nature communications

Article

https://doi.org/10.1038/s41467-023-35803-x

Uncovering the mechanism for aggregation in repeat expanded RNA reveals a reentrant transition

Received: 31 May 2022

Accepted: 29 December 2022

Published online: 19 January 2023

Check for updates

Ofer Kimchi [1] ✉, Ella M. King [2] & Michael P. Brenner [2,3]

RNA molecules aggregate under certain conditions. The resulting condensates are implicated in human neurological disorders, and can potentially be designed towards specified bulk properties in vitro. However, the mechanism for aggregation—including how aggregation properties change with sequence and environmental conditions—remains poorly understood. To address this challenge, we introduce an analytical framework based on multimer enumeration. Our approach reveals the driving force for aggregation to be the increased configurational entropy associated with the multiplicity of ways to form bonds in the aggregate. Our model uncovers rich phase behavior, including a sequence-dependent reentrant phase transition, and repeat parity-dependent aggregation. We validate our results by comparison to a complete computational enumeration of the landscape, and to previously published molecular dynamics simulations. Our work unifies and extends published results, both explaining the behavior of CAG-repeat RNA aggregates implicated in Huntington's disease, and enabling the rational design of programmable RNA condensates.

RNA molecules form structures through base-pairing interactions between complementary regions. Frequently, a given region of an RNA molecule will be complementary both to another region on the same molecule as well as to a different RNA molecule. How is the competition between forming intra- and inter-molecular contacts decided?

Predicting the outcome of this competition is a major open question, affecting a wide swath of both in vivo and in vitro phenomena. The effects of this competition are particularly stark in the context of biological condensates, in which RNA–RNA interactions play a major, largely understudied, role[1–6]. While typical condensates often involve RNA–protein contacts, purely RNA-based aggregation phenomena have been observed both in vitro and in vivo for certain transcripts associated with repeat expansion disorders[7].

The expansion of repeats in certain sections of DNA has been implicated in a significant number of (primarily) neurodegenerative disorders including Huntington's disease, myotonic dystrophy, and Fragile X syndrome[8–10]. While the proximate cause of many of these disorders is the effect of the expansion on the protein sequence, these expansions can lead to effects at the level of the RNA as well[11–17], including an aggregation transition[7,18]. In particular, RNA containing **CAG** or **CUG** repeats were found by Jain & Vale to phase separate depending on the number of repeats present in each molecule, led by **GC** stickers binding to one another[7]. Since all **GC** stickers are self-complementary, it is not immediately clear what leads RNA molecules in certain parameter regimes to form inter- vs. intra-molecular contacts at different rates. Aggregation was observed when the number of repeats per strand exceeded ~30, roughly the same number of repeats leading to diseases in humans[7]. This phenomenon was also observed and further studied in molecular dynamics (MD) simulations of the system by Nguyen et al.[19]. These simulations were able to explore the molecular details of the aggregation transition, at the cost of each simulation (at a different concentration or number of repeats per strand) requiring ~3 months of supercomputer time.

[1]Lewis-Sigler Institute, Princeton University, Princeton, NJ 08544, USA. [2]Physics Department, Harvard University, Cambridge, MA 02138, USA. [3]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. ✉e-mail: okimchi@princeton.edu

Current models are insufficient to explore the properties of the aggregation transition demonstrated by these studies. State-of-the-art models of associative polymers either do not include a competition between intra- and inter-molecular binding (as is more natural for rigid proteins and for heterotypic interactions) or (erroneously) assume it has no qualitative effects on the resulting system[20–22]. While intra-chain interactions are typically ignored, exceptions do exist. These include Dobrynin's 2004 study extending the Flory–Stockmayer approach to include intra-chain associations[23], and a recent publication by Weiner et al. which found that self-bonds play a crucial role in determining phase behavior in a lattice system with heterotypic binding motifs of varying lengths[24].

Here, we derive an analytical model to describe a system of polymers with self-complementary stickers. Eschewing mean-field-theory approaches that have dominated the field, we employ a multimerization-based framework that predicts the entire multimerization landscape in addition to the phase behavior, and thus naturally and explicitly considers the competition between intra- and inter-molecular contacts[25]. Quantitative consideration of this competition reveals that configurational entropy, arising from the multiplicity of ways to form bonds, is the driving force for aggregation in this system. Mapping out the complete phase diagram, we find that as a result of the competition between intra- and inter-molecular bonds, the system exhibits a tunable reentrant phase transition as a function of sequence or temperature. With very strong stickers (or low temperatures) the polymers fold into stable monomers and dimers, and are more likely to form aggregates at intermediate sticker strengths. We furthermore find that, for long enough linkers that enable adjacent stickers to bind, the parity of the number of stickers per strand affects not only the dimerization transition but the large-scale aggregation behavior as well. We validate our results by comparing them to a computational model that enumerates the complete landscape of intra- and inter-molecular structures that the RNA can form, and by comparing them to the results of the Jain & Vale and Nguyen et al. studies[7,19]. Our work provides a unified framework to explain both dimerization and aggregation phenomena in **CAG** repeat systems[17,19] and extends these to arbitrary sequences, temperatures, and concentrations, thus setting the stage for the construction of novel materials and new techniques based on programmable RNA condensates.

Results

Equilibrium behavior is predicted by an analytical model

We consider a nucleic acid sequence comprised of n identical stickers (Fig. 1a). The stickers are separated by n–1 equally spaced linkers that do not interact with the stickers. Each linker consists of l nucleotides. Stickers are self-complementary and bind through base pairing interactions, such that each sticker can be bound to at most one other. Each bonded sticker has a free energy contribution of F_b; however, bonds that create closed loops also have an entropic cost ΔS_{loop} that depends on the loop length l_{loop}. This is because nucleotides comprising a closed loop (such as a hairpin, internal, or multi-loop) are constrained in the conformations they can adopt. A simple model treating unbound nucleotides as a polymer random walk estimates that the entropic cost of forming loops scales logarithmically with the loop length (see the "Methods" section)[26,27]. Assuming a characteristic loop length l_{eff}, the effective strength of the sticker interactions is $F \equiv F_b - T\Delta S_{loop}(l_{eff})$ (see the "Methods" section).

In this work, we are concerned with the behavior resulting from such sequences interacting with one another. Two stickers that bind to one another may be on the same strand or on two different strands. Moreover, many strands can be connected to one another through a chain of such bonds. We call a group of m strands connected through a series of intermolecular bonds a multimer of size m, or an m-mer. There are many ways a multimer of size m can form: any combination of bonds that occur either intra- or inter-molecularly within a group of m strands, such that each strand is reachable from every other by following a series of intermolecular bonds, is an m-mer.

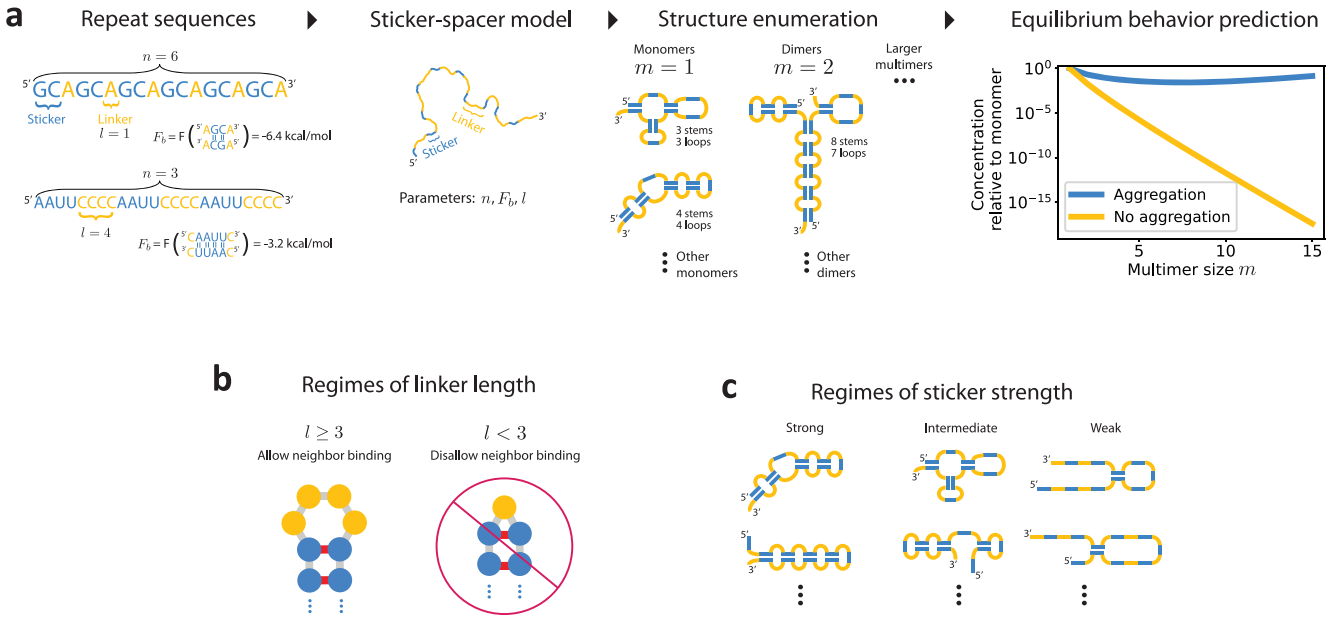


Fig. 1 | Model overview. a *Model procedure*: A repeat RNA or DNA sequence is converted to a sticker-spacer model, with stickers comprised of self-complementary regions. Possible structures, including multimers, are then enumerated by either computational or analytical methods. Partition functions are then calculated, leading to a complete description of the equilibrium behavior of the system, including the equilibrium concentrations of multimers. The system is in the aggregation regime when concentrations remain constant or increase with multimer size. **b** *Regimes of linker length*: The system can exhibit qualitatively different behavior depending on the length of the inert linkers. For long enough linkers, adjacent stickers can bind; for short linkers, they cannot because of hairpin size constraints. Structures visualized using *forna*[37]. **c** *Regimes of sticker strength*: For strong stickers, (almost) all of the sticker bonds are typically satisfied; for weak stickers almost none are; for intermediate strengths, the number of sticker bonds typically satisfied depends on a combination of the sticker strength and the multiplicity of structures in which a given number of stickers is bound.

We consider a system of M strands present in a container of volume V, such that their concentration is $c^{\mathrm{tot}} = M/V$. We take the thermodynamic limit of M and V going towards infinity with their ratio staying constant. We seek to predict how frequently multimers comprised of m strands form in this system, and how this frequency changes with m. We define c_m as the concentration of multimers of size m, such that

$$c^{\mathrm{tot}} = \sum_{m=1}^{\infty} m c_m. \tag{1}$$

There are two possible regimes for the system: For large m, c_m either decreases or increases with m (Fig. 1a). In the former case, the system is in the dilute phase, with only small multimers typically forming. In contrast, if c_m increases with m, large aggregates of the order of the system size dominate the landscape. The aggregation transition is defined as the crossover point between the regime in which very large multimers are suppressed, to that in which they are dominant.

In equilibrium, c_m is proportional to the ratio of the partition function of m-mers, Z_m, to the partition function of m monomers, $(Z_1)^m$ (see the "Methods" section). The partition functions are comprised of three terms:

$$Z_m = \mathrm{e}^{-\beta(m-1)\Delta F} \sum_{N_\mathrm{b}} g(n, m, N_\mathrm{b}) \mathrm{e}^{-\beta F N_\mathrm{b}}. \tag{2}$$

Here, the multiplicity factor $g(n, m, N_\mathrm{b})$ represents the number of distinct ways to make N_b bonds connecting m identical strands, each with n stickers. ΔF is the effective free energy cost of multimerization (see below) and $\beta = 1/k_\mathrm{B}T$ is the inverse thermal energy, where T is temperature. g can be calculated exactly (see the "Methods" section and Supplementary Note 1) and is qualitatively different depending on whether the linkers are long enough to allow adjacent stickers to bind to one another or not (Fig. 1b).

In order to fit experimental data on the prevalence of multiple nucleic acid strands binding to one another in vitro, nucleic acid models include a free energy penalty for multimerization. This leads to the term $(m-1)\Delta F$ in Eq. 2. This penalty is motivated by the enthalpic and entropic costs of nucleic acids binding, including ion effects and the translational and orientational entropies lost upon association[28–30]. This penalty scales linearly with the number of strands in a multimer, such that each additional strand added to a multimer carries the same penalty[31]. See the "Methods" section and Supplementary Note 2 for further discussion.

The sum in Eq. 2 can be approximated by its dominant term (a saddlepoint approximation). There are three regimes to consider, corresponding to strong, intermediate, and weak binding, in which the sum in Eq. 2 is dominated by large, intermediate, and small values of N_b, respectively (Fig. 1c). The value of $N_\mathrm{b} = N_\mathrm{b}^*$ that dominates the sum is that which maximizes a combination of the bond energy F and configurational entropy g. For example, the strong binding regime is characterized by bond energy considerations overwhelming configurational entropy effects, while the intermediate binding regime is characterized by a degree of balance between the two.

The model is validated by comparing to exact computational enumeration and previously published results

To validate the analytical model, we constructed a dynamic programming-based computational model that exactly enumerates Z_m in polynomial time (Supplementary Note 5.2). The analytical model described above makes three primary approximations compared to the computational model: (1) it assumes a constant entropy for all loops; (2) it considers only structures with a given number of bonds N_b (with a single next-order correction term); (3) it uses an approximate form for $g(n, m, N_\mathrm{b})$ (see the "Methods" section). The computational model makes none of these approximations, considering all (non-pseudoknotted; see Supplementary Note 5.1) structures that can form and including a loop-length-dependent loop entropy term.

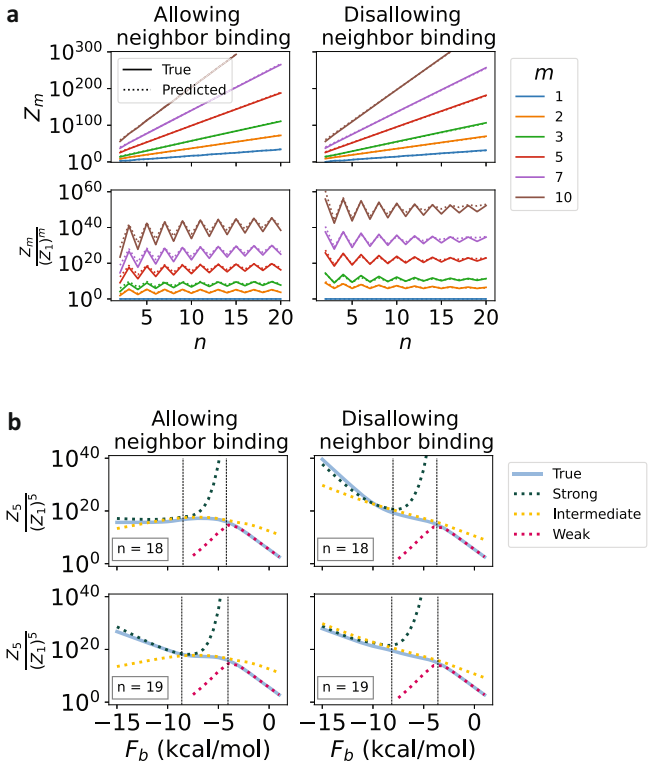


Fig. 2 | Analytical model demonstrates good agreement with computational results. a *As a function of number of stickers per strand*: Partition functions and partition function ratios are plotted with respect to n using the exact computational (solid) and simplified analytical (dotted) models. A single fitting parameter was used for the analytical models, fit to the monomer partition function (top row, blue). The slight discrepancy in the analytical prediction for large m and n disallowing neighbor bonds is primarily due to the heuristic approximation of $g(n, m, N_\mathrm{b})$ from $g(nm, 1, N_\mathrm{b})$ used. **b** As a function of binding strength: The ratio of the pentamer partition function to that of five monomers is plotted; similar results hold for any other multimer chosen. The analytical model predictions are separated into three regimes: strong (green), intermediate (yellow), and weak (red) binding. Vertical dashed lines separate where different regimes are expected to provide the best agreement and are calculated as the values of F_b such that $N_\mathrm{b}^* = N_\mathrm{b}^{\max} - 1$ and $N_\mathrm{b}^* = N_\mathrm{b}^{\min} + 3$. A single fitting parameter—the same one from panel (**a**)—is used.

Nevertheless, the analytical model closely approximates the exact computational model, as demonstrated in Fig. 2. The analytical model requires only one fitting parameter: the normalized effective loop length $l_{\mathrm{eff}}^{\mathrm{fit}}$ (see the "Methods" section). That parameter is fit separately to the regimes allowing and disallowing neighbor binding. Importantly, it is fit only once for each regime—to the monomer partition function with strong binding—and not separately for different values of n, m, or F_b. We demonstrate quantitative agreement between the analytical and computational models in Fig. 2, and in Supplementary Fig. 5.

We further sought to compare the model's predictions to previously published results, namely the MD simulations performed by Nguyen et al.[19]. Those simulations examined 64 **CAG**-repeat RNA strands with varying numbers of repeats per strand and of RNA concentrations. We considered the same system of **CAG** sequences, using the value $F_\mathrm{b} = -10$ employed in the MD simulations and no fitting parameters beyond the aforementioned single parameter fit to the computational model. We enumerated the monomer and dimer

https://doi.org/10.1038/s41467-023-35803-x

partition functions computationally, and used the analytical model to extrapolate up to $m = 64$, the number of strands used in the MD simulations. The extrapolation was performed by fitting the single parameter to our computational results for $m = 1$, and using Supplementary Eqs. S38 and S48 to obtain the results for $m > 2$. The primary difference between our model predictions and those of MD simulations is that the former is purely equilibrium, while the latter is decidedly not so, even after significant simulation time. (A secondary difference is that the former considers an infinite system of given concentration, while the latter considers a finite number of strands).

We plot the propensity of the system to form aggregates as a function of n and c^{tot} in Fig. 3. Following ref. [19], we define multimers of size $2 \le m \le 4$ as oligomers; however, this ensemble is dominated by dimers, with trimers and tetramers forming at very low fractions. We find that for certain concentrations, the system forms either monomers or dimers depending on the parity of n, in agreement with experimental results[17]; however, this parity does not significantly affect aggregation. We plot the results of Nguyen et al. on top of our predictions as colored points, finding excellent quantitative agreement between the two.

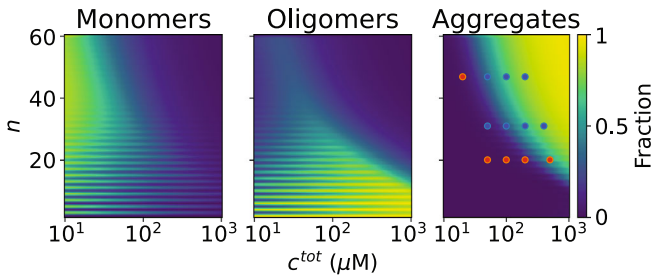


Fig. 3 | Landscape of CAG repeats. The equilibrium fraction of strands folded into monomers, oligomers (2–4-mers; primarily dimers), and aggregates are shown and compared to Nguyen et al.'s molecular dynamics (MD) simulation results. As the Nguyen et al. simulations used a sticker strength of $F_b = -10$ kcal/mol[19], we used the same sticker strength, with no fitting parameters to the simulations whatsoever. The MD simulation results are plotted as points in the aggregates panel, with blue points representing conditions for which aggregation was found, and red points for those in which it was not. We note that each of these points is a separate simulation taking 3 months of supercomputer time[19], in comparison to our analytical model for the entire landscape. In this system, neighbor binding is disallowed, monomers and dimers are in the strong binding regime, and multimers of $m \ge 3$ are in the intermediate regime. Aggregation is predicted for large concentrations and numbers of stickers per strand. Dimerization is less common as n increases, while dominant for small values of n, especially odd values.

