nature COMMUNICATIONS

ARTICLE

Received 7 Jun 2014 | Accepted 5 Jan 2015 | Published 11 Feb 2015

DOI: 10.1038/ncomms7203

Undesired usage and the robust self-assembly of heterogeneous structures

Arvind Murugan[1], James Zou[1] & Michael P. Brenner[1]

Inspired by multiprotein complexes in biology and recent successes in synthetic DNA tile and colloidal self-assembly, we study the spontaneous assembly of structures made of many kinds of components. The major challenge with achieving high assembly yield is eliminating incomplete or incorrectly bound structures. Here, we find that such undesired structures rapidly degrade yield with increasing structural size and complexity in diverse models of assembly, if component concentrations reflect the composition (that is, stoichiometry) of the desired structure. But this yield catastrophe can be mitigated by using highly non-stoichiometric concentrations. Our results support a general principle of 'undesired usage'—concentrations of components should be chosen to account for how they are 'used' by undesired structures and not just by the desired structure. This principle could improve synthetic assembly methods, but also raises new questions about expression levels of proteins that form biological complexes such as the ribosome.

[1] School of Engineering and Applied Sciences and Kavli Institute for Bionano Science and Technology, Harvard University, Cambridge, Massachusetts 02138, USA. Correspondence and requests for materials should be addressed to A.M. (email: amurugan@seas.harvard.edu).

A central feature of living systems is that they self-assemble their parts. Within the cell, an enormous variety of macromolecular complexes form spontaneously[1], with nearly every part having different shape and binding affinities. These complexes range from structures with a small number of different components (chaperones and proteasomes) to large structures with complex topologies (the nuclear pore complex and the ribosome). Although the assembly pathways of such complexes have been investigated[2–5], general design principles for binding energies, concentrations and other parameters of the self-assembly process are still mysterious. To minimize waste or errors, these complexes presumably assemble with high yield.

Material synthesis strategies have also recently started to use large numbers of different components, particularly with nanostructures either coated[6] with or entirely composed of DNA[7–10]. For example, work[7,8,11–13] on assemblies of short DNA strands has now been extended to a plethora of complex large shapes called tile or 'brick'[9,10] assemblies. Similar efforts are underway using rationally designed proteins[14,15] or colloidal particles[16] with specific interactions.

The major challenge with achieving high yield is eliminating competing undesired structures. Even when components are chosen so that the desired structure is the ground state, incomplete and incorrectly formed structures will occur. Incomplete structures can result from a lack of sufficient components to drive the reaction to completion. Incorrectly bound structures can result from inevitable nonspecific binding between components. Such crosstalking interactions are ubiquitous in both natural and synthetic systems. Nonspecific interactions between proteins arise from protein structure constraints[17,18], creating challenges for both signalling pathways[19] and self-assembly[12,20]. In DNA brick assembly[9,10], even with designer DNA sequences that precisely match the desired structure, there is still (weaker) binding affinity of each strand to other general components.

The number of undesired structures of varying size and composition grows exponentially with the number of component types in the desired structure[21–25]. A fundamental question is whether undesired structures can be sufficiently suppressed by choosing the binding energies[26,27] and concentrations[28] of different components.

Within biology, it is believed that the concentrations of individual proteins of multiprotein complexes obey the dosage balance hypothesis (DBH)[29], which asserts that the optimal expression levels are proportional to the stoichiometric ratio in which proteins form complexes. Likewise, the synthetic examples that we are aware of choose concentrations of individual components to match the stoichiometry of the desired structure. However, there is no evidence that stoichiometrically balanced concentrations maximize assembly yield.

We study these questions in a broad set of models relevant to protein complexes, DNA brick and colloidal assemblies. Individual structures are represented by their connectivity graphs (Fig. 1), which we then model with multivalent components with short ranged binding affinity. Our models build upon a framework introduced[12,13,30] in the context of DNA tiles.

Our analysis of these models suggests a simple new principle of 'undesired usage': control parameters such as component concentrations should be chosen based on how those components are 'used' by undesired structures, as opposed to just the desired structure. For example, if undesired structures 'use' component A more often than component B, A should be supplied in smaller amounts than B—even if the desired structure is made of one A and one B. We give a mathematical definition of 'usage' and show that this principle leads to regimes with markedly higher yields in diverse models of assembly, relevant to protein complexes, DNA-based self-assembly and colloidal assemblies. We first rigorously establish these results in an equilibrium model of self-assembly, introducing a perturbative Feynman diagram technique for computing yield that accounts for the

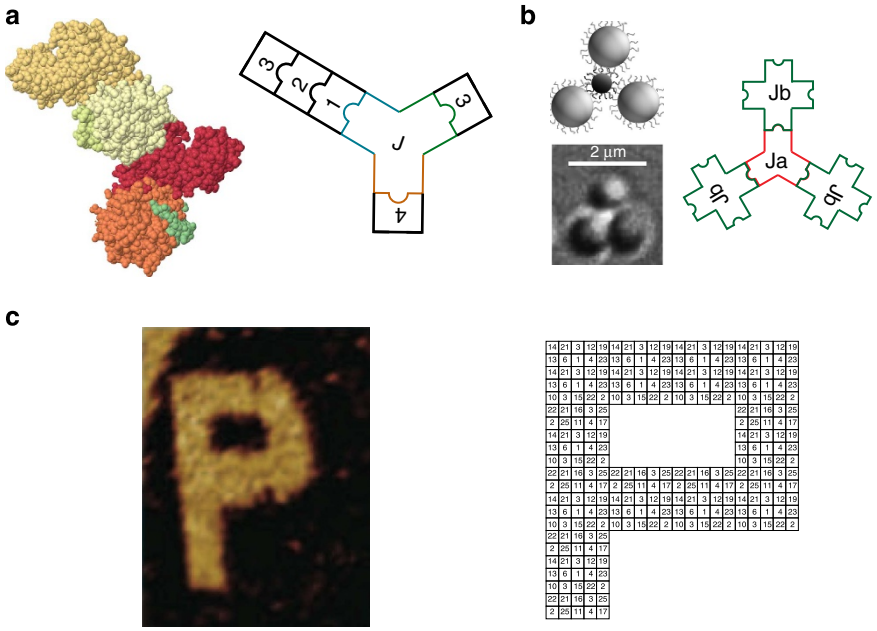

Figure 1 | Abstraction of multiprotein complexes, colloidal structures and DNA 'brick' assemblies. (**a**) The protein complex is a platelet-receptor complex involving thrombin, necessary for platelet aggregation[56] (classified by topology in ref. 1). (**b**) The colloidal structure is made of two kinds of DNA-coated particles[38] with valence imposed through geometry, whereas (**c**) the P-shaped structure is made of dozens of DNA strands called 'bricks'[9]. We model assembly using multiple component species with independent concentrations and demonstrate their role in suppressing undesired structures. Image permissions: reprinted (adapted) by permission from (**a**) RCSB PDB (http://www.rcsb.org) of PDB ID 1P8V (ref. 56). (**b**)The American Physical Society[38], copyright 2013, (**c**) Macmillan Publishers Ltd[9], copyright 2012.

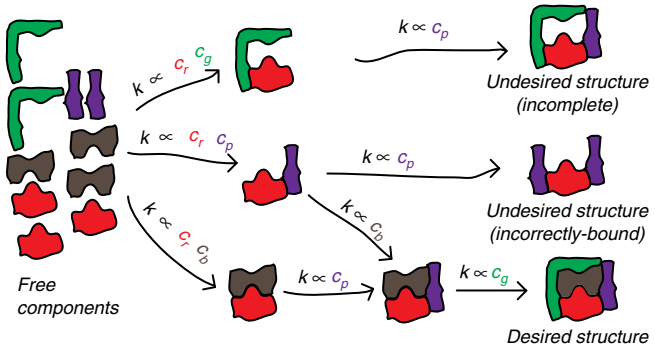

Figure 2 | Kinetic pathways leading to desired and undesired structures. Effective rate constants k (and hence fluxes) along different pathways depend on concentrations c_i of species in differing ways. Yield is improved by decreasing concentrations of species with high 'undesired usage', that is, species that increase flux along undesired pathways. Hence, optimal concentrations c_i may differ greatly from the stoichiometry of the desired structure. For the schematic selection of pathways shown here, yield might be improved if concentration c_b is higher than c_p,c_g as low c_p and c_g suppress incorrect and incomplete structures, respectively.

combinatorial explosion of competing structures. We then demonstrate that the principle of undesired usage can also alleviate kinetic yield catastrophes in non-equilibrium models, focusing on two models of recent interest in colloidal and DNA brick assembly. In the systems we study, the yield improvement from non-stoichiometric concentrations is typically larger when the stoichiometric concentrations yield is smaller.

