Article

Pattern recognition in the nucleation kinetics of non-equilibrium self-assembly

Constantine Glen Evans[1,2,3 ✉], Jackson O'Brien[4 ✉], Erik Winfree[1 ✉] & Arvind Murugan[4 ✉]

Inspired by biology's most sophisticated computer, the brain, neural networks constitute a profound reformulation of computational principles[1–3]. Analogous high-dimensional, highly interconnected computational architectures also arise within information-processing molecular systems inside living cells, such as signal transduction cascades and genetic regulatory networks[4–7]. Might collective modes analogous to neural computation be found more broadly in other physical and chemical processes, even those that ostensibly play non-information-processing roles? Here we examine nucleation during self-assembly of multicomponent structures, showing that high-dimensional patterns of concentrations can be discriminated and classified in a manner similar to neural network computation. Specifically, we design a set of 917 DNA tiles that can self-assemble in three alternative ways such that competitive nucleation depends sensitively on the extent of colocalization of high-concentration tiles within the three structures. The system was trained in silico to classify a set of 18 grayscale 30 × 30 pixel images into three categories. Experimentally, fluorescence and atomic force microscopy measurements during and after a 150 hour anneal established that all trained images were correctly classified, whereas a test set of image variations probed the robustness of the results. Although slow compared to previous biochemical neural networks, our approach is compact, robust and scalable. Our findings suggest that ubiquitous physical phenomena, such as nucleation, may hold powerful information-processing capabilities when they occur within high-dimensional multicomponent systems.

The success of life on Earth derives from its use of molecules to carry information, implement algorithms that control chemistry and respond intelligently to the environment. Genetic information encodes not only molecules with structural and chemical functionality, but also biochemical circuits that in turn process internal and external information relevant for cellular decision-making. Whereas some biological systems may, like modern modular engineering, isolate information processing from the physical subsystems being controlled[8], other critical decision-making may be embedded within and inseparable from processes such as protein synthesis, metabolism, self-assembly and structural reconfiguration. Understanding such physically entangled computation is necessary not only for understanding biology, but also for engineering autonomous molecular systems such as artificial cells, in which it is essential to pack as much capability as possible within limited space and energy budgets.

The interplay of structure and computation is particularly rich in molecular self-assembly. In biological cells, decisions about navigation, chemotaxis and phagocytosis are made through structural rearrangements of the cytoskeleton that integrate mechanical forces and chemical signals[9–12], but where and how information processing occurs remains elusive. In DNA nanotechnology[13], self-assembly of DNA tiles has been shown theoretically and experimentally to be capable of Turing-universal computation through simulation of cellular automata and Boolean circuits[14–16], but this digital model of computation lacks a clear analogue in biology.

Neural computation is an alternative form of naturally compact computation with several distinctive hallmarks[1–3]: mixed analogue and digital decision-making, recognition of high-dimensional patterns, reliance on the collective influence of many distributed weak interactions, robustness to noise and an inherent ability to learn and generalize. A paradigmatic neural network model is the Hopfield associative memory[17], which conceptualizes dynamics as a random walk on an energy landscape that has been sculpted by learning to contain attractor basins at each memory. Remarkably, neural network models map naturally onto models of well-mixed chemical networks[4,5], genetic regulatory networks[6] and signal transduction cascades[7]; such networks have been experimentally demonstrated both in cell-free systems and within living cells[18–21]. However, these well-mixed approaches still separate decision-making from downstream processes.

Neural information-processing principles embedded within molecular self-assembly have been harder to discern, and perhaps at first appear as a contradiction in terms. An early thermodynamic view of how free-energy minimization in molecular self-assembly could be akin to the Hopfield model did not lead to concrete realizations[22]. However,

[1]California Institute of Technology, Pasadena, CA, USA. [2]Evans Foundation for Molecular Medicine, Pasadena, CA, USA. [3]Maynooth University, Maynooth, Ireland. [4]University of Chicago, Chicago, IL, USA. ✉e-mail: cge@dna.caltech.edu; jdobrien07@gmail.com; winfree@caltech.edu; amurugan@uchicago.edu

a recent kinetic view of multicomponent systems that permit assembly of many distinct structures using the same components ('multifarious self-assembly')[23,24] revealed concrete connections to Hopfield associative memories[17] and models of hippocampal place cells[25] at the level of collective dynamics, even though individual molecules do not explicitly mimic the mechanistic behaviour of individual neurons.

Here we reformulate this connection as an intrinsic feature of heterogeneous nucleation kinetics and experimentally demonstrate its power for high-dimensional pattern recognition using DNA nanotechnology[13]. The phenomenon arises when the same components can form several distinct assemblies in different geometric arrangements (Fig. 1). Nucleation proceeds by spontaneous formation of a critical seed that subsequently grows into a structure[26]. Because the nucleation rate of a seed depends strongly on the bulk concentrations of components that occur in that seed, and many distinct seeds and pathways may be viable, the overall rate of formation of a given structure is a complex function of the concentration pattern. Further, because components are shared between structures, competition for resources[27] results in a winner-take-all (WTA) effect that accentuates the discrimination between concentration patterns.

Molecular system design

To explore these principles experimentally, we take advantage of the powerful foundation provided by DNA nanotechnology for programming molecular self-assembly. The well-understood kinetics and thermodynamics of Watson–Crick base pairing enables systematic sequence design[28,29] for DNA tiles that reliably self-assemble into periodic, uniquely addressed and algorithmically patterned structures with hundreds to thousands of distinct tile types[15,16,30–34]. These classes of self-assembly differ in the structures produced and in the nature of interactions: in periodic and uniquely addressed structures, each molecular component typically has a unique possible binding partner in each direction. For algorithmic patterns (as for multifarious assembly), some components have multiple possible binding partners, such that which one attaches at a given location is decided during self-assembly on the basis of which forms more bonds with neighbouring tiles.

We build on these ideas to create a molecular system capable of assembling multiple target structures (H, A and M in Fig. 2) from a shared set of interacting components by colocalizing them in different ways. The first stage of design begins with a set S of shared tiles that do not directly bind each other; then three sets of interaction-mediating tiles (also called H, A and M) are introduced for each of the respective desired structures. Each interaction tile in, for example, H, binds four specific S tiles together in a chequerboard arrangement that reflects neighbourhood constraints between shared S tiles in structure H. These H interaction tiles are unique to structure H and do not occur in the assembled A or M structures.

Tiles in a 1:1 stoichiometric mix of S + H, S + A or S + M will have no promiscuous interactions and will assemble H, A or M, respectively, as with previous work on uniquely addressable structures[32]. But a 1:1:1:1 mix of S + H + A + M, henceforth called our SHAM mix, can assemble three distinct structures. This additive construction of interaction-mediating tiles is analogous to Hebbian learning of multiple memories in Hopfield neural networks[17,23] (Extended Data Fig. 1). Furthermore, the use of interaction-mediating tiles avoids constraints from Watson–Crick complementarity, allowing almost arbitrary interactions to be engineered between S tiles. To avoid undesired consequences of the extensive promiscuous interactions present in the SHAM mix, later design stages optimized this initial layout using self-assembly proofreading principles to reduce errors[35,36] (Extended Data Figs. 2 and 3).

The resulting design in Fig. 2b has 168 tiles shared across all three shapes, 203 tiles shared across a pair and 546 tiles unique to a specific shape. Our experimental implementation used 42-nucleotide
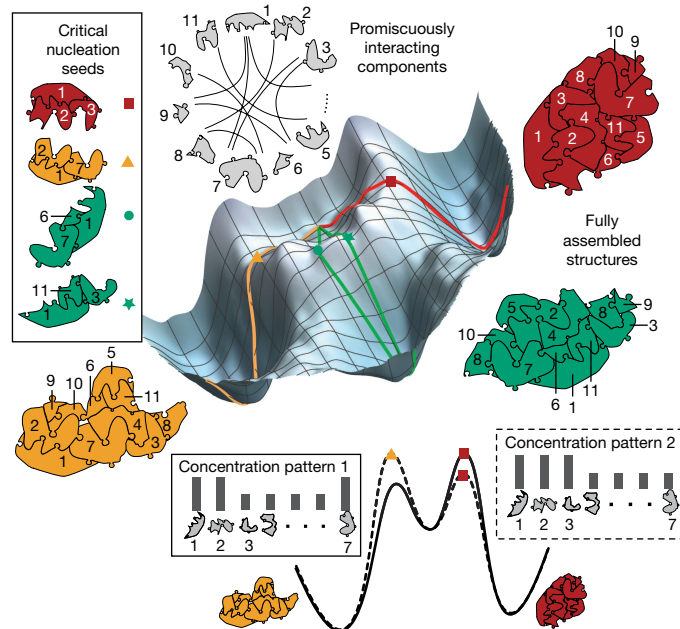

Fig. 1 | Conceptual framework for pattern recognition by nucleation. When one set of molecules can potentially assemble multiple distinct structures, the nucleation process that selects between outcomes is responsive to high-dimensional concentration patterns. Assembly pathways can be depicted on an energy landscape (schematic shown) as paths from a basin for unassembled components that proceed through critical nucleation seeds (barriers) to a basin for each possible final structure. Seeds that colocalize high-concentration components will lower the nucleation barrier for corresponding assembly pathways. The resulting selectivity of nucleation in high-dimensional self-assembly is sufficiently expressive to perform complex pattern recognition in a manner analogous to neural computation (Extended Data Fig. 1).

single-stranded DNA tiles[32] (Fig. 2a) with sequences designed using tools from previous work[16] to reduce unintended interactions and secondary structure and to ensure nearly uniform binding energies.

To test whether proofreading was sufficient to combat promiscuity and to test the unbiased yield of different structures, we annealed all tiles at equal concentration (60 nM) in solution over 150 hours from 48 to 45 °C. Atomic force microscopy (AFM) revealed a roughly equal yield of all three structures (Fig. 2c). Despite being a slow anneal, this uniform distribution is incompatible with an equilibrium Boltzmann distribution that would exponentially magnify differences in the area and perimeter (and thus free energy) of H, A and M; but it is compatible with kinetically controlled assembly in which nucleation rates are linearly proportional to a shape's area, as nucleation could occur anywhere within the shape. Furthermore, we did not observe significant chimeric structures or uncontrolled aggregation, indicating that proofreading was functioning as desired. However, many structures appeared to be incomplete—often missing tiles from two specific corners, perhaps due to asymmetric growth kinetics or lattice curvature[31]—or (in the case of A only) showed signs of spiral defect growth (Extended Data Fig. 3).

Colocalization controls nucleation

Understanding nucleation in multicomponent self-assembly has required extensions of classical nucleation theory[26] that have effectively guided the design of programmable DNA tile systems with well-defined assembly pathways[37–40]. Building on this work, here we examine how selection between target structures that differ in colocalization of tiles can be determined by nucleation kinetics and controlled by concentra-
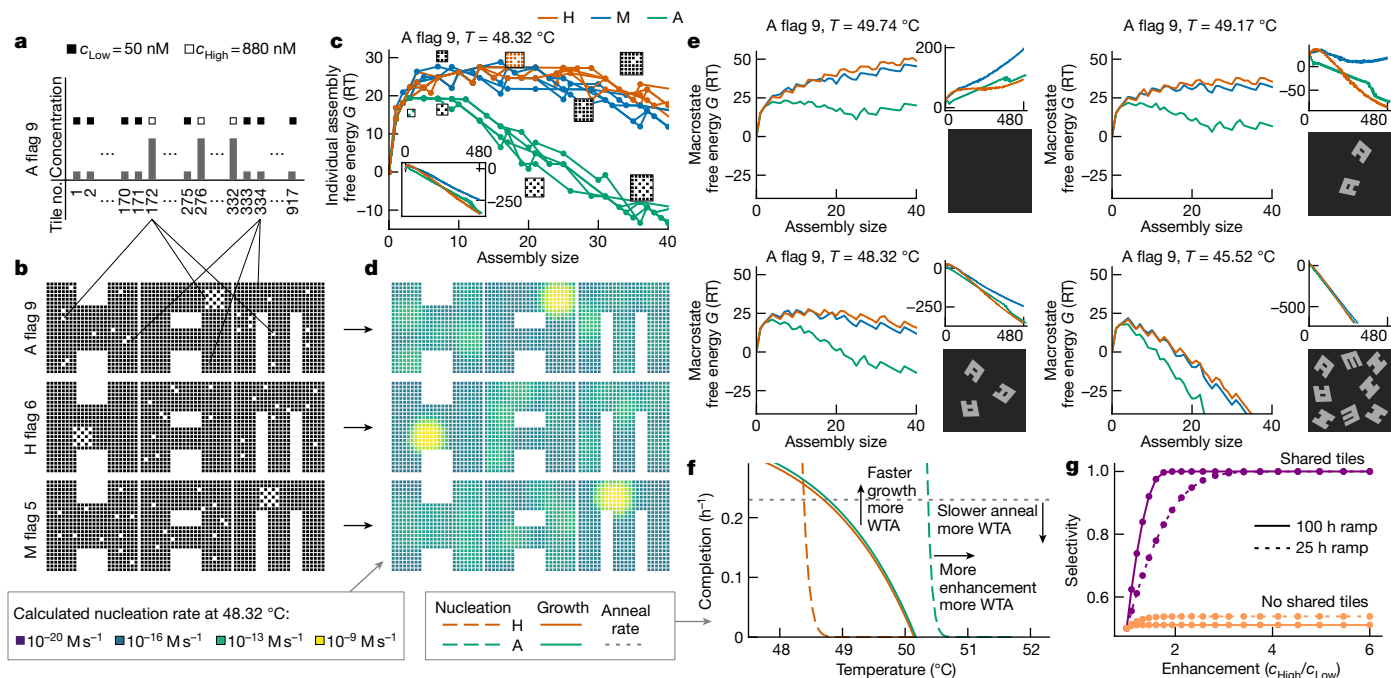
Fig. 2 | A multifarious mixture of 917 molecular species that can assemble into three distinct structures from one set of molecules. a, Here 42-nucleotide DNA strands self-assemble into two-dimensional (2D) structures by forming bonds with four complementary strands using four 10 or 11 nucleotide domains. The strands can be abstracted as square tiles, each named and shown with distinct binding domains identified by number, such that, for example, 708 is complementary to 708*. At nucleation and growth temperatures, attaching by two bonds or more is favourable whereas one is insufficient. **b**, One pool of 917 tile types assembles into three distinct shapes, H, A and M, through a multitude of pathways. Whereas each tile occurs at most once in each shape, the shared purple species recur in multiple shapes, in distinct spatial arrangements; for example, S149 is highlighted in red. **c**, Annealing an equal mix of all tiles results in a mixture of fully and partially assembled H, A and M, imaged by AFM. This is the same sample as SHAM60 in Fig. 6e. The inset illustrates the expected slant of the shapes due to SST geometry. Scale bars, 50 nm. **d**, A typical experiment mixes the desired concentrations of each tile type into a single tube, with some tiles swapped for fluorophore- and quencher-modified versions. The sample is heated to remove any pre-existing binding, cooled to a temperature slightly above where any growth is observed, then slowly annealed through a small range of temperatures while fluorescence is measured in a qPCR machine; samples are then imaged by AFM.

tion patterns. We model the free energy of a structure A with B total bonds as $G(A) = \sum_{i \in A} G_{mc}^i - BG_{se} - \alpha$, where α depends on the choice of reference concentration u_0, $G_{mc}^i = \alpha - \log c_i/u_0$ is the chemical potential (or equivalently, translational entropy) of tile i at concentration c_i and G_{se} is the energy of each bond in units of RT, the molar gas constant times temperature. $G(A)$ has competing contributions that scale with the structure's area and perimeter, and is hence maximized for certain partial assemblies called critical nucleation seeds. The formation of such seeds is often rate-limiting: once these seeds are assembled, subsequent growth is faster and mostly 'downhill' in free energy. If the nucleation rate η_{shape} for a given shape is dominated by a single critical nucleus A_s, we could use an Arrhenius-like approximation $\eta_{shape} \propto e^{-G(A_s)}$; in the case that multiple critical nuclei are significant, we must perform a sum.