A reentrant phase transition governs aggregation as a function of sticker strength

For very low temperatures or strong stickers, the ensemble of multimers is dominated by small structures such as dimers, in which all bonds can be satisfied. However, for intermediate sticker strengths, the configurational entropy gain of having a few unsatisfied bonds exceeds the energetic cost. This configurational entropy grows with multimer size, driving the system to aggregate. Finally, for very weak stickers or high temperatures, the structures melt. This phenomenon corresponds to a reentrant phase transition. We demonstrate this transition in our computational model in Fig. 4, enumerating up to $m = 15$. As shown in the figure, the two dilute phases at strong and weak

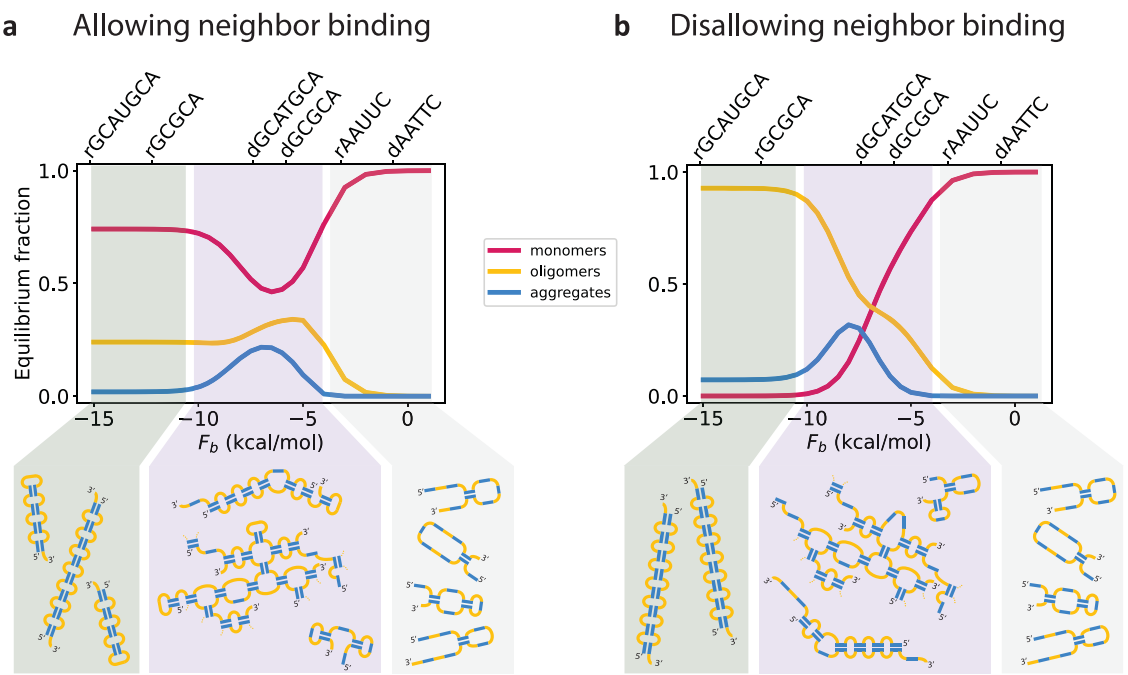


Fig. 4 | A reentrant transition as a function of sticker strength. Enumerating the exact partition functions up to $m = 15$ with the computational model, we find a reentrant transition with respect to F_b in both the regime allowing neighbor binding (panel **a**; $n = 8$, $l = 4$, $c^{tot} = 8$ mM is shown) and the regime disallowing neighbor binding (panel **b**; $n = 8$, $l = 1$, $c^{tot} = 4$ mM is shown). The high concentration used is a result of the lack of Mg^{2+} considered explicitly in the model; see the "Discussion" section. Aggregates (defined as $m \ge 5$-mers in accordance with ref. [19]) are most likely to form for intermediate sticker strengths, since very strong stickers lead to stable monomers (red) or dimers (dimers, trimers, and tetramers comprise the orange curve). Although aggregates are suppressed in both strong (green background; left) and weak (gray background; right) binding regimes, the molecular structures of monomers and dimers in these regimes are quite different: in the former, all or nearly all bonds are satisfied in a typical molecule, while very few bonds are typically satisfied in the latter regime. For this reason, the strong binding regime of the short linker case (i.e. disallowing neighbor binding) is predicted to contain a large concentration of dimers (which can satisfy all sticker bonds), and few monomers (which cannot). In the long linker case (i.e. allowing neighbor binding), for even values of n, monomers are also able to satisfy all bonds and are thus present at high concentrations in the strong binding regime. Top axis shows example sequences for RNA (r) and DNA (d), and their sticker strengths as calculated by the nearest-neighbor model, enabling a direct match from sequence to model predictions[29,30].

https://doi.org/10.1038/s41467-023-35803-x

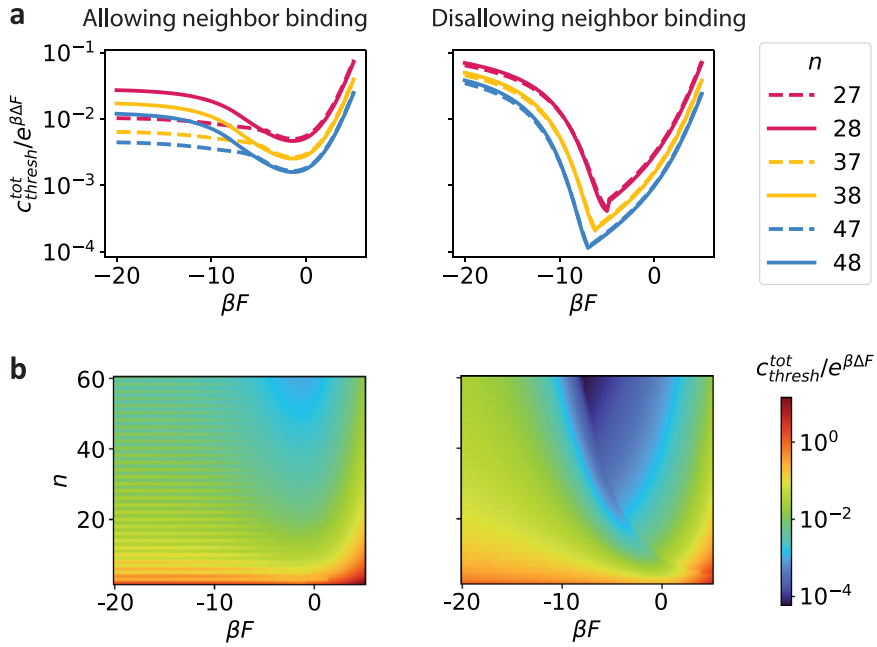


Fig. 5 | Phase diagram. a *Reentrant transition in the analytical model*: The analytical model enables enumeration up to arbitrarily large m, and reveals a reentrant transition. With high enough concentrations of RNA, aggregation is always possible; however, for certain concentrations, the analytical model predicts the system will undergo a reentrant phase transition in agreement with computational results (Fig. 4). Panel **a** shows slices for certain values of n through the complete phase diagram shown in panel (**b**). Parity of n affects aggregation phenomena for the system allowing neighbor binding (LHS). **b** *Complete phase diagram*: The complete phase diagram as predicted by enumeration up to arbitrarily large m with the analytical model is displayed. The normalized concentration needed to achieve aggregation is displayed as a function of n and βF. The reentrant transition is especially apparent for short linkers (RHS) as well as for long linkers with even values of n (LHS). Systems with long linkers typically require higher concentrations to aggregate than those with short linkers, since monomers are typically more stable in the former case. Discontinuities are due to the model's approximation of an abrupt transition from the strong to intermediate binding regimes for monomers. c^{tot}_{thresh} is made dimensionless by dividing by the multimerization cost $e^{\beta\Delta F}$ (see Supplementary Note 2).

binding regimes are quite different from one another. In the strong binding regime, (almost) all bonds are satisfied in a typical structure, mainly through intramolecular interactions or dimerization. In the weak binding regime, (almost) no bonds are typically satisfied.

We next explored whether this reentrant transition was merely a small m effect. We employed the analytical model, for which we can consider arbitrarily large values of m. Even when considering $m \to \infty$, we find a reentrant transition in the threshold concentration above which the system is expected to form aggregates, c^{tot}_{thresh} (see Supplementary Note 4), as shown in Fig. 5a. This transition is especially prominent for short linkers that disallow neighbor binding since the configurational entropy of dimers in this regime is quite limited (regardless of n, only one dimer configuration can satisfy all stickers). For longer linkers (allowing neighbor binding), this transition is most pronounced for even values of n for which monomers can satisfy all their own bonds, although it is apparent also for odd n, for which dimers can satisfy all bonds.

The behavior shown in Fig. 5 is in agreement with what we would expect from configurational entropy concerns alone (Supplementary Fig. 6). That the propensity of the system to aggregate occurs at more negative values of βF, and is more pronounced, for the case of disallowing neighbor binding than for the case of allowing neighbor binding, is predicted by the different forms of the configurational entropy in these two regimes. Similarly, larger values of n increase the propensity of the system to aggregate because of their effect on configurational entropy, rather than any enthalpic considerations (Supplementary Fig. 6).

Discussion

In this work, we have considered a simple model of competition between intra- and inter-molecular binding: a polymer with n identical evenly spaced self-complementary stickers. We have shown that the system is characterized by three parameters: n, the number of repeats per strand; βF, the effective strength of each bond accounting for the loop entropy cost; and $c^{tot}e^{-\beta\Delta F}$, a dimensionless concentration that accounts for multimerization cost.

Our model computes the prevalence of all possible multimers that can form, considering both intra-strand and inter-strand contacts. Our framework quantitatively recapitulates previously published MD simulation results, each data point of which required 3 months of supercomputer simulation time[19]. We substantially extend these results to arbitrary sequences, temperatures, and concentrations, and to arbitrarily large multimers (i.e. aggregates) in an analytical framework.

In this system, aggregation is not necessarily predicted as the regime where the most possible bonds are satisfied, as bonds can be satisfied by intramolecular as well as by intermolecular contacts. Instead, aggregation is predicted by the relative stability of the aggregate compared to smaller multimers. The stability of each structure is a function of three terms, as seen in Eq. 2: (1) the number of stickers bound (each contributes F to the free energy); (2) the number of strands in the structure (each contributes $\mu + \Delta F$, where μ is the chemical potential); and (3) the configurational entropy of the structure. This last term contributes $-\log(g)/\beta$ to the free energy, where g is the number of ways to satisfy the given number of bonds with the given number of strands in the structure.

This last term is the driving force for aggregation in this system. Aggregates are no more stable than dimers in terms of the first term, the possible number of stickers bound (both are able to satisfy all stickers). Aggregates are further penalized by the second term, the multimerization cost. If these two terms were the only terms in the free energy, we would not see any aggregates. It is the third term, the configurational entropy, that drives aggregation. Larger multimers are able to satisfy their bonds in many more configurations than a corresponding collection of smaller multimers, leading to an enormous

entropic benefit in forming aggregates. This has been described as a competition between configurational and translational entropies in other contexts[24,32]. In our system, the benefit due to g peaks when most, but not all, stickers are satisfied (Supplementary Fig. 6).

This behavior leads to a reentrant phase transition. For $-\beta F \gg 1$, the number of bonds satisfied is the primary consideration. Dimers are able to satisfy all their bonds, and the multiplicity benefit of aggregates is not sufficiently large when all bonds are satisfied, suppressing aggregation in this regime. Aggregation is also suppressed for very positive values of βF, which as a result of loop entropy costs can occur even when the sticker binding itself is favored (i.e. $F_b < 0$). However, for intermediate values of βF—when dimers prefer having some bonds left unsatisfied—the configurational entropy benefit of forming aggregates is overwhelming. Aggregates form at 1–2 orders of magnitude lower concentrations in this regime than in the strong binding regime.

The predicted aggregation transition of the system is completely described in Fig. 5b. We plot the (dimensionless) threshold concentration $c^{\mathrm{tot}}_{\mathrm{thresh}}$ as a function of n and βF. Aggregation is more prevalent for short linkers (disallowing neighbor binding) than for longer linkers (allowing neighbor binding). For short linkers, small structures are quite constrained in the number of ways they can satisfy all of their bonds, leading the differential configurational entropy benefit of aggregates to grow quite large. For longer linkers, smaller structures are more stable since the corresponding multiplicity is much larger. For similar reasons, the reentrant phase transition is most pronounced with short linkers. For long linkers, even values of n demonstrate a more pronounced reentrant transition than odd values, since their competition is between monomers—with no multimerization penalties—and aggregates. In all other cases, the reentrant transition is primarily due to competition between dimers and aggregates. For short linkers, the parity of n is found in our model to affect monomerization vs. dimerization in agreement with previously published results[17], but has almost no effect on aggregation properties. The reason is that for short linkers and strong stickers, dimers behave similarly regardless of the parity of n: both odd and even n can form a dimer satisfying all bonds with only one configuration.

Although there is a qualitative difference between short linkers of $l < 3$ and long linkers of $l \geq 3$, within each regime, increasing the linker length leads to larger values of ΔS and weaker binding. Decreasing the persistence length, for example by changing ionic conditions, would be expected to lead to a similar result. These effects and the predicted phase diagram as a whole (Fig. 5b) could be at least qualitatively tested experimentally by replicating the Jain & Vale experiments for multiple sequences with different sticker strengths and linker lengths and measuring the change in the concentration needed to form aggregates for the different conditions. The available published data is in good agreement with our predictions, in that larger values of n show a greater propensity for aggregation in both experiments and our model predictions[7].

Our results bear similarities to the so-called "magic number effect" whereby for heterotypic mixtures, aggregation is suppressed when the number of binding sites in one species is a small integer multiple of the other's[32,33]. In such systems, small stable clusters can form with all bonds satisfied. In our homotypic system, dimers can always exhibit a magic number-like effect for strong stickers, and in the regime in which neighbor binding is allowed, for even n, monomers can as well. In fact, a weak reentrant transition has been observed in some simulations of the magic number effect in heterotypic systems (see Fig. 3A of ref. [34]). Our results suggest that a reentrant transition may be a generic feature of the magic number effect and that the strength of the reentrant behavior may decay the more molecules are involved.

Our model has several limitations. To make the expression analytically tractable, our formalism makes a heuristic approximation for the multimer multiplicity factor g in the regime disallowing neighbor bonds. For similar reasons, we were unable to analytically explore the weak binding regime, applicable for systems where the loop entropy cost of forming stickers outweighs their energetic benefit. A limitation of our model's physiological applicability is that we did not explicitly consider magnesium. Magnesium can act as a bridge between negatively charged RNA molecules such that even in the absence of base pairing, Mg–RNA mixtures can form aggregates[18,35]. Experimental results thus rely on magnesium aiding the aggregation process[7]. However, the MD simulations to which we compare here do not explicitly consider magnesium[19] and the high concentrations required for the system to aggregate (e.g. Fig. 3) are the result. To first-order, the effects of magnesium could be accounted for in our model as modifying ΔF (along with F_b), which effectively modifies the concentrations, as concentrations only enter the model as $c^{\mathrm{tot}}e^{-\beta\Delta F}$. For clarity, we opted to leave ΔF unmodified; therefore, the high concentrations we consider should be significantly decreased for a system including magnesium.

While non-equilibrium effects are relevant in these systems, our analysis is entirely an equilibrium prediction. Indeed, kinetic trapping appears to be the biggest experimental hurdle to testing our reentrant phase predictions. At the same time, the results of decidedly out-of-equilibrium MD simulations[19] show excellent quantitative agreement with our equilibrium predictions (Fig. 3). For this reason, it is likely that out-of-equilibrium effects are not the dominant factor in repeat RNA aggregation behavior. In vivo RNA aggregates are even more fluid-like and dynamic than in vitro aggregates, for reasons that remain largely unclear but appear to be the result of active enzymes in the cell[7]. Future work may consider how such active processes affect the aggregation properties, and the connection between in vivo non-equilibrium steady states and the equilibrium steady state discussed here.

Given the radical simplicity of the model used here, there is a host of extensions to consider. For example: How does this model interact with complex coacervation, as when including polycations in the solution? How does a polymer pattern with multiple orthogonal stickers behave? How do multiple different polymers, with both *cis* and *trans* binding, interact with one another? And how do physiological RNA molecules use the principles explored here to control their aggregation properties?

Our work demonstrates that the competition between intra- and inter-molecular binding can lead to remarkable and (perhaps) unintuitive behavior. Our results mapping the control knobs for this phase behavior create a framework for the study of RNA–RNA interactions in in vivo biological condensates and set the stage for the construction of novel materials and new techniques based on programmable RNA condensates.

Methods

Partition functions determine equilibrium behavior

We consider a nucleic acid sequence comprised of n stickers separated by $n-1$ linkers (Fig. 1a). Stickers are self-complementary and bind through base pairing interactions, such that each sticker can be bound to at most one other sticker. The strength of the sticker interactions, F_b, is determined by the sequence of the stickers; for example, an RNA **GC** sticker with **A** nucleotide linkers in standard conditions has $F_b = -6.4$ kcal/mol (or, for DNA, −1.4), while a **GCGC** sticker has $F_b = -12.2$ kcal/mol (−5.8 for DNA). These are calculated using the classic nearest-neighbor model for RNA or DNA base-pairing interactions[29,30]. The linkers, each of which is of length l, are inert.

We seek to predict how frequently multimers comprised of m strands form, and how this frequency changes with m. Aggregation occurs in the parameter regime where the concentration of multimers comprised of m strands, c_m, increases with m. c_m is defined as the sum of all structures that have m strands connected by base pairing interactions. In equilibrium, c_m is proportional to the partition function of

m-mers, Z_m:

$$Z_m = \sum_{\sigma_m} e^{-\beta F(\sigma_m)}. \tag{3}$$

Here, σ_m is a structure comprised of m strands linked by base pairing, including potential intramolecular bonds; and $\beta = 1/k_B T$ where k_B is Boltzmann's constant and T is the temperature measured in Kelvin. $F(\sigma_m)$ is the free energy of the structure, given by[29]

$$F(\sigma_m) = F_b N_b(\sigma_m) + (m-1)\Delta G_{assoc} - T\sum_{loops} \Delta S_{loop}(l_{loop}), \tag{4}$$

where $N_b(\sigma_m)$ is the number of bonds in the structure, and ΔG_{assoc} is the hybridization penalty associated with intermolecular binding (discussed below). Each closed loop of length l_{loop} leads to an entropic penalty of $\Delta S_{loop}(l_{loop})$, associated with the decrease in three-dimensional configurations of the single-stranded region of the loop compared to a free chain, given by[26,27]

$$\Delta S_{loop}(l_{loop}) = k_B\left[\ln \upsilon_s + \frac{3}{2}\ln\left(\frac{3}{2\pi b\, l_{loop}}\right)\right], \tag{5}$$

where υ_s is the volume within which two nucleotides can bind, and b is the persistence length of single-stranded regions. This equation treats the single-stranded loop as an ideal chain. An excluded volume term υm^2 can be added to Eq. 4[20] but we assume υ is small enough that this term is negligible except for very large m (see Supplementary Note 4 for further discussion).

Given the partition functions Z_m for all m-mers, we can calculate the equilibrium concentrations of m-mers, c_m, for all m, by solving a set of m simultaneous equations. Z_m affects physical observables such as c_m only through the ratio Z_m/Z_1^m, describing, in essence, the propensity of m strands to form an m-mer as opposed to m monomers[25,31]:

$$\begin{aligned} c_m &= \frac{Z_m}{Z_1^m} c_1^m \\ \sum_m m c_m &= c^{tot} \end{aligned} \tag{6}$$

where the concentrations are made dimensionless by normalizing by a reference concentration (see Supplementary Note 2) and c^{tot} is the total concentration of strands added to solution. In short, this equation arises from $c_m = Z_m e^{m\beta\mu}$ where μ is the chemical potential and the fugacity $e^{\beta\mu} = c_1/Z_1$ in equilibrium[25].

Solutions to Eq. 6 have two typical regimes. In one, c_m decays exponentially with m. On the other, c_m grows with m (until excluded volume effects begin to dominate). The latter regime corresponds to aggregation (Fig. 1a).

An analytical model for the partition functions

The calculation of Z_m is too computationally intensive to perform directly, by explicitly enumerating all possible structures that can form, as the number of possible structures grows exponentially with n and m. In order to predict phase behavior for a wide range of sequences and experimental conditions, we develop an analytical framework for computing Z_m. This framework enables us to search a broad parameter space and tune phase behavior in the system. We validate our analytical model against a computational model that exactly calculates Z_m with a dynamic programming approach (Supplementary Note 5.2) thus providing an exact baseline model for comparison.

We rely on one major assumption to enable an analytical approach: we approximate the loop entropies as independent of loop length; or equivalently, we assume that the model is dominated by loops of one characteristic length, l_{eff}. This length depends on the length of the linkers in the system, l. This approximation is reasonable because of two factors. First, because of the logarithmic dependence of ΔS_{loop} on loop length (Eq. 5), moderate heterogeneities in loop length lead to only small differences in ΔS_{loop}. Second, because the typical number of loops in a multimer scales linearly with the size of the multimer (see Supplementary Note 3), we expect similar levels of heterogeneity in loop length independent of the size of the multimer. This approximation is expected to break down for very large n and weak binding ($F_b > 0$), in which case the few loops that typically form will likely have a broad distribution of lengths; this regime is not considered here.

With this approximation, for monomers, each bond provides constant free energy of $F = F_b - T\Delta S$, where $\Delta S = \Delta S_{loop}(l_{eff})$. Since the number of loops is given by $N_b - (m-1)$, we also define $\Delta F \equiv (\Delta G_{assoc} + T\Delta S)$. This quantity enters Eq. 6, such that it allows us to redefine a rescaled concentration $ce^{-\beta\Delta F}$ (also, see Supplementary Note 2). Without rescaling concentration, the partition function Z_m can thus be written as

$$\begin{aligned} Z_m &= e^{-\beta(m-1)\Delta F}\sum_{\sigma_m} e^{-\beta F N_b(\sigma_m)} \\ &= e^{-\beta(m-1)\Delta F}\sum_{N_b} g(n,m,N_b) e^{-\beta F N_b} \end{aligned} \tag{7}$$

where the multiplicity factor $g(n,m,N_b)$ represents the number of distinct ways to make N_b bonds connecting m identical strands, each with n stickers. This is identical to Eq. 2.

