Sequence dependence of biomolecular phase separation

Benjamin G. Weiner,[1] Yigal Meir,[1,2] and Ned S. Wingreen[3,4]

[1] *Department of Physics, Princeton University, Princeton, New Jersey 08544, USA*
[2] *Department of Physics, Ben Gurion University of the Negev, Beer-Sheva 84105, Israel*
[3] *Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA*
[4] *Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA*

Cells are home to a host of biomolecular condensates – phase-separated droplets that lack a membrane. In addition to nonspecific interactions, phase separation depends on specific binding motifs between constituent molecules. Nevertheless, few rules have been established on how these specific, heterotypic interactions drive phase separation. Using lattice-polymer simulations and mean-field theory, we find that the sequence of binding motifs strongly affects a polymer's ability to phase separate, influencing both phase boundaries and condensate properties (e.g. viscosity). Notably, sequence primarily acts by determining the conformational entropy of self-bonding by single polymers.

Eukaryotic cells contain a variety of phase-separated biomolecular condensates that organize intracellular processes ranging from ribosome assembly and metabolism to signaling and stress response [1–3]. How do the thermodynamic and material properties of these condensates emerge from their components, and how do cells regulate condensate formation and function? Unlike the droplets of simple molecules or homopolymers, intracellular condensates are typically composed of hundreds of molecular species, each with multiple interaction motifs. While the precise sequences of these motifs are believed to play a major role in determining condensates' phase diagrams and material properties, the nature of this relation has only begun to be explored [4, 5].

Previous studies have established important principles relating phase separation to the sequence of nonspecific interaction domains such as hydrophobic or electrostatic motifs [6–9]. However, in many cases condensate formation and function depend on specific interactions including residue-residue bonds, protein-protein bonds, and protein-RNA bonds, which are one-to-one and saturating [2]. Such one-to-one interactions between heterotypic domains are ubiquitous in biology, and recent studies have enumerated a large number of examples in both one-component [10] and two-component [11, 12] systems (e.g. cation-pi bonds between tyrosine and arginine in FUS-family proteins, protein-protein bonds in the SIM-SUMO system). Here, we address the important question: what is the role played by sequence when specific, heterotypic interactions are the dominant drivers of phase separation?

To address this question, we analyzed a model of polymers with specific, heterotypic interaction motifs using Monte Carlo simulations and mean-field theory. We found that motif sequence determines both the size of the two-phase region and dense-phase properties such as viscosity and polymer extension. Importantly, sequence acts primarily by controlling the entropy of self-bonds, suggesting a new paradigm for biological control of intracellular phase separation.

Specifically, we developed an FCC lattice model where each polymer consists of a sequence of "A" and "B" motifs which form specific, saturating bonds of energy ϵ (Fig. 1(a) and 1(b)). Monomers on adjacent lattice sites have nonspecific interaction energy J. We used Monte Carlo simulations in the Grand Canonical Ensemble (GCE): the 3D conformations of the polymers are updated using a predefined move-set, and polymers are inserted/deleted with chemical potential μ. (See Supplemental Material for details [13].) For each sequence, we first determined the critical point (temperature T_c, chemical potential μ_c, and density ϕ_c). Then for each $T < T_c$ we located the phase boundary, defined by the value μ^* for which the dilute and dense phases have equal thermodynamic weight. Around this value of μ, the simulated system transitions between the two phases, leading to a polymer number distribution $P(N)$ that has two peaks with equal weights (Fig. 1(c)) [14]. Multicanonical sampling was employed to adequately sample transitions [13].

How does a polymer's sequence of interaction motifs affect its ability to phase separate? We constructed phase diagrams for polymers with the six sequences shown in Fig. 1(a), all with $L = 24$ motifs ($a = b = 12$) arranged in repeating domains. Each simulation contains polymers of a single sequence, and the sequences differ only in their domain sizes ℓ. Figure 2(a) shows the resulting phase diagrams, which differ dramatically by domain size, e.g. the T_c values for $\ell = 2$ and $\ell = 12$ differ by 20%; if the former were in the physiological range around 300K, this 60K difference would render the condensed phase of $\ell = 2$ inaccessible in most biological contexts. Despite this wide variation, Fig. 2(b) shows that rescaling by T_c and ϕ_c causes the curves to collapse. This is expected near the critical point, where all sequences share the behavior of the 3D Ising universality class [14], but the continued nearly exact data collapse indicates that (T_c, ϕ_c) fully captures the sequence-dependence of the phase diagram.

