

RESEARCH ARTICLE

Analysis and design of disordered polypeptides with optimized sequence patterning properties

Arjun Singh^{*}, Ali I. Ukperaj^{*}, Gabriel F. Porto, Gregory L. Dignon^{†*}

Department of Chemical and Biochemical Engineering, Rutgers, The State University of New Jersey, Piscataway, New Jersey, United States of America

[†] These authors contributed equally to this work.

* gregory.dignon@rutgers.edu



OPEN ACCESS

Citation: Singh A, Ukperaj AI, Porto GF, Dignon GL (2026) Analysis and design of disordered polypeptides with optimized sequence patterning properties. PLoS Comput Biol 22(7): e1014462. <https://doi.org/10.1371/journal.pcbi.1014462>

Editor: Yaakov Koby Levy, Weizmann Institute of Science, ISRAEL

Received: February 20, 2026

Accepted: June 17, 2026

Published: July 8, 2026

Copyright: © 2026 Singh et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: All data used to generate figures in this work will have been deposited on an online database which can be accessed at <https://zenodo.org/records/20397905>. Source code for Monte Carlo sequence design software is also

Abstract

Intrinsically disordered proteins (IDPs) exhibit phase separation behavior that is closely linked to their degree of single-chain compaction, which in turn is governed by both amino acid composition and sequence patterning. Existing metrics such as sequence charge decoration (SCD) and sequence hydropathy decoration (SHD) describe these effects but are largely limited to describing differences between sequences of similar length and overall composition. In this work, we present a shuffle-based normalization scheme for SCD and SHD, enabling comparison of sequence patterning between very different IDP sequences. Leveraging this normalization scheme toward design space, we develop a Monte Carlo based sequence design algorithm that generates novel IDPs with desired patterning features. Our design framework is further strengthened by incorporating additional metrics such as sequence aromatic decoration (SAD), compositional RMSD, and a previously developed sequence based ΔG predictor. We validate our approach through coarse-grained MD simulations, showing that the designed sequences exhibit tunable phase behavior. This strategy lays the groundwork for rational design of IDPs for biomedical and biotechnology applications, as well as basic biophysical research.

Author summary

Intrinsically disordered proteins behave similar to polymers in solution, having no defined structure. Their behavior is dictated by the collection of shapes the protein adopts, known as its “conformational ensemble” which is tuned by its amino acid sequence and the solution environment. In this work, we have developed parameters to describe the patterning of charged and hydrophobic amino acids within these protein sequences, which are predictive of their ability to phase separate and form dense liquid-like droplets in solution. Importantly, the parameters we develop are motivated by physics and can be applied across a large number of amino acid sequences rapidly. This will enable researchers to rapidly predict

available on the same link, and on Github
<https://github.com/Dignon-Lab/Monte-Carlo>.

Funding: This study was supported by NIH Grant R35GM150589 to GLD (National Institutes of Health (NIH) National Institute for General Medical Sciences (NIGMS) <https://www.nigms.nih.gov/>). AS and AIU are paid their PhD stipends through this grant, and GLD summer salary is paid through this grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

the behavior of large libraries of protein sequences. We have additionally developed a software to design randomized amino acid sequences with desired amino acid composition and patterning properties. Finally, we have tested our design scheme and parameters by running simulations of designed IDP sequences and quantified each of their ability to phase separate.

Introduction

Biomolecular condensates are specialized cellular compartments that form without membranes, helping cells organize important molecules through a process called liquid-liquid phase separation (LLPS) [1]. Unlike traditional organelles, condensates form membraneless organelles, compartments which can assemble and disassemble in response to changing conditions and protein properties [2]. They play essential roles in organizing and regulating biological processes, such as DNA replication and repair, transcription, chromatin remodeling, RNA metabolism, ribosome biogenesis, protein quality control, cell division, intercellular adhesion, and signal transduction [3–9]. Disruptions in condensate dynamics have been linked to neurodegenerative diseases, cancer, and viral infections, with aberrant signaling and assembly potentially contributing to pathological states and pathways such as uncontrolled proliferation and aggregation [10–13]. Given their broad functional importance, understanding how biomolecular condensates form and operate is vital for unraveling their roles in both normal physiology and disease [14].