This multiplicity factor is most straightforward to consider for the case of monomers. We make the approximation that the contribution of pseudoknots to the partition function is negligible due to their high entropic cost (see Supplementary Note 5.1). Our goal is therefore to calculate the number of ways to form non-pseudoknotted structures containing N_b bonds given a strand of n stickers. For monomers, the multiplicity can be calculated exactly. However, the result depends on whether adjacent stickers are able to bind to one another or not. For a long enough linker length (≥3 nts for the case of RNA), neighboring stickers can bind; for shorter linker lengths (as, for example, for **CAG** repeats), they cannot (see Fig. 1b). As derived in Supplementary Note 1.1,

$$g(n,1,N_b) = \begin{cases} \frac{n!}{(n-2N_b)!\,(N_b+1)!\,N_b!} & \text{if adjacent stickers can bind} \\ \frac{(n-N_b)!\,(n-N_b-1)!}{(n-2N_b)!\,(n-2N_b-1)!\,(N_b+1)!\,N_b!} & \text{otherwise} \end{cases} \tag{8}$$

The top line (allowing neighbor binding) is simply calculated as the product of two factors: $\binom{n}{2N_b}$ (the number of ways to choose $2N_b$ bound stickers from n possibilities); and the Catalan number C_{N_b} (the number of non-pseudoknotted ways to construct bonds between the chosen stickers). The bottom line (disallowing neighbor bonds) requires a brief additional calculation to derive (Supplementary Note 1.1).

Calculating $g(n,m,N_b)$ from $g(n,1,N_b)$ also depends on whether or not adjacent stickers can bind (see Supplementary Note 1.2). While the exact calculation requires large numbers of sums with no closed-form solution, a close approximation is given by

$$g(n,m,N_b) \approx \begin{cases} \frac{g(nm,1,N_b)}{m} & \text{if adjacent stickers can bind} \\ \frac{g(nm+\alpha(m-1),1,N_b)}{m} & \text{otherwise} \end{cases} \tag{9}$$

where $\alpha \approx 0.42$, representing an additional heuristic for the case of disallowing neighbor binding compared to the case of allowing such binding. The value of $\alpha = 0.42$ used is a heuristic estimate that is an

especially good fit to the strong interaction regime, and other approximations may improve it (see Supplementary Fig. 1). The factor of $1/m$ corrects for overcounting due to symmetry (Supplementary Note 1.2.3; see also Supplementary Fig. 2)[36].

Given expressions for the multiplicity factor, the partition functions (Eq. 7) are now in principle computable. However, the full sum in that equation remains too computationally intensive to be useful. We, therefore, turn to a saddlepoint approximation: sums of exponentials are typically dominated by their maximum terms, and Eq. 7 is no exception.

In order to find the maximum term, there are three cases to consider, corresponding to physically meaningful distinctions (Fig. 1c). In one regime, the "strong binding" regime, the ensemble is dominated by structures that maximize the bond energy, and the sum is dominated by the last terms ($N_b = N_b^{max}$). In the second, the "intermediate binding" regime, the ensemble is dominated by structures that maximize a combination of the bond energy and configurational entropy measured by g, and the sum is dominated by an intermediate-term ($N_b = N_b^*$). In the third, the "weak binding" regime, the ensemble is dominated by structures that have almost no bonds, and the sum is dominated by the first terms ($N_b = N_b^{min}$). These three cases must be treated separately: in the strong and weak binding regimes, the discrete nature of the sum is crucial, while in the intermediate regime, the sum can be well-approximated by an integral. The boundary between these regimes occurs approximately when $N_b^* = N_b^{max} - 1$ or $N_b^* = N_b^{min} + 3$. For Figs. 3 and 5, we set the boundary between the strong and intermediate regimes at $N_b^* = N_b^{max} - \frac{1}{4}$ (allowing neighbor binding) and $N_b^* = \frac{n}{2} - 2$ (disallowing neighbor binding).

After computing the dominant term of the sum, the next-order correction to Z_m comes from either considering the next-dominant term (strong and weak regimes) or the curvature at the maximum (intermediate regime); see Supplementary Note 3 for more details.

When comparing between the analytical and computational models, we use a single fitting parameter l_{eff}^{fit}, which tunes the normalized effective loop length. That parameter is fit separately to the monomer partition functions allowing and disallowing neighbor binding, but is kept constant for all values of m. For different binding strengths, a different fraction of stickers will be bonded, leading to a different value of l_{eff}. Rather than having a separate fitting parameter for each parameter set, we only fit once (to monomers) in each of the two linker length regimes (allowing and disallowing neighbor binding). We then assume that l_{eff} changes linearly with the fraction of stickers bonded, leading to:

$$l_{eff} = \frac{nm}{2N_b^*} l_{eff}^{fit}. \tag{10}$$

We fit l_{eff}^{fit} to the strong binding regime (Fig. 2) for which $l_{eff} \approx l_{eff}^{fit}$. We find intuitively reasonable values for l_{eff}^{fit}. When using $l = 1$ (disallowing neighbor binding), we find $l_{eff}^{fit} = 4.3$ nucleotides. This value is in between the length of an internal loop formed by two individual linkers (4 nucleotides) and the length of a hairpin loop formed by two linkers and a sticker (5 nucleotides). When using $l = 4$ (allowing neighbor binding), we find $l_{eff}^{fit} = 7$ nucleotides. This value is also in between the length of an internal loop formed by two individual linkers (10 nucleotides) and the length of a hairpin loop formed by a single linker (5 nucleotides).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

All code used to generate the results and figures in this study can be found at https://github.com/ofer-kimchi/RNA-aggregation.

References

1. Van Treeck, B. & Parker, R. Emerging roles for intermolecular RNA–RNA interactions in RNP assemblies. *Cell* **174**, 791–802 (2018).
2. Polymenidou, M. The RNA face of phase separation. *Science* **360**, 859–860 (2018).
3. Fay, M. M. & Anderson, P. J. The role of RNA in biological phase separations. *J. Mol. Biol.* **430**, 4685–4701 (2018).
4. Roden, C. & Gladfelter, A. S. RNA contributions to the form and function of biomolecular condensates. *Nat. Rev. Mol. Cell Biol.* **22**, 1–13 (2020).
5. Van Treeck, B. et al. RNA self-assembly contributes to stress granule formation and defining the stress granule transcrip- tome. *Proc. Natl Acad. Sci. USA* **115**, 2734–2739 (2018).
6. Langdon, E. M. et al. mRNA structure determines specificity of a polyQ-driven phase separation. *Science* **360**, 922–927 (2018).
7. Jain, A. & Vale, R. D. RNA phase transitions in repeat expansion disorders. *Nature* **546**, 243–247 (2017).
8. Ellerby, L. M. Repeat expansion disorders: mechanisms and therapeutics. *Neurotherapeutics* **16**, 924–927 (2019).
9. Usdin, K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Res.* **18**, 1011–1019 (2008).
10. Hannan, A. J. Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **19**, 286–298 (2018).
11. Santhana Mariappan, S. V., Garcia, A. E. & Gupta, G. Structure dynamics of the DNA hairpins formed by tandemly repeated CTG triplets associated with myotonic dystrophy. *Nucleic Acids Res.* **24**, 775–783 (1996).
12. Kiliszek, A., Kierzek, R., Krzyzosiak, W. J. & Rypniewski, W. Atomic resolution structure of CAG RNA repeats: structural insights and implications for the trinucleotide repeat expansion diseases. *Nucleic Acids Res.* **38**, 8370–8376 (2010).
13. Figueroa, A. A., Cattie, D. & Delaney, S. Structure of even/odd trinucleotide repeat sequences modulates persistence of non-b conformations and conversion to duplex. *Biochemistry* **50**, 4441–4450 (2011).
14. Ni, ChengWei, Wei, Yu. Jie, Shen, Y. I. & Lee, I. R. Long-range hairpin slippage reconfiguration dynamics in trinucleotide repeat sequences. *J. Phys. Chem. Lett.* **10**, 3985–3990 (2019).
15. Sobczak, K., de Mezer, M., Michlewski, G., Krol, J. & Krzyzosiak, W. J. RNA structure of trinucleotide repeats associated with human neurological diseases. *Nucleic Acids Res.* **31**, 5469–5482 (2003).
16. Broda, M., Kierzek, E., Gdaniec, Z., Kulinski, T. & Kierzek, R. Thermodynamic stability of RNA structures formed by CNG trinucleotide repeats. Implication for prediction of RNA structure. *Biochemistry* **44**, 10873–10882 (2005).
17. Huang, J. & Delaney, S. Unique length-dependent biophysical properties of repetitive DNA. *J. Phys. Chem. B* **120**, 4195–4203 (2016).
18. Ma, Y., Li, H., Gong, Z., Yang, S., Wang, P. & Tang, C. Nucleobase clustering contributes to the formation and hollowing of repeat-expansion RNA condensate. *J. Am. Chem. Soc.* **144**, 4716–4720 (2022).
19. Nguyen, H. T., Hori, N. & Thirumalai, D. Condensates in RNA repeat sequences are heterogeneously organized and exhibit reptation dynamics. *Nat. Chem.* **14**, 775–785 (2022).
20. Semenov, A. N. & Rubinstein, M. Thermoreversible gelation in solutions of associative polymers. 1. Statics. *Macromolecules* **31**, 1373–1385 (1998).

21. Choi, JeongMo, Holehouse, A. S. & Pappu, R. V. Physical principles underlying the complex biology of intracellular phase transitions. *Annu. Rev. Biophys.* **49**, 107–133 (2020).
22. Lin, Y. et al. Narrow equilibrium window for complex coacervation of tau and RNA under cellular conditions. *eLife* **8**, 1–31 (2019).
23. Dobrynin, A. V. Phase diagram of solutions of associative polymers. *Macromolecules* **37**, 3881–3893 (2004).
24. Weiner, B. G., Pyo, A. G. T., Meir, Y. & Wingreen, N. S. Motif-pattern dependence of biomolecular phase separation driven by specific interactions. *PLoS Comput. Biol.* **17**, 1–17 (2021).
25. Curatolo, A. I., Kimchi, O., Goodrich, C. P. & Brenner, M. P. The assembly yield of complex, heterogeneous structures: a computational toolbox. Preprint at *bioRxiv* https://doi.org/10.1101/2022.06.26.497606 (2022).
26. Jacobson, H. & Stockmayer, W. H. Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.* **18**, 1600–1606 (1950).
27. Kimchi, O., Cragnolini, T., Brenner, M. P. & Colwell, L. J. A polymer physics framework for the entropy of arbitrary pseudoknots. *Biophys. J.* **117**, 520–532 (2019).
28. Xia, T. et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry* **37**, 14719–14735 (1998).
29. Turner, D. H. & Mathews, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.* **38**, 2009–2011 (2009).
30. SantaLucia, J. & Hicks, D. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 415–440 (2004).
31. Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E. & Pierce, N. A. Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.* **49**, 65–88 (2007).
32. Xu, B. et al. Rigidity enhances a magic-number effect in polymer phase separation. *Nat. Commun.* **11**, 4–11 (2020).
33. Freeman Rosenzweig, E. S. et al. The Eukaryotic CO2-concentrating organelle is liquid-like and exhibits dynamic reorganization. *Cell* **171**, 148-162.e19 (2017).
34. Zhang, Y., Xu, B., Weiner, B. G., Meir, Y. & Wingreen, N. S. Decoding the physical principles of two-component biomolecular phase separation. *eLife* **10**, 1–31 (2021).
35. Onuchic, P. L., Milin, A. N., Alshareedah, I., Deniz, A. A. & Banerjee, P. R. Divalent cations can control a switch-like behavior in heterotypic and homotypic RNA coacervates. *Sci. Rep.* **9**, 1–10 (2019).
36. Klein, E. D., Perry, R. W. & Manoharan, V. N. Physical interpretation of the partition function for colloidal clusters. *Phys. Rev. E* **98**, 1–12 (2018).
37. Kerpedjiev, P., Hammer, S. & Hofacker, I. L. Forna (force-directed RNA): simple and effective online RNA secondary structure diagrams. *Bioinformatics* **31**, 3377–3379 (2015).

Acknowledgements

We thank Sumit Majumder and Ankur Jain for sharing their expertise of this system and the central experiments, as well as Hung Nguyen and Naoto Hori for discussions of their molecular dynamics simulations. We thank Ned Wingreen and Yaojun Zhang for discussions of magic number systems and their connection to the present work. We also thank Krishna Shrinivas, Peter Clote, Megan Engel, Ben Weiner, and Rees Garmann for interesting and useful discussions. This work was supported by the Peter B. Lewis '55 Lewis-Sigler Institute/Genomics Fund through the Lewis-Sigler Institute of Integrative Genomics at Princeton University, and the National Science Foundation through the Center for the Physics of Biological Function (PHY-1734030) (O.K.); a National Science Foundation Graduate Research Fellowship under Grant No. DGE1745303 (E.M.K.); the Harvard Materials Research Science and Engineering Center (DMR 20-11754), the Office of Naval Research (ONR N00014-17-1-3029), and the Simons Foundation through the Simons Foundation Investigator Award (M.P.B.).

Author contributions

All the authors (O.K., E.M.K., M.P.B.) designed research. O.K. and E.M.K. carried out theoretical calculations, wrote Python code, and analyzed data. All the authors (O.K., E.M.K., M.P.B.) wrote the article.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-35803-x.

Correspondence and requests for materials should be addressed to Ofer Kimchi.

Peer review information *Nature Communications* thanks Gian Tartaglia and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Supplementary Notes

Contents

Supplementary Note 1 Calculating the multiplicity factor g

In this section we describe the calculation of $g(n, m, N_b)$, the number of distinct possible combinations of N_b bonds that can be formed with m strands, each with n stickers. We first derive $g(n, 1, N_b)$, i.e. the monomer case. We consider both the case in which adjacent stickers can bind to one another, and the case in which they cannot. We then consider multimers, deriving both exact and approximate expressions for $g(n, m, N_b)$. We also explain the origin of the $1/m$ term in $g(n, m, N_b)$, and its connection to symmetry factors.

1.1 The calculation of g for monomers

1.1.1 Allowing neighbor binding

The question we pose here is as follows: given n binding sites, each of which can be bound to at most one other binding site, how many non-pseudoknotted ways are there of constructing N_b bonds between the sites?

There are $\binom{n}{2N_b}$ ways of choosing the $2N_b$ bound sites out of the n possibilities. Then, there are C_{N_b} distinct ways of constructing N_b non-pseudoknotted bonds between $2N_b$ sites, where C_i is the i^{th} Catalan number, defined as

$$C_{N_b} = \frac{1}{N_b + 1}\binom{2N_b}{N_b}. \tag{S1}$$

Combining these expressions, we find that

$$g_a(n, 1, N_b) = \frac{n!}{(n - 2N_b)!\,(N_b + 1)!\,N_b!} \tag{S2}$$

where the subscript a refers to the fact that the expression is the result of allowing neighbor binding.

1.1.2 Disallowing neighbor binding

If the linker is too short to allow binding between adjacent stickers, the resulting expression is more complicated. To our knowledge, this case has not been explored previously, and so we derive it here.

We start with the result allowing neighbor binding, g_a derived in the previous section. We find g_d ("d" for "disallowing" neighbor binding) by taking g_a and subtracting out the number of structures with at least one pair of neighboring stickers bound.

There are $n-1$ possible neighbor bonds, and for each, there are $g_a(n-2, 1, N_b-1)$ ways to arrange the remaining $N_b - 1$ bonds. There thus appear to be $(n-1)g_a(n-2, 1, N_b-1)$ ways to make N_b bonds with n stickers, including at least one neighbor bond. Our answer appears to be given by $g_a(n, 1, N_b) - (n-1)g_a(n-2, 1, N_b-1)$.

However, there is an error in that calculation, since some structures have two neighbor bonds! These structures were counted twice: once when fixing the first neighbor bond, and once when fixing the second. We therefore need to add back in the number of structures that have at least two neighbor bonds. Using a similar reasoning to previously, we get that the number of such structures appears to be $\frac{1}{2}(n-2)(n-3)g_a(n-4, 1, N_b-2)$. (We derive this result more fully below, but it perhaps makes intuitive sense: the division by two corrects for the fact that it doesn't matter in what order you determine the two bonds). Our answer thus appears to be

$$g_a(n, 1, N_b) - (n-1)g_a(n-2, 1, N_b-1) + \frac{(n-2)(n-3)}{2}g_a(n-4, 1, N_b-2). \tag{S3}$$

You may already see where this is going. That calculation itself had a similar error to the first: we now overcounted the number of structures with at least 3 neighboring bonds. Continuing this procedure further, we find that

$$g_d(n, 1, N_b) = g_a(n, 1, N_b) + \sum_{i=1}^{N_b}(-1)^i\binom{n-i}{i}g_a(n-2i, 1, N_b-i). \tag{S4}$$

The result of this sum is given by

$$g_d(n, 1, N_b) = \frac{(n-N_b)!\,(n-N_b-1)!}{(n-2N_b)!\,(n-2N_b-1)!\,(N_b+1)!\,N_b!} \tag{S5}$$

This result is exact and was validated by comparing to explicit computational enumeration.

To derive the factor of $\binom{n-i}{i}$, consider for example the case of $i = 2$: how many ways are there to specify two neighbor bonds? One of the two neighbor bonds comes first. If that bond is between the first and second stickers, there are $n - 3$ ways to place the remaining neighbor bond. If that bond is between the second and third stickers, there are $n - 4$ ways to place the second neighbor bond. Continuing on, there is only one place to place the remaining neighbor bond if the first bond is between the $(n-3)^{\text{rd}}$ and $(n-2)^{\text{nd}}$ stickers. Thus, the total number of ways to place the two bonds is

$$\sum_{j=1}^{n-3}(n-2-j) = \frac{(n-2)(n-3)}{2} = \binom{n-2}{2}. \tag{S6}$$

The general expression of $\binom{n-i}{i}$ can be derived along similar lines.

1.2 The calculation of g for multimers

1.2.1 Allowing neighbor binding

If neighbors are allowed to bind, then the number of ways of constructing bonds between strands in a multimer is the same as the corresponding number for a monomer with the same total number of stickers. The result is therefore $g_a(n, m, N_b) = g_a(nm, 1, N_b)$.

One caveat here is that some of the bond combinations do not actually lead to a connected multimer. For example, $g_a(n, m, m-2)$ should always be zero – there is no way to make a multimer comprised of m strands with $m-2$ bonds – but $g_a(nm, 1, m-2)$ need not be zero. Correcting the expression for g to compute only connected multimers is the subject of a later section, but for the most part, this correction is negligible in the regimes we consider in this work.

Another caveat – of considering *distinct* structures, is also considered in a later section, and leads to a correction of $1/m$ in the expression above.

1.2.2 Disallowing neighbor binding

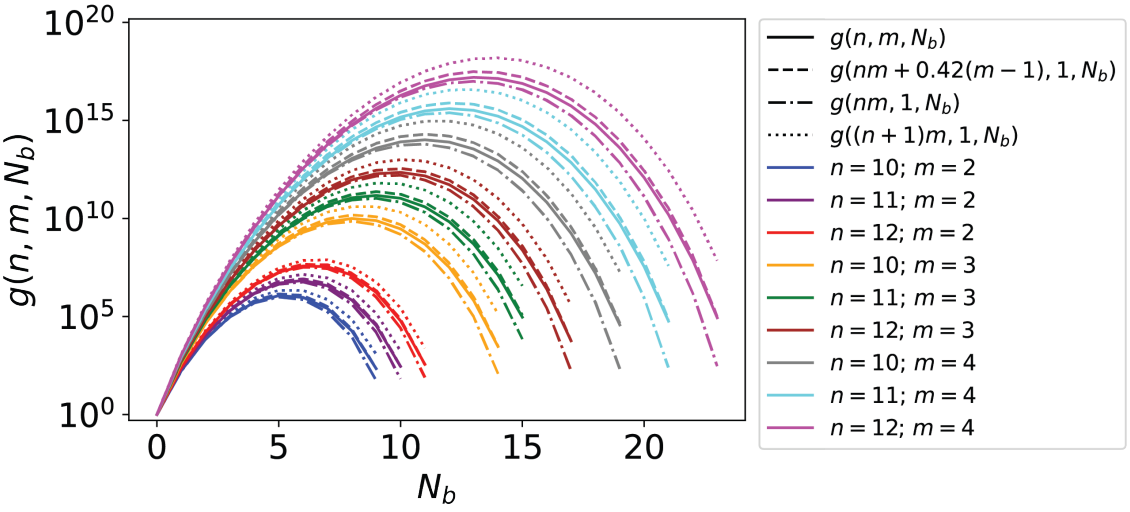
If neighbors are not allowed to bind, then $g_d(nm, 1, N_b)$ slightly underestimates $g_d(n, m, N_b)$, even ignoring the two caveats mentioned above. This is because the last sticker of one strand is actually allowed to bind to the first sticker of the next, though this binding would be disallowed in a monomer. If we consider $g_d(nm + (m-1), 1, N_b)$, that is a slight *over*estimate: although it corrects for the previous issue, the new "phantom" stickers are allowed to bind in the monomer approximation, although they are not for the real multimer. An intermediate estimate, of $g_d(nm+\alpha(m-1), 1, N_b)$ appears to provide a reasonably good approximation for the true value of g_d when $\alpha \approx 0.42$ (see Supplementary Fig. 1. The value of $\alpha = 0.42$ used is a heuristic estimate that especially well-fits the strong interaction regime, and other approximations may improve it.

