bioRxiv preprint doi: https://doi.org/10.1101/2022.05.12.491634; this version posted May 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

Ofer Kimchi[1, 2, *], Ella M. King[3], and Michael P. Brenner[3, 4]

[1]*Lewis-Sigler Institute, Princeton University, Princeton, New Jersey, 08544, USA*
[2]*Initiative for the Theoretical Sciences, Graduate Center, City University of New York, New York, NY 10016, USA*
[3]*Physics Department, Harvard University, Cambridge, MA, 02138, USA*
[4]*School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, 02138, USA*
[*]*okimchi@princeton.edu*

Abstract

Repeat expanded RNA molecules aggregate under certain conditions both *in vitro* and *in vivo*. Understanding the mechanism for aggregation—including how aggregation properties change with sequence and environmental conditions—would explain and predict the behavior of RNA-based biomolecular condensates, and enable the rational design of RNA-based materials. Here, we introduce an analytical framework to predict aggregation for any repeat RNA sequence, accounting for both intra- and inter-molecular bonding. By enumerating the equilibrium landscape of multimers, we reveal the driving force for aggregation: 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, and enables the 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, 2, 3, 4, 5, 6]. While these interactions are often coupled to RNA-protein contacts in typical condensates, 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, 9, 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, 12, 13, 14, 15, 16, 17], including an aggregation transition [7, 18]. In particular, RNA containing **CAG** or **CUG** repeats have been found to phase separate depending on the number of repeats present in each molecule, led by **GC** stickers binding to one another. 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 $\sim$3 months of supercomputer time.

Current models are insufficient to explore the properties of the aggregation transition demonstrated by these studies. With very few exceptions [20], 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 [21, 22, 23]. These models typically employ mean-field theory approaches that do not naturally distinguish between intra- and inter-molecular bonds [24].

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.12.491634; this version posted May 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