When such analyses are applied to homogeneous crystals with uniform concentration $c_i = c$ of components, critical nuclei are simply those with the appropriate balance of size and perimeter. Heterogeneous concentration patterns require a more nuanced analysis: critical seeds can now be arbitrarily shaped, potentially offsetting a larger perimeter penalty by incorporating tiles with higher bulk concentration. Therefore, we implemented a stochastic sampling algorithm to estimate the nucleation rate of a structure with an uneven pattern of concentrations (Extended Data Fig. 4).

Consider the examples in Fig. 3 where the concentrations of some shared tiles in the SHAM mix have been enhanced. These high-concentration tiles are colocalized in structure A but scattered

across H and M. Consequently, such a pattern will lower kinetic barriers for the nucleation of A while maintaining high barriers for H and M. The typical area K over which colocalization promotes nucleation can be estimated from the size of critical seeds predicted by classical nucleation theory and is generally larger at higher temperatures[26]. Hence, we expect a trade-off between speed and complexity of pattern recognition (Fig. 3e), with more subtle discrimination at higher temperatures (large K)—at the expense of slower experiments—and lower discriminatory power at lower temperatures (small K).

To experimentally characterize the basis of selectivity, we systematically tested a series of 37 concentration patterns, which we call 'flags' because each one uses high concentrations in a chequerboard localized somewhere in one of the shapes (three examples are shown in Fig. 3b). We did not enhance concentrations of tiles unique to shapes, to avoid additional thermodynamic bias towards any one structure. We ramped the temperature down slowly, from 48 to 46 °C (the expected range for nucleation, a few degrees below the melting temperatures) to provide robustness to variations in nucleation temperatures among flags in different locations and to probe for slow off-target nucleation. To monitor nucleation and growth in real time, we designed distinct fluorophore–quencher pairs on adjacent tiles in four locations on each shape, using tiles not shared between shapes. Each pair quenches when the local region of that specific structure assembles (Fig. 4a).

Experimental results illustrating selective nucleation are shown in Fig. 4c for three example flag concentration patterns. When the pattern localizes high-concentration species in a structure, for example,

Fig. 3 | Theory shows selective nucleation when high-concentration tiles are colocalized in one shape more than in others. a, One pattern (A flag 9) enhancing the concentration of shared tiles colocalized in A but relatively dispersed in H and M. **b**, A flag 9, plotted by tile locations in each shape along with example flag patterns that have colocalization in H and M. **c**, For A flag 9, free energies of assemblies along predicted nucleation pathways for each shape (Extended Data Fig. 4). Several example assemblies are shown; the green and red ones are critical seeds for the A and H pathways, respectively. **d**, Regions predicted to participate in nucleation by the simulation for three concentration patterns (lighter colours correspond to higher participation). **e**, Macrostate free energies for sets of partial assemblies of increasing size (number of tiles) and predicted AFM results at several temperatures spanning the melting temperature. Small plots show the full-size range, thus illustrating the independence of the nucleation barrier kinetics and the complete assembly

thermodynamics. **f**, For on-target (A, green) and off-target (H, red) shapes, nucleation rates (dashed) and growth rates (solid) are plotted as a function of temperature, according to the simplified model of Extended Data Fig. 4f. Rates are given relative to the time to completely consume the lowest-concentration tile; the horizontal dotted line indicates the rate of annealing between the on-target to off-target nucleation temperatures. Owing to the higher nucleation temperature for the on-target shape, when annealing time scales are comparable to or slower than growth time scales, depletion of shared tiles during a temperature anneal can lead to a WTA effect. Slower annealing and faster growth can increase the WTA effect. **g**, In this model, WTA leads to higher selectivity (on-target versus total nucleation) compared to systems with no shared components; for slower anneals, selectivity increases for systems with shared components, but decreases for systems with no shared components.

H, the fluorophore in the expected nucleation region of that structure quenched first and rapidly. After a delay, fluorophore signals from other parts of the same structure also dropped, indicating growth. Fluorophores on off-target structures showed minimal to no quenching until late in the experiment. AFM images from samples at the end of the experiment confirm that fluorophore quenching corresponded to selective self-assembly of complete or partial shapes. Of the 37 flag positions, roughly half showed robust selective nucleation and growth (Fig. 4d,e), while other positions were either not selective or did not grow well, for reasons we have not been able to determine.

In multifarious systems, we expect enhanced selectivity because of a competitive suppression of nucleation. Using an annealing protocol that spends sufficient time at temperatures in which A can nucleate and grow significantly, but H cannot nucleate (Fig. 3f), we expect a WTA effect in which the assembly of A depletes shared tiles S and thus actively suppresses nucleation of H. As shown in Fig. 4f, we see evidence for this effect in most experiments, suggesting that WTA dynamics is amplifying small differences in nucleation kinetics.

Pattern recognition by nucleation

Our work thus far shows that the space of all concentration patterns, which includes patterns not experimentally tested, consists of regions that result in the selective assembly of each of H, A and M, respectively (Fig. 5a). These regions together represent a phase diagram for this

self-assembling system[23] that reflects the decisions it makes to classify concentration patterns. Whereas phase boundaries of traditionally studied physical systems are usually low dimensional and not fruitfully interpreted as decision boundaries, in multicomponent heterogeneous systems such as ours, the phase diagram is naturally high dimensional. More generally, phase boundaries in disordered many-body systems tend to be complex and thus implicitly solve complex pattern recognition problems, a perspective that also underlies Hopfield's associative memory in neural networks[17,41].

Here nucleation is solving a particular pattern recognition problem based on which molecules are colocalized in different structures. Similar colocalization-based decision boundaries arise in neural place cells studied by the Mosers[24,25,42,43] and are complex enough to solve pattern recognition problems and permit statistically robust learning. Having demonstrated that multifarious self-assembly can solve a specific pattern recognition problem, could different molecules be designed to solve other tasks such as recognizing or classifying images? Here the grayscale value of each pixel position in the 30×30 images is taken to represent the concentration of a distinct molecule. Instead of synthesizing new molecules with new interactions to solve the above challenge, we show that the design problem is solvable with our existing molecules by an optimized choice of a pixel-to-tile map θ that specifies which existing tile should correspond to which pixel position (Fig. 5b). In addition to saving DNA synthesis costs, this approach helps demonstrate that a random molecular design can be exploited,
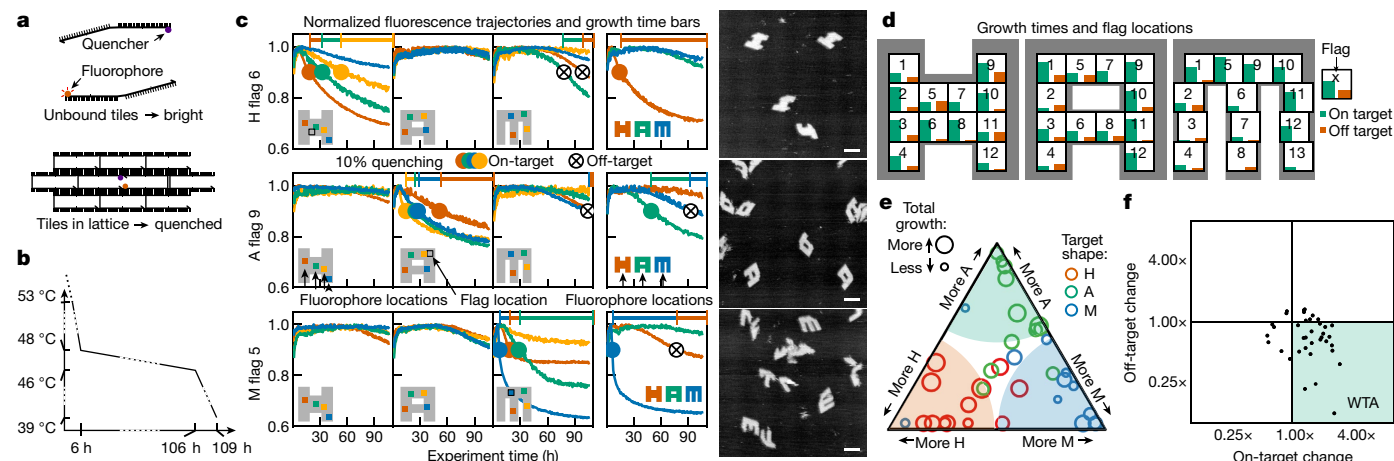
Fig. 4 | Selective nucleation in experiments with shape-specific localized concentration patterns of shared tiles. a, Pairs of alternative tiles with a fluorophore and quencher (Fig. 2d) have their fluorescence quenched when incorporated together in an assembly; small assemblies of just a few strands do not effectively quench (Extended Data Fig. 5). **b**, Samples were annealed with a temperature protocol that cooled from 71 °C (well above melting temperature) to 48 °C over roughly 6 hours, cooled to 46 °C over 100 hours and finally cooled to 39.5 °C over 3 hours (Extended Data Fig. 6). **c**, Experimental results for the three flag patterns shown in Fig. 3. The positions of fluorophore–quencher tile pairs used in each of the four samples are shown by the inset icons. Points where fluorescence signals dropped by 10% below their maximum (to which signals were normalized) are shown with coloured dots for on-target nucleation and with ⊗ for off-target nucleation. 'Growth times' measure the period from '10% quenching' to the end of the experiment, shown as horizontal bars. Sample AFM images from one of the samples are shown for each flag. Scale bars, 100 nm. **d**, Total growth times for on-target versus off-target nucleation are summarized for all 37 flag patterns. Each numbered box indicates the location of the corresponding 5 × 5 chequerboard flag; good performance is indicated by a tall green bar and a short red bar. **e**, The same data shown as a ternary plot, with proximity to triangle corners indicating relative fractions of growth time and circle size indicating overall growth time. **f**, Average change in quenching (a measure of nucleation) of on- and off-target structures with flag patterns compared to equimolar SHAM mixes (Extended Data Fig. 7). For most patterns, increasing shared tile concentrations reduces the absolute off-target nucleation, supporting a WTA effect.

ex post facto, to solve a specific computational problem by modifying how the problem is mapped onto physical components, as done in reservoir computing[44].

We specified our design problem by picking arbitrary images as training sets shown in Fig. 5d. Note that images in one class share no more resemblance than images across classes, for example, class H is Hodgkin, Hopfield, Horse and so on, although the number of pixels and grayscale histogram were standardized across images (Methods). In this way, the number of distinct images per class (six in the experiments presented below) tests the flexibility of decision surfaces inherent to this self-assembling molecular system as a classifier.

We then used an optimization algorithm (Fig. 5c and Methods) on θ that sought to maximize nucleation of the on-target structure for the concentration pattern corresponding to each image while also minimizing off-target nucleation. That is, our algorithm sought to map high-concentration pixels in each image (for example, Mitscherlich) to colocalized tiles in the corresponding on-target structure (here, M) to enhance nucleation, while mapping those same pixels to scattered tiles in undesired structures (here, A and H). Note that this map θ is simultaneously optimized for all images and not independently for each image. Hence no map θ might be able to perfectly satisfy all the above requirements simultaneously for all images in all classes; analogous to associative memory capacity[17,23,41], performance drops as one attempts to train more patterns (Extended Data Fig. 8).

For pattern recognition experiments, we enhanced concentrations of tiles in the SHAM mix in accordance with each of the 18 training images (using the optimized θ) and annealed each of the 18 mixes with a 150 hour ramp from 48 to 45 °C. As verified by AFM imaging and real-time fluorescence quenching, we found that the 18 training images yielded correct nucleation, in the sense that there was more of the correct shape than any other shape and in all but five cases was highly (more than 80%) selective (Fig. 6).

We also tested 12 degraded images and six alternate handwriting images (Fig. 5e), with the same trained pixel-to-tile map θ. Pattern recognition was successful for random speckle distortions and all but one partly obscured image. Generalization, the ability to recognize related images not present in a training set, is a critical aspect of learning in neural networks. A given architecture can be naturally robust to certain families of distortions (for example, convolutional networks can handle translation) but not others (for example, dilation). As nucleation is a cooperative process, often dominated by one or a few critical seeds involving just a handful of tiles, flipping of random uncorrelated pixels and obscuring parts of an image that do not involve those critical pixel combinations will not inhibit nucleation, demonstrating robustness. On the other hand, only three of the six alternate handwritten digits were correctly recognized by self-assembly, indicating a lack of robustness to this type of variation without further training.

Discussion

The phenomena underlying pattern recognition by multifarious self-assembly may be exploited by complex evolved or designed systems (Extended Data Fig. 10). Beyond self-assembly, molecular folding processes could potentially recognize patterns in the concentrations of cofactors or subcomponents if folding kinetics can select between distinct stable states[45]. Similarly, the phase boundaries for multicomponent condensates governing genetic regulation[46] may also contain inherent information-processing capabilities. In such cases, the 'pixel-to-tile' map would instead correspond to a layer of phosphorylation or binding circuitry that activates or deactivates specific components on the basis of the levels of upstream information-bearing molecular signals. Within artificial cells[47], multicomponent nucleation may be an especially compact way to implement decision-making within the limited space constraints.

To better understand the information-processing potential of nucleation, we may treat this physical process as a machine learning model. A key issue is how the complexity of decision surfaces, quantified
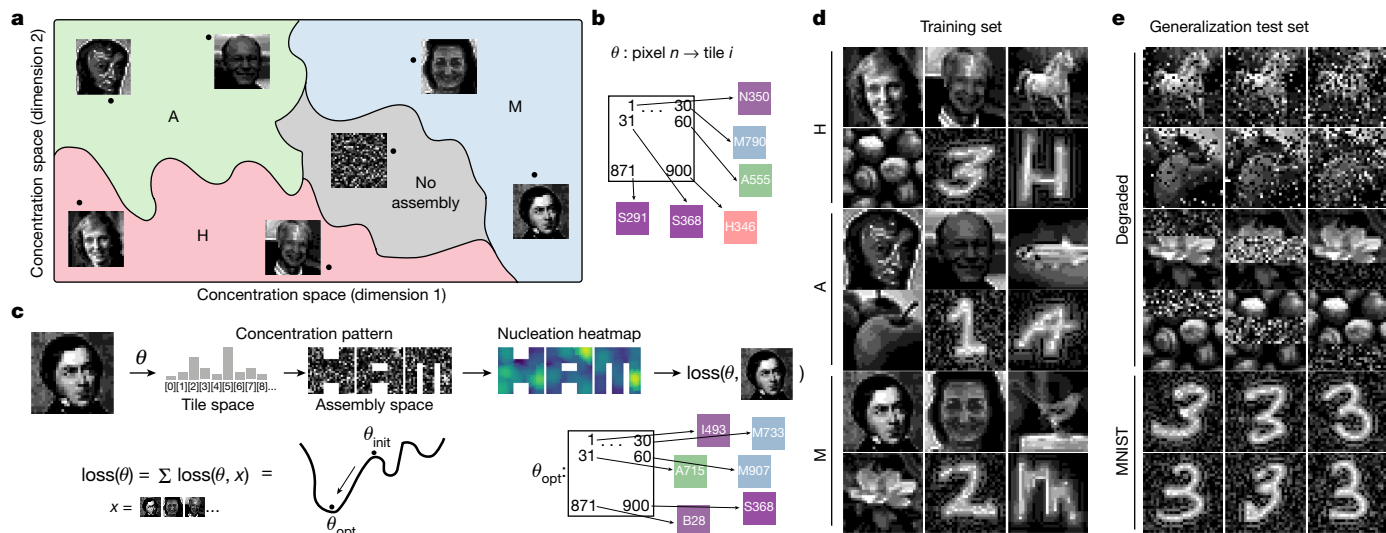
Fig. 5 | Design of self-assembly phase diagrams to solve pattern recognition problems. a, Phase diagram shows desired outcomes of kinetically controlled self-assembly in different regions of $N = 917$ dimensional concentration space (2D schematic shown). Each grayscale image represents a vector of tile concentrations. **b**, θ specifies which pixel location corresponds to which tile. **c**, Given a map θ, any image can be converted to a tile concentration vector by associating the grayscale value of pixel location n with the concentration of the corresponding tile $i = \theta(n)$. We compute the loss for a given pixel-to-tile map θ using simulations to estimate the nucleation rates of desired and undesired structures for each image and summing over a training set. Stochastic optimization in θ space gives a putative optimal θ_{opt} that we used for experiments.

d, Images used for training. **e**, Extra images used to test generalization power. Sources and names of individuals are from left to right as follows in **d** (for details, see Supplementary Information section 2.7). In **a**, **c** and **e**, some of the images are also shown, credits are as for **d**. **d**, Top row, D. Hodgkin, Keystone/Getty Images; J. Hopfield, Princeton University; Horse, Pixabay. Second row: Hazelnuts, Pixabay; Harom, MNIST; H, EMNIST. Third row: A. Avogadro, C. Sentier/University of Pennsylvania; L. Abbott, himself; Anchovy, NOAA/NMFS/SEFSC Pascagoula Laboratory. Fourth row: Apples, M. Shemesh; Aon, MNIST; A, EMNIST. Fifth row: E. Mitscherlich, William Sharpe/Smithsonian Institute; M.-B. Moser, BI Basmo/Kavli Institute of Systems Neuroscience; Mockingbird, Pixabay. Sixth row: Magnolia, D. Richardson; Mbili, MNIST; M, EMNIST.

in terms of computational power or learning capacity, depends on underlying physical aspects of self-assembly such as the number of molecular species, binding specificity and geometry[48,49]. Our work already suggests that temperature mediates a trade-off between speed, accuracy and complexity of pattern recognition; at higher temperatures, nucleation seeds are larger, allowing discrimination on the basis of higher-order correlations in the concentration patterns, but the physical process is also correspondingly slower. The trade-off derives from how computation here exploits the inherently stochastic nature of nucleation; monomers must make many unsuccessful attempts at forming a critical seed for both on- and off-target structures, with repeated disassembly before discovering the seed for the correct pattern recognition outcome. Relating such backtracking to stochastic search algorithms for NP-complete problems, as has been done for well-mixed chemistry[50], might characterize the computational power of stochastic nucleation.

