; Gang, O.; Kumar, S. K. Combinatorial-entropy-driven aggregation in DNA-grafted nanoparticles. *ACS Nano* **2020**, *14*, 5628–5635.

14. Mao, R.; Mittal, J. Self-assembly of DNA-functionalized nanoparticles guided by binding kinetics. *J. Phys. Chem. B* **2020**, *124*, 11593–11599.

15. Mao, R.; Pretti, E.; Mittal, J. Temperature-controlled reconfigurable nanoparticle binary superlattices. *ACS Nano* **2021**,

16. McManus, J. J.; Charbonneau, P.; Zaccarelli, E.; Asherie, N. The physics of protein self-assembly. *Curr. Opin. Colloid Interface Sci.* **2016**, *22*, 73–79.

17. Liu, H.; Kumar, S. K.; Sciortino, F. Vapor-liquid coexistence of patchy models: Relevance to protein phase behavior. *J. Chem. Phys.* **2007**, *127*, 084902.

18. Allen, R. J.; Valeriani, C.; ten Wolde, P. R. Forward flux sampling for rare event simulations. *J. Phys. Condens. Matter* **2009**, *21*, 3102–3123.

19. DeFever, R. S.; Sarupria, S. Contour forward flux sampling: Sampling rare events along multiple collective variables. *J. Chem. Phys.* **2019**, *150*, 024103.

20. Sarwar, H.; Haji-Akbari, A. Studying rare events using forward-flux sampling: Recent breakthroughs and future outlook. *J. Chem. Phys.* **2020**, *152*, 060901.

21. Matthews, C.; Weare, J.; Kravtsov, A.; Jennings, E. Umbrella sampling: a powerful method to sample tails of distributions. *Monthly Notices of the Royal Astronomical Society* **2018**, *480*, 4069–4079.

22. Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

23. Bolhuis, P. G.; Swenson, D. W. H. Transition Path Sampling as Markov Chain Monte Carlo of Trajectories: Recent Algorithms, Software, Applications, and Future Outlook. *Adv. Theory Simul.* **2021**, *4*, 2000237.

24. Wang, F.; Landau, D. P. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Phys. Rev. Lett.* **2001**, *86*, 2050–2053.

25. Brown, W. M.; Martin, S.; Pollock, S. N.; Coutsias, E. A.; Watson, J.-P. Algorithmic dimensionality reduction for molecular structure analysis. *J. Chem. Phys.* **2008**, *129*, 064118.

26. Ferguson, A. L. Machine learning and data science in soft materials engineering. *J. Condens. Matter Phys.* **2017**, *30*, 043002.

27. Bittracher, A.; Banisch, R.; Schütte, C. Data-driven computation of molecular reaction coordinates. *J. Chem. Phys.* **2018**, *149*, 154103.

28. Jackson, N. E.; Webb, M. A.; de Pablo, J. J. Recent advances in machine learning towards multiscale soft materials design. *Curr. Opin. Colloid Interface Sci.* **2019**, *23*, 106–114.

29. Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361–390.

30. Coli, G. M.; Dijkstra, M. An Artificial Neural Network Reveals the Nucleation Mechanism of a Binary Colloidal AB13 Crystal. *ACS Nano* **2021**, *15*, 4335–4346.

31. Baldi, P. Autoencoders, Unsupervised Learning, and Deep Architectures. Proceedings of ICML Workshop on Unsupervised and Transfer Learning. Bellevue, Washington, USA, 2012; pp 37–49.

32. Chen, W.; Ferguson, A. L. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *J. Comput. Chem.* **2018**, *39*, 2079–2102.

33. Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 13023–13028.

34. Lemke, T.; Peter, C. EncoderMap: Dimensionality Reduction and Generation of Molecule Conformations. *J. Chem. Theory Comput.* **2019**, *15*, 1209–1215.

35. Snodin, B. E. K.; Randisi, F.; Mosayebi, M.; Šulc, P.; Schreck, J. S.; Romano, F.; Ouldridge, T. E.; Tsukanov, R.; Nir, E.; Louis, A. A.; Doye, J. P. K. Introducing improved structural properties and salt dependence into a coarse-grained model of DNA. *J. Chem. Phys.* **2015**, *142*, 234901.

36. Tishby, N.; Pereira, F. C.; Bialek, W. The information bottleneck method. Proceedings of the 37-th Annual Allerton Conference on Communication. 1999; pp 368–377.

37. Saxe, A. M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B. D.; Cox, D. D. On the information bottleneck theory of deep learning. *J. Stat. Mech.: Theory Exp.* **2019**, *2019*, 124020.

38. Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **2010**, *52*, 99–105.

39. Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **2011**, *134*, 174105.

40. Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525–5542.

41. Ouldridge, T. Coarse-grained modelling of DNA and DNA self-assembly. Ph.D. thesis, Oxford University, UK, 2011.

42. Dunn, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *J. Cybern.* **1973**, *3*, 32–57.

43. Wehmeyer, C.; Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.* **2018**, *148*, 241703.

44. Noé, F.; Clementi, C. Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2015**, *11*, 5002–5011.

45. Hummer, G. From transition paths to transition states and rate coefficients. *J. Chem. Phys.* **2004**, *120*, 516–523.

46. Grünwald, M.; Geissler, P. L. Patterns without Patches: Hierarchical Self-Assembly of Complex Structures from Simple Building Blocks. *ACS Nano* **2014**, *8*, 5891–5897.

47. Lutsko, J. F. How crystals form: A theory of nucleation pathways. *Sci. Adv.* **2019**, *5*, eaav7399.

48. Lechner, W.; Dellago, C.; Bolhuis, P. G. Role of the Prestructured Surface Cloud in Crystal Nucleation. *Phys. Rev. Lett.* **2011**, *106*, 085701.

49. Jungblut, S.; Singraber, A.; Dellago, C. Optimising reaction coordinates for crystallisation by tuning the crystallinity definition. *Mol. Phys.* **2013**, *111*, 3527–3533.

50. Campo, M.; Speck, T. Polydisperse hard spheres: crystallization kinetics in small systems and role of local structure. *J. Stat. Mech.: Theory Exp.* **2016**, *2016*, 084007.

51. Frenkel, D.; Smit, B. *Understanding molecular simulation: From algorithms to applications*; Academic Press, 2001.