Results

Undesired usage must balance desired usage. We consider self-assembly of heterogeneous structures made of n components; each component is one of m species types and has multiple distinct binding sites. Besides the desired structure, the m species of components can assemble numerous incomplete or incorrect undesired structures. (See Fig. 2 where $m = n = 4$.) Incomplete structures are pieces of the desired structure that do not have all the necessary components, whereas incorrectly bound structures contain weak 'crosstalking' interactions. Such undesired structures vary in size, shape and composition and can markedly reduce yield.

We define yield Y as the number of desired structures produced relative to all undesired structures. That is,

$$Y = \frac{X_{\rm d}}{X_{\rm d} + X_{\rm u}} \qquad (1)$$

where $X_{\rm d}(c_i)$, $X_{\rm u}(c_i)$ are the numbers of desired and undesired structures produced by assembly and which depend on species concentrations c_i. $X_{\rm u}$ can be written as a sum over all undesired structures a, $X_{\rm u} = \sum_a X_{\rm a}$.

We define the 'usage' $\upsilon_{\rm a}^i$ of species i by structure a:

$$\upsilon_{\rm a}^i \equiv \partial_{\log c_i} \log X_{\rm a}. \qquad (2)$$

$\upsilon_{\rm a}^i$ reflects how the production rate of structure a depends on the concentration of component i. For example, consider an experiment carried out using concentrations c_i and then perturb around these concentrations, $c_i \to f_i c_i$. If the perturbations are small, we can Taylor expand $\log X_{\rm a}$ to write $X_{\rm a}(f_i c_i) \approx f_1^{\upsilon_{\rm a}^1} f_2^{\upsilon_{\rm a}^2} \cdots f_m^{\upsilon_{\rm a}^m} X_{\rm a}(c_i) + \ldots$. Thus, large usage $\upsilon_{\rm a}^i > 0$ implies that increasing the concentration of species i will greatly increase the production of structure a. The gradient of yield with respect to $\log c_i$ can be written in terms of usage,

$$\vec{\nabla} \mathbf{Y} \;\propto\; \vec{\upsilon}_{\rm d} - \langle \vec{\upsilon}_{\rm a} \rangle_{\rm u} \qquad (3)$$

where the average $\langle \vec{\upsilon}_{\rm a} \rangle_{\rm u} = \sum_{\rm a} \frac{X_{\rm a}}{X_{\rm u}} \vec{\upsilon}_{\rm a}$ is over all undesired structures a, each weighted by the amount $X_{\rm a}$ of it produced. The proportionality constant in equation (3) is always positive.

This equation defines the principle of undesired usage—yield is improved by lowering the concentration of component i whose average undesired usage $\langle \upsilon_{\rm a}^i \rangle_{\rm u}$ is higher than its correct usage $\upsilon_{\rm d}^i$ (and vice-versa). Changing concentrations in such a manner produces different distributions of undesired structures $X_{\rm a}/X_{\rm u}$, whose average undesired usage $\langle \upsilon_{\rm a}^i \rangle_{\rm u}$ is closer to the correct usage $\upsilon_{\rm d}^i$. The resulting optimal concentrations can be very different from the stoichiometry in the desired structure.

Our analysis of specific models will show that the resulting yield improvements are marked. In what follows, we apply this principle to equilibrium assembly as well as to two paradigmatic kinetic examples. In our model of equilibrium assembly, c_i and $X_{\rm a}$ in equation (2) will refer to steady-state concentrations, and we analytically show that the usage $\upsilon_{\rm a}^i$ is the number of occurrences of species i in structure a. In our kinetic models, $\upsilon_{\rm a}^i$ quantifies the dependence of the final amount $X_{\rm a}$ of structure a on initial concentrations c_i that deplete with time. Thus, equation (3) can be applied to both steady-state or initial concentrations and to equilibrium or kinetic assembly.

Equilibrium assembly and yield catastrophes. We begin by studying equilibrium assembly. We assume that the interactions between binding sites in the desired structure are strong and of energy $s\, k_{\rm B}T < 0$. For example, in Fig. 3, components 6 and 7 and components 4 and J_1 have strong binding with each other. In many systems, such as DNA bricks designed with random sequences or protein assemblies, the strong interaction energy will typically vary across the structure. Optimal concentrations will depend on such variation in binding energies and can be computed using the framework we introduce below; for simplicity, we focus on the case of a single strong binding energy scale $s\, k_{\rm B}T$ and discuss generalizations in Supplementary Note 1.

In any natural or synthetic system, there is always some level of nonspecific interactions that we call 'crosstalk'[12,18,20,31–33]. To model crosstalk, we assume that all non-desired binding sites interact weakly with an energy $w\, k_{\rm B}T < 0$ that is distributed randomly as $\rho(w)$. Thus, components 6 and 2 interact through such a weak crosstalk interaction. The 'male' site of component 3 interacts weakly with all components since it is unbound in the desired structure.

We assume for now that each component i is supplied at a fixed chemical potential μ_i, so the concentration of free components has the constant value $c_i = e^{\beta\mu_i}$, with $\beta = \frac{1}{k_{\rm B}T}$. This steady-state model mimics the assembly of the ribosome and other macromolecular complexes whose protein components are being continually produced, or assembly in a large sea of components whose concentrations change very little during assembly. Recent works[10,34,35] suggest that, in some temperature regimes, the experiments of[9,10] can be described in such a manner. See kinetic models below for complementary possibilities.

Yield at equilibrium can be obtained by summing the partition function over all undesired structures; we developed a method adapted from Feynman diagrams to perform such numerical computations efficiently (see Methods section). We find that yield is determined by an energy–entropy balance, with the number of the most stable competing structures growing as $\sim n^2$ and each such crosstalk-containing structure suppressed by an energetic factor of $e^{-\beta(w-s)}$. Building structures of size n with any yield at all requires the energetic suppression to dominate which in turn implies a bound on the crosstalk energy w. Extending such analytic arguments (detailed in Supplementary Note 2 and

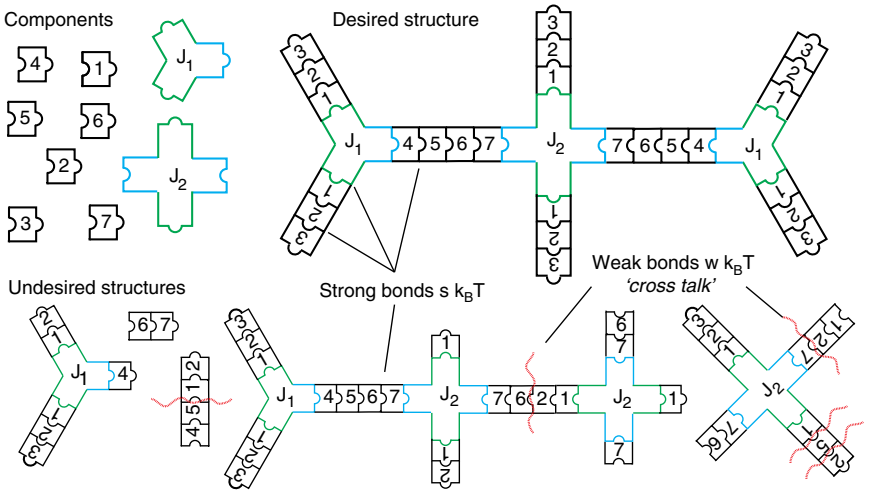

Figure 3 | Crosstalk leads to an exponential number of undesired competing structures. In our model of equilibrium assembly, desired structures of size n are built from m species of bivalent and multivalent components, supplied at chemical potentials $\mu_1,\mu_2,\ldots,\mu_m$. ($n=29, m=9$ here.) All bonds present in the millipede-like desired structure are strong (energy $s<0$). All other bonds correspond to weak 'crosstalk' interactions (wavy red lines, energy $w<0$). Undesired structures made of weak bonds are energetically suppressed by factors of $e^{-\beta(w-s)}$, but the number of such structures is exponentially large in n. Yield is determined by this energy–entropy balance.