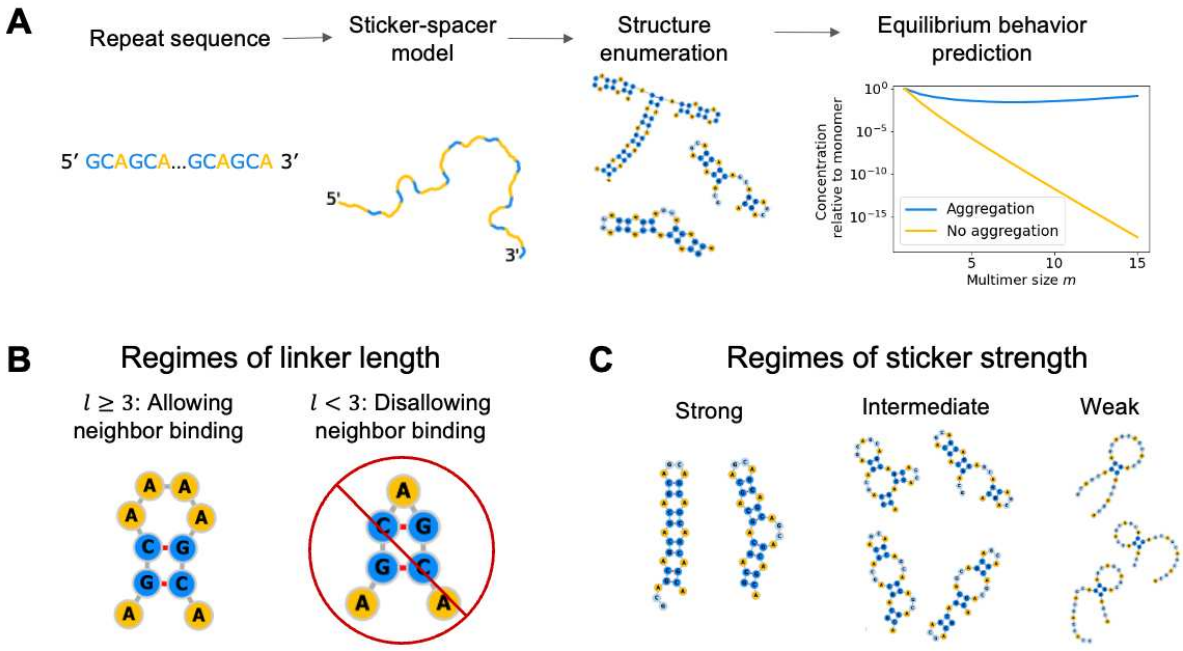

Figure 1: **Model overview**. **A: Model procedure** A repeat RNA or DNA sequence is converted to a sticker-spacer model, with stickers comprised of self-complementary regions. Possible structures, including multimers, are then enumerated by either computational or analytical methods. Partition functions 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. **C: Regimes of sticker strength** For strong stickers, (almost) all of the stickers are typically satisfied; for weak stickers almost none are; for intermediate strengths, the number of stickers typically satisfied depends on a combination of the sticker strength and the multiplicity of structures in which a given number of stickers is bound. Structures visualized using *forna* [26].

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 to a computational model that enumerates the complete landscape of intra- and inter-molecular structures that the RNA can form, and by comparing 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.12.491634; this version posted May 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Results

Equilibrium behavior is predicted by an analytical model

We consider a nucleic acid sequence comprised of n stickers separated by $n-1$ equally spaced linkers, present at concentration c^{tot} (Fig. 1A). 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}. Assuming a characteristic loop length l_{eff}, the effective strength of the sticker interactions is $F \equiv F_b - T\Delta S_{\text{loop}}(l_{\text{eff}})$ (see Methods).

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 (Fig. 1A). 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 Methods). The partition functions are comprised of three terms:

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

Here, 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. ΔF is the effective free energy cost of multimerization (see Methods) and $\beta = 1/k_B T$ is the inverse thermal energy, where T is temperature. g can be calculated exactly (see Methods and Section S1) 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).

The sum in (1) 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 (1) is dominated by large, intermediate, and small values of N_b, respectively (Fig. 1C). The value of N_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 (Section S5.2). 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 (with a single next-order correction term); 3) it uses an approximate form for $g(n, m, N_b)$ (see Methods). The computational model makes none of these approximations, considering all (non-pseudoknotted; see Section S5.1) structures that can form and including a loop-length-dependent loop entropy term.

Nevertheless, the analytical model closely approximates the exact computational model, as demonstrated in Fig. 2. The analytical model requires only one fitting parameter: the effective loop length l_{eff}. 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, Fig. S2 and Fig. S3.

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 and concentrations. We considered the same system of **CAG** sequences, using the value $F_b = -10$ 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 partition functions computationally, and used the analytical model to extrapolate up to $m = 64$. 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 $2 \leq m \leq 4$ as oligomers; however, this ensemble is dominated by dimers, with trimers

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.12.491634; this version posted May 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