Viewing nucleation as a machine learning model raises the question of whether there is a natural physical implementation of learning. Here we trained decision boundaries in silico using ideas from reservoir computing[44,51]; molecules with a fixed set of interactions could nevertheless solve an arbitrary problem by changing the mapping between inputs and fixed components (Extended Data Fig. 1). The analogy between Hopfield associative memories and multifarious self-assembly, especially those based on random colocalization[23–25,42,43], suggests a way to go beyond fixed components to a scenario in which interactions between components are learned in a Hebbian manner by a natural physical process. Notably, interactions between shared tiles in our system are mediated by shape-specific molecules. If these interaction-mediating tiles could be physically created or activated in response to environmental inputs, for example, through proximity-based ligation, molecular systems could autonomously learn new self-assembling behaviours from examples[52] without the need for computer-based

learning. Alternatively, the natural evolution of hydrophobic residues to stabilize multi-protein complexes may have the necessary properties for inducing multifarious pattern recognition[53].

The connection between self-assembly and neural network computation raises many questions for further exploration, the broadest being a variant on Anderson's observation that 'more is different'[54]. Anderson was referring to the fact that systems containing many copies of the same simple component can show emergent phenomena, such as fluid dynamics, that are best understood at a higher level. Biology also explores another sense of 'more is different': it often makes use of a few copies of a great many different types of component[8]. Here new phenomena naturally emerge in the 'large N limit': robustness, programmability and information processing. These phenomena are best explored in information-rich model systems devoid of the distracting complexities of biology. DNA nanotechnology provides one such platform that already hints at such 'more types is different' phenomena. For example, self-assembled few-component DNA structures are often sensitive to sequence details and molecular purity, thus taking years to refine experimentally, whereas DNA origami[55] and uniquely addressed tile systems[32–34] use hundreds to thousands of components and usually work on the first try, even with unpurified strands, imprecise stoichiometry and no sequence optimization. Such observations suggest heterogeneity as a defining principle for biological self-assembly[56].

Our work adds sophisticated information processing as a new emergent phenomenon in which self-assembly, in the multicomponent limit, gains programmable and potentially learnable phase boundaries to solve specific pattern recognition problems, analogous to earlier results for large N neural networks[41]. This neural network inspired perspective may help us recognize information processing in high-dimensional molecular systems that is deeply entangled within physical processes, whether in biology or in molecular engineering:
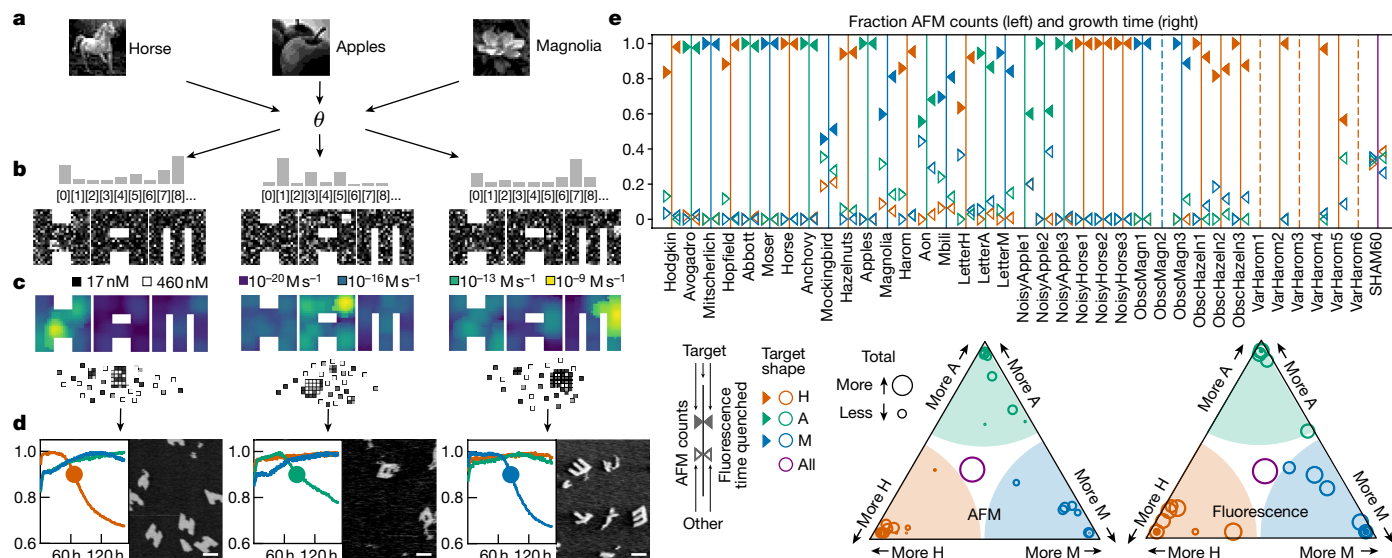
Fig. 6 | Pattern recognition results with a pre-existing multifarious system. **a**–**c**, All images (three shown) are converted (**a**) through a single pixel-to-tile map θ to vectors of tile concentrations (**b**), which are shown mapped onto tile locations in each shape, with nucleation rate predictions (**c**). **d**, Normalized fluorescence over time in hours (one label per shape; other label configurations shown in Extended Data Fig. 9) during a 150 hour temperature ramp from 48 to 45 °C, and final AFM images. Scale bars, 100 nm. **e**, Summary of results for both fluorescence and AFM for all 36 images, and a uniform 60 nM tile concentration control sample. Above, colours of vertical lines indicate the target shape for each pattern, whereas triangular markings of each colour indicate the relative fraction of growth time (on right) or fraction of shapes counted in AFM images (on left) for the corresponding shape (solid markings indicate target shape). Dashed lines indicate samples with no significant quenching or observed shapes. Below, ternary plots summarize the same results, with proximity to triangle corners indicating relative fractions of growth time (right) or counted shapes (left) and circle size indicating overall growth time (right) or total number of shapes (left).

multicomponent liquid condensates, multicomponent active matter and other systems might have similar programmable and learnable phase boundaries.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06890-z.

1. Hertz, J., Krogh, A. & Palmer, R. G. *Introduction to the Theory of Neural Computation* (CRC, 1991).
2. Dayan, P. & Abbott, L. F. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (MIT, 2005).
3. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT, 2016).
4. Rössler, O. E. A synthetic approach to exotic kinetics (with examples). In *Physics and Mathematics of the Nervous System* (eds Conrad, M., Güttinger, W. & Cin, M.) 546–582 (Springer, 1974).
5. Hjelmfelt, A., Weinberger, E. D. & Ross, J. Chemical implementation of neural networks and Turing machines. *Proc. Natl Acad. Sci. USA* **88**, 10983–10987 (1991).
6. Mjolsness, E., Sharp, D. H. & Reinitz, J. A connectionist model of development. *J. Theor. Biol.* **152**, 429–453 (1991).
7. Bray, D. Protein molecules as computational elements in living cells. *Nature* **376**, 307–312 (1995).
8. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
9. Fletcher, D. A. & Mullins, R. D. Cell mechanics and the cytoskeleton. *Nature* **463**, 485–492 (2010).
10. Holy, T. E. & Leibler, S. Dynamic instability of microtubules as an efficient way to search in space. *Proc. Natl Acad. Sci. USA* **91**, 5682–5685 (1994).
11. Lee, C.-Y. et al. *Coccidioides* endospores and spherules draw strong chemotactic, adhesive, and phagocytic responses by individual human neutrophils. *PLoS ONE* **10**, e0129522 (2015).
12. Floyd, C., Levine, H., Jarzynski, C. & Papoian, G. A. Understanding cytoskeletal avalanches using mechanical stability analysis. *Proc. Natl Acad. Sci. USA* **118**, e2110239118 (2021).
13. Seeman, N. C. & Sleiman, H. F. DNA nanotechnology. *Nat. Rev. Mater.* **3**, 17068 (2018).
14. Rothemund, P. W. K. & Winfree, E. The program-size complexity of self-assembled squares. In *Proc. Thirty-Second Annual ACM Symposium on Theory of Computing* (eds Yao, F. & Luks, E.) 459–468 (Association for Computing Machinery, 2000).
15. Rothemund, P. W. K., Papadakis, N. & Winfree, E. Algorithmic self-assembly of DNA Sierpinski triangles. *PLoS Biol.* **2**, e424 (2004).
16. Woods, D. et al. Diverse and robust molecular algorithms using reprogrammable DNA self-assembly. *Nature* **567**, 366 (2019).
17. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl Acad. Sci. USA* **79**, 2554–2558 (1982).
18. Qian, L., Winfree, E. & Bruck, J. Neural network computation with DNA strand displacement cascades. *Nature* **475**, 368–372 (2011).
19. Cherry, K. M. & Qian, L. Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks. *Nature* **559**, 370–376 (2018).
20. Okumura, S. et al. Nonlinear decision-making with enzymatic neural networks. *Nature* **610**, 496–501 (2022).
21. Rizik, L., Danial, L., Habib, M., Weiss, R. & Daniel, R. Synthetic neuromorphic computing in living cells. *Nat. Commun.* **13**, 5602 (2022).
22. Conrad, M. Self-assembly as a mechanism of molecular computing. In *Images of the Twenty-First Century. Proc. Annual International Engineering in Medicine and Biology Society* (eds Kim, Y. & Spelman, F. A.) 1354–1355 (IEEE, 1989).
23. Murugan, A., Zeravcic, Z., Brenner, M. P. & Leibler, S. Multifarious assembly mixtures: systems allowing retrieval of diverse stored structures. *Proc. Natl Acad. Sci. USA* **112**, 54–59 (2015).
24. Zhong, W., Schwab, D. J. & Murugan, A. Associative pattern recognition through macro-molecular self-assembly. *J. Stat. Phys.* **167**, 806–826 (2017).
25. Moser, E. I., Kropff, E. & Moser, M.-B. Place cells, grid cells, and the brain's spatial representation system. *Ann. Rev. Neurosci.* **31**, 69–89 (2008).
26. Frenkel, D. & Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications* (Academic, 2002).
27. Genot, A. J., Fujii, T. & Rondelez, Y. Computing with competition in biochemical networks. *Phys. Rev. Lett.* **109**, 208102 (2012).
28. Seeman, N. C. De novo design of sequences for nucleic acid structural engineering. *J. Biomol. Struct. Dyn.* **8**, 573–581 (1990).
29. Zadeh, J. N. et al. NUPACK: analysis and design of nucleic acid systems. *J. Comput. Chem.* **32**, 170–173 (2011).
30. Winfree, E., Liu, F., Wenzler, L. A. & Seeman, N. C. Design and self-assembly of two-dimensional DNA crystals. *Nature* **394**, 539–544 (1998).
31. Yin, P. et al. Programming DNA tube circumferences. *Science* **321**, 824–826 (2008).
32. Wei, B., Dai, M. & Yin, P. Complex shapes self-assembled from single-stranded DNA tiles. *Nature* **485**, 623–626 (2012).
33. Ke, Y., Ong, L. L., Shih, W. M. & Yin, P. Three-dimensional structures self-assembled from DNA bricks. *Science* **338**, 1177–1183 (2012).
34. Ong, L. L. et al. Programmable self-assembly of three-dimensional nanostructures from 10,000 unique components. *Nature* **552**, 72–77 (2017).
35. Winfree, E. & Bekbolatov, R. Proofreading tile sets: error correction for algorithmic self-assembly. In *DNA Computing (Lecture Notes in Computer Science)* Vol. 2943 (eds Chen, J. & Reif, J.) 126–144 (Springer, 2004).

36. Evans, C. G. & Winfree, E. Optimizing tile set size while preserving proofreading with a DNA self-assembly compiler. In *DNA Computing and Molecular Programming (Lecture Notes in Computer Science)* Vol. 11145 (eds Doty, D. & Dietz, H.) 37–54 (Springer, 2018).
37. Schulman, R. & Winfree, E. Programmable control of nucleation for algorithmic self-assembly. *SIAM J. Comput.* **39**, 1581–1616 (2009).
38. Schulman, R. & Winfree, E. Synthesis of crystals with a programmable kinetic barrier to nucleation. *Proc. Natl Acad. Sci. USA* **104**, 15236–15241 (2007).
39. Jacobs, W. M. & Frenkel, D. Self-assembly of structures with addressable complexity. *J. Am. Chem. Soc.* **138**, 2457–2467 (2016).
40. Sajfutdinow, M., Jacobs, W. M., Reinhardt, A., Schneider, C. & Smith, D. M. Direct observation and rational design of nucleation behavior in addressable self-assembly. *Proc. Natl Acad. Sci. USA* **115**, E5877–E5886 (2018).
41. Amit, D., Gutfreund, H. & Sompolinsky, H. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.* **55**, 1530–1533 (1985).
42. Battaglia, F. P. & Treves, A. Attractor neural networks storing multiple space representations: a model for hippocampal place fields. *Phys. Rev. E* **58**, 7738–7753 (1998).
43. Monasson, R. & Rosay, S. Transitions between spatial attractors in place-cell models. *Phys. Rev. Lett.* **115**, 098101 (2015).
44. Tanaka, G. et al. Recent advances in physical reservoir computing: a review. *Neural Netw.* **115**, 100–123 (2019).
45. Dunn, K. E. et al. Guiding the folding pathway of DNA origami. *Nature* **525**, 82–86 (2015).
46. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A phase separation model for transcriptional control. *Cell* **169**, 13–23 (2017).
47. Pirzer, T. & Simmel, F. C. Tiny robots made from biomolecules. *Europhys. News* **53**, 24–27 (2022).
48. Minev, D., Wintersinger, C. M., Ershova, A. & Shih, W. M. Robust nucleation control via crisscross polymerization of highly coordinated DNA slats. *Nat. Commun.* **12**, 1741 (2021).
49. Wintersinger, C. M. et al. Multi-micron crisscross structures grown from DNA-origami slats. *Nat. Nanotechnol.* **18**, 281–289 (2023).
50. Winfree, E. Chemical reaction networks and stochastic local search. In *DNA Computing and Molecular Programming (Lecture Notes in Computer Science)* Vol. 11648 (eds Thachuk, C. & Liu, Y.) 1–20 (Springer, 2019).
51. Wright, L. G. et al. Deep physical neural networks trained with backpropagation. *Nature* **601**, 549–555 (2022).
52. Lakin, M. R. & Stefanovic, D. Supervised learning in adaptive DNA strand displacement networks. *ACS Synth. Biol.* **5**, 885–897 (2016).
53. Hochberg, G. K. A. et al. A hydrophobic ratchet entrenches molecular complexes. *Nature* **588**, 503–508 (2020).
54. Anderson, P. W. More is different: broken symmetry and the nature of the hierarchical structure of science. *Science* **177**, 393–396 (1972).
55. Rothemund, P. W. K. Folding DNA to create nanoscale shapes and patterns. *Nature* **440**, 297–302 (2006).
56. Sartori, P. & Leibler, S. Lessons from equilibrium statistical physics regarding the assembly of protein complexes. *Proc. Natl Acad. Sci. USA* **117**, 114–120 (2020).