Supplementary Fig. 5), we find $-\tilde{w} < -s + A\log n + B$, where $\tilde{w} = -\log\langle \rho(w) e^{-\beta w}\rangle$ is the exponential average of crosstalk energies w between all species, assumed to be distributed as $\rho(w)$. A and B are $O(1)$ numbers that depend on the topology of the desired structure. Note that stronger interactions correspond to more negative energies in our convention. If crosstalk strength w exceeds this bound, the different species of components are indistinguishable and the system effectively behaves as having only one species. Hence, we will restrict ourselves to crosstalk below this bound (which we call the '$\log n$ crosstalk threshold') and examine how the maximum yield depends on the size and structural complexity of the desired structures.

Equal chemical potentials lead to a yield catastrophe. For simplicity, we begin by considering structures where all components are distinct, that is, the number m of distinct species used is also the size n of the structure. We first assume that the chemical potentials of all $m = n$ species are stoichiometric, so $\mu_i = \mu$. Figure 4a shows the yield (red curve) as a function of μ, for a linear structure of size $n = 8$. Low chemical potentials favour incomplete structures, whereas high potentials favour large aggregates held together by weak bonds; yield is maximized for an intermediate value of μ.

The red data points in Fig. 4b show how this maximum yield depends on n, the size of the linear structure. To fairly compare yields for structures of different n, we need to increase the difference $g = w - s$ between strong and weak bond energies as $A\log(n) + G$, so that g stays above the $\log n$ crosstalk threshold by a fixed amount G. We set $A = 2.0$ for linear structures. Strikingly, the yield degrades exponentially with increasing n, indicating that by $n \approx 35$ the maximum yield is at most $\sim 1\%$. The yield degrades because the number of competing structures increases markedly with increasing size n; although each individual competing structure has higher energy, the combinatorial explosion of possibilities with growing n strongly limits the yield.

This yield catastrophe also occurs with increasing structural complexity. Using the Feynman method, we numerically computed the yield of branched structures for fixed n, but with increasing numbers of arms (Fig. 4c). Again, we find that the yield decreases exponentially with the number of arms; the reason is that the number of competing structures increases markedly with the number of arms (calculated in Supplementary Note 1).

The yield catastrophe presents a fundamental constraint: if the concentrations of components are stoichiometrically matched to the structure, then there is a limit to how large and how complex a structure can be robustly assembled.

Unequal chemical potentials alleviate the yield catastrophe. We now allow the components to have unequal chemical potentials μ_i. We numerically vary μ_i independently to optimize yield, using gradient descent on our Feynman construction for partition functions. Strikingly, the optimal values of μ_i are highly non-uniform across the structure, with exterior pieces having much higher μ_i (and hence concentrations) than interior pieces. Moreover, the optimal μ_is lead to a nearly complete recovery from the yield catastrophe. Figure 4 shows the effect on our two model calculations: Fig. 4b,c (black dots) show that the optimal μ_is lead to a yield that is independent of the length n of linear structures and the number of arms a in branched structures.

To understand why the optimal μ_is are unequal, we analyse undesired usage. In equilibrium, the 'usage' υ_a^i of species i by structure a, defined in equation (2), simply reduces to the number of times species i occurs in structure a (see derivation in Methods section).

Figure 5 shows the implications of usage graphically; we plot the undesired usage of components $i = 1,\ldots 8$ averaged over competing structures grouped as larger and smaller than the desired structure. With equal chemical potentials (Fig. 5a), only 60% of small undesired structures contain component **1** but almost 90% of them contain **5**. This asymmetry arises because, for example, the dominant small structures of length four are **1234**, **2345**, **3456**, **4567**, **5678** (others, such as **1278**, are suppressed by weak bonds). Component **5** occurs in most of these structures while **1** occurs in only one. Large structures have unequal usage for similar reasons.

In the desired structure, each component is only used once. Hence, increasing the supply of **1** (that is, increasing μ_1) will suppress small structures (whose usage of **1** is 0.6) relative to the desired structure while only somewhat enhancing large structures (whose usage of **1** is 1.1). The suppression of small structures dominates and yield is improved. Similarly, decreasing μ_5 will

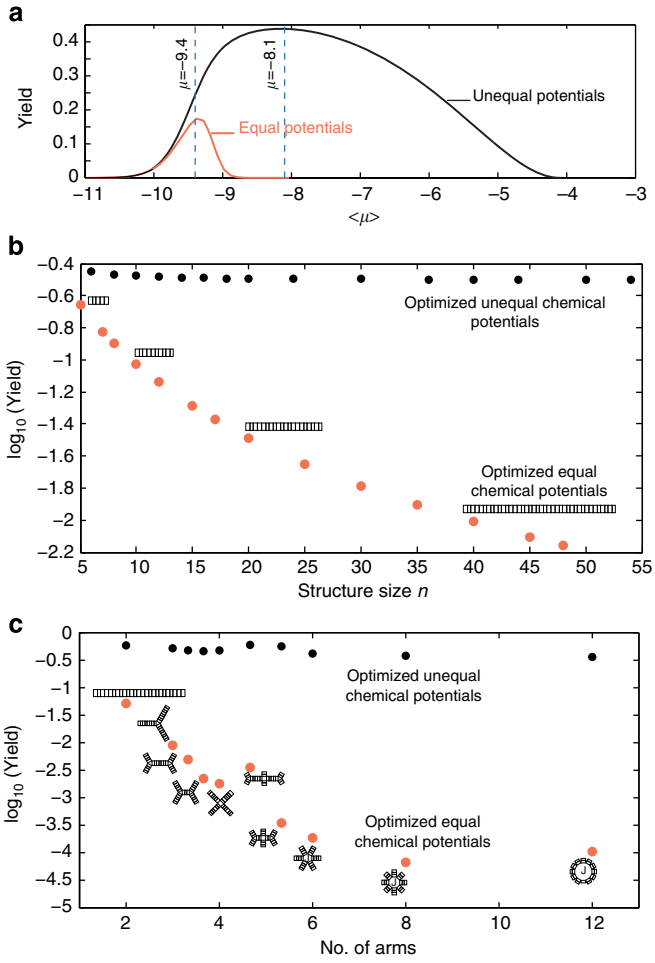

Figure 4 | Alleviation of the yield catastrophe. (**a**) Yield is maximized at an intermediate average chemical potential that favours neither large aggregates nor small incomplete structures. Yield is shown for a linear structure of size $n = m = 8$ as a function of the average chemical potential $\beta\langle\mu\rangle = \log(\sum_i c_i)/n$. (**b**,**c**) Yield falls (red) with (**b**) size and (**c**) increasing number of arms if the chemical potentials of different components are equal. The yield improvement owing to optimized unequal potentials is greater for larger and highly branched structures. The resulting optimized yield (black) is relatively independent of shape and size. ($n = m$ for all structures. In **a**, $w = -2k_BT, g = w - s = -10k_BT$. In **b**, to compare different n fairly, we scaled $w - s = 3 + 2\log n$ to keep $w - s$ a constant amount over the $2\log n$ crosstalk threshold. All shapes in **c** have $n = 25$, with $g = 12k_BT, w = -2k_BT$.)

improve yield by greatly suppressing large structures (usage ≈ 2) and somewhat enhancing small structures (usage ≈ 0.9).

