Multifarious assembly mixtures: Systems allowing retrieval of diverse stored structures

Arvind Murugan[a,b,1,2], Zorana Zeravcic[a,b,1,2], Michael P. Brenner[a,b], and Stanislas Leibler[c,d]

[a]School of Engineering and Applied Sciences and [b]Kavli Institute for Bionano Science and Technology, Harvard University, Cambridge, MA 02138; [c]Simons Center for Systems Biology, School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540; and [d]Laboratory of Living Matter, The Rockefeller University, New York, NY 10065

Edited by José N. Onuchic, Rice University, Houston, TX, and approved November 17, 2014 (received for review July 22, 2014)

Self-assembly materials are traditionally designed so that molecular or mesoscale components form a single kind of large structure. Here, we propose a scheme to create "multifarious assembly mixtures," which self-assemble many different large structures from a set of shared components. We show that the number of multifarious structures stored in the solution of components increases rapidly with the number of different types of components. However, each stored structure can be retrieved by tuning only a few parameters, the number of which is only weakly dependent on the size of the assembled structure. Implications for artificial and biological self-assembly are discussed.

programmed assembly | stored structures | complex materials

A classical example of self-assembly is crystallization. At low temperatures the crystalline phase is typically stable and thus grows spontaneously from solution through homogeneous nucleation. If several competing crystalline phases are allowed by microscopic interactions, the efficient production of a desired phase often requires heterogeneous nucleation from a seed of this phase, with precise annealing protocols (1, 2). More complex microscopic interactions may lead to a glassy phase with many competing structures; however, it is generally impossible to control local compositions or microscopic interactions to obtain a particular structure.

Recently, there has been a dramatic change in macromolecular and colloidal assembly techniques, made possible by the use of biopolymers, such as DNA, to create a large variety of intercomponent interactions. Indeed, biomolecules offer exquisite control of microscopic interactions that allows self-assembly of diverse large structures. Examples range from nanoparticle assemblies (3–6), which can also form macroscopic crystals (7–10), to structures using DNA itself as a building material. In the latter case, DNA origami uses short DNA strands to controllably fold a long backbone strand into different well-controlled structures (11, 12), whereas short strands of DNA by themselves can also build up complex 3D objects (13–15). Similar efforts are underway, using rationally designed proteins by creating complementary binding sites on their surfaces (16–18). Until now, however, experimental and theoretical studies have been limited to devising interactions for the assembly of a single structure. This is contrasted with biological systems, where many different self-assembled structures can be formed within the cell cytoplasm. These assembled structures can in fact share some of their components and can be dynamically induced independently from one another (19).

Here, we propose a previously unidentified mechanism for the self-assembly of many different structures from one large set of shared components. Each structure is multifarious, i.e., is made of many different types of components. Such self-assembling systems, which we propose calling "multifarious assembly mixtures," are stable and yet responsive. This means that the mixtures do not form structures spontaneously, but can be controllably induced to assemble a specific structure. Different structures are "encoded" through the choice of molecular interactions and thus "stored" in the mixture, to then be "retrieved" by changing only a small number of parameters.

The theoretical framework introduced below allows calculation of the capacity of these systems, i.e., of how many different independent structures can be stored and retrieved in a mixture of N species of components. In the traditional approach of self-assembly without shared components, if each structure S is composed of the same number N_S of different species, only N/N_S different structures can be self-assembled. In contrast, in multifarious assembly mixtures, many more distinct structures can be stored. Any stored structure can be retrieved with a (super)critical nucleation seed. Multiple seeds can induce the simultaneous assembly of multiple corresponding structures. Moreover, we show that each different structure can be retrieved by changing only a small number of chemical potentials or interspecies interactions, where the number of tuned components is the number of components in a (super)critical nucleation seed of the desired structure. Classical nucleation theory implies that the size of this seed is only weakly dependent on the size of the structure that is built.

Model

Consider N species of interacting components in a solution kept at a constant temperature T. In principle, each species can have a different chemical potential μ_α, $(\alpha = 1, \ldots, N)$, but for simplicity we assume for now that the chemical potentials have the same value μ. We want the components to be able to self-assemble into one of m distinct, multifarious structures, $S = 1, \ldots, m$, on demand (Fig. 1*A*).

A typical multifarious structure S is built of N_S component species. In general, each species α in the structure S has its own multiplicity n_α^S. In contrast to traditional studies of self-assembly, e.g., of crystals, where the same component species appears in many copies in the assembled structure, for multifarious assembly

Significance

Self-assembly has recently emerged as a powerful technique for synthesizing structures on the nano- and microscales. The basis of this development is the use of biopolymers, like DNA, to design specific interactions between multiple species of components, allowing the spontaneous assembly of complex structures. Our work addresses a fundamental limitation of the existing approaches to self-assembly: Namely, every target structure must have its own dedicated set of components, which are programmed to assemble only that very structure. In contrast, in biological systems, the same set of components can assemble many different complexes. Inspired by this, we extend the self-assembly framework to mixtures of shared components capable of assembling distinct structures on demand.

Author contributions: A.M., Z.Z., M.P.B., and S.L. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]A.M. and Z.Z. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: zorana@seas.harvard.edu or amurugan@seas.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1413941112/-/DCSupplemental.