Article

Methods

Multifarious DNA tile system design

Previous theoretical proposals[23,24,57] for multifarious mixtures require each component to accept multiple strongly binding partners at each binding site. However, in DNA tile assembly, each binding site can usually only bind its Watson–Crick complement, not an arbitrary set of other domains. Hence, we used an alternate approach: we laid out three structures made of entirely unique, abstract tiles, designed a merging algorithm to reuse tiles in multiple locations if consequences for unintentional binding between other tiles was minimal, and then designed DNA sequences reflecting the resulting abstract layout of tiles.

The three target shapes were drawn on a 24 × 24 single-stranded tile (SST) molecular canvas[32], at an abstract level without sequences. Each location in each shape was initially a unique tile, with four abstract binding sites referred to as 'glues' in place of binding domains with sequences: after sequence design, 'matching' glues correspond to domains with complementary sequences. Edges of the shapes used a special 'null glue' with no valid binding partner. In total, this initial design had 2,706 glues and 1,456 tiles.

The three shapes were then processed through a 'merging' algorithm that attempted to reuse the same tiles in different shapes. Each step of the algorithm randomly chose two tiles in two different shapes, with null glues on the same sides of each tile, if any. It then considered a modified set where the two tiles were identical, by making them use the same four glues, and propagating the changes in the glues to all other places they occurred within all shapes, starting with the neighbouring tiles (for example, Extended Data Fig. 2c). Such a change could create undesired growth pathways, for example, allowing chimera of multiple shapes. Thus, the algorithm then checked the modified set for two criteria taken from algorithmic self-assembly (Extended Data Fig. 2a,b). The self-healing criterion requires that, for any correct subassembly of any shape, whereas attachments of the wrong tile for a particular location may take place by one bond, only the correct tile can attach by two or more bonds[58]. The second-order sensitivity criterion for proofreading requires that, for any correct subassembly of any shape, if an incorrect attachment by one bond takes place, the incorrectly attached tile will not create a neighbourhood where an additional incorrect tile can attach by two bonds, and thus the initial error will be likely to fall off[35,36]. If the modified set satisfied these two criteria, which are trivially satisfied when every tile and bond is unique to a particular location, then the merging algorithm accepted the modified set and continued to another step with a different pair of randomly chosen tiles. Thus, we ensured that there is at least a minimum barrier to continued incorrect growth in a regime where tile attachment by two or more bonds is favourable, and attachment by one bond is unfavourable, which is the case close to the melting temperature of most DNA tile assembly systems[59,60].

The algorithm repeatedly merged tiles that satisfied the two criteria until no further acceptable merges were possible. As each merge could affect the acceptability of later merges by changing the glues around each tile, to guide the algorithm towards a sequence of merges it was more likely to be compatible with, the algorithm was initially restricted to considering pairs of tiles from an alternating 'chequerboard' subset, which, apart from edges, were likely to be merge-able. After exhausting acceptable merges from this subset, the algorithm attempted merges using all tiles in the system. After repeating this stochastic algorithm multiple times, and selecting the system with the smallest number of tiles, the final resulting system had 698 binding domain and 917 tiles, with 371 of tiles shared between at least two shapes (Extended Data Fig. 2d).

After the assignment of abstract binding domains to each tile by the merging algorithm, the sequences for the binding domains, and thus tiles themselves, were generated using the sequence design software of Woods et al.[16]. Tiles used a standard SST motif, with alternating 10 and 11 nt binding domains, designed to have similar binding strengths

as predicted using a standard thermodynamic model[16,29,61]. Following Woods et al.[16], we set a target range of −8.9 to −9.2 kcal mol^{-1} for a single domain at 53 °C, which was between the melting temperature and growth temperature for their system. Null binding domains on the edges of shapes, not intended to bind to any other tiles, were assigned poly-T sequences.

Models of nucleation

To model the dependence of the nucleation rates of the three shapes on patterns of unequal concentration, we developed a simple nucleation model based on the stochastic generation of possible nucleation pathways and critical nuclei, which we call the Stochastic Greedy Model (SGM). The model estimates nucleation rates by analysing stochastic paths generated in a greedy manner by making single-tile additions starting from a particular monomer in the system. At each step, all favourable attachments are added and then an unfavourable attachment is performed with probability weighted by the relative free-energy differences of the available tile attachment positions. When multiple favourable attachments are available, the most favourable attachment is made deterministically. This procedure is repeated for many paths over all possible initial positions within the shape considered, and the barrier (highest free-energy state visited in 'growing' a full structure) is recorded for each path. A nucleation rate is estimated by assuming an equilibrium occupation of this barrier state (Arrhenius' approximation[26]) and summing over the kinetics of the available attachments from this state (see Extended Data Fig. 4 and Supplementary Information section 2.2 for a detailed discussion). The approximations here could be improved by running fully reversible simulations, for example, using xgrow and the kinetic Tile Assembly Model[59,62] augmented with Forward Flux Sampling[63].

Fluorophore labels and DNA synthesis

Sites for fluorophore and quencher modifications were chosen to avoid edges, modify only unshared tiles and provide a reasonable distribution of locations on each shape. Fluorophores were chosen for spectral compatibility and temperature stability[64]. ROX, ATTO550 and ATTO647N were paired with Iowa Black RQ, and FAM was paired with Iowa Black FQ. Both fluorophore and quencher modifications were made on the 5′ ends of tiles; to sufficiently colocalize fluorophores and quenchers, one tile in the label pair used a reversed orientation (Fig. 4a). Fluorophore labels are discussed in detail in Supplementary Information section 3.

Tiles without fluorophore or quencher modifications were ordered unpurified (desalted) and normalized to 400 μM in TE buffer (Integrated DNA Technologies). Tiles with fluorophore or quencher modifications were ordered purified by high-performance liquid chromatography (HPLC) and normalized to 100 μM. Given that unpurified synthetic oligonucleotides typically have less than 40 to 60% of the molecules being full length, it is remarkable (although consistent with Woods et al.[16]) that this did not prevent successful pattern recognition by nucleation.

Experimental overview

The basic workflow for the main experiments was as follows: for a chosen set of concentration patterns (flag or image), samples were prepared on a 96-well plate using an acoustic liquid handler to mix strand stocks in the necessary proportions; vortexed, spun and transferred to PCR tubes for the days-long anneal in the quantitative PCR (qPCR) machine; then samples were deposited on mica for AFM imaging. Fluorescence from the qPCR machine and AFM images were subsequently analysed.

Mixing and growth

Individual tiles were mixed, in the concentration patterns used for experiments, using an Echo 525 acoustic liquid handler (Beckman Coulter). Samples used TEMg buffer (TE buffer with 12.5 mM MgCl$_2$)

in a total volume of roughly 20 µl. Flag experiments used a 50 nM base concentration of unenhanced tiles and an 880 nM concentration of enhanced concentration tiles, whereas pattern recognition experiments used tiles with nominal concentrations between 16.67 and 450 nM, which were then quantized into ten discrete values to simplify mixing and conserve material (Supplementary Information section 2.8).

For each concentration pattern in the flag experiments and pattern recognition of trained images, four samples were prepared, each with the same concentration pattern of tiles, but with tiles in different locations replaced by their fluorophore–quencher-modified alternates: one sample for each shape with tiles for all four fluorophore labels on only that shape, to monitor growth of multiple regions on each shape, and an additional sample with one fluorophore on each shape: ROX, ATTO550 ('five') and ATTO647N ('six') on H, A and M structures, respectively. To reduce the total number of samples, only the lattermost sample type was prepared for pattern recognition of test images. Fluorophore and quencher-modified tile locations always had tiles mixed at the lowest concentration used in the experiment.

After transferring samples to PCR tubes, samples were grown in an mx3005p qPCR machine (Agilent), to provide a program of controlled temperature over time while monitoring fluorescence. Growth protocols began with a ramp from 71 to 53 °C over 40 min to ensure any potentially pre-existing complexes were melted, and then a slower ramp from 53 °C to an initial growth temperature at 1 °C h^{-1}. At this point, three different protocols were used. For constant temperature flag growth experiments, the growth temperature was 47 °C and this was held for 51 h. For temperature ramp flag growth, the initial growth temperature was 48 °C, which was reduced over the course of 100 h to 46 °C. For pattern recognition, a ramp from 48 to 45 °C over 150 h was used. For constant temperature experiments, fluorescence readings were taken every 12 min and for other experiments, every 30 min. After the growth period, temperature was lowered to 39 at 1 °C per 26 min. See Supplementary Information sections 5 and 6 for temperature protocols plotted as a function of time. The experimental timescales and temperatures were chosen not to test the potential speed of selective nucleation, but rather to provide robustness to unknown nucleation temperatures and to convincingly show that nucleation of incorrect structures is limited over long timescales. Thus, on-target nucleation often took place during a comparatively short time and temperature in the experiment, with the remaining time spent either above the expected nucleation temperature or waiting to observe potential off-target nucleation. We also did not try to optimize the system's speed: the WTA mechanism suggests that significantly faster timescales are possible, and smaller assemblies would reduce the time needed for growth after nucleation. Because of the small sample size and long experiment duration, great care to avoid evaporation was necessary. Once protocols were finished, samples were stored at room temperature until ready for AFM imaging.

Imaging

AFM imaging was performed using a FastScan AFM (Bruker) in fluid tapping mode directly after annealing was completed. In contrast to previous studies[32–34] in which uniquely addressed SST shapes were gel purified before imaging, we did not do so here, thus we were able to observe assembly intermediates. To achieve better images, two techniques were combined: sample warming to prevent non-specific clumping of structures, and washing with Na-supplemented buffer to prevent smaller material, such as unbound, single DNA tile strands, from adhering to the mica surface. Each sample was diluted 50 times into TEMg buffer with an added 100 mM NaCl, then warmed to roughly 40 °C for 15 min. Next, 50 µl of the sample mix was deposited on freshly cleaved mica, then left for 2 min. As much liquid as possible was pipetted off the mica and discarded, then immediately replaced with Na-supplemented buffer again and mixed by pipetting up and down.

This washing process of buffer removal and addition was repeated twice with added-Na buffer, then once with TEMg buffer to remove remaining Na, before imaging was performed in TEMg buffer. As adhesion of DNA to mica is dependent on the ratio of monovalent and divalent cations in the imaging buffer, this process was meant to ensure that unbound tiles were removed during the washing process where Na and Mg were present, whereas imaging itself took place with only Mg so that the lattice structures would be more strongly adhered to the surface resulting in better image quality.

Fluorescence and AFM data analysis

Fluorophore signals are known to be affected by extraneous factors such as temperature, pH, secondary structure and the local base sequence near the fluorophore[64], which complicates quantitative interpretation of absolute fluorescence levels. Our own control experiments also illustrated effects due to partial assembly intermediates as well as due to the total amount of single-stranded DNA in solution (Supplementary Information section 3). For this reason, the fluorescence of each fluorophore was normalized to the maximum raw fluorescence value of that fluorophore in that particular sample, and the time at which the fluorescence signal decreased by 10% was then used as a measure of the extent of nucleation that appears less sensitive to these artefacts (Extended Data Fig. 5). The duration between the point of 10% quenching and the end of the growth segment of the experiment was defined as the 'growth time' for that fluorophore label; the growth time was defined as 0 in the event of quenching never reaching 10%. For concentration patterns with four samples with different fluorophore arrangements, the total growth time of a shape was defined as the average of the growth time of the five total fluorophore labels on the shape across the four samples (four in the shape-specific sample and one in the each-shape sample), whereas for concentration patterns with only one sample, the growth time of the corresponding fluorophore label was used. As the position of the fluorophore within the shape, relative to where nucleation occurs, has a substantial influence on growth time measurements, the considerable variability in these measurements relative to the true nucleation kinetics must be acknowledged.

For flag experiments, AFM imaging was done only for qualitative confirmation of the selective nucleation and growth indicated by fluorescence results. For pattern recognition and equal-concentration experiments, however, shapes in AFM images were uniformly quantified. At least one sample of each of the patterns had three 5 × 5 µm images taken under comparable conditions. The sample corresponding with each image was blinded, and structures were counted independently by each of the four authors, classifying structures as either 'nearly complete' or 'clearly identifiable' examples of each of the three shapes. For the purposes of analysing pattern-dependent nucleation and growth, no clear distinction between the number of nearly complete and clearly identifiable shapes was found, and so the two categories were summed. Counts were averaged across the three images, then averaged across the counts of the four authors, to obtain a count per shape per 25 µm^2 region for each pattern. Each author used their own, subjective, interpretation of 'nearly complete' and 'clearly identifiable' structures, and the total number of structures counted in each image differed by up to ±50% for different authors. However, the ratios of different shapes in each image counted by each author remained within 5% of the mean ratios for most images, and across all images no author had a bias of more than ±4% towards identifying a particular shape more or less often than average. Results are detailed in Supplementary Information section 6.3.

To measure the selectivity of patterns, the fraction of on-target shape growth time and AFM counts, compared to the sum of shape growth times and AFM counts, was used. The total growth times, and total AFM counts, of the on-target shapes were used to measure overall shape growth.

Article

Pattern recognition training

Images for pattern recognition were adapted from several sources (Fig. 5d). Each image was rescaled to 30×30, discretized to ten grayscale values and adjusted so that the number of pixels with each value was consistent across all images. Each pixel's grayscale value, $0 \leq p_n \leq 1$, was converted to the concentration c_i for the corresponding tile t_i where $i = \theta(n)$ using an exponential formula, $c_i = c e^{3 p_n \ln 3}$, where the base concentration is $c = 16.67$ nM. The intention of the numbers used was to make the average tile concentration 60 nM for each image. As each image had 900 pixels and there are 917 tiles in the system, 17 tiles did not have their concentrations set by any pixel; these tile concentrations were uniformly set to the lowest concentration, and the assignment of these tiles was used to ensure that fluorophore label locations did not vary in concentration.

The tile-pixel assignment was optimized through a simple hill-climbing algorithm, starting from a random assignment, where random modifications to the assignment map are attempted at each step and accepted if the move increases the efficacy of the map. This efficacy was quantified through a heuristic function that accounts for relative nucleation rates, location of nucleation sites (with preference given to locations that succeeded in the flag experiments shown in Fig. 4d) and satisfaction of constraints related to the fluorescent reporters. Because the nucleation algorithm described above, the SGM, is computationally expensive, a simplistic model of nucleation we call the Window Nucleation Model (WNM) was used to evaluate relative nucleation rates for most of the optimization steps. The WNM is based on the Boltzmann-weighted sum of concentrations over a $k \times k$ window swept over each structure, similar to the model used in Zhong et al.[24]. The more detailed but computationally costly SGM was then used for an additional several hours in hopes of improving the mapping. The WNM, along with all constraints about nucleation location and fluorescent reporters, was also used to explore the capacity of this map-training procedure in Extended Data Fig. 8. Details of the pattern recognition training and the window-based nucleation model are discussed in Supplementary Information sections 2.4 and 2.5.