Modifying chemical potentials in this way changes the distribution of competing structures produced and hence usage will have to be recomputed. The process terminates when the undesired usage of different components matches the stoichiometry in the desired structure (see equation (3) and Fig. 5b). Increasing μ_1 any further will promote large structures more than it suppresses small structures. Similarly decreasing μ_5 any further will promote small structures and harm yield. In this way, the optimal supply of components strikes a balance between different groups of competing structures.

Undesired usage analysis also explains the optimal profiles for general branched structures shown in Fig. 6; for example, the dominant competing structures for the 3-armed star in Fig. 6a consist of junction ***J*** with partially built arms. These partial arms will almost always contain component **1** and not contain component **3** unless it is a complete arm. Partial arms in which component **3** occurs without **1** involve crosstalking interactions and contribute less to the partition function. In the millipede-like structure, junction J_2 is to have lower concentration than J_1, since J_2, with four sites, is more prone to self-aggregation. See Supplementary Figs 1–3 and Supplementary Note 1 for a sampling of such competing structures, further usage analysis of structures shown here and the weak dependence of optimized yield on topology of structures in Fig. 4c.

In Supplementary Fig. 4, we show how the optimal concentration profile changes if the strong binding energy s varies across the structure; we show that our results on the yield catastrophe and its alleviation continue to hold.

In summary, we see marked yield improvement (Fig. 4) by using unequal chemical potentials as dictated by undesired usage (equation (3)). Thus, unequal potentials provide sufficient 'control knobs' to suppress competing structures whose number is exponential in size n.

Undesired usage analysis alleviates kinetic yield catastrophes. Thus far, we have analysed the mitigation of the equilibrium yield catastrophe using the principle of undesired usage. However, assembly in many systems is limited by kinetics. Unlike equilibrium yield, the details of kinetic pathways differ from system to system. Nonetheless, equation (3) shows that undesired usage analysis still applies, although the definition of usage depends on the details of the kinetics.

Here we examine complementary exemplars of kinetic issues in models of two synthetic systems—DNA brick and irreversible colloidal assemblies. In both cases, we show that non-stoichiometric concentrations markedly alleviate yield catastrophes. In the kinetic problems we study here, concentrations c_i will refer to the initial amounts of components supplied, since free components are significantly depleted over the course of assembly.

Incomplete incompatible structures. Kinetic yield catastrophes can occur even in the absence of crosstalk. Our first model assumes that a desired structure can nucleate and grow from several inequivalent seeds (Fig. 7a). If the rate of nucleation is comparable or fast compared with the rate of subsequent growth of seeds into full structures, many inequivalent seeds will nucleate and grow in parallel. Such parallel growth can lead to a kinetic yield catastrophe in the form of a 'depletion trap'; all components are locked up in incomplete incompatible structures.

The problem of incomplete structures has been argued to be central to both DNA brick assembly[9,34–36] and protein complexes[27,37]. Indeed, before the experiments of ref. 9, such depletion traps were thought to make DNA brick-like approaches unlikely to succeed. Recent simulations[34] show that in a narrow regime of parameter space, assembly is successful because nucleation is sufficiently slow compared with growth, that is, each nucleated structure typically completes growth before the nucleation of another seed. However, outside of this regime, multiple structures nucleate rapidly, deplete monomers and might incorrectly bind each other[34]. Hence, it is important to develop strategies that expand the range of conditions for successful assembly; other materials, assembly conditions and larger target structures may not satisfy the criterion of slow nucleation and fast growth.

Undesired usage analysis implies that unequal concentrations can alleviate depletion traps. We first demonstrate this in a simple nucleation-and-growth model of a ring (Fig. 7a), where the ring is made of n bivalent components that bind only to their correct partners. We assume that the interactions are cooperative (for example, through allostery) so that the ring grows from critical nucleating seeds of size a or greater: structures smaller than size a

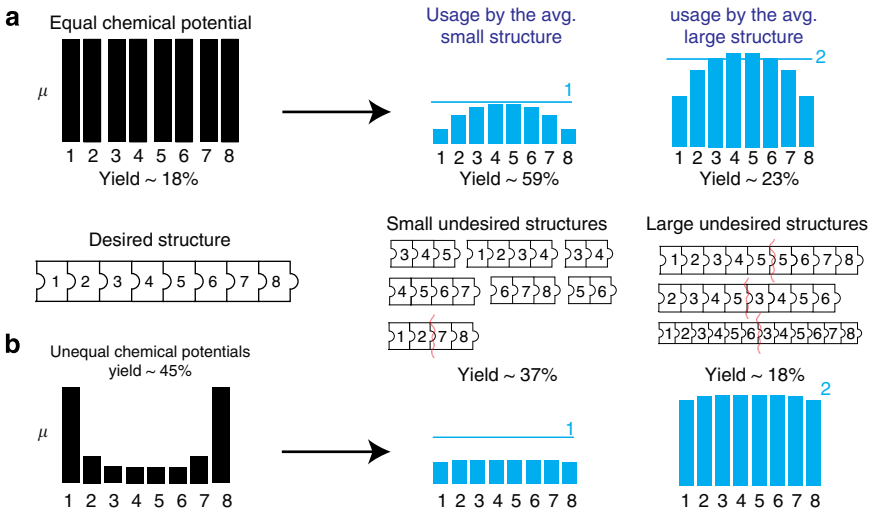

Figure 5 | Exploiting undesired usage. (**a**) Interior components have higher undesired usage than desired usage (and vice-versa for exterior components), when assembly is carried out with equal chemical potentials. For example, the undesired usage of **5** by large structures is higher (≈ 2) than its desired usage ($=1$). (**b**) Hence, as dictated by undesired usage (equation (3)), decreasing μ_5 and increasing μ_1 suppresses large and small structures relative to the desired structure, respectively. At the optimal U-shaped profile shown in **b**, the undesired usage of all components matches usage in the desired structure. Avg., average.

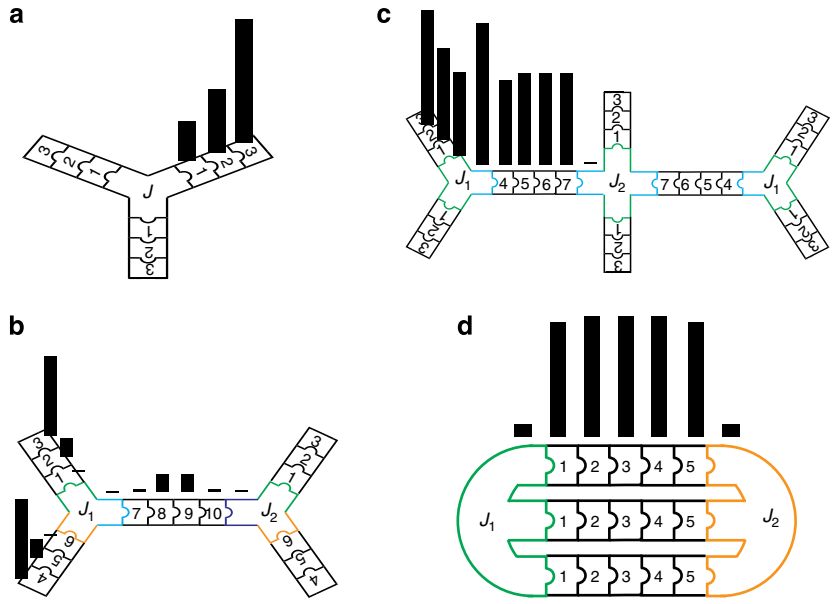

Figure 6 | Optimal chemical potential profiles. Highly non-stoichiometric potentials optimize yield for (**a**) a star, (**b**) an H, (**c**) a millipede-like and (**d**) a double-looped Θ-like structure; for example, $c_1 \sim \frac{1}{100} c_3$ in **a**. Yield improvement over using equal potentials (with the same average concentration of free components) is $\approx 2\times$, $\approx 36\times$, $2{,}800\times$ and $1.8\times$ for the star, H, millipede and Θ structures, respectively. (Numeric details in Supplementary Note 1.)

quickly dissolve into monomers, but a nucleating seed of size a grows irreversibly[30]. Assembly starts with an initial supply of components of concentration $c_i, i = 1,\ldots,n$ that is depleted over time. We compute yield by numerically solving the master equation with initial conditions c_i (see Methods section).

