Accelerated simulations of RNA phase separation: a systematic study of non-redundant tandem repeats

Dilimulati Aierken [1,2,*] and Jerelle A. Joseph [1,2,†]

[1] *Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544, USA*

[2] *Omenn–Darling Bioengineering Institute, Princeton University, Princeton, NJ 08544, USA*

(Dated: December 23, 2023)

RNA promotes phase separation of proteins to form biomolecular condensates. Notably, pure RNA systems composed of RNA repeat sequences can phase separate and form liquid-like droplets. Significantly, such RNA tandem repeats have been implicated in several neurological and neuromuscular disorders. Thus, investigating the principles governing RNA repeat condensation can provide insights into RNA repeat expansion disorders. However, it is still challenging to characterize the microscopic structures and dynamics of RNA condensates. To address this, we have implemented and optimized a coarse-grained model to study the phase behavior of RNA. Our simulations reproduce key experimental findings on RNA–RNA phase separation. Importantly, we achieve multifold speed up in the simulation time compared to previous work. Leveraging this efficiency, we study the phase separation propensity of all 20 non-redundant trinucleotide repeat sequences. Our study aligns with findings from *in vivo* experiments, emphasizing that canonical base pairing and G–U wobble pairs play a dominant role in influencing the condensation of RNA repeat sequences and regulating the composition of ensuing condensates. Our implementation of this RNA model should make it feasible to efficiently probe sequence–phase behavior relationships for RNAs, and, when combined with coarse-grained protein models, elucidate the rules governing composition and dynamics of normal and aberrant biomolecular condensates.

RNA serves as a ubiquitous component of biomolecular condensates [1–5] – membraneless compartments within living cells formed through phase separation. Moreover, studies have demonstrated that pure RNA systems can undergo phase separation [6–8]. Notably, RNA-driven condensates play crucial roles in both normal cellular function and dysfunction [2, 9, 10]. Particularly intriguing are RNA-based condensates comprised of tandem repeat RNA sequences, which have been associated with various neurological and neuromuscular disorders [6]. Consequently, there exists an urgent need to decode how RNA sequence, structure, and dynamics modulate the material properties and functions of biomolecular condensates.

Recent advancements in experimental techniques, including smFISH [11], SHAPE [12], and proximity labeling [13], have offered valuable insights into the roles of RNA polymers in biomolecular condensates. However, investigating the effects of RNA at submolecular resolutions remains a formidable challenge. Computer-based approaches, particularly explicit chain molecular dynamics simulations [14–18], have proven to be excellent complements to these experimental methods, striving to bridge the gap in our ability to map RNA sequences to their phase behaviors. In recent efforts to enhance the representation of RNA polymers in simulations, explicit considerations for base-pairing propensities [14, 17, 18] and the formation of secondary structures [14] have been incorporated. The single-interaction-site (SIS) RNA model, for instance, represents each nucleotide with a unique bead, and its parameters are derived through a top-down approach leveraging experimental structures of repeat sequences [14]. Remarkably, this model successfully reproduces the length-dependent phase behavior observed experimentally in repeat RNA sequences. However, more sophisticated models like the SIS model often come with an increased computational cost. As an illustration, the initial implementation of the SIS RNA model requires approximately 100 days of sampling to acquire robust statistical data on the phase behavior of RNA. The current challenge, therefore, lies in finding a balance: how can we model RNA phase separation with both experimental accuracy and computational efficiency?

Here, our objective is to establish an accurate and efficient method for investigating the sequence-dependent phase behavior of pure RNA systems. Our starting point is the SIS RNA model, initially implemented in OpenMM [20], which inherently employs 'implicit' force fields – eliminating the need for explicit definitions of analytical forces. Here, we have broadened the applicability of the RNA model to LAMMPS [21], a widely utilized scientific molecular simulation software with exceptional parallelization capabilities. To achieve this, we derive the analytical forces for the many-body potential, employing the formalism from Ref. [22] and seamlessly integrate it into LAMMPS with optimizations. Additionally, we derive the analytical pressure tensor for the model, enabling simulations involving variations in mechanical forces on the system. As a validation of successful implementation, we replicated previously reported simulation results of $(CAG)_{47}$ in OpenMM (Supplementary Fig-

* d.aierken@princeton.edu

† Corresponding author: jerellejoseph@princeton.edu

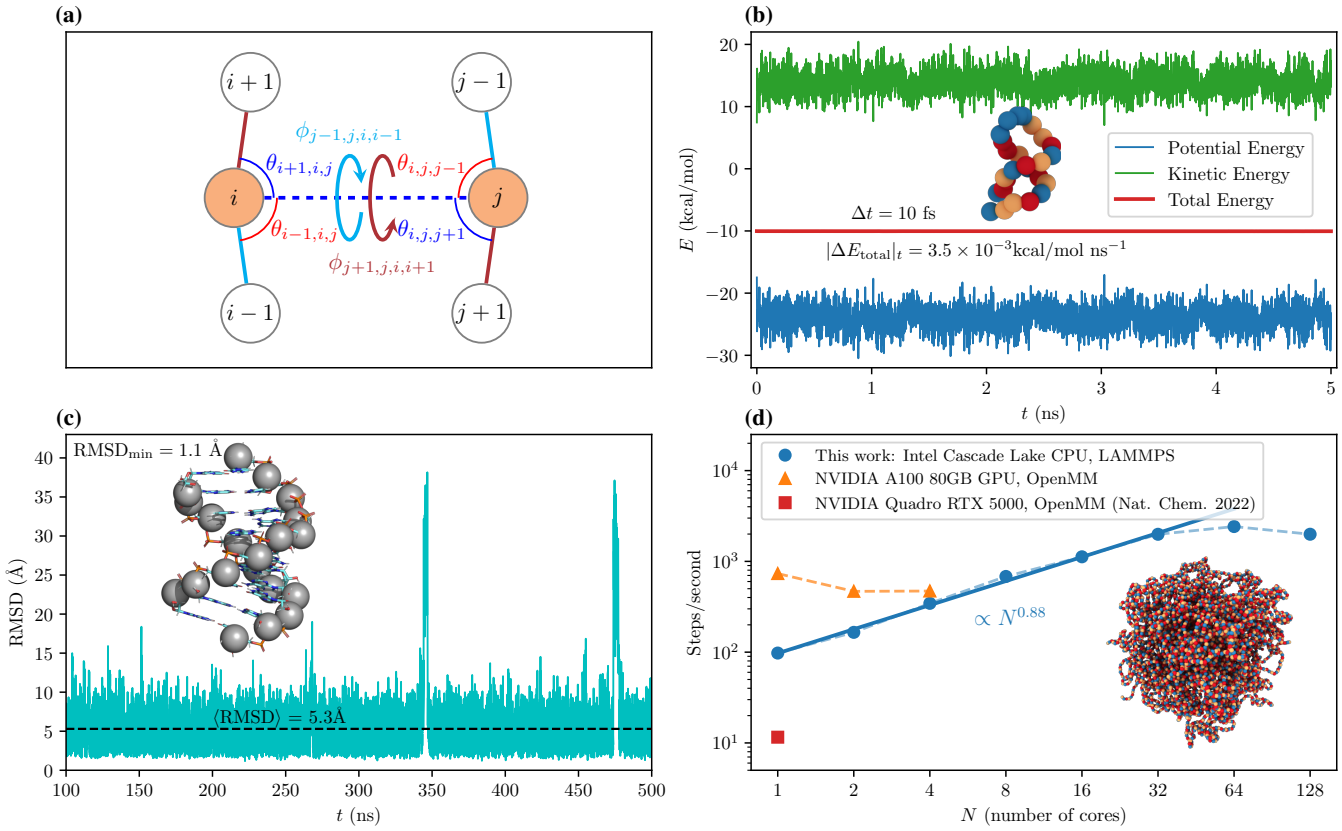

FIG. 1. **Characterization of the RNA model implementation in LAMMPS. (a)** Schematic representation of angles and dihedrals that are considered when two nucleotides i and j form base pairs in the model. The A-form right-handed antiparallel helical structures are energetically favored. **(b)** Energy trajectories from the microcanonical simulation of a single chain CAG repeat. The simulation is started from a folded structure to make sure the force calculations are properly implemented. Setting the timestep size of 10 fs, the average total energy drift (3.5×10^{-3} kcal/mol ns^{-1}) is about 4 orders of magnitude smaller than the potential energy fluctuations (~ 10 kcal/mol) demonstrating the stability of the model. **(c)** Revalidation of the structural property of a single chain that is modified from $(\text{CAG})_2$ duplex (PDB 3NJ6) [19] by adding an AAAA tetraloop and two terminal A nucleotides. In the canonical simulation at 293 K, the average root-mean-square distance (RMSD) from the crystal structure is about 5.3 Å with a minimum of 1.1Å, which matches the previously reported average of 5.4 Å in the original model implementation. For clarity, the bonds, AAAA tetraloop, and the two terminal A nucleotides are not shown in the included superposition of the minimum RMSD structure. **(d)** To evaluate the scalability of parallelization in LAMMPS, we simulated a condensed system of 64 $(\text{CAG})_{47}$ chains with Langevin thermostat at 293 K for 10^5 steps. This compact system requires an extensive evaluation of base-pair potentials at each timestep, thus representing the slowest simulation speed. The scaling exponent is 0.88 for up to 32 CPUs. Data for the original implementation and our implementation of the original OpenMM code on a newer GPU are also provided for comparison.

ure S3). Notably, we achieved comparable results in less than 3 days using only 16 CPUs, representing a two-orders-of-magnitude speedup.

Leveraging the efficiency gains of the RNA model, we investigate all 20 non-redundant trinucleotide RNA repeats (excluding homopolymers) that have also been examined in experiments [23]. Specifically, we conduct direct coexistence molecular dynamics simulations on the RNA repeat sequences, including binary mixtures of trinucleotide repeats. In line with experimental findings, our simulations unveil a robust positive correlation between the propensity to form RNA condensates and the ability to establish base pairs, specifically canonical Watson–Crick and G—U wobble base pairs. Collectively, this work significantly enhances our capability to scrutinize RNA phase behavior with both precision and speed, opening avenues for closely investigating RNA sequence effects in condensates.

The implementation in LAMMPS is stable

The SIS model is a 1-bead-per-nucleotide coarse-grained model designed to capture the properties of RNA in a sequence-dependent fashion. Initially implemented in OpenMM, the authors reported that it takes about 3 months to sample 10^8 timesteps for a system composed of 64 $(\text{CAG})_{47}$ chains. Since our main objective is to probe

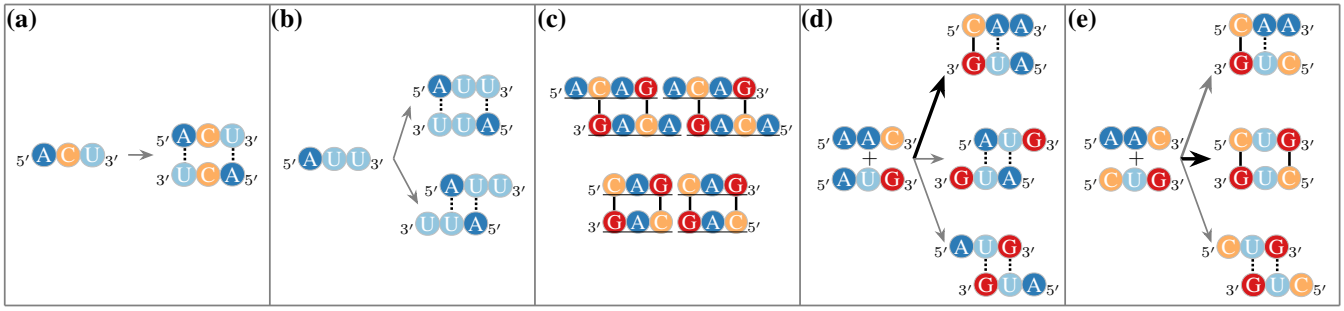

FIG. 2. **Representation of ideal minimum energy base-pairing modes for different trinucleotide repeat systems.** Base-pairing modes are defined by considering a simplified sticker–spacer representation and does not consider atomistic orientational degrees of freedom. The solid lines represent the strong C–G interaction (−5.00 kcal/mol), while dotted lines represent weaker A–U and G–U interactions (−3.34 kcal/mol) in the current model. The black bold arrows indicate the pairing of lower energies. **(a)** ACU repeats are self-complementary, forming two weaker A–U base pairs with C acting as a spacer. **(b)** AUU repeats are also self-complementary, forming two weaker A–U base pairs with either the first U or the second U acting as a 'spacer' nucleotide. **(c)** The ideal pairing pattern of ACAG repeats involves non-consecutive C–G pairs, while CAG repeats feature consecutive groups of two C–G base pairs. **(d)** The mixed system of AAC and AUG repeats has three possible pairing modes. The pairing between AAC and AUG has lower energy than AUG base pairing with itself. **(e)** The mixed system of AAC and CUG has three possible pairing options, two of which involve CUG base pairing with itself. The pairing mode that includes two C–G pairs has the lowest energy.

the phase behavior of RNA-based systems of similar and larger sizes, we first opted to boost the efficiency of this model by leveraging the parallelization features in the LAMMPS software. OpenMM employs its own toolkit to derive forces from potentials. While LAMMPS can perform derivatives numerically (e.g., from tabulated potentials), analytical potentials are much more desirable due to their boost in computational efficiency. Consequently, we derive and implement analytical forces for the SIS RNA model in LAMMPS (Supplementary Section A 3).

To achieve this, we first assess the functional forms of interactions governing bonds, excluded volumes, and base-pairing in the model. Bonded interactions, encompassing bonds and angles, are modeled via harmonic potentials that are readily available in LAMMPS, facilitating direct implementation by assigning corresponding parameters. Similarly, the excluded volume potential in the model utilizes the soft repulsive Weeks–Chandler–Andersen potential [24], which can be implemented in LAMMPS by shifting and cutting off the standard 12–6 Lennard-Jones potential. However, there is no existing form in LAMMPS for the base-pairing interactions, constituting a 6-body potential dependent on 4 angles, 2 dihedrals, and 1 distance [Figure 1(a)]. To address this gap, we implement this potential as a new extra pairwise potential in LAMMPS; i.e., the forces involved in this 6-body potential are derived analytically and encoded. Our procedure is detailed in the Supplementary Section A 3. Specifically, the underlying base-pairing potential encompasses a total of 66 different force components in the 3-dimensional Cartesian coordinate system. In comparison, typical Lennard-Jones type pairwise interactions commonly employed in coarse-grained models only necessitate consideration of 6 force components.