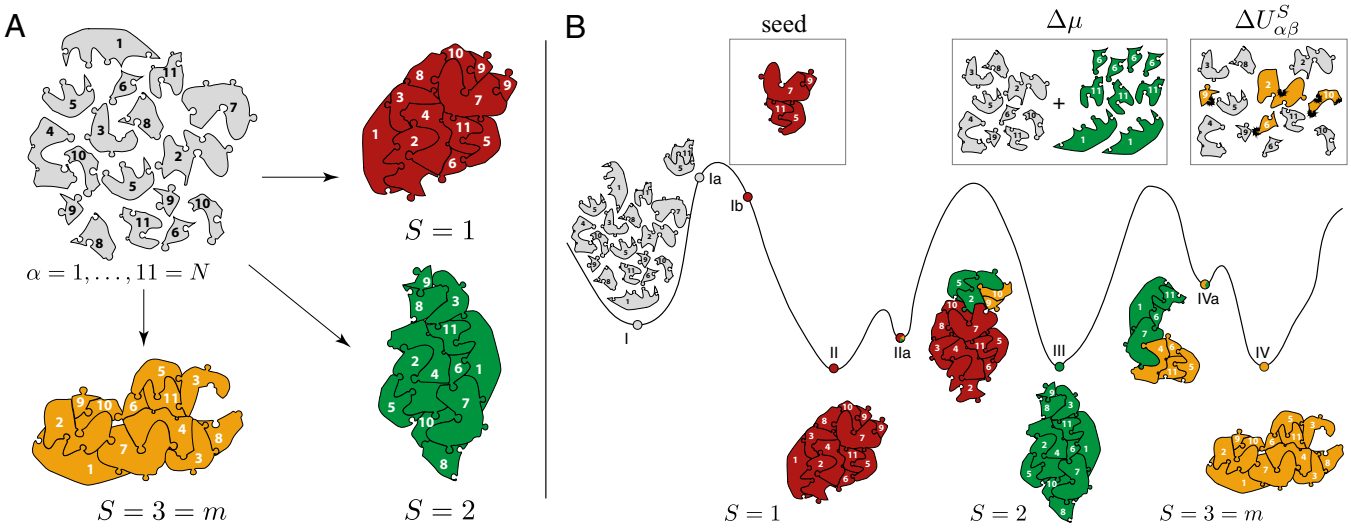

Fig. 1. (*A*) Schematic depiction of the basic idea of assembly of different desired multifarious structures ($S=1, S=2$, or $S=3=m$) by using the same set of components. In general, the multiplicity of different component species within a structure can be nontrivial; i.e., the number of component species, N_S, composing structure S can be different from the size of that structure, M_S, $N_S \neq M_S$. For example, in structure $S=1$, the multiplicity of species 9 is $n_9^1=2$. Similarly, $n_8^2=2$ and $n_3^3=2$ in structures $S=2$ and $S=3$, respectively. (*B*) Free energy landscape and chimeric states. (I) A solution of N different species of components, with interactions designed for assembly of desired structures $S=1$ (II), $S=2$ (III), and $S=3=m$ (IV). The desired stored structures are not the only free energy minima; chimeric structures, i.e., hybrids between different stored structures, can also exist (IIa and IVa). *Insets* show assembly of the stored structures can be triggered by manipulating a small number of components: (*Left*) introducing a supercritical seed, a subcluster of the desired stored structure; (*Center*) increasing the average concentration of components that can make a supercritical seed by tuning their chemical potentials; and (*Right*) increasing the specific binding energy of components that can make a supercritical seed.

mixtures we are interested in the case of small values of n_α^S. Indeed, for simplicity, we assume here that all component species have a single copy in every stored structure, $n_\alpha^S=1$, so that the number of species N_S used in the structure equals the size of the structure $N_S=M_S$. Additionally, we make a simplifying assumption that all of the structure sizes M_S have the same value M, and so $N_S=M_S=M$.

Both cellular systems and recent DNA-mediated assembly experiments show that a single structure S can be robustly assembled if each pair of neighboring components of species α and β $(\alpha,\beta\in\{1,\ldots,N\})$ interact through a specific binding interaction. Our next simplifying assumption is that all these interaction energies are equal, $U_{\alpha\beta}^S=-E$, and we also set all nonspecific interactions to zero. The binding interactions between different components are mediated through a discrete number of "binding sites," with a species α having a valence z_α. For simplicity we assume that all components have the same valence z.

How might we choose an interaction energy matrix $U_{\alpha\beta}^{\text{tot}}$ so that the components are capable of assembling different desired structures $S=1,\ldots,m$ (Fig. 1*A*)? The simplest general prescription that can work for arbitrary structures is to assume that two species α and β bind specifically with energy $-E$ if and only if at least one of the desired structures S requires this binding. Such a matrix $U_{\alpha\beta}^{\text{tot}}$ then has the potential for "storing" each structure S as a local free energy minimum (Fig. 1*B*). This matrix can be written as

$$U_{\alpha\beta}^{\text{tot}}=\begin{cases}-E & \text{if } \alpha,\beta \text{ interact specifically } \left(U_{\alpha\beta}^S=-E\right) \text{ in any } S,\\ 0 & \text{otherwise.}\end{cases}$$

This form of energy matrix implies that component species can be promiscuous in their interactions. Indeed, because a given species α binds specifically to its partners in each of the stored structures, the total number of specific binding partners for species α can be large.

