Biophysical Journal
Article

Biophysical Society

Connecting Coil-to-Globule Transitions to Full Phase Diagrams for Intrinsically Disordered Proteins

Xiangze Zeng,[1,2] Alex S. Holehouse,[2,3] Ashutosh Chilkoti,[4] Tanja Mittag,[5] and Rohit V. Pappu[1,2,*]
[1]Department of Biomedical Engineering and [2]Center for Science & Engineering of Living Systems, Washington University in St. Louis, St. Louis, Missouri; [3]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri; [4]Department of Biomedical Engineering, Duke University, Durham, North Carolina; and [5]Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, Tennessee

ABSTRACT Phase separation is thought to underlie spatial and temporal organization that is required for controlling biochemical reactions in cells. Multivalence of interaction motifs, also known as stickers, is a defining feature of proteins that drive phase separation. Intrinsically disordered proteins with stickers uniformly distributed along the linear sequence can serve as scaffold molecules that drive phase separation. The sequence-intrinsic contributions of disordered proteins to phase separation can be discerned by computing or measuring sequence-specific phase diagrams. These help to delineate the combinations of protein concentration and a suitable control parameter, such as temperature, that support phase separation. Here, we present an approach that combines detailed simulations with a numerical adaptation of an analytical Gaussian cluster theory to enable the calculation of sequence-specific phase diagrams. Our approach leverages the known equivalence between the driving forces for single-chain collapse in dilute solutions and the driving forces for phase separation in concentrated solutions. We demonstrate the application of the theory-aided computations through calculation of phase diagrams for a set of archetypal intrinsically disordered low-complexity domains. We also leverage theories to compute sequence-specific percolation lines and thereby provide a thermodynamic framework for hardening transitions that have been observed for many biomolecular condensates.

SIGNIFICANCE Intrinsically disordered proteins that have the requisite valence of adhesive linear motifs can drive phase separation and give rise to membraneless biomolecular condensates. Knowledge of how phase diagrams vary with amino acid sequence and changes to solution conditions is essential for understanding how proteins contribute to condensate assembly and dissolution. In this work, we introduce a new, to our knowledge, two-pronged computational approach to predict sequence-specific phase diagrams. This approach starts by extracting key parameters from simulations of single-chain coil-to-globule transitions. We use these parameters in our numerical implementation of the Gaussian cluster theory for polymer solutions to construct sequence-specific phase diagrams. The method is efficient and demonstrably accurate and should pave the way for high-throughput assessments of phase behavior.

INTRODUCTION

Intrinsically disordered proteins (IDPs) that have the requisite valence of adhesive linear motifs—stickers—can drive phase separation and give rise to membraneless biomolecular condensates (1–5). Knowledge of how phase diagrams vary with amino acid sequence and changes to solution conditions is an essential component of understanding how proteins contribute to condensate assembly and their regulated dissolution (1–4,6). Comparative assessments of sequence-specific phase diagrams are also invaluable for enabling the design of novel disordered low-complexity domains (LCDs) that give rise to synthetic condensates for engineering applications (7–13). In this work, we introduce a two-pronged computational approach to predict sequence-specific phase diagrams that describe phase separation. We start by extracting key parameters from simulations of single-chain coil-to-globule transitions. We use these in our numerical implementation of the Gaussian cluster theory (GCT) for polymer solutions (14,15) to construct sequence-specific phase diagrams.

Our approach leverages the known equivalence between the driving forces for coil-to-globule transitions of

Submitted May 13, 2020, and accepted for publication June 11, 2020.

*Correspondence: pappu@wustl.edu

Editor: Scott Showalter.
https://doi.org/10.1016/j.bpj.2020.06.014

individual polymer molecules in dilute solutions and phase separation that arises from collective interactions among polymer molecules in concentrated solutions (10,14–19). For systems with an upper critical solution temperature (UCST) (Fig. 1), the coil-to-globule transition is characterized by indifferent solvent conditions at the θ temperature (T_θ), poor solvent conditions below T_θ, and good solvent conditions above T_θ (11). At finite concentrations, for $T < T_\theta$, chains overlap with one another, and because of cohesive interactions among chain molecules, there exist temperature-dependent threshold volume fractions above which the system separates into two phases, namely a dilute solution that coexists with a dense phase. For a temperature $T < T_\theta$, we denote the volume fractions of the coexisting dilute and dense phases as ϕ_{b1} and ϕ_{b2}, respectively (Fig. 1). Although the dilute arm of the binodal, namely ϕ_{b1} as a function of T, can be obtained via a range of experimentally tractable approaches, accurate estimates of the dense phase concentration—i.e., ϕ_{b2} as a function of T—are plagued with several challenges. These include the need for high concentrations of protein, artifacts from fluorescent reporters, the impact of finite size effects, and slow equilibration within dense phases (7). As a result, despite their overarching importance for characterizing condensates, measurements of system-specific coexistence curves (binodals) are available for only a handful of LCDs (3,8–10,20).

The driving forces for temperature-dependent coil-to-globule transitions and phase separation are governed by the three-way interplay among chain-chain, chain-solvent, and solvent-solvent interactions (16). The consequences of this three-way interplay are captured via an effective two-body interaction coefficient B, which has units of inverse volume. The value of B is positive below T_θ for a system with UCST behavior and negative for a system above T_θ for a system with a lower critical solution temperature (LCST). These features of B signify net attractions among monomer units below T_θ for a UCST system and above T_θ for a LCST system. The three-body interaction coefficient w (Fig. 2) is always positive, and to first approximation, it is taken to be independent of temperature (16); the magnitude of w is determined, in part, by the local persistence length or thickness of the chain. The steepness of coil-to-globule transitions and the temperature-dependent values of ϕ_{dense} are governed by the ratio of the two-/three-body coefficient (14,15).

Disordered LCDs (as well as folded domains) can be mapped onto "stickers-and-spacers" architectures (10,21–25). Stickers are attractive groups (26,27) that include short linear motifs or individual residues that form reversible physical (noncovalent) cross-links with one another; conversely, the excluded volumes of spacers control the extent of compaction of individual chains and the cooperativity of phase transitions that give rise to two coexisting phases (25). Recent studies have shown that the correspondence between coil-to-globule transitions and phase separation, which is well established for flexible homopolymers, is also applicable for describing the phase behavior of disordered LCDs (18,28). For this correspondence to apply, the stickers have to be uniformly distributed along the linear sequence (10). Intriguingly, a significant fraction of naturally occurring disordered LCDs, specifically prion-like domains, and designed sequences have the requisite sequence features that should enable correspondence between the driving forces that control coil-to-globule transitions and phase separation (9–11,29–32).

We reasoned that it should be possible to compute complete phase diagrams using quantitative analysis of sequence-specific coil-to-globule transitions in conjunction with a requisite theoretical framework that directly connects coil-to-globule transitions to phase separation. Accordingly, we deployed a two-pronged strategy for calculating sequence-specific phase diagrams of disordered LCDs. First, all-atom or coarse-grained simulations are used to obtain temperature-dependent profiles for coil-to-globule

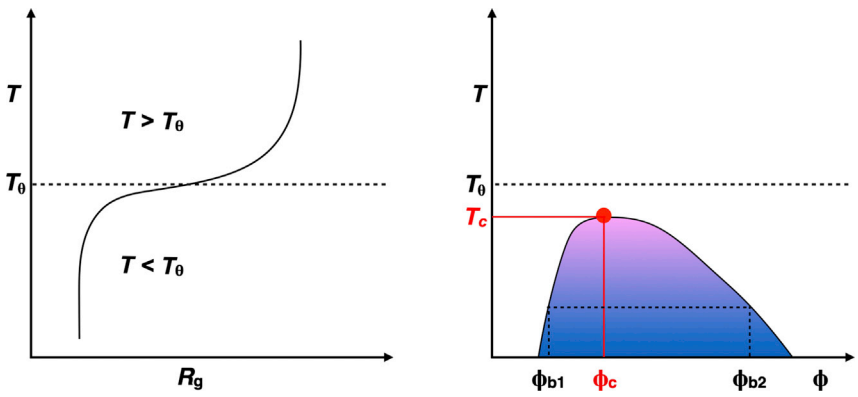

FIGURE 1 Equivalence between single-chain coil-to-globule transitions and phase separation. The schematic depicts expectations for a system that shows UCST behavior. Coil-to-globule transitions whereby coils form at temperatures above T_θ (*dashed line*) and globules form below T_θ. At T_θ, the radius of gyration scales as $N^{0.5}$ with chain length N. There is a system-specific collapse temperature $T_{collapse}$ below which $R_g \sim N^{0.33}$. The crossover regime, whose width is defined by the ratio of the two-/three-body interaction coefficient, corresponds to the interval $T_{collapse} < T < T_\theta$, where the scaling exponent for R_g lies between 0.33 and 0.5. Expected phase diagram for a flexible polymer that shows UCST behavior. The critical point defined by the critical temperature (T_c) and critical volume fraction (ϕ_c) is the point at which the binodal and spinodal coincide. The binodal is shown as the contour that delineates the two-phase regime (shown in *colored shading*, where the colors become cooler away from the critical point). For a given temperature $T < T_\theta$ and $T < T_c$, there exists a volume fraction above which the system separates into two coexisting phases, a dilute and dense phase characterized by volume fractions $\phi_{b1}(T)$ and $\phi_{b2}(T)$. The difference between $\phi_{b1}(T)$ and $\phi_{b2}(T)$ decreases as T approaches T_c. To see this figure in color, go online.

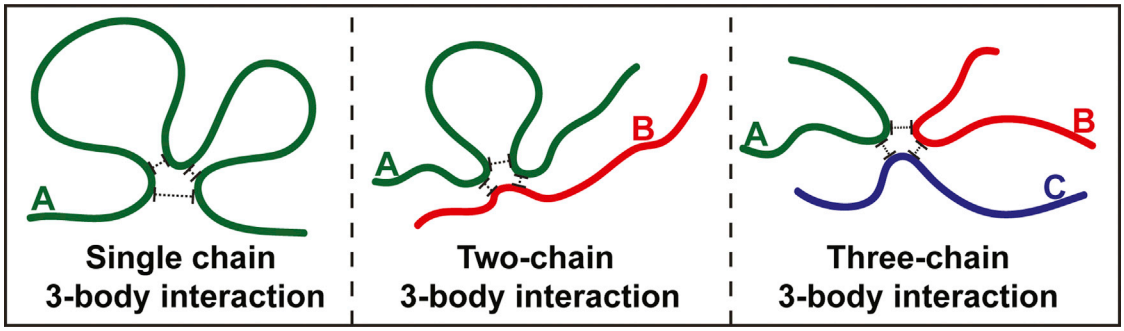

FIGURE 2 Depicting how three-body interactions are realized. Three-body interactions are realizable in at least three different ways: within a single polymer (*left*), between pairs of polymers (*middle*), or among three polymers (*right*). The interactions are always repulsive, and this derives from the finite thickness and steric impenetrability of polymers. To see this figure in color, go online.

transitions, and the results are analyzed to extract T_θ, B, and w. Next, these parameters are used in our numerical adaptation of the GCT for polymer solutions (14,15) to construct full phase diagrams in the two-parameter (ϕ,T) space.

The GCT was developed for generic homopolymers, each comprising N monomers connected by $N-1$ bonds of length l_0. The distributions of intramolecular and intermolecular distances are described in terms of distances between statistical Kuhn segments instead of monomers. At their specific theta temperatures, any homopolymer can be reduced to a freely jointed chain of Kuhn segments, where the latter refers to the length scale beyond which the chemical details of the repeating units can be ignored (33). If n_s is the number of bonds per Kuhn segment, then the number of Kuhn segments is defined as $n_K = (N-1)/n_s$ and the mean-square length of each Kuhn segment is written as $l_K^2 = n_s C_\infty l_0^2$, where C_∞ is the characteristic ratio of the chain (33). The GCT assumes Gaussian forms for the distributions of intrachain and interchain distances between Kuhn segments (14,15). The main parameters of the Gaussian distributions are T_θ, $B(T)$, and w. Mean-field free energies, written in terms of moments of the Gaussian distributions, describe the temperature dependence of intrachain and interchain interactions. Minimization of the sum of single- and multichain free energies allows us to extract temperature-dependent values of $\phi b_1^{(T)}$ and $\phi b_2^{(T)}$, respectively. The locus of these temperature-dependent values defines the coexistence curve, or "binodal"; likewise, the locus of points at which the second derivative of the sum of single- and multichain free energies is zero helps define the instability curve, or "spinodal." In this way, we can compute system-specific full phase diagrams by minimization of the overall free energy for a polymer solution based on parameters obtained from single-chain coil-to-globule transitions.

The remainder of the manuscript is organized as follows: first, we summarize the main tenets of the GCT. Next, we detail how simulations of coil-to-globule transitions of individual LCDs are used to extract parameters for use in the GCT and demonstrate how we use our numerical implementation of the GCT to compute sequence-specific, temperature-dependent binodals and spinodals. We then illustrate the accuracy of our two-pronged approach by computing phase diagrams for LCDs in which temperature-dependent binodals have been measured. We further illustrate the methodology by computing full phase diagrams for distinct synthetic systems that have UCSTs and LCSTs, respectively. In doing so, we compare the calculated binodals with recent experimental measurements of binodals for the A1-LCD system and designed variants of this sequence (10). We conclude with a discussion of how the GCT-aided two-pronged method might be useful for high-throughput characterizations of phase diagrams of disordered LCDs.

MATERIALS AND METHODS

Systems of interest

We demonstrate the application of GCT-aided calculation of phase diagrams using simulations of coil-to-globule transitions for a series of systems. We start with the 137-residue LCD from the hnRNP-A1 protein (34). Recently, Martin et al. (10) characterized the phase behavior of this system and three designed variants that differ from the wild-type (WT) sequence in terms of the number (valence) of aromatic stickers. The sequences of interest are referred to as WT-A1; Aro^+, which was designed to have higher valence of aromatic residues than WT-A1; and Aro^- and $Aro^{-\ -}$, which have lower numbers of aromatic stickers than the WT (10). For these systems, we use coarse-grained simulations of coil-to-globule transitions that are based on PIMMS (10,35), which is a lattice-based simulation engine.

Polyglutamine (polyQ) expansions are associated with Huntington's disease and at least eight other neurodegenerative disorders (36). Considerable effort has focused on characterizing the conformational equilibria of individual polyQ chains and the phase equilibria of these systems (19,37–52). Homopolymers of polyQ forms globules in aqueous solvents; additionally, experiments and simulations point to UCST phase behavior for polyQ (37,45,52,53). Evidence has also been presented for the prospect of disordered, oligomeric or liquid-like phases serving as precursors of fibril formation (37,38,52,54,55). Here, we use the ABSINTH implicit solvation model (56) to perform all-atom simulations of coil-to-globule transitions. We use these simulations to demonstrate how the GCT can be used to construct phase diagrams for a biologically relevant system of homopolypeptides.

Finally, systems that undergo thermoresponsive phase behavior are of considerable interest for the design of biomaterials (9,29–31). As the third set of systems, we demonstrate the applicability of the two-pronged approach for constructing temperature-dependent phase diagrams of two

systems, viz. $(QGQSPYG)_9$ and $(TPKAMAP)_9$. These synthetic systems are designed to have UCST and LCST behaviors, respectively.