To assess the stability of our implementation, we conduct simulations in the microcanonical ensemble. In these simulations, total energy is conserved while potential energy and kinetic energy undergo interconversion. The correctness of force implementation and the handling of critical terms can be evaluated by observing the drift in total energy over time. Specifically, we prepare a folded RNA chain, ensuring extensive evaluation of base-pair interactions. As illustrated in Figure 1(b), with a discrete timestep of 10 fs, the energy drift per nanosecond (equivalent to 10^5 timesteps) is four orders of magnitude smaller than the potential energy fluctuations. This observation confirms the stability of our implementation.

Previous simulation results are reproduced with multifold speedup

With the stable implementation of the force field, we proceed to replicate quantitative results for individual RNA chains. During the original model parameterization, a single chain was generated by modifying the $(CAG)_2$ duplex (PDB 3NJ6)[19]. The root-mean-square distance (RMSD) from the PDB crystal structure was then determined. Subsequently, we conduct simulations of the same chain and assess the RMSD [Figure1(c)]. The computed average RMSD value is 5.3 Å, with a minimum RMSD of 1.1 Å. Notably, these RMSD values closely align with previously reported findings, specifically an average of 5.4 Å, and a minimum of approximately 1 Å. These results significantly bolster confidence in the accuracy of our LAMMPS implementation, particularly in capturing the properties of individual RNA chains.

We also assess the model's ability to replicate the phase separation propensities of CAG repeat sequences. Ex-

perimental evidence indicates that such sequences display length and concentration-dependent phase separation behaviors. For our evaluation, we choose two systems: $(CAG)_{20}$ at 50 μM and $(CAG)_{47}$ at 200 μM, which are expected to be in the dilute and condense regimes, respectively. Consistent with both experimental observations and prior simulations, we quantitatively reproduce the phase separation behaviors of these two systems (Supplementary Figure S3). Furthermore, our computed droplet concentrations align well with the reported distribution of droplet concentration previously documented for $(CAG)_{47}$ at 200 μM [14].

In addition to capturing the single-chain properties and phase separation propensities of RNA sequences, a primary objective of our work was to enhance the efficiency of sampling RNA in phase-separating systems. Notably, the original OpenMM implementation of $(CAG)_{47}$ at 200μM on a single GPU required approximately 3 months to simulate about 10^8 timesteps. In comparison, our LAMMPS implementation achieves similar sampling in under 3 days using 16 CPUs. This significant speedup (approximately 2 orders of magnitude) will, therefore, substantially enhance our capacity to investigate RNA phase separation through such many-body potentials.

Our LAMMPS implementation of RNA is scalable

The scalability of our implementation is a crucial factor for the broader applicability of the RNA model. While GPU usage has seen significant growth in recent years due to their efficiency over CPUs, CPUs remain more widely accessible on laptops, desktops, research clusters, and supercomputers. Additionally, LAMMPS is designed to be highly parallel, employing advanced domain decomposition techniques. To evaluate the scalability of our implementation, we simulate a condensed system of $(CAG)_{47}$ with 64 chains. In the case of condensed systems, numerous force calculations are required, representing the slowest simulation speed. Using the same initial conditions, we vary the number of CPUs from 1 to 128 [Figure 1(d)]. Up to 32 CPUs, the speedup scales well in the linear regime (with a scaling exponent of ~ 0.88). Unsurprisingly, when the number of CPUs equals or exceeds the number of chains in the system, the domain decomposition and inter-CPU communication increase wall time, leading to a reduction in simulation speed. However, the scalability is expected to hold well if the number of chains or the chain length increases. In contrast, the OpenMM implementation, when tested on state-of-the-art GPUs, exhibits poor scaling beyond 1 GPU [triangles in Figure 1(d)]. This may be due to the multi-GPU implementation of OpenMM, where force evaluation decomposition is used instead of domain decomposition [25]. Overall, the multi-fold speedup achieved by our LAMMPS implementation of the model will enables us to investigate RNA phase behavior with greater efficiency.

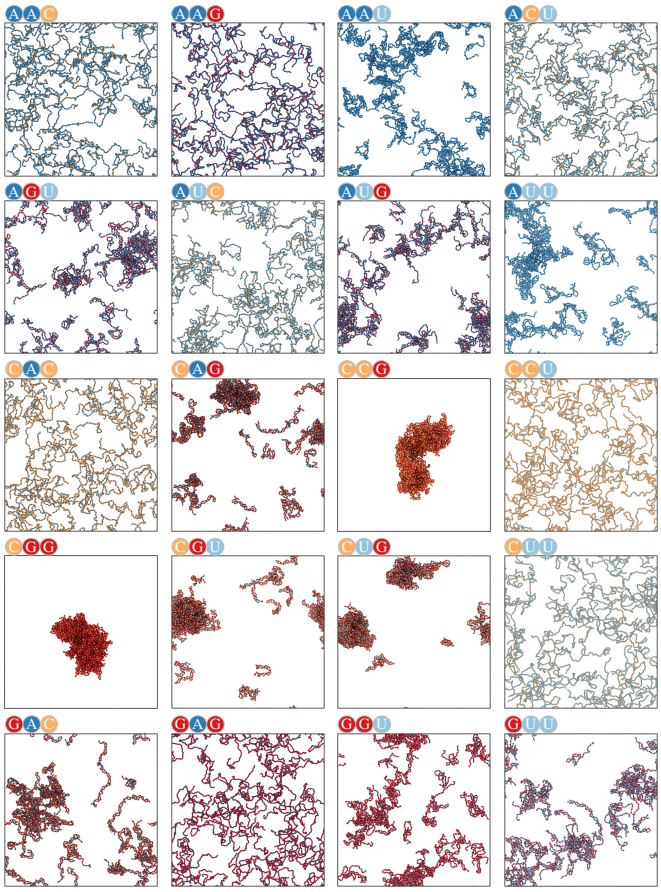

FIG. 3. **Phase behavior of all 20 non-redundant trinucleotide repeats.** Snapshots shown are taken from the end of the molecular dynamics simulations. (Adenine: blue; Uracil: light blue; Cytosine: Orange; and Guanine: Red). Each simulation is prepared by placing 64 chains of 40 trinucleotide repeats evenly in a periodic box. After using an annealing procedure (described in the text), we simulated each system for 2 μs at 293 K. Corresponding sequences are labeled at the left top of each snapshot. The redundancy is explained more explicitly in Supporting Figure S5.

Phase behavior of all 20 non-redundant trinucleotide repeats

Capitalizing on the improved efficiency of the RNA implementation in LAMMPS, we employ the model to comprehensively investigate the phase behavior of the complete set of non-redundant trinucleotide repeats. Specifically, there are a total of 64 trinucleotide repeat combinations, with the CAG repeat being just one among them. Excluding homopolymers and accounting for redundancy, this leads to a set of 20 non-redundant trinucleotide repeats (Supplementary Figure S5).

To systematically investigate the phase behavior of the 20 non-redundant trinucleotide repeats, we adopt the same methodology as employed for $(CAG)_{47}$. We simulate systems composed of 40 repeats, known to exhibit

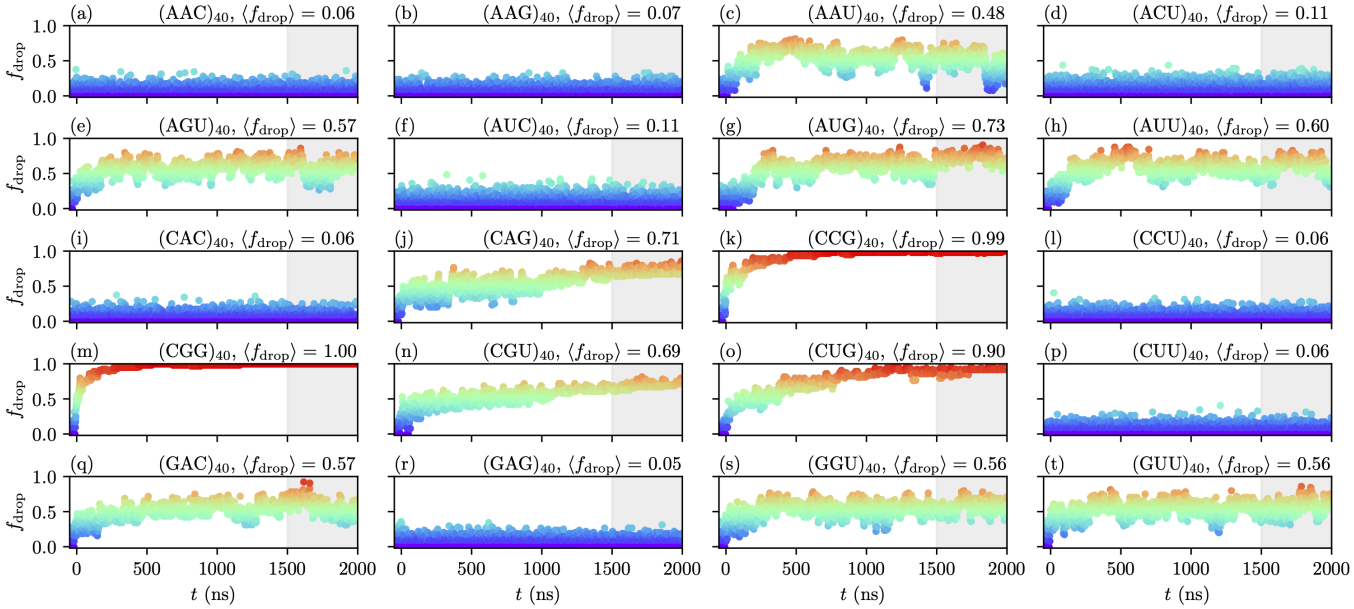

FIG. 4. **Time evolution of the fraction of droplets of all 20 non-redundant trinucleotide repeat systems.** Each system is composed of 64 chains and simulated with periodic conditions. The systems are prepared at 200 μM with chains evenly distributed and melted at 373 K. Then the systems are cooled down to 293 K within 40 ns. Finally, the systems are simulated at 293 K for 2 μs. The droplet is defined as $\geq$ 5 chains forming clusters (cutoff 20 Å). Data from the last 500 ns (shaded gray) is used to calculate the average values of f_{drop} in the system. The data points are colored based on f_{drop} values.

phase separation[6, 14] at 200 μM, and 200 repeats at the same number density for a more direct comparison with experiments [23]. The ideal base-pairing modes for the systems we simulate are summarized in Figure 2(a)–(e). In systems consisting of a single type of repeat, and when self-complementary bases are present, there are at most two classes of ideal base-pairing modes [Figure 2(a),(b)]. For repeats such as ACU, where C acts as a 'spacer' nucleotide, binding negligibly to either A or U, there is only one base-pairing mode [Figure 2(a)]. Consequently, the phase separation propensity is governed by the interaction strength of the constituent base pair. Thus, for repeats encoding weak base pair interactions (e.g., A–U in ACU, AUC) and distinct spacer nucleotides, no phase separation is observed (Figure 3). In contrast, repeats with spacer nucleotides forming strong base-pair interactions (e.g., C–G in CAG, GAC) demonstrate a greater propensity for phase separation. This dependency of phase separation on base-pair energy is also supported by a recently proposed analytical sticker–spacer model for RNA repeats [26].

However, for repeats such as AUU and AGU, which form the same number of weaker A–U or G–U base pairs (as in ACU-type sequences) but lack a residue that distinctly acts as a 'spacer', different phase separation propensities are observed. This distinction arises from the fact that such sequences encode two ideal base-pairing modes [Figure 2(b)]. The additional entropic contribution effectively reduces the free energy of phase separation compared to the scenario with a nucleotide that distinctly acts as a spacer [Figure 2(a)]. Consequently, $(AUU)_{40}$, $(AGU)_{40}$, $(AUG)_{40}$, $(AUU)_{40}$, $(GGU)_{40}$, and $(GUU)_{40}$ exhibit higher phase separation propensities, as illustrated in Figure 3, compared to ACU-type repeats.

Remarkably, phase separation is highly favored when trinucleotide repeat sequences encode strong C–G base pairs; exemplified by $(CAG)_{40}$, $(CGU)_{40}$, $(CUG)_{40}$, $(GAC)_{40}$, $(CCG)_{40}$, and $(CGG)_{40}$, as illustrated in Figure 3. Furthermore, repeat sequences comprising only C and G nucleotides lead to the formation of a single condensate encompassing all available chains at a concentration of 200 μM and after 2 μs of sampling.