A major factor influencing condensate formation is the distinct behavior of intrinsically disordered proteins (IDPs), which lack a stable three-dimensional structure and instead fluctuate among a diverse ensemble of conformations [15]. This conformational ensemble is governed by the IDP's amino acid sequence and composition, which in turn determines how it interacts with other molecules and contributes to phase separation and condensate dynamics [16,17]. Compositional features such as net charge, mean hydropathy, and charge density offer valuable insights into the conformational preferences and phase behavior of IDPs [16,18]. Beyond composition, the arrangement of charged or aromatic residues within the amino acid sequence further modulates chain compaction and phase separation propensity [19–22]. Understanding these principles opens the door to rational sequence design, not only to map how sequence governs behavior, but also to engineer IDPs for specific functions or subcellular localizations [23–26]. Moreover, short synthetic sequences can serve as minimal models for LLPS, enabling mechanistic studies under controlled conditions [27].

Computational and theoretical studies have shown that the degree of single-chain compaction in soluble IDPs strongly correlates with their phase separation propensity—more compact chains generally phase separate more readily [21,28–30]. This compaction is influenced by overall protein composition as well as the patterning of particular amino acid chemistries within the sequence. For example, blocky distributions of like-charged or hydrophobic residues promote intra-chain attractions, leading

to collapse, and in turn, enhance intermolecular interactions that drive LLPS [20,21,31]. Quantitative metrics such as sequence hydropathy decoration (SHD) and sequence charge decoration (SCD) capture these effects and help predict phase behavior [24,32,33]. Our goal here is to refine the definitions of various patterning parameters and utilize them in design of unique IDP sequences.

In this work, we address limitations in how sequence-based patterning parameters—such as sequence charge decoration (SCD) and sequence hydropathy decoration (SHD), are currently defined and applied. Previous studies have shown advances where modifications to a sequence charge patterning or arrangement of hydrophobic/aromatic residues has a significant impact on their phase behavior, validated by both simulations and experiments [34–36]. While these metrics capture how charged and hydrophobic residues are arranged along a sequence, it is difficult to compare the degree of “blockiness” between sequences that are of different length or composition. Thus, studies applying SCD and SHD are typically limited to mutagenesis or randomization of individual sequences. In this work, we look to address this issue by proposing a normalization scheme, akin to the z-score approach from Pappu lab’s NARDINI parameter set [37,38] and further propose empirical methods of approximating the sampled distribution as a function of compositional terms. This method is then leveraged to develop a Monte Carlo algorithm for the directed design of novel sequences with desired patterning characteristics. To enhance the robustness of this design framework, we later incorporate additional parameters including sequence aromatic decoration (SAD), compositional RMSD, and a sequence-based ΔG predictor [39]. Finally, we validate the designed sequences across diverse protein sequences using HPS-Urry simulations, demonstrating the algorithm’s ability to generate proteins with tunable phase behavior, laying the groundwork for applications such as biomolecular condensate-based drug delivery and determination of compositional control within condensates [2,40].

Results

Generalized parameters to quantify patterning within a fixed composition

Previous work has shown that the critical temperature of IDP phase separation can be roughly predicted by the degree of collapse as a single chain [21,29,30,41], generally demonstrating that IDPs with a more collapsed conformation will be more capable of phase separating, while those more extended will be incapable of phase separation on their own. It has also been demonstrated that the patterning of charged and hydrophobic amino acids is predictive of IDP chain dimensions [19,20,32] and phase separation [21,35,36]. In this work, we focus specifically on the patterning parameters sequence hydropathy decoration (SHD) and sequence charge decoration (SCD). SHD is calculated from the scaled hydropathy values and sequence separation of all residues in an amino acid sequence:

$$SHD = \frac{1}{N} \sum_i^{N-1} \sum_{j>i}^N \frac{(\lambda_i + \lambda_j)}{j-i} \quad (1)$$

where N is the sequence length and λ_i is the scaled hydropathy value of residue i according to the Urry hydropathy scale [42] (Table A in S1 Appendix). SCD is calculated from the relative charge and sequence separation of formally charged residues in an amino acid sequence:

$$SCD = \frac{1}{N} \sum_i^{N-1} \sum_{j>i}^N q_i q_j (j-i)^{1/2} \quad (2)$$

where N is the sequence length and q_i is the formal charge of residue i .

Both parameters are demonstrated to be very effective at predicting IDP conformations [32,43] as well as phase separation and materials properties [33,35]. We note, however, that there is a limitation to using such metrics since they

effectively describe the degree of patterning, but only within sequences of identical length and composition. When used for sequences having drastically different composition or size, we observe widely varying values of SCD and SHD. To demonstrate this disconnect, we consider seven disordered sequences of different sizes and compositions from biology (Table B in [S1 Appendix](#)). To evaluate the effect of sequence composition on the range of accessible SCD and SHD values, we take an amino acid sequence and shuffle it 1 million times, producing only sequences with identical composition, but random ordering of amino acids. This can be considered as a representative ensemble of constant-composition sequences, from which we can approximate the underlying probability density of patterning parameters and the “blockiness” of sequences. From these sequence ensembles, we calculate SCD and SHD and plot the probability distributions ([Fig 1A,B](#)). We note that for each individual sequence, the probability distributions have no overlap in SHD and only moderate overlap in SCD, making direct comparison of SHD or SCD values between different proteins challenging to interpret. For instance, the randomized sequence of FUS LC that is the least “blocky” has a greater SHD value than the most “blocky” randomized sequence of hnRNPA2 LC. Thus, we note the limitation of SCD and SHD for quantifying blockiness of disparate sequences, particularly those with differing size and composition, and highlight the need for a universal normalization scheme.