Coarse-grained and all-atom simulations of coil-to-globule transitions

PIMMS is a lattice-based simulation engine that uses a single bead per amino acid residue (10,35). Simulations of coil-to-globule transitions of WT-A1, Aro^+, Aro^-, and Aro^{--} were performed using a stickers-and-spacers representation of these sequences. Here, aromatic residues are stickers, and all other residues are spacers. In our model, we set $k_B = 1$ and T is in the interval $30 \leq T \leq 60$. Martin et al. (10) showed that sticker-sticker energies of -14, sticker-spacer energies of -5, and spacer-spacer energies of -2 provide a minimalist model that reproduces the variation of R_g with sticker valence in dilute solutions and experimentally measured binodals in the space of ϕ and T, with suitable conversions between the reduced temperature units and Kelvin (10). In this work, we use results from PIMMS simulations as the touchstone for comparative assessments of GCT-derived phase diagrams, and we use unitless, reduced temperatures, setting $k_BT = 1$ for comparative calculations on the four systems, WT-A1, Aro^+, Aro^-, and Aro^{2-}.

The all-atom simulations for polyQ sequences denoted as Q_n ($n = 20$, 30, 40, 50, 60, 70) and the two sequences $(QGQSPYG)_9$ and $(TPKAMAP)_9$ were performed using version 2.0 of the CAMPARI modeling suite (http://campari.sourceforge.net/). The N- and C-termini are capped using N-acetyl and N′-methylamide groups, respectively. All simulations use the ABSINTH implicit solvent model and force field paradigm. The parameters were derived based on the OPLS-AA/L force field, and all the parameters are implemented in the abs_3.2_opls.prm parameter set. Parameters for proline residues were based on the work of Radhakrishnan et al. (57). The cutoff for the short-range Lennard-Jones potential is set to be 10 Å. The cutoff for electrostatic interactions between different sites on neutral groups is 14 Å. There are no cutoffs among charge groups that carry a net charge, and these pertain to solution ions as well as charge groups on the side chains. In systems with charged residues, specifically $(TPKAMAP)_9$, nine Cl^- ions, modeled using the parameters developed by Mao and Pappu (58), were included to ensure electroneutrality. In addition, ion pairs were included to mimic an excess NaCl concentration of 1 mM. For the temperature dependence of macroscopic dielectric constant of water, we use the form adapted by Wuttke et al. (59).

ABSINTH uses experimentally derived reference free energies of solvation for model compounds (56,60). The reference free energies of solvation have weak temperature dependencies for polar compounds. Accordingly, we use the reference values at 298 K for the model compounds that mimic side chains of Gln, Ser, Gly, Pro, and Tyr. However, to model LCST behavior for $(TPKAMAP)_9$, we adapt the generalization introduced by Wuttke et al. (59). Specifically, we set the reference free energies of solvation for the backbone amide and the model compounds that mimic the side chains to be temperature dependent. The parameters for the enthalpy and heat capacity that are used in the Gibbs-Helmholtz equation to estimate temperature-dependent free energies of solvation are summarized in Table S1.

Standard Metropolis Monte Carlo (MMC) simulations aided by thermal replica exchange (61) were used to sample conformations of polyQ systems and the two synthetic sequences. Move sets for MMC simulations include side-chain torsions, concerted rotations, pivot moves, translations of ions, and translations and rotations of individual molecules (10,57,60,62). Details regarding the number of replicas and the temperature schedule for each system are listed in Table S2. Swaps between neighboring replicas were proposed once every 5×10^4 MMC simulation steps. Simulations were initiated using a randomly generated self-avoiding conformation for each polypeptide. All simulations use a spherical droplet, and the droplet is large enough to ensure against confinement artifacts. Multiple independent replica exchange simulations were performed—three for $(TPKAMAP)_9$ and four for the other systems—and each independent replica exchange simulation comprises a total of 6×10^7 MMC simulation steps.

Theory

GCT for a single chain

The workflow for constructing full phase diagrams using simulations of single-chain coil-to-globule transitions and the use of GCT based on parameters extracted from the single-chain simulations is summarized in Fig. 3. We first connect the GCT formalism to results from single-chain simulations and show how this connection allows us to estimate T_θ, the two-body interaction coefficient B, and the three-body interaction coefficient w. The GCT prescribes specific forms for the distribution of intrachain and interchain distances between chain segments (14,15). The segments of interest are Kuhn segments. Formally, a real chain of N residues can be reduced to an effective freely jointed chain of n_K Kuhn segments. For polypeptides, the length of a Kuhn segment can be calculated using $l_0 = 3.8$ Å as the length of a virtual bond between adjacent α-carbon atoms for all-atom simulations and $l_0 = 1$ lattice unit for PIMMS simulations. We set n_K to be five for polyQ and the WT-A1 system as well as variants of this system. This choice is a robust estimate for systems that have the local flexibility of generic polypeptides, which refers to systems in which proline residues make up less than 10% of the sequence (63). For the synthetic systems, we set n_K to be 7, which is identical to the number of residues within each repeat. Systematic calibrations showed that the calculated phase diagrams are insensitive to the choice made for n_K.

The GCT assumes a Gaussian form for $\rho(r)$, the distribution of intersegment distances within a single chain (14,15), which is written as

$$\rho(r) \propto \exp\left\{ -\frac{3r^2}{2\left\langle R_g^2 \right\rangle} \right\}. \tag{1}$$

Here, $\langle R_g^2 \rangle$is the ensemble-averaged value of the mean-square radius of gyration for a single chain. To simplify the algebra, the theory introduces three parameters, viz. the reduced temperature τ, the parameter σ that quantifies the standard deviation of the Gaussian distribution for $\rho(r)$, and the contraction ratio α_s:

$$\tau = \left(\frac{T - T_\theta}{T} \right), \tag{2}$$

$$\sigma^2 = \frac{\left\langle R_g^2 \right\rangle}{n_K l_K^2}, \tag{3}$$

and

$$\alpha_s^2 = 6\sigma^2. \tag{4}$$

Notice that $\tau < 0$ for $T < T_\theta$. Allegra and Ganazzoli (64) showed that the strain ratio for low wavenumbers (k) in Fourier space, which translates to large separations r in real space, can be defined as $\alpha_k = (\pi^2 k^2 / q^2 + \pi^2 k^2)$. Accordingly, the strain ratio in Fourier space and the contraction ratio in real space depend on a single auxiliary parameter q, known as the compression parameter (14,15). Here, q tends to zero as T approaches T_θ, whereas q approaches ∞ in the nonphysical limit at which the chain is compressed down to a single point. Based on the relationship between $\rho(r)$ and its Fourier transform, which is also a Gaussian distribution in k-space, we rewrite $\rho(r)$ and r in terms of q; moments such as $\langle R_g^2 \rangle$ can be evaluated in terms of q. Allegra and co-workers have shown that (64,65)

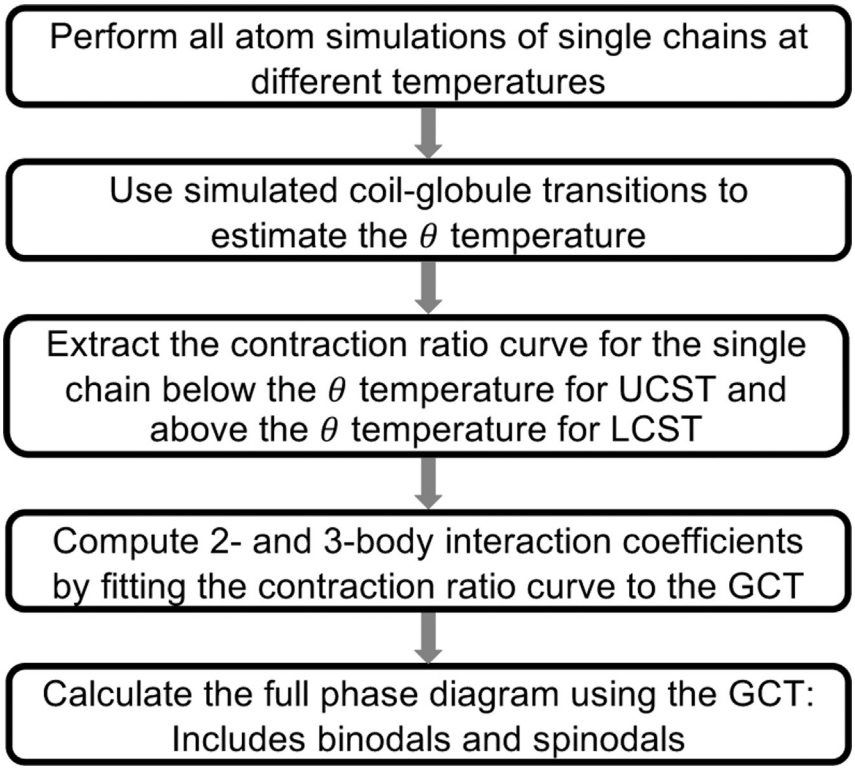

FIGURE 3 Workflow summarizing the overall two-pronged approach that leverages the GCT formalism. We use the prescribed approach to connect results from single-chain simulations of coil-to-globule transitions to calculations of phase diagrams using the GCT.

$$\left\langle R_g^2 \right\rangle = \frac{n_K l_K^2 L(q)}{2q}; \tag{5}$$

here, $L(q)$ is the Langevin function $L(q) = (\coth q - 1/q)$ (33). From Eqs. 3, 4, and 5, it follows that σ, $R_g = \sqrt{\langle R_g^2 \rangle}$, and the contraction ratio α_s are functions of q.

The mean-field free energy $A_1(q;\tau)$ of a single chain at a relative temperature τ is written in terms of the compression parameter q. Given that q is a continuous parameter that quantifies the dependence of chain dimensions on T, we have an analytical route to relate the temperature-dependent chain dimensions to the free energy of the chain. At equilibrium, the free energy of a single chain is minimized by setting $(dA_1/dq) = 0$.The desired free energy $A_1(q;\tau)$ is a sum of the effects of intrachain interactions $A_{\text{intra}}(q;\tau)$ and the elastic term $A_{\text{el}}(q)$ that quantifies contributions of conformational entropy. Accordingly,

$$A_1 = A_{\text{intra}}(q;\tau) + A_{\text{el}}(q). \tag{6}$$

At equilibrium, the free energy (see Appendix) of a single chain at fixed τ is minimized by setting $(dA_1/dq) = 0$. This yields a relationship between τ and the compression parameter q, which allows us to estimate q by solving the transcendental equation below:

$$\tau B\sqrt{n_K} = -\frac{2^{\frac{7}{2}}[\sigma^2(q)]^{\frac{5}{2}}}{3}\left\{\frac{w}{2(3)^{\frac{3}{2}}[\sigma^2(q)]^4} + 3\left[\frac{\coth q}{2} + \frac{q}{2\sinh^2 q} - \frac{1}{q}\right]\left[\frac{1}{q\sinh^2 q} + \frac{\coth q}{q^2} - \frac{2}{q^3}\right]^{-1}\right\}. \tag{7}$$

Extraction of T_θ, B, and w by comparing GCT to results from single-chain simulations

Equation 7 provides the central connection between simulations of coil-to-globule transitions and parameters of the GCT. The protocol for its usage is as follows:

1) Coarse-grained or all-atom simulations are used to extract the system-specific profile for $\langle R_g^2 \rangle$as a function of simulation temperature T. This information is converted into a profile for the contraction ratio α_s vs. T using Eqs. 3 and 4. Additionally, for a given temperature, we extract internal scaling profiles, which plot $\langle\langle R_{ij}\rangle\rangle$ vs. $|j - i|$. Here, $\langle\langle R_{ij}\rangle\rangle$refers to ensemble-averaged mean distance between all pairs of residues i and j that are $|j - i|$ apart along the linear sequence.
2) If the system of interest is a perfect homopolymer or can be reduced to a perfect homopolymer, then simulations of coil-to-globule transitions are performed for a series of chain lengths or numbers of repeats. In this scenario, T_θ is estimated as the intercept along the temperature axis at which plots of $\langle R_g^2 \rangle n_K^{0.5}$ vs. T intersect with one another. This is reflective of the tricritical nature of the θ-point (14,15,66,67). Alternatively, given temperature-dependent profiles for plot $\langle\langle R_{ij}\rangle\rangle$ vs. $|j - i|$, we identify T_θ as the temperature for which$\langle\langle R_{ij}\rangle\rangle = R_0|j - i|^{0.5}$ (68).
3) Next, we fit a smooth curve to the profile of α_s vs. T. From this profile, we estimate T^*, defined as the temperature at which $T^* = T|_{\alpha_s = 0.5(1+\min(\alpha_s^{\text{sim}}))}$. Here, $\min(\alpha_s^{\text{sim}})$ is the minimal value of the contraction ratio from the simulations. For the UCST system, $\min(\alpha_s^{\text{sim}})$ is the contraction ratio at the lowest sampled temperature. For the LCST system, it is the value of the contraction ratio at the highest sampled temperature.
4) From the GCT, it follows that the transition temperature has to satisfy $(T^* - T_\theta)/T^* B n_K^{0.5} = a$. Accordingly, given prior knowledge of T_θ and T^*, we can estimate the value of B. Raos and Allegra have shown that $a \approx -0.35$ for generic homopolymers (14). We use this to set

bounds on the value of a and develop a numerical procedure (Fig. S1) to estimate T^*.

5) The relationship in Eq. 7 is used to rewrite the compression parameter q as a function of τ. This, in conjunction with the definitions introduced in Eqs. 3 and 4, leads to numerical profiles for the contraction ratio as a function of $\tau B n_K^{0.5}$.
6) We then estimate w using a parameter scan in which the constraint is the requirement that the theoretical profile for the contraction ratio versus $\tau B n_K^{0.5}$ be coincident with the profile derived from the single-chain simulations.

In practice, we identify a putative transition region for the location of T^*, choose different values of T^* in this region, and simultaneously choose values for B and w using the bounds based on the constraint prescribed in step 4. We then identify the combination that minimizes the difference between the simulated contraction ratio and the one calculated using Eq. 7 from the GCT. We have developed two distinct approaches for estimating B and w, and these approaches are described in the pseudocode shown in Fig. S1. It helps to have roughly 10 distinct temperatures that span the transition region and the region where the contraction ratio plateaus as the chain forms a globule. A detailed sensitivity analysis for choosing parameters from simulations of coil-to-globule transition profiles is presented in the Supporting Materials and Methods.