In the remainder of this section, we will derive an exact expression for $g_d(n, m, N_b)$. This derivation proceeds along similar lines to the derivation of the monomer case.

There are $g_a(n, m, N_b) = g_a(nm, 1, N_b)$ ways of making N_b non-pseudoknotted bonds with nm stickers, including neighbor pairings. There are $m(n-1)$ ways to fix one neighbor bond, and $g_a(nm-2, 1, N_b-1)$ ways to arrange the remaining bonds given one fixed bond. Therefore, it would appear that the result is $g_a(nm, 1, N_b) - m(n-1)g_a(nm-2, 1, N_b-1)$. However, some of the ways of rearranging the remaining bonds *themselves* have a neighbor bond, and so we counted those structures twice: once when fixing the first neighbor bond, and once when fixing the second. Following this through, as in the monomer case, we have that

$$g_d(n, m, N_b) = g_a(nm, 1, N_b) + \sum_{i=1}^{N_b}(-1)^i\, t(n, m, i) g_a(nm-2i, 1, N_b-i) \tag{S7}$$

where $t(n, m, i)$ is the number of ways to fix i neighbor bonds, given m strands, each of length n. We found previously that $t(n, 1, i) = \binom{n-i}{i}$. For general m,



Supplementary Figure 1: **Approximating $g_d(n, m, N_b)$ for $m > 1$.** The exact calculation of $g_d(n, m, N_b)$ (S8) is shown alongside the heuristic approximations $g_d(nm + \alpha(m-1), 1, N_b)$ for $\alpha = 0, 0.42$, and 1. The range of n and m chosen was limited by the computation time of the full sum. $\alpha = 0.42$ shows good agreement with the full sum over this range especially for large values of N_b, which are most relevant in the strong interaction regime. A better fit may be found by modulating α as a function of F_b, or perhaps with a different model.

$$t(n, m, i) = \binom{m}{1} t(n, 1, i) + \binom{m}{2} \sum_{j=1}^{i-1} t(n, 1, j) t(n, 1, i-j) + \binom{m}{3} \sum_{j=1}^{i-2} \sum_{k=1}^{i-2} t(n, 1, j) t(n, 1, k) t(n, 1, i-j-k) + \ldots \quad \text{(S8)}$$

since there are $t(n, 1, i)$ ways to fix i neighbor bonds on a single strand and $\binom{m}{1}$ ways to pick a single strand out of m strands; there are $t(n, 1, j)$ ways to fix j neighbor bonds on a single strand, $t(n, 1, i-j)$ ways to fix the remaining bonds on another strand, and $\binom{m}{2}$ ways to choose the two strands; and so on.

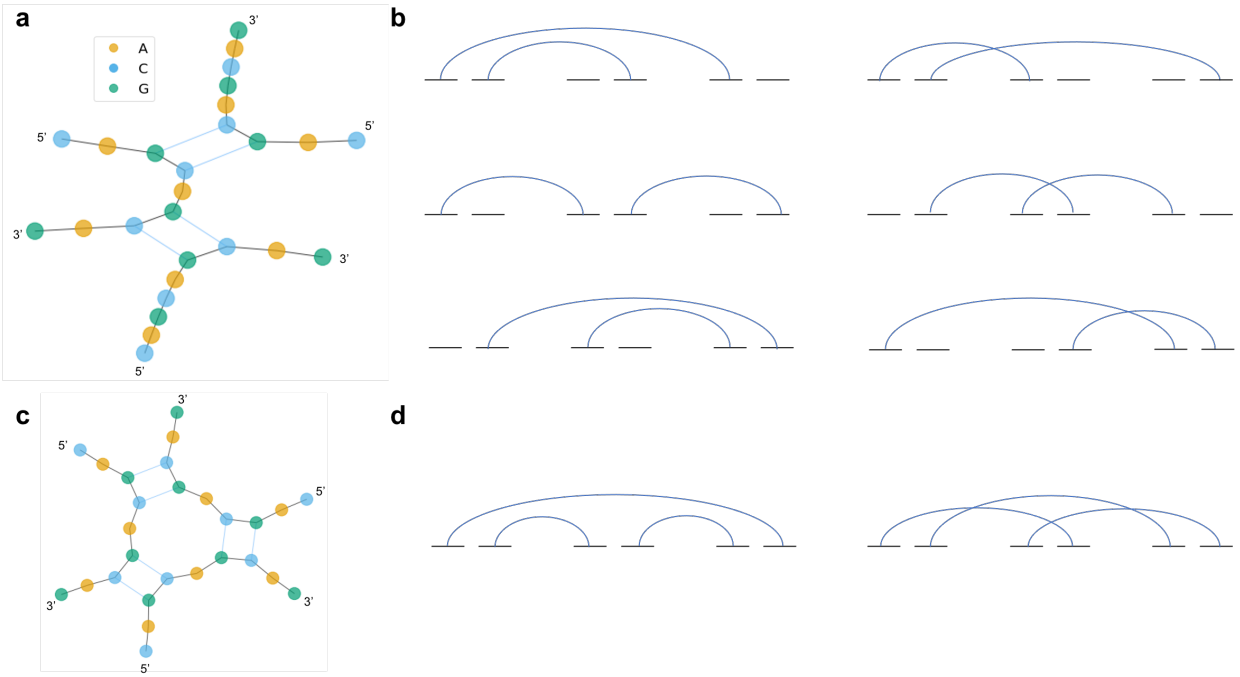
While the sum in $t(n, m, i)$ can be written more succinctly in terms of generating functions, we are not aware of any simple closed-form formula for $g_d(n, m, N_b)$. We therefore rely on the heuristic described previously for our analytical calculations.

1.2.3 A note about symmetry factors and g

In this work, we have used $g(n, m, N_b)$ to describe the number of *distinct* possible combinations of N_b bonds that can be formed with m strands, each with n stickers. It is this italicized word *distinct* that leads to the factor of $1/m$ in $g(n, m, N_b)$ that was missing in the previous sections – it ensures identical structures are counted only once. However, this factor leads to results that are apparently non-sensical: it is possible for g to result in a fractional output. As we will explain in this section, this is not an error, and in fact accounts for the symmetry-factor-based entropic cost of forming symmetric structures.

We begin by considering the expression for g_a derived previously (identical arguments can be made for g_d). $g_a(nm, 1, N_b)$ describes the number of – not necessarily distinct – possible combinations of N_b bonds that can be formed with m strands, each with n stickers, treating an m-mer the same as a monomer m times longer.

g_a overcounts multimers comprised of m strands by a factor of m. To demonstrate this, let us a consider one particular structure, for example, the structure shown in Supplementary Fig. 2a. The panel shows a particular trimer structure of $n = 3$ CAG repeat strands. In panel b, we show that this trimer can be depicted in six (or 3!) different



Supplementary Figure 2: **Repeat RNA enumeration multiplicity**. **a:** A trimer of $n = 3$ CAG repeat strands as depicted by LandscapeFold is shown. **b:** Structures can be depicted by treating each sticker as a site (black lines) that can be connected to at most one other site (blue arcs show connections). The structure from panel a can be depicted in six ways corresponding to permutations of the strands. Three of the ways (right column) appear pseudoknotted (they have intersecting arcs) though the structure contains no pseudoknots. **c:** Another trimer of $n = 3$ CAG repeats is shown. **d:** This trimer can only be depicted in two ways in the abstraction, one of which appears pseudoknotted. See main text for discussion.

ways in the enumeration performed by g_a, corresponding to different permutations of the strands. However, three of these (shown on the right) have intersecting arcs and therefore are not considered as part of g_a's enumeration. For 4-mers, there are 24 ways to enumerate each structure, but only 4 of these are considered within g_a as the rest appear as pseudoknots. Thus, in order to count the contribution of each structure to the partition function only once, we need to divide the contribution of each m-mer structure to Z by m. This corresponds to the number of cyclic permutations of the strands [1].

Some structures though are only enumerated once in g_a. In fact, structures that contain an R-fold symmetry are enumerated m/R times. However, these should should indeed be counted with a partition function penalty of $1/R$ (or a free energy penalty of $k_BT\log(R)$) [1]. This symmetry factor correction arises from the entropic difference between structures with and without these symmetries [2]. An example of such a structure is shown in panels c-d of Supplementary Fig. 2. The structure shown in panel c has 3-fold symmetry, and has only one possible depiction (panel d). Its free energy is thus effectively given by $3\Delta F_b - T\Delta S_{\text{loop}}(6) + 3\Delta G_{\text{assoc}} + k_BT\log(3)$. This latter term is effectively added by dividing its contribution to the partition function by 3 (the symmetry number). Thus, by dividing the contribution of each m-mer structure to Z by m we simultaneously correct for the overcounting performed by our enumeration procedure and account for the entropic penalty of symmetric structures.

1.2.4 An exact calculation of g for multimers

We will describe the exact calculation for $g(n, m, N_b)$ here. For most purposes, the exact calculation of g is overkill: $g(n, m, N_b)$ is well approximated by $g(nm, 1, N_b)/m$ (or a slight modification thereof when disallowing neighbor stickers from binding). However, in certain regimes – in particular for very small n or very positive values of F – this approximation is no longer valid. Moreover, we need to calculate g exactly if we are to make the claim that the approximation we use is appropriate.

The issue we address in this section is that of *connected* multimers. To give an example (neglecting the factor of $1/m$ for the moment): $g_a(2,2,2)$ should be equal to 1; there is one way to make a dimer with two bonds, given two strands each with two stickers. (Accounting for the symmetry factor, the structure has a two-fold symmetry, so that $g_a/m = 1/2$). However, $g_a(4,1,2) = 2$ is twice as large, since it also considers the structure which has each sticker bound to its neighbor on the same strand (a structure that is also two-fold symmetric). This second structure is not a dimer in truth – it is actually two monomers. This problem gets worse with larger m (for very small n). For example, $g_a(2,3,3)$ should be equal to 1: there is only one way to make a trimer with three bonds between three strands of $n = 2$ (and in fact, this trimer is shown in panels c and d of Supplementary Fig. 2). However, $g(6,1,3) = 5$ since it also considers the structure comprised of 3 monomers, as well as the three permutations of the same structure comprised of one dimer and one monomer.

Typically, corrections of this sort are negligible, since the vast majority of possible structures are connected. However, in the regime that very few stickers are typically bound, or that all stickers are bound with a very small value of n (such as $n = 2$), this is no longer the case.

An exact calculation of $g(n,m,N_b)$ therefore subtracts out the contribution of disconnected structures. We thus need to calculate all integer partitions of m, and for each, the number of ways to connect m strands (excluding connections that appear as pseudoknots when drawing the strands in a line, as those are already not counted). For example, for $m = 4$, there is 1 way to connect the strands as 4 monomers, 6 ways to connect them as 2 monomers and a dimer, 4 ways to connect them as a monomer and a trimer, and 2 ways to connect them as 2 dimers. We then need to enumerate, for each of these combinations, the different ways of splitting up the N_b bonds among the different sub-structures.

Thus, in order to calculate the connected multiplicity factor $g_{\text{conn}}(n,4,N_b)$, we calculate:

$$\begin{aligned} g_{\text{conn}}(n,4,N_b) = g_{a/d}(n,4,N_b) - \Bigg[& \sum_{i=0}^{N_b-j-k-l} \sum_{j=0}^{N_b-k-l} \sum_{k=0}^{N_b-l} \sum_{l=0}^{N_b} g(n,1,i)g(n,1,j)g(n,1,k)g(n,1,l) + \\ & 6 \sum_{i=0}^{N_b-j-k} \sum_{j=0}^{N_b-k} \sum_{k=0}^{N_b} g(n,1,i)g(n,1,j)g_{\text{conn}}(n,2,k) + \\ & 4 \sum_{i=0}^{N_b-j} \sum_{j=0}^{N_b} g(n,1,i)g_{\text{conn}}(n,3,j) + \\ & 2 \sum_{i=0}^{N_b-j} \sum_{j=0}^{N_b} g_{\text{conn}}(n,2,i)g_{\text{conn}}(n,2,j) \Bigg]. \end{aligned} \tag{S9}$$

Here, $g_{a/d}$ is the unconnected multiplicity as defined in Supplementary Notes 1.2.1 and 1.2.2 for the cases of allowing and disallowing neighbor binding, respectively (the former case simply yields $g_a(4n,1,N_b)$; the latter is more complicated as discussed in the referenced section). This unconnected multiplicity is corrected by subtracting the contribution of 4 monomers that have among them N_b total bonds, the contribution of 2 monomers and 1 dimer, and so on.

Finally, to get the true multiplicity factor, $g_{\text{conn}}(n,m,N_b)$ needs to be divided by m to account for symmetries as discussed in Supplementary Note 1.2.3.

Supplementary Note 2 Concentration units and their connection to ΔF

Throughout the text, we have treated concentrations as dimensionless. In other words, the concentrations we use are normalized by some reference concentration ρ. We have not defined ρ explicitly, because its value does not affect the physical observables we consider here. As explained in Ref. [1], any reference concentration used also enters into the definition of ΔG_{assoc}, as

$$\Delta G_{\text{assoc}} = \Delta G^{\text{pub}}_{\text{assoc}} - \frac{1}{\beta} \log\left(\frac{\rho}{1 \text{ mol/Liter}}\right) \tag{S10}$$

where $\Delta G^{\text{pub}}_{\text{assoc}}$ is the published value for the free energy cost of dimerization. This value is 4.09 kcal/mol for the free energy cost of RNA-RNA association [3, 4, 5]; 1.96 for the free energy cost of DNA-DNA association [6]; and 3.1 for the free energy cost of RNA-DNA association [7, 8].

The effect is that factors of ρ cancel out in the relevant equations. Namely,

$$\begin{aligned}
\frac{c_m/\rho}{(c_1/\rho)^m} &= \frac{Z_m}{Z_1^m} \\
\frac{c_m}{c_1^m}\rho^{m-1} &= \frac{Z'_m e^{-\beta(m-1)\Delta G_{\text{assoc}}}}{Z_1'^m} \\
\frac{c_m}{c_1^m}\rho^{m-1} &= \frac{Z'_m e^{-\beta(m-1)\Delta G^{\text{pub}}_{\text{assoc}}}}{Z_1^m}\left(\frac{\rho}{1\ M}\right)^{m-1}
\end{aligned} \tag{S11}$$

where $Z'_m \equiv Z_m e^{\beta(m-1)\Delta G_{\text{assoc}}}$ is Z_m not including the free energy cost of multimerization.

In fact, the natural units with which to measure concentration are $c/(e^{\beta \Delta G^{\text{pub}}_{\text{assoc}}} \times 1\ M)$, since the above equation could be further simplified to

$$\frac{c_m/e^{\beta \Delta G^{\text{pub}}_{\text{assoc}}}}{(c_1/e^{\beta \Delta G^{\text{pub}}_{\text{assoc}}})^m} = \frac{Z'_m}{Z_1^m} \tag{S12}$$

where c is now made dimensionless by measuring it in units of M, and the right-hand side is now independent of $G^{\text{pub}}_{\text{assoc}}$.

We chose to keep ΔG_{assoc} explicitly in the equations, in order to emphasize that changes to its value (for example, by changing ionic conditions) can affect aggregation properties. However, you will notice that both concentrations and $\Delta G^{\text{pub}}_{\text{assoc}}$ always appear in the combination $c/e^{\beta \Delta G^{\text{pub}}_{\text{assoc}}}$. Since we change units early on to $\Delta F \equiv \Delta G^{\text{pub}}_{\text{assoc}} - T\Delta S_{\text{loop}}$, that combination becomes $c/e^{\beta \Delta F}$.

Supplementary Note 3 Calculating the partition function analytically

The partition function is given by

$$Z_m = \sum_{N_b} g(n, m, N_b) e^{-\beta F N_b} e^{-\beta(m-1)\Delta F}. \tag{S13}$$

In principle, we can of course compute this sum. In practice however, such a computation is too computationally intensive for reasonable purposes. We therefore need to approximate the sum. We do so by a saddlepoint approximation, approximating the sum as being dominated by a particular term. We allow for a maximum of a second order correction in our approximations in order to balance computational feasibility and accuracy of the model. Indeed, these approximations appear to be entirely sufficient to describe the system with a high degree of accuracy (see Supplementary Fig. 3.

In this section, we will describe the different regimes of this saddlepoint approximation.

3.1 Allowing neighbor bonds

In this case, for $n \gg 1$, a good approximation for g is

$$g(n, m, N_b) = \frac{(nm)!}{m\ (nm - 2N_b)!\ (N_b + 1)!\ N_b!} \tag{S14}$$

(See Supplementary Note 1.2.4 for a description of the calculation for very small values of n). In order to proceed, we need an estimate of which term in the sum is dominant.

3.1.1 (Almost) all stickers are typically bound

The maximum possible number of stickers bound is

$$N_b^{\max} = \text{floor}\left(\frac{nm}{2}\right). \tag{S15}$$

Finding an approximation for g therefore depends on whether nm is even or odd.

Even nm
In this case, $N_b^{\max} = nm/2$. We then have

$$g(n, m, N_b^{\max}) = \frac{(nm)!}{m\ \left(\frac{nm}{2} + 1\right)\ \left[\left(\frac{nm}{2}\right)!\right]^2} \tag{S16}$$

where we have written the factorial in a way that will make the final result cleaner. We approximate the factorial with Stirling's approximation, $x! \approx \sqrt{2\pi x}(x/e)^x$, yielding

$$g(n, m, N_b^{\max}) \approx \frac{2^{nm}}{(nm + 2)\sqrt{\frac{n\pi}{8}} m^{3/2}} \tag{S17}$$

Odd nm
In this case, $N_b^{\max} = (nm - 1)/2$. We then have

$$g(n, m, N_b^{\max}) = \frac{nm(nm - 1)!}{m\ \left(\frac{nm+1}{2}\right)\ \left[\left(\frac{nm-1}{2}\right)!\right]^2} \tag{S18}$$

where, again, the factorials have been broken up to make the final expression cleaner. Using Stirling's approximation, we get

$$g(n, m, N_b^{\max}) \approx \frac{2^{nm-1}}{\frac{nm+1}{n}\sqrt{\frac{(nm-1)\pi}{8}}} \tag{S19}$$

We can then repeat this procedure to get the next order correction, meaning the term corresponding to $N_b^{\max} - 1$ bonds. All combined, we find that

$$Z_m \approx \begin{cases} e^{\left(\log(2)-\frac{\beta F}{2}\right)nm-\beta(m-1)\Delta F}\frac{1}{m}\left(\frac{1}{nm+2}\sqrt{\frac{8}{nm\pi}}+\sqrt{\frac{nm}{8\pi}}e^{\beta F}+\mathcal{O}\left(e^{2\beta F}\right)\right) & \text{if } nm \text{ is even} \\ e^{\left(\log(2)-\frac{\beta F}{2}\right)(nm-1)-\beta(m-1)\Delta F}\frac{1}{m}\left(\frac{nm}{nm+1}\sqrt{\frac{8}{(nm-1)\pi}}+\frac{nm}{3}\sqrt{\frac{nm-1}{8\pi}}e^{\beta F}+\mathcal{O}\left(e^{2\beta F}\right)\right) & \text{otherwise} \end{cases} \tag{S20}$$

in this regime.

For very negative βF, the first term is dominant. For even values of n, that implies that Z_m/Z_1^m is independent of F for very negative values. Furthermore, for intermediate values of βF and even n, Z_m/Z_1^m is actually larger than for very negative βF. This behavior is at the heart of the reentrant transition described in the main text, and can be seen in the exact computationally-enumerated partition function in Supplementary Fig. 4.