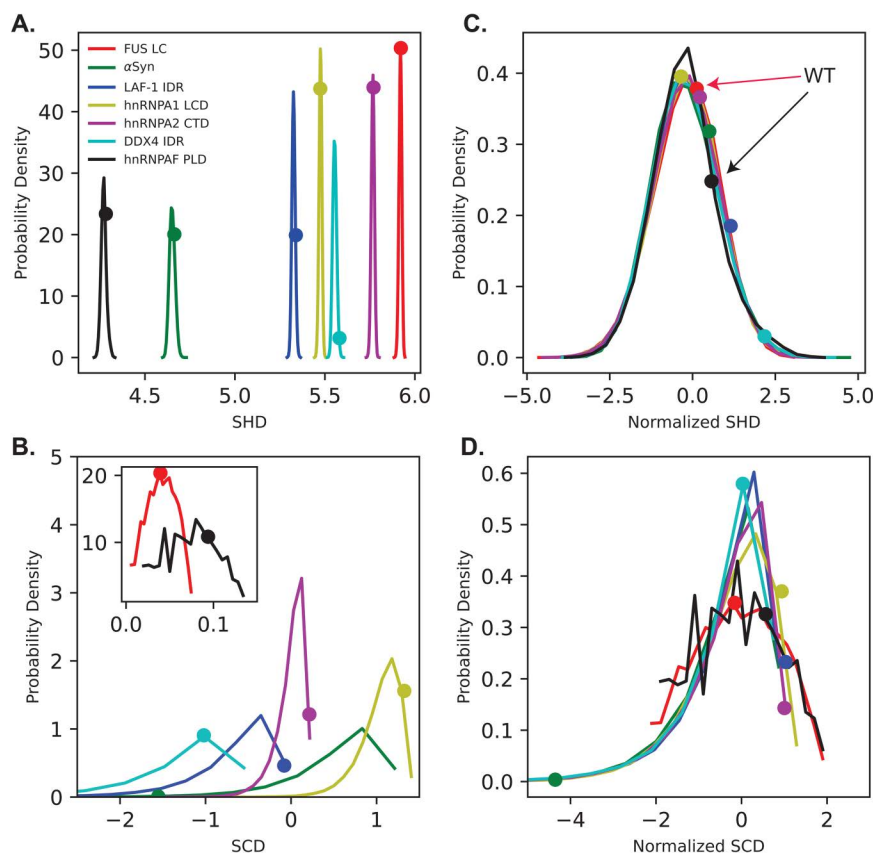


Fig 1. The probability density of patterning parameter values based on real IDP sequences that have been shuffled 1 million times each. **A)** SHD values of shuffled IDP sequences each follow normal distributions with different mean values, and relatively small spread. **B)** SCD values from shuffled sequences show more spread and overlap than SHD (FUS and hnRNPAF shown in inset since they have very few charged residues), **C)** collapsed SHD_{norm} and **D)** SCD_{norm} values match well with the standard normal distribution, with SCD_{norm} having an extended tail toward negative numbers. In all panels, circles indicate the WT sequence value on each distribution.

<https://doi.org/10.1371/journal.pcbi.1014462.g001>

$$SHD_{norm} = \frac{SHD - \mu_{SHD}}{\sigma_{SHD}} \quad (3)$$

$$SCD_{norm} = \frac{SCD - \mu_{SCD}}{\sigma_{SCD}} \quad (4)$$

We find that the probability distributions are roughly gaussian in shape, allowing each distribution to be collapsed to the standard normal distribution (Fig 1C,D). This normalization serves to identify the significance of a value within each parameter's respective fixed-composition distributions, giving a sense of whether this protein sequence is well-mixed or segregated for something of its composition. For example, the SCD_{norm} of WT LAF-1 RGG is on the right arm of the probability distribution, while WT α -synuclein is far to the left of the main density of the distribution. This indicates that the charged residues of LAF-1 RGG and α -synuclein are more evenly distributed and more blocky than would be expected if randomly arranged, respectively. We note that while raw SCD values tend to have wider overlapping distributions and SHD has narrow, non-overlapping distributions, this is simply due to the number of sample distance comparisons within the given sequences, having all residue pairs included in SHD calculations, but only charged residues considered for SCD calculations.

We then ask the question of whether we can bypass the need for shuffling sequences to obtain the probability density of accessible SCD and SHD values. Toward this goal of approximating μ_{SHD} and σ_{SHD} for a given composition, we must first identify descriptors that don't change upon shuffling, being only composition-dependent and not sequence-dependent. Starting with SHD, we find that the shuffled ensemble probability distribution can be predicted by just three composition-based parameters, i.e., the mean hydropathy (commonly denoted $\langle \lambda \rangle$, but referred to here as μ_λ), standard deviation of hydropathy (σ_λ) of all residues in the sequence, and chain length (N). To test the predictability, we designed 48 mock sequences ranging in size from 50 to 200 residues, with μ_λ between 0.4 to 0.7, and with σ_λ values up to 0.5 depending on what was reasonably attainable for the given μ_λ . For example, a polyglycine sequence will have $\mu_\lambda \sim 0.5$ and $\sigma_\lambda = 0$, while a sequence comprising half glutamate and half tryptophan will have roughly the same μ_λ , but will have a larger σ_λ value ~ 0.5 . Taking these 48 sequences, we shuffled each one million times, and calculated the μ_{SHD} and σ_{SHD} values to provide real distributions which can then be predicted by the composition-based parameters (Fig 2).