Why does the sequence of binding motifs have such a strong effect on phase separation? Importantly, sequence determines the entropy of intra-polymer bonds,

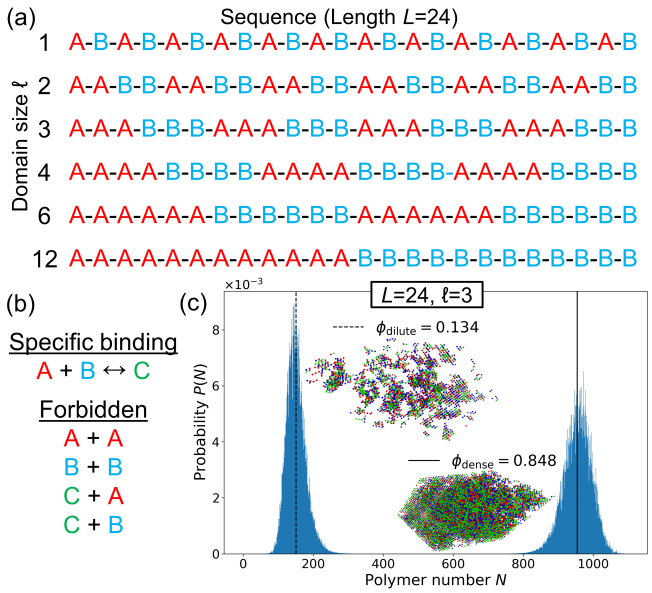

FIG. 1. Lattice model for phase separation by polymers with one-to-one interacting motifs. (a) Each polymer is defined by its sequence of motifs, which come in types "A" (red) and "B" (blue). The class of sequences shown consists of repeated domains of As and Bs, labeled by their domain size ℓ. (b) In lattice simulations, an A and a B motif on the same lattice site form a specific, saturating bond (green) with binding energy ϵ. Monomers of any type on adjacent lattice sites have an attractive nonspecific interaction energy J. We set $J = 0.05\epsilon$ unless otherwise noted. (c) Simulations are conducted in the Grand Canonical Ensemble (GCE), where the polymer number N fluctuates. In the two-phase region the polymer number distribution has two peaks, and when these have equal weight, their mean polymer numbers define the concentrations of the phase boundaries. Simulation parameters: $\ell = 3$, binding energy $\beta\epsilon = 0.9287$. *Inset:* Snapshots of the GCE simulation at densities corresponding to the dilute and dense phase boundaries.

i.e. the facility of a polymer to form bonds with itself, as quantified by $g(s)$, the number of configurations with s self-bonds (which can be extracted from Monte Carlo simulations of a single polymer (Fig. 2(c))). Sequences with small domain sizes have many more conformations available to them at all s (see [13] for a semi-log plot). Intuitively, a sequence like $\ell = 2$ allows a polymer to make many local bonds, whereas a sequence like $\ell = 12$ cannot form multiple bonds without folding up globally, which is entropically unfavorable. Consequently, it is more favorable for polymers like $\ell = 12$ to phase separate so each polymer can form trans-bonds with others, leading to a high T_c value.

This intuition can be captured by a simple mean-field theory that incorporates the single-polymer properties $g(s)$ and motif numbers a, b. We make two mean-field simplifications: 1) every polymer has the mean number of trans-bonds $\bar{t}$ (i.e. for every polymer i, $t_i = \bar{t}$), and 2) each polymer interacts with the mean-field background