GCT for phase separation used to calculate full phase diagrams

The GCT assumes that in a cluster comprising ν distinct polymers, the distribution of distances for polymer i vis-à-vis all other polymers is a Gaussian of the form (14,15)

$$W(\{R_i\}) \propto \exp\left\{ -\eta_\nu \sum_{i<j=1}^{\nu} \left(\mathbf{R}_i - \mathbf{R}_j\right)^2 \right\}. \tag{8}$$

In Eq. 8, $\mathbf{R}_i$ and $\mathbf{R}_j$ denote the centers-of-mass of polymers i and j; the parameter $\eta_\nu = (3/\langle\Delta_\nu^2\rangle\nu)$ is related to the ensemble-averaged mean-square distance between chains, $\langle\Delta_\nu^2\rangle$, in a cluster of ν distinct molecules. As with the single-chain case, the GCT introduces a few key variables. These are defined in terms of $\langle R_{g,\nu}^2\rangle$, the ensemble-averaged mean-square radius of gyration of a single chain within a cluster of ν molecules; $\langle S_\nu^2\rangle$, the ensemble-averaged mean-square radius of gyration of a cluster of ν molecules; and $\alpha_{s,\nu}$, the contraction ratio of a single chain within the cluster. Accordingly,

$$\sigma_\nu = \sqrt{\frac{\left\langle R_{g,\nu}^2\right\rangle}{n_K l_K^2}};\ \Delta_\nu = \sqrt{\frac{\left\langle S_\nu^2\right\rangle}{n_K l_K^2}};\ \alpha_{s,\nu} = \sqrt{\frac{6\left\langle R_{g,\nu}^2\right\rangle}{n_K l_K^2}}. \tag{9}$$

For a given value of τ, the free energy A_ν of a cluster of ν molecules, with $\langle\Delta_\nu^2\rangle$ as the ensemble-averaged mean-square distance between chains in the cluster and q as the value of the compression parameter for a single chain, is the sum of three terms:

$$\begin{aligned} A_\nu\left(q, \left\langle\Delta_\nu^2\right\rangle; \tau\right) &= A_{\text{inter}}\left(q, \left\langle\Delta_\nu^2\right\rangle; \tau\right) + \nu A_{\text{el}}(q) \\ &\quad + A_{\text{conf}}\left(\left\langle\Delta_\nu^2\right\rangle\right). \end{aligned} \tag{10}$$

The contact free energy $A_{\text{inter},\nu}$ combines contributions from two- and three-body contacts among chain segments. Following the derivation shown in the Appendix, for specific values of ϕ and τ, there should be a single value of q that minimizes the system free energy. This is estimated from a generalization of the expression in Eq. 7, now rewritten for a single chain in a polymer solution as

$$\begin{aligned} (\tau B + wF\phi)\sqrt{n_K} &= -\frac{2^{\frac{7}{2}}[\sigma^2(q)]^{\frac{5}{2}}}{3} \\ &\left\{ \frac{w}{2(3)^{\frac{3}{2}}\left[\sigma^2(q)\right]^4} + \right. \\ &\left. 3\left[\frac{\coth q}{2} + \frac{q}{2\sinh^2 q} - \frac{1}{q}\right]\left[\frac{1}{q\sinh^2 q} + \frac{\coth q}{q^2} - \frac{2}{q^3}\right]^{-1} \right\}. \end{aligned} \tag{11}$$

To compute the phase diagram, we need an expression for the chemical potential μ_P of the polymer as a function of values of ϕ. The chemical potential is the partial derivative of a_{vol} with respect to the number of polymer molecules; here, a_{vol} denotes the free energy per unit volume in the polymer solution (see Appendix). The chemical potential takes the form

$$\begin{aligned} \mu_P &= \left(\tau B\sqrt{n_K}F + \frac{wF}{2^{\frac{3}{2}}\sigma^3}\right)\phi\sqrt{n_K} + \frac{12w\left(F\phi\sqrt{n_K}\right)^2}{(3)^{\frac{5}{2}}} \\ &\quad + \ln\left(E\phi\sqrt{n_K}\right) + \frac{\tau B\sqrt{n_K}}{2^{\frac{5}{2}}\sigma^3} + \frac{w}{2(3)^{\frac{5}{2}}\sigma^6} \\ &\quad + \frac{3}{2}\left[\ln\left(\frac{\sinh q}{q}\right) - \frac{qL(q)}{2}\right] + \left\{\frac{3}{2}\ln\frac{3}{\pi} - \frac{3}{2} - \frac{3}{2}\ln n_K\right\}. \end{aligned} \tag{12}$$

For fixed values of τ, B, and w, the chemical potential is a function of the compression parameter q and the volume fraction ϕ. Because there is a one-to-one numerical mapping between q and ϕ, it follows that we can calculate μ_P for different values of ϕ given prior knowledge of the two- and three-body interaction coefficients B and w, respectively.

Calculation of full phase diagrams

The mapping between q and ϕ is extracted numerically from Eq. 11. For this mapping we use values of B and w derived from the analysis of single-chain coil-to-globule transitions. Next, for fixed τ, we calculate the locations of spinodals by computing the values of ϕ that are solutions of the equation $(\partial\mu_P/\partial\phi)_{\tau,B,w} = 0$. This yields two solutions, viz. $\phi_{s1}(\tau)$ and $\phi_{s2}(\tau)$. Next, to generate the coexistence curve, we compute the locations of coexisting dilute and dense phases, viz. $\phi_{b1}(\tau)$ and $\phi_{b2}(\tau)$. For fixed τ, the two coexisting phases represent points for which the chemical potential (μ_P) and osmotic (Π) pressure are equalized along the horizontal tie line that connects the points ϕ_{b1} and ϕ_{b2}. Accordingly, it follows that

$$\begin{aligned} \mu_P\big|_{\phi=\phi_{b1}} &= \mu_P\big|_{\phi=\phi_{b2}}. \\ &\text{and} \\ \Pi\big|_{\phi=\phi_{b1}} &= \Pi\big|_{\phi=\phi_{b2}} \end{aligned} \tag{13}$$

The osmotic pressure is calculated using the expression

$$\frac{\Pi}{k_B T} = \frac{\phi}{n_K n_s v_c} + \frac{F\phi^2 n_K}{2n_s v_c}\left(\tau B\sqrt{n_K} + \frac{w}{2^{\frac{3}{2}}\sigma^3\sqrt{n_K}}\right) + \frac{8wF^2\phi^3}{3^{\frac{5}{2}}n_s v_c}. \tag{14}$$

Here, $v_c = l_0^3\overline{v}_c$ and $\overline{v}_c = \sqrt{w}(2\pi C_\infty/3)^{\frac{3}{2}}$. For fixed τ, B, and w, we use the expressions in Eqs. 12, 13, and 14 and numerically identify the points ϕ_{b1} and ϕ_{b2} that lie on the same tie line. The locus of points $\phi_{b1}(\tau)$ and

$\phi_{b2}(\tau)$ obtained for different values of τ yield the desired binodal in the (ϕ,τ) or (ϕ,T) plane.

RESULTS

We demonstrate the application of the two-pronged approach wherein we obtain system-specific parameters from coarse-grained or all-atom simulations of coil-to-globule transitions of individual chains and incorporate them into the GCT to obtain full phase diagrams. Here, full phase diagrams refer to system-specific binodals and spinodals that are calculated and shown in the two-parameter space with volume fraction ϕ along the abscissa and simulation temperature T along the ordinate. Note that T is unitless for phase diagrams computed using PIMMS because it is written in terms of reduced temperatures used by Martin et al.(10). For phase diagrams computed using all-atom simulations, the temperature is in units of degree-Kelvin.

Phase diagrams for *Aro*$^{+}$, WT-A1, *Aro*$^{-}$, and *Aro*$^{2-}$ variants

First, we extract the theta temperature by analyzing the internal scaling profiles obtained from PIMMS simulations of coil-to-globule transitions for each of the four systems. This analysis is summarized in Fig. S2. Additionally, Fig. S3 shows plots of $R_g n_K^{-0.5}$ as a function of τ for each of the four constructs. The system-specific estimates for T_θ are summarized in each of the panels of Fig. 4. Each panel in Fig. 4 also shows a system-specific profile for the contraction ratio plotted as a function of $\tau B\sqrt{n_K}$. Fig. 5 *A* shows a comparison between binodals obtained using GCT and those obtained directly from PIMMS simulations. Fig. 5 *B* shows a direct comparison between the experimentally measured binodals and those calculated using our two-pronged approach. To facilitate this comparison, we used the approach of Martin et al. (10) and converted the reduced temperature to Kelvin using $T = 6.25T_{PIMMS}$; concentrations in molar units were computed using $c = \rho\phi$, where ϕ is the mole fraction to be converted into molar units and $\rho = 61.869$ mM. The results are shown in Fig. 5 *B*. They indicate that the binodals obtained using GCT, based on the parameters extracted from single-chain simulations, are in good agreement with binodals obtained from direct simulation of phase separation using PIMMS and with experimentally measured binodals. The critical temperature, as well as the width of the two-phase regimes, decreases with decreasing sticker valence, and these trends are accurately captured by the GCT-derived binodals.

Direct simulations of phase separation enable the extraction of sequence- or architecture-specific phase diagrams (10,18,22,28). However, identifying the critical point is challenging because the amplitudes of conformational fluctuations become commensurate with chain dimensions as T approaches T_c. Similarly, as ϕ approaches ϕ_c, the amplitudes of concentration fluctuations become equivalent or even larger than the simulation volume. It is also challenging to

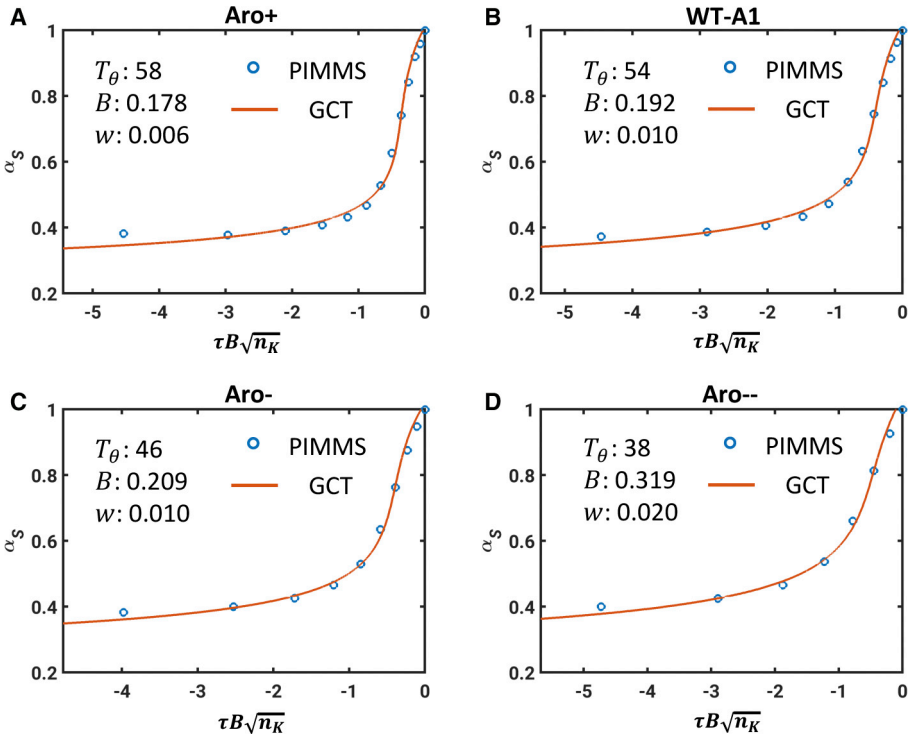

FIGURE 4 Profiles of contraction ratios for *Aro*$^{+}$, WT-A1, *Aro*$^{-}$, and *Aro*$^{2-}$. In each panel, the blue circles are results from PIMMS simulations of single chains, and the red curves are contraction ratios obtained using Eq. 9 that best fit the simulation results. The best-fit, system-specific parameters for T_θ, B, and w are shown as legends in each of the four panels. B has units of inverse volume, and w has units of the square of the inverse volume. To see this figure in color, go online.

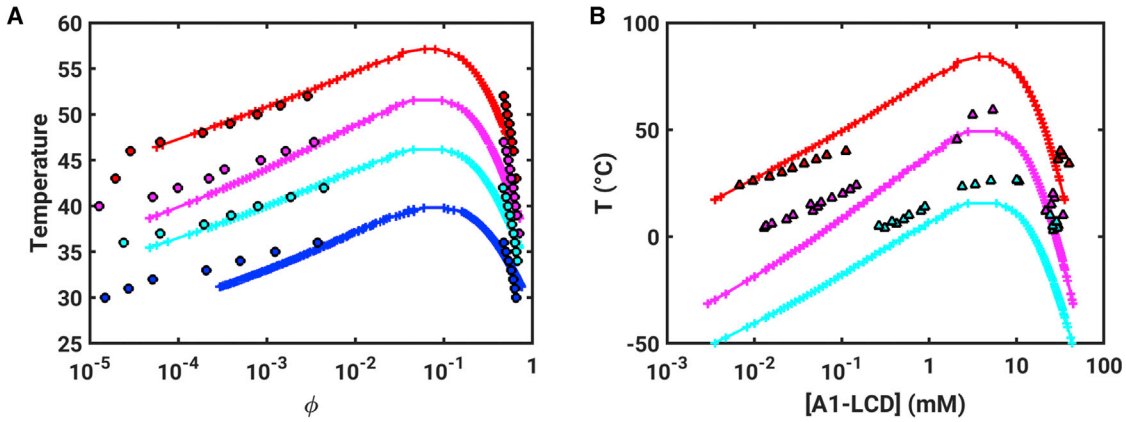

FIGURE 5 Comparison of binodals obtained using direct simulations based on PIMMS and numerical implementation of the GCT formalism that uses parameters extracted from single-chain simulations. (*A*) The PIMMS results are shown as circles, and the results using GCT are shown as + marks. Here, the binodals for the different systems are shown using symbols of different colors: red (Aro^{+}), magenta (WT-A1), cyan (Aro^{-}), and blue ($Aro^{-\ -}$). The binodals obtained using GCT were shifted upward along the ordinate by 10.5 units and along the abscissa by −0.3 units to bring them onto the same scale as the PIMMS-derived binodals. (*B*) A comparison of binodals calculated using GCT to the three binodals measured experimentally by Martin et al. (10). The coexisting concentrations obtained from experiments (in molar units) are shown as triangles (*red* for Aro^{+}, *magenta* for WT-A1, and *cyan* for Aro^{-}). The GCT-derived values are shown using the + symbol, and lines join these symbols as guides to the eye. To see this figure in color, go online.

estimate the low- and high-concentration arms of spinodals because this requires direct observation of the onset of instabilities. These challenges are readily overcome in the GCT-aided calculation of phase diagrams.