Data availability

AFM images, fluorescence trajectories, DNA sequences and simulation results are available at https://www.dna.caltech.edu/Supplementary-Material/MultifariousSST/.

Code availability

Algorithms for tile set design, sequence design, nucleation rate prediction and pixel-to-tile map optimization are available at https://www.dna.caltech.edu/SupplementaryMaterial/MultifariousSST/.

57. Bupathy, A., Frenkel, D., & Sastry, S. Temperature protocols to guide selective self-assembly of competing structures. *Proc. Natl Acad. Sci. USA* **119**, e2119315119 (2022).
58. Winfree, E. in *Nanotechnology: Science and Computation* (eds Junghuei, C. et al.) 55–78 (Springer, 2006).
59. Winfree, E. *Simulations of Computing by Self-Assembly* Technical Report CaltechCSTR:1998.22 (California Institute of Technology, 1998).
60. Evans, C. G. & Winfree, E. Physical principles for DNA tile self-assembly. *Chem. Soc. Rev.* **46**, 3808–3829 (2017).
61. SantaLucia, J. & Hicks, D. The thermodynamics of DNA structural motifs. *Ann. Rev. Biophys. Biomol. Struct.* **33**, 415–440 (2004).
62. Evans, C. G., Schulman, R. & Winfree, E. The xgrow simulator. *GitHub* https://github.com/DNA-and-Natural-Algorithms-Group/xgrow.
63. Allen, R. J., Warren, P. B. & Ten Wolde, P. R. Sampling rare switching events in biochemical networks. *Phys. Rev. Lett.* **94**, 018104 (2005).
64. You, Y., Tataurov, A. V. & Owczarzy, R. Measuring thermodynamic details of DNA hybridization using fluorescence. *Biopolymers* **95**, 472–486 (2011).
65. Weibrecht, I. et al. Proximity ligation assays: a recent addition to the proteomics toolbox. *Expert Rev. Proteomics* **7**, 401–409 (2010).
66. Schaus, T. E., Woo, S., Xuan, F., Chen, X. & Yin, P. A DNA nanoscope via auto-cycling proximity recording. *Nat. Commun.* **8**, 696 (2017).
67. Hopfield, J. J. Neurodynamics of mental exploration. *Proc. Natl Acad. Sci. USA* **107**, 1648–1653 (2010).

Author contributions C.G.E., E.W. and A.M. conceived the study. C.G.E. and E.W. designed the molecules. C.G.E., J.O.B., E.W. and A.M. wrote simulation code, designed the experiments and performed the experiments, analysed the data and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06890-z.
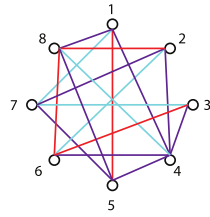**Correspondence and requests for materials** should be addressed to Constantine Glen Evans, Jackson O'Brien, Erik Winfree or Arvind Murugan.
Peer review information *Nature* thanks Friedrich Simmel and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
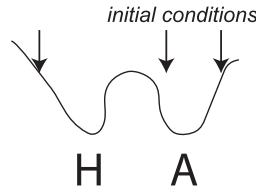**Reprints and permissions information** is available at http://www.nature.com/reprints.

Hopfield associative memory

Multifarious self-assembly

Wiring diagram

Memory retrieval

initial conditions

H   A

concentration patterns

monomers

H   A

Energy function

$$E(x) = -\frac{1}{2}\sum_{i,j} w_{i,j} x_i x_j - \sum_i b_i x_i$$

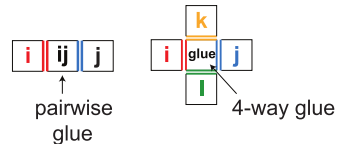$$G(A) = -\frac{1}{2}\sum_{p,p'}\sum_{i,j} J_{i,j}^{\delta(p,p')} x_p^i x_{p'}^j - \sum_p \sum_i \Theta_i x_p^i$$

Hebbian rule

neuron 1
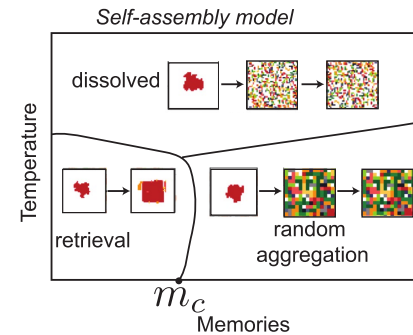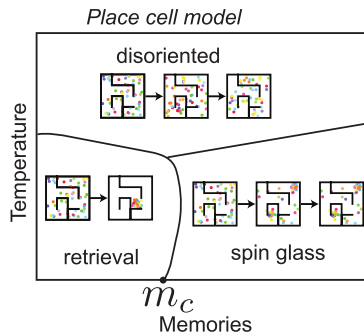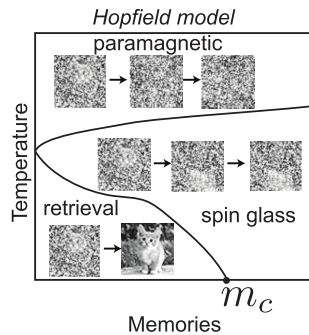neuron 2
neuron 3

* *   * time

"fire together, wire together"

* make synapse ij stronger

H ⋰⋰ A ⋰⋰ M ⋰⋰

*   *   * time
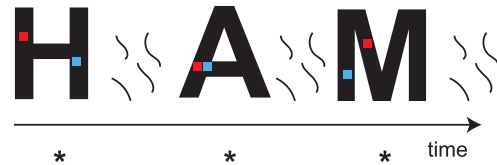
"get together, glue together"

| i | ij | j |

pairwise glue

k
| i | glue | j |
l
4-way glue

* make more glue ij

Phase diagram

Hopfield model
paramagnetic

Temperature

retrieval   spin glass

m_c
Memories

Place cell model
disoriented

Temperature

retrieval   spin glass

m_c
Memories

Self-assembly model

dissolved

Temperature

retrieval   random aggregation

m_c
Memories

Capacity

$$m_c \sim N$$

$$m_c \sim N^{1-1/d}$$
(d-dim place cell networks)

$$m_c \sim N^{1-1/d}$$
(d-dim self-assembly)

Extended Data Fig. 1 | Parallels and differences between neural network models and self-assembly models as exemplars of collective behaviour. In this rough metaphor, a neuron corresponds to a tile. While Hopfield networks allow full connectivity, multifarious self-assembly (like place cell networks) restricts connectivity to a superposition of grids with different unit permutations. The state of a Hopfield network consists of the set of active neurons, while the state of an assembly consists of the set of tiles present and their arrangement, which is restricted to be connected. We use $x_i \in \{-1, +1\}$ for the activity of neuron i, and $x_p^i \in \{0, 1\}$ for the occupancy of tile i in position p. The energy of a state is a quadratic function governed by synaptic weights w_{ij} and biases b_i for neural activities, or for assemblies, by directional binding energies $J_{i,j}^\delta$ for tiles i and j at positions p and p' that are neighbors in direction δ, along with (inverted) tile chemical potentials Θ_i. An environment presents a sequence of outside influences driving system state, either stimulating neural activity or spatially organizing tiles. Learning in Hopfield networks occurs any time neurons are simultaneously active. For self-assembly, learning an interaction requires tiles i,j to be located next to each other; we envision a hypothetical proximity-based ligation process[65,66] that creates interaction mediating glues ij for molecules i,j that spend time together in spatial proximity. Qualitative system behaviors depend on the number of memories being stored and the operating temperature, including phases where system state randomizes (paramagnetic/disoriented/dissolved), gets locked in a spurious local minimum (spin glass/random aggregation), or successfully retrieves learned memories. Due in large part to the restrictions on connectivity, the capacity of place cell networks and multifarious self-assembly is less than for the Hopfield model. See Supplementary Information section 1 for details and discussion.
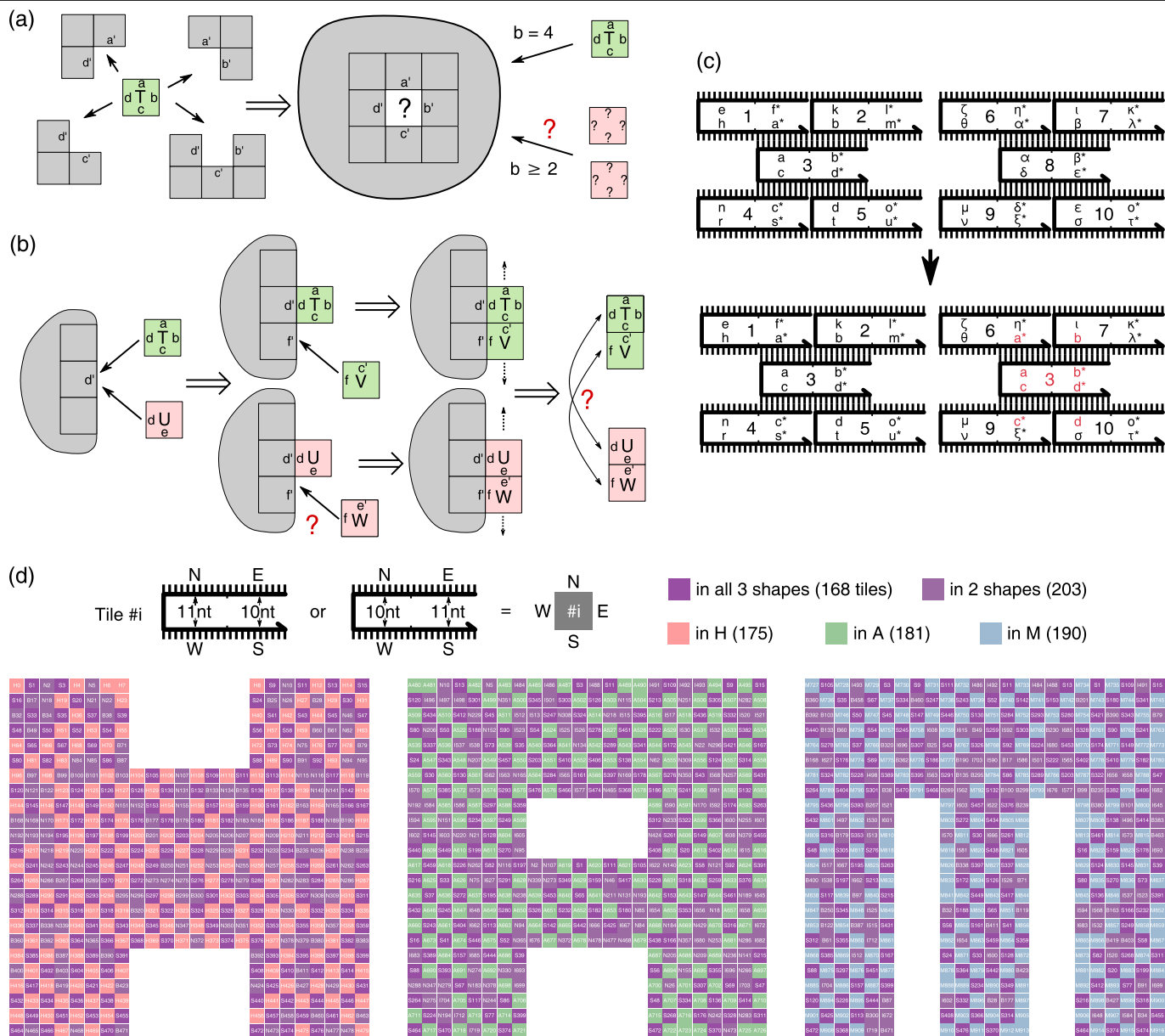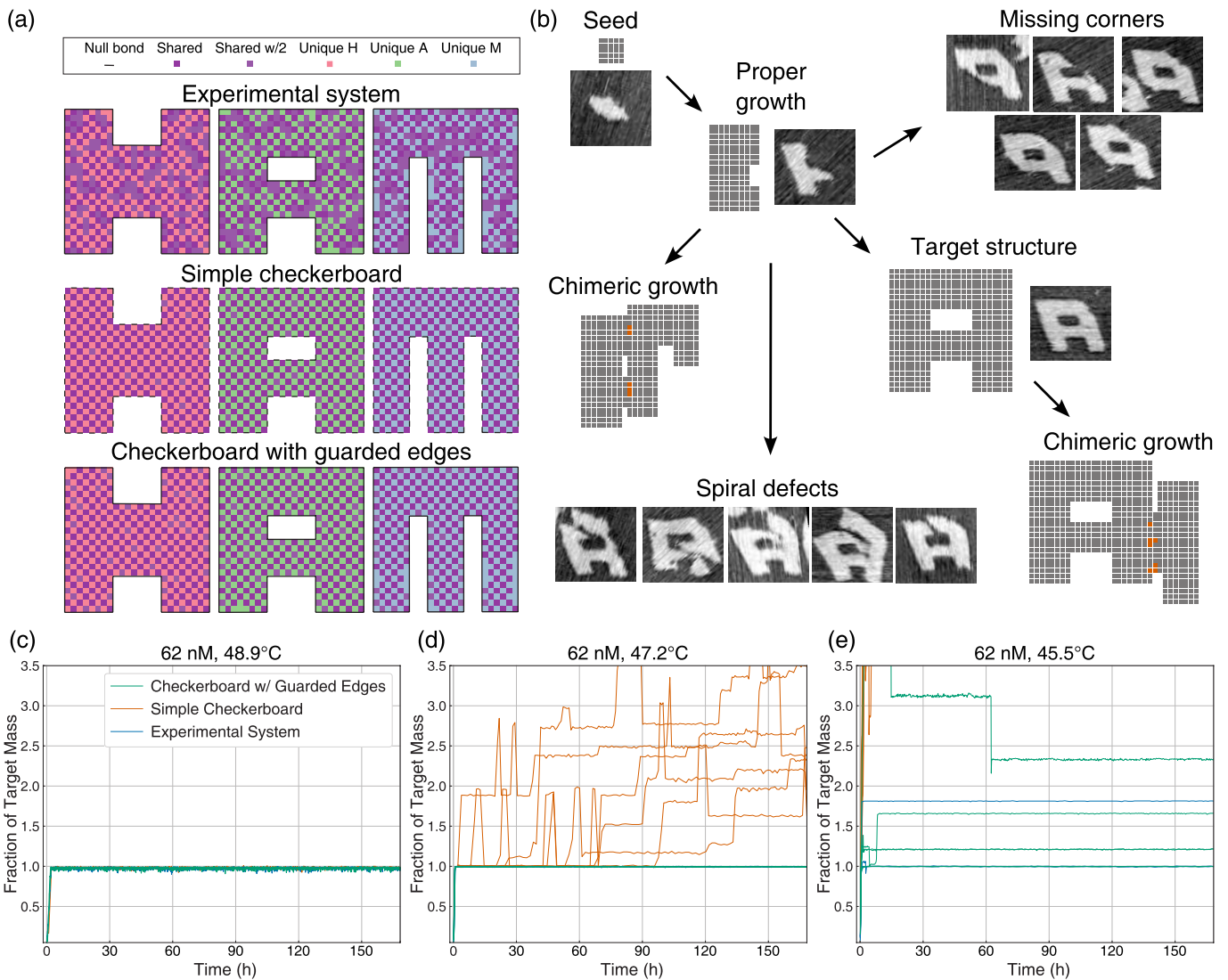
Extended Data Fig. 2 | Proofreading tile set design and tile assignment map. Extensive promiscuous interactions present in the SHAM mix could in principle lead to unintended chimeric structures and other malformed assemblies. To reduce or prevent such behaviors, our design incorporates self-assembly proofreading principles, so called because they enhance quick rejection of mis-assembled tiles. Much like with neural networks[67], random arrangement of tiles (such as the initial checkerboard layout in the first stage of our design process) provides a statistical proofreading[23] in the sense that problematic interactions are unlikely to arise. Further optimization of the tile set (in our second stage) ensures that two types of problematic interactions do not occur, thereby conferring algorithmic proofreading[35] and self-healing properties[58]. This tile set optimization is derived from prior work[36]. **a,** Our systems are designed to grow in a regime where a tile attaching by at least two bonds is favorable, but a tile attaching by one bond is not ('threshold 2'). Motivated by self-healing tile systems[58], we seek a tile set where no correct partial assembly should ever allow an undesired tile to attach by two or more bonds, though undesired attachments by one bond are allowed, such that any favorable attachment to a partial assembly will be correct. **b,** In addition to tiles attaching favourably by 2 bonds to growing facets, new facets in the system will only be created by tiles attaching unfavourably by one bond, and then being stabilized by further, favorable growth. At a site where tile T would correctly attach by one

bond, a tile U might be able to attach incorrectly by the same bond. T would correctly be stabilized by the subsequent attachment of V by two bonds, but U might be stabilized as well if there is a tile W that can attach to it and shares the same glue as V. Thus, if for every pair of tiles that can bind to each other (e.g., $T + V$), there is no other pair of binding tiles (e.g., $U + W$) that share two glues on the same edges of the tiles, then any tile that attaches by one bond to an assembly will either be the correct tile, or will not allow a subsequent stable attachment, and will likely detach quickly. This is equivalent to 'second-order sensitivity' with all directions treated as inputs, functioning as a form of self-assembly proofreading[35,36]. **c,** We created a multifarious tile system by first starting with three shapes constructed entirely of unique tiles, then repeatedly attempting to 'merge' tiles in different shapes by constraining the sequences of their domains to be identical, and checking whether each merge of two tiles results in a tile system that does not have any tile pairs violating criteria in **a** and **b**. **d,** From multiple trials of the merging process, each initially favoring a checkerboard arrangement before attempting more general merges, we selected the smallest result containing 917 tiles. DNA sequences for tiles in the system were designed with the single-stranded tile (SST) motif[31], with two alternating tiles motifs of 10 nt and 11 nt domains (full shape layouts and tile sequences are shown in Supplementary Information sections 3.3 and 4.1).