Phase separation propensities of trinucleotide repeats agree well with experiments

To further quantitatively assess the phase behaviors of non-redundant trinucleotides, we track the time-dependent formation of droplets in each system, Figure 4(a)–(t). Utilizing a metric similar to previous work, a 'cluster' is defined as two nucleotides in separate chains that are less than 20 Å apart; a droplet is then characterized as a cluster comprising $\geq$ 5 chains. Figure 4 provides a summary of the computed average droplet fraction for each system near equilibrium. Repeats lacking complementary base pairs (e.g., AAC) exhibit average droplet fractions below 0.08, essentially behaving as well-mixed fluids. For systems that encode distinct spacer nucleotides and weak base pairs (A–U, G–U), the average droplet fraction is approximately 0.1, aligning with previous hypotheses that sticker strength alone is insuf-

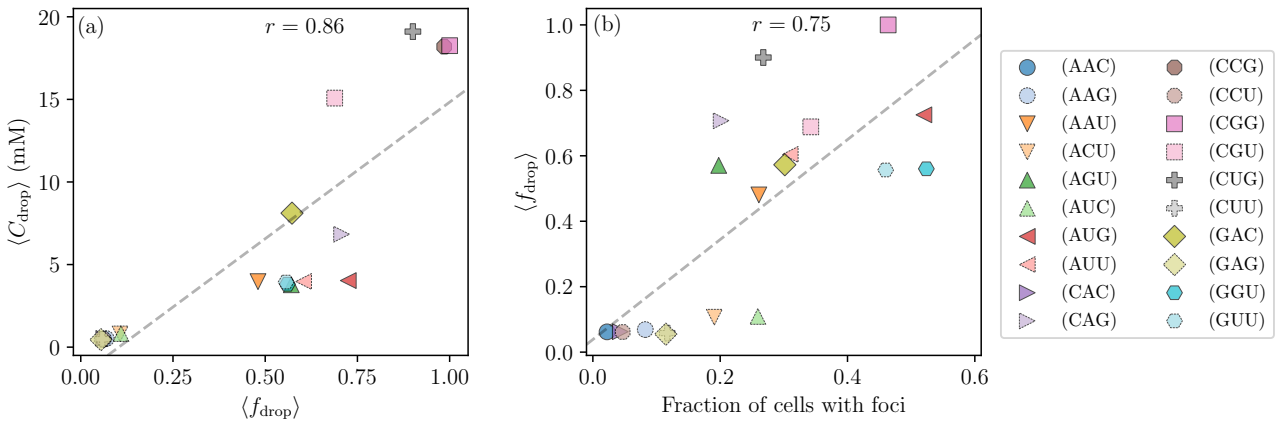

FIG. 5. **Comparison of condensate properties in simulations and with experiments. (a)** Scatter plot of the average fraction and concentration of droplets of trinucleotide repeats, both obtained from simulations. The detailed results are provided in Supplementary Figure S6 for the concentration of droplets. The Pearson correlation coefficient is 0.86, indicating a high correlation between phase separation propensity and the concentration of the resulting condensate. **(b)** Scatter plot of the average fraction of droplets from our simulations against the fraction of cells with foci from *in vivo* experiments. The positive Pearson correlation of 0.75 indicates a strong positive correlation between simulation and experimental quantities.

ficient to drive homotypic RNA phase separation. The average droplet fraction increases to more than 50% for the remaining repeat sequences. Notably, for $(CCG)_{40}$ and $(CGG)_{40}$, the droplet fraction quickly approaches 1, and the dynamics are significantly reduced. To complement these results, we also evaluate the average droplet concentrations (Supplementary Figure S6) for each system. Interestingly, there is a strong positive correlation (Pearson correlation coefficient of 0.86) between average droplet fraction and concentrations, Figure 5(a).

Finally, we compare the average fraction of droplets observed in our simulations with *in vivo* experiments. Notably, Isiktas and colleagues systematically expressed different RNA trinucleotide repeats inside cells and measured the number of cells that expressed RNA foci (i.e., droplets) in a sequence-dependent fashion [23]. We consider the fraction of RNA-positive cells as an analogue to the average fraction of droplets, as illustrated in Figure 5(b). Remarkably, our simulated values align well with the experimental findings (Pearson correlation coefficient of 0.75). This result further underscores the hypothesis that by considering the binding modes encoded via Watson–Crick pairs and the G–U wobble base pair, one can faithfully predict the phase behavior of trinucleotide RNA repeats.

Trinucleotide repeats show greater phase separation propensities than comparable quadranucleotide repeats

Experimental studies have demonstrated that odd and even repeat sequences exhibit markedly different phase separation propensities [23]. Specifically, based on the *in vivo* experiments of $(CAG)_{200}$ versus $(ACAG)_{120}$ [23], we compare the phase separation propensities of these systems. For this comparison, we utilize $(CAG)_{40}$ and $(ACAG)_{30}$, which are comparable to the experimental systems, featuring sequences of the same length with odd and even repeats, respectively. In agreement with the experiments, $(CAG)_{40}$ forms droplets [Figure 6(a)] and no phase separation is observed for $(ACAG)_{30}$ [Figure 6(b)]. The average fraction of droplets is 0.71 for the phase-separated $(CAG)_{40}$, but only 0.18 for $(ACAG)_{30}$. Moreover, the average fraction of droplets consistently drops to zero for $(ACAG)_{30}$, indicating the instability of the small droplets formed.

Analyzing the ideal pairing modes provides insight into this distinct behavior, Figure 2(c). The ratio of C–G base pairs to spacer (A) is 2:1 for CAG repeats, featuring two consecutive C–G pairs [bottom panel, Figure 2(c)]. In contrast, this ratio is 1:1 for ACAG repeats, which lack consecutive C–G pairs [top panel, Figure 2(c)]. Combining this simulation result with the experimental findings and the analytical model supports the hypothesis that consecutive groups of at least two base pairs are required for homotypic RNA condensation.

Base-pairing strength governs the composition of binary mixtures of trinucleotide repeats

In addition to pure systems with only one type of repeat sequence, we also investigate binary mixtures of two different trinucleotide repeats to understand how competing interactions can influence phase separation propensities and the composition of resulting condensates. For instance, while pure $(AUG)_{40}$ forms droplets with an average droplet fraction of 0.73, mixing this sequence with $(AAC)_{40}$ in a 1:1 ratio results in an average droplet fraction of 0.25 [Figure 6(c)]. Interestingly,

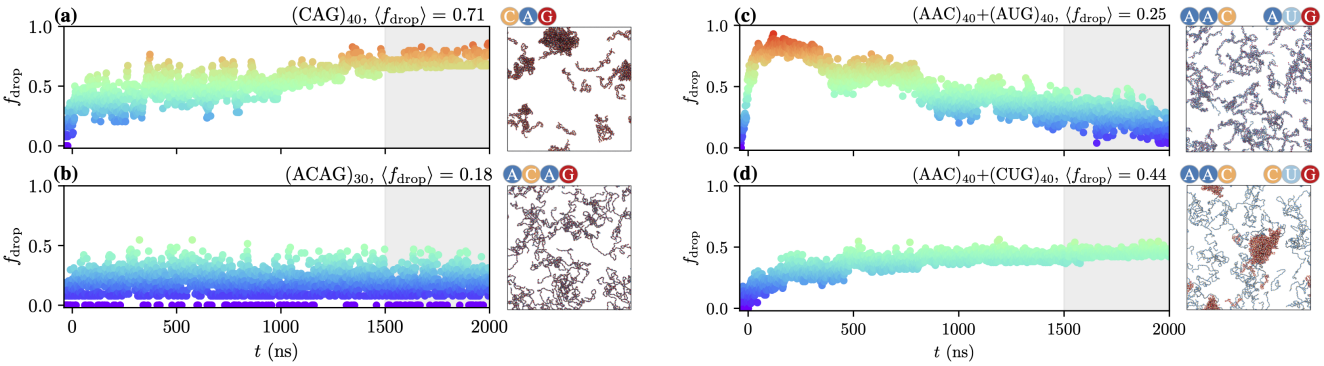

FIG. 6. **Effect of repeat units and mixing on RNA condensation.** All systems are prepared at 200 μM with chains evenly distributed and melted at 373 K. Then the systems are cooled down to 293 K within 40 ns. Finally, the systems are simulated at 293 K for 2 μs. The droplet is defined as ≥ 5 chains forming clusters (cutoff 20 Å). Data from the last 500 ns (shaded gray) is used to calculate expected average values of the fraction of the droplet f_{drop}. A simulation snapshot corresponding to the final state of each system is also shown. **(a, b)** RNA condensation of $(\mathrm{CAG})_{40}$ (**a**) versus $(\mathrm{ACAG})_{30}$ (**b**). Each system is composed of 64 chains and simulated with periodic conditions. **(c, d)** Comparison of RNA condensation processes for 1:1 mixtures of $(\mathrm{AAC})_{40}$ with $(\mathrm{AUG})_{40}$ (**c**) and $(\mathrm{CUG})_{40}$ (**d**). Each system is composed of $64 + 64 = 128$ chains and simulated with periodic conditions. In the $(\mathrm{AAC})_{40}$–$(\mathrm{AUG})_{40}$ mixture a droplet fraction of only about 0.25 is observed. The system is mostly composed of dimers formed between $(\mathrm{AAC})_{40}$ and $(\mathrm{AUG})_{40}$. The large fraction of droplets at the early stage may be due to the rapid cooling scheme used in simulations. In the $(\mathrm{AAC})_{40}$–$(\mathrm{CUG})_{40}$ mixture, the droplets formed almost exclusively contain $(\mathrm{CUG})_{40}$.

droplets in this mixture are effectively unstable, but a significant fraction comprises $(\mathrm{AAC})_{40}$–$(\mathrm{AUG})_{40}$ dimers. We further examined an equimolar binary mixture of $(\mathrm{CUG})_{40}$ and $(\mathrm{AAC})_{40}$. In this case, $(\mathrm{CUG})_{40}$ sequences form droplets that almost entirely exclude $(\mathrm{AAC})_{40}$, [Figure 6(d)].

The base-pairing strength of different pairing modes [Figure 2(d, e)] provides insights into the composition of RNA condensates. For the mixture of $(\mathrm{AAC})_{40}$ and $(\mathrm{AUG})_{40}$, the mixed pairing strength [top panel, Figure 2(d)] is stronger than the pairing of $(\mathrm{AUG})_{40}$ by itself. Consequently, the lower pairing energy drives the system to mix, even though self-pairing of $(\mathrm{AUG})_{40}$ is entropically favored for a pure system. On the other hand, $(\mathrm{CUG})_{40}$ self-complementary pairing remains the strongest for the $(\mathrm{AAC})_{40}$–$(\mathrm{CUG})_{40}$ mixture, resulting in homotypic $(\mathrm{CUG})_{40}$ droplets.

These findings lead us to further hypothesize that, in general, the compatibility of base-pairing strengths among trinucleotide repeats dictates the composition of RNA droplets in heterotypic systems. Notably, mixing $(\mathrm{CUG})_{40}$ with $(\mathrm{CAG})_{40}$ – which feature compatible C–G base pairs in all three homotypic base-pairing modes – results in droplets that contain roughly equal amounts of each sequence, Supplementary Figure S8. This observation holds broader implications for the role of RNA base pairing in regulating the composition of RNA-based condensates and aligns with previous *in vivo* experiments [27], as well as recent RNA lattice simulations [28], demonstrating that different RNA sequences can lead to distinct condensates.

DISCUSSION

We have optimized an RNA coarse-grained model and successfully implemented it in LAMMPS. Beyond reproducing the single and multi-chain properties of $(\mathrm{CAG})_n$ repeat systems [14], our implementation significantly reduces simulation time by two orders of magnitude. Furthermore, we demonstrate the stability and scalability of our LAMMPS implementation. Leveraging the increased efficiency of the model, we systematically explore the phase behavior of all 20 non-redundant trinucleotide repeats and compare the results with experimental data. Overall, we observe a strong positive correlation between simulations and *in vivo* experiments [23]. A notable exception occurs for $(\mathrm{ACU})_{40}$ and $(\mathrm{AUC})_{40}$, where no apparent phase separation is observed in simulations, despite the presence of RNA foci in *in vivo* experiments for $(\mathrm{ACU})_{200}$ and $(\mathrm{AUC})_{200}$. To address potential finite size effects, we simulate $(\mathrm{ACU})_{200}$ under the same conditions (Supplementary Figure S9), which produces comparable results to the smaller system. Our analysis suggests that the disparities could be attributed to various factors. Technically, this result may indicate a need for further calibration of the A–U interaction strength in the model, as well as the comparable G–U wobble base-pair. Another plausible explanation for the discrepancy is the involvement of proteins inside cells, which often facilitate the formation of RNA foci, anticipated in the *in vivo* experiments.

Expanding our investigation to different RNA sequences, we have compared the phase behavior of odd versus even repeats, specifically CAG and ACAG repeats. While the CAG repeats undergo phase separa-

tion, the addition of an extra 'A' spacer for the ACAG repeats disrupts the separation, consistent with experiments [23]. This finding underscores the importance of consecutive groups of at least two base pairs for homotypic RNA phase separation [26]. Additionally, our study of binary mixtures of trinucleotide repeats reveals that droplet formation is highly contingent on the base-pair binding strengths relative to the single-component solutions. Further examination of base-pairing modes, as illustrated in Figure 2, provides insights into how interaction strength (energetic contribution) and different pairing modes (entropic contribution) govern the phase behavior of RNA repeats. We anticipate that this behavior may also be applicable to similar biopolymers. However, it is crucial to note that the studied sequences represent only a small fraction of the vast sequence space. Typically, mutations at a few sites in such biopolymers can significantly alter their phase behaviors. Hence, our optimized implementation of the coarse-grained RNA model opens avenues for the exploration of a more diverse range of sequences.

It is crucial to acknowledge that the interaction strengths in the current model represent effective free energies under the assumption of charge neutralization. In principle, these interaction strengths can be rescaled, and the charge can be incorporated to investigate the phase separation behavior under monovalent salt concentrations [29] using Debye–Hückel theory. This adjustment can be seamlessly implemented in LAMMPS. Furthermore, by meticulously accounting for electrostatic effects, it opens avenues for quantitatively studying the interactions of such RNA sequences with proteins [27, 30–35]. Moreover, in the current implementation, only Watson–Crick canonical base pairs and G–U wobble base pairs are considered, as they are the most prevalent base-pairing types according to an analysis of experimental RNA structures. Notably, G-rich sequences also yield stable RNA condensates, likely sustained by G-quadruplex structures [36, 37], underscoring the significance of other base-pair interactions. Therefore, non-canonical base pairs should be incorporated in subsequent coarse-grained RNA models to effectively explore RNA condensation.

Overall, the evolving approach in developing efficient RNA phase separation models, which explicitly account for sequence-dependent effects will enhance our ability to correlate RNA sequences with macroscopic condensate features. When integrated with experimental data, these biophysical models contribute to advancing our understanding of spatiotemporal organization within cells, shedding light on mechanisms involved in abnormal condensation, and elucidating principles for designing soft materials based on RNA compartmentalization.