In addition to the free energy minima corresponding to the desired structures, other undesired local minima might emerge. These correspond to chimeric structures, or "chimeras," made of chunks of different desired structures that can bind together due to the promiscuity implied by Eq. **1**. The stability of the stored structures is determined by the size of the free energy barriers between the different minima. For instance, if the barriers are low, chimeras will form spontaneously, even if their local free energy minima lie higher than those of the desired structures (Fig. 1*A*). Similarly, the free energy barriers between the solution of unbound components and other minima determine the solution's characteristic time t_*, beyond which stored structures nucleate spontaneously and the process of the controlled retrieval of stored structures is compromised. Thus, t_* is the functional "lifetime" of the multifarious assembly mixture.

Storage Capacity

How many different multifarious structures, each of size M, can one store by using N different species of components with well-chosen interspecies interactions defined by Eq. **1**? If each species contributed to only a single structure, the maximum capacity would be N/M. By sharing species between structures, however, a much larger number of structures can be stored before chimeras start to dominate. To find this increased capacity, consider components attaching to the boundary of a growing seed. The promiscuous interactions implied by Eq. **1** might allow the seed to bind different sets of components, resulting in chimeras. Therefore, let us compute the number of species that can specifically bind to a given boundary site of the seed. Because each component in the bulk of a stored structure has z nearest neighbors, for an incoming component to bind stably, it must form specific bonds with $z/2$ components on the seed's boundary. Due to the promiscuous nature of Eq. **1**, each of these $z/2$ boundary components can bind specifically to a set of $\sim m(M/N)$ other species. [To see this, note that if each structure of size M is randomly constituted from the N species, a given species will occur in $\sim m(M/N)$ of the m stored structures and typically have a different partner in each of them. Hence, a typical species will have $\sim m(M/N)$ specifically binding partners.] For randomly constituted m structures, each set contains a fraction mM/N^2 of all of the N component species. The intersection of these $z/2$ sets, of the