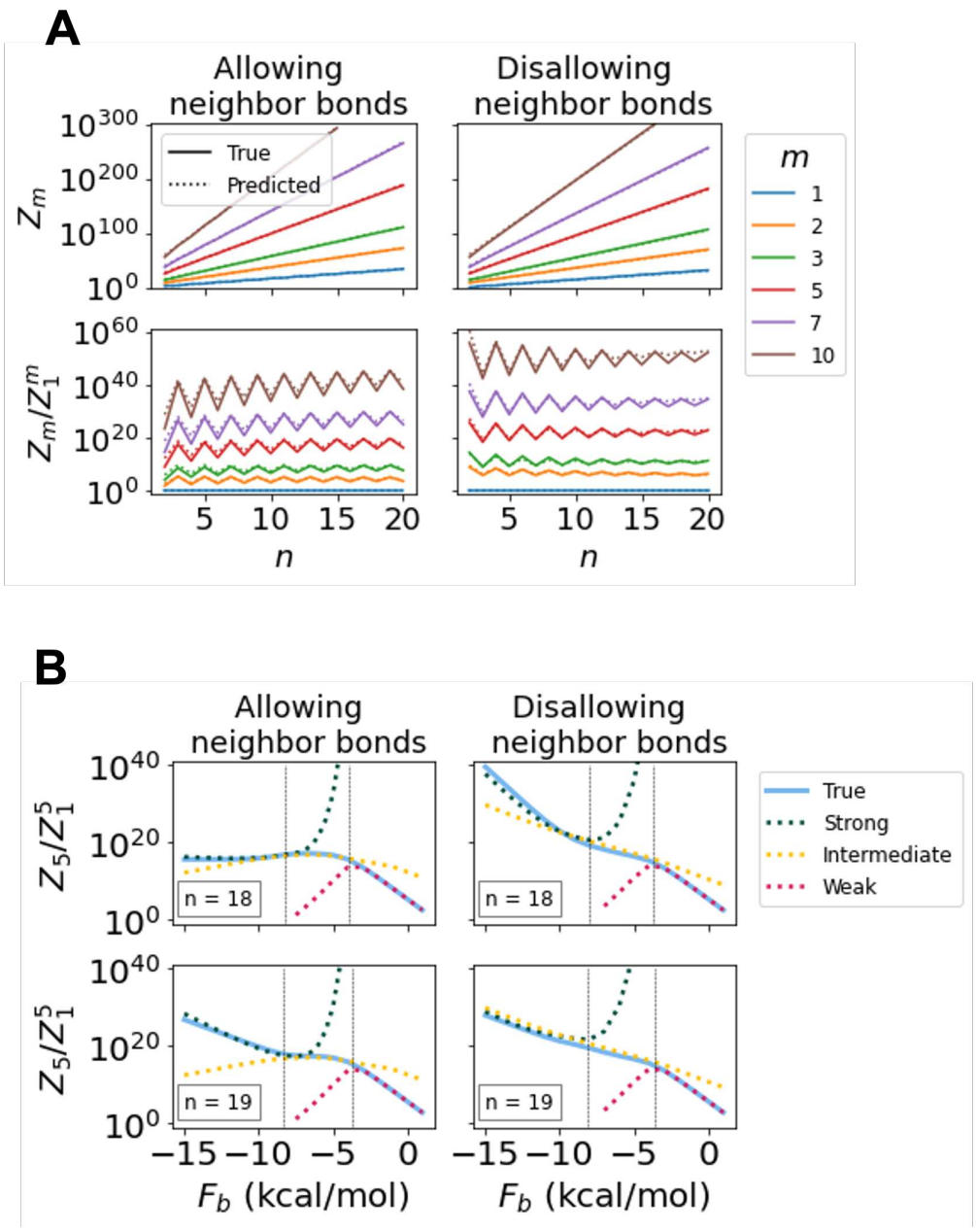

Figure 2: **Analytical model demonstrates good agreement with computational results**. **A: As a function of number of stickers per strand** Partition functions and partition function ratios are plotted with respect to n using the exact computational (solid) and simplified analytical (dotted) models. 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 $g(n, m, N_b)$ from $g(nm, 1, N_b)$ 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 $N_b^\star = N_b^{\max} - 1$ and $N_b^\star = N_b^{\min} + 3$. A single fitting parameter – the same one from panel A – is used.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.12.491634; this version posted May 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

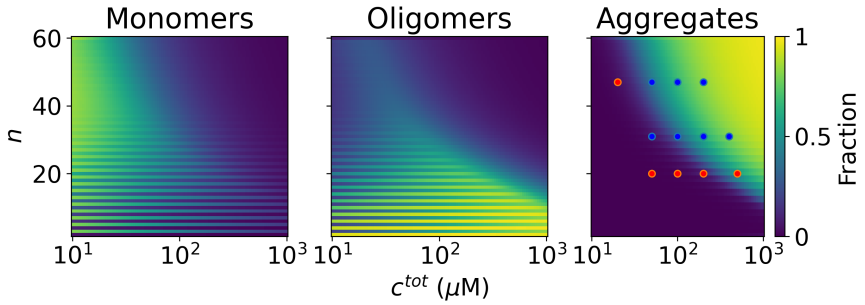

Figure 3: **Landscape of CAG repeats**. The equilibrium fraction of strands folded into monomers, oligomers (2-4-mers; primarily dimers), and aggregates is shown and compared to Nguyen *et al.*'s molecular dynamics (MD) simulation results. As the Nguyen *et al.* simulations used a sticker strength of $F_b = -10$ 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 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 $m \geq 3$ 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.

and tetramers forming at extremely 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 agreement between the two.