METHODS

The coarse-grained model of RNA

In the single-interaction-site (SIS) model [14], each RNA nucleotide is represented by a single bead. As the center of coarse-graining is not specified in the original implementation, we utilized the C3' atom on ribose to determine the location of the coarse-grained bead. This choice is deemed reasonable based on the RMSD comparison against the CAG repeat crystal structure, Figure 1(a), and our analysis of the force field parameters against bioinformatics data [38], Supplementary Figure S2. In this model, the total potential energy (U) is decoupled into three coarse-grained effective potentials:

$$U = U_{\mathrm{BA}} + U_{\mathrm{EV}} + U_{\mathrm{BP}}. \tag{1}$$

A schematic representation of the potential terms is presented in Supplementary Figure S1. Here, U_{BA} describes the bonded interactions of RNA nucleotides, consisting of bond and angle restraints between the connected beads. The specific form of the potential is:

$$U_{\mathrm{BA}} = \frac{1}{2}\sum_i k_{\mathrm{bond}}(r_i - r_0)^2 + \frac{1}{2}\sum_j k_{\mathrm{angle}}(\alpha_j - \alpha_0)^2. \tag{2}$$

All parameter values for both the previous and subsequent potentials are detailed in Supplementary Table S1. The excluded volume interaction U_{EV} is given by the Weeks–Chandler–Andersen potential:

$$U_{\mathrm{EV}} = \sum_{i,j} \Theta\left(\sigma - r_{ij}\right)\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma}{r_{ij}}\right)^{6} + 1\right]. \tag{3}$$

Θ is the Heavyside step function and if the two beads are on the same chain and involved in bond or angle potentials, then this term is not computed. For the short-ranged interactions that mimic the canonical Watson–Crick A-form base-pairing between beads i and j, the base-pair potential, U_{BP} is given as:

$$U_{\mathrm{BP}} = U_{\mathrm{bp}}^{i,j} \exp\left(U_{\mathrm{bp,bond}} + U_{\mathrm{bp,angle}} + U_{\mathrm{bp,dihedral}}\right). \tag{4}$$

Here $U_{\mathrm{bp}}^{i,j}$ is the interaction strength between two base-pairs that depends on nucleotide types of bead i and j. In the current model, only canonical A–U, C–G and G–U wobble pairs are considered. The specific forms for each term are listed as follows:

$$\begin{aligned}
U_{\mathrm{bp,bond}} &= -k_r\left(r_{ij} - r_{\mathrm{bp},0}\right)^2, \\
U_{\mathrm{bp,angle}} &= -k_\theta\left(\theta_{i,j,j-1} - \theta_1\right)^2 - k_\theta\left(\theta_{i-1,i,j} - \theta_1\right)^2, \\
&\quad - k_\theta\left(\theta_{i,j,j+1} - \theta_2\right)^2 - k_\theta\left(\theta_{i+1,i,j} - \theta_2\right)^2, \\
U_{\mathrm{bp,dihedral}} &= -k_\phi\left[1 + \cos\left(\phi_{j-1,j,i,i-1} + \phi_1\right)\right] \\
&\quad - k_\phi\left[1 + \cos\left(\phi_{j+1,j,i,i+1} + \phi_2\right)\right].
\end{aligned}$$

The interaction strength $U_{\mathrm{bp}}^{i,j}$ (A–U) $= -3.33\,\mathrm{kcal\ mol^{-1}}$, $U_{\mathrm{bp}}^{i,j}$ (G–U) $= -3.33\,\mathrm{kcal\ mol^{-1}}$, and $U_{\mathrm{bp}}^{i,j}$ (G–C) =

$-5.0\,\mathrm{kcal\,mol^{-1}}$. These interaction strengths were selected to correspond to the number of hydrogen bonds formed between the base pairs, comprising two H-bonds for A–U and G–U, and three for G–C. We confirm that the other parameters of the base-pairing interaction potential align well with the A-form structure by analyzing approximately 500 RNA A-form antiparallel structures (Supplementary Table S2) from the Nucleic Acid Knowledgebase (NAKB) [38]. Our analysis reveals no significant differences in the distributions of bond distances, angles, and dihedrals, Supplementary Figure S2. As a result, in our LAMMPS implementation, all other bonded parameters for the A–U, G–C, and G–U pairs remain consistent with the original model.

Implementation and simulations of RNA in LAMMPS

The force field is implemented in LAMMPS stable release version of 23June2022. Corresponding potential forms for the bonded potential U_{BA} and the excluded volume potential U_{EV} are already available in LAMMPS. Subsequently, we implement the base-pair potential U_{BP} and derive the analytical forces, as outlined in Supplementary Section A 3. We created a new `base/rna` keyword for accessing the potential in LAMMPS. Here, adenine (A) is defined as type 1, cytosine (C) is type 2, guanine (G) is type 3, and uracil (U) is type 4. A sample LAMMPS input scrip is provided in Supplementary List S1.

For the microcanonical simulation of the single chain (AGGCAGCAGAAAAGACGACCCA), a timestep of 10 fs is employed. The potential energy and the kinetic energy are computed every 1 ps. The average drift of the total energy $|\Delta E_{\mathrm{total}}|_t$ is calculated by dividing the absolute overall drift from the initial total energy $\left|E_{\mathrm{total}}(t+\Delta t) - E_{\mathrm{total}}(t)\right|$ by the duration of time Δt:

$$|\Delta E_{\mathrm{total}}|_t = \frac{\left|E_{\mathrm{total}}(t+\Delta t) - E_{\mathrm{total}}(t)\right|}{\Delta t}. \tag{5}$$

For all the canonical ensemble simulations of RNA repeats, the timestep is set the same as the simulations in the microcanonical ensemble (10 fs). In multichain simulations, each system is composed of 64 chains, except where otherwise specified, and simulated with periodic boundary conditions. The systems are prepared at 200 μM concentration with chains evenly distributed in the initial configuration. For $(\mathrm{AUC})_{200}$, we prepared the system with 27 chains at the same number density.

For all NVT simulations, a Langevin thermostat with low friction is used (damping time = 100,000 timesteps). First, the chains are melted at 373 K for 5 ns. This ensures that most of the chains are in an unfolded state. Then the temperature is reduced by 10 K, and the system is simulated with the Langevin thermostat for 5 ns. Repeating this cooling procedure, the system is annealed to 293 K within 40 ns. Finally, the systems are simulated at 293 K for 2 μs. Configurations for trajectory files are recorded every 0.1 ns. For estimating the average quantities near equilibrium, the data from every nanosecond from the last 500 ns are used.

Testing the scalability of the model

For the benchmarking results presented in Fig. 1(d), we prepare a compressed condensate of $(\mathrm{CAG})_{47}$ with 64 chains and simulate the system for 10000 timesteps with a Langevin thermostat at 293 K. A timestep of 10 fs is used for our GPU and CPU tests. Then we obtain the simulation speed by dividing the total number of steps by wall clock time.

In the original implementation [14] 10^8 timesteps were simulated in about 100 days using OpenMM on an NVIDIA Quadro RTX 5000 graphics card. Therefore, we estimated the speed as 10^6 steps per day or 11.6 steps per second for comparison.

For further testing, we use the original OpenMM code and simulate the system using NVIDIA A100 Tensor Core 80GB graphics cards on Della cluster (Princeton Research Computing). We use OpenMM version 7.4 for code compatibility and vary the number of GPUs to analyze the scalability.

Finally, to test the scalability of our implementation, we simulate the same system in LAMMPS using Intel Cascade Lake CPUs on Della cluster (Princeton Research Computing), and vary the number of CPUs from 1 to 128.

Comparison of single chain with experimental crystal structure

For comparison with the crystal structure (PDB code: 3NJ6), the modified singe chain (AGGCAGCAGAAAAGACGACCCA) is simulated using the same Langevin thermostat at 293 K for 500 ns. The structure is recorded every 10 ps. The RMSD values are calculated against the C3' atoms in the 3NJ6 reference structure. Specifically, the calculation is performed using the Python API from PyMOL [39] version 2.5, and data from the last 400 ns is used to calculate the average.

Droplet measurements

To be consistent with the previous definition of different types of clusters and volumes in simulations [14], we used the same definition of droplets (clusters ≥ 5 chains with cutoff 20 Å). We obtained the cluster data with gyration tensors using OVITO [40] version 3.9. Thus, the fraction of droplet is obtained as

$$f_{\mathrm{drop}} = \frac{N_{\mathrm{drop}}}{N_{\mathrm{total}}}, \tag{6}$$

where N_{drop} is the total number of chains in droplets and N_{total} is the total number of chains. For the concentrations of droplets, the gyration tensor is used, which is diagnonalized for a droplet as

$$S_{\text{D}} = \begin{pmatrix} \lambda_x^2 & 0 & 0 \\ 0 & \lambda_y^2 & 0 \\ 0 & 0 & \lambda_z^2 \end{pmatrix}, \tag{7}$$

where the principle moments λ_x^2, λ_y^2 and λ_z^2 describes effective extension of the cluster in the principal axial directions. Thus, the volume of the cluster can be estimated using an ellipsoid as $V = 4\pi\sqrt{3}\lambda_x\lambda_y\lambda_z$. Finally, the droplet concentration is calculated as the average concentration of clusters with ≥ 5 chains.

ACKNOWLEDGEMENTS

The simulations reported on in this manuscript was substantially performed using the Princeton Research Computing resources at Princeton University, which is a consortium of groups led by the Princeton Institute for Computational Science and Engineering (PICSciE) and Office of Information Technology. We thank Nathaniel Hess for helpful discussions regarding the implementation of custom potentials in LAMMPS and the manuscript. We also thank Dr. Alina Emelianova, Ananya Chakravarti and Prof. Athanassios Panagiotopoulos for helpful feedback on the manuscript. This research was supported by departmental start-up funds allocated to J.A.J. through the Department of Chemical and Biological Engineering and the Omenn–Darling Bioengineering Institute at Princeton University. This research was also partially supported by NSF through the Princeton University (PCCM) Materials Research Science and Engineering Center DMR-2011750.

CODE AND DATA AVAILABILITY

The source code for running the RNA model in LAMMPS, input scripts, as well as instructions for compiling can be found at https://doi.org/10.5281/zenodo.10428292 and at Joseph Group GitHub repository: https://github.com/josephresearch/RNA_lammps.The data supporting the findings of this manuscript will be available at https://doi.org/10.34770/byjm-r290 in DataSpace at Princeton University.

[1] C. Roden and A. S. Gladfelter, RNA contributions to the form and function of biomolecular condensates, Nat. Rev. Mol. Cell Biol. **22**, 183 (2021).

[2] S. Maharana, J. Wang, D. K. Papadopoulos, D. Richter, A. Pozniakovsky, I. Poser, M. Bickle, S. Rizk, J. Guillén-Boixet, T. M. Franzmann, M. Jahnel, L. Marrone, Y.-T. Chang, J. Sterneckert, P. Tomancak, A. A. Hyman, and S. Alberti, RNA buffers the phase separation behavior of prion-like RNA binding proteins, Science **360**, 918 (2018).

[3] M. M. Fay and P. J. Anderson, The Role of RNA in Biological Phase Separations, J. Mol. Biol. **430**, 4685 (2018).

[4] L. Guo and J. Shorter, It's Raining Liquids: RNA Tunes Viscoelasticity and Dynamics of Membraneless Organelles, Mol. Cell **60**, 189 (2015).

[5] K. Rhine, V. Vidaurre, and S. Myong, RNA Droplets, Annu. Rev. Biophys. **49**, 247 (2020).

[6] A. Jain and R. D. Vale, RNA phase transitions in repeat expansion disorders, Nature **546**, 243 (2017).

[7] B. Van Treeck, D. S. W. Protter, T. Matheny, A. Khong, C. D. Link, and R. Parker, RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome, Proc. Natl. Acad. Sci. **115**, 2734 (2018).

[8] G. M. Wadsworth, W. J. Zahurancik, X. Zeng, P. Pullara, L. B. Lai, V. Sidharthan, R. V. Pappu, V. Gopalan, and P. R. Banerjee, RNAs undergo phase transitions with lower critical solution temperatures, Nat. Chem. **15**, 1693 (2023).

[9] W. Shao, X. Bi, Y. Pan, B. Gao, J. Wu, Y. Yin, Z. Liu, M. Peng, W. Zhang, X. Jiang, W. Ren, Y. Xu, Z. Wu, K. Wang, G. Zhan, J. Y. Lu, X. Han, T. Li, J. Wang, G. Li, H. Deng, B. Li, and X. Shen, Phase separation of RNA-binding protein promotes polymerase binding and transcription, Nat. Chem. Biol. **18**, 70 (2022).

[10] J. A. Riback, J. M. Eeftens, D. S. W. Lee, S. A. Quinodoz, A. Donlic, N. Orlovsky, L. Wiesner, L. Beckers, L. A. Becker, A. R. Strom, U. Rana, M. Tolbert, B. W. Purse, R. Kleiner, R. Kriwacki, and C. P. Brangwynne, Viscoelasticity and advective flow of RNA underlies nucleolar form and function, Mol. Cell **83**, 3095 (2023).

[11] N. Tsanov, A. Samacoits, R. Chouaib, A.-M. Traboulsi, T. Gostan, C. Weber, C. Zimmer, K. Zibara, T. Walter, M. Peter, *et al.*, smifish and fish-quant–a flexible single rna detection approach with super-resolution capability, Nucleic Acids Res. **44**, e165 (2016).

[12] M. J. Smola, G. M. Rice, S. Busan, N. A. Siegfried, and K. M. Weeks, Selective 2-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis, Nature Protocols **10**, 1643 (2015).

[13] W. Qin, K. F. Cho, P. E. Cavanagh, and A. Y. Ting, Deciphering molecular interactions by proximity labeling, Nature methods **18**, 133 (2021).

[14] H. T. Nguyen, N. Hori, and D. Thirumalai, Condensates in RNA repeat sequences are heterogeneously organized and exhibit reptation dynamics, Nat. Chem. **14**, 775 (2022).

[15] A. R. Tejedor, A. Garaizar, J. Ramírez, and J. R. Espinosa, 'RNA modulation of transport properties and stability in phase-separated condensates, Biophys. J. **120**, 5169 (2021).

[16] J. A. Joseph, J. R. Espinosa, I. Sanchez-Burgos, A. Garaizar, D. Frenkel, and R. Collepardo-Guevara, Thermodynamics and kinetics of phase separation of protein-RNA mixtures by a minimal model, Biophys. J **120**, 1219 (2021).