With equal concentrations, we find that the depletion trap strongly degrades yield for large n; this degradation occurs because the nucleation rate of seeds is independent of structure size n while the subsequent growth rate of the seed into the full structure falls as $1/n$. Further, the number of parallel nucleation pathways increases as n. Hence, yield falls exponentially with increasing structure size n, giving a kinetic yield catastrophe (Fig. 7b).

With unequal concentrations, this depletion trap is strongly alleviated. Through gradient descent optimization, we find that the optimal concentration profile is highly non-uniform (Fig. 7c), with high concentrations for a region of size $\sim a$ and low concentrations everywhere else. (By symmetry, this high concentration region can be anywhere around the ring.) The unequal concentrations greatly enhance nucleation for one of the many assembly parallel paths shown in Fig. 7a, suppressing the rest. Hence, even with slow growth, the components needed for growth are not significantly depleted by competing pathways, with a marked effect on yield (red dots in Fig. 7b). We validated these same conclusions in a 2-dim 'P'-shaped structure (Fig. 7d), made of $n = 208$ tiles and resembling DNA brick assemblies[9].

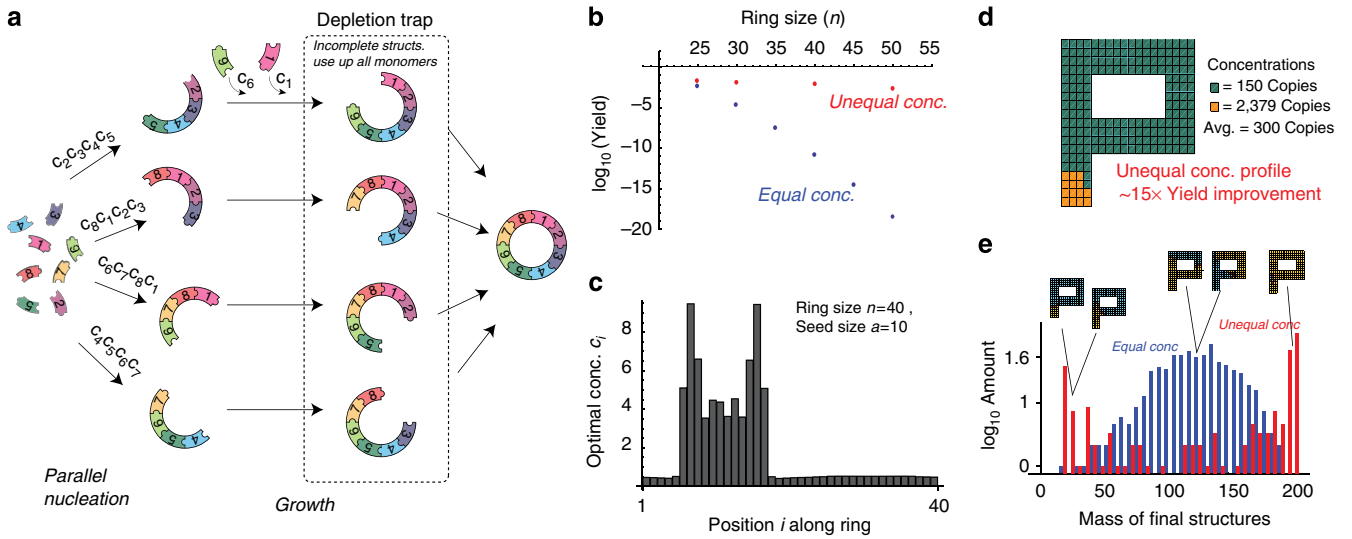

Figure 7 | Non-stoichiometric concentrations can avoid depletion traps created by rapid parallel nucleation of distinct seeds. (**a**) If nucleation is faster than subsequent growth into full rings, all free components can be locked up in the set of incomplete structures shown in the dotted box. These structures cannot grow further since they lack free components and the ability to bind each other. (**b**) Depletion traps are more severe for larger structure sizes n, resulting in a kinetic yield catastrophe; yield with equal concentrations falls exponentially with n (blue dots). (**c**) The unequal concentrations shown for $n = 40$ selectively choose one nucleation path in **a** over others. Hence, the depletion trap is mitigated and yield improvement is marked (red dots in **b**). (**d**) Yield of a 2-dim P-structure made of 208 distinct species of tiles, resembling DNA brick assemblies[9], is also improved by unequal concentrations. (**e**) With equal concentrations, a depletion trap results in incomplete structures of typical size ~100–150 (blue bars). With the unequal profile in **d**, the depletion trap is resolved; final structures are either of full size 208 or remain small seeds of size ~20 (red bars).

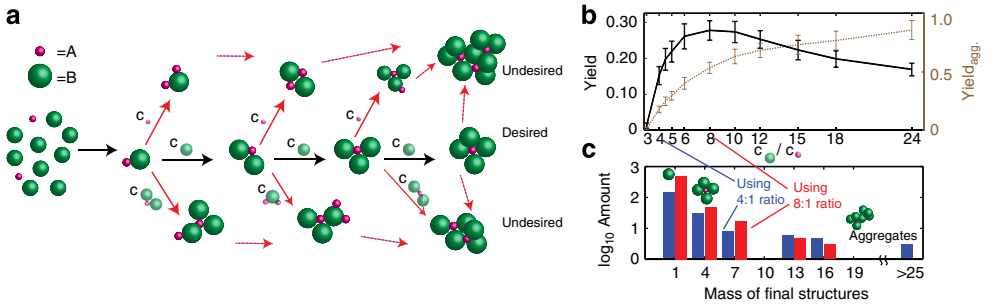

Figure 8 | Non-stoichiometric concentrations can suppress undesired aggregation in irreversible colloidal assembly. (**a**) In the experiments of[38], incomplete structures can incorrectly glue to each other (red arrows), greatly reducing yield. However, red pathways leading to undesired aggregates (agg.) have higher usage of particle A than the black pathway leading to the desired AB_4 structure. Hence, a high ratio of initial concentrations c_B/c_A markedly increases yield; incomplete structures are rapidly completed by B particles along the black path, with low probability of taking any red branches. (**b**) Using a stochastic simulation of these pathways, we find that yield peaks at $c_B{:}c_A = 8{:}1$. (**c**) If $c_B{:}c_A = 4{:}1$ (stoichiometric ratio) instead, the distribution of assembled structures is shifted towards larger aggregates (blue bars) and hence yield is lower. In **b**, we also plot an alternative notion of yield, $\text{yield}_{\text{agg.}}$, relevant to the experiments in ref. 38, which considers only large aggregates as undesirable and disregards excess A or B particles, which are easily washed out. $\text{Yield}_{\text{agg.}}$ increases indefinitely with c_B/c_A.

We assume that each tile in the structure is a unique species with four distinct binding sites that bind only to their correct partners. We simulated a nucleation-and-growth model similar to the ring above (see Methods section); seeds of size $a = 18$ could nucleate anywhere in the structure and then grow irreversibly, with the rates of both processes determined by the concentrations of components.

Starting assembly with equal amounts (300 copies) of each species, we find a severe depletion trap, resulting in a wide distribution of incomplete structures of typical size ~100–150 tiles (blue bars in Fig. 7e). Only nine structures are of size within 15% of the complete $n = 208$ P-structure. On the other hand, the highly unequal profile (with the same average number 300 as earlier) shown as a heatmap in Fig. 7d, gives a yield of ~135 structures, a 15-fold improvement.