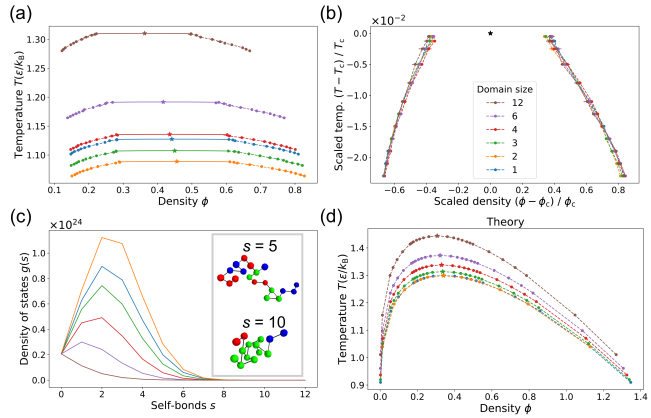

FIG. 2. The sequence of binding motifs strongly affects a polymer's ability to phase separate. (a) Binodal curves defining the two-phase region for the six sequences of length $L = 24$ shown in Fig. 1(a). Stars indicate the critical points and the solid curves are fits to scaling relations for the 3D Ising universality class. Each point shows the mean ± SD for three replicates. (Uncertainties are too small to see for most points.) (b) When rescaled by the critical temperature T_c and critical density ϕ_c, the phase boundaries in (a) collapse, even far from the critical point. Color key applies to all panels. (c) The tendency to phase separate is inversely related to the entropy of self-interactions, as quantified by the density of states $g(s)$, i.e. the number of ways a given sequence can form s bonds with itself. Inset: Snapshots of $\ell = 3$ polymer with $s = 5$ (top) and $s = 10$ (bottom). (d) Phase boundaries from a mean-field theory that accounts for sequence only through $g(s)$.

of motifs, leading to the following free energy density (see [13] for full expressions and derivation):

$$f \equiv \frac{F}{k_B T V} = f_{\text{steric}}(\bar{s}, \bar{t}) + f_{\text{trans}}(\bar{s}, \bar{t}) + \beta\chi\phi^2 - \frac{\phi}{L}\left(\log \sum_s g(s) e^{ws}\right) + \frac{\phi}{L} w\bar{s} - \frac{\phi}{L}\beta\epsilon\left(\bar{s} + \frac{\bar{t}}{2}\right), \quad (1)$$

where V is the number of lattice sites and χ is the nonspecific-interaction parameter. f_{steric} is the translational contribution from the number of ways to place polymers without overlap and f_{trans} is the entropy of forming $\bar{t}$ trans-bonds given $\bar{s}$ self-bonds, derived from the combinatorics of pairing independent motifs. w is the self-bond weight chosen to self-consistently enforce $\sum_i s_i/N = \bar{s}$. This allows us to estimate the entropy of $\bar{s}$ without assuming that $s_i = \bar{s}\ \forall\ i$. In the thermodynamic limit the partition function is dominated by the largest term, so we minimize Eq. 1 with respect to $\bar{s}$ and $\bar{t}$ at each ϕ to yield $f(\phi)$ and determine the phase diagram.

Figure 2(d) shows mean-field phase diagrams. In spite of the approximations, the theory captures the main patterns observed in the full Monte Carlo simulations. Specifically, sequences with larger motif domains have larger two-phase regions and these extend to higher temperatures. Rescaling by T_c and ϕ_c also causes the mean-

field phase boundaries to collapse [13]. Intriguingly, the mean-field theory does not correctly place the $\ell = 1$ sequence in the T_c hierarchy. The single-polymer density of states $g(s)$ suggests that $\ell = 1$ should be similar to $\ell = 2$, but its T_c is closer to to $\ell = 4$. We trace this discrepancy to trans-bond correlations in the dense phase: the $\ell = 1$ sequence tends to form segments of multiple bonds rather than independent bonds [13]. Overall, the success of the theory demonstrates that sequence mainly governs phase separation through the entropy of self-interactions. We capture this dependence, as well as corrections due to dense-phase correlations, in a simple "condensation parameter" described below.

Do these conclusions still hold if the motifs are not arranged in regular domains, and how do polymer length and motif stoichiometry affect phase separation? To address these questions, we located the critical points for three new types of sequences: 1) Length $L = 24$ sequences with $a = b = 12$ in scrambled order, 2) domain sequences with $L \neq 24$, and 3) sequences with $L = 24$ but $a \neq b$.

Figure 3(a) shows T_c and ϕ_c for the scrambled $L = 24$ sequences and for domain sequences of various lengths. T_c and ϕ_c are negatively correlated across all sequence types, because for low-T_c sequences, trans-bonds – and consequently, phase separation – only become favorable at higher polymer density.