[17] J. A. Joseph, A. Reinhardt, A. Aguirre, P. Y. Chew, K. O. Russell, J. R. Espinosa, A. Garaizar, and R. Collepardo-Guevara, Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy, Nat. Comput. Sci. **1**, 732 (2021).

[18] R. M. Regy, G. L. Dignon, W. Zheng, Y. C. Kim, and J. Mittal, Sequence dependent phase separation of protein-polynucleotide mixtures elucidated using molecular simulations, Nucleic Acids Res. **48**, 12593 (2020).

[19] A. Kiliszek, R. Kierzek, W. J. Krzyzosiak, and W. Rypniewski, Atomic resolution structure of CAG RNA repeats: Structural insights and implications for the trinucleotide repeat expansion diseases, Nucleic Acids Res. **38**, 8370 (2010).

[20] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, OpenMM 7: Rapid development of high performance algorithms for molecular dynamics, PLOS Comput. Biol. **13**, e1005659 (2017).

[21] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. In 'T Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton, LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales, Comput. Phys. Commun. **271**, 108171 (2022).

[22] W. C. Swope and D. M. Ferguson, Alternative expressions for energies and forces due to angle bending and torsional energy, J. Comput. Chem. **13**, 585 (1992).

[23] A. U. Isiktas, A. Eshov, S. Yang, and J. U. Guo, Systematic generation and imaging of tandem repeats reveal base-pairing properties that promote RNA aggregation, Cell Rep. Methods **2**, 100334 (2022).

[24] J. D. Weeks, D. Chandler, and H. C. Andersen, Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids, J. Chem. Phys. **54**, 5237 (1971).

[25] D. Pavlov, V. Galigerov, D. Kolotinskii, V. Nikolskiy, and V. Stegailov, Gpu-based molecular dynamics of fluid flows: Reaching for turbulence, The International Journal of High Performance Computing Applications , 10943420231213013 (2023).

[26] O. Kimchi, E. M. King, and M. P. Brenner, Uncovering the mechanism for aggregation in repeat expanded RNA reveals a reentrant transition, Nat. Commun. **14**, 332 (2023).

[27] E. M. Langdon, Y. Qiu, A. Ghanbari Niaki, G. A. McLaughlin, C. A. Weidmann, T. M. Gerbich, J. A. Smith, J. M. Crutchley, C. M. Termini, K. M. Weeks, S. Myong, and A. S. Gladfelter, mRNA structure determines specificity of a polyQ-driven phase separation, Science **360**, 922 (2018).

[28] A. Z. Lin, K. M. Ruff, F. Dar, A. Jalihal, M. R. King, J. M. Lalmansingh, A. E. Posey, N. A. Erkamp, I. Seim, A. S. Gladfelter, and R. V. Pappu, Dynamical control enables the formation of demixed biomolecular condensates, Nat. Commun. **14**, 7678 (2023).

[29] H. Maity, H. T. Nguyen, N. Hori, and D. Thirumalai, Odd–even disparity in the population of slipped hairpins in RNA repeat sequences with implications for phase separation, Proc. Natl. Acad. Sci. **120**, e2301409120 (2023).

[30] R. S. Fisher and S. Elbaum-Garfinkle, Tunable multiphase dynamics of arginine and lysine liquid condensates, Nat. Commun. **11**, 4628 (2020).

[31] T. Kaur, M. Raju, I. Alshareedah, R. B. Davis, D. A. Potoyan, and P. R. Banerjee, Sequence-encoded and composition-dependent protein-RNA interactions control multiphasic condensate morphologies, Nat Commun **12**, 872 (2021).

[32] I. Alshareedah, M. M. Moosa, M. Raju, D. A. Potoyan, and P. R. Banerjee, Phase transition of RNA-protein complexes into ordered hollow condensates, Proc. Natl. Acad. Sci. **117**, 15650 (2020).

[33] P. R. Banerjee, A. N. Milin, M. M. Moosa, P. L. Onuchic, and A. A. Deniz, Reentrant Phase Transition Drives Dynamic Substructure Formation in Ribonucleoprotein Droplets, Angew. Chem. **129**, 11512 (2017).

[34] M. Farag, W. M. Borcherds, A. Bremer, T. Mittag, and R. V. Pappu, Phase separation of protein mixtures is driven by the interplay of homotypic and heterotypic interactions, Nat. Commun. **14**, 5527 (2023).

[35] D. W. Sanders, N. Kedersha, D. S. W. Lee, A. R. Strom, V. Drake, J. A. Riback, D. Bracha, J. M. Eeftens, A. Iwanicki, A. Wang, M.-T. Wei, G. Whitney, S. M. Lyons, P. Anderson, W. M. Jacobs, P. Ivanov, and C. P. Brangwynne, Competing Protein-RNA Interaction Networks Control Multiphase Intracellular Organization, Cell **181**, 306 (2020).

[36] Y. Zhang, M. Yang, S. Duncan, X. Yang, M. A. S. Abdelhamid, L. Huang, H. Zhang, P. N. Benfey, Z. A. E. Waller, and Y. Ding, G-quadruplex structures trigger RNA phase separation, Nucleic Acids Res. , gkz978 (2019).

[37] W. Guo, D. Ji, A. B. Kinghorn, F. Chen, Y. Pan, X. Li, Q. Li, W. T. S. Huck, C. K. Kwok, and H. C. Shum, Tuning Material States and Functionalities of G-Quadruplex-Modulated RNA-Peptide Condensates, J. Am. Chem. Soc. **145**, 2375 (2023).

[38] C. L. Lawson, H. M. Berman, L. Chen, B. Vallat, and C. L. Zirbel, The Nucleic Acid Knowledgebase: A new portal for 3D structural information about nucleic acids, Nucleic Acids Res. , gkad957 (2023).

[39] W. L. DeLano *et al.*, Pymol: An open-source molecular graphics tool, CCP4 Newsl. Protein Crystallogr **40**, 82 (2002).

[40] A. Stukowski, Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool, Modelling Simul. Mater. Sci. Eng. **18**, 015012 (2009).

Appendix A: Supplementary Material

1. Model Parameters

The schematic representation of potential terms are shown in Figure S1 and the parameters are listed in Table S1.

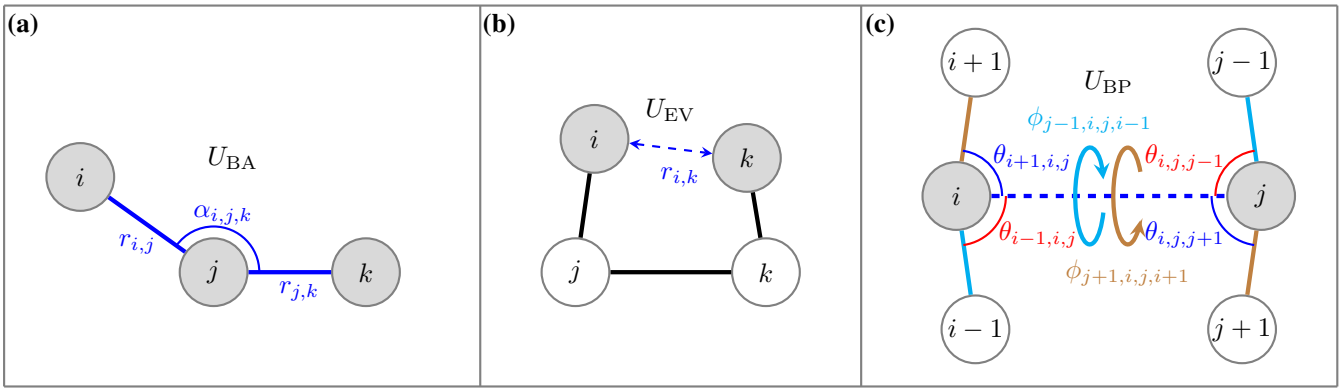

FIG. S1. **Schematic representation of distances and angles that are used in the RNA model**. **(a)** The harmonic bonded interaction has two parts that rely on the bond lengths and bond angles fluctuating around equilibrium distances and equilibrium angles. **(b)**. The excluded volume potential is a soft purely repulsive potential that is calculated for RNA beads between different chains or beads on the same chain that are at least two beads apart. **(c)** A 6-body potential is used for base-pair interaction. It considers the distance between nucleotide beads that are forming a base pair, the four angles and two dihedral angles that are formed by four neighboring beads. This term is calculated for beads from different chains or beads that are 4 beads apart on the same chain.

TABLE S1. Parameters for the coarse-grained RNA model

Parameter	Value	Parameter	Value	Parameter	Value
k_{bond}	$15.0\,\mathrm{kcal\,mol^{-1}\,\AA^{-2}}$	r_0	$5.9\,\mathrm{\AA}^2$	k_{angle}	$10.0\,\mathrm{kcal\,mol^{-1}rad^{-2}}$
α_0	$2.618\,\mathrm{rad}$	σ	$10\,\mathrm{\AA}$	ϵ	$2.0\ \mathrm{kcal\,mol^{-1}}$
$U_{\mathrm{bp}}^{i,j}$ (C-G)	$-5.00\,\mathrm{kcal\,mol^{-1}}$	$U_{\mathrm{bp}}^{i,j}$ (A-U)	$-3.33\,\mathrm{kcal\,mol^{-1}}$	$U_{\mathrm{bp}}^{i,j}$ (G-U)	$-3.33\,\mathrm{kcal\,mol^{-1}}$
k_r	$3.0\,\mathrm{kcal\,mol^{-1}\,\AA^{-2}}$	$r_{\mathrm{bp},0}$	$13.8\,\mathrm{\AA}^2$	k_θ	$1.5\,\mathrm{kcal\,mol^{-1}rad^{-2}}$
θ_1	$1.8326\,\mathrm{rad}$	θ_2	$0.9425\,\mathrm{rad}$	k_ϕ	0.5
ϕ_1	$1.8326\,\mathrm{rad}$	ϕ_2	$1.1346\,\mathrm{rad}$		

2. Bioinformatic Survey of RNA Structures

We first coarse-grain atomistic structures (C3' as heavy atom center) to calculate base-pair distances and the relative angles. As shown in Fig. S2(a), the distribution of base-pairing distance for all three A–U, C–G, and G–U pairs are indeed close to the reference distance $r_0 = 13.8\,\text{Å}$ in the model. For the bond and dihedral angles formed by these pairs, we also combine these terms into three groups and the plot corresponding scatter plots (Fig. S2 (b), (c), (d)). In addition, we also define the corresponding energy functional forms of these three groups as follows:

$$f_1(\theta_{i,j,j-1}, \theta_{i-1,j,j}) = \exp\left[-k_\theta(\theta_{i,j,j-1} - \theta_1)^2 - k_\theta(\theta_{i-1,i,j} - \theta_1)^2\right], \tag{S1}$$

$$f_2(\theta_{i,j,j+1}, \theta_{i+1,j,j}) = \exp\left[-k_\theta(\theta_{i,j,j+1} - \theta_2)^2 - k_\theta(\theta_{i+1,i,j} - \theta_2)^2\right], \tag{S2}$$

$$f_3(\phi_{j-1,j,i,i-1}, \phi_{j+1,j,i,i+1}) = \exp\left[-k_\phi(1 + \cos(\phi_{j-1,j,i,i-1} + \phi_1)) - k_\phi(1 + \cos(\phi_{j+1,j,i,i+1} + \phi_2))\right]. \tag{S3}$$

Thus, the base-pair energy can be expressed as $U_{\text{BP}} = U_{\text{bp}}^{i,j} \exp(U_{\text{bp,bond}}) f_1 f_2 f_3$. As shown in Fig. S2, the distributions of experimental values align well with the corresponding functional forms computed using the same references.