Incomplete sticky structures. Colloidal particles form aggregates easily because they are isotropically sticky and pick up multiple partners, making it difficult to build predictable finite structures. Such a yield catastrophe was recently observed[38] and mitigated using highly non-stoichiometric concentrations. Finite clusters of type AB_4 were built from two kinds of DNA-coated colloidal particles A and B, designed so as to bind each other irreversibly but not bind their own kind. The radii of A and B were chosen so that exactly four Bs bind to each A. Figure 8a shows a selection of kinetic pathways leading to the desired AB_4 structure (black) and to several undesired aggregates (red). The main hurdle to high assembly yield is ensuring that incomplete structures (that is, AB, AB_2 and AB_3) bind only to particle B and not to other incomplete structures, A, or larger aggregates. The experiments[38] found large aggregates and a very low yield of AB_4 if assembly starts with the

stoichiometric ratio 4:1 of B:A. On the other hand, supplying a large excess of B greatly enhanced the yield of AB_4.

We were able to reproduce this behaviour with the stochastic simulation of a simple model of irreversible aggregation (see Methods section). Yield is maximized at the highly non-stoichiometric ratio of 8:1 for B:A (Fig. 8b). Figure 8c (red bars) shows that the 8:1 supply shifts the distribution of assembled structures towards lower mass, compared with a 4:1 supply (blue bars). An excess of *B* suppresses the probability of taking the red pathways, relative to the correct black pathway (Fig. 8a). The depletion of free A particles with time helps further suppress undesired pathways—otherwise, the desired structure AB_4 can pick up free As and aggregate. Hence, a high ratio of B:A helps rapidly complete incomplete structures through addition of monomers and prevent them from sticking to each other. The results here generalize Flory's classic theory of irreversible condensation[39] that involved only a small number of species.

These complementary examples from DNA, protein and colloidal assembly demonstrate that undesired usage analysis can greatly enhance yield in examples where assembly is kinetically controlled. Unlike our equilibrium examples, the character of kinetic yield catastrophes depends on modelling assumptions; for example, numerical prefactors in nucleation and growth rates, reversibility of growth and so on, which vary across different systems. Nevertheless, using unequal concentrations markedly alleviates kinetic yield catastrophes, especially when kinetic traps are strong. We leave a study of unequal concentrations using kinetic models tailored to particular systems such as colloids or DNA brick assemblies (for example, along the lines of refs 25,34) to future work.

Discussion

We have studied strategies to suppress undesired structures that can catastrophically reduce the assembly yield of complex heterogeneous structures. The central lesson of our work is that the supply of different components should account for their 'usage' by undesired structures. The resulting optimal concentrations can differ greatly from the stoichiometry of the desired structure itself.

We have shown that undesired usage analysis is a useful consideration to apply to a range of assembly experiments involving DNA, colloids and proteins[9,10,14,15,40], in both equilibrium and highly kinetic conditions, even if the precise degree of kinetic yield improvement is hard to predict without knowledge of the detailed kinetics. We leave the extensions to more complex kinetic models[41–43] such as hierarchical assembly to future work. Our framework and results have precedence in earlier work on the impact of concentrations in specific models of micelles[44], polymer condensation[39] and protein complexes[37]. Chen and Kao[28] studied a DNA tile model with crosstalk and found that optimal concentrations are proportional to the square root of stoichiometric ratios. They assume that correct bonds are irreversible. In contrast, we used a fully reversible crosstalk model that describes intrinsically more error-prone assembly and showed that even catastrophically low yields can be restored by non-stoichiometric concentrations. As a result, while our results are similar in spirit, yield improvements found in ref. 28 are weaker than those reported here. Our results also differ because our physical partition function treatment accounts for all possible erroneous structures while ref. 28 used a heuristic local definition of errors.

The principle of undesired usage can be generalized to other control parameters beyond concentrations. For example, unequal binding energies are expected to alleviate yield catastrophes in our linearly branched models and in depletion trap models[27]. These results differ from earlier analyses[26] that only considered competing structures of the same size as the desired structure (that is, not bulk assembly). The design of structure itself[45] can be subject to undesired usage analysis; given a physical structure, how should it be composed of different species to minimize assembly of competing structures? Finally, in practice, some structures might be more undesirable than others, owing to their being more difficult to separate or more toxic in a cell. Our framework can be generalized to accommodate such varying functional costs (Supplementary Note 4), providing a way to produce the least undesirable mix of structures appropriate in a given functional context.

Biology has many examples of heterogeneous multiprotein/RNA complexes[1], ranging from simple linear complexes to ribosomes and the nuclear pore complex. Our work predicts that optimal expression levels of components of large protein complexes can be highly non-uniform; for example, if a protein component is particularly likely to form toxic aggregates, that protein should be expressed at lower levels than other complex components.

Phrased this way, the implications of our work seem intuitive and perhaps even obvious—and yet they seemingly contradict a common formulation of the DBH[29,46–48]. DBH states that the optimal expression levels for complex-forming proteins is stoichiometric and that deviations have large fitness costs. Such fitness costs are believed to be a strong evolutionary constraint on gene duplication events, aneuploidy, gene family sizes and dominance.

Instead, our results suggest that the optimal expression level baseline may differ greatly from stoichiometry due to undesired structures. However, the rest of DBH—that changes from this optimal level have large fitness costs—would still apply about this modified baseline.

We note that all the strong evidence for DBH involves changes in expression levels[29,46]; such evidence includes experiments that enhance or repress expression of select proteins and statistical studies of gene duplication events. These data say little about what the optimal ratios are in the first place.

In fact, the database consensus for expression levels of proteins involved in complexes is highly uneven across those complexes (ref. 49 and Supplementary Note 4 and Supplementary Fig. 6). Our proposal also finds support in the work of refs 20,31, which found that highly promiscuous proteins in yeast have lower expression levels. The production of proteins comes at a cost (energy, material and time) to the cell and highly unequal production of proteins needed in equal amounts seems wasteful, unless some other benefit were conferred[49,50]. However, expression levels need to be measured more accurately to draw further conclusions.

Finally, it is intriguing to ask how the organization of operons—sets of bacterial genes whose expression is regulated together—is related to potential undesired complexes[51,52]. There is evidence for such connections between genome structure and the physics of protein complexes in other contexts[53,54]. We leave a detailed study of these questions to future work.

Methods

Equilibrium assembly and Feynman expansion. Our model has m species of multivalent particles with distinguishable binding sites as shown in Fig. 3. The desired structure defines the interaction energy between pairs of binding sites on two different particles. If a pair of sites is bound in the desired structure, the binding is strong and of energy sk_BT; else, the binding energy is wk_BT (a weak 'crosstalk' interaction). Each species is supplied at a steady chemical potential μ_i (or equivalently concentration $c_i = e^{\beta\mu_i}$ where $\beta = \frac{1}{k_BT}$. We define $g = w - s$.

At equilibrium, yield can be written in terms of partition functions of desired and undesired structures; in equation (1), $X_d = Z_d$ and $X_u = \sum_a X_a = \sum_a Z_a$ where Z_a is the partition function of structure a,

$$Z_a = e^{-\beta(ps+qg-\sum_{k\in a}\mu_{i_k})} \quad (4)$$

where p is the total number of bonds in a, q is the number of crosstalking bonds and $g = w - s > 0$ is the difference between the strong and weak bond energies. $\sum_{k\in a}\mu_{i_k}$ is the sum of chemical potentials μ_{i_k} over all particles k (of species type i_k) present in structure a.

Since $X_a = Z_a$ at equilibrium and $c_i = e^{\beta\mu_i}$, equation (4) implies that usage $v_a^i \equiv \partial_{\log c_i} \log X_a$ at equilibrium is simply given by the number of occurrences of species i in structure a.