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

For very low temperatures or strong stickers, the ensemble 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. 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. 4A, enumerating up to $m = 15$.

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, $c^{\text{tot}}_{\text{thresh}}$ (see Section S4), as shown in Fig. 4B. 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, 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.

Discussion and Conclusions

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 3 parameters: n, the number of repeats per strand; βF, the effective strength of each bond accounting for the loop entropy cost; and $c^{\text{tot}}e^{-\beta\Delta F}$, 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

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.12.491634; this version posted May 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

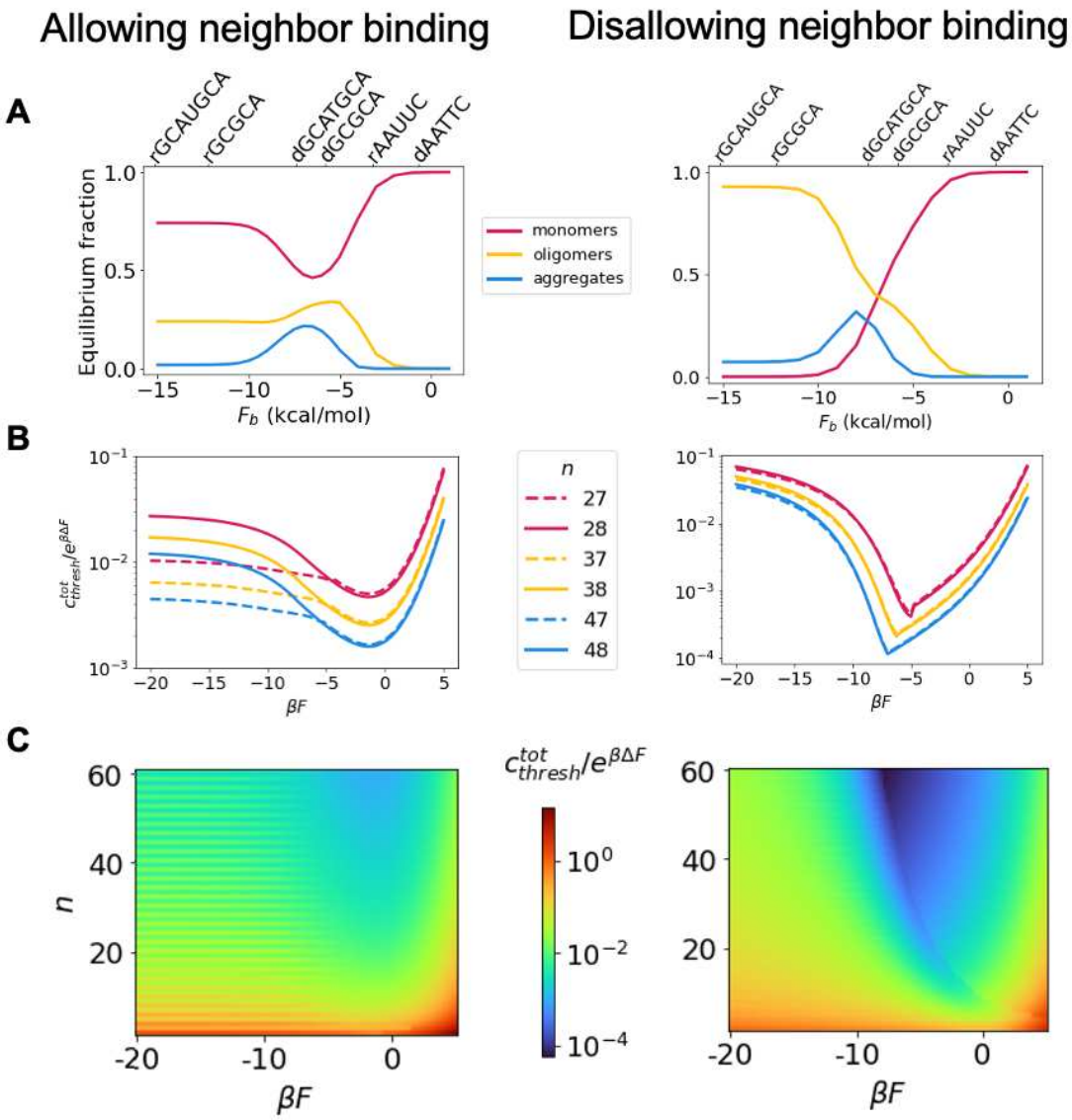

Figure 4: **Phase diagram shows a reentrant transition**. **A: Reentrant transition in computational model** Enumerating the exact partition functions up to $m = 15$, we find a reentrant transition with respect to F_b in both the regime allowing neighbor binding (LHS; $n = 8$, $l = 4$, $c^{\text{tot}} = 8$ mM is shown) and the regime disallowing neighbor binding (RHS; $n = 8$, $l = 1$, $c^{\text{tot}} = 4$ mM is shown). The high concentration is a result of the lack of Mg^{+2} considered explicitly in the model; see Discussion. Aggregates (defined as $m \geq 5$-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. Top axis shows example sequences for RNA (r) and DNA (d), and their sticker strengths as calculated by the nearest neighbor model [27, 28]. **B: Reentrant transition in analytical model** The analytical model enables enumeration up to arbitrarily large m, and reveals a reentrant transition with respect to the concentrations at which the system forms aggregates. Parity of n affects aggregation phenomena for the system allowing neighbor binding. **C: 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 long linkers (RHS) as well as for short 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 in panels B and C are due to the model's approximation of an abrupt transition from the strong to intermediate binding regimes for monomers. $c^{\text{tot}}_{\text{thresh}}$ is made dimensionless by dividing by the multimerization cost $e^{\beta\Delta F}$; see Section S2.