Details regarding the width of the metastable region and its dependence on T are direct determinants of the mechanisms of phase separation (16). For a given $T < T_c$, if $\phi_0(T)$ is the initial value of ϕ and $\phi_{b1}(T) < \phi_0(T) < \phi_{s1}(T)$, then the system is in the metastable regime. The rate of nucleation of phase separation from an initial homogeneous one-phase state will be governed by two parameters, viz. $\ln[(\phi_{b2}(T)/\phi_{b1}(T))]$, which is a measure of the width of the two-phase regime, and $\ln[(\phi_0(T)/\phi_{b1}(T))]$, which is a measure of the supersaturation (69). An analogous relationship exists for the initial conditions $\phi_{s2}(T) < \phi_0(T) < \phi_{b2}(T)$, in which the starting conditions are in the high-concentration metastable regime. For simplicity, we will focus here on the low-concentration metastable regime because this is the most relevant scenario for biological systems. If $\phi_{s1}(T) < \phi_0(T) < \phi_{s2}(T)$, then the dynamics of phase separation follow spinodal decomposition (16,70). The dynamics of phase separation, given a prior functional for the free energy of a solution, can be modeled using the Cahn-Hilliard formalism (71). The quench depth relative to the binodal and spinodal directly determines the dynamics of phase separation. Knowledge of the quench depth helps one discern whether the dynamics can be modeled using a Cahn-Hilliard formalism (72,73) or whether advances in chemical-kinetics-approaches-based modernizations of nucleation theories (74,75) can be brought to bear on modeling the dynamics of phase separation.

Calculations of full phase diagrams should aid in the assessment of comparative dynamics for similar quench depths across different systems and for different quench depths of the same system. Fig. 6, *A–D* show the GCT-derived binodals and spinodals for WT-A1. For this analysis, we define $\tau_c = (T - T_c)/T_c$, where T_c is the system-specific critical temperature. The predicted spinodal is manifest only for larger volume fractions ($\phi > 10^{-2}$), which would correspond to initial concentrations that are in the millimolar range for WT-A1. Accordingly, concentrations corresponding to initial conditions in typical in vitro experiments are likely to lie between the binodal and spinodal thereby implicating nucleation as the dominant mechanism for phase separation. Further, the low endogenous concentrations that one expects in vivo are also likely to correspond to the metastable regime. The sizable gap between the binodal and the spinodal that we report for each of the four constructs (Fig. 6) implicates primary, secondary, and possibly heterogeneous nucleation as the dominant mechanisms for phase separation in vivo (55,74–76). The width of the metastable region determines the interplay between nucleation-dependent phase separation and spinodal decomposition (14,15,65). Accordingly, clusters that form in the metastable regime are likely to be key contributors to nucleating the growth of dense phases (15). Indeed, super-resolution methods have made headway in identifying such clusters as signatures of first-order phase transitions in vivo (77).

Fig. S4 shows that the gap between the low arms of the spinodal and binodal, which we can compute as $\Delta\phi_m = [\phi_{s1}(T) - \phi_{b1}(T)]$, varies with $(T_c - T)$ for Aro^{+}, WT-A1, Aro^{-}, and $Aro^{-\ -}$, respectively. This width approaches zero as T approaches T_c. In general, for $(T_c - T) > 2$, $\Delta\phi_m$ is on the order of 10^{-2}, and for a given value of T, $\Delta\phi_m$ is smallest for Aro^{+} and largest for Aro^{2-}. Additionally, $\Delta\phi_m$ varies most steeply with $(T_c - T)$ for $Aro^{-\ -}$, followed by Aro^{-}, WT-A1, and Aro^{+}, respectively. These

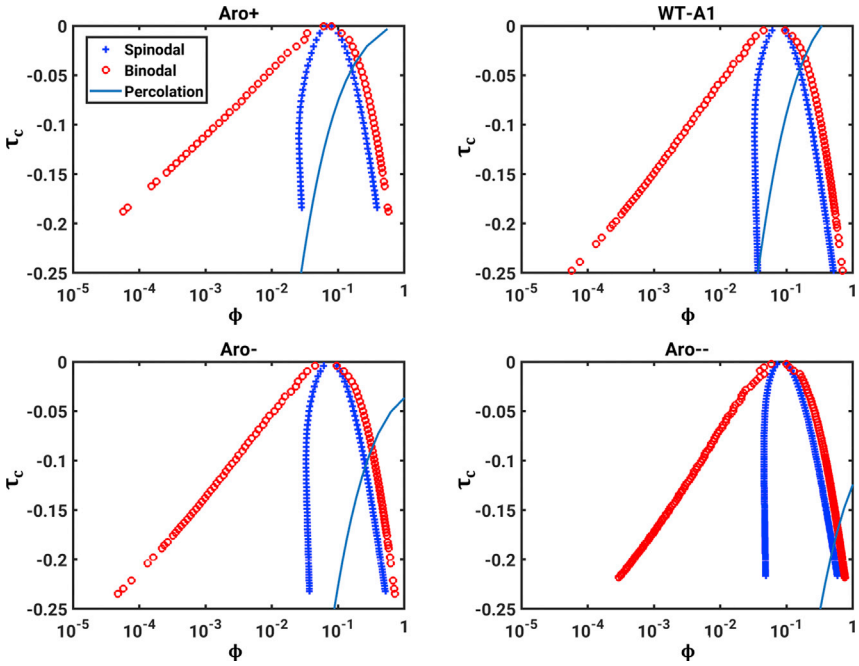

FIGURE 6 Assessments of binodals, spinodals, and percolation lines for the WT-A1 and three designed variants. To see this figure in color, go online.

results indicate that the width of the metastability region decreases and shows a weaker dependence on temperature as the system becomes a stronger driver of phase separation.

Connecting phase diagrams to hardening transitions of LCDs

Condensates formed by LCDs undergo "hardening transitions" as a function of time (6,21,78,79). This refers to the development of irregular morphologies, drastic slowing of the exchange of molecules between condensates and their surroundings, and increased inability of condensates to fuse and grow. Heteropolymers with stickers-and-spacers architectures also undergo a networking transition known as "percolation" that is enabled by noncovalent (physical) cross-links among stickers. Percolation is essentially gelation, and it occurs above a system-specific concentration known as the percolation threshold that is designated as ϕ_{perc} (22). Above ϕ_{perc}, the collection of multivalent macromolecules in solution form a system-spanning network; if $\phi_{sat} < \phi_{perc} < \phi_{dense}$, then multivalent macromolecules will undergo "phase-separation-aided percolation" (25). Systems that conform to stickers-and-spacers architectures undergo phase-separation-aided percolation. Accordingly, dense phases are condensate-spanning networks characterized by intracondensate cross-links among stickers that are aided by features of spacers.

In the physical literature, hardening represents the formation of dynamically arrested phases that arise because of the coupling between phase separation and percolation (80). At a given temperature, the likelihood that a condensate will undergo a hardening transition will be governed by the magnitude of the gap between the percolation line and the high-concentration arm of the binodal, which we quantify in terms of the difference between ϕ_{dense} and ϕ_{perc} and designate as ϕ_{gap}. Large values of ϕ_{gap} lead to increased chances of forming randomly networked assemblies because of physical cross-links among stickers of polymers that overlap with one another in dense phases. Large values of ϕ_{gap} also imply that the cross-links are significantly stronger than k_BT, and this leads to long-lived cross-links that cause dynamical arrest and irregular morphologies. These features were illustrated in recent results obtained by Boeynaems et al. (35), who demonstrated dynamically arrested phase separation in a model protein-RNA system (35). Irregular, void-rich condensates that do not rearrange rapidly or fuse with one another are signatures of dynamical arrest. The timescales associated with the making and breaking of physical cross-links are governed by the free energy barrier for bond dissociation. This and the extent of cross-linking, both of which contribute to ϕ_{gap}, are the causes of dynamical arrest.

Building on the calculations of sequence-specific binodals and spinodals we also estimated the percolation line, i.e., ϕ_{perc} as a function of temperature. For this, we used a recent generalization (23,24) of the mean-field theory developed by Semenov and Rubinstein (27). Specifically, we set $\phi_{perc}(T; \tau) = [\lambda(T;\tau)n_a^2]^{-1}$, where n_a is the number of aromatic stickers and $\lambda(T; \tau) = -B\tau$ is a simple estimate for the temperature dependence of the intersticker interaction strengths. The A1-LCD has 19 aromatic stickers. In the designed variants, the valence of aromatic stickers is either increased (to 24 in Aro^+) or decreased (to 13 in Aro^- and 6 in $Aro^{-\ -}$) (10). The calculated, system-specific

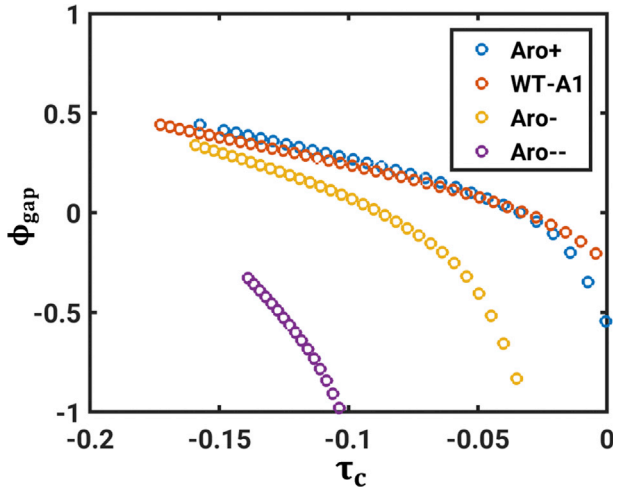

FIGURE 7 ϕ_{gap} vs. τ_c for each of WT-A1 LCD and the three designed variants. To see this figure in color, go online.

percolation lines are overlaid on the binodals for the WT A1-LCD and three designed variants as shown in each of the panels in Fig. 6.

We plotted ϕ_{gap} vs. τ_c to compare the likelihoods of hardening transitions across different systems (Fig. 7). This enables comparisons across different sequences with different critical points. Note that only large positive ϕ_{gap}-values are relevant for the dynamical arrest of condensates. We find that ϕ_{gap} decreases as τ_c approaches zero, which happens as the temperature approaches the critical point (Fig. 7); for fixed τ_c, the value of ϕ_{gap} is larger for Aro^{+}, followed by WT, Aro^{-}, and $Aro^{-\ -}$, respectively. This implies that stronger driving forces for phase separation, which in this case are governed by the increased valence of stickers, also lead to larger gaps between the percolation lines and the high-concentration arms of binodals. Therefore, at identical values of τ_c, it follows that sequences encoding stronger driving forces for phase separation will also have larger values of ϕ_{gap}, and such sequences are likely to be more susceptible to hardening transitions.

Phase diagrams for polyQ and synthetic IDPs

The results shown in Fig. 5 illustrate the accuracy of the GCT-derived binodals as compared to binodals obtained using PIMMS for four closely related systems. Next, we computed full phase diagrams using parameters extracted from all-atom simulations that were based on the ABSINTH implicit solvation model. We present results for three separate systems, viz. polyQ and two synthetic IDPs, one with UCST behavior and another showing LCST behavior. The results for the polyQ system are shown in Fig. 8.

Crick et al. previously measured the temperature dependence of the solubility limit for different constructs with polyQ tracts (37). The measured saturation concentrations, which are in the low micromolar range, decrease to be in the submicromolar range for longer polyQ molecules. These measured concentrations translate to volume fractions that are $\sim 10^{-5}$ and are what we obtain for the low arms of binodals (Fig. 8). Further, Crick et al. showed that polyQ has UCST behavior with a critical temperature that is above the boiling point of water—a feature that is recapitulated by the simulations and the GCT-aided calculations. Simulation results, summarized in Fig. 8 *A*, were used to identify the theta temperature for this system. Here, $T_\theta \approx 405$ K. Fig. 8 *B* shows the full phase diagrams, including binodals and spinodals for phase separation. These data are shown for the Q_{70} system. Fig. 8, *C* and *D* show that the critical temperature (T_c) increases with increasing polyQ length, whereas the critical volume fraction (ϕ_c) decreases with increasing polyQ length. This is in accord with general trends expected from mean-field theories for the phase separation of homopolymers (16).

Finally, Fig. 9 summarizes results obtained for the two synthetic IDPs with UCST and LCST behavior. Recent attention has focused on IDPs derived from resilin- and elastin-like polypeptides that are designed to have UCST versus LCST behavior, respectively (31,81). We show results for two archetypal systems. The θ temperatures for these systems were estimated by analyzing the temperature-dependent internal scaling profiles. Fig. 9, *A–C*, respectively, show the profile of the normalized R_g plotted as a function of simulation temperature, the contraction ratio as a function of $\tau B\sqrt{n_K}$, and the resultant phase diagram for the $(QGQSPYG)_9$ system that shows UCST behavior. Fig. 9, *D–F* show the corresponding results for the $(TPKAMAP)_9$ system that has LCST phase behavior. Systems that are perfect repeats of specific peptide sequences may be thought of as homopolymers on length scales that are longer than one repeat. Accordingly, these systems are amenable to a joint computational approach that combines simulations of single chains with GCT to extract computed phase diagrams. The results demonstrate that the GCT is applicable for calculating phase diagrams of systems with LCST behavior. All that is required is that simulations at the single-chain level demonstrate a collapse transition at higher temperatures, which we observe for the $(TPKAMAP)_9$ system if we use temperature-dependent reference free energies of solvation as prescribed by Wuttke et al. (59); see Fig. 8 *D*.

DISCUSSION

In this work, we have adapted and deployed the GCT of Raos and Allegra (14,15) to compute full phase diagrams for archetypal LCDs using parameters extracted from single-chain simulations of coil-to-globule transitions. Our motivation for adapting and deploying the GCT-aided approach comes from recent observations showing that mean-field theories based on the Flory-Huggins theory (82,83) can be reliably used to describe experimentally

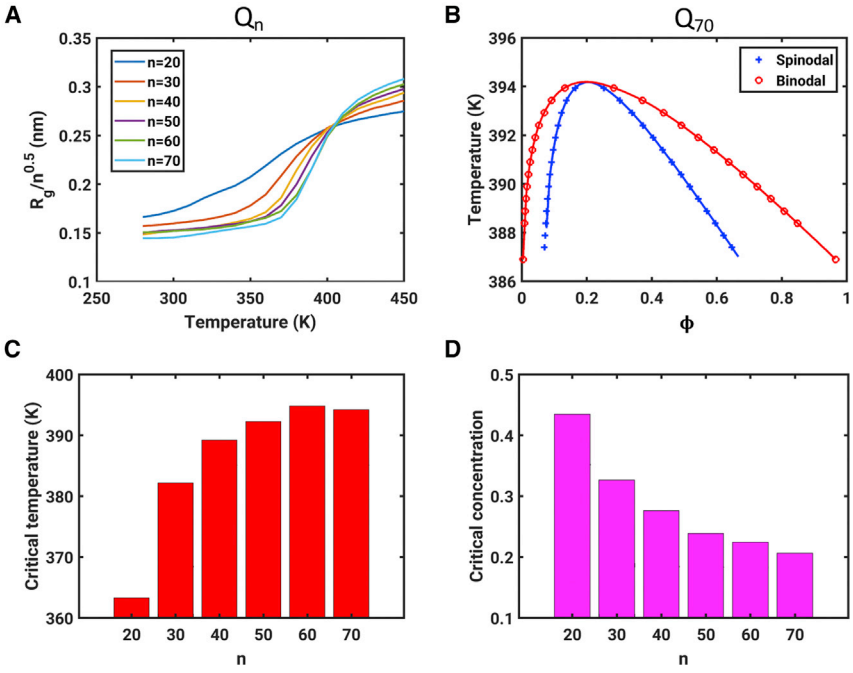

FIGURE 8 Results for the polyQ system. (*A*) Normalized R_g for different chain lengths as a function of simulation temperature *T*. The θ temperature ($T_\theta \approx 405$ K) corresponds to the point at which the curves coincide with one another. Note that previous simulations based on the ABSINTH model estimated a value of 390 K for the T_θ of polyQ (44,60). The discrepancy between published and current estimates comes from our current usage of a temperature-dependent functional form for the macroscopic dielectric constant of water (59). (*B*) Full phase diagrams for the Q_{70} system. The binodal and spinodal coincide at the critical point. The smooth red and blue curves join the dots along the binodal and spinodal, respectively, and they are used to guide the eye. (*C*) A bar plot showing the increase of the critical temperature (T_c) with polyQ length. (*D*) A bar plot showing the decrease of the critical volume fraction (ϕ_c) with increasing polyQ length. To see this figure in color, go online.

measured binodals for LCDs (3,8–10). Bioinformatics analyses have identified a uniform linear distribution of stickers as a defining feature of LCDs that are implicated in liquid-liquid phase separation (10,84,85). This suggests that, to zeroth order, LCDs that are drivers of liquid-liquid phase separation can be treated as effective homopolymers. Additionally, recent theoretical, computational, and experimental work has shown that there is a one-to-one correspondence between the driving forces for chain collapse in dilute solutions and phase separation in concentrated solutions of LCDs (10,18,28,68,86,87). The theoretical work of Lin and Chan, based on a random phase-approximation-based mean-field approach to incorporate the effects of sequence patterning and cation-π as well as π-π interactions, showed an inverse correlation between the extent of collapse in polyampholytes and the critical temperature (86). Overall, the observations to date are suggestive of the swapping of intrachain interactions for interchain interactions as being a key determinant of phase separation (17). Therefore, we reasoned that a theory designed to connect single-chain coil-to-globule transitions to phase separation should be suitable for numerical adaptation. We selected the GCT for our two-pronged approach because, to the best of our knowledge, this is one of the few theories that describes both single-chain conformational equilibria and multichain phase equilibria using a single set of parameters. A detailed analysis of the robustness of the parameterization procedure is presented in the Supporting Materials and Methods (see Figs. S5 and S6).