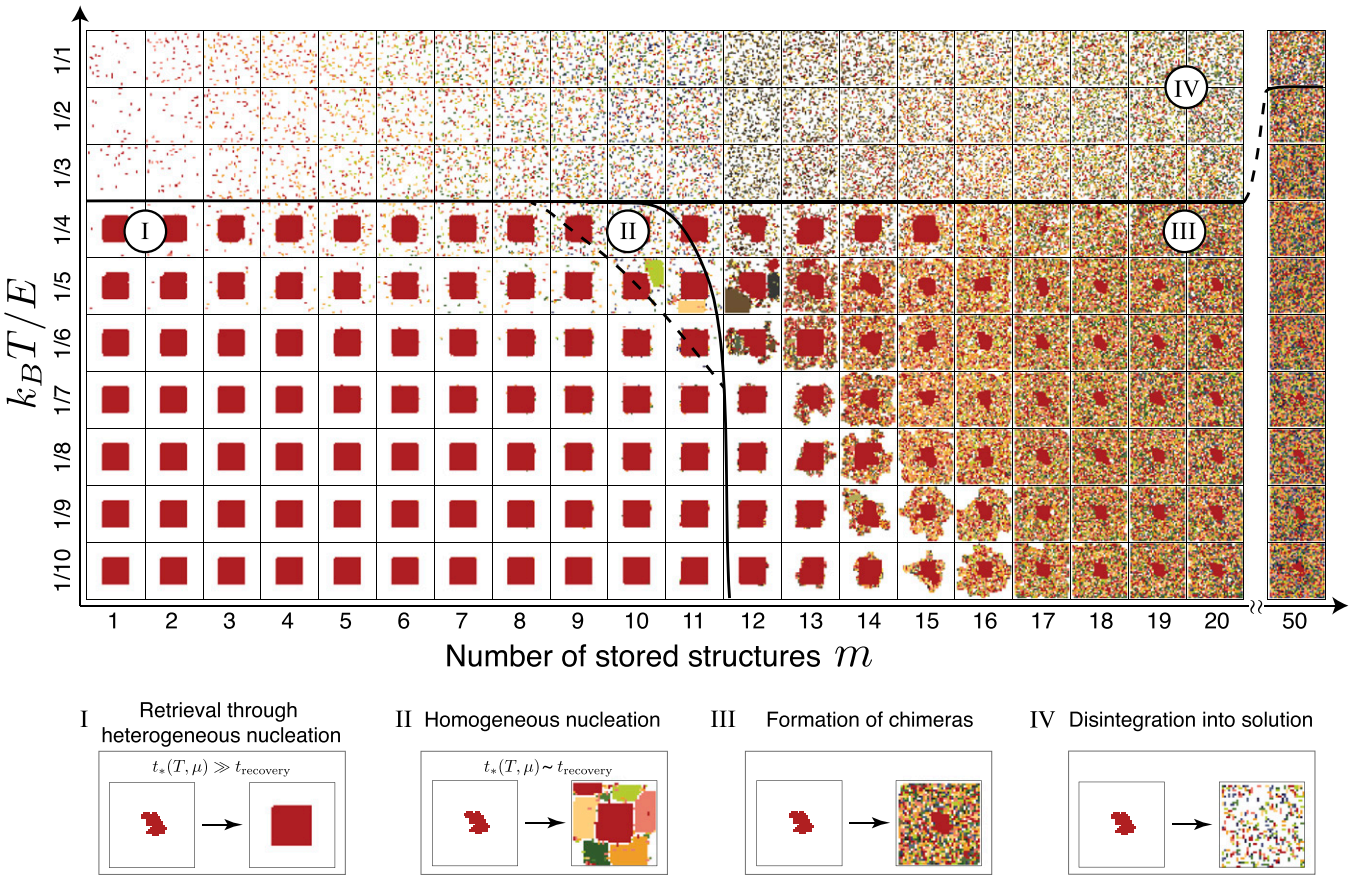

Fig. 2. Diagram of the different simulation outcomes as a function of the number of stored structures m and temperature k_BT/E, starting from a particular supercritical seed (shown at *Bottom*). We use different colors to visualize different stored structures, with the seeded structure colored in dark red. *Bottom Row* distinguishes the four regimes identified in the diagram. In regime I the desired structure is retrieved through heterogeneous nucleation because the solution remains stable in the time required for assembly. The solution in this regime is a functional multifarious assembly mixture. Regime II is characterized by homogeneous nucleation of all structures due to reduced stability of the solution (*SI Text*). In regime III, formation of structures is dominated by chimeras. Finally, in regime IV, any initial seed is disintegrated into the solution (*SI Text* and Fig. S1). These simulations were run for a fixed length of time, 2×10^6 lattice sweeps, and with fixed chemical potential, $\mu = 1.80E$, for all species. The value of μ mostly influences the extent of regimes I and II. In each plotted snapshot only neighboring tiles that have specific binding between them are plotted, and hence tiles without any bonds are omitted. Note that in a system with fixed concentrations, rather than μ, most components would clump to the seed in regime III, whereas in regime I they would disperse in the solution independently of the structure nucleated from the seed.

size $N(mM/N^2)^{z/2}$, determines the species that can specifically bind to all of the $z/2$ boundary components. When this number is larger than 1, many different species can attach to a given boundary site on a growing seed, resulting in a proliferation of chimeras. Hence, the largest number m of structures that can be stored is

$$m_c \sim \left(\frac{N}{M}\right) N^{(z-2)/z}. \qquad [1]$$

For $z>2$, the exponent $(z-2)/2$ is positive and this equation implies that the capacity m_c can be much larger than the traditional estimate of the capacity N/M. It is instructive to understand why $z=2$ structures i.e., linear chains, cannot share components. Binding to an end of a growing chain requires forming a bond with just one component. If that component is promiscuous, the seed can always grow in a nonunique chimeric manner. Hence the promiscuity of individual species, implied by Eq. **1**, must be countered by the requirement on incoming particles to form multiple (i.e., $z/2>1$) bonds. (A detailed description of our calculations is in *SI Text*.)

Retrieval

The above argument shows that the number of structures that can be stored and stabilized with N components is large. For this to be useful, we need to be able to retrieve each of them easily. The retrieval can be done in three different ways. One can introduce a nucleation seed, i.e., a part of a stored structure, into the solution. Alternatively, one can enhance the formation of such a seed by increasing the chemical potential of its components by an appropriate amount $\Delta\mu$ or by strengthening the interactions $U^S_{\alpha\beta}$ by ΔU for bonds found in such a seed. These methods enhance the nucleation of one stored structure without nucleating others, despite all stored structures being made of the same set of components. Such selective nucleation is possible only for multifarious structures; it relies on the fact that small contiguous subsets of distinct structures typically have distinct compositions. Such subsets can be used as selective nucleating seeds or to selectively lower the nucleation barrier for one structure, using the other two methods described above.

The critical question is, How many different species have to be tuned in this way to successfully retrieve a particular stored structure? The answer follows directly from general nucleation theory, which specifies a critical nucleation radius r_* in terms of the chemical potential μ and bond energy E (20). The minimal seed size N_* needed to recover a structure is set by r_*; smaller seeds dissolve back into components whereas larger seeds are supercritical and grow into stored structures. We can make the multifarious assembly mixture responsive to smaller seeds by