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.12.491634; this version posted May 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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: 1) the number of stickers bound (each contributes F to the free energy); 2) the number of strands in the structure (each contributes $\mu + \Delta F$); 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 possible number of stickers bound (both are able to satisfy all stickers). Aggregates are further penalized by the multimerization cost. However, 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 [29, 20]. In our system, the benefit due to g peaks when an intermediate number of stickers is satisfied.

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. 4C. We plot the (dimensionless) threshold concentration $c^{\text{tot}}_{\text{thresh}}$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 striking 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 a competition between dimers and aggregates. For short linkers, the parity of n is found in our model to affect monomerization vs. dimerization in aggreement 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.

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 [30, 29]. 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. [31]). 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 decays 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, 32]. Experimental results thus rely on magnesium aiding the aggregation process [7]. However, the MD simulations we compare to here do not explicitly consider Magnesium [19] and the high concentrations required for the system to aggregate (e.g Fig. 3) are

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.12.491634; this version posted May 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 $c^{\text{tot}}e^{-\beta\Delta F}$. 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. Moreover, *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]. Thus, our equilibrium results may even be more relevant *in vivo* than *in vitro*.

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 patterned 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) [27, 28]. The linkers 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 over 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:

$$Z_m = \sum_{\sigma_m} e^{-\beta F(\sigma_m)}. \tag{2}$$

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 [27]

$$F(\sigma_m) = F_b N_b(\sigma_m) + (m-1)\Delta G_{\text{assoc}} - T\sum_{\text{loops}} \Delta S_{\text{loop}}(l_{\text{loop}}) \tag{3}$$