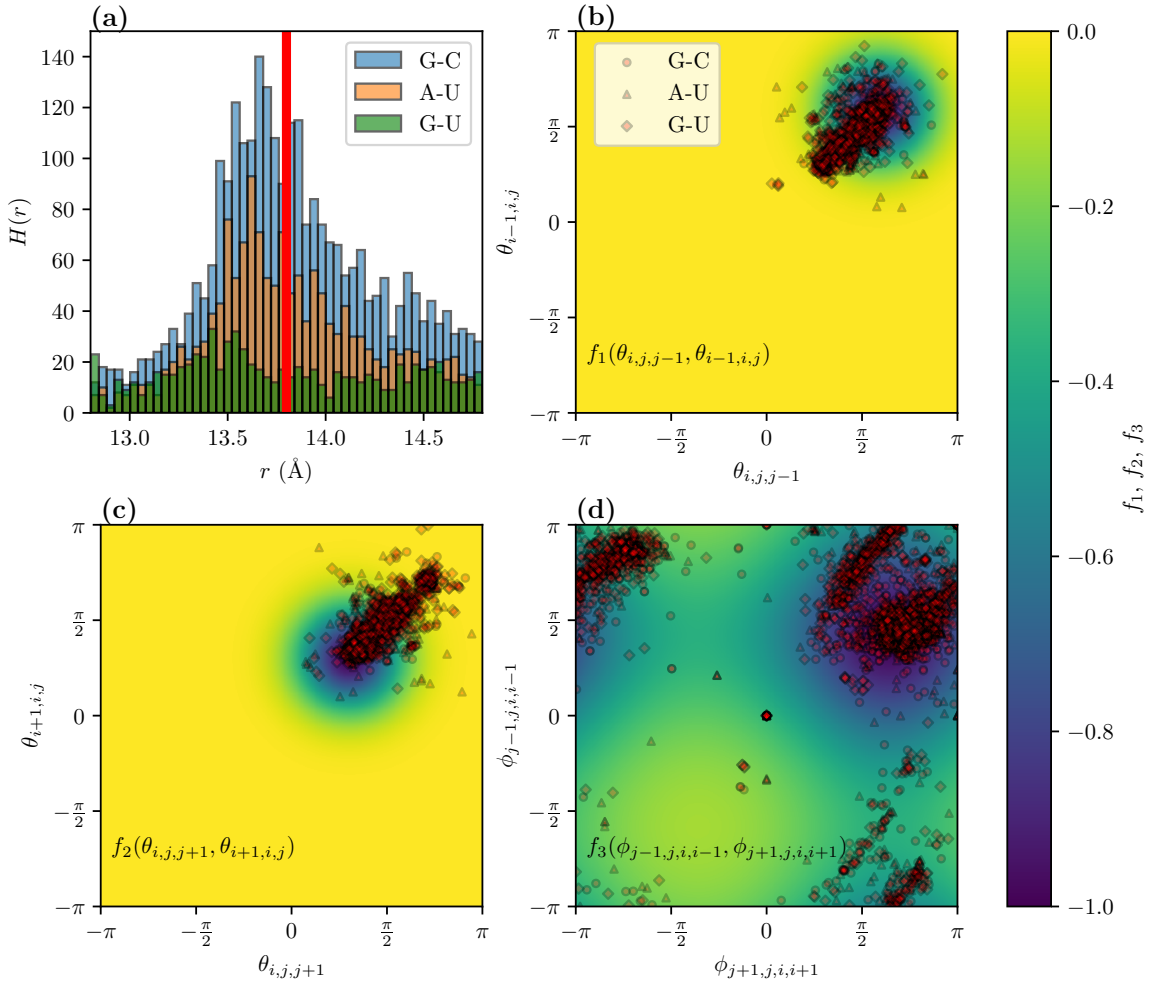

FIG. S2. **Evaluating current parameters of base-pairing potential against experimental structures.** **(a)** The base-pair distance (C3' $\cdots$ C3') distributions for the G–C, A–U, and G–U base-pairs peak close to the equilibrium distance in the model (red line, $r_0 = 13.8\,\text{Å}$). **(b)** Heatmap of base-pair angles $\theta_{i,j,j-1}$ and $\theta_{i-1,j,j}$, and the related base-pair functional form $f_1(\theta_{i,j,j-1}, \theta_{i-1,j,j})$. The scatter plot of the experimental values coincides with the minima of the potential for all three base pairs. **(c)** Heatmap of base-pair angles $\theta_{i,j,j+1}$ and $\theta_{i+1,j,j}$, and the related base-pair functional form $f_2(\theta_{i,j,j+1}, \theta_{i+1,j,j})$. The bioinformatic data coincides well with the corresponding functional form. **(d)** Heatmap of the base-pair dihedral angles $f\phi_{j-1,j,i,i-1}$ and $\phi_{j+1,j,i,i+1}$, and the related base-pair functional form $f_3(\phi_{j-1,j,i,i-1}, \phi_{j+1,j,i,i+1})$. The overlay of bioinformatic data shows good agreement with the energy landscape of the functional form.

TABLE S2. **Surveyed PDB IDs of RNAs in A-form**. Data obtained from the Nucleic Acid Knowledgebase (NAKB) with options of (1) antiparallel double helices, (2) A-form RNA and (3) without proteins.

1AJF	1AL5	1ATO	1ATV	1ATW	1BZ2	1BZT	1BZU	1CSL	1D0U	1D4R	1DQF	1DQH	1DUQ	1EBQ	1EBR
1EKA	1EKD	1ELH	1ESH	1F6X	1F79	1F7F	1F7I	1FYP	1G2J	1G3A	1GUC	1HS1	1HS2	1HS3	1HS4
1HS8	1I3X	1I3Y	1I46	1I4B	1I9X	1IDV	1JZC	1JZV	1K4B	1K5I	1KFO	1L1W	1L3Z	1LC6	1LMV
1LNT	1MFK	1MIS	1MNX	1MUV	1MV1	1MV2	1MV6	1MWG	1N8X	1NA2	1NC0	1NLC	1NZ1	1O3Z	1OQ0
1OW9	1PBM	1PJY	1QBP	1QC0	1QCU	1QES	1QET	1RNA	1RXA	1RXB	1SA9	1SAQ	1SDR	1SLO	1SLP
1SY4	1SYZ	1SZY	1T0D	1T0E	1TJZ	1TUT	1U2A	1WTS	1WTT	1WVD	1XHP	1XV0	1Y6S	1Y6T	1Y73
1Y90	1Y95	1Y99	1YFV	1YLG	1YN2	1YNC	1YY0	1YZD	1Z30	1Z79	1Z7F	1ZIH	1ZX7	1ZZ5	205D
255D	259D	280D	283D	2A04	2A0P	2AHT	2AO5	2B7G	2B8R	2CD1	2CD3	2CD5	2DD2	2ES5	2ET3
2FQN	2G1G	2G3S	2G91	2G92	2GBH	2GIP	2GVO	2H49	2HNS	2JR4	2JRG	2JSG	2JUK	2JWV	2JXQ
2JXS	2K41	2K5Z	2K65	2K66	2K7E	2KD8	2KEZ	2KF0	2KOC	2KPC	2KPD	2KRQ	2KRV	2KRZ	2KVN
2KXZ	2KY0	2KY1	2KYD	2KYE	2L2J	2L6I	2L8F	2L8U	2L8W	2L9E	2LAC	2LBQ	2LK3	2LP9	2LPA
2LPS	2LPT	2LV0	2LWK	2M1O	2MER	2MS5	2MVS	2MVY	2MXJ	2MXK	2N4L	2NBY	2NCR	2O32	2O3W
2O3Y	2O81	2O83	2OE5	2OE6	2OE8	2Q1O	2Q1R	2QEK	2R1S	2R20	2R21	2R22	2RLU	2RO2	2RPT
2U2A	2V6W	2V7R	2VAL	2VUQ	2W89	2XSL	2Y95	333D	353D	397D	3BNL	3BNT	3CGP	3CGQ	3CGR
3CGS	3CJZ	3FTM	3GLP	3GVN	3HGA	3JXQ	3JXR	3LOA	3MEI	3ND4	3NJ6	3NJ7	3OK2	3P4B	3P4C
3P4D	3R1C	3R1D	3R1E	3S49	3SZX	3TD0	402D	405D	409D	413D	420D	433D	434D	435D	438D
439D	462D	464D	466D	472D	4A4R	4A4S	4E48	4E59	4E5C	4E6B	4J50	4JAB	4JRT	4K27	4K31
4KYY	4MCE	4MCF	4MS9	4MSB	4MSR	4NFO	4NFP	4NFQ	4O41	4P3T	4P3U	4P43	4PCJ	4RBQ	4RBY
4RBZ	4RC0	4U34	4U35	4U37	4U38	4U3L	4U3O	4U3P	4U3R	4U47	4U78	4XW0	4YN6	4ZC7	5AY2
5AY3	5AY4	5BTM	5C5W	5D8T	5DA6	5DER	5DHB	5DHC	5DO5	5EW4	5EW7	5HBW	5HBX	5HBY	5HN2
5HNJ	5HNQ	5KRG	5KVJ	5L00	5LQO	5LQT	5LR3	5LR4	5LR5	5LSN	5MWI	5NXT	5T3K	5TDJ	5TGP
5TKO	5UED	5UEE	5UEF	5UEG	5UF3	5UZ6	5V0H	5V0J	5V0O	5V1K	5V1L	5V2H	5V2R	5V9Z	5VH7
5VH8	5VR4	5WQ1	5ZEM	6AAS	6AAU	6AZ4	6BG9	6BGB	6BMD	6C8D	6C8E	6C8I	6C8J	6C8K	6C8M
6C8N	6C8O	6CAB	6CXZ	6CY2	6CY4	6E7L	6EZ0	6H0R	6HBX	6HC5	6HMO	6HU6	6I1V	6I1W	6IA2
6JBG	6L0Y	6MXQ	6N8F	6N8H	6N8I	6NOA	6U6J	6U79	6U89	6U8U	6UGI	6UGJ	6VA1	6VA2	6VA3
6VA4	6VEM	6VU1	6VZC	6WY2	6WY3	6XUR	6XUS	6XWJ	6XWW	6XXB	6YMC	6Z18	6ZPF	6ZQ9	6ZR1
6ZRS	6ZX5	6ZX8	7A3Y	7A9N	7A9O	7A9Q	7A9S	7A9T	7ECJ	7ECK	7ECL	7ECM	7ECN	7ECO	7ECP
7EDT	7EDU	7JU1	7KUB	7KUC	7KUD	7KUK	7KUM	7KUN	7KUO	7KUP	7LNE	7LNF	7LNG	7QTN	7QUA
7U87	7U88	7U89	7U8A	7U8B	7UME	7VFT	7Y2B	7Y2P	7Y8P	8AMG	8AMI	8AMJ	8AMK	8AML	8AMM
8AMN	8BWT	8CLM	8CLR	8FCS	8I43	8SWG	8SWO	8SX5	8SX6	8SXL	8SY1				

3. Implementation of RNA model in LAMMPS

As mentioned in the main text, the force calculation for base pairing requires the careful evaluation of forces due to multiple angle references.

Below is one example of of our derivation for calculating the $\hat{\boldsymbol{x}}$ component of the force $\boldsymbol{f}_i$ on bead i due to the dihedral angle $\phi_{j+1,j,i,i+1}$ (Supplementary Figure S1(c)). First, the displacement vectors are defined from the position vectors of coarse-grained beads $\boldsymbol{r}_{j+1} = (x_{j+1}, y_{j+1}, z_{j+1})$, $\boldsymbol{r}_j = (x_j, y_j, z_j)$, $\boldsymbol{r}_i = (x_i, y_i, z_i)$, and $\boldsymbol{r}_{i+1} = (x_{i+1}, y_{i+1}, z_{i+1})$:

$$\boldsymbol{r}_{j+1,j} = \boldsymbol{r}_{j+1} - \boldsymbol{r}_j = (x_{j+1} - x_j, y_{j+1} - y_j, z_{j+1} - z_j), \tag{S4a}$$

$$\boldsymbol{r}_{i,j} = \boldsymbol{r}_i - \boldsymbol{r}_j = (x_i - x_j, y_i - y_j, z_i - z_j), \tag{S4b}$$

$$\boldsymbol{r}_{i+1,i} = \boldsymbol{r}_{i+1} - \boldsymbol{r}_i = (x_{i+1} - x_i, y_{i+1} - y_i, z_{i+1} - z_i). \tag{S4c}$$

Then, the normal vectors of the surfaces that are formed by these vectors are defined as

$$\boldsymbol{t} = \boldsymbol{r}_{j+1,j} \times \boldsymbol{r}_{i,j}, \tag{S5a}$$

$$\boldsymbol{u} = \boldsymbol{r}_{i+1,i} \times \boldsymbol{r}_{i,j}. \tag{S5b}$$

Thus, the dihedral angle can be calculated using the cos function

$$\cos(\phi_{j+1,j,i,i+1}) = \frac{\boldsymbol{t} \cdot \boldsymbol{u}}{|\boldsymbol{t}||\boldsymbol{u}|}, \tag{S6}$$

and thus force can be calculated using the base pair potential (Eq.4):

$$\begin{aligned}
\hat{\boldsymbol{x}} \cdot \boldsymbol{f}_i = -\hat{\boldsymbol{x}} \cdot \boldsymbol{\nabla}_i U_{\mathrm{BP}} &= -\left(\frac{\mathrm{d}U_{\mathrm{BP}}}{\mathrm{d}\phi_{j+1,j,i,i+1}}\right)\left(\frac{\mathrm{d}\phi_{j+1,j,i,i+1}}{\mathrm{d}\cos(\phi_{j+1,j,i,i+1})}\right)\hat{\boldsymbol{x}} \cdot \boldsymbol{\nabla}_i \cos(\phi_{j+1,j,i,i+1}) \\
&= -\left(\frac{\mathrm{d}U_{\mathrm{BP}}}{\mathrm{d}\phi_{j+1,j,i,i+1}}\right)\left(\frac{\mathrm{d}\cos(\phi_{j+1,j,i,i+1})}{\mathrm{d}\phi_{j+1,j,i,i+1}}\right)^{-1}\left[\hat{\boldsymbol{x}} \cdot \left(\hat{\mathbf{x}}\frac{\partial}{\partial x_i} + \hat{\mathbf{y}}\frac{\partial}{\partial y_i} + \hat{\mathbf{z}}\frac{\partial}{\partial x_j}\right)\cos(\phi_{j+1,j,i,i+1})\right] \\
&= -U_{\mathrm{BP}} k_\phi \sin(\phi_{j+1,j,i,i+1} + \phi_2)\left(\frac{-1}{\sin\phi_{j+1,j,i,i+1}}\right)\frac{\partial\cos(\phi_{j+1,j,i,i+1})}{\partial x_i} \\
&= \frac{U_{\mathrm{BP}} k_\phi \sin(\phi_{j+1,j,i,i+1} + \phi_2)}{\sin(\phi_{j+1,j,i,i+1})}\left(\frac{\partial\cos(\phi_{j+1,j,i,i+1})}{\partial x_i}\right) \\
&= \frac{U_{\mathrm{BP}} k_\phi \sin(\phi_{j+1,j,i,i+1} + \phi_2)}{\sin(\phi_{j+1,j,i,i+1})}\left(\frac{\partial\cos(\phi_{j+1,j,i,i+1})}{\partial t_x}\frac{\partial t_x}{\partial x_k} + \frac{\partial\cos(\phi_{j+1,j,i,i+1})}{\partial t_y}\frac{\partial t_y}{\partial x_k} + \frac{\partial\cos(\phi_{j+1,j,i,i+1})}{\partial t_z}\frac{\partial t_z}{\partial x_k}\right. \\
&\qquad \left. + \frac{\partial\cos(\phi_{j+1,j,i,i+1})}{\partial u_x}\frac{\partial u_x}{\partial x_k} + \frac{\partial\cos(\phi_{j+1,j,i,i+1})}{\partial u_y}\frac{\partial u_y}{\partial x_k} + \frac{\partial\cos(\phi_{j+1,j,i,i+1})}{\partial u_z}\frac{\partial u_z}{\partial x_k}\right) \\
&= \frac{U_{\mathrm{BP}} k_\phi \sin(\phi_{j+1,j,i,i+1} + \phi_2)}{\sin(\phi_{j+1,j,i,i+1})}\left[(z_{j+1} - z_j)\left(\frac{u_y}{|\mathbf{t}||\mathbf{u}|} - \frac{t_y}{|\mathbf{t}|^2}\cos(\phi_{j+1,j,i,i+1})\right)\right. \\
&\qquad - (y_{j+1} - y_j)\left(\frac{u_z}{|\mathbf{t}||\mathbf{u}|} - \frac{t_z}{|\mathbf{t}|^2}\cos(\phi_{j+1,j,i,i+1})\right) + (z_{i+1} - z_j)\left(\frac{t_y}{|\mathbf{t}||\mathbf{u}|} - \frac{u_y}{|\mathbf{u}|^2}\cos(\phi_{j+1,j,i,i+1})\right) \\
&\qquad \left. -(y_{i+1} - y_j)\left(\frac{t_z}{|\mathbf{t}||\mathbf{u}|} - \frac{u_z}{|\mathbf{u}|^2}\cos(\phi_{j+1,j,i,i+1})\right)\right].
\end{aligned} \tag{S7}$$