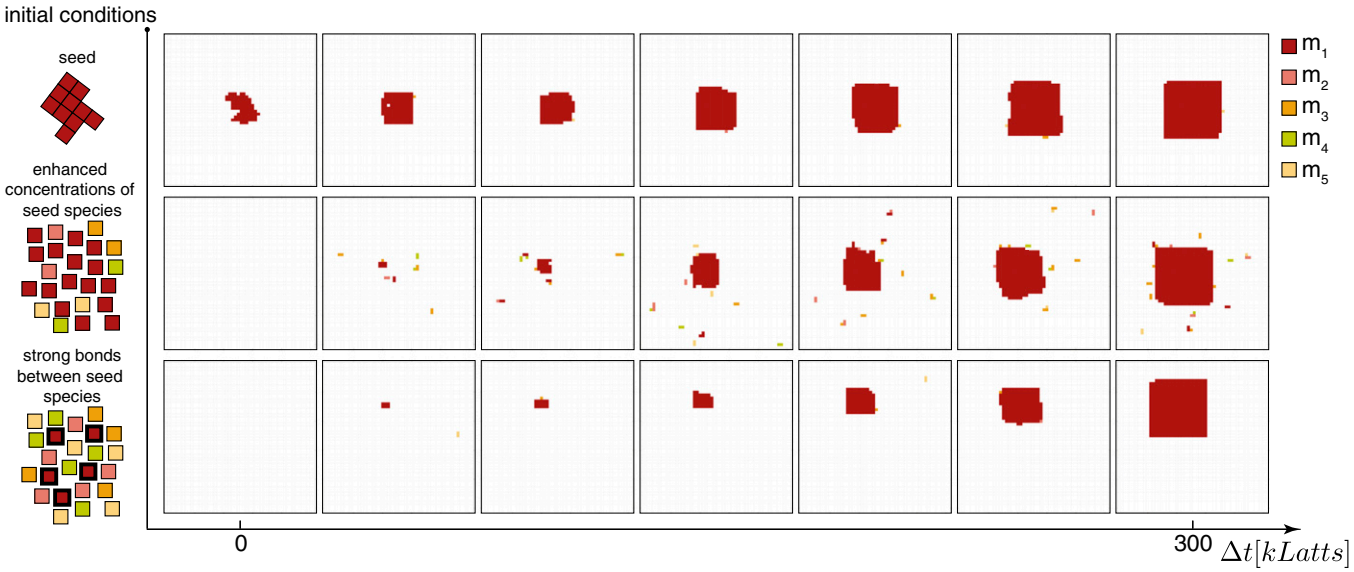

Fig. 3. Configurations observed during the simulations of retrieval of the desired structure (dark red) in a solution of tiles, whose interactions encode five different structures ($m=5$). Each type of simulation was run at a fixed temperature $k_BT/E=0.15$. The retrieval uses three different triggers: (*Top Row*) a nucleation seed, i.e., a subcluster of the desired structure, as appears in the first snapshot (we used chemical potential $\mu=1.85E$ for all tile species and observed the retrieval of the desired stored structure progress during the time window between 10^5 and 3×10^5 lattice sweeps); (*Middle Row*) enhanced concentrations of a small number of tile species that can make the seed used in the top row, by using $\mu=1.35E$ for these tile species and $\mu=1.85E$ otherwise (we observed the retrieval of the desired structure in the time between 4×10^5 and 7×10^5 lattice sweeps); (*Bottom Row*) stronger binding energies $U_{\alpha\beta}=-2E$ between the small number of tile species that can make the seed shown in the *Top Row* (we used chemical potential $\mu=1.85E$ for all tile species and observed the retrieval progress in the time between 3×10^5 and 6×10^5 lattice sweeps). In each of the three rows the snapshots were taken within the time interval $\Delta t=3\times10^5$ lattice sweeps, starting at the time when retrieval begins.

lowering the critical nucleation radius, for example, by lowering the ratio μ/E of chemical potential to bond energy (*SI Text* and Figs. S2–S4).

However, lowering the critical nucleation radius also lowers the barrier to spontaneous homogenous nucleation. As noted above, critical seeds can spontaneously assemble on a characteristic timescale t_* and grow into random stored structures, without any external input. Thus, at a minimum, we need t_* to be much longer than the retrieval time, i.e., time necessary for a supercritical seed to grow into a full structure. Nucleation theory, adapted to multifarious d-dimensional structures, determines t_* as

$$\log\left(\frac{t_*}{\tau}\right)=\frac{F_*}{k_BT}-\log(q(m,M)), \qquad [2]$$

where $F_*\sim\gamma r_*^{d-1}$ is the free energy barrier, γ is the free energy per area required for creating the critical seed, and τ is a timescale connected with microscopic processes. The second term on the right-hand side arises because we must account for the multiplicity $q(m,M)$ of distinct nucleation paths leading to the m different stored structures. For small m, we can estimate $q(m,M)\sim mM$ to account for critical seeds from different parts of the m stored structures of size M each (*SI Text*).

For a fixed t_*/τ, Eq. **2** can be solved for the nucleation radius r_* and hence the minimal number of components N_* that must be tuned to retrieve a structure. If all of the components have a typical size a, this number N_* is of order

$$N_*\sim\left(\frac{r_*}{a}\right)^d=\left(\frac{k_BT}{a^{d-1}\gamma}\right)^{d/(d-1)}Q\left(m,M,\frac{t_*}{\tau}\right), \qquad [3]$$

where Q depends only weakly (logarithmically) on m, M, and t_*/τ. We thus conclude that because N_* is determined by the nucleation barrier, it is essentially independent of the size of the structure, M, that is being retrieved. Note that the above equations show an unavoidable tradeoff: Increasing the lifetime of the multifarious assembly mixture t_* necessarily increases the nucleation radius r_* and hence increases N_*. Thus, a more stable multifarious assembly mixture requires a larger seed for recovering stored structures.

Simple Lattice Model

To study different regimes of self-assembly of multifarious structures, we have considered assembly based on Eq. **1**, on a simple $2d$ square lattice. Individual components are square tiles that can be one of $N=400$ species. All m stored structures consist of $M=N$ tiles, each tile being of different species positioned inside a 20×20 square block. More precisely, we assume that all of the species are present in all of the structures, and each species appears only once in each structure, $n_\alpha^S=1$ for all α and all S, so that $N=N_S=M_S=M$ for all structures $S=1,\ldots,m$. In other words, each stored structure is simply a different random permutation of the tiles inside the square block. We assume that each tile component can bind up to $z=4$ neighbors through specific binding interactions given by Eq. **1** and that all species of tiles have the same chemical potential μ. We run grand canonical Monte Carlo simulations with different numbers m of stored structures on a square lattice of total size 40×40, for different values of temperature T and chemical potential μ (*SI Text* and Fig. S5 provide a detailed description of our simulations).