3.1.2 Calculating $N_b^\star$

For the regime in which some stickers are typically bound (next section) we need to calculate the dominant term of Z_m, corresponding to $N_b^\star$. This corresponds to the value of N_b such that

$$\frac{\partial g(n,m,N_b)e^{-\beta F N_b}}{\partial N_b} = \frac{\partial}{\partial N_b}\left[\frac{(nm)!\, e^{-\beta F N_b}}{m\,(nm-2N_b)!\,(N_b+1)!\,N_b!}\right] = 0. \tag{S21}$$

Calculating the derivative and simplifying, we arrive at

$$\frac{1}{N_b^\star+1} + \beta F + 2\psi^{(0)}(N_b^\star+1) = 2\psi^{(0)}(nm-2N_b^\star+1) \tag{S22}$$

where $\psi^{(m)}(z)$ is the polygamma function of order m, defined in terms of the gamma function as

$$\psi^{(m)}(z) = \frac{d^{m+1}}{dz^{m+1}}\log\Gamma(z). \tag{S23}$$

For large values of z, $\psi^{(0)}(z+1)$ is approximately given by $\log(z)$. This approximation directly applies in the intermediate binding regime: the argument of the left-hand-side polygamma function is large when not in the weak binding regime, and the argument of the right-hand-side function is large when not in the strong binding regime. Assuming the $N_b^\star \gg 1$, we also neglect the first term, arriving at

$$\beta F + 2\log(N_b^\star) = 2\log(nm-2N_b^\star). \tag{S24}$$

Solving this equation, we have

$$N_b^\star = \frac{nm}{2+e^{\beta F/2}} \tag{S25}$$

3.1.3 Some stickers are typically bound

As derived above, we find

$$N_b^\star \approx \frac{nm}{2+e^{\beta F/2}}. \tag{S26}$$

When $e^{\beta F/2} \ll 1$, the previous regime applies, and the discreteness of the number of stickers matters. In the regime with which we are concerned here, the fact that $N_b^\star$ is not an integer is insignificant, and in fact, as we will see, the sum can be well approximated as an integral.

This regime applies when each of the factorial terms in (S14) is greater than unity. We therefore approximate the factorial with Stirling's approximation, $x! \approx \sqrt{2\pi x}(x/e)^x$, yielding

$$g(n,m,N_b) \approx \frac{1}{2\pi m N_b(N_b+1)}\left(\frac{nm}{nm-2N_b}\right)^{nm+\frac{1}{2}}\left(\frac{nm-2N_b}{nm}\right)^{2N_b} \tag{S27}$$

Plugging in $N_b^\star$ for N_b, we have

$$g(n,m,N_b) \approx \frac{1}{2\pi m N_b^\star (N_b^\star+1)} \left(1+2e^{-\beta F/2}\right)^{nm+\frac{1}{2}} \left(e^{\beta F/2}\right)^{2N_b^\star} \tag{S28}$$

Thus, the dominant term of Z_m, (which we call $Z_m^\star$) is approximately

$$Z_m^\star \approx \frac{1}{2\pi m N_b^\star (N_b^\star+1)} \left(1+2e^{-\beta F/2}\right)^{nm+\frac{1}{2}} e^{-\beta(m-1)\Delta F} \tag{S29}$$

which can be written more tellingly as

$$Z_m^\star \approx \frac{e^{\beta\Delta F}\left(1+2e^{-\beta F/2}\right)^{1/2}}{2\pi m N_b^\star (N_b^\star+1)} \left[\left(1+2e^{-\beta F/2}\right)^{n} e^{-\beta\Delta F}\right]^m \tag{S30}$$

Written this way, connections between this expression and that found in the previous regime are apparent. In particular, it is clear why $\log Z$ increases linearly with both n and m, and the factor of 2^{nm} makes an appearance here as in the previous regime. However, understanding the behavior of Z_m/Z_1^m involves understanding the behavior of the prefactor to the bracketed term.

We also consider the next-order correction to Z_m. This is found using the saddlepoint approximation. For an exponential integrand $e^{f(x)}$ that has a maximum at $x^\star$ (and therefore $f'(x^\star)=0$), the following becomes a very good approximation since exponentials are so sharply peaked:

$$\int_{-\infty}^{\infty} e^{f(x)}dx \approx \int_{-\infty}^{\infty} e^{f(x^\star)+xf'(x^\star)+\frac{x^2}{2}f''(x^\star)}dx = e^{f(x^\star)}\sqrt{\frac{2\pi}{|f''(x^\star)|}}. \tag{S31}$$

For our purposes, $f(N_b) = \log\left(g(n,m,N_b)\right) - \beta F N_b - \beta(m-1)\Delta F$. What we have done so far – finding $Z_m^\star$ – is equivalent in this language to finding $e^{f(x^\star)}$. Since $|f''(N_b^\star)|$ is generally quite small ($\mathcal{O}(1)$ as a general rule, < 6 in all cases we examined), the error introduced by the approximation of the sum as an integral is negligible. In this case, the curvature term $f''(x^\star)$ leads to

$$Z_m \approx Z_m^\star \sqrt{\frac{2\pi}{4\psi^{(1)}(n-2N_b^\star+1)+2\psi^{(1)}(N_b^\star+1)-\frac{1}{\left(N_b^\star+1\right)^2}}}. \tag{S32}$$

We compare this analytical formula to the computational results for intermediate binding ($F_b = -6$ kcal/mol) in Supplementary Fig. 5. Here we use the parameter l_{eff} fit to data from Fig. 2 along with (10).

3.1.4 Very few stickers are typically bound

This is the regime in which the typical number of stickers in a multimer of size m is equal to or only slightly larger than $m-1$. This is the most computationally difficult regime to explore, since in this regime, $g(n,m,N_b)$ is not well approximated by $g(nm,1,N_b)/m$. Instead, we need to employ the exact calculation of $g(n,m,N_b)$ described in Supplementary Note 1.2.4. We consider only the first three terms of the sum in this regime, enumerating the contribution of $N_b = m-1$ up to $m-3$. We do not consider this regime further analytically, but do employ it in comparison to the computational predictions (e.g. Fig. 2B).

3.2 Disallowing neighbor bonds

In this case, a good approximation for $g(n,m,N_b)$ is given by

$$g(n,m,N_b) \approx \frac{(q-N_b)!\,(q-N_b-1)!}{m\,(q-2N_b)!\,(q-2N_b-1)!\,(N_b+1)!\,N_b!} \tag{S33}$$

where $q = nm+\alpha(m-1)$, with $\alpha \approx 0.42$.

We will follow a similar procedure in this section as that taken in the previous section, but the calculations are slightly more cumbersome here because of the increased complexity of g.

3.2.1 (Almost) all stickers are typically bound

In this regime, we find that we get a good approximation for Z by considering the monomer case, and substituting in q for n after simplifying.

With neighbor bonds disallowed, the maximum possible number of stickers bound for monomers is

$$N_b^{\max} = \text{floor}\left(\frac{n-1}{2}\right). \tag{S34}$$

Finding an approximation for g therefore again depends on whether n is even or odd. The procedure here is analogous – with slightly more involved calculations – to the procedure for calculating Z allowing for neighbor binding.

Odd n

The final term in the sum for Z_1 for odd n is given by $N_b = \frac{n-1}{2}$. After simplifying the expression for g in this case, we find that there is only one possible way to arrange this number of bonds in a monomer. The next term corresponds to a value of N_b given by $\frac{n-3}{2}$. Combined, we find that in this regime,

$$Z_1 = e^{-\frac{\beta F}{2}(n-1)}\left(1 + \frac{(n+3)(n+1)^2(n-1)}{192}e^{\beta F}\right). \tag{S35}$$

Even n

The final term in the sum for Z_1 for even n is given by $N_b = \frac{n}{2} - 1$. Simplifying the factorials and including the second-to-last term as well, we find that for even n in this regime,

$$Z_1 = e^{-\frac{\beta F}{2}(n-2)}\left(\frac{n(n+2)}{8}\right)\left(1 + \frac{(n+4)(n+2)n(n-2)}{1152}e^{\beta F}\right). \tag{S36}$$

To convert from Z_1 to Z_m, we first multiply by $\frac{e^{-\beta(m-1)\Delta F}}{m}$. Second, we (heuristically) replace each instance of n in these expressions by q. Finally, for even values of nm with $m > 1$, multimers are described by this equation with the exception that they are able to form one further bond (such that all stickers are bonded). We can account for this by adding a term

$$\frac{e^{-\beta(m-1)\Delta F}}{m}e^{-\frac{\beta F}{2}q} \tag{S37}$$

to the partition function for even nm. This term is primarily useful for $m = 2$, for which the multiplicity factor of one is accurate (for $m = 2$, there is indeed only one way to fulfill all bonds, when disallowing neighbor binding). For $m > 2$, the replacement of n by q partially accounts for the lack of a further multiplicity factor; furthermore, we ultimately find that multimers for $m > 2$ are fairly well approximated by the intermediate regime – considered in the following sections – for the free energies we consider. Put together, we have

$$Z_{m>1} \approx \begin{cases} \frac{e^{-\beta(m-1)\Delta F}}{m}e^{-\frac{\beta F}{2}(q-1)}\left(1 + \frac{(q+3)(q+1)^2(q-1)}{192}e^{\beta F}\right) & \text{if } nm \text{ is odd} \\ \frac{e^{-\beta(m-1)\Delta F}}{m}\left[e^{-\frac{\beta F}{2}(q-1)}\left(1 + \frac{(q+3)(q+1)^2(q-1)}{192}e^{\beta F}\right) + e^{-\frac{\beta F}{2}q}\right] & \text{if } n \text{ is odd, } m \text{ is even} \\ \frac{e^{-\beta(m-1)\Delta F}}{m}\left[e^{-\frac{\beta F}{2}(q-2)}\left(\frac{q(q+2)}{8}\right)\left(1 + \frac{(q+4)(q+2)q(q-2)}{1152}e^{\beta F}\right) + e^{-\frac{\beta F}{2}q}\right] & \text{if } n \text{ is even.} \end{cases} \tag{S38}$$

where $q = nm + \alpha(m-1)$, with $\alpha \approx 0.42$.

3.2.2 Calculating $N_b^\star$

For the regime in which some stickers are typically bound (next section) we need to calculate the dominant term of Z_m, corresponding to $N_b^\star$. This corresponds to the value of N_b such that

$$\frac{\partial g(n,m,N_b)e^{-\beta F N_b}}{\partial N_b} = \frac{\partial}{\partial N_b}\left[\frac{((q-N_b)!)^2\ (q-2N_b)\ e^{-\beta F N_b}}{m\ (q-N_b)\ ((q-2N_b)!)^2\ (N_b+1)\ (N_b!)^2}\right] = 0. \tag{S39}$$

Calculating the derivative and simplifying, we arrive at

$$\frac{2}{q-2N_b^\star}+\frac{1}{N_b^\star+1}-\frac{1}{q-N_b^\star}+\beta F-4\psi^{(0)}(q-2N_b^\star+1)+2\psi^{(0)}(q-N_b^\star+1)+2\psi^{(0)}(N_b^\star+1)=0. \quad \text{(S40)}$$

Recognizing that in the intermediate regime, $q-2N_b^\star>1$, $N_b^\star+1>1$, and $q-N_b^\star>1$, we can make the approximation of neglecting the first three terms. We then make the same approximation as previously, treating $\psi^{(0)}(z+1)\approx\log(z)$. Solving the resulting equation and recognizing that $N_b^\star<\frac{q}{2}$ to remove the extraneous solution, we arrive at

$$N_b^\star\approx\frac{q}{2}\left(1-\frac{e^{\beta F/4}}{\sqrt{4+e^{\beta F/2}}}\right). \quad \text{(S41)}$$

3.2.3 Some stickers are typically bound

As derived above, we find that for monomers

$$N_b^\star\approx\frac{n}{2}\left(1-\frac{e^{\beta F/4}}{\sqrt{4+e^{\beta F/2}}}\right). \quad \text{(S42)}$$

where, as previously, we substitute q for n when considering multimers.

The dominant term of Z for monomers corresponds to

$$g\approx\frac{(n-2N_b^\star)\left[(n-N_b^\star)!\right]^2}{(n-N_b^\star)(N_b^\star+1)\left[(n-2N_b^\star)!\right]^2\left[(N_b^\star)!\right]^2}. \quad \text{(S43)}$$

Using Stirling's approximation, this expression becomes

$$g\approx\frac{1}{2\pi mN_b^\star(N_b^\star+1)}\left(\frac{n-N_b^\star}{n-2N_b^\star}\right)^{2(n-N_b^\star)}\left(\frac{n}{N_b^\star}-2\right)^{2N_b^\star}. \quad \text{(S44)}$$

As previously, we consider not only the dominant term of Z but also the curvature term of the saddlepoint approximation. This term can be most clearly written in terms of the following function

$$h(x)=x^{-2}-2\psi^{(1)}(x) \quad \text{(S45)}$$

as

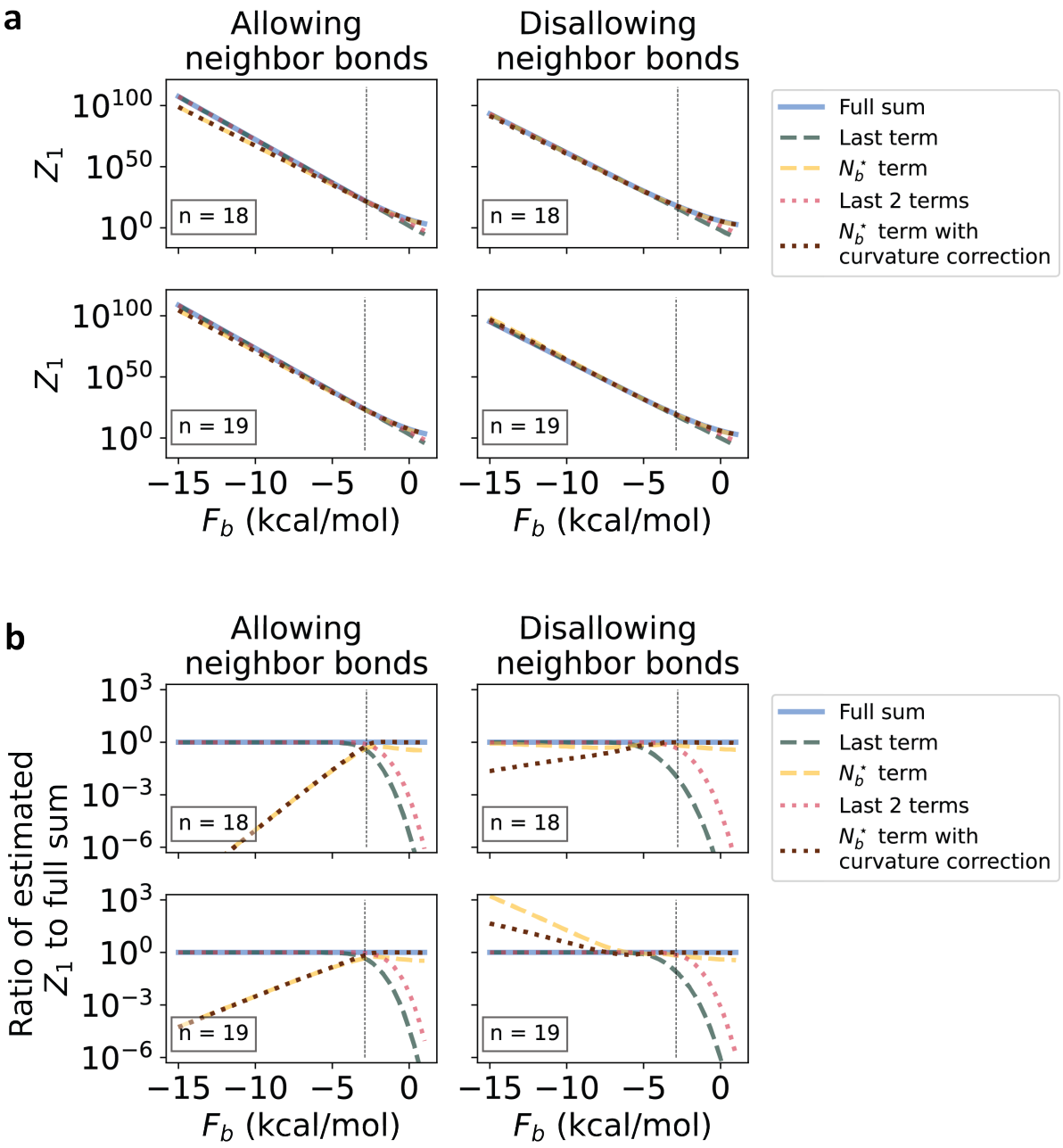
$$\sqrt{\frac{2\pi}{-4h(n-2N_b^\star)+h(n-N_b^\star)-h(N_b^\star+1)}}. \quad \text{(S46)}$$

Putting this all together, in this regime, we approximate Z_m as

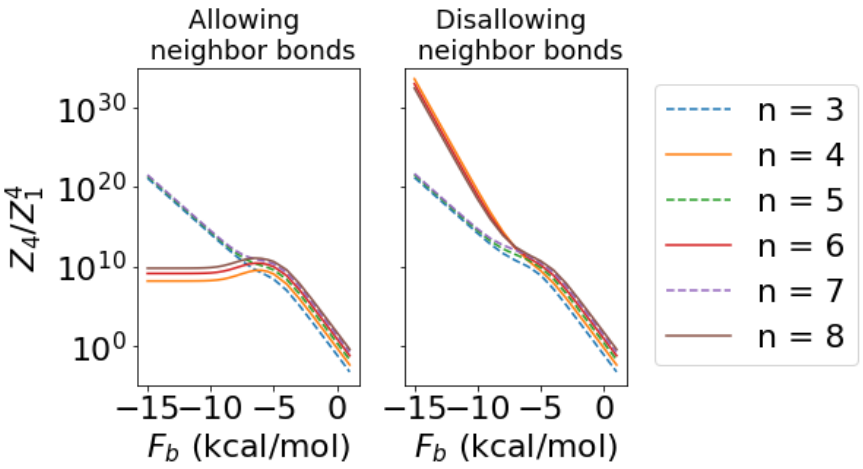
$$Z_m\approx Z_m^\star\sqrt{\frac{2\pi}{-4h(q-2N_b^\star)+h(q-N_b^\star)-h(N_b^\star+1)}} \quad \text{(S47)}$$

where

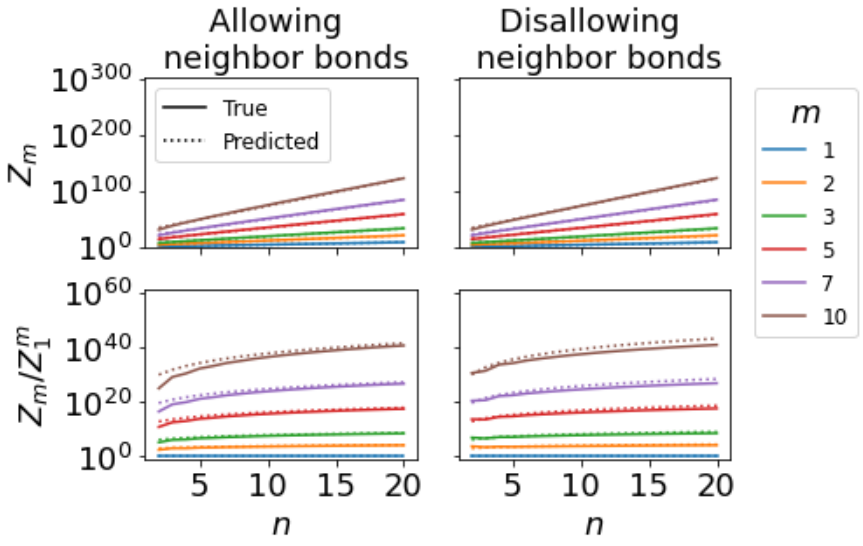
$$\begin{aligned}Z_m^\star=&\frac{e^{-\beta(FN_b^\star+(m-1)\Delta F)}}{2\pi mN_b^\star(N_b^\star+1)}\left(\frac{q-N_b^\star}{q-2N_b^\star}\right)^{2(q-N_b^\star)}\left(\frac{q}{N_b^\star}-2\right)^{2N_b^\star}\\ N_b^\star=&\frac{q}{2}\left(1-\frac{e^{\beta F/4}}{\sqrt{4+e^{\beta F/2}}}\right).\end{aligned} \quad \text{(S48)}$$



Supplementary Figure 3: **Accuracy of saddlepoint approximation**. The full sum calculating Z_1 (with the Stirling approximation made for the factorials) is plotted alongside four versions of the saddlepoint approximation: the first-order (green and yellow, dashed) and second-order-corrected (pink and magenta, dotted) approximations in both the strong and intermediate regimes. Panel **A** shows the full partition function; panel **B** shows the ratio of the estimated partition function to the full sum. A dashed vertical line plots the expected crossover point between the strong and intermediate regimes. Indeed, the strong approximation (last term) and intermediate approximation ($N_b^\star$ term) provide good agreement with the full sum in their respective regimes, with the second-order-corrected approximations improving the accuracy.



Supplementary Figure 4: **Partition functions as a function of F_b.** Z_4/Z_1^4 is plotted as a function of F_b using the exact computationally-enumerated partition functions. The surprising behavior predicted by the analytical model and discussed in the main text can be seen clearly, namely that at the interface between the strong and intermediate binding regimes for even n, the partition function ratio is larger than it is in the very strong binding regime. This behavior is seen regardless of the value of m chosen (here, $m = 4$).