Extended Data Fig. 3 | Suppression of chimeric growth through tile set design. a, We use simulations to contrast assembly errors in three distinct tile sets: the proofreading tile set with an inert boundary used in experiments, described in Fig. 2(a, top); a simple checkerboard tile set with a strictly alternating shared and unique tile pattern for each shape, where unique tiles can be seen as mediating different interactions between shared tiles (a, middle); and an edge-guarded checkerboard in which we additionally enforce inert bonds around each shape's perimeter (a, bottom). For each tile set, we performed kinetic growth simulations, starting from a pre-formed 5 × 5 seed taken from a location within H. Simulations were performed using the kinetic Tile Assembly Model as implemented by xgrow (with chunk fission)[62] with uniform tile concentrations corresponding to 62 nM and parameters estimated in Supplementary Information section 2.1. b, Schematic illustrates various desired and undesired growth pathways for A, along with representative AFM images taken from the A flag 1 experiment (Supplementary Information section 5.3.13). Two distinct kinds of chimeric structures were seen in simulation as the result of promiscuous interactions: chimeric structures can grow either before full assembly of the target structure (e.g., part-A, part-M) or emerge spontaneously from the edge of a properly formed structure (e.g. full-A, part-H). Chimeras like those illustrated along the lower path are held together by just a few bonds and sometimes can quickly break apart (tiles with unintended bonds are shown in red); these result in sharp drops in simulated assembly size, as the simulation discards one subassembly when disconnected. Note that chimeric growth was not observed experimentally, possibly as a result of effective experimental system design; however, many observed structures failed to complete the upper right and/or lower left corners, or appeared to have suffered a spiral growth defect. A possible explanation for the missing corners, which is also seen in H and M, is supported by coarse-grained molecular dynamics simulations of SST lattice curvature (Supplementary Information section 3.4). Spiral defects were not seen in H or M and are presumably due to the interior hole in A. c–e, The size of the assembly (in units of the size of the fully formed H) is shown as a function of time. For higher temperature 48.9 °C (c), no chimeras are observed on the simulated timescales for any tile set. For intermediate temperature 47.2 °C (d), all 6 checkerboard trajectories still result in chimeras, while no errors are observed on the timescale probed for the guarded checkerboard or experimentally-implemented proofreading tile set. For lower temperature 45.5 °C (e), chimeras are seen in all runs for checkerboard structures (red traces), 4 of the 6 runs for guarded checkerboard structures (green traces) and 1 of the 6 runs for proofreading structures.
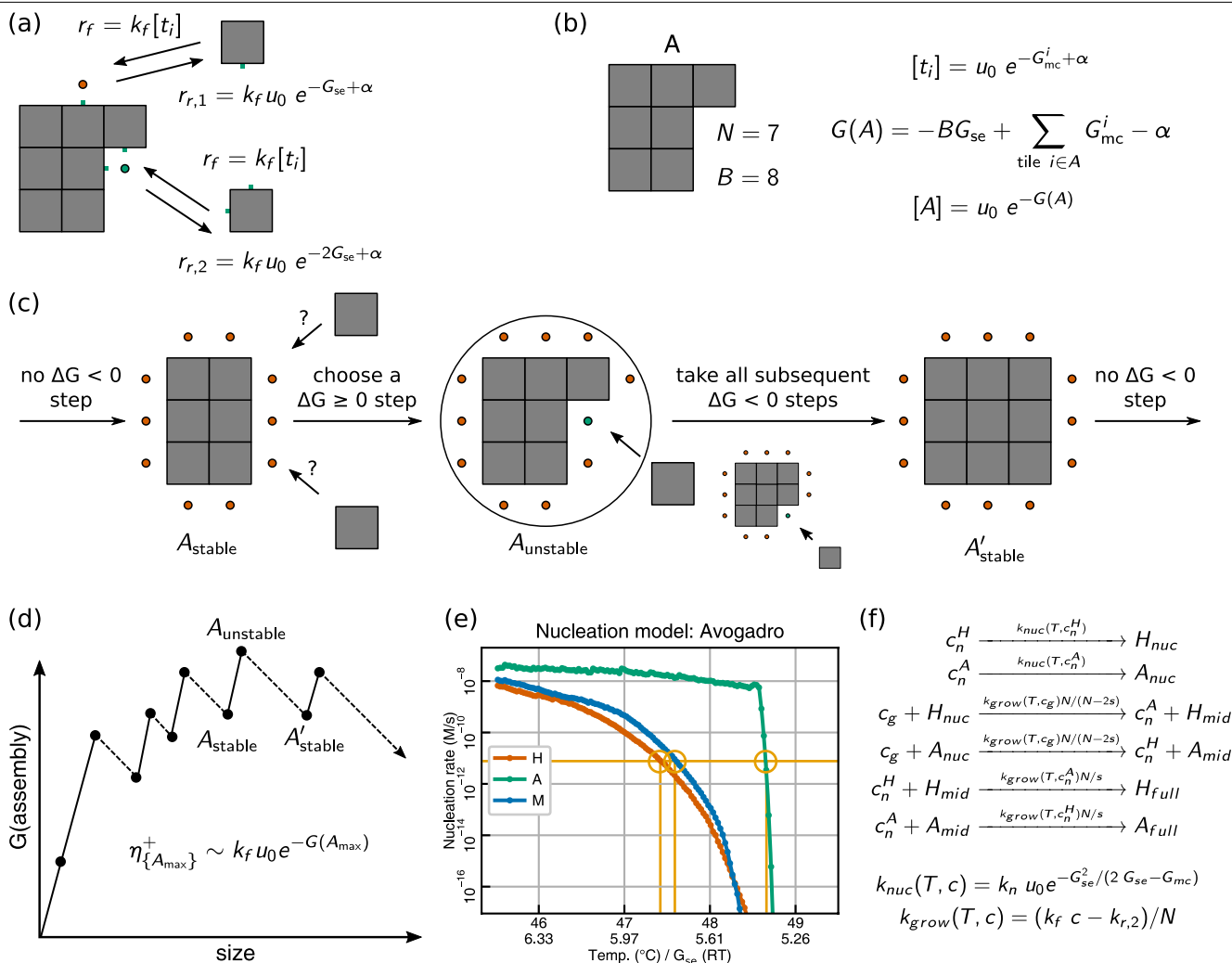
Extended Data Fig. 4 | Stochastic Greedy Model of nucleation, based on repeated stochastic simulations. a, The frequently-used kinetic Tile Assembly Model (kTAM)[59,60] has rates for tile attachment and detachment events based on tile and assembly diffusion and total binding strength of correct attachments a tile can make at a lattice site. Here $u_0 = 1$ M. **b**, These rates can be used to derive a free energy for any tile assembly in a system, and, assuming fixed monomer concentrations, an equilibrium concentration for any assembly. Schulman & Winfree[37] showed that the equilibrium concentration of the highest-energy assembly along a nucleation trajectory under this assumption provides an upper bound for nucleation rate through that trajectory, with or without fixed monomer concentrations. However, in a large system, considering all possible intermediate assemblies and all pathways, including many that are extremely unlikely, would be infeasible. Thus, we developed the Stochastic Greedy Model (SGM) to generate stochastically-chosen paths of tile attachments. **c**, Starting from a single tile (chosen with probability proportional to relative concentration), whenever the assembly is in a state A_{stable} where there is no tile attachment that would be favorable (have $\Delta G < 0$), one of the possible unfavorable (with $\Delta G \geq 0$) attachments is stochastically chosen, resulting in a higher-G state $A_{unstable}$. Then, all subsequent possible $\Delta G < 0$ attachments are made, resulting in the next A'_{stable} state; for our system of unique tiles for each site in the lattice, this sequence of favorable steps has a unique resulting assembly. **d**, The process repeats until all tiles in a shape are attached, which results in a trajectory with a maximum-G assembly that can be used to bound the rate of nucleation, η, through that particular trajectory. **e**, By using this process to collect many trajectories, and then repeating the entire process for each of the three shapes in the system, we can estimate nucleation rates dependent upon temperature, with the assumption that tile monomer concentrations do not deplete, and that the trajectories found are a reasonable representation of likely trajectories. For comparison between model predictions and experimental data in Extended Data Figs. 6d and 9b, we determined the temperature at which the model predicted the nucleation rate exceeded a threshold (orange line), to compare with when fluorescence quenching exceeded a threshold. For details on the SGM model, see Supplementary Information section 2.2. **f**, To study the winner-take-all effect, we use a simplified chemical reaction network (CRN) model for the case of systems with shared tiles (shown here) and a similar model for systems without shared tiles (described in Supplementary Information section 2.3). Here, c_n^H represent tiles in the flag area of shape H, which have initially higher concentrations; c_n^A are the corresponding tiles in the flag area of shape A, which have normal concentrations; and c_g represent tiles involved in growth from the nucleated seed H_{nuc} to the almost-complete structure H_{mid}; and similarly for structure A. A more detailed model based on (but simpler than) the SGM gives qualitatively similar results, as detailed in Supplementary Information section 2.3.
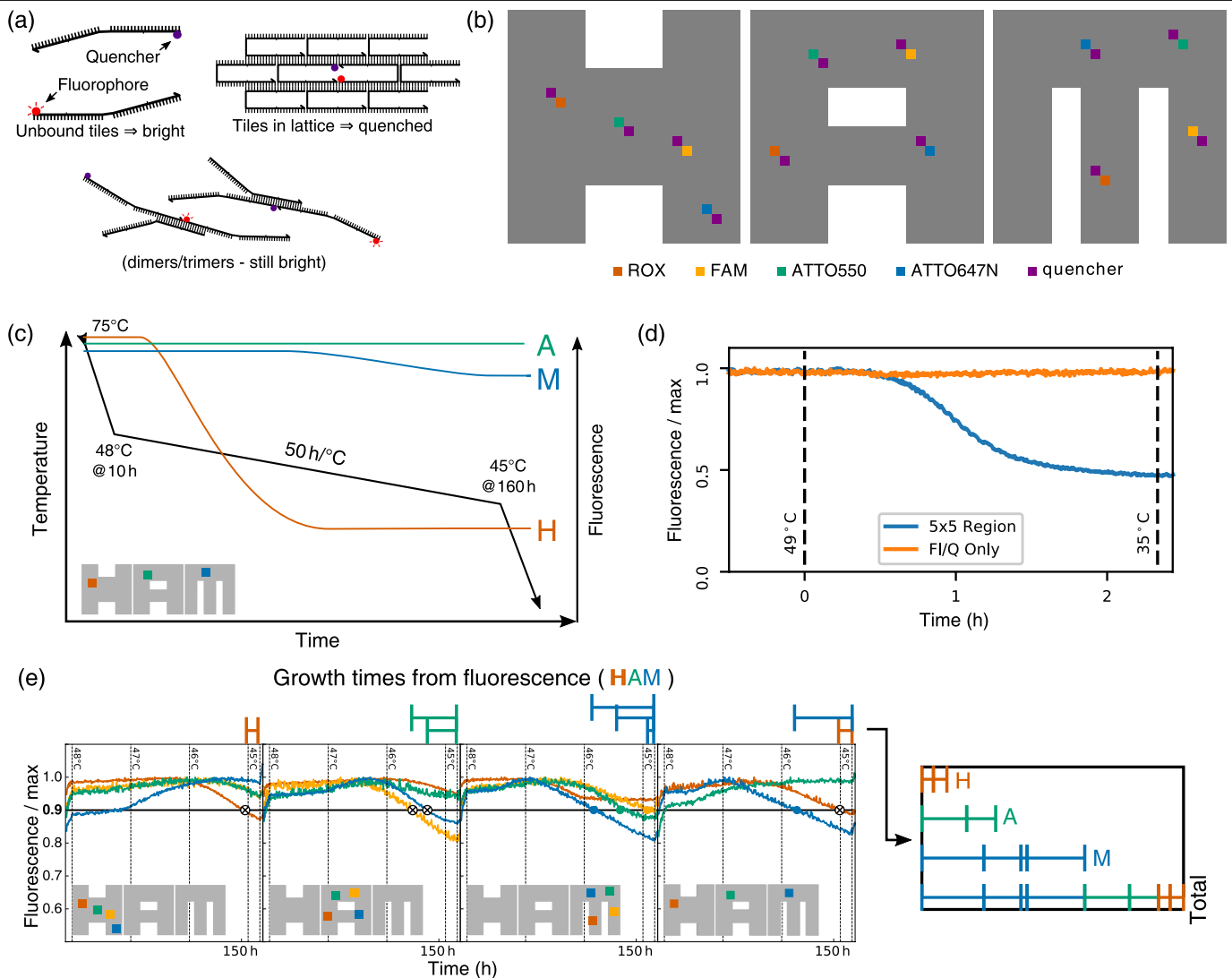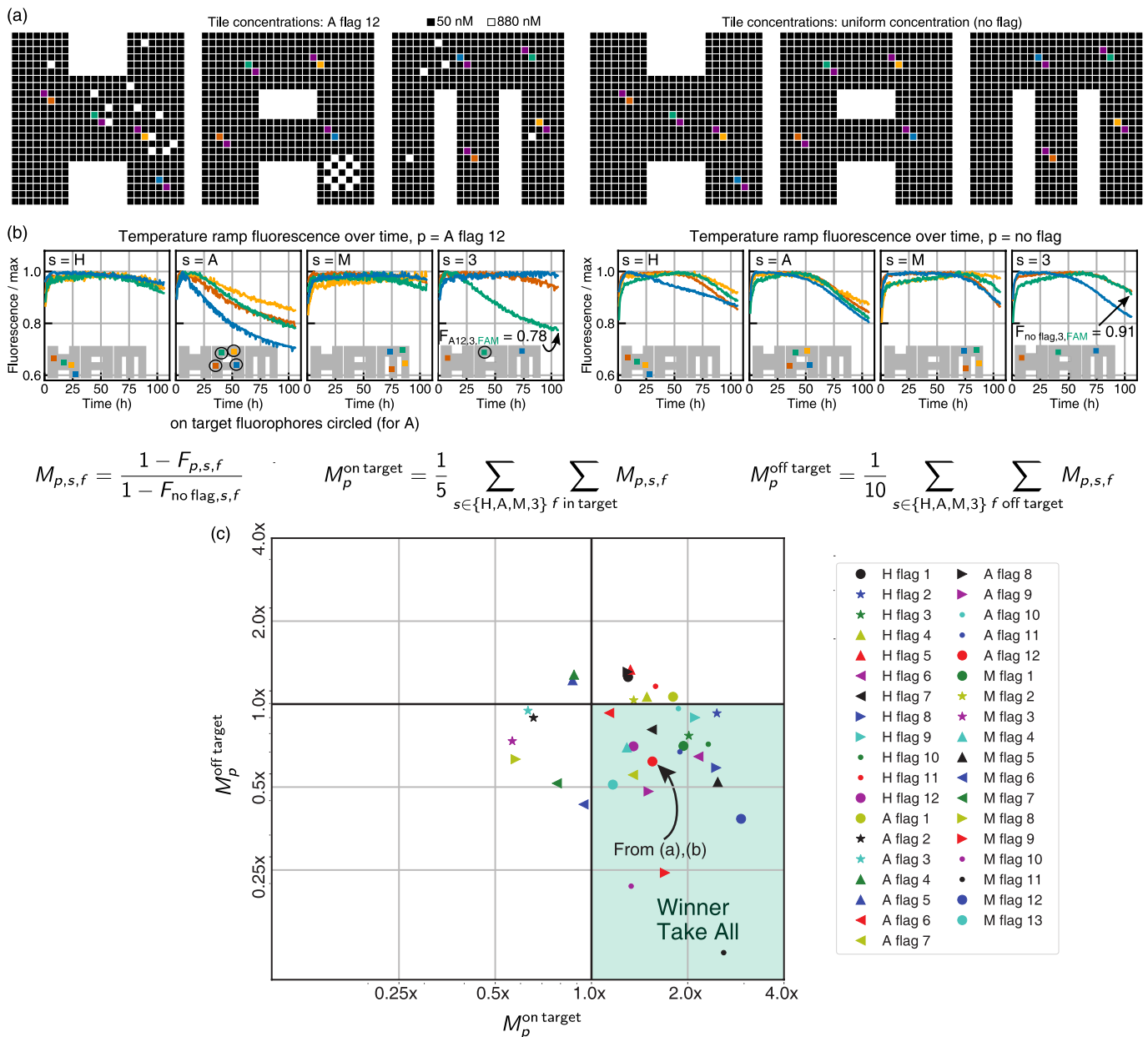
Extended Data Fig. 5 | Fluorophore quenching as a measure of nucleation and growth. a, Fluorescent labels used a fluorophore-quencher pair placed on the 5′ ends of two modified tiles unique to one shape, where they were colocated, but had no complementary binding domains, ensuring that dimers could not form, and trimers would not closely colocate the fluorophore and quencher. To constrain the pair to be close enough to quench in a well-formed lattice, one of the two tiles had its orientation and crossover position swapped compared to the unmodified tile for the location. **b**, Positions and types of all fluorophore/quencher pairs available for use. For one sample, one position for each of four types of fluorophores could be chosen, and tile pairs for those locations replaced by their modified counterparts. Thus different samples could probe different arrangements of up to four locations; four arrangements were used in experiments (e.g., in **e**). **c**, Expected behavior of fluorophore labels on shapes as one shape nucleates and grows. **d**, Fluorescence data for non-quenching (fluorophore tile only, orange) and quenching (5 × 5 lattice around fluorophore