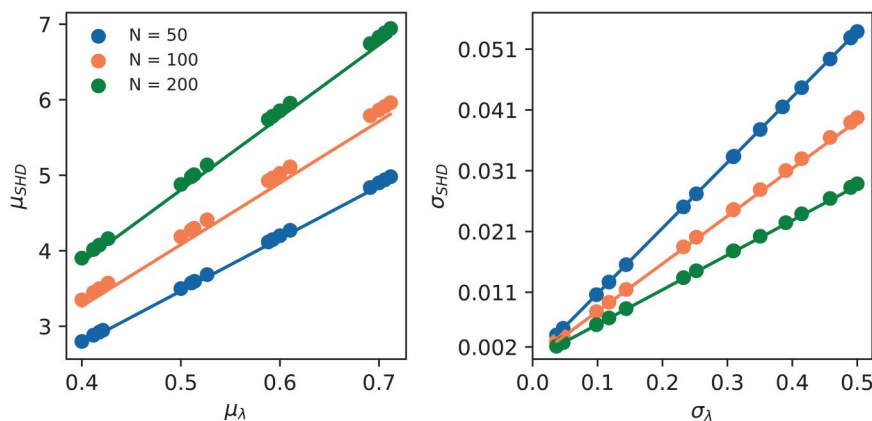


Fig 2. Predicting probability density of SHD from composition-dependent parameters. A) The average SHD for a constant-composition ensemble depends on the average hydropathy and length of the sequence. B) The standard deviation of SHD values for constant-composition ensemble depends on standard deviation of hydropathy and length.

<https://doi.org/10.1371/journal.pcbi.1014462.g002>

We first compare μ_{SHD} with μ_{λ} and find that there is a positive linear correlation, having a slope dependent on the length of the sequence (Fig 2A). We then compare σ_{SHD} with σ_{λ} and find another linear correlation with an N-dependent slope (Fig 2B). Notably, μ_{SHD} is larger for longer sequences, indicating that the value of SHD increases for longer and more hydrophobic sequences (Fig 2A). In contrast, σ_{SHD} is smaller with increasing N, indicating smaller fluctuations occurring for greater sampling (Fig 2B). Intriguingly, we find that the average hydrophathy is not predictive of σ_{SHD} on its own, and the standard deviation of hydrophathy is not predictive of μ_{SHD} (Fig A in S1 Appendix). Using this data set, we derive a simple empirical formula for both the mean and standard deviation of possible SHD values based solely on sequence composition within the bounds of the tested sequences,

$$\mu_{SHD} = 2.761\mu_{\lambda} \cdot N^{0.2354} \quad (5)$$

$$\sigma_{SHD} = 0.6500\sigma_{\lambda} \cdot N^{-0.4592} \quad (6)$$

This enables prediction of the probability distribution of SHD values for any amino acid sequence without the need for hundreds of thousands of shuffles, and provides a useful normalization scheme for the SHD patterning parameter.

We repeated this process for the SCD parameter, aiming to identify composition-dependent but sequence-independent features that could be used to estimate μ_{SCD} and σ_{SCD} . We similarly generated a library of synthetic sequences with varied charge compositions, having varied values of fraction of charged residues (FCR), net charge per residue (NCPR), and chain length (N). We observed that both the mean (μ_{SCD}) and standard deviation (σ_{SCD}) of the SCD distribution showed approximately linear relationships with FCR (Fig 3A,B), indicating that FCR alone provides a useful predictor in these cases. On the other hand, the relationships between SCD statistics and NCPR were non-linear (Fig 2C,D), suggesting more complex dependencies when considering net charge effects. Using these trends, we determined empirical equations that combine FCR, NCPR, and N to estimate the mean and standard deviation of the SCD distribution for a given sequence composition:

$$\mu_{SCD} = 0.2555 \cdot N^{1.5092} \cdot (NCPR^2 - 0.0054 \cdot FCR) \quad (7)$$

$$\sigma_{SCD} = (0.0531 \cdot N^{1.1216}) \cdot FCR \cdot (NCPR^2 - 1.3640 \cdot NCPR^4 + N^{-0.3991}) \quad (8)$$