Supplementary Figure 5: **Partition functions in the intermediate regime**. Z_m and Z_m/Z_1^m are plotted as in Fig. 2 with $F_b = -6$ kcal/mol. A single parameter for each column is fit to the results in Fig. 2 as described in the discussion of that figure, and extrapolated to $F_b = -6$ kcal/mol using (10). As predicted, there is no even/odd discrepancy for n in this intermediate regime of binding.

Supplementary Note 4 Predicting the aggregation threshold

Once we have the partition functions, we can then compute the equilibrium concentration of each multimer. These are found by solving the following set of equations:

$$\begin{aligned} c_m &= \frac{Z_m}{Z_1^m} c_1^m \\ \sum_m m c_m &= c^{\text{tot}} \end{aligned} \tag{S49}$$

where the concentrations are made dimensionless by normalizing by a reference concentration (see Supplementary Note 2) and c^{tot} is the total concentration of strands added to solution.

Aggregation is predicted when c_m grows with m. However, these equations have c_m dependent on c_1. If m has a finite maximum value m_{max}, we can simply solve the m_{max} equations for c_1, plug that in, and observe the dependence on m of c_m. Indeed, this is how we make predictions of c_m in our computational model. There is also some finite value to m_{max} in an experiment given by the total number of molecules in the solution. However, that number is far too large to treat *via* this approach, and is reasonably treated as infinite. How can we find the aggregation threshold allowing for arbitrarily large clusters?

In this section, we will demonstrate how this question can be addressed. We will first consider the case allowing neighbor binding, and then the case disallowing neighbor binding. For each, the simplest case to consider is that in which monomers have some of their stickers typically bound (i.e. the intermediate regime considered above), and that is where we will begin.

Throughout this section, we will consider only the dominant term of the sum for Z_m, without any higher-order corrections. Some accuracy is therefore compromised for the sake of computational feasibility, but errors are expected to be minor.

4.1 General framework

We start with a simple example to demonstrate the framework of the calculation. Consider a partition function

$$Z_m = \kappa m^{-\lambda} \gamma^m. \tag{S50}$$

Partition functions for this system are often approximately in this form. The concentration of m-mers c_m is given in terms of c_1 by

$$c_m = Z_m \left(\frac{c_1}{Z_1}\right)^m = \kappa m^{-\lambda} \left(\frac{c_1}{\kappa}\right)^m. \tag{S51}$$

The total concentration is given by

$$c^{\text{tot}} = \sum_m m c_m = \kappa \sum_m m^{-\lambda+1} \left(\frac{c_1}{\kappa}\right)^m. \tag{S52}$$

We now recognize that the right-hand-side has a maximum possible value it can reach for $c_1 \leq \kappa$, given by $\kappa\zeta(\lambda - 1)$ where ζ is the Riemann zeta function:

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}. \tag{S53}$$

if $c^{\text{tot}} > \kappa\zeta(\lambda - 1)$, no value of $c_1 \leq \kappa$ will satisfy (S52). In that case, c_1/κ must be > 1, and thus c_m increases with m for large enough m. The aggregation threshold is thus defined by

$$c^{\text{tot}}_{\text{thresh}} = \kappa\zeta(\lambda - 1). \tag{S54}$$

For concentrations greater than $c^{\text{tot}}_{\text{thresh}}$, the system is expected to form aggregates; for smaller concentrations, c_m exponentially decreases with m.

To be clear, for $c^{\text{tot}} > c^{\text{tot}}_{\text{thresh}}$ (i.e. $c_1 > \kappa$) the sum in (S52) will actually diverge, an outcome that is both unphysical and unreasonable mathematically given that the left hand side is finite. The solution is that we are neglecting here

any excluded volume interactions which will contribute a term e^{-vm^2} in the summand and ensure the sum does not actually diverge, and physically mean that we will still expect to form finite-sized clusters in the aggregation regime. However, since we expect v to be small, the integrand will still increase with m within a certain regime for $c_1/\kappa > p$, where p is only fractionally larger than unity. We will continue to neglect these effects here, since they are not, for our purposes, instructive.

4.2 Allowing neighbor bonds

4.2.1 Monomers have some stickers typically bound

The number of stickers typically bound is approximately

$$N_s(m) \approx \frac{2nm}{2 + e^{\beta F/2}} \tag{S55}$$

Here we are treating values of n and βF such that $n - 2 > N_s(1) > 6$ (or so) such that the intermediate regime approximation applies for monomers.

A question arises: if monomers are in the intermediate regime, how about multimers of size m? Because $N_s(m)$ scales linearly with m, a constant fraction of sites are expected to be bound, regardless of m. Therefore, the intermediate regime will always apply for multimers, if it applies for monomers.

We found previously that the partition function in this regime is approximately given by

$$Z_m = \frac{e^{\beta\Delta F}\left(1 + 2e^{-\beta F/2}\right)^{1/2}}{2\pi m N_b^\star(N_b^\star + 1)}\left[\left(1 + 2e^{-\beta F/2}\right)^n e^{-\beta\Delta F}\right]^m. \tag{S56}$$

The bracketed term cancels exactly in the equation for Z_m/Z_1^m, yielding

$$\frac{Z_m}{Z_1^m} = \frac{1}{m^2}\left[\frac{e^{\beta\Delta F}\left(1 + 2e^{-\beta F/2}\right)^{1/2}\left(2 + e^{\beta F/2}\right)^2}{2\pi n\left(n + 2 + e^{\beta F/2}\right)}\right]^{1-m} \frac{n + 2 + e^{\beta F/2}}{nm + 2 + e^{\beta F/2}}. \tag{S57}$$

We can then combine equations (S49) with this to get

$$c^{\text{tot}} = \frac{e^{\beta\Delta F}\left(1 + 2e^{-\beta F/2}\right)^{1/2}\left(2 + e^{\beta F/2}\right)^2}{2\pi n}\sum_{m=1}^{\infty}\frac{1}{m\left(nm + 2 + e^{\beta F/2}\right)}\left[x(c_1)\right]^m \tag{S58}$$

where we have defined $x(c_1)$ to be

$$x(c_1) = \frac{2\pi n\left(n + 2 + e^{\beta F/2}\right)c_1}{e^{\beta\Delta F}\left(1 + 2e^{-\beta F/2}\right)^{1/2}\left(2 + e^{\beta F/2}\right)^2} \tag{S59}$$

We will discuss in the next section a way to use this equation as is, but for now, in order to proceed will will make the assumption that $nm \gg 2 + e^{\beta F/2}$. This is equivalent to the assumption that $N_b^\star \gg 1$, which is entirely reasonable in this regime. This allows us to write

$$c^{\text{tot}} = \sum_m m\, c_m \approx \frac{e^{\beta\Delta F}\left(1 + 2e^{-\beta F/2}\right)^{1/2}\left(2 + e^{\beta F/2}\right)^2}{2\pi n^2}\sum_{m=1}^{\infty}\frac{\left[x(c_1)\right]^m}{m^2} \tag{S60}$$

Since the prefactor to the sum is independent of m, the statement that c_m increases with m (our definition of aggregation) is equivalent to the statement that $x(c_1) > 1$. Once $x(c_1) > 1$, then for large enough m, the summand will be increasing with m.

We have now refined our problem of finding the aggregation threshold as finding the set of parameters for which $x(c_1) > 1$. However, we still don't have an estimate for c_1! How then can we find for what set of parameters $x(c_1) > 1$?

Let us rewrite (S60) in a more suggestive form:

$$\frac{2\pi n^2 c^{\text{tot}}}{e^{\beta\Delta F}\left(1 + 2e^{-\beta F/2}\right)^{1/2}\left(2 + e^{\beta F/2}\right)^2} = \sum_{m=1}^{\infty}\frac{\left[x(c_1)\right]^m}{m^2} \tag{S61}$$

The key comes from recognizing that for $x(c_1) \leq 1$, the right hand side has a maximum value it can reach, given by the Riemann zeta function $\zeta(2) = \pi^2/6$. If the left hand side is greater than that value, $x(c_1)$ *must be* > 1. (And again – the reason the sum nonetheless evaluates to a finite value is because of the excluded volume interactions we are omitting here). Since the left hand side is comprised only of terms we control directly – c^{tot}, n, βF, and $\beta\Delta F$ – we have found our condition for aggregation in this regime:

$$\frac{n^2 c^{\text{tot}}}{e^{\beta\Delta F}\left(1+2e^{-\beta F/2}\right)^{1/2}\left(2+e^{\beta F/2}\right)^2} > \frac{\pi}{12}. \tag{S62}$$

4.2.2 Monomers have (almost) all stickers typically bound

If $\frac{2n}{2+e^{\beta F/2}} \gtrsim n-2$, monomers typically have all or almost all their stickers bound. In that case, Z_1 is best approximated by that regime. However, for large enough m, Z_m will be best approximated by the intermediate regime. This is because an approximately constant fraction of the stickers is typically bound, and for larger m, that fraction will correspond to a greater number of unbound stickers. We define $m^\star + 1$ to be the smallest value of m for which the intermediate regime applies.

Even n

For even n, we found previously that Z_1 is approximately given by

$$Z_1 = e^{\left(\log(2)-\frac{\beta F}{2}\right)n}\left(\frac{1}{n+2}\sqrt{\frac{8}{n\pi}}\right). \tag{S63}$$

Since when n is even, nm is also always even, we have

$$Z_{m\leq m^\star} = e^{\left(\log(2)-\frac{\beta F}{2}\right)nm-\beta(m-1)\Delta F}\frac{1}{m}\left(\frac{1}{nm+2}\sqrt{\frac{8}{nm\pi}}\right). \tag{S64}$$

This gives a ratio

$$\frac{Z_{m\leq m^\star}}{Z_1^m} = m^{-3/2}\frac{n+2}{nm+2}\left[e^{-\beta\Delta F}(n+2)\sqrt{\frac{n\pi}{8}}\right]^{m-1} \tag{S65}$$

For $m > m^\star$, we have

$$Z_{m>m^\star} = \frac{e^{\beta\Delta F}\left(1+2e^{-\beta F/2}\right)^{1/2}}{2\pi m\left(\frac{nm}{2+e^{\beta F/2}}\right)\left(\frac{nm}{2+e^{\beta F/2}}+1\right)}\left[\left(1+2e^{-\beta F/2}\right)^n e^{-\beta\Delta F}\right]^m \tag{S66}$$

yielding the ratio

$$\frac{Z_{m>m^\star}}{Z_1^m} = \frac{e^{\beta\Delta F}\left(1+2e^{-\beta F/2}\right)^{1/2}}{2\pi m\left(\frac{nm}{2+e^{\beta F/2}}\right)\left(\frac{nm}{2+e^{\beta F/2}}+1\right)}\left[e^{-\beta\Delta F}\left(1+\frac{e^{\beta F/2}}{2}\right)^n(n+2)\sqrt{\frac{n\pi}{8}}\right]^m. \tag{S67}$$

We follow the same protocol as in the previous section, but our sum is now broken up into two parts:

$$\begin{aligned} c^{\text{tot}} &= \sum_{m=1}^{m^\star} m\frac{Z_{m\leq m^\star}}{Z_1^m}c_1^m + \sum_{m=m^\star+1}^{\infty} m\frac{Z_{m>m^\star}}{Z_1^m}c_1^m \\ &= e^{\beta\Delta F}\sqrt{\frac{8}{n\pi}}\sum_{m=1}^{m^\star}\frac{1}{(nm+2)\sqrt{m}}\left[y(c_1)\right]^m + \\ &\quad \frac{e^{\beta\Delta F}\left(1+2e^{-\beta F/2}\right)^{1/2}\left(2+e^{\beta F/2}\right)}{2\pi n}\sum_{m=m^\star+1}^{\infty}\frac{1}{m\left(\frac{nm}{2+e^{\beta F/2}}+1\right)}\left[y(c_1)\left(1+\frac{e^{\beta F/2}}{2}\right)^n\right]^m \end{aligned} \tag{S68}$$

where we have defined

$$y(c_1) = e^{-\beta\Delta F}(n+2)\sqrt{\frac{n\pi}{8}}\, c_1 \tag{S69}$$

If we make the approximation that $n \gg 2$ (for the first sum) and $\frac{nm^\star}{2+e^{\beta F/2}} \gg 1$ (for the second), the equations become a bit cleaner. Note that we could also not make such an approximation, and instead further split each sum in two; the first new sum (for smaller values of m) would then remain unapproximated and need to be computed explicitly, and the second (for larger values) will be approximated with the aforementioned approximations. For clarity and simplicity, we do not split the sums up further here. We thus arrive at the slighly simpler equation

$$c^{\text{tot}} = e^{\beta\Delta F}\sqrt{\frac{8}{n^3\pi}}\sum_{m=1}^{m^\star} m^{-3/2}\left[y(c_1)\right]^m + \frac{e^{\beta\Delta F}\left(1+2e^{-\beta F/2}\right)^{1/2}\left(2+e^{\beta F/2}\right)^2}{2\pi n^2}\sum_{m=m^\star+1}^{\infty} m^{-2}\left[y(c_1)\left(1+\frac{e^{\beta F/2}}{2}\right)^n\right]^m. \tag{S70}$$

The first sum will always be finite; aggregation is predicted when $y(c_1) > \left(1+\frac{e^{\beta F/2}}{2}\right)^{-n}$ (and therefore the second sum diverges). This inequality is substituted by an equality at the aggregation threshold itself, thus defining the value of c_1 at the threshold. By plugging in this value into the first sum, we find that the aggregation is predicted when $c^{\text{tot}} > c^{\text{tot}}_{\text{thresh}}$, defined by

$$c^{\text{tot}}_{\text{thresh}} = e^{\beta\Delta F}\sqrt{\frac{8}{n^3\pi}}\sum_{m=1}^{m^\star} m^{-3/2}\left[\left(1+\frac{e^{\beta F/2}}{2}\right)^{-n}\right]^m + \frac{e^{\beta\Delta F}\left(1+2e^{-\beta F/2}\right)^{1/2}\left(2+e^{\beta F/2}\right)^2}{2\pi n^2}\sum_{m=m^\star+1}^{\infty} m^{-2}. \tag{S71}$$

The result of the second sum is given by $\psi^{(1)}(m^\star+1)$ where ψ is the polygamma function defined previously. The result of the first sum can either be written in terms of the Lerch transcendent Φ and polylogarithm function $Li_{3/2}$, or simply evaluated directly. The latter approach needs no further explanation, and for many purposes is the most straightforward approach to take. To demonstrate the former approach, we define the two functions here:

$$\begin{aligned}\Phi(z,s,\alpha) =& \sum_{k=0}^{\infty}\frac{z^k}{(k+\alpha)^s}\\ Li_s(z) =& \Phi(z,s,0) = \sum_{k=0}^{\infty}\frac{z^k}{k^s}\end{aligned} \tag{S72}$$

and use them to write

$$\sum_{m=1}^{m^\star} m^{-3/2}\left[\left(1+\frac{e^{\beta F/2}}{2}\right)^{-n}\right]^m = Li_{3/2}\left(\left(1+\frac{e^{\beta F/2}}{2}\right)^{-n}\right) - \left(1+\frac{e^{\beta F/2}}{2}\right)^{-n(1+m^\star)}\Phi\left(\left(1+\frac{e^{\beta F/2}}{2}\right)^{-n}, \frac{3}{2}, m^\star+1\right). \tag{S73}$$

The benefit of writing the sum in this way is that these two functions have been deeply categorized and explored mathematically (and that therefore, functions in Python, Mathematica, and similar programs can be used to evaluate them very efficiently). Putting it together, we find that we predict aggregation in this regime when

$$\frac{c^{\text{tot}}}{e^{\beta\Delta F}} > \frac{\left(1+2e^{-\beta F/2}\right)^{1/2}\left(2+e^{\beta F/2}\right)^{2}}{2\pi n^2}\psi^{(1)}(m^\star+1)+ \sqrt{\frac{8}{n^3\pi}}\Bigg[Li_{3/2}\left(\left(1+\frac{e^{\beta F/2}}{2}\right)^{-n}\right)- \left(1+\frac{e^{\beta F/2}}{2}\right)^{-n(1+m^\star)}\Phi\left(\left(1+\frac{e^{\beta F/2}}{2}\right)^{-n},\frac{3}{2},m^\star+1\right)\Bigg]. \quad \text{(S74)}$$

A reasonable estimate of $m^\star$ is the value of m for which $N_s(m^\star) = nm^\star - 2$, yielding

$$m^\star = \frac{2+4e^{-\beta F/2}}{n}. \quad \text{(S75)}$$

Plugging this result back into the previous equation we find a prediction for the concentration threshold that depends only on the values of n, βF, and $\beta\Delta F$.

Odd n

We follow the same procedure for odd values of n. However, now, $m < m^\star$, the value of Z_m depends on whether m is even or odd. We therefore now have three cases. Besides $Z_{m\leq m^\star,\,\text{even}}$ and $Z_{m>m^\star}$ which are equivalent to the corresponding expressions for the even n case, we have

$$\begin{aligned} Z_1 &= e^{\left(\log(2)-\frac{\beta F}{2}\right)(n-1)}\left(\frac{n}{n+1}\sqrt{\frac{8}{(n-1)\pi}}\right) \\ Z_{m\leq m^\star,\,\text{odd}} &= e^{\left(\log(2)-\frac{\beta F}{2}\right)(nm-1)-\beta(m-1)\Delta F}\frac{1}{m}\left(\frac{nm}{nm+1}\sqrt{\frac{8}{(nm-1)\pi}}\right). \end{aligned} \quad \text{(S76)}$$

The three cases are

$$\begin{aligned} \frac{Z_{m\leq m^\star,\,\text{odd}}}{Z_1^m} &= \left[2e^{-\frac{\beta F}{2}-\beta\Delta F}\left(\frac{n+1}{n}\right)\left(\frac{(n-1)\pi}{8}\right)^{1/2}\right]^m \frac{n}{nm+1}\left(\frac{8}{(nm-1)\pi}\right)^{1/2}\frac{e^{\frac{\beta F}{2}+\beta\Delta F}}{2} \\ \frac{Z_{m\leq m^\star,\,\text{even}}}{Z_1^m} &= \left[2e^{-\frac{\beta F}{2}-\beta\Delta F}\left(\frac{n+1}{n}\right)\left(\frac{(n-1)\pi}{8}\right)^{1/2}\right]^m \frac{1}{m(nm+2)}\left(\frac{8}{nm\pi}\right)^{1/2}e^{\beta\Delta F} \\ \frac{Z_{m>m^\star}}{Z_1^m} &= \left[\left(1+\frac{e^{\beta F/2}}{2}\right)^n 2e^{-\frac{\beta F}{2}-\beta\Delta F}\left(\frac{n+1}{n}\right)\left(\frac{(n-1)\pi}{8}\right)^{1/2}\right]^m \frac{e^{\beta\Delta F}\left(1+2e^{-\beta F/2}\right)^{1/2}}{2\pi m\left(\frac{nm}{2+e^{\beta F/2}}\right)\left(\frac{nm}{2+e^{\beta F/2}}+1\right)} \end{aligned} \quad \text{(S77)}$$

In keeping with the previous procedure, we define $z(c_1)$ to be

$$z(c_1) = 2e^{-\frac{\beta F}{2}-\beta\Delta F}\left(\frac{n+1}{n}\right)\left(\frac{(n-1)\pi}{8}\right)^{1/2}c_1. \quad \text{(S78)}$$

The total concentration is then given by

$$c^{\text{tot}} = \sum_{m=1,3,5\ldots}^{m^\star} \frac{nm}{nm+1}\left(\frac{8}{(nm-1)\pi}\right)^{1/2} \frac{e^{\frac{\beta F}{2}+\beta\Delta F}}{2}\left[z(c_1)\right]^m + \sum_{m=2,4,6\ldots}^{m^\star} \frac{1}{nm+2}\left(\frac{8}{nm\pi}\right)^{1/2} e^{\beta\Delta F}\left[z(c_1)\right]^m + \sum_{m=m^\star+1}^{\infty} \frac{e^{\beta\Delta F}\left(1+2e^{-\beta F/2}\right)^{1/2}}{2\pi\left(\frac{nm}{2+e^{\beta F/2}}\right)\left(\frac{nm}{2+e^{\beta F/2}}+1\right)}\left[\left(1+\frac{e^{\beta F/2}}{2}\right)^n z(c_1)\right]^m. \quad \text{(S79)}$$