Starting from a particular supercritical seed (of linear size $r>r_*$) of one of the m stored structures, Fig. 2 shows a diagram of the different outcomes of our simulations, as a function of the number of stored structures, m, and the temperature T (or more precisely, k_BT/E, where E is the specific binding energy), for a fixed μ. We visualize the different stored structures with different colors, with the desired structure colored in dark red. For low m and T, the supercritical seed indeed grows into the desired structure. In this regime of parameter space (regime I), the solution behaves as a useful multifarious assembly mixture: The mixture is stable for a long time t_* and stored structures can be retrieved through heterogeneous nucleation. [Our results show that the simulation dynamics obey the prediction of classical

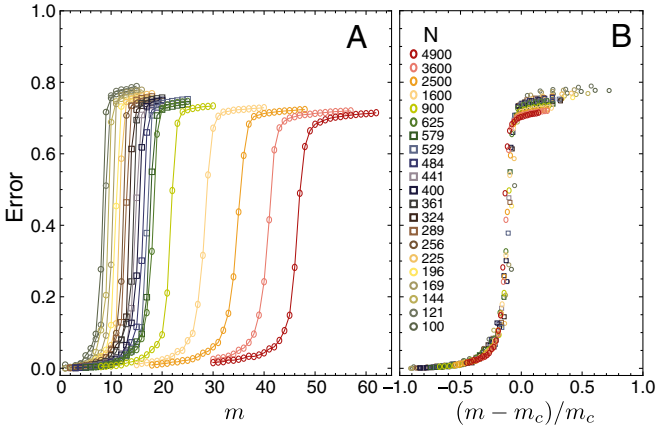

Fig. 4. Scaling of the storage capacity. (*A*) Measure of the difference between the desired structure and the obtained structure (see *SI Text* and Fig. S6 for the definition and Fig. S7) as a function of the number of stored structures m, for different numbers of tile species N. Stored structures contain $M=N$ tiles, each of different species. Each point is an ensemble average result of ~100 different simulation runs (the curves saturate at ≈ 0.75; details in *SI Text*). (*B*) Collapse of the curves when the number of stored structures m is rescaled as $(m-m_c)/m_c$, with $m_c \sim N^{\kappa}$ and $\kappa = 0.47 \pm 0.02$ (scaling analysis in *SI Text*). All of the simulations were run for values of $N \in [100, 4, 900]$, with $\mu = 1.85E$ and $k_BT/E = 0.15$, for a fixed time of $t_{\text{run}} = 2 \times 10^6$ lattice sweeps. A diagram of the different simulation outcomes as a function of the number of stored structures and temperature for $\mu = 1.85E$ can be found in *SI Text*.

nucleation theory. For instance, within the recovery regime the timescale for appearance of a supercritical cluster, t_*, is much longer than the timescale for recovery, t_{recovery}. Thus, even though the Monte Carlo dynamics do not reflect the dynamics of a realistic self-assembly system (e.g., ref. 21), they do substantiate the predictions of nucleation theory and expose different regimes of self-assembly of multifarious structures.] For a higher number of stored structures m (and at higher temperatures T) another behavior appears (regime II). It is characterized by the spontaneous homogeneous nucleation of all stored structures from the solution: In this regime, the multifarious assembly mixture is too short lived to allow the structure retrieval; i.e., t_* becomes comparable to the time taken for a supercritical seed to grow into a full desired structure, t_{recovery}. At even higher values of m we find yet another regime of behavior (regime III), where chimeric structures dominate. Finally, at high temperatures T, and for all values of m, we encounter regime IV, where any initial seed disintegrates into small clusters of individual components. The extent of different regimes depends of course on the chosen model parameters. In particular, the chemical potential μ influences the extent of regimes I and II (*SI Text* and Figs. S1, S8, and S9).

Simulations presented in Fig. 3 confirm that, in regime I, the assembly of a structure can be triggered not only with supercritical nucleating seeds, but also by enhancing the chemical potential of a small set of tile species or by increasing the bond energies between the tile species from such a set.

Numerical simulations are also a way to gauge the capacity of a multifarious assembly mixture to store structures and to compare it with the theoretical predictions presented above. To do this, we have introduced the entire target structure as a supercritical seed and have examined it after a fixed simulation time chosen to be shorter than the mixture's lifetime t_*. We have assessed the quality of retrieval by measuring the error, i.e., the fraction of the final assembled structure that differs from the initial target structure (*SI Text* and Fig. S6). Fig. 4*A* depicts the error as a function of the number of stored structures m, for different numbers of particle species N (structure sizes being $M=N$), at fixed temperature T and chemical potential μ. There is a transition at a critical value $m=m_c$, above which the error rises rapidly. We show that the error curves for different N collapse onto each other when plotted against $(m-m_c)/m_c$, Fig. 4*B*, where m_c increases with increasing N as $m_c \equiv N^{\kappa}$ with $\kappa = 0.47 \pm 0.02$. This is in good agreement with the prediction of Eq. **1** that the memory capacity scales as $m_c \sim N^{0.5}$, for the square lattice model with $z=4$ nearest neighbors (*SI Text*).