In Fig. 3(b) we show T_c as a function of length for different domain sizes and observe that the T_c hierarchy is preserved across sequence lengths. Thus domain size is a robust predictor of T_c via its relationship with self-bond entropy. The dashed curve in Fig. 3(c) shows T_c for scrambled sequences with unequal motif stoichiometry. T_c decreases as the motif imbalance grows because the dense phase is crowded with unbonded motifs, making phase separation less favorable. How does this effect relate to the role of $g(s)$? Scrambled sequences are clustered near the $\ell = 3$ sequence in (T_c, ϕ_c) space (Fig. 5(a)), so we generated sequences by starting with $\ell = 3$ and randomly mutating B motifs into A motifs (Fig. 5(c), solid curve). The $\ell = 3$ mutants follow the same pattern as the scrambled sequences, indicating that self-bond entropy and stoichiometry are nearly independent inputs to T_c.

The mean-field theory of Eq. 1 also captures the behavior of these more general sequences, as shown in Fig. 3(d). [15] This reinforces the picture that phase separation is mainly governed by the relative entropy of intra- and inter-polymer interactions. The former is captured by $g(s)$ and the latter is described by pairing motifs without reference to their arrangement on the polymer (although see [13] regarding the $\ell = 1$ sequence and correlations). To capture these effects in a single number, we propose a condensation parameter X ("Chi") which correlates with a sequence's ability to phase separate (see

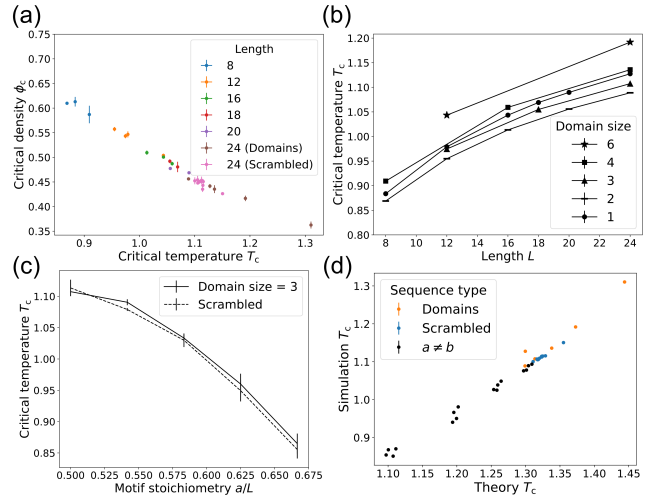

FIG. 3. Ability to phase separate is determined by the sequence of binding motifs for polymers of different lengths, patterns, and motif stoichiometries. (a) T_c and ϕ_c for $L = 24$ polymers with scrambled sequences and domain sequences of various lengths. Mean ± SD over three replicates. (Temperature uncertainties are too small to see in (a), (b), and (d).) (b) T_c as a function of length for sequences with different domain sizes. Mean ± SD over three replicates. (c) T_c as a function of motif stoichiometry a/L. The solid curve corresponds to $\ell = 3$ sequences where a number of B motifs are randomly mutated to A motifs, and the dashed curve shows scrambled sequences. Each point shows mean ± SD over four sequences. (d) T_c from Monte Carlo simulations versus mean-field theory for domain sequences, scrambled sequences, and sequences with unequal motif stoichiometry. Mean ± SD over three replicates.

[13] for a heuristic derivation):

$$X \equiv -\log\left(\frac{1}{(r_A)^b (r_B)^a} \sum_s \frac{g(s)}{(4P_{\text{corr}})^{s/2}}\right), \qquad (2)$$

where $r_{A/B} = (a/b)/L$ is the fraction of motifs that are A/B and P_{corr} is a simple, easily calculable measure of trans-bond correlations (see [13] for details). As shown in Fig. 4(a), this accurately captures the phase separation hierarchy of T_c, including the correlation-enhanced T_c of the $\ell = 1$ sequence.