Use of the GCT approach requires a priori numerical characterization of single-chain coil-to-globule transition profiles. These can be extracted from all-atom or coarse-grained simulations, which are accessible via different approaches. For example, a recent study (88), which draws on our presentation at the 64th annual meeting of the Biophysical Society (89), demonstrated the use of the Raos-Allegra theory for computing phase diagrams for the protein FUS using data for coil-to-globule transitions obtained using the transferrable coarse-grained model of Dignon et al. (18,28,90). Computational methods based on the ABSINTH implicit solvation model and novel refinements to explicit solvent models have enabled increasingly accurate and efficient simulations of conformational equilibria of a range of IDPs (39,40,59,63,91–107). Many of these approaches can use experimental data as inputs to generate accurate coarse-grained descriptions for conformational ensembles of IDPs (108,109). These methods can be brought to bear for accurate characterization of coil-to-globule transitions. Here, we focused on temperature as the control parameter; however, one can conceive of other intensive parameters such as pH (110) or the chemical potentials of solution ions or small molecules as control parameters (8,11,20,111).

Our method of extracting parameters from single-chain coil-to-globule transitions and computing phase diagrams using a numerical approach based on the GCT is similar in spirit to the field-theoretic approaches that have been adapted from the polymer literature for the study of LCD phase separation (112,113). The essential differences are that the GCT, unlike field-theoretic approaches, remains primarily a mean-field theory. This implies that effects of fluctuations are ignored in the GCT formalism. This is likely to create issues in describing how the width of the two-phase regime varies as the critical temperature is approached (114). However, as a rapid and potentially easy to deploy

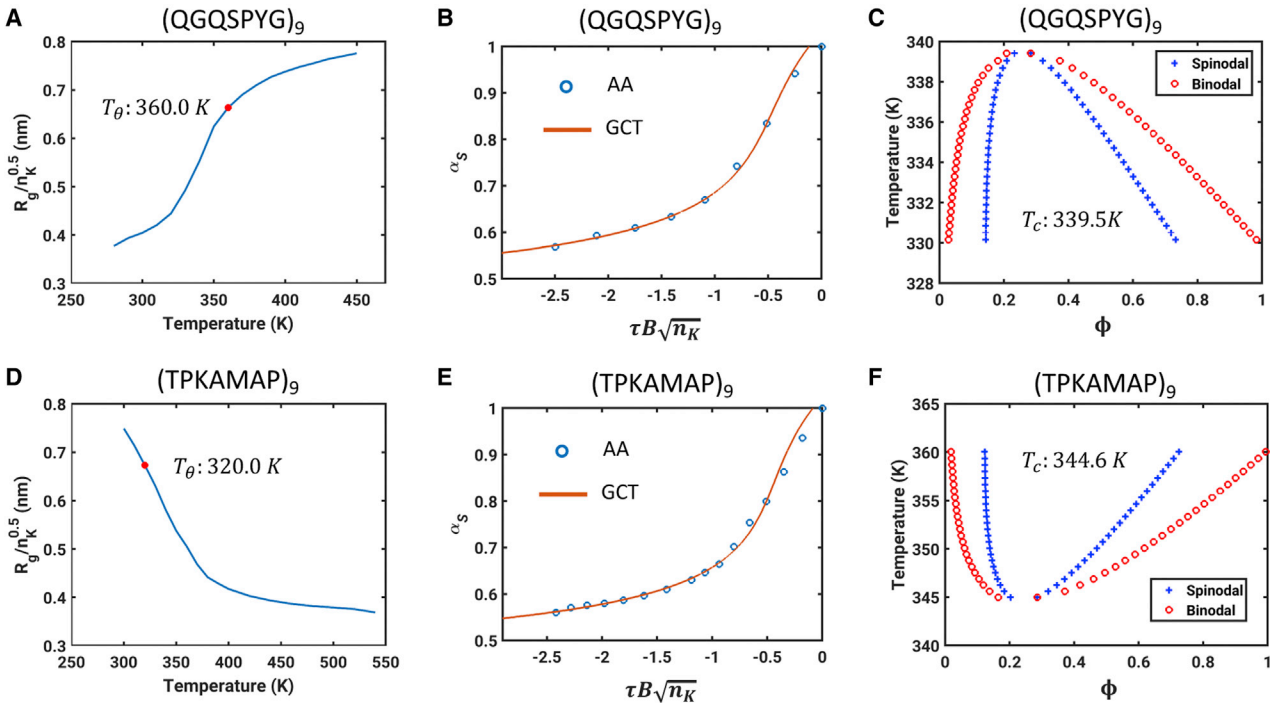

FIGURE 9 Results for synthetic systems that show UCST behavior (*top row*) and LCST behavior (*bottom row*). (*A*) Normalized R_g plotted as a function of simulation temperature for $(QGQSPYG)_9$. The intercept along the abscissa from the red circle corresponds to the location of the theta temperature. (*B*) Contraction ratio plotted against $\tau B\sqrt{n_K}$ for $(QGQSPYG)_9$. (*C*) Binodal and spinodal for $(QGQSPYG)_9$. The critical point is at which the binodal and spinodal coincide and the estimated UCST is $T_c \approx 340$ K. The symbols represent the temperatures for which the points on the binodal and spinodal were obtained from the GCT theory. (*D*) Plot that is equivalent to (*A*) for the system $(TPKAMAP)_9$ that has LCST behavior. This system undergoes a robust collapse transition above the θ temperature, which is the intercept along the abscissa that passes through the red circle. (*E*) Plot of the contraction ratio for $(TPKAMAP)_9$ as a function of $\tau B\sqrt{n_K}$. Note that B is positive for UCST systems and negative for LCST systems. Accordingly, $\tau B\sqrt{n_K}$ is always negative. (*F*) Full phase diagram, including the binodal and spinodal for $(TPKAMAP)_9$. The LCST (~345 K) corresponds to the temperature at which the binodal and spinodal coincide. To see this figure in color, go online.

strategy, the GCT should afford a useful approach for high-throughput comparative assessments of phase diagrams, providing we have results from simulations of coil-to-globule transitions. This GCT-driven approach is likely to be useful for the design of repetitive sequences with bespoke phase behavior—a problem of considerable interest in the design of novel biomaterials as well as synthetic condensates. It can be combined with field-theoretic approaches to obtain improved descriptions of near critical regions and of the effects of density fluctuations engendered by conformational fluctuations.

Efficiency considerations

There are three major categories of methods that enable the numerical calculation of sequence- and/or architecture-specific phase diagrams of intrinsically disordered regions (IDRs) and chimeras of IDRs and folded domains. Lattice-based methods were introduced by Harmon et al. (25,115), to demonstrate the relevance of the stickers-and-spacers model for chimeric systems. The underlying physics of bond fluctuation models used by Harmon et al. was formalized in the LASSI engine for modeling phase separation and percolation of IDRs and chimeric systems comprising IDRs tethered to folded domains (22). An analogous lattice-based approach, known as PIMMS, that mainly uses a single bead per amino acid resolution (although this is not an absolute constraint) has been deployed for simulations of phase behaviors of IDRs in different contexts (10,35,116). LASSI and PIMMS are analogous engines, although the latter is intended to enable "on the fly" discovery of stickers and spacers because it can be deployed with a transferrable or a learned model. In contrast, LASSI requires an a priori delineation of stickers versus spacers. LASSI, however, affords the incorporation of anisotropic interactions, and it can, in theory, be generalized to incorporate stickers with higher coordination numbers.

Off-lattice models include those developed by Dignon et al. (18,28,90). Their approach and code base are gaining traction and usage (88,117) because they are built on a transferrable model for the interactions between coarse-grained beads that mimic individual amino acids or groups of residues. An alternative to the transferrable model of Dignon et al. is the machine-learning-aided off-lattice approach developed by Ruff et al. (41) that derives system-specific

parameters by bootstrapping against all-atom simulations for individual molecules and pairs of molecules. Finally, the field-theoretic strategy pioneered by Fredrickson and co-workers (118) has gained usage for modeling phase behavior of polyelectrolytic and polyampholytic IDPs (112,113). The field-theoretic approach requires numerical solution of dynamical equations that incorporate free energy functionals derived from simulations of single chains.

In our two-pronged approach that leverages the GCT, the main bottleneck remains the simulation of single-chain coil-to-globule transitions. Once we have this information, all that remains is calculation of the chemical potential and osmotic pressure system for each combination of ϕ and T. The total time for the calculation of the phase diagram is therefore linearly related to the number of ϕ and T pairs of that are of interest. A typical calculation, such as the one that yields the phase diagrams shown in Fig. 6 for the WT-A1 system, takes an average of 58 s on a single 4 Intel i5-7600 processor. The code used in this study is provided as Data S1.

Summary and prognosis

Our use of the GCT rests on the validity of the assumption that the driving forces for single-chain collapse in dilute solutions are the same as the ones that drive interchain associations in concentrated solutions. We refer to this as the "strong coupling regime" and note that the correspondence required for the use of GCT need not always prevail (20,114). For example, a direct prediction of the GCT, which we have not called out here explicitly, is the fact that the chain conformations in the dense phase are likely to be distinct from those in the coexisting dilute phase. Well below the θ temperature, chains in dilute solutions should form globules, whereas chains in the dense phase are likely to behave like random walks because intermolecular interactions weaken the driving force for collapse (15). Therefore, systems characterized by limited conformational change across the phase boundary are poor candidates for the use of GCT. This will be true as we approach the critical temperature at which the two-phase regime narrows, giving rise to ultradilute droplet phases, which has been observed in recent studies (20).

Predictions from the GCT can be tested by explicit measurements of binodals and by characterizing the dimensions of individual chains in both the dilute and dense phases as a function of temperature or any other control parameter of interest. This should be feasible using advances in small angle x-ray scattering as well as multiparameter single-molecule fluorescence spectroscopies (102,119,120). A problem of considerable interest would be the use of GCT to analyze published results for low-concentration arms of binodals and derive coil-to-globule transition profiles in the dilute solutions and reconstruct full phase diagrams using parameters from the inferred coil-to-globule transitions. This follow-up work is currently underway.

APPENDIX: MATHEMATICAL DETAILS THAT LEAD TO EQS. 7 AND 11

From the assumption of a Gaussian distribution for intrasegment distances within a chain, it follows that $A_{\text{intra}}(q;\tau)$ is written in terms of q, τ, the two-body interaction coefficient B (where $B > 0$), and the three-body interaction coefficient w as

$$A_{\text{intra}}(q;\tau) = \frac{\tau B\sqrt{n_{\text{K}}}}{2^{\frac{5}{2}}\sigma^3} + \frac{w}{2(3)^{\frac{5}{2}}\sigma^6}. \tag{15}$$

The parameter τB is the second virial coefficient that reflects the interplay of pairwise segment-segment, segment-solvent, and solvent-solvent interactions. Note that for UCST systems, τB is negative for $T < T_\theta$ because B here is positive; conversely, for LCST systems, B is negative and the second virial coefficient τB is negative for $T > T_\theta$. The parameter w is a direct measure of the third virial coefficient defined by three-way interactions of polymer segments (Fig. 2).

The elastic free energy is written in terms of the compression parameter q as

$$A_{\text{el}} = \frac{3}{2}\left[\ln\left(\frac{\sinh q}{q}\right) - \frac{qL(q)}{2}\right]. \tag{16}$$

These expressions lead to Eq. 7, which is the lynchpin of the theory for extracting parameters from simulated coil-to-globule transition profiles for single chains.