Finally, we have also assessed the tradeoff, expressed in Eq. **3**, between the stability of the multifarious assembly mixture, i.e., its lifetime t_*, and the minimal size N_* of a seed needed for retrieval (Fig. 5). The minimal seed size N_* increases slowly with increasing t_* and remains a small fraction of the total number of components (400, in this case) in a stored structure. The number of stored structures m has only a modest effect on N_*, in agreement with Eq. **3** (*SI Text* and Fig. S8).

Discussion

To conclude, we have demonstrated that it is possible to store multiple structures in a solution of components with designed interactions between them. Using N different component species, we can store as many as $\sim (N/M)\, N^{(z-2)/z}$ different multifarious structures of size M and of average coordination number z. In an extended region of parameter values (e.g., temperature, chemical potentials, and binding energies), such a multifarious assembly mixture with many stored structures is both stable and responsive; each of the multifarious structures can be selectively grown (retrieved) by modifying chemical potentials or binding energies of only a small fraction of the N component types or by introducing an appropriate seed.

The model that we have explored is very similar to the way associative neural networks, such as Hopfield's classical networks (22), store multiple memories in a distributed way. In these models, a neural network is programmed to have multiple stable states, i.e., memories, using a prescription for neuronal connections that is very similar in spirit to the pooled energy matrix in Eq. **1**. It has been shown (23) that if the number of programmed memories is sufficiently small, each memory is indeed a stable state and can be recovered through initial conditions in a robust manner. However, if the number of stored memories exceeds the capacity of the network, recovery is spoiled by the presence of many "spurious memories"—undesired stable states—resulting in regimes (24) similar to those shown in Fig. 2. A distinctive feature of multifarious assembly mixtures, however, is that we require the stability of the unassembled mixture itself for a long time t_*, in addition to the stability of the stored structures (*SI Text*).

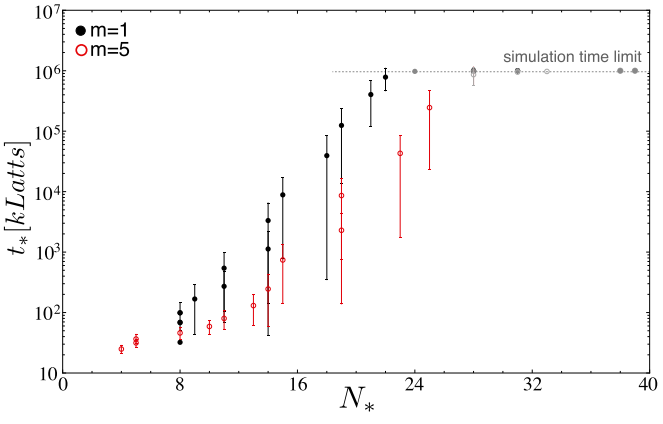

Fig. 5. Stability of the solution. The characteristic time t_* is plotted here as function of the number of tiles in a critical seed, N_*, which was varied in simulations by changing the value of $\mu \in [1.50E, 1.90E]$. The number of tiles needed to trigger assembly depends weakly (approximately logarithmically) on t_* and remains small compared with the structure size of $M=400$. Note that N_* depends very weakly on the number of stored structures m. The simulations were run at fixed temperature $k_BT/E = 0.15$.

In our simulated lattice model, different stored structures have identical components rearranged in random permutations. Thus, stored structures are assumed to be independent, or "orthogonal," as in the case of stored memories in the original Hopfield model (22). An important extension of the present model would be to study stored structures with built-in correlations, such as the presence of shared modules. After all, the controlled assembly of chimeric structures could be useful. It is also important to stress that designing specific binding interactions between different components based on superposition (Eq. **1**) is not the only way to create functional multifarious assembly mixtures. Although it is arguably the simplest prescription that works for generic structures, other nonlinear prescriptions can be tailored for particular structures by exploiting structural motifs (e.g., creating multifarious assembly mixtures with higher capacity or longer lifetimes). Such tailored interactions have been used to store and retrieve a particular set of structures composed of a small number of component species in recent work on DNA programmed assembly (25). In a similar vein, the ability of a protein sequence to code for multiple stored internal structures has been studied in the context of protein folding (26).

Beyond immediate applications to artificial systems with controllable binding specificity, the present model proposes a new paradigm to understand molecular aggregates in biology. Of course, cellular self-assembly is generically an out-of-equilibrium phenomenon in terms of its regulation and error corrections. The simple form of Eq. **1** describing equilibrium assembly is not intended to model directly biological systems, where the interactions between components such as proteins evolved over long periods of time under various constraints. In addition to involving chemical energy sources, such as ATP hydrolysis, these interactions often include allosteric effects, which should be taken into account in more realistic models. Despite these limitations, our calculations show that instead of creating new proteins for every individual structure, it is more efficient if individual proteins are used in a multiplicity of structures, as is the case in many cellular assemblies, ranging from transcription factors (27, 28) to ribonucleoproteins such as spliceosomes (29). Our calculations also indicate that such versatility can be quite high, increasing rapidly with the number of different component species in the pool. Still, different structures can be selectively assembled by reprogramming molecular interactions, e.g., by a simple modulation of the expression levels (corresponding to changing of chemical potentials in our model) or of the specific binding energies via posttranslation modifications, of a small number of selected components. This is indeed what seems to happen often in cellular assembly. Cellular self-assembly also takes place in a crowded and structured environment, where detailed kinetics, competition for components belonging to different structures, and geometrical constraints play a very important role. Such effects are missing in the present model. We hope, however, that the theoretical framework presented here, properly generalized to far-from-equilibrium situations, can form a basis for quantitative studies of functioning, regulation, and evolution of biological assembly.