Are domain sequences special? The space of possible sequences is much larger than can be explored via Monte Carlo simulations. However, we can use the condensation parameter to quickly estimate T_c for any sequence. By approximating $\sum_s g(s) \approx g(1)$ and estimating $g(1)$ from a loop-entropy calculation, we can obtain X without simulation. We then estimate T_c for new sequences using a linear fit of T_c to X for the domain sequences. (See [13] for details.) Figure 4(b) shows the distribution of critical temperatures calculated in this way for 20,000 random sequences with $a = b = 12$. Strikingly, the distribution is sharply peaked at low T_c, comparable to the T_c values for the $\ell = 2$ and $\ell = 3$ domain sequences. Sequences

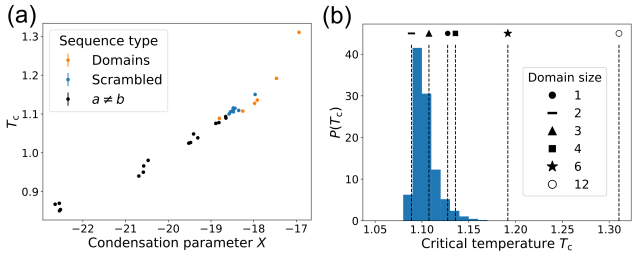

FIG. 4. The sequence dependence of phase separation is captured by a "condensation parameter" X, which takes into account the entropy of self-interactions, motif stoichiometry, and dense-phase correlations. (a) T_c values from simulations as a function of X. Mean ± SD over three replicates. (b) Distribution of T_c values for 20,000 random sequences of length $L = 24$ with $a = b$, calculated from X values and the linear T_c versus X relation for domain sequences. Domain sequence T_c values are marked.

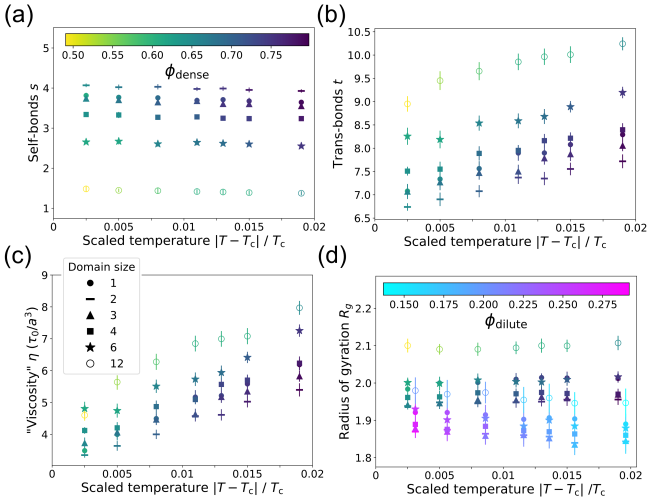

FIG. 5. The structure of the dense phase depends on the sequence of binding motifs. (a) Number of self-bonds s in the dense phase as a function of reduced temperature for domain sequences (symbols as in (c)). Each point shows s (mean ± SD) over all configurations with $\phi \in [\phi_{\text{dense}} - 0.01, \phi_{\text{dense}} + 0.01]$. Color bar: droplet density. (b) Number of trans-bonds t (bonds with other polymers) versus temperature as in (a). (c) Viscosity (Eq. 3) of the dense phase, shown as in (a). Symbol key applies to all panels. (d) Radius of gyration R_g of polymers in the dense phase (shown as in (a)) and in the dilute phase. Each dilute-phase point shows R_g (mean ± SD) over all configurations with $\phi \in [\phi_{\text{dilute}} - 0.01, \phi_{\text{dilute}} + 0.01]$. The ϕ_{dilute} points have the same scaled temperature as the adjacent ϕ_{dense} points, but are shifted for clarity. Color bar: dilute phase density.

with larger domains (e.g. $\ell = 6$ and $\ell = 12$) have anomalously high T_c values, suggesting that they are unlikely to evolve without selection.

The sequence of specific-interaction motifs influences not only the formation of droplets, but also their physical properties. Figures 5(a) and 5(b) show the number of self-bonds and trans-bonds in the dense phase. The sequences have very different T_c values, but the data collapse of Fig. 2(b) allows us to compare their degree of bonding relative to scaled temperature $|T - T_c|/T_c$. Density fluctuates in the GCE, so each point in Fig. 5 is averaged over configurations with ϕ within 0.01 of the phase boundary. Figure 5(a) shows that the single-polymer properties are still relevant in the dense phase: smaller domain sizes lead to more self-bonds, and the hierarchy matches the ordering of $g(s)$ (Fig. 2(c)). Figure 5(b) shows that larger domains lead to more trans-bonds, even though the droplets are less dense. As temperature is reduced – and thus density is increased – the number of trans-bonds increases, whereas the number of self-bonds is relatively stable. Interestingly, we conclude that even though the phase boundaries collapse to the same curve, different sequences lead to droplets with very different internal structures.