Feynman method. Finding the yield requires summing the partition function $X_u = \sum_a X_a = \sum_a Z_a$ over all competing structures a of varying shapes and sizes as shown in Fig. 3; the list can be quite large even for a simple structure. We markedly simplified such tedious calculations by developing a perturbative method for computing the partition functions of linearly branched and/or looped structures based on rules adapted from Feynman diagrams. Feynman methods have been used before in the computation of partition functions of polymers[55]. In our context, Feynman rules give us a one-step method of summing over structures of all sizes that are consistent with a given topology. The Feynman rules here are: (1) A linear segment made of any number of bivalent components between two junctions is associated with the matrix D_i^j.

$$D_i^j = c_i \left(\frac{1}{\mathbb{I} - B}\right)_{ij} \text{ where } B_{ij} = e^{-\beta E_{ij}} c_j \quad (5)$$

where E_{ij} is the interaction energy between species i and j, binding in the order $i-j$, and c_j are the concentrations of components. (2) Every binding site p of a junction J is associated with a row vector $v_{J,p}^i = e^{-\beta E_{pi}}$ (column vector if site p is female) where E_{pi} is the binding energy of p to bivalent species i. In addition, each junction J is associated with a concentration factor c_J. 3. Free ends of linear segments are associated with a special vector $\vec{f}$ that can account for solvent–component interactions. We take $\vec{f} = (1, 1, \ldots, 1)$ in this paper.

To find the partition function summed over all structures with a given topology, we first carry out matrix multiplication of the row and column vectors associated with junctions (or free ends) and the matrix D_i^j associated with the linear segment between such junctions, giving a scalar factor of the type $\vec{v} \cdot D \cdot \vec{u}$. The total partition function is the product of all such scalar factors associated with a given topology.

The Feynman method is particularly efficient when the number of bivalent species is large, as long as the number of species with valence larger than two stays small. Supplementary Fig. 1 and Supplementary Note 1 contain detailed derivations and examples of applying these Feynman rules.

Kinetic depletion trap. *Ring.* The ring structure is modelled as made of n bivalent particles, each of a distinct species. The left binding site on species $i-1$ will bind only to the right binding site of i; for example, in Fig. 7a, the left site of species 5 will only bind the right site of 6. Correct binding is irreversible and no incorrect binding is allowed. Particle n is assumed to bind to 1, allowing assembly of a ring.

We denote concentration of species i by positive real numbers $C_i(t)$, $i = 1 \ldots n$. The concentration of a clockwise segment of the ring from i to j is denoted $X_{ij}(t)$; we assume modular (mod n) arithmetic for indices. Structures of size less than a critical nucleation size a are assumed to quickly dissolve back into monomers. Structures of size exactly a are created through nucleation at a rate k_n. Additional monomer are added to the structure at the two ends with rate k_g. For simplicity we assume that all the correct partners bind with the same rate k_g. The master equation for $X_{i,\, i+a-1}(t)$ is:

$$\begin{aligned}\partial_t X_{i,i+a-1} =& k_n C_i C_{i+1} \cdots C_{i+a-1} \\ & - k_g C_{i-1} X_{i,i+a-1} - k_g X_{i,i+a-1} C_{i+a}\end{aligned} \quad (6)$$

Structures of size larger than critical size a and smaller than n grow by picking up monomers on either side, that is, for $n-1 > j-i > a-1$,

$$\partial_t X_{ij} = -k_g C_{i-1} X_{ij} - k_g X_{ij} C_{j+1} + k_g C_i X_{i+1,j} + k_g X_{i,j-1} C_j \quad (7)$$

Structures of size n are assumed to close up and form full rings, which are stable and inert.We numerically solved these equations in Mathematica with fixed values for different initial concentrations $C_i(t = 0)$ and found the value of $X_{ij}(t)$ at large times when no further changes occur. We varied n between 25 and 50 with fixed $a = 10$. Supplementary Note 3 contains numerical details.

P-shaped structure. We carried out discrete-time stochastic simulations for nucleation and growth of the P-shaped structure, which is made of $n = 208$ distinct square tiles; each tile has four distinct binding sites that bind only to their correct partners in Fig. 7(d). We start the simulation with an integer $c_i(t = 0)$ copies of species i. At each discrete time step, we randomly choose to either nucleate a new seed of size $a = 18$ or grow existing seeds. The probability of nucleating a seed S of critical size $a = 18$ is $\frac{1}{\lambda_N}\prod_{j\in S} c_j(t)$ where the product is over the $a = 18$ species $j \in S$ found in the seed S and λ_N is a normalization constant. The probability of a pre-existing seed growing by picking up the correct tile of species i at a boundary site is $\frac{1}{\lambda_G} c_i(t)$. We reduce the numbers $c_i(t+1) = c_i(t) - 1$ for species i, which participate in nucleation or growth of seeds and run the simulation until no further nucleation or growth takes place (that is, tiles are fully depleted).

The ratio of the constants λ_N, λ_G in the above probabilities of nucleation and growth is a free parameter that is not set by normalization and determines the relative speed of nucleation and growth processes. We report results for rapid nucleation (that is, a high ratio of λ_N to λ_G), a limit in which depletion traps are severe. See Supplementary Note 3 for more details of the above Gillespie algorithm, including normalization and numeric details.

Kinetic aggregation of colloids. In our discrete-time stochastic simulations, we begin with a mixture of c_A particles of type A and c_B particles of type B such that c_A,c_B are positive integers with $c_A + c_B = 1{,}000$. Any structure composed of i_A A and i_B B particles is said to be of type $i = (i_A, i_B)$. At each time step, we pick two random structures (including free particles), uniformly and without replacement, out of the mixture; let us say they are of type $i = (i_A, i_B)$ and type $j = (j_A, j_B)$. Then, we glue them together—producing a new structure of type $k = (i_A + j_A, i_B + j_B)$—with a probability given by a kernel $K(i, j) \in [0,1]$ that depends on the mass and composition of the two structures. Hence, reactions (that is., gluing) between structures of type i and j happen at a rate $c_i c_j K(i, j)$, where c_i, c_j are the numbers of the type i and j structures. The two original structures of type i and j are removed from the solution. Kernel $K(i, j)$ implements the rule that A and B only bind to each other and not to themselves. Note that as an approximation, we track structures only by their overall composition (that is, numbers of A and B) and do not track the precise arrangement of the A, B particles within the structure. This approximation and corresponding details of the kernel $K(i, j)$ are described in Supplementary Note 3. We evolve the system in discrete time steps until no further gluing occurs, giving a final mix of structures that cannot react any further.

References

1. Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.* **2,** e155 (2006).
2. Mizushima, S. & Nomura, M. Assembly mapping of 30S ribosomal proteins from *E. coli*. *Nature* **226,** 1214–1218 (1970).
3. Talkington, M. W. T., Siuzdak, G. & Williamson, J. R. An assembly landscape for the 30S ribosomal subunit. *Nature* **438,** 628–632 (2005).
4. Mulder, A. M. *et al.* Visualizing ribosome biogenesis: parallel assembly pathways for the 30S subunit. *Science* **330,** 673–677 (2010).
5. Trylska, J., McCammon, J. A. & Hamacher, K. a. y. Dependency map of proteins in the small ribosomal subunit. *PLoS Comput. Biol.* **2,** e10 (2006).
6. Mirkin, C. A., Letsinger, R. L., Mucic, R. C. & Storhoff, J. J. A DNA-based method for rationally assembling nanoparticles into macroscopic materials. *Nature* **382,** 607–609 (1996).
7. Fu, T. J. & Seeman, N. C. DNA double-crossover molecules. *Biochemistry* **32,** 3211–3220 (1993).
8. Seeman, N. C. Nucleic acid junctions and lattices. *J. Theor. Biol.* **99,** 237–247 (1982).
9. Wei, B., Dai, M. & Yin, P. Complex shapes self-assembled from single-stranded DNA tiles. *Nature* **485,** 623–626 (2012).
10. Ke, Y., Ong, L. L., Shih, W. M. & Yin, P. Three-dimensional structures self-assembled from DNA bricks. *Science* **338,** 1177–1183 (2012).
11. Winfree, E., Liu, F., Wenzler, L. A. & Seeman, N. C. Design and self-assembly of two-dimensional DNA crystals. *Nature* **394,** 539–544 (1998).
12. Winfree, E. *Algorithmic Self-Assembly of DNA* (California Institute of Technology, 1998).
13. Adleman, L. M. Towards a Mathematical Theory of Self-Assembly. Technical Report No. 00-722 (University of Southern California, USA, 2000).
14. King, N. P. *et al.* Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **336,** 1171–1174 (2012).
15. Lai, Y.-T., King, N. P. & Yeates, T. O. Principles for designing ordered protein assemblies. *Trends Cell Biol.* **22,** 653–661 (2012).
16. Biancaniello, P. L., Kim, A. J. & Crocker, J. C. Colloidal interactions and self-assembly using DNA hybridization. *Phys. Rev. Lett.* **94,** 058302 (2005).
17. Segel, L. A. & Perelson, A. S. Shape space: an approach to the evaluation of cross-reactivity effects, stability and controllability in the immune system. *Immunol. Lett.* **22,** 91 (1989).
18. Johnson, M. E. & Hummer, G. Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks. *Proc. Natl Acad. Sci. USA* **108,** 603–608 (2011).
19. Laub, M. T. & Goulian, M. Specificity in two-component signal transduction pathways. *Annu. Rev. Genet.* **41,** 121–145 (2007).
20. Zhang, J., Maslov, S. & Shakhnovich, E. I. Constraints imposed by non-functional protein--protein interactions on gene expression and proteome size. *Mol. Syst. Biol.* **4,** 210 (2008).