ACKNOWLEDGMENTS. We thank our colleagues for discussions and their comments on the manuscript, in particular John Hopfield, David Huse, Olivier Rivoire, and Joel Lebowitz. Z.Z. acknowledges support from the George F. Carrier Fellowship. M.P.B. acknowledges funding by the National Science Foundation through the Harvard Materials Research Science and Engineering Center (DMR-0820484), the Division of Mathematical Sciences (DMS-1411694), and by Grant RFP-12-04 from the Foundational Questions in Evolutionary Biology Fund. M.P.B. is an investigator of the Simons Foundation.

1. Lifshitz I, Slyozov V (1961) The kinetics of precipitation from supersaturated solid solutions. *J Phys Chem Solids* 19(1-2):35–50.
2. Binder K, Stauffer D (1976) Statistical theory of nucleation, condensation and coagulation. *Adv Phys* 25:343–396.
3. Mirkin CA, Letsinger RL, Mucic RC, Storhoff JJ (1996) A DNA-based method for rationally assembling nanoparticles into macroscopic materials. *Nature* 382(6592):607–609.
4. Alivisatos AP, et al. (1996) Organization of 'nanocrystal molecules' using DNA. *Nature* 382(6592):609–611.
5. Valignat MP, Theodoly O, Crocker JC, Russel WB, Chaikin PM (2005) Reversible self-assembly and directed assembly of DNA-linked micrometer-sized colloids. *Proc Natl Acad Sci USA* 102(12):4225–4229.
6. Biancaniello PL, Kim AJ, Crocker JC (2005) Colloidal interactions and self-assembly using DNA hybridization. *Phys Rev Lett* 94(5):058302.
7. Park SY, et al. (2008) DNA-programmable nanoparticle crystallization. *Nature* 451(7178): 553–556.
8. Nykypanchuk D, Maye MM, van der Lelie D, Gang O (2008) DNA-guided crystallization of colloidal nanoparticles. *Nature* 451(7178):549–552.
9. Kim AJ, Biancaniello PL, Crocker JC (2006) Engineering DNA-mediated colloidal crystallization. *Langmuir* 22(5):1991–2001.
10. Macfarlane RJ, et al. (2011) Nanoparticle superlattice engineering with DNA. *Science* 334(6053):204–208.
11. Rothemund PW (2006) Folding DNA to create nanoscale shapes and patterns. *Nature* 440(7082):297–302.
12. Douglas SM, et al. (2009) Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature* 459(7245):414–418.
13. Wei B, Dai M, Yin P (2012) Complex shapes self-assembled from single-stranded DNA tiles. *Nature* 485(7400):623–626.
14. Ke Y, Ong LL, Shih WM, Yin P (2012) Three-dimensional structures self-assembled from DNA bricks. *Science* 338(6111):1177–1183.
15. Wang T, et al. (2011) Self-replication of information-bearing nanoscale patterns. *Nature* 478(7368):225–228.
16. King NP, et al. (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* 336(6085):1171–1174.
17. Lai Y-T, King NP, Yeates TO (2012) Principles for designing ordered protein assemblies. *Trends Cell Biol* 22(12):653–661.
18. King NP, et al. (2014) Accurate design of co-assembling multi-component protein nanomaterials. *Nature* 510(7503):103–108.
19. Kühner S, et al. (2009) Proteome organization in a genome-reduced bacterium. *Science* 326(5957):1235–1240.
20. Oxtoby DW (1998) Nucleation of first-order phase transitions. *Acc Chem Res* 31(2):91–97.
21. Reinhardt A, Frenkel D (2014) Numerical evidence for nucleated self-assembly of DNA brick structures. *Phys Rev Lett* 112(23):238103.
22. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA* 79(8):2554–2558.
23. Hertz J, Krogh A, Palmer RG (1991) *Introduction to the Theory of Neural Computation,* Santa Fe Institute Series (Addison-Wesley Longman, Boston).
24. Amit DJ, Gutfreund H, Sompolinsky H (1985) Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys Rev Lett* 55(14):1530–1533.
25. Barish RD, Schulman R, Rothemund PW, Winfree E (2009) An information-bearing seed for nucleating algorithmic self-assembly. *Proc Natl Acad Sci USA* 106(15): 6054–6059.
26. Fink TM, Ball RC (2001) How many conformations can a protein remember? *Phys Rev Lett* 87(19):198103.
27. Ihmels J, et al. (2002) Revealing modular organization in the yeast transcriptional network. *Nat Genet* 31(4):370–377.
28. Reményi A, Schöler HR, Wilmanns M (2004) Combinatorial control of gene expression. *Nat Struct Mol Biol* 11(9):812–815.
29. Wahl MC, Will CL, Lührmann R (2009) The spliceosome: Design principles of a dynamic RNP machine. *Cell* 136(4):701–718.