where $N_b(\sigma_m)$ is the number of bonds in the structure, and ΔG_{assoc} is the hybridization penalty associated with intermolecular binding (see Section S2). Each closed loop of length l_{loop} leads to an entropic penalty of $\Delta S_{\text{loop}}(l_{\text{loop}})$, associated with the decrease in three-dimensional configurations of the single-stranded region of the loop compared to a free chain, given by [33, 34]

$$\Delta S_{\text{loop}}(l_{\text{loop}}) = k_B\left[\ln v_s + \frac{3}{2}\ln\left(\frac{3}{2\pi b\, l_{\text{loop}}}\right)\right] \tag{4}$$

where v_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 vm^2 can be added

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.12.491634; this version posted May 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

to (3) [21] but we assume v is small enough that this term is negligible except for very large m (see Section S4 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, 35]:

$$\begin{aligned} c_m &= \frac{Z_m}{Z_1^m} c_1^m \\ \sum_m m c_m &= c^{\text{tot}} \end{aligned} \tag{5}$$

where the concentrations are made dimensionless by normalizing by a reference concentration (see Section S2) 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 (5) have two typical regimes. In one, c_m decays exponentially with m. In 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 (Section S5.2) 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. With this approximation, for monomers, each bond provides a constant free energy of $F = F_b - T\Delta S$, where $\Delta S = \Delta S_{\text{loop}}(l_{\text{eff}})$. Since the number of loops is given by $N_b - (m-1)$, we also define $\Delta F \equiv (\Delta G_{\text{assoc}} + T\Delta S)$. This quantity enters (5), such that it allows us to redefine a rescaled concentration $ce^{-\beta\Delta F}$ (also, see Section S2). Under this approximation, the partition function Z_m can thus be written as:

$$\begin{aligned} Z_m &= \sum_{\sigma_m} e^{-\beta F N_b(\sigma_m)} \\ &= \sum_{N_b} g(n, m, N_b) e^{-\beta F N_b} \end{aligned} \tag{6}$$

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 multiplicity factor is most straightforward to consider for the case of monomers. Since the contribution of pseudoknots to the partition function is negligible due to their high entropic cost (Section S5.1), our goal is 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 ($\geq$ 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 Section S1.1,

$$g(n, 1, N_b) = \begin{cases} \frac{n!}{(n-2N_b)!\ (N_b+1)!\ N_b!} & \text{if adjacent stickers can bind} \\ \frac{(n-N_b)!\ (n-N_b-1)!}{(n-2N_b)!\ (n-2N_b-1)!\ (N_b+1)!\ N_b!} & \text{otherwise} \end{cases} \tag{7}$$

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

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.12.491634; this version posted May 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ways to construct bonds between the chosen stickers). The bottom line (disallowing neighbor bonds) requires a brief additional calculation to derive (Section S1.1).

Calculating $g(n, m, N_b)$ from $g(n, 1, N_b)$ also depends on whether or not adjacent stickers can bind (see Section S1.2). While the exact calculation requires large numbers of sums with no closed-form solution, a close approximation is given by

$$g(n, m, N_b) \approx \begin{cases} \frac{g(nm,1,N_b)}{m} & \text{if adjacent stickers can bind} \\ \frac{g(nm+\alpha(m-1),1,N_b)}{m} & \text{otherwise} \end{cases} \tag{8}$$

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 factor of $1/m$ corrects for overcounting due to symmetry (Section S1.2.3; see also Fig. S1) [36].

Given expressions for the multiplicity factor, the partition functions ((6)) 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 (6) 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 ($N_b = N_b^\star$). 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 $N_b^\star = N_b^{\max} - 1$ or $N_b^\star = N_b^{\min} + 3$. For Fig. 4, we set the boundary between the strong and intermediate regimes at $N_b^\star = N_b^{\max} - \frac{1}{4}$ (disallowing neighbor binding) and $N_b^\star = \frac{n}{2} - 2$ (allowing 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 Section S3 for more details.