As previously, aggregation is predicted when the final sum diverges, or when $z(c_1) > \left(1+\frac{e^{\beta F/2}}{2}\right)^{-n}$. In order to simplify the mathematics, we make the approximation that $n \gg 2$. This allows us to write each summand as $m^{-\lambda}\gamma^m$ for some values of λ and γ:

$$c^{\text{tot}} = \frac{e^{\frac{\beta F}{2}+\beta\Delta F}}{2}\sqrt{\frac{8}{n\pi}}\sum_{m=1,3,5\ldots}^{m^\star} m^{-1/2}\left[z(c_1)\right]^m + \sqrt{\frac{8}{n^3\pi}}e^{\beta\Delta F}\sum_{m=2,4,6\ldots}^{m^\star} m^{-3/2}\left[z(c_1)\right]^m + \frac{e^{\beta\Delta F}\left(1+2e^{-\beta F/2}\right)^{1/2}}{2\pi\left(\frac{n}{2+e^{\beta F/2}}\right)^2}\sum_{m=m^\star+1}^{\infty} m^{-2}\left[\left(1+\frac{e^{\beta F/2}}{2}\right)^n z(c_1)\right]^m. \quad \text{(S80)}$$

To address the first sum, we let $m_o = (m+1)/2$, such that

$$\begin{aligned}\sum_{m=1,3,5\ldots}^{m^\star} m^{-1/2}\left[z\right]^m &= \frac{1}{z\sqrt{2}}\sum_{m_o=1}^{\frac{m^\star+1}{2}}\left(m_o-\frac{1}{2}\right)^{-1/2}\left[z^2\right]^{m_o} \\ &= \frac{z}{\sqrt{2}}\left[\Phi\left(z^2,\frac{1}{2},\frac{1}{2}\right) - z^{m^\star+1}\Phi\left(z^2,\frac{1}{2},\frac{m^\star}{2}+1\right)\right]\end{aligned} \quad \text{(S81)}$$

(where we have written $z(c_1)$ as z for notational convenience and clarity). For notational clarity, we will denote this combination by $s_o(z, m^\star)$. We make a similar substitution for the second sum (which we will denote by $s_e(z, m^\star)$), letting $m_e = m/2$:

$$\begin{aligned}\sum_{m=2,4,6\ldots}^{m^\star} m^{-3/2}\left[z\right]^m &= 2^{-3/2}\sum_{m_e=1}^{\frac{m^\star}{2}} m_e^{-3/2}\left(z^2\right)^{m_e} \\ &= 2^{-3/2}\left[Li_{3/2}(z^2) - z^{m^\star+2}\Phi\left(z^2,\frac{3}{2},\frac{m^\star}{2}+1\right)\right].\end{aligned} \quad \text{(S82)}$$

The aggregation threshold occurs when $z(c_1) = \left(1+\frac{e^{\beta F/2}}{2}\right)^{-n}$, meaning that

$$\frac{c^{\text{tot}}_{\text{thresh}}}{e^{\beta\Delta F}} = e^{\frac{\beta F}{2}}\sqrt{\frac{2}{n\pi}}\, s_o\left(\left(1+\frac{e^{\beta F/2}}{2}\right)^{-n}, m^\star\right) + \sqrt{\frac{8}{n^3\pi}}\, s_e\left(\left(1+\frac{e^{\beta F/2}}{2}\right)^{-n}, m^\star\right) + \frac{\left(1+2e^{-\beta F/2}\right)^{1/2}}{2\pi\left(\frac{n}{2+e^{\beta F/2}}\right)^2}\psi^{(1)}(m^\star+1). \quad \text{(S83)}$$

Given our previous estimate of $m^\star = \frac{2+4e^{-\beta F/2}}{n}$, this expression only depends on the parameters n, βF, and $\beta\Delta F$, and is therefore our final expression for the aggregation threshold in this regime.

4.3 Disallowing neighbor bonds

The procedure here follows that outlined in the previous section, with the relevant partition functions substituted for one another. Because the procedure is so similar, we will move faster through these calculations.

As we saw, the partition function for multimers in the intermediate regime appears in all of the calculations. If we define a function $f(\beta F)$ such that

$$f(\beta F) = 1 - \frac{e^{\beta F/4}}{\sqrt{4 + e^{\beta F/2}}} \tag{S84}$$

then the dominant term in that partition function is that corresponding to $N_b^\star = qf(\beta F)/2$. Recognizing that in the intermediate binding regime, $N_b^\star \gg 1$, the equation for Z_m in this regime (considering only the dominant term) is

$$Z_m = \frac{2}{\pi}\left[qf(\beta F)\right]^{-2}\frac{e^{-\beta(m-1)\Delta F}}{m}\left(\left[2e^{-\beta F/2}\left(\frac{1-f(\beta F)}{1-\frac{f(\beta F)}{2}}\right)\left(\frac{1}{f(\beta F)}-1\right)\right]^{f(\beta F)}\left[\frac{1-\frac{f(\beta F)}{2}}{1-f(\beta F)}\right]^2\right)^q. \tag{S85}$$

For clarity, we denote by $\mathcal{F}(\beta F)$ the expression raised to the power of q, such that the expression above can be written a bit more cleanly:

$$Z_m = \frac{2}{\pi}\left[qf(\beta F)\right]^{-2}\frac{e^{-\beta(m-1)\Delta F}}{m}\left[\mathcal{F}(\beta F)\right]^q. \tag{S86}$$

4.3.1 Monomers have some stickers typically bound

In this regime, we have

$$\frac{Z_m}{Z_1^m} = \frac{n^2}{mq^2}\left[\frac{\pi}{2}\left(f(\beta F)\right)^2 n^2 e^{-\beta\Delta F}\left(\mathcal{F}(\beta F)\right)^\alpha\right]^{m-1}. \tag{S87}$$

where $q = nm + \alpha(m-1)$. If we approximate the q in the denominator as $m(n+\alpha)$ (assuming that $nm \gg \alpha$, a reasonable assumption given that $\alpha < 1$), the prefactor becomes

$$\frac{1}{m^3}\frac{n^2}{(n+\alpha)^2}. \tag{S88}$$

Defining

$$x(c_1) = \frac{\pi}{2}\left(f(\beta F)\right)^2 n^2 e^{-\beta\Delta F}\left(\mathcal{F}(\beta F)\right)^\alpha c_1 \tag{S89}$$

we have

$$c^{\text{tot}}\frac{(n+\alpha)^2}{n^2}x(1) = \sum_{m=1}^{\infty}\frac{1}{m^2}\left[x(c_1)\right]^m. \tag{S90}$$

We therefore predict that aggregation occurs when

$$\left(f(\beta F)\right)^2\left(\mathcal{F}(\beta F)\right)^\alpha(n+\alpha)^2\frac{c^{\text{tot}}}{e^{\beta\Delta F}} > \frac{\pi}{3}. \tag{S91}$$

4.3.2 Monomers have (almost) all stickers typically bound

In the corresponding section in which we allowed neighbor bonds, we split up the respective sums into those for $m \le m^\star$ and $m > m^\star$, with $m^\star$ defined as the value of m for which the multimer partition function begins to be better approximated by the intermediate regime than by the regime in which all stickers are typically bound. We could certainly do the same split here; however, we find that typical values of $m^\star$ are so small (typically 3), and the intermediate regime such a good approximation, as to make it reasonable to forgo this split. We instead simply consider all multimers with $m > 2$ as being well-approximated by the intermediate regime. Nevertheless, if a different

split is needed, it follows along the same lines as when allowing neighbor bonds (with the exception that even and odd m should to be considered separately for all n and not only for odd n).

Even n

For even n, the dominant term in the monomer partition function is

$$Z_1 = \frac{n(n+2)}{8} e^{-\frac{\beta F}{2}(n-2)} \tag{S92}$$

Thus, for large m,

$$\frac{Z_{m>m^\star}}{Z_1^m} = \frac{2}{\pi}\left(\frac{1}{nm+\alpha(m-1)}\right)^2 (f(\beta F))^{-2} \frac{e^{\beta\Delta F}}{m} (\mathcal{F}(\beta F))^{-\alpha} \left((\mathcal{F}(\beta F))^{n+\alpha} e^{-\beta\Delta F} \frac{8}{n(n+2)} e^{\frac{\beta F}{2}(n-2)}\right)^m. \tag{S93}$$

For small m,

$$\frac{Z_{m\leq m^\star}}{Z_1^m} = e^{\beta F\frac{\alpha}{2}} \frac{e^{\beta\Delta F}}{m}\left[e^{-\beta F(1+\alpha/2)}\left(\frac{8}{n(n+2)}\right)e^{-\beta\Delta F}\right]^m. \tag{S94}$$

Approximating $m^\star(n+\alpha) \gg \alpha$ (i.e. $n \gg \alpha/m^\star$), the total concentration is

$$\begin{aligned} c^{\text{tot}} &= \sum_{m=1}^{m^\star} m\frac{Z_{m\leq m^\star}}{Z_1^m} c_1^m + \sum_{m=m^\star+1}^{\infty} m\frac{Z_{m>m^\star}}{Z_1^m} c_1^m \\ &= e^{\beta F\frac{\alpha}{2}} e^{\beta\Delta F} \sum_{m=1}^{m^\star}\left[y(c_1)\left(\mathcal{F}(\beta F)e^{\frac{\beta F}{2}}\right)^{-n-\alpha}\right]^m + \frac{2\left(\mathcal{F}(\beta F)\right)^{-\alpha} e^{\beta\Delta F}}{\pi(n+\alpha)^2\left(f(\beta F)\right)^2} \sum_{m=m^\star+1}^{\infty} \frac{1}{m^2}\left[y(c_1)\right]^m. \end{aligned} \tag{S95}$$

where we have defined

$$y(c_1) = (\mathcal{F}(\beta F))^{n+\alpha} e^{-\beta\Delta F} \frac{8}{n(n+2)} e^{\frac{\beta F}{2}(n-2)} c_1. \tag{S96}$$

The aggregation transition is defined by $y(c_1) = 1$, or $c_1 = 1/y(1)$. Thus, aggregation is predicted when

$$c^{\text{tot}} > e^{\beta F\frac{\alpha}{2}} e^{\beta\Delta F} \sum_{m=1}^{m^\star}\left[\left(\mathcal{F}(\beta F)e^{\frac{\beta F}{2}}\right)^{-(n+\alpha)}\right]^m + \frac{2\left(\mathcal{F}(\beta F)\right)^{-\alpha} e^{\beta\Delta F}}{\pi(n+\alpha)^2\left(f(\beta F)\right)^2}\psi^{(1)}(m^\star+1). \tag{S97}$$

The first sum can also be calculated directly, yielding

$$c^{\text{tot}} > e^{\beta F\frac{\alpha}{2}} e^{\beta\Delta F} \frac{\left(\mathcal{F}(\beta F)e^{\frac{\beta F}{2}}\right)^{-(n+\alpha)}\left[1-\left(\mathcal{F}(\beta F)e^{\frac{\beta F}{2}}\right)^{-(n+\alpha)m^\star}\right]}{1-\left(\mathcal{F}(\beta F)e^{\frac{\beta F}{2}}\right)^{-(n+\alpha)}} + \frac{2\left(\mathcal{F}(\beta F)\right)^{-\alpha} e^{\beta\Delta F}}{\pi(n+\alpha)^2\left(f(\beta F)\right)^2}\psi^{(1)}(m^\star+1). \tag{S98}$$

A reasonable estimate of $m^\star$ is the value of m for which $N_s(m^\star) = nm^\star - 2$, yielding

$$m^\star = \frac{2}{n(1-f(\beta F))} = \frac{2\sqrt{4+e^{\beta F/2}}}{ne^{\beta F/4}}. \tag{S99}$$

Plugging this result back into the previous equation we find a prediction for the concentration threshold that depends only on the values of n, βF, and $\beta\Delta F$.

As discussed, for the free energies we consider, we typically find $m^\star = 2$. We will now redo the calculation for that particular value. If we set $m^\star = 2$, we can separate out the $m = 1$ and $m = 2$ terms from the sum. We then have

$$
\begin{aligned}
c^{\text{tot}} - c_1 - 2\frac{Z_2}{Z_1^2}c_1^2 &= \sum_{m=3}^{\infty} m\frac{Z_m}{Z_1^m}c_1^m \\
&= \frac{2}{\pi}\left(f(\beta F)\right)^{-2} e^{\beta\Delta F}\left(\mathcal{F}(\beta F)\right)^{-\alpha}\sum_{m=3}^{\infty}\left(\frac{1}{(n+\alpha)m-\alpha}\right)^2 [y(c_1)]^m
\end{aligned}
\tag{S100}
$$

where

$$
2\frac{Z_2}{Z_1^2} = \left(\frac{8}{n(n+2)}\right)^2 e^{-\beta\Delta F-\beta F(2+\alpha/2)}\left(1+\mathcal{O}(e^{\beta F})\right) \tag{S101}
$$

and we have defined the same $y(c_1)$ as for the general $m^\star$ case.

$$
y(c_1) = \left(\mathcal{F}(\beta F)\right)^{n+\alpha} e^{-\beta\Delta F}\frac{8}{n(n+2)}e^{\frac{\beta F}{2}(n-2)}c_1. \tag{S102}
$$

Approximating $3(n+\alpha) \gg \alpha$ (i.e. $n \gg \alpha/3$), the $(n+\alpha)^2$ term can be taken out of the sum.

Since the aggregation transition is defined by $y(c_1) = 1$, or $c_1 = 1/y(1)$, aggregation is predicted when

$$
\begin{aligned}
[(n+\alpha)f(\beta F)]^2 \Bigg[&\left(\mathcal{F}(\beta F)\right)^{\alpha}\frac{c^{\text{tot}}}{e^{\beta\Delta F}} - \\
&\frac{n(n+2)}{8}\left(\mathcal{F}(\beta F)\right)^{-n} e^{-\frac{\beta F}{2}(n-2)} - \left(\mathcal{F}(\beta F)\right)^{-2n-\alpha} e^{-\beta F(n+\alpha/2)}\Bigg] > \frac{\pi}{3} - \frac{5}{2\pi}.
\end{aligned}
\tag{S103}
$$

Odd n

For odd n, the dominant term in the monomer partition function is

$$
Z_1 = e^{-\frac{\beta F}{2}(n-1)}. \tag{S104}
$$

The approach follows the same procedure as for even n. Instead of $y(c_1)$ though, we have

$$
z(c_1) = \left(\mathcal{F}(\beta F)\right)^{n+\alpha} e^{-\beta\Delta F}e^{\frac{\beta F}{2}(n-1)}c_1. \tag{S105}
$$

We also have three cases to consider:

$$
\begin{aligned}
\frac{Z_{m>m^\star}}{Z_1^m} &= \frac{2}{\pi}\left(f(\beta F)\right)^{-2} e^{\beta\Delta F}\left(\mathcal{F}(\beta F)\right)^{-\alpha}(n+\alpha)^{-2}\frac{1}{m^3}z(1)^m \\
\frac{Z_{m\leq m^\star,\,\text{odd}}}{Z_1^m} &= \frac{e^{\beta\Delta F}}{m}e^{\frac{\beta F}{2}(1+\alpha)}\left[e^{-\beta\Delta F}e^{-\frac{\beta F}{2}(1+\alpha)}\right]^m \\
\frac{Z_{m\leq m^\star,\,\text{even}}}{Z_1^m} &= \frac{e^{\beta\Delta F}}{m}e^{\frac{\beta F}{2}\alpha}\left[e^{-\beta\Delta F}e^{-\frac{\beta F}{2}(1+\alpha)}\right]^m.
\end{aligned}
\tag{S106}
$$

The total concentration is

$$
\begin{aligned}
c^{\text{tot}} = e^{\beta\Delta F}e^{\frac{\beta F}{2}(1+\alpha)}&\sum_{m=1,3,5\ldots}^{m^\star}\left[\left(\mathcal{F}(\beta F)\right)^{-(n+\alpha)} e^{-\frac{\beta F}{2}(n+\alpha)}z(c_1)\right]^m + \\
&e^{\beta\Delta F}e^{\frac{\beta F}{2}\alpha}\sum_{m=2,4,6\ldots}^{m^\star}\left[\left(\mathcal{F}(\beta F)\right)^{-(n+\alpha)} e^{-\frac{\beta F}{2}(n+\alpha)}z(c_1)\right]^m + \\
&\frac{2}{\pi}\left(f(\beta F)\right)^{-2} e^{\beta\Delta F}\left(\mathcal{F}(\beta F)\right)^{-\alpha}(n+\alpha)^{-2}\sum_{m=m^\star+1}^{\infty}\frac{1}{m^2}z(c_1)^m.
\end{aligned}
\tag{S107}
$$

Aggregation is predicted when $z(c_1) > 1$, or

$$c^{\text{tot}} > e^{\beta\Delta F} e^{\frac{\beta F}{2}\alpha} \frac{w\left(1 + e^{\beta F/2} - w^{\frac{m^\star}{2}}\left[1 + w^{1/2} e^{\beta F/2}\right]\right)}{1-w} + \frac{2}{\pi}\left(f(\beta F)\right)^{-2} e^{\beta\Delta F}\left(\mathcal{F}(\beta F)\right)^{-\alpha}(n+\alpha)^{-2}\psi^{(1)}(m^\star+1) \quad \text{(S108)}$$

where

$$w = \left(\mathcal{F}(\beta F) e^{\frac{\beta F}{2}}\right)^{-(n+\alpha)}. \quad \text{(S109)}$$

Setting $m^\star$ to (S99), this result depends only on the values of n, βF, and $\beta\Delta F$.
With this result in hand, we can also do the calculation setting $m^\star = 2$. For odd n, the dimer term yields

$$2\frac{Z_2}{Z_1^2} = e^{-\beta\Delta F - \beta F(1+\alpha/2)}\left(1 + \mathcal{O}(e^{\beta F})\right). \quad \text{(S110)}$$

With the substitution of z for y, the equation for c^{tot} is the same as for even n:

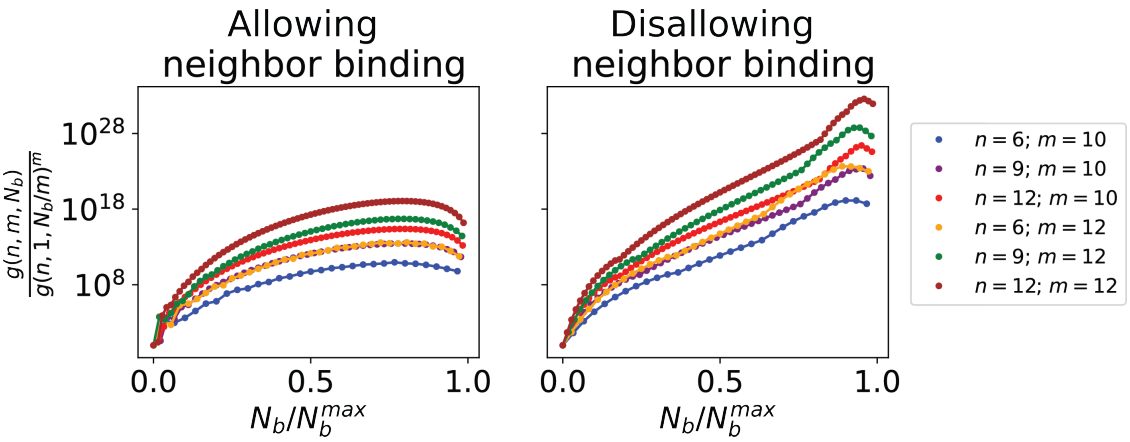
$$c^{\text{tot}} - c_1 - 2\frac{Z_2}{Z_1^2}c_1^2 = \frac{2}{\pi}\left(f(\beta F)\right)^{-2} e^{\beta\Delta F}\left(\mathcal{F}(\beta F)\right)^{-\alpha}(n+\alpha)^{-2}\sum_{m=3}^{\infty}\frac{1}{m^2}\left[z(c_1)\right]^m \quad \text{(S111)}$$

The aggregation threshold is in this case defined by $c_1 = 1/z(1)$, such that aggregation is predicted when

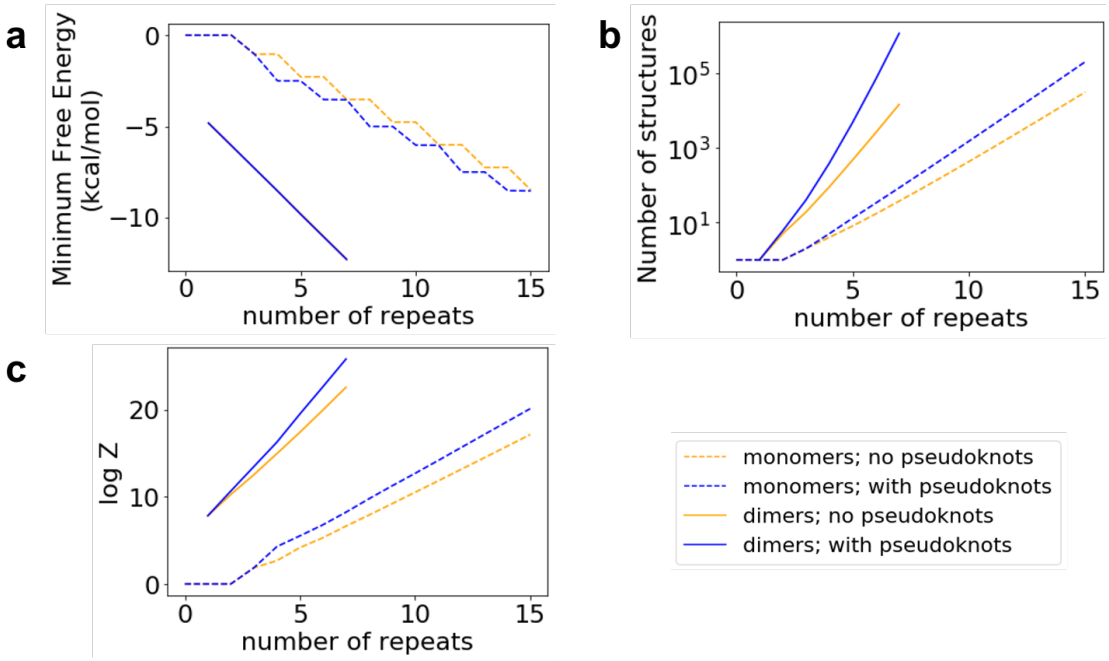
$$\left[(n+\alpha)f(\beta F)\right]^2\left[\left(\mathcal{F}(\beta F)\right)^{\alpha}\frac{c^{\text{tot}}}{e^{\beta\Delta F}} - \left(\mathcal{F}(\beta F)\right)^{-n} e^{-\frac{\beta F}{2}(n-1)} - \left(\mathcal{F}(\beta F)\right)^{-2n-\alpha} e^{-\beta F(n+\alpha/2)}\right] > \frac{\pi}{3} - \frac{5}{2\pi}. \quad \text{(S112)}$$

4.3.3 Monomers have almost no stickers bound

Because of the computational difficulty of computing partition functions in this regime, and the lack of a simple analytical formula for g, we do not consider it here. We believe this regime – in which monomers have almost no stickers bound, but the multiplicity of possible binding combinations drives binding for multimers – to be an interesting potential area for future research.