and quencher tiles, blue) controls for the ATTO647N fluorophore/quencher pair on A. Here, the temperature ramps linearly from 49 °C to 35 °C at a rate of 0.1 °C/min, with all tiles at 50 nM, and each sample has its fluorescence normalized to its maximum value independently. **e**, An example of fluorescence growth time measurements (Mockingbird; see Supplementary Information section 6.4.9). Each fluorophore signal, in each sample, is independently normalized to its maximum value during the experiment, and the time between the point where the signal goes below 0.9 ('10% quenching') and the end of the experiment is measured ('growth time'). These times are then summed for all fluorophores, in all four samples, on each shape, resulting in a growth time for each shape, and, when normalized to the sum of all growth times, a relative growth time for each shape. See Methods and Supplementary Information section 3 for design and characterization of the fluorescence readout method, as well as an estimate of the melting temperature of H.
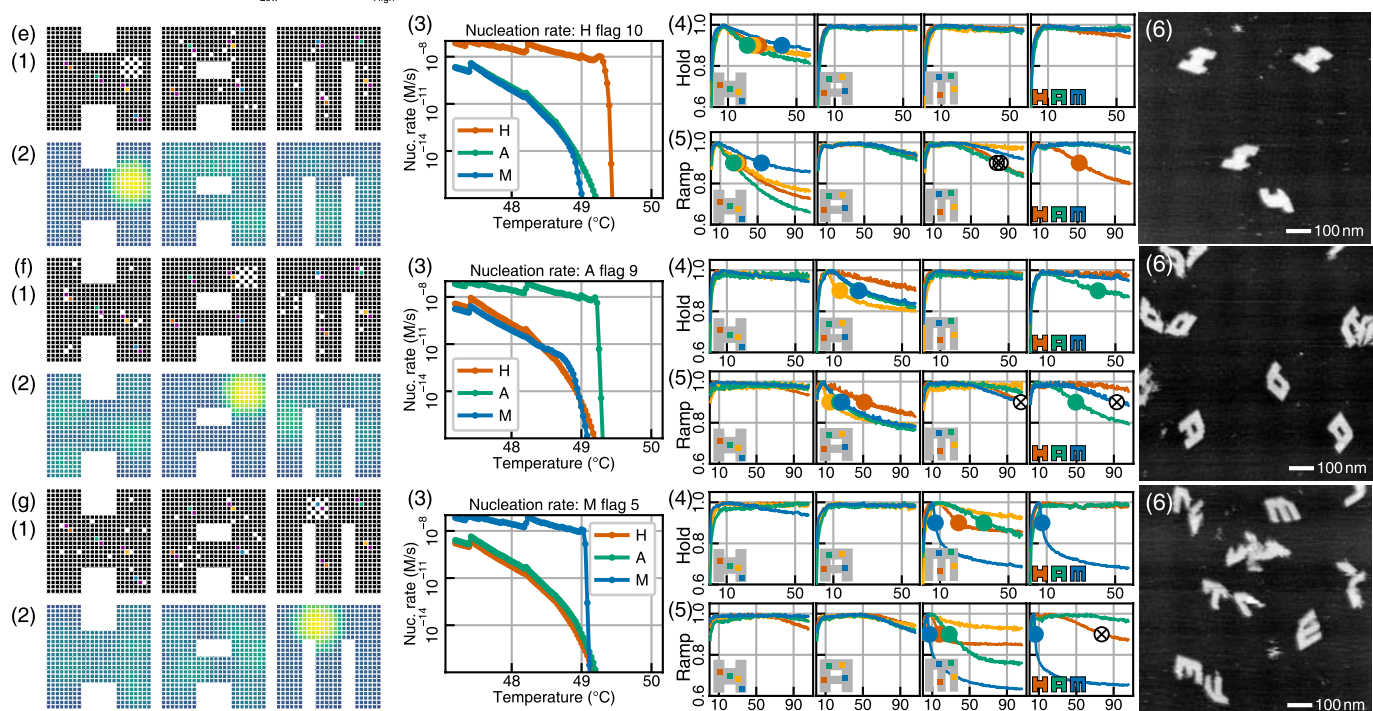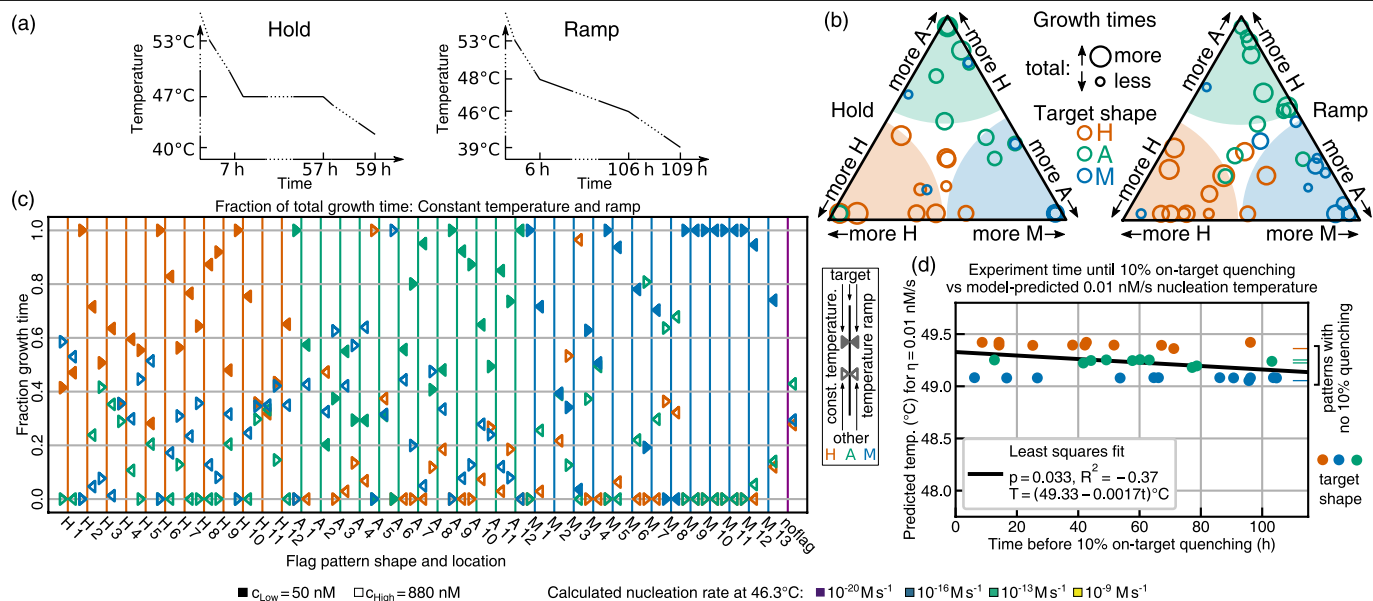
Extended Data Fig. 6 | Nucleation and growth with 'flag' patterns of enhanced concentration. a–c, 37 different concentration patterns with enhanced concentrations of shared tiles in 5 × 5 regions were prepared, each with four different standard sets of fluorophores in four samples, and grown using two temperature protocols (**a**): a ramp focusing on 48 °C to 46 °C over 100 hours, and a hold at 47 °C. Using growth times as described in Extended Data Fig. 5, fluorescence data for many samples in both experiments showed preference for the desired shapes (**b, c**), but with considerable variation in selectivity and total amount of growth. **d**, No statistically significant correlation was found between the nucleation model prediction for temperature of on-target

nucleation and the time of on-target shape quenching in the temperature ramp experiment. Although the nucleation model overestimates the nucleation temperature and its nucleation rate estimates may be far off, our interest here is in the qualitative features and difference between the shapes. **e–g**, Details of three patterns, with concentration patterns (1), weighted critical nucleus free energy starting from particular tiles (2), nucleation-model-estimated nucleation rates (3), temperature hold (4) and temperature ramp (5) experiment fluorescence results, and (6) AFM images from the temperature hold experiments. Information for all individual flag patterns is available in the Supplementary Information section 5.

(a)

Tile concentrations: A flag 12   ■ 50 nM   □ 880 nM

Tile concentrations: uniform concentration (no flag)

(b)

Temperature ramp fluorescence over time, p = A flag 12

on target fluorophores circled (for A)

Temperature ramp fluorescence over time, p = no flag

$$M_{p,s,f} = \frac{1 - F_{p,s,f}}{1 - F_{\text{no flag},s,f}} \qquad M_p^{\text{on target}} = \frac{1}{5} \sum_{s \in \{H,A,M,3\}} \sum_{f \text{ in target}} M_{p,s,f} \qquad M_p^{\text{off target}} = \frac{1}{10} \sum_{s \in \{H,A,M,3\}} \sum_{f \text{ off target}} M_{p,s,f}$$

(c)

Extended Data Fig. 7 | Evidence of winner-take-all in flag experiments.
a, An example flag pattern (A flag 12), and uniform 50 nM concentration 'no flag' pattern. **b**, Fluorescence normalized to maximum readings, from the temperature ramp experiment (Extended Data Fig. 6 and Supplementary Information section 5). The fluorescence at the end of the experiment, $F_{p,s,f}$, of fluorophore f in sample s of pattern p is used along with the corresponding fluorescence value for the no flag pattern, $F_{\text{no flag},s,f}$, to calculate the ratio $M_{p,s,f}$. This ratio corresponds to the relative amount of quenching for that fluorophore in the flag pattern compared to the no flag pattern. The ratios are averaged across the 5 on-target fluorophores (circled in **b**) in samples for the flag pattern to obtain an average on-target ratio, and across the 10 off-target fluorophores to obtain an average off-target ratio. **c**, The on- and off-target ratios are plotted for each flag pattern. For winner-take-all behavior, on-target quenching is expected to be higher with a flag pattern than with no flag, resulting in $M_p^{\text{on target}} > 1$, while off-target quenching is expected to be reduced, resulting in $M_p^{\text{off target}} < 1$.
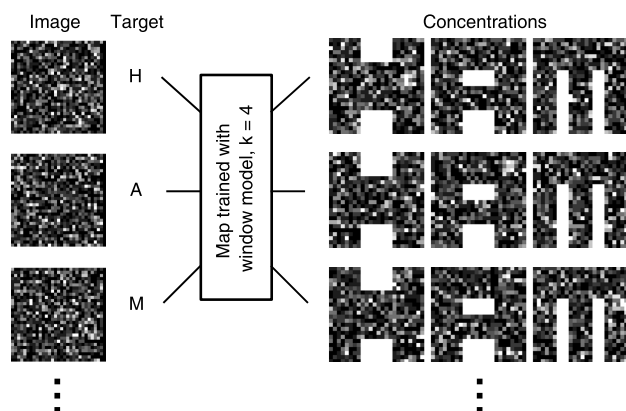
Article

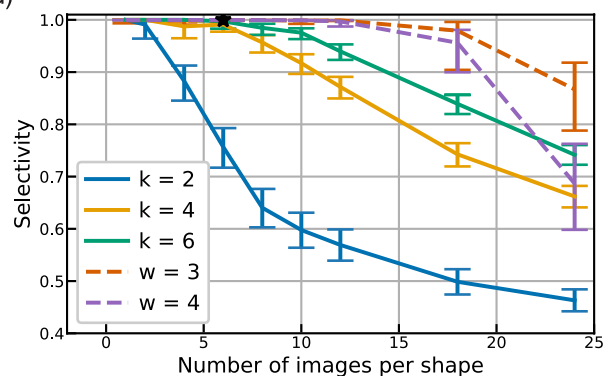## (a) Concentrations for 1 image per shape, k = 4
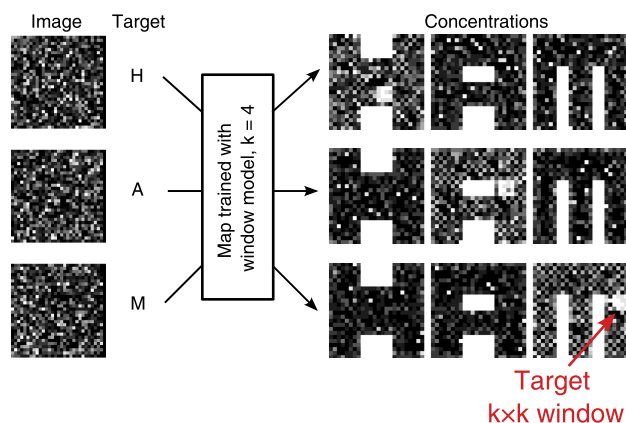
(b) Concentrations for 6 images per shape, k = 4
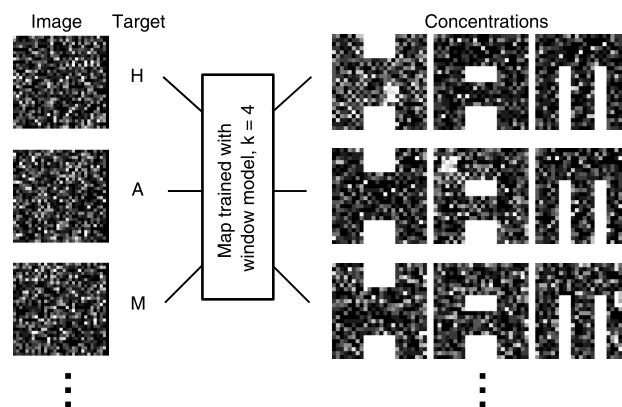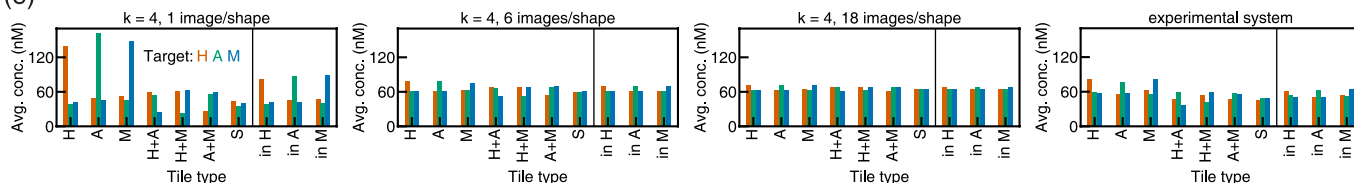