When comparing between the analytical and computational models, a single fitting parameter, corresponding to the (constant) effective loop length, l_{eff}, is used. 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:

$$l_{\text{eff}} = \frac{nm}{2N_b^\star} l_{\text{eff}}^{\text{fit}} \tag{9}$$

where $l_{\text{eff}}^{\text{fit}} = 4.3$ nucleotides when disallowing neighbor binding (for which we used a linker length of 1 nucleotide in the computational model) and 7 nucleotides when allowing neighbor binding (we used a linker length of 4 nucleotides).

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.

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, Megan Engel, Ben Weiner, and Rees Garmann for interesting and useful discussions.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.12.491634; this version posted May 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

References

[1] Briana Van Treeck and Roy Parker. Emerging Roles for Intermolecular RNA-RNA Interactions in RNP Assemblies. Cell, 174(4):791–802, 2018.

[2] Magdalini Polymenidou. The RNA face of phase separation. Science, 360(6391):859–860, 2018.

[3] Marta M. Fay and Paul J. Anderson. The Role of RNA in Biological Phase Separations. Journal of Molecular Biology, 430(23):4685–4701, 2018.

[4] Christine Roden and Amy S. Gladfelter. RNA contributions to the form and function of biomolecular condensates. Nature Reviews Molecular Cell Biology, 2020.

[5] Briana Van Treeck, David S.W. Protter, Tyler Matheny, Anthony Khong, Christopher D. Link, and Roy Parker. RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. Proceedings of the National Academy of Sciences of the United States of America, 115(11):2734–2739, 2018.

[6] Erin M. Langdon, Yupeng Qiu, Amirhossein Ghanbari Niaki, Grace A. McLaughlin, Chase Weidmann, Therese M. Gerbich, Jean A. Smith, John M. Crutchley, Christina M. Termini, Kevin M. Weeks, Sua Myong, and Amy S. Gladfelter. mRNA structure determines specificity of a polyQ-driven phase separation. Science, 10(1), 2018.

[7] Ankur Jain and Ronald D. Vale. RNA phase transitions in repeat expansion disorders. Nature, 546(7657):243–247, 2017.

[8] Lisa M. Ellerby. Repeat Expansion Disorders: Mechanisms and Therapeutics. Neurotherapeutics, 16(4):924–927, 2019.

[9] Karen Usdin. The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. Genome Research, 18(7):1011–1019, 2008.

[10] Anthony J. Hannan. Tandem repeats mediating genetic plasticity in health and disease. Nature Reviews Genetics, 19(5):286–298, 2018.

[11] S. V.Santhana Mariappan, Angel E. Garcia, and Goutam Gupta. Structure dynamics of the DNA hairpins formed by tandemly repeated CTG triplets associated with myotonic dystrophy. Nucleic Acids Research, 24(4):775–783, 1996.

[12] Agnieszka Kiliszek, Ryszard Kierzek, Wlodzimierz J. Krzyzosiak, and Wojciech Rypniewski. Atomic resolution structure of CAG RNA repeats: Structural insights and implications for the trinucleotide repeat expansion diseases. Nucleic Acids Research, 38(22):8370–8376, 2010.

[13] Amalia Ávila Figueroa, Douglas Cattie, and Sarah Delaney. Structure of even/odd trinucleotide repeat sequences modulates persistence of non-b conformations and conversion to duplex. Biochemistry, 50(21):4441–4450, 2011.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.12.491634; this version posted May 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[14] Cheng Wei Ni, Yu Jie Wei, Yang I. Shen, and I. Ren Lee. Long-Range Hairpin Slippage Reconfiguration Dynamics in Trinucleotide Repeat Sequences. Journal of Physical Chemistry Letters, 10(14):3985–3990, 2019.

[15] Krzysztof Sobczak, Mateusz de Mezer, Gracjan Michlewski, Jacek Krol, and Wlodzimierz J. Krzyzosiak. RNA structure of trinucleotide repeats associated with human neurological diseases. Nucleic Acids Research, 31(19):5469–5482, 2003.

[16] Magdalena Broda, Elzbieta Kierzek, Zofia Gdaniec, Tadeusz Kulinski, and Ryszard Kierzek. Thermodynamic stability of RNA structures formed by CNG trinucleotide repeats. Implication for prediction of RNA structure. Biochemistry, 44(32):10873–10882, 2005.