Supplementary Figure 6: **Configurational entropy drives aggregation**. The ratio of the configurational entropy of an m-mer to the configurational entropy of m monomers with the same total number of stickers bound, N_b, is shown as a function of N_b normalized by the maximum value it can take, $N_b^{\max} = nm/2$. This ratio is defined as $g(n, m, N_b)/g(n, 1, N_b/m)^m$. This ratio can serve as a proxy for the propensity of the system to aggregate based on configurational entropy considerations alone. The typical number of bonds satisfied in the system, or where on the x-axis the system will typically lie, is determined by the sticker strength $-\beta F$. The configurational entropy ratio is maximized when sticker strength is large enough such that most, but not all, stickers are bound. For the case of disallowing neighbor binding, aggregation is most likely when nearly all stickers are bound. That the maximum of the ratio plotted does not occur at the maximum value of N_b leads to the reentrant phase transition explored in this work.



Supplementary Figure 7: **Repeat RNA pseudoknots**. We consider the landscape of structures formed by RNA molecules consisting of n CAG repeats for values of $n \leq 15$ (monomers; dashed) or $n \leq 7$ (dimers; solid lines). We either disallow (orange) or allow (blue) both intra- and inter-molecular pseudoknots. **a:** The minimum free energy (MFE) structure for monomers is pseudoknotted. The energy gap between the pseudoknotted and the non-pseudoknotted MFE structure is between $\sim 1.25 - 1.5$ kcal/mol for even values of n (and smaller for odd n) but does not appear to grow with n. The dimeric MFE structure is always non-pseudoknotted. **b:** The number of structures enumerated grows significantly when allowing pseudoknots. For $n = 7$ dimers, $\sim 14,000$ non-pseudoknotted structures are enumerated compared to > 1.1 million pseudoknotted structures. **c:** The partition function is affected by the inclusion or exclusion of pseudoknotted structures, but this effect does not appear particularly significant for our purposes (see Supplementary Fig. 8).

Supplementary Note 5 Computational enumeration

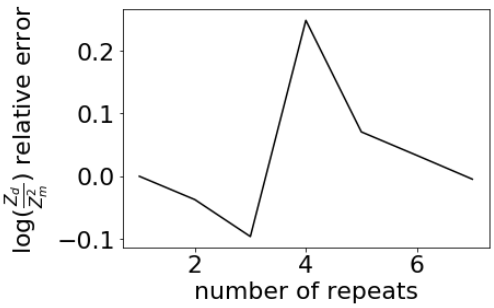
5.1 Complete enumeration is used to probe the contribution of pseudoknots to Z

How much do pseudoknots affect the landscape of structures? To answer this question, we used LandscapeFold [9] to enumerate all monomeric and dimeric structures that can form with n CAG repeats ($n \leq 15$ for monomeric structures; $n \leq 7$ for dimeric structures). The results are shown in Supplementary Fig. 7.

In panel A we show that the minimum free energy (MFE) structure is a pseudoknot for monomeric structures for $n \geq 4$. However, for odd n, its free energy is almost equal to the non-pseudoknotted MFE structure. Moreover, the energy gap between MFE pseudoknotted and non-pseudoknotted structures appears constant as a function of n (aside from the even/odd discrepancy). For dimers, the MFE structure is always non-pseudoknotted.

More significant is the effect of including pseudoknots on the landscape as a whole. In panel B we show that the number of pseudoknotted structures vastly outweighs the number of non-pseudoknotted for large n, especially for dimers. This multiplicity affects the partition function calculation (panel C). While all four partition functions here grow roughly exponentially with the number of repeats, the slope of the exponentials is higher when accounting for pseudoknots.

To quantify the effect of disallowing pseudoknots on our model results, we consider the predicted yields of the dimers while allowing and disallowing pseudoknots. We refer to the ratio Z_2/Z_1^2 as r here. We define the relative error in r due to disallowing pseudoknots as $(\log(r_n) - \log(r_p))/\log(r_p)$ where r_n describes the results of the calculation disallowing pseudoknots, and r_p the results allowing pseudoknots. Supplementary Fig. 8 shows how the relative error changes as a function of the number of repeats n in the RNA. We find that the error is mostly within $\sim 10\%$ (aside



Supplementary Figure 8: **The relative error due to disallowing pseudoknots**. We consider the landscapes of structures formed by RNA molecules consisting of n CAG repeats. As in Supplementary Fig. 7, we consider two landscapes: one in which we disallow pseudoknots; the other which includes them in the enumeration. We quantify the relative error of the ratio Z_2/Z_1^2 due to disallowing pseudoknots as described in the main text, and show the results as a function of n. We find that the relative error lies mostly within 10% and does not appear to grow with n. In fact, the relative error due to disallowing pseudoknots for the longest RNA we considered here, $n = 7$, is $< 1\%$. We therefore do not consider pseudoknotted structures further in our model.

from $n = 4$ which has a $\sim 25\%$ error). Especially, we find that the error does not appear to grow with longer repeat lengths: the error for the longest RNA we considered with pseudoknots, $n = 7$, is $< 1\%$.

5.2 Description of the dynamic programming algorithm

5.2.1 General notes on the procedure

In this section we describe how we calculate Z_m computationally by enumerating all possible structures that an arbitrary set of strands can form. By neglecting pseudoknots, we can employ a dynamic programming methodology that allows us to perform the enumeration in polynomial time (although the enumeration of certain classes of pseudoknots using dynamic programming approaches is possible; see e.g. Ref. [10]).

We exactly calculate the loop entropies for each structure. Non-pseudoknotted loops of length s have an entropic penalty given by [11, 9]

$$\Delta S_{\text{loop}}(s) = k_B \left[\ln v_s + \frac{3}{2} \ln \left(\frac{\gamma}{\pi s} \right) \right] \tag{S113}$$

where $v_s = 0.02$ nucleotides^3 is the volume within which two nucleotides can bind, and $\gamma = 3/2b$ where b is the persistence length of single-stranded RNA (we used $b = 2.4$ nucleotides here for consistency with Ref. [9]). The length of the loop s is calculated by taking the number of phosphodiester bonds in the loop. For example, hairpin loops comprised of s nucleotides have a length of $s + 1$, while internal loops comprised of s nucleotides have a length of $s + 2$.

Let us first consider monomers. Given the no-pseudoknot approximation, if a binding event between sites i and j is present, all the sites between $i + 1$ and $j - 1$ can only pair to one another, and therefore can be treated like an RNA molecule comprised of $j - i - 1$ binding sites. If another bond is formed between nucleotides $i' > i$ and $j' < j$, there are two loops that form: one of length $j' - i'$, and one of length $j' - i'$ and the other of length $i' - i + j' - j$. Thus, the length of the loop formed by the bond (i, j) depends on the structure in the middle $j - i - 1$ nucleotides. For example, if $i = 5$ and $j = 10$, if there are no binding events in those middle binding sites, the length of the loop will be 14 nucleotides long, while if there is a binding event between sites 6 and 8, the loop will only be 7 nucleotides long.

To keep track of this feature, we define for each structure a quantity we call the phantom "outer loop": how long would a loop be that connects a binding site at position 0 with a binding site at position $n + 1$? This phantom outer loop does not enter into the free energy calculation for n repeats, but does enter into the calculation when we use the results from n repeats to calculate the landscape of a longer RNA with this number of repeats in the "middle" section.

Moving on to multimers, we consider a set of m RNA strands, each comprised of n_i CAG repeats (with i ranging from 1 to m). The procedure is effected by calculating the landscape for the set of sequences of lengths (n_1, n_2, ...,

$n_m - 1$), and then adding to that the ensembles of structures that can form for each possible binding event involving the final binding site.

5.2.2 The recursive relation underlying the algorithm

We keep track of three quantities: (1) $Z(n_1, n_2, ..., n_m)$, the partition function for the set of sequences being considered; (2) $Z_s(n_1, n_2, ..., n_m)$, the partition function of the ensemble of structures with a phantom outer loop of length s; (3) $\tilde{Z}(n_1, n_2, ..., n_m)$, the value taken by Z when the entropy costs of the various phantom outer loops are taken into account. These three quantities are defined as:

$$Z_s(n_1, n_2, ..., n_m) = \sum_{\sigma_s} \exp(-\beta F_{\sigma_s}) \tag{S114}$$

$$Z(n_1, n_2, ..., n_m) = \sum_{\sigma} \exp(-\beta F_{\sigma}) = \sum_{s} Z_s(n_1, n_2, ..., n_m) \tag{S115}$$

$$\tilde{Z}(n_1, n_2, ..., n_m) = \sum_{\sigma} e^{-\beta(F_\sigma + F_b - T\Delta S_{\text{loop}}(s_\sigma))} = \sum_{s} Z_s(n_1, n_2, ..., n_m) e^{-\beta F_b + \Delta S_{\text{loop}}(s)/k_B} \tag{S116}$$

where σ is a structure defined by a set of binding events, F_σ is its free energy and s_σ is its phantom outer loop length; σ_s is a structure with a phantom outer loop of length s and F_{σ_s} is its free energy. As can be seen from these equations, Z and $\tilde{Z}$ can both be written in terms of Z_s.

Extending our previous definition of the phantom outer loop, for a set of strands it is defined as the loop formed by a binding event between a binding site at position 0 of the first sequence and position $n_m + 1$ of the final sequence. If no closed loop is formed by such a binding event (e.g. for a dimer if there are no other intermolecular binding events) there is no explicit entropy cost to the phantom outer loop forming ($\Delta S_{\text{loop}} = 0$). Instead, the penalty will be given by ΔG_{assoc}; we will add in appropriate factors of ΔG_{assoc} later in this section. For such phantom outer loops, it is useful to consider their lengths to be infinite in the following formulae (despite the fact that $\Delta S_{\text{loop}}(\infty) \neq 0$). This allows us in formulae which have terms such as $s - 2$ and $s - 3$ to consider these phantom outer loops to be unchanged by such subtractions.

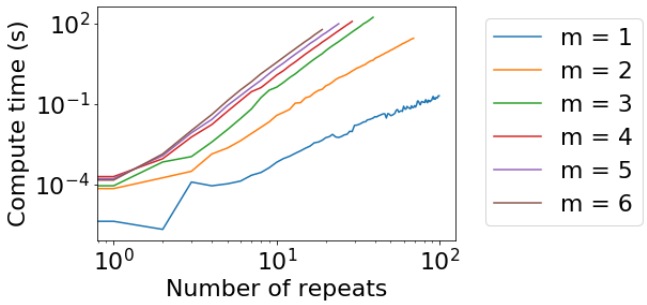
We first consider the ensemble of structures in which the last binding site is unbound, and then consider each possible site i to which the last binding site can bind. When the final binding site is bound to a site i, the set of RNA molecules is effectively split in two parts, since by disallowing pseudoknots we disallow any binding events between sites to the right of i and sites to the left. For multimers of $m \geq 3$, some non-pseudoknotted structures are disallowed by this assumption; however, as discussed in Supplementary Note 1.2.3, this actually works to our benefit since these are always identical to structures previously enumerated and we ultimately wish to enumerate each structure only once.

We describe our sum over binding sites i as the combination of a sum over the strand m' and a separate sum over the binding sites i' in that strand. We also consider the case of $m' = m$ separately since when disallowing neighbor bonds, the final site cannot bind to the one immediately preceding it (because of constraints on hairpin loop length).

$$\begin{aligned} Z_s(n_1, n_2, ..., n_m + 1) =& Z_{s-3}(n_1, n_2, ..., n_m) + \\ & \sum_{m'=1}^{m-1} \sum_{i'=1}^{n_i} Z_{s-2}(n_1, n_2, ..., n_{m'-1}, i' - 1) \tilde{Z}(n_{m'} - i', n_{m'+1}, ..., n_m) + \\ & \sum_{i'=1}^{n_m - 1} Z_{s-2}(n_1, n_2, ..., n_{m-1}, i' - 1) \tilde{Z}(n_m - i). \end{aligned} \tag{S117}$$

Here, the last sum includes n_m if allowing neighbor bonds. We also require the base cases: $Z_\infty(n_1, n_2, ..., n_{m-1}, 0) = Z(n_1, n_2, ..., n_{m-1})$; $Z_2(0) = Z_5(1) = Z_8(2) = 1$ (the last case is not included when allowing neighbor bonds).

With this procedure, we are in principle able to calculate the landscapes of arbitrary multimers. The scaling of compute time with repeat length is shown in Supplementary Fig. 9. Given the recursive nature of the algorithm, we display only the additional computation time necessary to compute Eqns. S114-S116 for a set of m strands comprised of $n_i = n$ repeats given that the landscapes for a set of m strands comprised of $n_i = n - 1$ repeats and for a set of $m - 1$ strands comprised of n repeats have both already been computed.



Supplementary Figure 9: **Compute time for multimer landscape calculation.** The computation times for multimers of m strands each comprised of n repeats is shown. Given the recursive nature of the algorithm, we display only the additional computation time necessary to compute Z for a multimer comprised of n repeats given that the landscape for a multimer comprised of $n-1$ repeats has already been computed.

5.2.3 Computing Z for single complexes and correcting for symmetries

In the previous subsections, we showed how to compute the landscape of all structures that can be formed by a set of m strands of different lengths. However, that landscape includes both m-mers as well as monomers and $m-1$-mers, dimers and $m-2$-mers, etc. In this subsection we consider how to use these landscapes to compute Z for single complexes comprised of m strands. This is analogous to the correction described in Supplementary Note 1.2.4. We also correct for symmetries as discussed in Supplementary Note 1.2.3.

For $m=2$, there are two types of structures: dimers and pairs of monomers. Therefore,

$$Z_2(n_1,n_2) = \frac{1}{2}\Big(Z(n_1,n_2) - Z(n_1)Z(n_2)\Big). \tag{S118}$$

The factor of $1/2$ corrects for that every (asymmetric) dimer structure is counted twice in $Z(n_1,n_2)$ (see Supplementary Note 1.2.3). The subtraction accounts for that $Z(n_1,n_2)$ includes not only dimers but pairs of monomers as well.

For $m=3$, there are three types of structures: trimers, 3 monomers, and 1 monomer and 1 dimer. We therefore have

$$\begin{aligned} Z_3(n_1,n_2,n_3) = \frac{1}{3}Z(n_1,n_2,n_3) - \frac{1}{3}\Big(&Z(n_1)Z(n_2)Z(n_3) + \\ &2Z_2(n_1,n_2)Z(n_3) + 2Z_2(n_1,n_3)Z(n_2) + 2Z_2(n_2,n_3)Z(n_1)\Big). \end{aligned} \tag{S119}$$

The factor of 2 before each factor of Z_2 accounts for the fact that $Z(n_1,n_2,n_3)$ overcounted the dimer & monomer structures in the same way Z_2 did (i.e. by a factor of 2). The factor of 2 here ensures we properly subtract the contribution of dimers & monomers from $Z(n_1,n_2,n_3)$.

For $m=4$ the calculation is somewhat complicated by our no-pseudoknot assumption. The system can form a tetramer, 4 monomers, 1 trimer and 1 monomer, 2 monomers and 1 dimer, or 2 dimers. However not every set of two dimers can form, since the pair of dimers $(n_1,n_3),(n_2,n_4)$ looks like a pseudoknot in our model and was therefore not enumerated. Our model therefore yields

$$\begin{aligned} Z_4(n_1,n_2,n_3,n_4) = \frac{1}{4}Z(n_1,n_2,n_3,n_4) - \frac{1}{4}\Big(&Z(n_1)Z(n_2)Z(n_3)Z(n_4) + \\ &3Z_3(n_1,n_2,n_3)Z(n_4) + 3Z_3(n_1,n_2,n_4)Z(n_3) + 3Z_3(n_1,n_3,n_4)Z(n_2) + \\ &3Z_3(n_2,n_3,n_4)Z(n_1) + 2Z_2(n_1,n_2)Z(n_3)Z(n_4) + 2Z_2(n_1,n_3)Z(n_2)Z(n_4) + \\ &2Z_2(n_1,n_4)Z(n_2)Z(n_3) + 2Z_2(n_2,n_3)Z(n_1)Z(n_4) + 2Z_2(n_2,n_4)Z(n_1)Z(n_3) + \\ &2Z_2(n_3,n_4)Z(n_1)Z(n_2) + 4Z_2(n_1,n_2)Z_2(n_1,n_2) + 4Z_2(n_1,n_4)Z_2(n_2,n_3)\Big). \end{aligned} \tag{S120}$$

The results for larger m proceed in a similar fashion.

At this point it is straightforward to include the penalty for multimerization, which (as discussed in Supplementary Note 2) we set to $\Delta G_{\text{assoc}} = 4.09$ kcal/mol $- k_B T \log(\rho/1 \text{ mol/L})$. Z_2 is multiplied by one factor of $\exp(-\beta \Delta G_{\text{assoc}})$; Z_3 by two factors; Z_4 by three, etc.

Supplementary References

[1] Robert M. Dirks, Justin S. Bois, Joseph M. Schaeffer, Erik Winfree, and Niles A. Pierce. Thermodynamic analysis of interacting nucleic acid strands. SIAM Review, 49(1):65–88, 2007.

[2] Guangnan Meng, Natalie Arkus, Michael P. Brenner, and Vinothan N. Manoharan. The free-energy landscape of clusters of attractive hard spheres. Science, 327(5965):560–563, 2010.

[3] Tianbing Xia, David H. Mathews, and Douglas H. Turner. Thermodynamics of RNA secondary structure formation. In Dieter Soll, Susumu Nishimura, and Peter B. Moore, editors, RNA, chapter 2, pages 21–48. Pergamon, 1 edition, 2001.

[4] Mirela Andronescu, Zhi Chuan Zhang, and Anne Condon. Secondary structure prediction of interacting RNA molecules. Journal of Molecular Biology, 345(5):987–1001, 2005.

[5] Douglas H. Turner and David H. Mathews. NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Research, 38(SUPPL.1):2009–2011, 2009.

[6] John SantaLucia and Donald Hicks. The Thermodynamics of DNA Structural Motifs. Annual Review of Biophysics and Biomolecular Structure, 33(1):415–440, 2004.

[7] Naoki Sugimoto, Shu-ichi Nakano, Misa Katoh, Akiko Matsumura, Hiroyuki Nakamuta, Tatsuo Ohmichi, Mari Yoneyama, and Muneo Sasaki. Thermodynamic Parameters To Predict Stability of RNA/DNA Hybrid Duplexes. Biochemistry, 34(35):11211–11216, 1995.

[8] Douglas H. Turner. Conformational Changes. In Victor A Bloomfield, Donald M Crothers, and Ignacio Tinoco, editors, Nucleic Acids: Structures, Properties, and Functions, chapter 8, pages 271–291. University Science Books, Sausalito, CA, 2000.

[9] Ofer Kimchi, Tristan Cragnolini, Michael P. Brenner, and Lucy J. Colwell. A Polymer Physics Framework for the Entropy of Arbitrary Pseudoknots. Biophysical Journal, 117(3):520–532, 2019.

[10] Elena Rivas and Sean R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. Journal of molecular biology, 285(5):2053–2068, 1999.

[11] Homer Jacobson and Walter H. Stockmayer. Intramolecular reaction in polycondensations. I. The theory of linear systems. The Journal of Chemical Physics, 18(12):1600–1606, 1950.