A similar derivation procedure applies to calculations of all other 65 force components. For the pressure calculation, LAMMPS uses the pressure tensor that takes into account the contribution of the kinetic energy tensor and the virial tensor. Thus, for example, the contribution of the 4 forces $\boldsymbol{f}_{j+1}$, $\boldsymbol{f}_j$, $\boldsymbol{f}_i$ and $\boldsymbol{f}_{i+1}$ from the dihedral angle $\phi_{j+1,j,i,i+1}$ to the xx component of the virial tensor is

$$\begin{aligned}
p_{xx} &= \frac{1}{V}\left[x_{j+1}\left(\hat{\boldsymbol{x}} \cdot \boldsymbol{f}_{j+1}\right) + x_j\left(\hat{\boldsymbol{x}} \cdot \boldsymbol{f}_j\right) + x_i\left(\hat{\boldsymbol{x}} \cdot \boldsymbol{f}_i\right) + x_{i+1}\left(\hat{\boldsymbol{x}} \cdot \boldsymbol{f}_{i+1}\right)\right] \\
&= \frac{1}{V}\left\{x_{j+1}\left(\hat{\boldsymbol{x}} \cdot \boldsymbol{f}_{j+1}\right) + x_j\left(\hat{\boldsymbol{x}} \cdot \boldsymbol{f}_j\right) + x_i\left(\hat{\boldsymbol{x}} \cdot \boldsymbol{f}_i\right) + x_{i+1}\left[\hat{\boldsymbol{x}} \cdot \left(\hat{\mathbf{0}} - \boldsymbol{f}_{j+1} - \boldsymbol{f}_j - \boldsymbol{f}_i\right)\right]\right\} \\
&= \frac{1}{V}\left[\left(x_{j+1} - x_{i+1}\right)\left(\hat{\boldsymbol{x}} \cdot \boldsymbol{f}_{j+1}\right) + \left(x_j - x_{i+1}\right)\left(\hat{\boldsymbol{x}} \cdot \boldsymbol{f}_j\right) + \left(x_i - x_{i+1}\right)\left(\hat{\boldsymbol{x}} \cdot \boldsymbol{f}_i\right)\right].
\end{aligned} \tag{S8}$$

Similarly, the contributions to the pressure tensor are calculated for all the forces involved in the base-pair interaction.

The following are the potential energy definitions in a LAMMPS input script.

```
units real
dimension 3
boundary p p p
atom_style full
#### harmonic bond between bonds
bond_style harmonic
#    coeff for harmonic bond, specify 3:
#    * bond type
#    * K/2
#    * equilibrium distance
bond_coeff 1 7.5 5.9

#### harmonic angle potential between bonds
angle_style harmonic
#    coeff for harmonic angle, specify 3:
#    * angle type
#    * K/2
#    * equilibrium angle in degrees
angle_coeff 1 5.0 150.0

#### overlay of WCA-LJ potential and Base-Pair Potential
pair_style hybrid/overlay lj/cut 10.0 base/rna 18.0

#  pair_coeff for WCA-LJ potential, specify 2:
#    * energy
#    * sigma
pair_coeff * * lj/cut 2.0 8.9089871814
pair_modify pair lj/cut shift yes

#  pair_coeff for Base Pair Interactions, specify 9:
#    * energy Ubp: A-1, C-2, G-3, U-4
#    * harmonic K_r
#    * Equilibrium distance r_0
#    * harmonic K_theta
#    * dihedral K_phi
#    * Equilibrium angle theta1
#    * Equilibrium angle theta2
#    * Equilibrium angle phi1
#    * Equilibrium angle phi2
pair_coeff 1 1 base/rna 0.00000 3.0 13.8 1.5 0.5 1.8326 0.9425 1.8326 1.1345 #AA
pair_coeff 1 2 base/rna 0.00000 3.0 13.8 1.5 0.5 1.8326 0.9425 1.8326 1.1345 #AC
pair_coeff 1 3 base/rna 0.00000 3.0 13.8 1.5 0.5 1.8326 0.9425 1.8326 1.1345 #AG
pair_coeff 1 4 base/rna 3.33333 3.0 13.8 1.5 0.5 1.8326 0.9425 1.8326 1.1345 #AU
pair_coeff 2 2 base/rna 0.00000 3.0 13.8 1.5 0.5 1.8326 0.9425 1.8326 1.1345 #CC
pair_coeff 2 3 base/rna 5.00000 3.0 13.8 1.5 0.5 1.8326 0.9425 1.8326 1.1345 #CG
pair_coeff 2 4 base/rna 0.00000 3.0 13.8 1.5 0.5 1.8326 0.9425 1.8326 1.1345 #CU
pair_coeff 3 3 base/rna 0.00000 3.0 13.8 1.5 0.5 1.8326 0.9425 1.8326 1.1345 #GG
pair_coeff 3 4 base/rna 3.33333 3.0 13.8 1.5 0.5 1.8326 0.9425 1.8326 1.1345 #GU
pair_coeff 4 4 base/rna 0.00000 3.0 13.8 1.5 0.5 1.8326 0.9425 1.8326 1.1345 #UU

special_bonds lj 0.0 0.0 1.0 # no excluded volume calculation for bonded beads
```

Listing S1. The part of the potential definitions in LAMMPS scripts.

4. Validation of results for CAG repeats

After calibration and verification for single chains, as described in the main text, we simulate multichain systems of RNA $(CAG)_n$ repeats to test their length-dependent phase behavior. Specifically, we simulate two CAG repeat systems, $(CAG)_{47}$ at concentration $200\,\mu$M, and $(CAG)_{20}$ at concentration $50\,\mu$M. Each system is composed of 64 chains and simulated with periodic boundary conditions. The systems are prepared with an even distribution of chains and melted at 373 K. Then the systems are cooled down to 293 K within 40 ns and simulated at 293 K to observe for phase separation. The results are shown in Figure S3 (a, b). We have successfully reproduced the phase behavior for these two systems: $(CAG)_{47}$ at concentration 200μM condenses but $(CAG)_{20}$ at concentration 50μM does not. More importantly, the result of $(CAG)_{47}$ at concentration 200μM are obtained in less than 3 days (~ 60 hours) using 16 CPUs, whereas the original implementation took about 3 months using a GPU for the same system. As shown in Figure S3(c), the measured average concentration of the droplet matches well with the previous reports that the distribution centers around 8 mM [14].

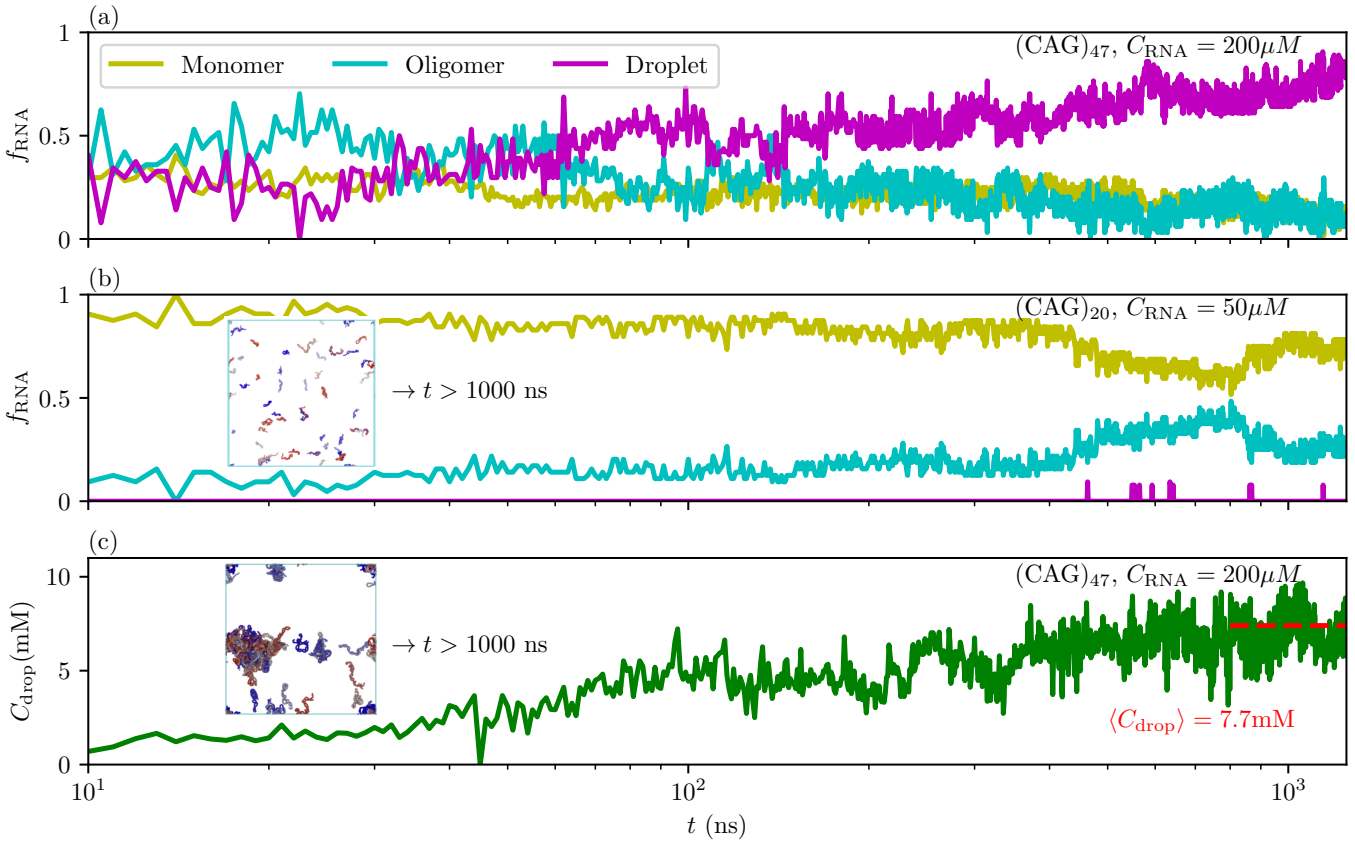

FIG. S3. **Reproduction of condensation results for CAG repeats from previous implementation [14]. (a)** Time evolution of $(CAG)_{47}$ monomer, oligomer and droplet populations at 200μM, where monomer is a single chain, oligomer is defined as intermediate clusters with 2–4 chains, and droplets are large clusters with ≥ 5 chains. We observed consistent dynamics where droplets constituted the predominant composition in $(CAG)_{47}$. **(b)** Time evolution of different structural ensembles in $(CAG)_{20}$ at 50μM. As in the original simulations, we do not observe phase separation. The less populated oligomers are primarily dimers. **(c)** Concentration of the droplet in $(CAG)_{47}$ from the same simulation as in **(a)**. For the two systems, simulation snapshots ($t > 1000$ ns) are provided in **(b)** and **(c)**.

As a final validation of our implementation in LAMMPS, we compare the energy trajectories, Figure S4. As demonstrated, the energy trajectories are consistent. The slight difference in trajectories stem from random thermal fluctuations.

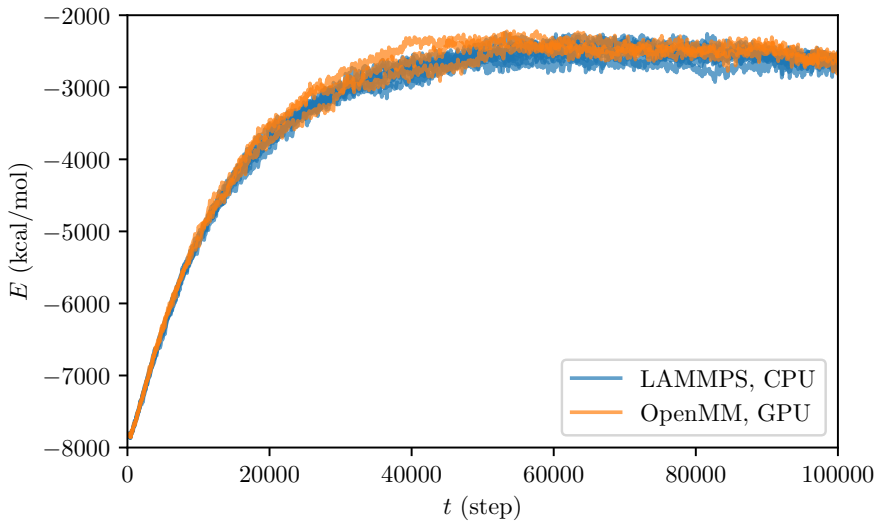

FIG. S4. **Consistency of potential energy calculations in different implementations.** The potential energy trajectories from benchmarks are shown for both CPU simulations in LAMMPS and GPU simulations in OpenMM. Data are obtained from the same simulations that are summarized in Figure1(d).

5. Redundancy of RNA trinucleotide repeats

To systematically study the phase behavior of all trinucleotide repeats, we have also determined the redundancy. As illustrated in Fig. S5, there are 24 non-redundant RNA trinucleotide repeats, and each group of three trinucleotide repeats describes the same chain when repeat numbers are large enough. Since the current model does not consider other interaction types, we have also excluded the homopolymers (poly(rA), poly(rC), poly(rG) and poly(rU)). Therefore, we have simulated 20 non-redundant trinucleotide repeats for this study.