FREE ENERGY FOR A SYSTEM WITH MULTIPLE CHAIN MOLECULES

Using the assumption of a Gaussian distribution for $W(R_i)$, it follows that

$$A_{\text{inter},\nu} = A_{2,\nu} + A_{3,\nu}, \tag{17}$$

where

$$A_{2,\nu} = \frac{\nu\tau B\sqrt{n_{\text{K}}}}{2^{\frac{5}{2}}\sigma_\nu^3}\left[1 + (\nu-1)\left(1 + \frac{\Delta_\nu^2}{2n_{\text{K}}l_{\text{K}}^2\sigma_\nu^2}\right)^{-\frac{3}{2}}\right] \tag{18}$$

and

$$A_{3,\nu} = \nu\frac{w}{2(3)^{\frac{5}{2}}\sigma_\nu^6}\left[1 + 3(\nu-1)\left(1 + \frac{\Delta_\nu^2}{2n_{\text{K}}l_{\text{K}}^2\sigma_\nu^2}\right)^{-\frac{3}{2}} + (\nu-1)(\nu-2)\left(1 + \frac{2\Delta_\nu^2}{2n_{\text{K}}l_{\text{K}}^2\sigma_\nu^2}\right)^{-3}\right]. \tag{19}$$

νA_{el} is the elastic free energy of individual chains in the cluster, whereas A_{conf} is the free energy cost associated with confining ν molecules into a cluster. The confinement term accounts for the loss of translational entropy. It also includes a combinatorial Gibbs-like correction term that accounts for

the indistinguishability of chains within a cluster. The expression for A_{conf} derived by Raos and Allegra is as follows:

$$A_{\text{conf}}\left(\left\langle \Delta_{\nu}^{2}\right\rangle\right) = \frac{3}{2}(\nu-2)\ln\nu + \frac{3}{2}(\nu-1)\left[\ln\left(\frac{3l_0^2}{\pi\nu\Delta_{\nu}^{2}}\right)-1\right] + \ln\nu!. \tag{20}$$

Phase separation is an infinitely cooperative process (121), and accordingly, the cluster free energy is minimized in the limit $\nu \to \infty$. In this limit, it is useful to work in terms of the volume fraction ϕ. Raos and Allegra show that the free energy per chain in a polymer solution a_{ch} is written in the limit $\nu \to \infty$ as

$$a_{\text{ch}} = \lim_{\nu\to\infty}\frac{A_{\nu}}{\nu} = (\tau B + wF\phi)\frac{\sqrt{n_{\text{K}}}}{2^{\frac{5}{2}}\sigma^{3}} + \frac{w}{2(3)^{\frac{5}{2}}\sigma^{6}} + \frac{\tau B n_{\text{K}} F\phi}{2} + \frac{4wn_{\text{K}}(F\phi)^{2}}{(3)^{\frac{5}{2}}} + \ln\left(\frac{E\phi}{N}\right) + \frac{3}{2}\left[\ln\left(\frac{\sinh q}{q}\right) - \frac{qL(q)}{2}\right] + \left\{\frac{3}{2}\ln\left(\frac{3}{\pi}\right) - \frac{5}{2}\right\}. \tag{21}$$

Here, $E = \left(4\pi/3(2)^{\frac{3}{2}}n_s\sqrt{w}\right)(2\pi C_{\infty}/3)^{-\frac{3}{2}}$ and $F = E(n_5 C_{\infty})^{\frac{3}{2}}$. The expression in Eq. 17 is then used to estimate the free energy per unit volume in the polymer solution denoted as a_{vol}. Here, $a_{\text{vol}} = a_{\text{ch}}(\phi/(N-1)l_0^3\sqrt{w})(3/2\pi C_{\infty})^{\frac{3}{2}}$.

SUPPORTING MATERIAL

Supporting Material can be found online at https://doi.org/10.1016/j.bpj.2020.06.014.

AUTHOR CONTRIBUTIONS

Conceptualization, X.Z. and R.V.P.; all-atom simulations and GCT implementation, X.Z.; PIMMS simulations, A.S.H.; Analysis, X.Z. and R.V.P.; writing, R.V.P.; editing, X.Z., A.S.H., A.C., T.M., and R.V.P.; funding, A.C., T.M., and R.V.P.

ACKNOWLEDGMENTS

We thank Mina Farag for helpful discussions and critical inputs.

This work was supported by grants DMR 1729783 from the US National Science Foundation (to A.C. and R.V.P.), RGP0034/2017 from the Human Frontier Science Program (to R.V.P.), 5R01NS056114 from the US National Institutes of Health (to R.V.P.), and the St. Jude Children's Research Collaborative on Membraneless Organelles (to T.M. and R.V.P.).

REFERENCES

1. Banani, S. F., H. O. Lee, …, M. K. Rosen. 2017. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* 18:285–298.
2. Brangwynne, C. P., P. Tompa, and R. V. Pappu. 2015. Polymer physics of intracellular phase transitions. *Nat. Phys.* 11:899–904.
3. Shin, Y., J. Berry, …, C. P. Brangwynne. 2017. Spatiotemporal control of intracellular phase transitions using light-activated optoDroplets. *Cell.* 168:159–171.e14.
4. Martin, E. W., and T. Mittag. 2018. Relationship of sequence and phase separation in protein low-complexity regions. *Biochemistry.* 57:2478–2487.
5. Shin, Y., and C. P. Brangwynne. 2017. Liquid phase condensation in cell physiology and disease. *Science.* 357:eaaf4382.
6. Alberti, S., S. Saha, …, A. A. Hyman. 2018. A user's guide for phase separation assays with purified proteins. *J. Mol. Biol.* 430:4806–4820.
7. Posey, A. E., A. S. Holehouse, and R. V. Pappu. 2018. Phase separation of intrinsically disordered proteins. *Methods Enzymol.* 611:1–30.
8. Brady, J. P., P. J. Farber, …, L. E. Kay. 2017. Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proc. Natl. Acad. Sci. USA.* 114:E8194–E8203.
9. Simon, J. R., N. J. Carroll, …, G. P. López. 2017. Programming molecular self-assembly of intrinsically disordered proteins containing sequences of low complexity. *Nat. Chem.* 9:509–515.
10. Martin, E. W., A. S. Holehouse, …, T. Mittag. 2020. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science.* 367:694–699.
11. Ruff, K. M., S. Roberts, …, R. V. Pappu. 2018. Advances in understanding stimulus-responsive phase behavior of intrinsically disordered protein polymers. *J. Mol. Biol.* 430:4619–4635.
12. Faltova, L., A. M. Küffner, …, P. Arosio. 2018. Multifunctional protein materials and microreactors using low complexity domains as molecular adhesives. *ACS Nano.* 12:9991–9999.
13. Linsenmeier, M., M. R. G. Kopp, …, P. Arosio. 2019. Dynamics of synthetic membraneless organelles in microfluidic droplets. *Angew. Chem. Int. Ed. Engl.* 58:14489–14494.
14. Raos, G., and G. Allegra. 1996. Chain collapse and phase separation in poor-solvent polymer solutions: a unified molecular description. *J. Chem. Phys.* 104:1626–1645.
15. Raos, G., and G. Allegra. 1997. Macromolecular clusters in poor-solvent polymer solutions. *J. Chem. Phys.* 107:6479–6490.
16. Rubinstein, M., and R. H. Colby. 2003. Polymer Physics. Oxford University Press, New York.
17. Panagiotopoulos, A. Z., V. Wong, and M. A. Floriano. 1998. Phase equilibria of lattice polymers from histogram reweighting Monte Carlo simulations. *Macromolecules.* 31:912–918.
18. Dignon, G. L., W. Zheng, …, J. Mittal. 2018. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. USA.* 115:9929–9934.
19. Pappu, R. V., X. Wang, …, S. L. Crick. 2008. A polymer physics perspective on driving forces and mechanisms for protein aggregation. *Arch. Biochem. Biophys.* 469:132–141.
20. Wei, M. T., S. Elbaum-Garfinkle, …, C. P. Brangwynne. 2017. Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. *Nat. Chem.* 9:1118–1125.
21. Wang, J., J.-M. Choi, …, A. A. Hyman. 2018. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell.* 174:688–699.e16.
22. Choi, J.-M., F. Dar, and R. V. Pappu. 2019. LASSI: a lattice model for simulating phase transitions of multivalent proteins. *PLoS Comput. Biol.* 15:e1007028.
23. Choi, J.-M., A. S. Holehouse, and R. V. Pappu. 2020. Physical principles underlying the complex biology of intracellular phase transitions. *Annu. Rev. Biophys.* 49:107–133.
24. Choi, J.-M., A. A. Hyman, and R. V. Pappu. 2020. Generalized models for bond percolation transitions of associative polymers. *arXiv*, arXiv:2004.03278.
25. Harmon, T. S., A. S. Holehouse, …, R. V. Pappu. 2017. Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *eLife.* 6:e30294.

26. Rubinstein, M., and A. V. Dobrynin. 1997. Solutions of associative polymers. *Trends Polym. Sci. (Regul. Ed.).* 5:181–186.
27. Semenov, A. N., and M. Rubinstein. 1998. Thermoreversible gelation in solutions of associative polymers. 1. Statics. *Macromolecules.* 31:1373–1385.
28. Dignon, G. L., W. Zheng, …, J. Mittal. 2018. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* 14:e1005941.
29. Dzuricky, M., S. Roberts, and A. Chilkoti. 2018. Convergence of artificial protein polymers and intrinsically disordered proteins. *Biochemistry.* 57:2405–2414.
30. Roberts, S., T. S. Harmon, …, A. Chilkoti. 2018. Injectable tissue integrating networks from recombinant polypeptides with tunable order. *Nat. Mater.* 17:1154–1163.
31. Quiroz, F. G., and A. Chilkoti. 2015. Sequence heuristics to encode phase behaviour in intrinsically disordered protein polymers. *Nat. Mater.* 14:1164–1171.
32. Pierce, W. K., C. R. Grace, …, T. Mittag. 2016. Multiple weak linear motifs enhance recruitment and processivity in SPOP-mediated substrate ubiquitination. *J. Mol. Biol.* 428:1256–1271.
33. Flory, P. J., and M. Volkenstein. 1969. Statistical mechanics of chain molecules. *Biopolymers.* 8:699–700.
34. Molliex, A., J. Temirov, …, J. P. Taylor. 2015. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell.* 163:123–133.
35. Boeynaems, S., A. S. Holehouse, …, A. D. Gitler. 2019. Spontaneous driving forces give rise to protein-RNA condensates with coexisting phases and complex material properties. *Proc. Natl. Acad. Sci. USA.* 116:7889–7898.
36. Walker, F. O. 2007. Huntington's disease. *Semin. Neurol.* 27:143–150.
37. Crick, S. L., K. M. Ruff, …, R. V. Pappu. 2013. Unmasking the roles of N- and C-terminal flanking sequences from exon 1 of huntingtin as modulators of polyglutamine aggregation. *Proc. Natl. Acad. Sci. USA.* 110:20075–20080.
38. Posey, A. E., K. M. Ruff, …, R. V. Pappu. 2018. Profilin reduces aggregation and phase separation of huntingtin N-terminal fragments by preferentially binding to soluble monomers and oligomers. *J. Biol. Chem.* 293:3734–3746.
39. Newcombe, E. A., K. M. Ruff, …, D. M. Hatters. 2018. Tadpole-like conformations of huntingtin exon 1 are characterized by conformational heterogeneity that persists regardless of polyglutamine length. *J. Mol. Biol.* 430:1442–1458.
40. Warner, J. B., IV, K. M. Ruff, …, H. A. Lashuel. 2017. Monomeric huntingtin exon 1 has similar overall structural features for wild-type and pathological polyglutamine lengths. *J. Am. Chem. Soc.* 139:14456–14469.
41. Ruff, K. M., T. S. Harmon, and R. V. Pappu. 2015. CAMELOT: a machine learning approach for coarse-grained simulations of aggregation of block-copolymeric protein sequences. *J. Chem. Phys.* 143:243123.
42. Ruff, K. M., S. J. Khan, and R. V. Pappu. 2014. A coarse-grained model for polyglutamine aggregation modulated by amphipathic flanking sequences. *Biophys. J.* 107:1226–1235.
43. Williamson, T. E., A. Vitalis, …, R. V. Pappu. 2010. Modulation of polyglutamine conformations and dimer formation by the N-terminus of huntingtin. *J. Mol. Biol.* 396:1295–1309.
44. Vitalis, A., N. Lyle, and R. V. Pappu. 2009. Thermodynamics of beta-sheet formation in polyglutamine. *Biophys. J.* 97:303–311.
45. Vitalis, A., X. Wang, and R. V. Pappu. 2008. Atomistic simulations of the effects of polyglutamine chain length and solvent quality on conformational equilibria and spontaneous homodimerization. *J. Mol. Biol.* 384:279–297.
46. Wetzel, R. 2012. Physical chemistry of polyglutamine: intriguing tales of a monotonous sequence. *J. Mol. Biol.* 421:466–490.
47. Kar, K., M. Jayaraman, …, R. Wetzel. 2011. Critical nucleus size for disease-related polyglutamine aggregation is repeat-length dependent. *Nat. Struct. Mol. Biol.* 18:328–336.
48. Thakur, A. K., M. Jayaraman, …, R. Wetzel. 2009. Polyglutamine disruption of the huntingtin exon 1 N terminus triggers a complex aggregation mechanism. *Nat. Struct. Mol. Biol.* 16:380–389.
49. Slepko, N., A. M. Bhattacharyya, …, R. Wetzel. 2006. Normal-repeat-length polyglutamine peptides accelerate aggregation nucleation and cytotoxicity of expanded polyglutamine proteins. *Proc. Natl. Acad. Sci. USA.* 103:14367–14372.
50. Bhattacharyya, A. M., A. K. Thakur, and R. Wetzel. 2005. Polyglutamine aggregation nucleation: thermodynamics of a highly unfavorable protein folding reaction. *Proc. Natl. Acad. Sci. USA.* 102:15400–15405.
51. Chen, S., F. A. Ferrone, and R. Wetzel. 2002. Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation. *Proc. Natl. Acad. Sci. USA.* 99:11884–11889.
52. Walters, R. H., and R. M. Murphy. 2009. Examining polyglutamine peptide length: a connection between collapsed conformations and increased aggregation. *J. Mol. Biol.* 393:978–992.
53. Crick, S. L., M. Jayaraman, …, R. V. Pappu. 2006. Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions. *Proc. Natl. Acad. Sci. USA.* 103:16764–16769.
54. Lee, C. C., R. H. Walters, and R. M. Murphy. 2007. Reconsidering the mechanism of polyglutamine peptide aggregation. *Biochemistry.* 46:12810–12820.
55. Vitalis, A., and R. V. Pappu. 2011. Assessing the contribution of heterogeneous distributions of oligomers to aggregation mechanisms of polyglutamine peptides. *Biophys. Chem.* 159:14–23.
56. Vitalis, A., and R. V. Pappu. 2009. ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* 30:673–699.
57. Radhakrishnan, A., A. Vitalis, …, R. V. Pappu. 2012. Improved atomistic Monte Carlo simulations demonstrate that poly-L-proline adopts heterogeneous ensembles of conformations of semi-rigid segments interrupted by kinks. *J. Phys. Chem. B.* 116:6862–6871.
58. Mao, A. H., and R. V. Pappu. 2012. Crystal lattice properties fully determine short-range interaction parameters for alkali and halide ions. *J. Chem. Phys.* 137:064104.
59. Wuttke, R., H. Hofmann, …, B. Schuler. 2014. Temperature-dependent solvation modulates the dimensions of disordered proteins. *Proc. Natl. Acad. Sci. USA.* 111:5213–5218.
60. Choi, J.-M., and R. V. Pappu. 2019. Improvements to the ABSINTH force field for proteins based on experimentally derived amino acid specific backbone conformational statistics. *J. Chem. Theory Comput.* 15:1367–1382.
61. Sugita, Y., and Y. Okamoto. 1999. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314:141–151.
62. Vitalis, A., and R. V. Pappu. 2009. Methods for Monte Carlo simulations of biomacromolecules. *Annu. Rep. Comput. Chem.* 5:49–76.
63. Das, R. K., and R. V. Pappu. 2013. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. USA.* 110:13392–13397.
64. Allegra, G., and F. Ganazzoli. 1985. Polymer collapse in dilute solution: equilibrium and dynamical aspects. *J. Chem. Phys.* 83:397–412.
65. Raos, G., G. Allegra, and F. Ganazzoli. 1994. The collapse of chains with different architectures. *J. Chem. Phys.* 100:7804–7813.
66. Stephen, M. J. 1975. Collapse of a polymer chain. *Phys. Lett. A.* 53:363–364.
67. De Gennes, P. G. 1975. Collapse of a polymer chain in poor solvents. *Journale de Physique Letters.* 36:55–57.
68. Holehouse, A. S., K. Garai, …, R. V. Pappu. 2015. Quantitative assessments of the distinct contributions of polypeptide backbone amides versus side chain groups to chain expansion via chemical denaturation. *J. Am. Chem. Soc.* 137:2984–2995.