These structural differences will affect the physical properties of the dense phase. The timescales of a droplet's internal dynamics will determine whether it behaves more like a solid or a liquid. We might expect denser droplets to have slower internal dynamics, so the $\ell = 1$ and $\ell = 2$ sequences would be the most solid-like. However, the extra inter-polymer bonds at large ℓ will slow down the dynamics. To disentangle these effects, we estimate the viscosity and polymer-diffusivity by modeling the dense phase as a viscoelastic polymer melt with reversible cross-links formed by trans-bonds. Then the viscosity is expected to scale as [16]

$$\eta \sim G\tau = \left(k_{\text{B}}T\frac{\phi}{m^3 L}\right)\left(\tau_b \bar{t}^2\right), \tag{3}$$

where m is the monomer length, G is the elastic modulus ($k_{\text{B}}T$ times the number density of polymers), and τ is the relaxation time of the polymer melt. The relaxation time depends on the trans-bonds per polymer $\bar{t}$ and the bond lifetime $\tau_b = \tau_0 \exp(\beta\epsilon)$, where τ_0 is a microscopic time which we take to be sequence-independent. Figure 5(c) shows the inferred dense-phase viscosity. Sequences with large domains have more viscous droplets due to the strong dependence on inter-polymer bonds, which are the main repository of elastic "memory" in the melt. By the same arguments leading to Eq. 3, diffusivity scales as $1/\bar{t}$ [13], so polymers with large domains will also diffuse more slowly within droplets.

The motif sequence also affects the polymer radius of gyration in both phases (Fig. 5(d)). In the dense phase, polymers with large domains adopt expanded conformations which allow them to form more trans-bonds. Polymers of all sequences are more compact in the dilute phase, where there are fewer trans-bonds and nonspecific interactions with neighbors. Thus self-bonds cause

polymers to contract, while trans-bonds cause them to expand.

In summary, we developed a simple lattice-polymer model to study how the sequence of specific-interaction motifs affects phase separation. We found that motif sequence determines the size of the two-phase region by setting the relative entropy of intra- and inter-polymer bonds. In particular, large domains disfavor self-bonds and thus favor phase separation. This is consistent with recent experimental [17] and theoretical [6, 8] studies on coacervation (phase separation driven by electrostatics) where small charge-domains lead to screening of the attractive forces driving aggregation. However, electrostatic interactions (generic, longer-range, promiscuous) are qualitatively very different from the interactions in our model (specific, local, saturating). This points to a different underlying mechanism: in the former sequence changes the electrostatic energetics of the dense phase, but in the latter sequence controls the conformational entropy of the dilute phase. Thus specific interactions provide a distinct physical paradigm for the control of intracellular phase separation.

We then analyzed how sequence influences condensates' physical properties such as viscosity and diffusivity. Previous studies have related such physical properties to overall amino acid composition [10, 18] and particular RNA-binding domains [19], and simulations of amphiphilic polymers have linked sequence to aggregate morphology [9]. We found that motif sequence strongly affects both droplet density and inter-polymer connectivity, and, in particular, that sequences with large domains form more viscous droplets with slower internal diffusion. All sequences expand in the dense phase to form more trans-bonds, and notably, small-domain sequence are always more compact than large-domain sequences. This contrasts with results on single polyampholytes, where "blocky" sequences with large domains are more compact [20, 21].

Taken together, these results suggest that motif sequence provides cells with a means to tune the formation and properties of intracellular condensates. For example, motif stoichiometry could be an active regulatory target – a cell could dissolve droplets by removing just a few binding motifs per polymer through post-translational modifications. The negative correlation between T_c and ϕ_c provides another regulatory knob: if a particular condensate density is required at fixed temperature, this can be achieved by either tuning the binding strength or modifying the sequence. However, the physics underlying condensation via specific interactions also implies biological constraints: the same trans-bonds that drive condensation for high-T_c sequences also lead to high viscosity.

We have used a simple model of biological condensates to show how the sequence of specific-interaction motifs affects phase separation, thus linking the microscopic details of molecular components to the emergent properties relevant for biological function. However, many open questions remain: how are distinct droplets maintained by networks of specific-interaction motifs? How do specific interactions shape the phases of more complex condensates with many components and a variety of interaction types? We hope our work encourages further research across a range of theoretical and experimental systems.