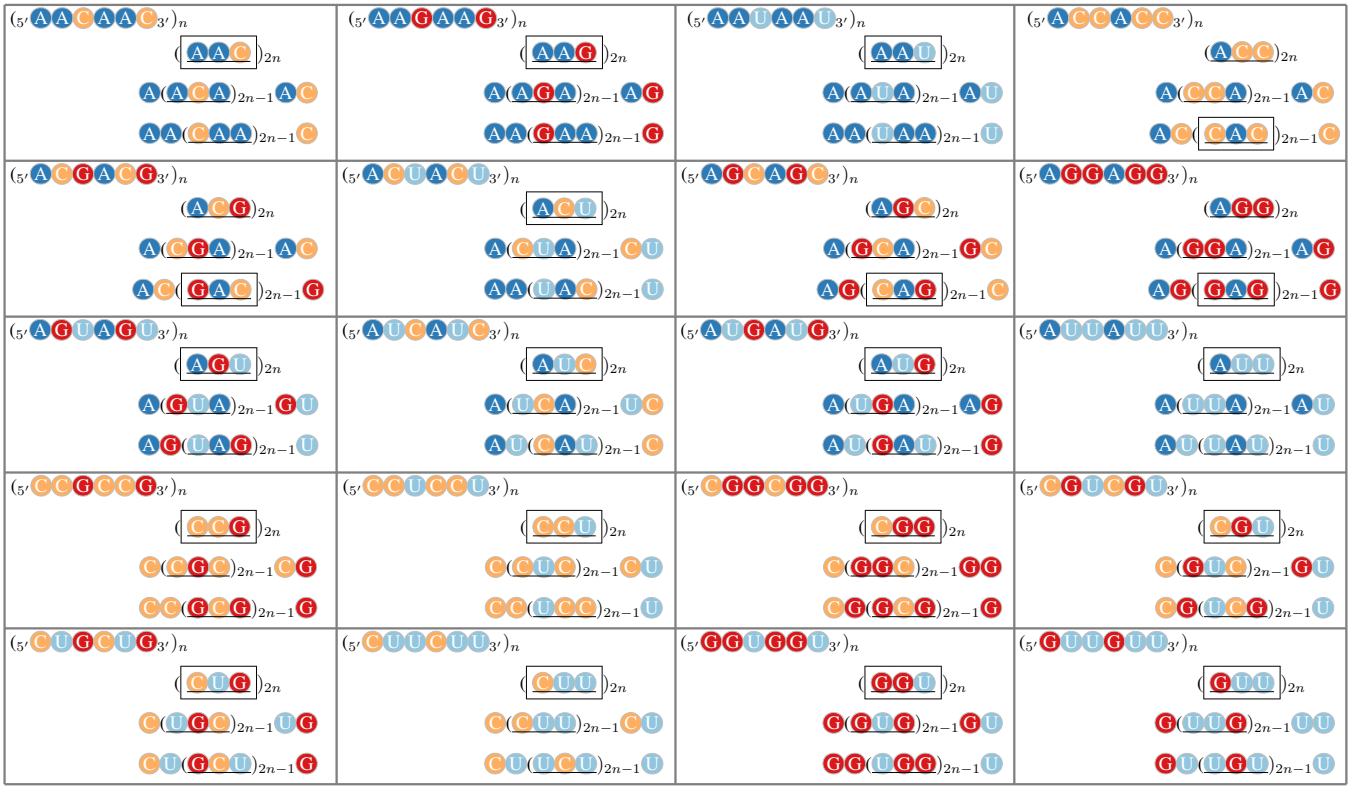

FIG. S5. **Illustration of redundancy in trinucleotide repeats.** Considering the directionality from 5' to 3', there are 64 (4^3) combinations. Excluding the 4 homopolymers, there are 60 different trinucleotide sequences. However, as illustrated in each box, when the trinucleotide repeat is long enough (large n), each three trinucleotide (underlined) repeat sequences represent the same chain. Therefore, there are only 20 non-redundant trinucleotide repeats. The simulated systems for this paper are shown in square frames.

6. Droplet concentrations and sizes

Here we also include the time evolution of droplet concentrations (Fig. S6) and the size of the largest droplet (Fig. S7)

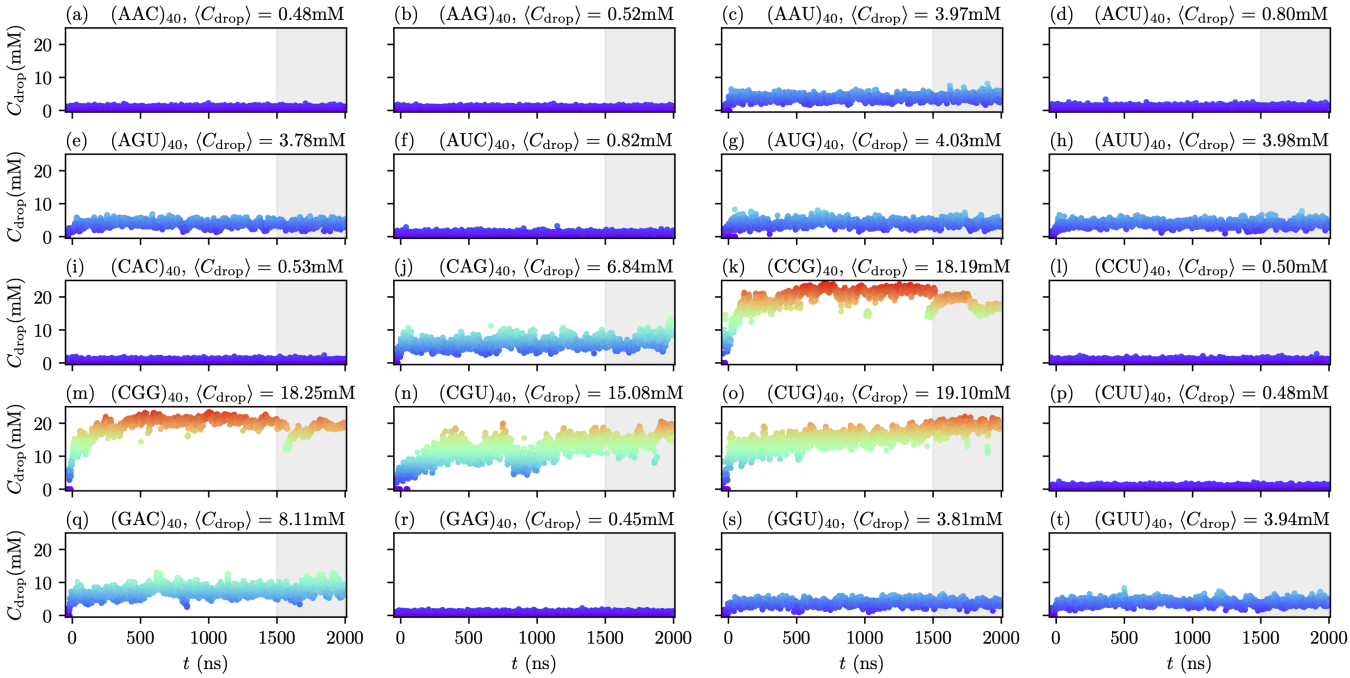

FIG. S6. **Time evolution of the concentration of droplets C_{drop} of all 20 non-redundant trinucleotide repeats.** The concentration is calculated from the number of chains in the droplets and the volume of the droplets, the latter being estimated from the gyration tensor of the clusters. Data from the last 500 ns (shaded gray) is used to calculate average values of C_{drop} in the system. The data points are colored based on C_{drop} values.

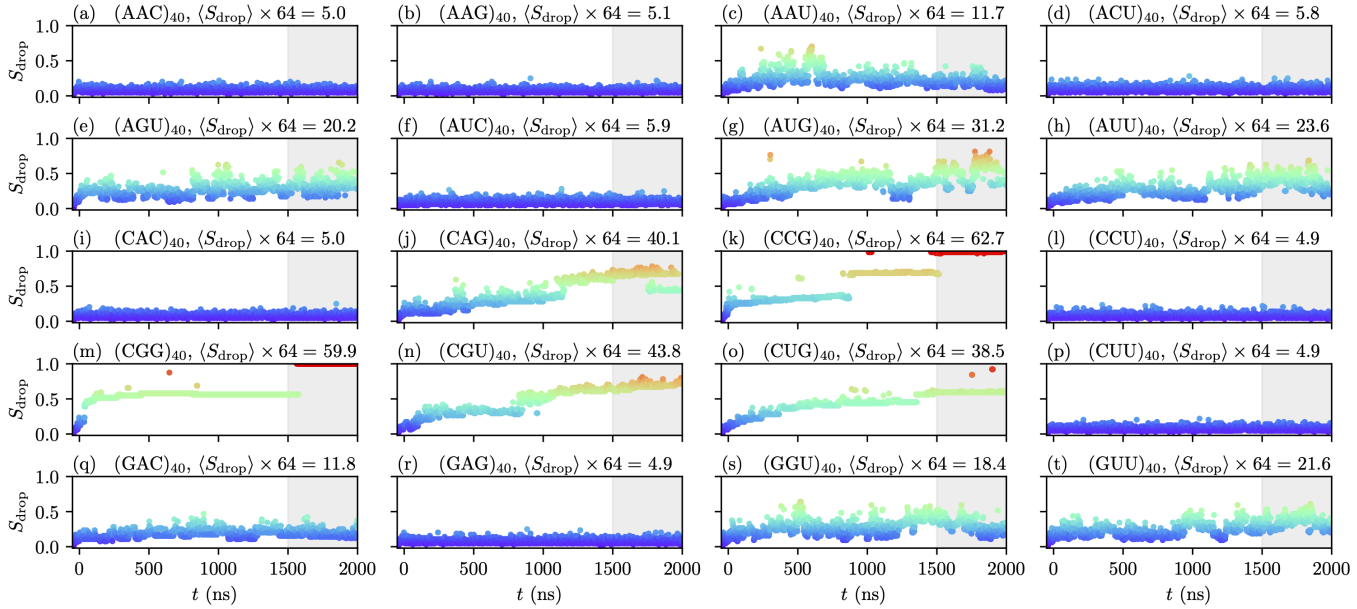

FIG. S7. **Time evolution of the largest droplet fraction S_{drop} of all 20 non-redundant trinucleotide repeats.** Data from the last 500ns (shaded gray) is used to calculated average values of S_{drop} in the system. The data points are colored based on S_{drop} values.

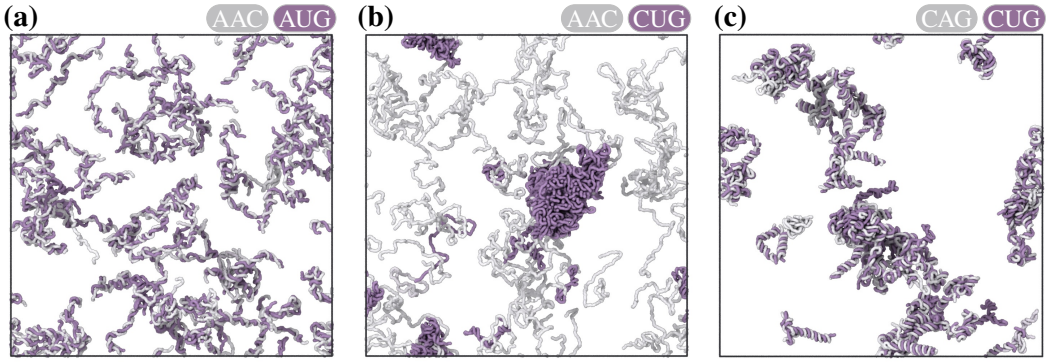

FIG. S8. **Binary mixtures of RNA systems at 1:1 ratio. (a)** $(\mathrm{AAC})_{40}$ and $(\mathrm{AUG})_{40}$. **(b)** $(\mathrm{AAC})_{40}$ and $(\mathrm{CUG})_{40}$. **(c)** $(\mathrm{CAG})_{40}$ and $(\mathrm{CUG})_{40}$. Each mixture contains 128 chains (64 each) at a total concentration of 200 μM. Systems are simulated for 2 μs.

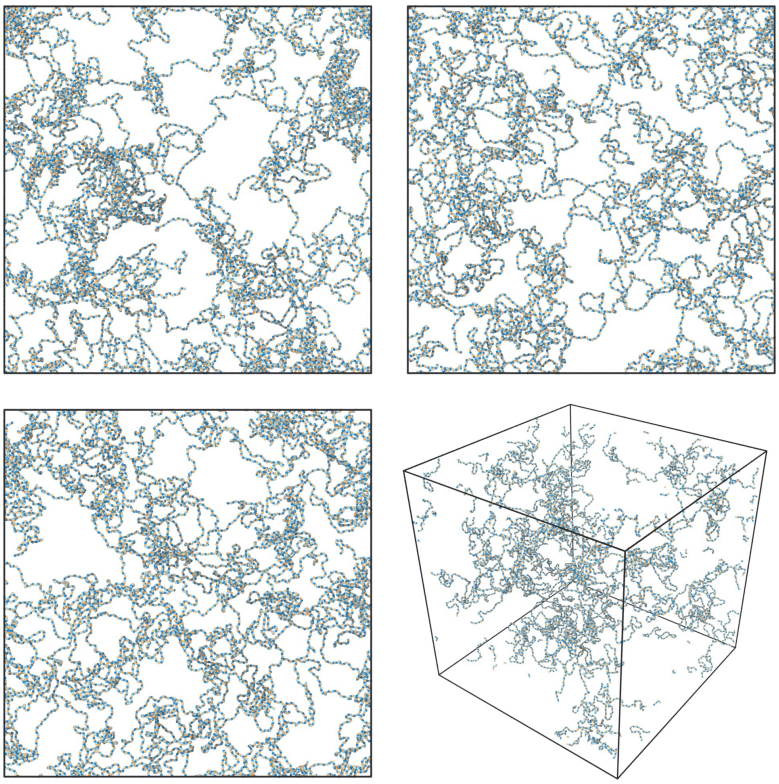

FIG. S9. **Simulation results of (ACU)$_{200}$.** The system is composed of 27 chains and simulated with periodic boundary conditions. The systems are prepared at a similar concentration as (ACU)$_{40}$ at 200 μM with chains evenly distributed and melted at 373 K. Then the systems are cooled down to 293 K within 40 ns. Finally, the systems are simulated at 293 K for 2 μs. Consistent with ACU)$_{40}$, within 2 μs, the system does not phase separate.