69. Oxtoby, D. W., and D. Kashchiev. 1994. A general relation between the nucleation work and the size of the nucleus in multicomponent nucleation. *J. Chem. Phys.* 100:7665–7671.

70. Cahn, J. W. 1961. On spinodal decomposition. *Acta Metall.* 9:795–801.

71. Cahn, J. W., and J. E. Hilliard. 1958. Free energy of a nonuniform system. I. Interfacial free energy. *J. Chem. Phys.* 28:258–267.

72. Berry, J., C. P. Brangwynne, and M. Haataja. 2018. Physical principles of intracellular organization via active and passive phase transitions. *Rep. Prog. Phys.* 81:046601.

73. Berry, J., S. C. Weber, …, C. P. Brangwynne. 2015. RNA transcription modulates phase transition-driven nuclear body assembly. *Proc. Natl. Acad. Sci. USA.* 112:E5237–E5245.

74. Michaels, T. C. T., A. Šarić, …, T. P. J. Knowles. 2018. Chemical kinetics for bridging molecular mechanisms and macroscopic measurements of amyloid fibril formation. *Annu. Rev. Phys. Chem.* 69:273–298.

75. Cohen, S. I. A., M. Vendruscolo, …, T. P. J. Knowles. 2012. From macroscopic measurements to microscopic mechanisms of protein aggregation. *J. Mol. Biol.* 421:160–171.

76. Šarić, A., A. K. Buell, …, D. Frenkel. 2016. Physical determinants of the self-replication of protein fibrils. *Nat. Phys.* 12:874–880.

77. Narayanan, A., A. Meriin, …, I. I. Cisse. 2019. A first order phase transition mechanism underlies protein aggregation in mammalian cells. *eLife.* 8:e39695.

78. Patel, A., H. O. Lee, …, S. Alberti. 2015. A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell.* 162:1066–1077.

79. Alberti, S., A. Gladfelter, and T. Mittag. 2019. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell.* 176:419–434.

80. Sciortino, F., R. Bansil, …, P. Alstrøm. 1993. Interference of phase separation and gelation: a zeroth-order kinetic model. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics.* 47:4615–4618.

81. Varanko, A. K., J. C. Su, and A. Chilkoti. 2020. Elastin-like polypeptides for biomedical applications. *Annu. Rev. Biomed. Eng.* 22:343–369.

82. Flory, P. J. 1942. Thermodynamics of high polymer solutions. *J. Chem. Phys.* 10:51–61.

83. Huggins, M. L. 1941. Solutions of long chain compounds. *J. Chem. Phys.* 9:440.

84. Nott, T. J., E. Petsalaki, …, A. J. Baldwin. 2015. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell.* 57:936–947.

85. Schmidt, H. B., A. Barreau, and R. Rohatgi. 2019. Phase separation-deficient TDP43 remains functional in splicing. *Nat. Commun.* 10:4890.

86. Lin, Y.-H., and H. S. Chan. 2017. Phase separation and single-chain compactness of charged disordered proteins are strongly correlated. *Biophys. J.* 112:2043–2046.

87. Lin, Y.-H., J. P. Brady, …, K. Ghosh. 2020. A unified analytical theory of heteropolymers for sequence-specific phase behaviors of polyelectrolytes and polyampholytes. *J. Chem. Phys.* 152:045102.

88. Chou, H.-Y., and A. Aksimentiev. 2020. Single-protein collapse determines phase equilibria of a biological condensate. *J. Phys. Chem. Lett.* 11:4923–4929.

89. Zeng, X., and R. V. Pappu. 2020. Combined theoretical and computational approach for calculating sequence-specific phase diagrams of thermoresponsive intrinsically disordered homopolypeptides. *Biophys. J.* 118 (*Suppl 1*):304a.

90. Dignon, G. L., W. Zheng, …, J. Mittal. 2019. Temperature-controlled liquid-liquid phase separation of disordered proteins. *ACS Cent. Sci.* 5:821–830.

91. Sherry, K. P., R. K. Das, …, D. Barrick. 2017. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. *Proc. Natl. Acad. Sci. USA.* 114:E9243–E9252.

92. Das, R. K., Y. Huang, …, R. V. Pappu. 2016. Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. *Proc. Natl. Acad. Sci. USA.* 113:5616–5621.

93. Das, R. K., K. M. Ruff, and R. V. Pappu. 2015. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* 32:102–112.

94. Das, R. K., S. L. Crick, and R. V. Pappu. 2012. N-terminal segments modulate the α-helical propensities of the intrinsically disordered basic regions of bZIP proteins. *J. Mol. Biol.* 416:287–299.

95. Staller, M. V., A. S. Holehouse, …, B. A. Cohen. 2018. A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. *Cell Syst.* 6:444–455.e6.

96. Fuertes, G., N. Banterle, …, E. A. Lemke. 2017. Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proc. Natl. Acad. Sci. USA.* 114:E6342–E6351.

97. Ripaud, L., V. Chumakova, …, M. S. Hipp. 2014. Overexpression of Q-rich prion-like proteins suppresses polyQ cytotoxicity and alters the polyQ interactome. *Proc. Natl. Acad. Sci. USA.* 111:18219–18224.

98. Gibbs, E. B., and S. A. Showalter. 2016. Quantification of compactness and local order in the ensemble of the intrinsically disordered protein FCP1. *J. Phys. Chem. B.* 120:8960–8969.

99. Metskas, L. A., and E. Rhoades. 2015. Conformation and dynamics of the troponin I C-terminal domain: combining single-molecule and computational approaches for a disordered protein region. *J. Am. Chem. Soc.* 137:11962–11969.

100. Robustelli, P., S. Piana, and D. E. Shaw. 2018. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. USA.* 115:E4758–E4766.

101. Best, R. B., W. Zheng, and J. Mittal. 2014. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* 10:5113–5124.

102. Zheng, W., G. H. Zerze, …, R. B. Best. 2018. Inferring properties of disordered chains from FRET transfer efficiencies. *J. Chem. Phys.* 148:123329.

103. Best, R. B. 2017. Computational and theoretical advances in studies of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* 42:147–154.

104. Zerze, G. H., R. B. Best, and J. Mittal. 2015. Sequence- and temperature-dependent properties of unfolded and disordered proteins from atomistic simulations. *J. Phys. Chem. B.* 119:14622–14630.

105. Best, R. B., W. Zheng, and J. Mittal. 2015. Correction to balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* 11:1978.

106. Ganguly, D., and J. Chen. 2015. Modulation of the disordered conformational ensembles of the p53 transactivation domain by cancer-associated mutations. *PLoS Comput. Biol.* 11:e1004247.

107. Lee, K. H., and J. Chen. 2017. Optimization of the GBMV2 implicit solvent force field for accurate simulation of protein conformational equilibria. *J. Comput. Chem.* 38:1332–1341.

108. Milles, S., N. Salvi, …, M. R. Jensen. 2018. Characterization of intrinsically disordered proteins and their dynamic complexes: from in vitro to cell-like environments. *Prog. Nucl. Magn. Reson. Spectrosc.* 109:79–100.

109. Salvi, N., A. Abyzov, and M. Blackledge. 2017. Atomic resolution conformational dynamics of intrinsically disordered proteins from NMR spin relaxation. *Prog. Nucl. Magn. Reson. Spectrosc.* 102–103:43–60.

110. Fossat, M. J., and R. V. Pappu. 2019. *q*-canonical Monte Carlo sampling for modeling the linkage between charge regulation and conformational equilibria of peptides. *J. Phys. Chem. B.* 123:6952–6967.

111. Holehouse, A. S., and S. Sukenik. 2020. Controlling structural bias in intrinsically disordered proteins using solution space scanning. *J. Chem. Theory Comput.* 16:1794–1805.

112. McCarty, J., K. T. Delaney, …, J.-E. Shea. 2019. Complete phase diagram for liquid-liquid phase separation of intrinsically disordered proteins. *J. Phys. Chem. Lett.* 10:1644–1652.

113. Lin, Y., J. McCarty, …, S. Han. 2019. Narrow equilibrium window for complex coacervation of tau and RNA under cellular conditions. *eLife.* 8:e42571.

114. Muthukumar, M. 1986. Thermodynamics of polymer solutions. *J. Chem. Phys.* 85:4722–4728.

115. Harmon, T. S., A. S. Holehouse, and R. V. Pappu. 2018. Differential solvation of intrinsically disordered linkers drives the formation of spatially organized droplets in ternary systems of linear multivalent proteins. *New J. Phys.* 20:045002.

116. Guillen-Boixet, J., A. Kopach, …, T. M. Franzmann. 2020. RNA-induced conformational switching and clustering of G3BP drive stress granule assembly by condensation. *Cell.* 181:346–361.e17.

117. Espinosa, J. R., J. A. Joseph, …, R. Collepardo-Guevara. 2020. Liquid network connectivity regulates the stability and composition of biomolecular condensates with many components. *Proc. Natl. Acad. Sci. USA.* 117:13238–13247.

118. Fredrickson, G. H., V. Ganesan, and F. Drolet. 2002. Field-theoretic computer simulation methods for polymers and complex fluids. *Macromolecules.* 35:16–39.

119. Aznauryan, M., L. Delgado, …, B. Schuler. 2016. Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proc. Natl. Acad. Sci. USA.* 113:E5389–E5398.

120. Schuler, B., A. Soranno, …, D. Nettels. 2016. Single-molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins. *Annu. Rev. Biophys.* 45:207–231.

121. Sugar, I. P. 1987. Cooperativity and classification of phase transitions. Application to one- and two-component phospholipid membranes. *J. Phys. Chem.* 91:95–101.

Biophysical Journal, Volume *119*

Supplemental Information

Connecting Coil-to-Globule Transitions to Full Phase Diagrams for Intrinsically Disordered Proteins

Xiangze Zeng, Alex S. Holehouse, Ashutosh Chilkoti, Tanja Mittag, and Rohit V. Pappu

GaussianClusterV2.0.zip is a zip archive of the commented MATLAB code and the data files used to extract system-specific phase diagrams using our adaptation of the Gaussian Cluster Theory. All relevant details are found within the README file in the archive.

Figure S1: Pseudo code for extracting *B,* and *w*

Two approaches were developed to extract *B,* and *w* based on analysis of $\sqrt{\langle R_g^2 \rangle}$ versus *T* that we obtain from simulations. In Approach 1, we set $\frac{\left(T^* - T_\theta\right)}{T^*} Bn_K^{0.5} = -0.35$ irrespective of the value of *w.* In Approach, the choice of $\frac{\left(T^* - T_\theta\right)}{T^*} Bn_K^{0.5}$ is coupled to the choice made for *w*, although the equality used in Approach 1 provides bounds on the value for $\frac{\left(T^* - T_\theta\right)}{T^*} Bn_K^{0.5}$. The relation between $\frac{\left(T^* - T_\theta\right)}{T^*} Bn_K^{0.5}$ and *w* is obtained by numerically calculating the saddle point of the contraction ratio profile from GCT at different values for *w*. These two approaches give very similar results and the choice of approaches does not affect our conclusions. The pseudo code for approaches 1 and 2 is referenced as **Figure S1** in the main text.

A. Pseudo code for Approach 1:

for T^* in $[T_{min}, T_\theta]$, where T_{min} is the lowest simulation temperature:

do

get *B* based on $\frac{\left(T^* - T_\theta\right)}{T^*} Bn_K^{0.5} = -0.35$;

convert the simulated of $\sqrt{\langle R_g^2 \rangle}$ versus *T* curve to a contraction ratio profile as a function $\tau Bn_K^{0.5}$ and denote it as $\alpha_s^{sim}\left(\tau Bn_K^{0.5}\right)$;

Next, perform a parameter scan for *w*;

For *w* in $[w_{min}, w_{max}]$ ([0.001, 0.04] has proven to be sufficient)

do

Get the contraction ratio profile $\alpha_s^{GCT}\left(\tau Bn_K^{0.5}\right)$

Compute the difference between $\alpha_s^{sim}\left(\tau Bn_K^{0.5}\right)$ and $\alpha_s^{GCT}\left(\tau Bn_K^{0.5}\right)$;

Since the minimum of $\alpha_s^{GCT}\left(\tau Bn_K^{0.5}\right)$ always larger than the minimum of $\alpha_s^{sim}\left(\tau Bn_K^{0.5}\right)$, a scaling factor λ is used to best $\alpha_s^{sim}\left(\tau Bn_K^{0.5}\right)$ and $\alpha_s^{GCT}\left(\tau Bn_K^{0.5}\right)$ for λ in [0.3, 0.9]

do

compute difference $\alpha_s^{\mathrm{sim}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)$ and $\lambda\left[\alpha_s^{\mathrm{GCT}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)-1\right]+1$

done

The minimum difference between $\alpha_s^{\mathrm{sim}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)$ and $\lambda\left[\alpha_s^{\mathrm{GCT}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)-1\right]+1$ is chosen as the difference between $\alpha_s^{\mathrm{GCT}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)$ and $\alpha_s^{\mathrm{sim}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)$

done

done

The parameters (B, w) that yield the minimal difference between $\alpha_s^{\mathrm{GCT}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)$ and $\alpha_s^{\mathrm{sim}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)$ is used for the system.

B. Pseudo code for Approach 2:

w in $[w_{\min}, w_{\max}]$

do

Get $\alpha_s^{\mathrm{GCT}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)$ for the given w;

Get the position for the saddle point, namely the value of $\frac{\left(T^*-T_\theta\right)}{T^*}Bn_{\mathrm{K}}^{0.5}=a$ for T^* in $[T_{\min}, T_\theta]$,

do

Compute B based the value of $\frac{\left(T^*-T_\theta\right)}{T^*}Bn_{\mathrm{K}}^{0.5}=a$

Compute $\alpha_s^{\mathrm{sim}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)$ for a given B and w;

Use the method in Approach 1 to calculate the difference between $\alpha_s^{\mathrm{GCT}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)$ and $\alpha_s^{\mathrm{sim}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)$;

done

done

Again, the parameters (B, w) that yield the minimal difference between $\alpha_s^{\mathrm{GCT}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)$ and $\alpha_s^{\mathrm{sim}}\left(\tau Bn_{\mathrm{K}}^{0.5}\right)$ is used for the system.