[1] A. A. Hyman, C. A. Weber, and F. Jülicher, Liquid-liquid phase separation in biology, Annual Review of Cell and Developmental Biology **30**, 39 (2014).

[2] S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen, Biomolecular condensates: organizers of cellular biochemistry, Nature Reviews Molecular Cell Biology **18**, 285 (2017).

[3] S. Boeynaems, S. Alberti, N. L. Fawzi, T. Mittag, M. Polymenidou, F. Rousseau, J. Schymkowitz, J. Shorter, B. Wolozin, L. Van Den Bosch, P. Tompa, and M. Fuxreiter, Protein phase separation: a new phase in cell biology, Trends in Cell Biology **28**, 420 (2018).

[4] C. P. Brangwynne, P. Tompa, and R. V. Pappu, Polymer physics of intracellular phase transitions, Nature Physics **11**, 899 (2015).

[5] S. Alberti, A. Gladfelter, and T. Mittag, Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates, Cell **176**, 419 (2019).

[6] Y.-H. Lin, J. D. Forman-Kay, and H. S. Chan, Sequence-specific polyampholyte phase separation in membraneless organelles, Physical review letters **117**, 178101 (2016).

[7] S. Das, A. N. Amin, Y.-H. Lin, and H. S. Chan, Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters, Physical Chemistry Chemical Physics **20**, 28558 (2018).

[8] J. McCarty, K. T. Delaney, S. P. Danielsen, G. H. Fredrickson, and J.-E. Shea, Complete phase diagram for liquid–liquid phase separation of intrinsically disordered proteins, The journal of physical chemistry letters **10**, 1644 (2019).

[9] A. Statt, H. Casademunt, C. P. Brangwynne, and A. Z. Panagiotopoulos, Model for disordered proteins with strongly sequence-dependent liquid phase behavior, The Journal of Chemical Physics **152**, 075101 (2020).

[10] J. Wang, J.-M. Choi, A. S. Holehouse, H. O. Lee, X. Zhang, M. Jahnel, S. Maharana, R. Lemaitre, A. Pozniakovsky, D. Drechsel, *et al.*, A molecular grammar governing the driving forces for phase separation of prion-like rna binding proteins, Cell **174**, 688 (2018).

[11] J. A. Ditlev, L. B. Case, and M. K. Rosen, Who's in and who's out—compositional control of biomolecular condensates, Journal of molecular biology **430**, 4666 (2018).

[12] B. Xu, G. He, B. G. Weiner, P. Ronceray, Y. Meir, M. C. Jonikas, and N. S. Wingreen, Rigidity enhances a magic-number effect in polymer phase separation, Nature communications **11**, 1 (2020).

[13] See supplemental material for methods and additional discussion.

[14] A. Z. Panagiotopoulos, V. Wong, and M. A. Floriano, Phase equilibria of lattice polymers from histogram reweighting monte carlo simulations, Macromolecules **31**, 912 (1998).

[15] The magnitude of the theoretical T_c depends on the nonspecific-interaction parameter χ; see [13].

[16] M. Rubinstein and A. N. Semenov, Dynamics of entangled solutions of associating polymers, Macromolecules **34**, 1058 (2001).

[17] C. W. Pak, M. Kosno, A. S. Holehouse, S. B. Padrick, A. Mittal, R. Ali, A. A. Yunus, D. R. Liu, R. V. Pappu, and M. K. Rosen, Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein, Molecular cell **63**, 72 (2016).

[18] S. C. Weber, Sequence-encoded material properties dictate the structure and function of nuclear bodies, Current Opinion in Cell Biology **46**, 62 (2017).

[19] M.-T. Wei, S. Elbaum-Garfinkle, A. S. Holehouse, C. C.-H. Chen, M. Feric, C. B. Arnold, R. D. Priestley, R. V. Pappu, and C. P. Brangwynne, Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles, Nature Chemistry **9**, 1118 (2017).

[20] R. K. Das and R. V. Pappu, Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues, Proceedings of the National Academy of Sciences **110**, 13392 (2013).

[21] L. Sawle and K. Ghosh, A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins, The Journal of chemical physics **143**, 08B615_1 (2015).