21. Halverson, J. D. & Tkachenko, A. V. DNA-programmed mesoscopic architecture. *Phys. Rev. E* **87,** 062310 (2013).
22. Tkachenko, A. V. Theory of programmable hierarchic self-assembly. *Phys. Rev. Lett.* **106,** 255501 (2011).
23. Winfree, E. & Bekbolatov, R. Proofreading tile sets: error correction for algorithmic self-assembly. *DNA Comput.* **2943,** 126–144 (2004).
24. Fujibayashi, K., Zhang, D. Y., Winfree, E. & Murata, S. Error suppression mechanisms for DNA tile self-assembly and their simulation. *Nat. Comput.* **8,** 589–612 (2009).
25. Hedges, L. O., Mannige, R. V. & Whitelam, S. Growth of equilibrium structures built from a large number of distinct component types. *Soft Matter* **10,** 6404–6416 (2014).
26. Hormoz, S. & Brenner, M. P. Design principles for self-assembly with short-range interactions. *Proc. Natl Acad. Sci. USA* **108,** 5193–5198 (2011).
27. Deeds, E. J., Bachman, J. A. & Fontana, W. Optimizing ring assembly reveals the strength of weak interactions. *Proc. Natl Acad. Sci. USA* **109,** 2348–2353 (2012).
28. Chen, H.-L. & Kao, M.-Y. in *Proceedings of the 16th International Conference on DNA Computing and Molecular Programming* 13–24 (Heidelberg, 2010).
29. Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl Acad. Sci. USA* **109,** 14746–14753 (2012).
30. Adleman, L., Cheng, Q., Goel, A., Huang, M.-D. & Wasserman, H. a. l. in *Sixth International Conference on Difference Equations and Applications* (Taylor and Francis, 2001).
31. Heo, M., Maslov, S. & Shakhnovich, E. Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proc. Natl Acad. Sci. USA* **108,** 4258–4263 (2011).
32. Milenkovic, O. & Kashyap, N. On the design of codes for DNA computing. *Coding Cryptogr.* **3969,** 100–119 (2006).
33. Sacanna, S., Irvine, W. T. M., Chaikin, P. M. & Pine., D. J. Lock and key colloids. *Nature* **464,** 575–578 (2010).
34. Reinhardt, A. & Frenkel, D. Numerical evidence for nucleated self-assembly of DNA brick structures. *Phys. Rev. Lett.* **112,** 238103 (2014).
35. Gothelf, K. V. LEGO-like DNA structures. *Science* **338,** 1159–1160 (2012).
36. Rothemund, P. W. K. & Andersen, E. S. Nanotechnology: the importance of being modular. *Nature* **485,** 584–585 (2012).
37. Bray, D. & Lay, S. Computer-based analysis of the binding steps in protein complex formation. *Proc. Natl Acad. Sci. USA* **94,** 13493–13498 (1997).
38. Schade, N. B. *et al.* Tetrahedral colloidal clusters from random parking of bidisperse spheres. *Phys. Rev. Lett.* **110,** 148303 (2013).
39. Flory, P. J. Molecular size distribution in linear condensation polymers. *J. Am. Chem. Soc.* **58,** 1877–1885 (1936).
40. Rothemund, P. W. K. Folding DNA to create nanoscale shapes and patterns. *Nature* **440,** 297–302 (2006).
41. Haxton, T. K. & Whitelam, S. Do hierarchical structures assemble best via hierarchical pathways? *Soft Matter* **9,** 6851–6861 (2013).
42. Whitelam, S. & Jack., R. L. The Statistical Mechanics of Dynamic Pathways to Self-assembly. arXiv (2014).
43. Whitelam, S., Schulman, R. & Hedges, L. Self-assembly of multicomponent structures in and out of equilibrium. *Phys. Rev. Lett.* **109,** 265506 (2012).
44. Leibler, L., Orland, H. & Wheeler, J. C. Theory of critical micelle concentration for solutions of block copolymers. *J. Chem. Phys.* **79,** 3550 (1983).
45. Adleman, L. e. n. *et al.* in *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing* 23–32 (New York, NY, USA, 2002).
46. Papp, B., Pál, C. & Hurst, L. D. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424,** 194–197 (2003).
47. Veitita, R. A., Bottani, S. & Birchler, J. A. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends Genet.* **24,** 390–397 (2008).
48. Veitita, R. A. Nonlinear effects in macromolecular assembly and dosage sensitivity. *J. Theor. Biol.* **220,** 19–25 (2003).
49. Warner, J. R. The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.* **24,** 437–440 (1999).
50. Warner, J. R. & McIntosh, K. B. How common are extraribosomal functions of ribosomal proteins? *Mol. Cell.* **34,** 3–11 (2009).
51. Dean, D. & Nomura, M. Feedback regulation of ribosomal protein gene expression in Escherichia coli. *Proc. Natl Acad. Sci. USA* **77,** 3590–3594 (1980).
52. Nomura, M., Yates, J. L., Dean, D. & Post, L. E. Feedback regulation of ribosomal protein gene expression in Escherichia coli: structural homology of ribosomal RNA and ribosomal protein mRNA. *Proc. Natl Acad. Sci. USA* **77,** 7084–7088 (1980).
53. Callahan, B., Thattai, M. & Shraiman, B. I. Emergent gene order in a model of modular polyketide synthases. *Proc. Natl Acad. Sci. USA* **106,** 19410–19415 (2009).
54. Marsh, J. A. *et al.* Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell* **153,** 461–470 (2013).
55. de Gennes, P. G. Statistics of branching and hairpin helices for the dAT copolymer. *Biopolymers* **6,** 715–729 (1968).
56. Dumas, J. J. Crystal structure of the GpIb -thrombin complex essential for platelet aggregation. *Science* **301,** 222–226 (2003).

Acknowledgements

We thank Miranda Holmes-Cerfon, Miriam Huntley, Sarah Kostinksi, Nicholas Schade, Eugene Shakhnovich, William Shih, Tom Witten, Zorana Zervavcic and members of the Brenner group for their helpful discussions. This research was funded by the National Science Foundation through the Harvard Materials Research Science and Engineering Center (DMR-0820484, DMR-1435964), 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.

Author contributions

A.M., J.Z. and M.P.B. designed the study; A.M. carried out the calculations and analysis; and A.M. and M.P.B. wrote the manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Murugan, A. *et al.* Undesired usage and the robust self-assembly of heterogeneous structures. *Nat. Commun.* 6:6203 doi: 10.1038/ncomms7203 (2015).