Table S1. Parameters for the temperature dependent free energy of solvation

Unit of interest	Reference temperature T_0 (K)	Reference free energy of solvation $\Delta G_{FES}(T_0)$ kcal/mol	Enthalpy of solvation ΔH_{FES} kcal/mol	Heat capacity of solvation Δc_P cal /mol-K
Backbone	298	-10.1	-17.0	4.00
Thr	298	-5.0	-12.5	4.66
Pro	298	2.0	-5.4	7.05
Lys	298	-100.9	-111.8	5.08
Ala	298	1.9	-3.3	4.96
Met	298	1.4	-10.0	5.59

Table S2. Details of all-atom simulation parameters for each system

System	Number of replicas	Temperature schedule	Number of independent simulations
Q_{20}	18	280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 410, 420, 430, 440, 450 K	4
Q_{30}			
Q_{40}			
Q_{50}			
Q_{60}			
Q_{70}			
$(QGQSPYG)_9$			
$(TPKAMAP)_9$	18	300, 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 420, 440, 460, 480, 500, 520, 540 K	3

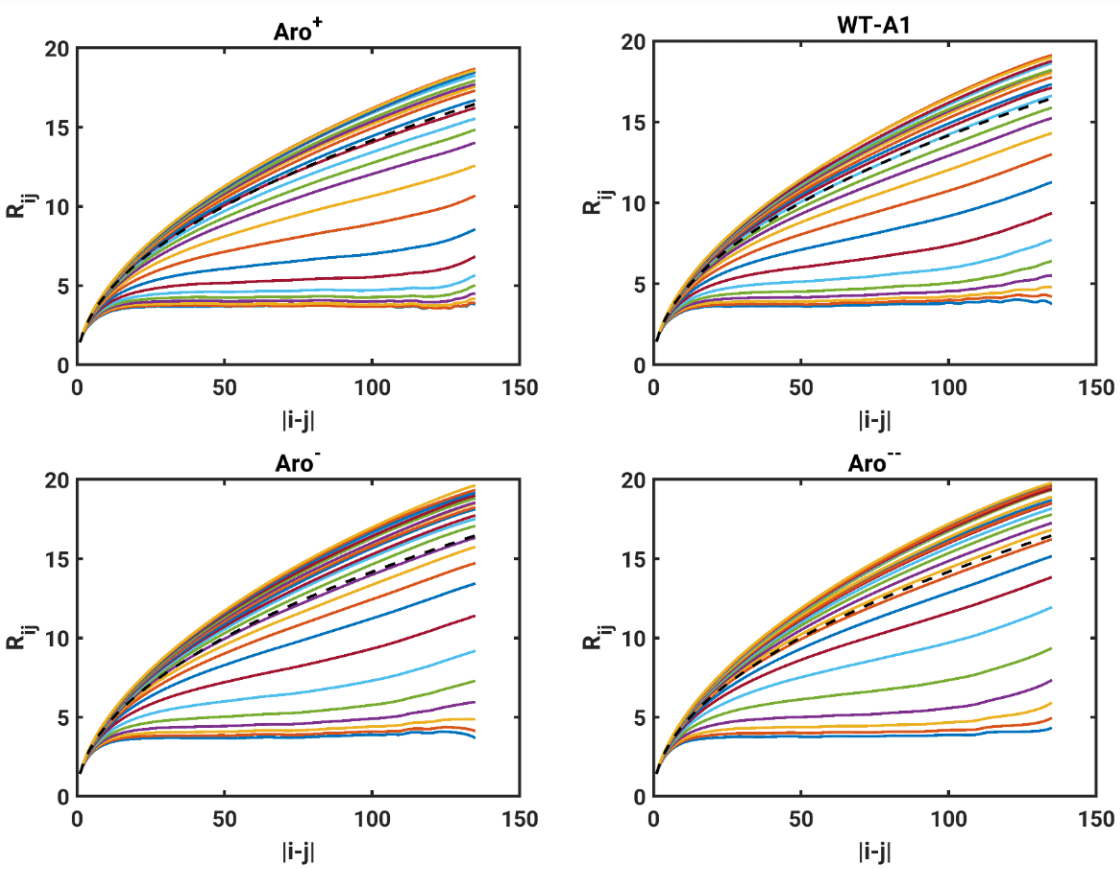

Figure S2. Internal scaling distance for Aro^{+}, WT-A1, Aro^{-}, and Aro^{--} variants. Dashed line denotes the reference internal scaling profile: $\left\langle\left\langle R_{ij}\right\rangle\right\rangle = R_0\left|j-i\right|^{0.5}$. Here, R_0 is the average value obtained across all the temperatures when we set |j–i| = 1.

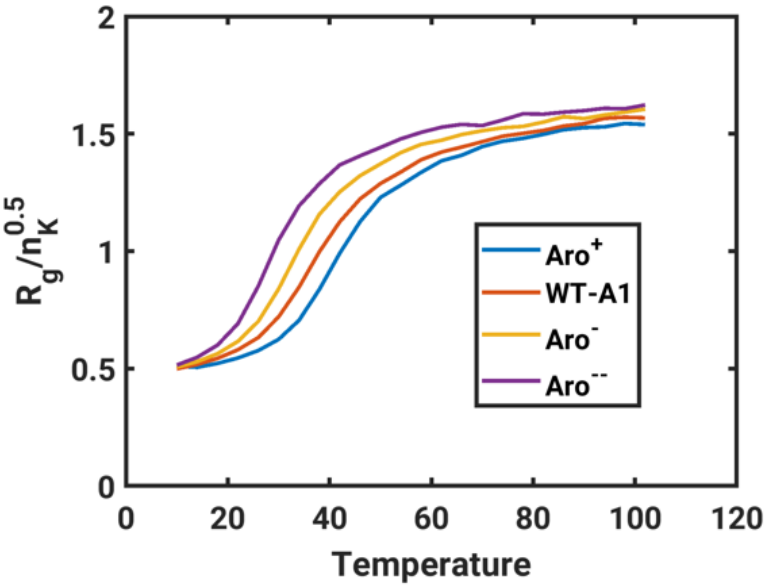

Figure S3. Plot of normalized R_g for Aro$^+$, WT-A1, Aro$^-$, and Aro^{--}.

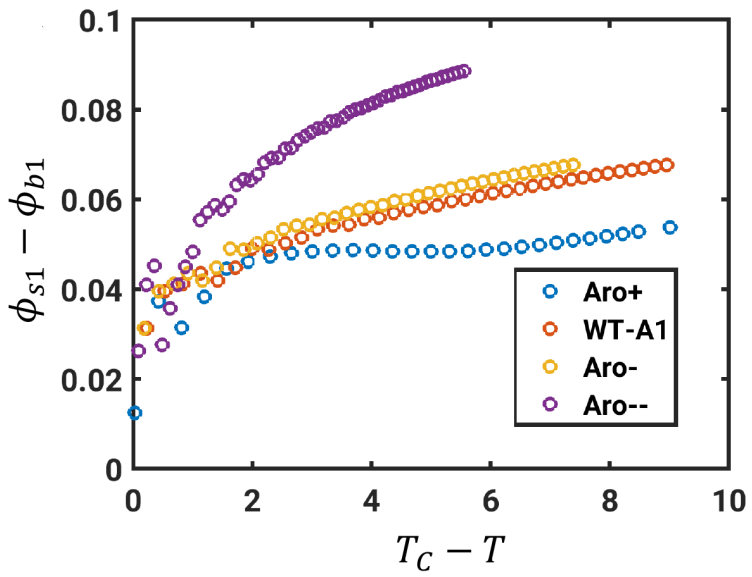

Figure S4: Analysis of how the width of the metastable regime $\Delta\phi_m(T)$ varies with $(T - T_c)$ for each of the four sequences, Aro$^+$, WT-A1, Aro$^-$, and Aro^{--}.

Sensitivity analysis for estimates of *B* and *w*

WT-A1 is used to test the robustness of the fitting result. There are 11 sampled temperatures below T_θ for A1-WT. The robustness is tested as following:

a) A subset of the 11 sampled temperatures which contains N randomly non-repeating temperatures is chosen at random.
b) This subset is used to fit *B* and *w*.
c) For each N in [5, 10], repeat a) and b) 10 times, and then get the mean value of *B* and *w*: B_N^{mean} and w_N^{mean}, and the corresponding standard deviation B_N^{std} and w_N^{std}.
d) Plot B_N^{mean} and w_N^{mean} as a function of N. B_N^{std} and w_N^{std} are plotted as the error bar. A complete analysis from a) to d) is denoted as one trial.
e) Repeat a)-d) 4 times, namely, perform 4 trial analysis.

Results are shown in **Figure S5**. Black dashed line indicates *B* or *w* fitted by using all the data. The fitting result becomes convergent when N>6.

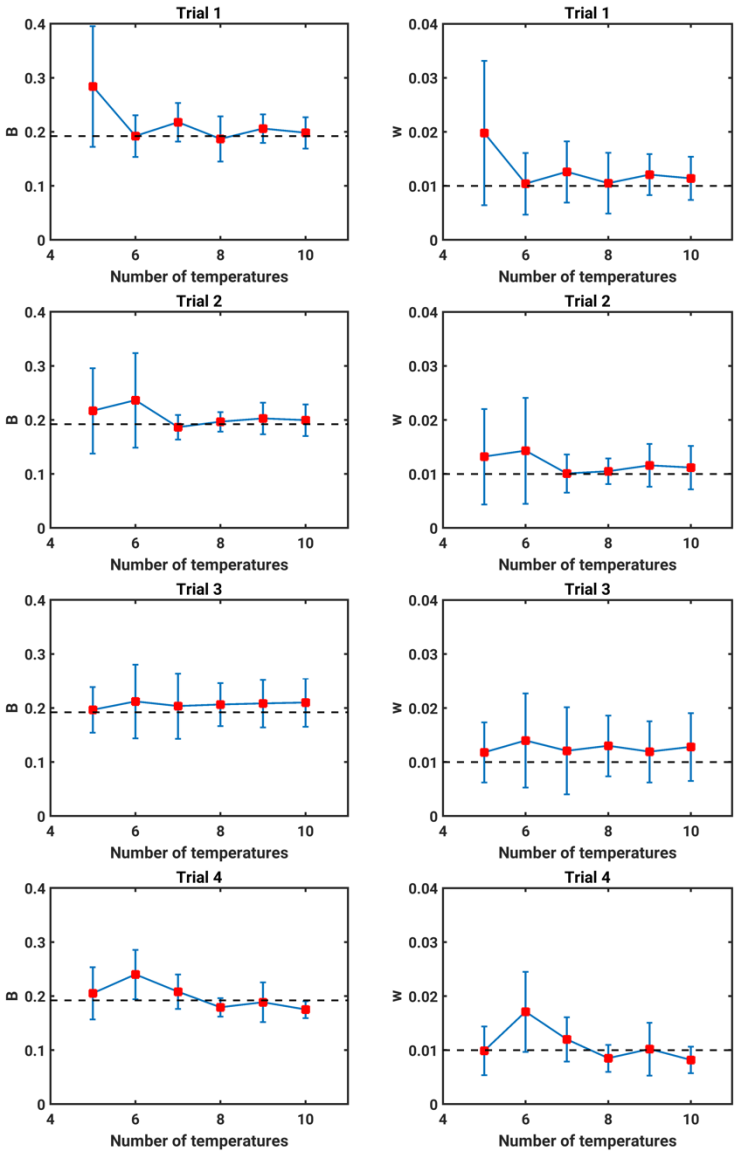

Figure S5: Sensitivity analysis of for estimates of *B* and *w*. Red square indicates the mean values of *B* or *w* for the subset with same number of sampled temperatures. Error bar indicates the stand deviation. Black dashed line indicates *B* or *w* fitted by using all the data.

Fitting result by using different T^*

We assess the impact of different values of T^* on the fitting procedure used to obtain B and w. The set of parameters $\{T^*, B, w]$ that best fits the contraction ratio profile from simulation (α_S^{sim}) is chosen as the final parameter to calculate the phase diagram. The searching range for T^* is $\left[T|_{\alpha_S^{sim}=0.5(1+\min(\alpha_S^{sim}))} - 10, T|_{\alpha_S^{sim}=0.5(1+\min(\alpha_S^{sim}))} + 10\right]$. Here, $\min\left(\alpha_s^{sim}\right)$ is the minimum value of the contraction ratio from the simulations. For the UCST system, $\min\left(\alpha_s^{sim}\right)$ is the contraction ratio at the lowest sampled temperature. For the LCST system, it is the value of the contraction ratio at the highest sampled temperature. **Figure S6** shows the fitting result by using different T^* for WT-A1 system. Approach 2 in Figure S1 is used for the fitting process.

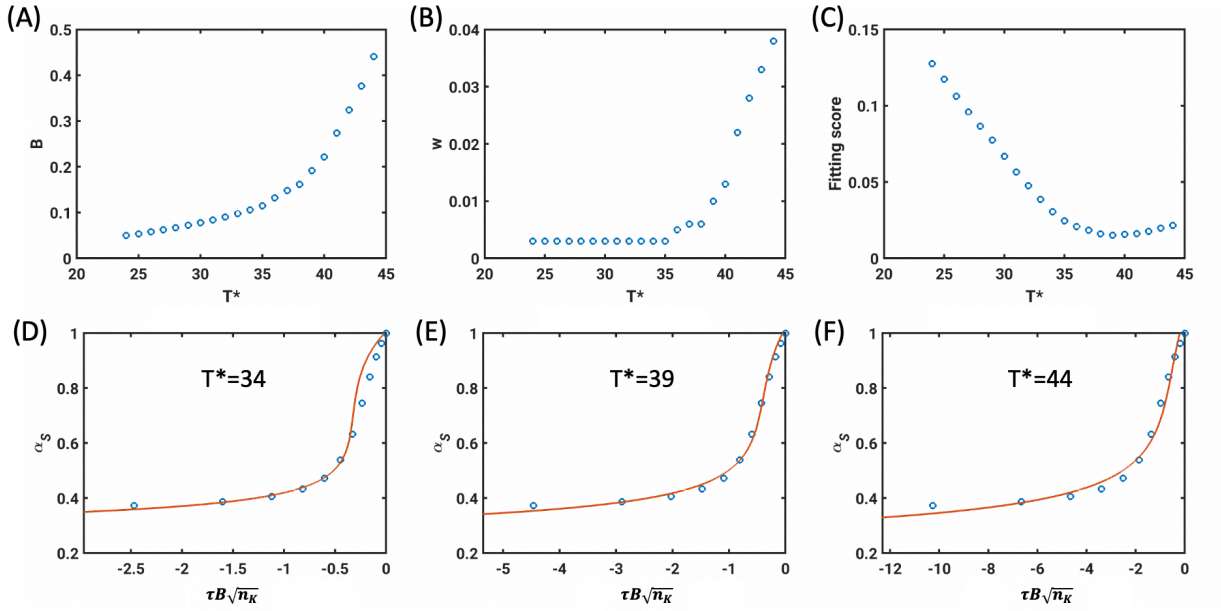

Figure S6: Fitting result by using different T^* for WT-A1 system. (A)-(C) B, w and the fitting score as a function of T^*. The fitting function is defined as the mean absolute difference between α_S^{simul} and α_S^{GCT}, detailed information about calculating the difference between α_S^{simul} and α_S^{GCT} can be found in **Figure S1**. $T^* = 39$ has the minimum fitting score, thus B and w fitted from $T^* = 39$ are chosen as the final parameter to calculate the phase diagram.**(D)-(F)** Comparison of α_S^{simul} and α_S^{GCT} at different T^*.