## (c) Concentrations for 18 images per shape, k = 4

(d)

(e)

Extended Data Fig. 8 | Pattern recognition capacity. To analyze the pattern-recognition capabilities of the designed tile set, the map-training algorithm (see Supplementary Information section 2.4) was run for increasingly larger sets of random images. **a-c**, Example images mapped to concentration patterns for sets with 1, 12, and 18 trained images per shape, with the intended target shape for each image indicated. Following the same procedure as used for the experimental system, with the same weighting of locations, 30 × 30 images with 10 possible grayscale values and matching histograms were mapped exponentially to tile concentrations in the 917 tile system; however, all images were generated randomly. Training was done using only the Window Nucleation Model with a window size k of either 2, 4, or 6, with a limit of 400,000 steps (Supplementary Information section 2.5). For each number of images per shape considered, ten repetitions of training (starting from random assignments) were performed (to account for variability of the training algorithm) for each of three different sets of images (to account for variability in sets of images). **d**, As the number of images in the set increases, the selectivity of nucleation using the trained map decreases. For larger k, the pixel-tile map can exploit higher-order correlations and can thus accommodate more images. For each fully-trained system, nucleation rates were calculated using the Stochastic Greedy Model, described in Supplementary Information section 2.2, at $G_{se} = 5.4$, which roughly corresponds to a temperature of 48.6 °C, and with concentrations comparable to the experimental system. Selectivity was calculated as the nucleation rate of the target shape for each image divided by the total nucleation rate of all three shapes for that image, averaged over all images in the system, and over all 30 systems (10 repetitions for each of 3 sets of images) for each point, with 90% confidence intervals shown. Star shows selectivity calculated from nucleation model results for the experimentally-implemented system. Alternatively, dashed lines show results (at $G_{se} = 5.5$) for maps constructed by a simpler training method that assigns the highest w^2 previously-unassigned pixels in each training image to a unique $w \times w$ region in the target shape, detailed in Supplementary Information section 2.6. These maps have at least as much capacity as the model-trained maps within the time constraints of these tests, suggesting a robustness to training method. **e**, As the number of images increases, pattern recognition must increasingly rely on patterns of concentrations of shared tiles, rather than choosing a pixel-to-tile map that places high-concentration pixels on tiles unique to the target shape. Histograms show average concentrations of tiles in different shapes or combinations of shape (including the average across tile categories) for images in training cases **a–c**, and the experimental system. The change can also be seen in the concentration maps of **a–c**, with the sharp checkerboard of high concentration tiles in target shapes in **a** becoming less apparent in **b** and **c**.
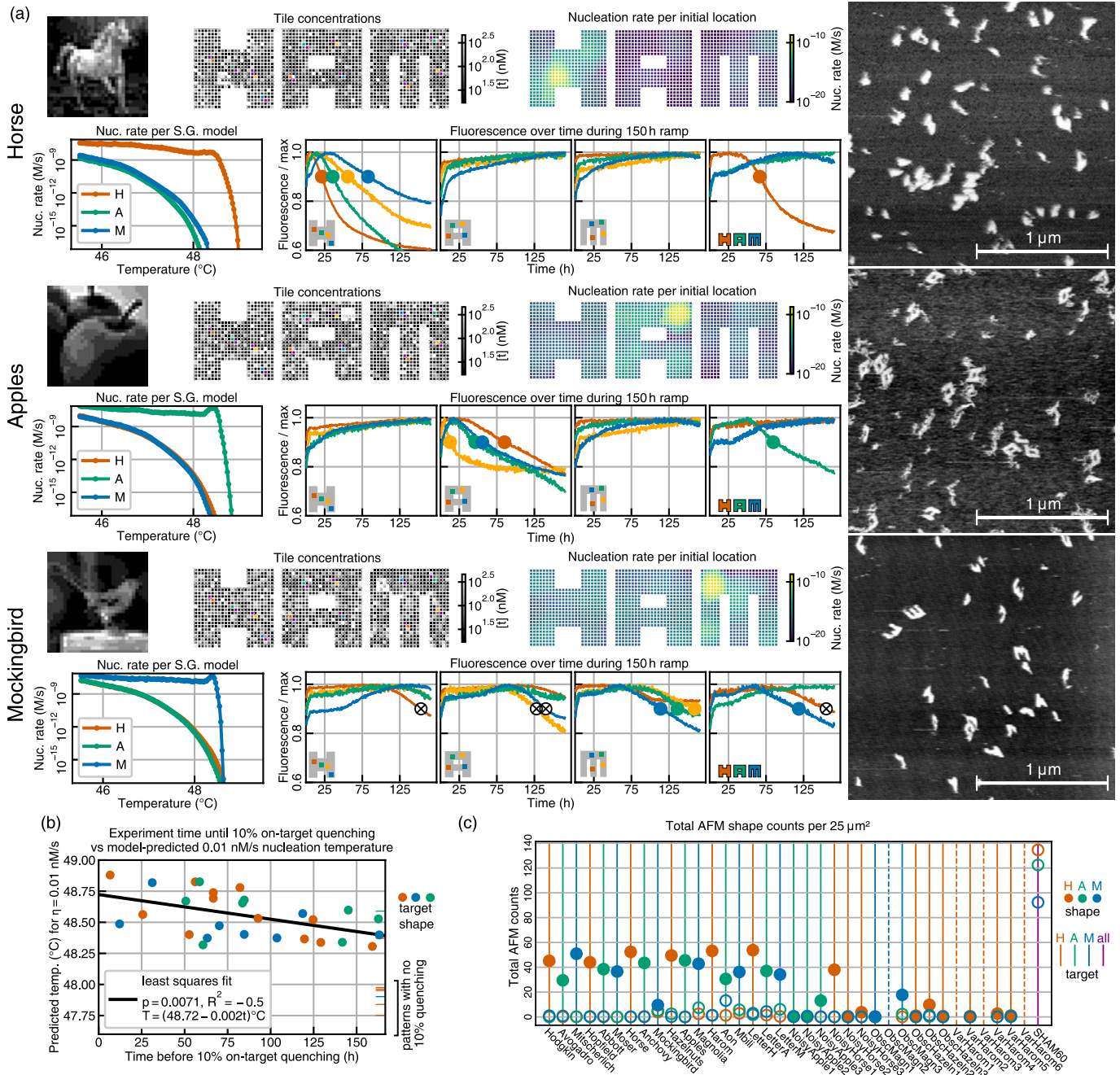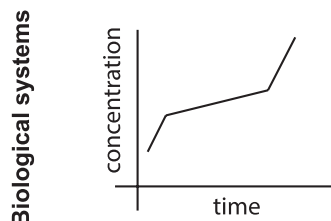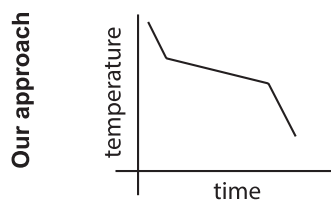
Extended Data Fig. 9 | Classification of images viewed as concentration patterns. 36 different concentration patterns, derived from a mapping of 36 grayscale images, were run using a ramp between 48 °C to 45 °C over approximately 150 hours. **a**, Three pattern examples, with source image, concentration pattern, nucleation model nucleation rate starting from particular tiles, nucleation model nucleation rates, fluorescence results, and AFM images. **b**, Across all patterns there was some correlation between the on-target nucleation temperature predicted by the nucleation model and on-target shape quenching time. **c**, Total AFM shape counts for each sample. Information for all patterns is available in the Supplementary Information section 6.
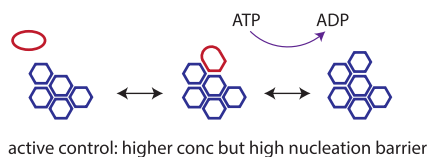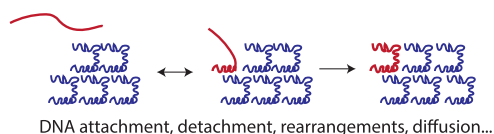
Article

**A) Triggering pattern recognition**

B) Timescales of pattern recognition

C) WTA via shared components

Our approach

System specific

temperature

time

DNA attachment, detachment, rearrangements, diffusion...

Biological systems

concentration

time

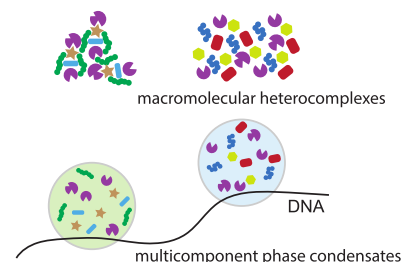ATP   ADP

active control: higher conc but high nucleation barrier

Broader trends

$$t \sim \tau e^{\sqrt{K}}$$

speed of pattern recognition vs complexity
(K = critical nucleus area)
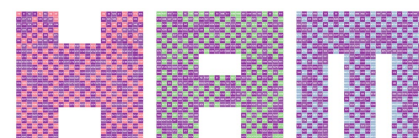
■ components shared across structures (40%)
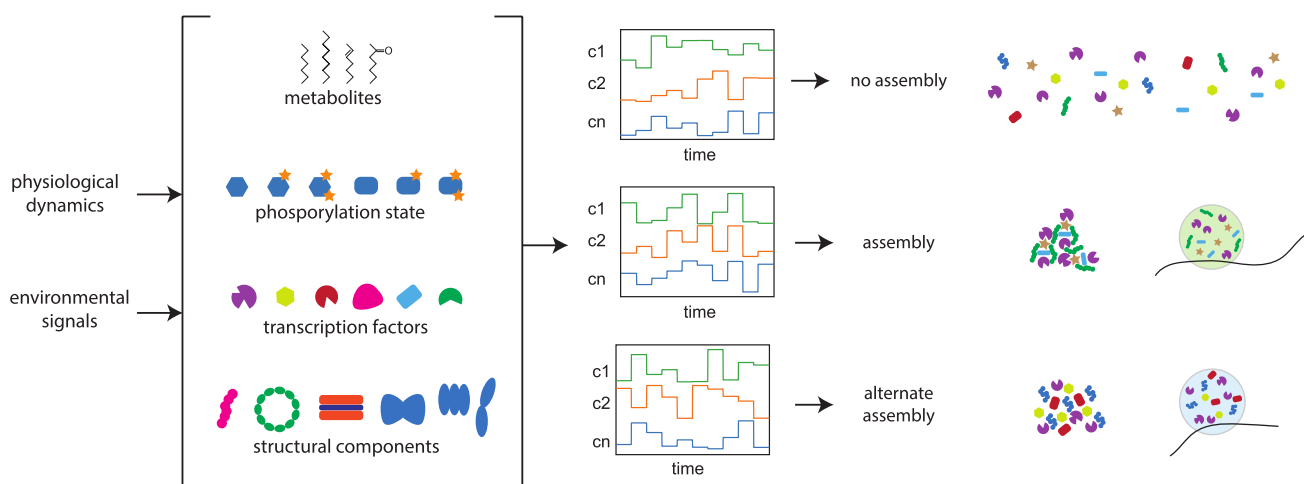■ components unique to each structure (60%)

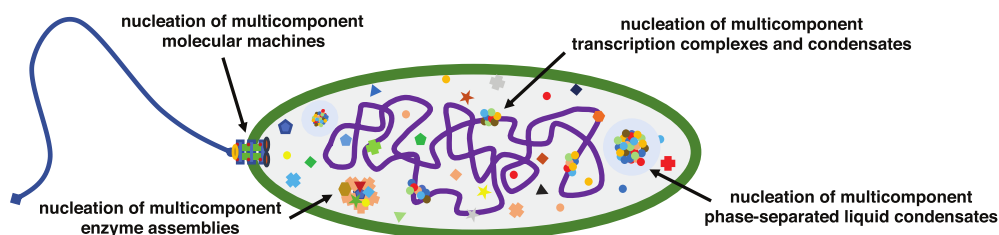macromolecular heterocomplexes

DNA

multicomponent phase condensates

D) Pattern recognition by multicomponent nucleation kinetics (through co-localization of mutually reinforcing interactions)

metabolites

physiological dynamics

phosporylation state

environmental signals

transcription factors

structural components

c1
c2
cn
time

→ no assembly

c1
c2
cn
time

→ assembly

c1
c2
cn
time

→ alternate assembly

E) Where pattern recognition might occur in natural or synthetic cells

nucleation of multicomponent molecular machines

nucleation of multicomponent transcription complexes and condensates

nucleation of multicomponent enzyme assemblies

nucleation of multicomponent phase-separated liquid condensates

Extended Data Fig. 10 | Parallels and differences between pattern recognition in our synthetic approach and in potential biological systems. While we studied pattern recognition using a specific set of molecules (DNA) in an in vitro system, the concepts behind our work have potential relevance to biological systems built out of molecules of different nature and in different conditions. **a**, Pattern recognition was triggered in our system by lowering temperature over time which drives the self-assembly process forward. Instead, in the cell, concentrations of molecular components can rise over time (e.g., through gene expression), leading to nucleation and self-assembly and thus pattern recognition. **b**, Timescale of pattern recognition is controlled by system-specific aspects and also general trends. System specific: DNA-specific processes such as tile attachment, detachment and restructuring set the timescale of nucleation and growth. Living systems can use active mechanisms to control nucleation timescales in addition to the concentration effects described here. Broader trends: our theoretical work supports a general relationship between the speed of pattern recognition (e.g., by working at a lower temperature), the size of critical nuclei and thus the complexity of pattern recognition. **c**, The winner-take-all effect in our work enhanced selectivity by exploiting the depletion of shared components. Biomolecular systems, such as macromolecular complexes[56] and multicomponent phase condensates[46] are thought to share components as well, potentially enabling a winner-take-all effect in cells. **d**, In the biological context, the inputs could represent physiological or environmental signals encoded in the relative concentrations of many species of molecules. Some patterns of enhanced concentration may not lead to self-assembly or phase condensation if those components with enhanced concentration are not colocalized on a structure or reinforce a nucleation pathway for a condensate; but an alternative pattern of high concentrations could lead to assembly of one of several assemblies or condensates. **e**, Such sensitivity of kinetic pathways to concentration patterns can be exploited for complex decision-making in numerous aspects of cellular physiology, or may provide compact and robust control mechanisms for cell-scale molecular robots. See also Supplementary Information section 1.5 for how a pixel-to-tile map could be physically incarnated.