[17] Ji Huang and Sarah Delaney. Unique Length-Dependent Biophysical Properties of Repetitive DNA. Journal of Physical Chemistry B, 120(18):4195–4203, 2016.

[18] Yingxue Ma, Haozheng Li, Zhou Gong, Shuai Yang, Ping Wang, and Chun Tang. Nucleobase Clustering Contributes to the Formation and Hollowing of Repeat-Expansion RNA Condensate. Journal of the American Chemical Society, 144(11):4716–4720, 2022.

[19] Hung T. Nguyen, Naoto Hori, and D. Thirumalai. Condensates in RNA Repeat Sequences are Heterogeneously Organized and Exhibit Reptation-like Dynamics. bioRxiv, page 2021.02.20.432119, 2021.

[20] Benjamin G. Weiner, Andrew G.T. Pyo, Yigal Meir, and Ned S. Wingreen. Motif-pattern dependence of biomolecular phase separation driven by specific interactions. PLoS Computational Biology, 17(12):1–17, 2021.

[21] Alexander N. Semenov and Michael Rubinstein. Thermoreversible gelation in solutions of associative polymers. 1. Statics. Macromolecules, 31(4):1373–1385, 1998.

[22] Jeong Mo Choi, Alex S. Holehouse, and Rohit V. Pappu. Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions. Annual Review of Biophysics, 49:107–133, 2020.

[23] Yanxian Lin, James McCarty, Jennifer N. Rauch, Kris T. Delaney, Kenneth S. Kosik, Glenn H. Fredrickson, Joan Emma Shea, and Songi Han. Narrow equilibrium window for complex coacervation of tau and RNA under cellular conditions. eLife, 8:1–31, 2019.

[24] Michael Rubinstein and Ralph H. Colby. Polymer physics. Oxford University Press, Oxford, 2003.

[25] Agnese I Curatolo, Ofer Kimchi, Carl P Goodrich, and Michael P. Brenner. Using automatic differentiation to compute assembly yield of complex, heterogeneous components. In Preparation.

[26] Peter Kerpedjiev, Stefan Hammer, and Ivo L. Hofacker. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics, 31(20):3377–3379, 2015.

[27] 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.

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

[29] Bin Xu, Guanhua He, Benjamin G. Weiner, Pierre Ronceray, Yigal Meir, Martin C. Jonikas, and Ned S. Wingreen. Rigidity enhances a magic-number effect in polymer phase separation. Nature Communications, 11(1):4–11, 2020.

[30] Elizabeth S. Freeman Rosenzweig, Bin Xu, Luis Kuhn Cuellar, Antonio Martinez-Sanchez, Miroslava Schaffer, Mike Strauss, Heather N. Cartwright, Pierre Ronceray, Jürgen M. Plitzko, Friedrich Förster, Ned S. Wingreen, Benjamin D. Engel, Luke C.M. Mackinder, and Martin C. Jonikas. The Eukaryotic CO2-Concentrating Organelle Is Liquid-like and Exhibits Dynamic Reorganization. Cell, 171(1):148–162.e19, 2017.

[31] Yaojun Zhang, Bin Xu, Benjamin G. Weiner, Yigal Meir, and Ned S. Wingreen. Decoding the physical principles of two-component biomolecular phase separation. eLife, 10:1–31, 2021.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.12.491634; this version posted May 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[32] Paulo L. Onuchic, Anthony N. Milin, Ibraheem Alshareedah, Ashok A. Deniz, and Priya R. Banerjee. Divalent cations can control a switch-like behavior in heterotypic and homotypic RNA coacervates. Scientific Reports, 9(1):1–10, 2019.

[33] 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.

[34] 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.

[35] 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.

[36] Ellen D. Klein, Rebecca W. Perry, and Vinothan N. Manoharan. Physical interpretation of the partition function for colloidal clusters. Physical Review E, 98(3):1–12, 2018.