In addition to charge and hydrophathy patterning, there is also known to be an influence of aromatic residue patterning on IDP conformations and phase separation [22]. We thus develop a third patterning parameter, focused specifically on aromatic residues within the amino acid sequence, termed sequence aromaticity decoration (SAD) and following a similar functional form to SHD:

$$SAD = \sum_i^{N-1} \sum_{j>i}^N \frac{(a_i \cdot a_j)}{j-i} \quad (9)$$

where a_i and a_j are the aromaticity of residues i and j . Aromatic residues are assigned a value of 1 and non-aromatic residues a value of 0. We find that upon the shuffling of random sequences, AD also follows a roughly gaussian distribution, and can be predicted using the following empirical equations:

$$\mu_{SAD} = 1.6188 \cdot \mu_A^{2.0390} \cdot N^{0.2110} \quad (10)$$

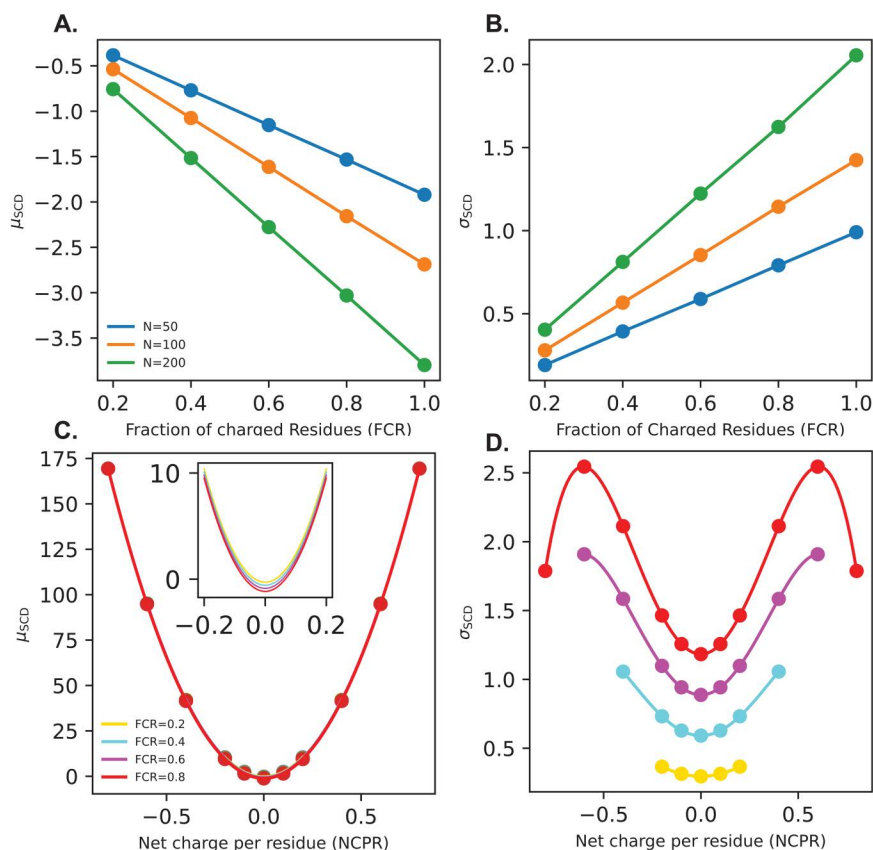


Fig 3. Prediction of SCD distribution from sequence-independent parameters. Each point represents statistics calculated from 1 million shuffled sequences per composition, with solid lines indicating empirical fits. A-B) Mean and standard deviation of SCD vs. FCR for sequences with net charge of zero. C-D) Mean and standard deviation of SCD vs. NCPR for sequences of size $N=100$. Inset in panel C highlights smaller influence of FCR on average SCD.

<https://doi.org/10.1371/journal.pcbi.1014462.g003>

$$\sigma_{SAD} = 1.4177 \cdot \sigma_A^{2.4598} \cdot N^{-0.4404} \quad (11)$$

where μ_{SAD} and σ_{SAD} are the mean and standard deviation of aromaticity in the sequence, respectively. Intriguingly, both mean and standard deviation of the SAD distribution are dependent on both mean and standard deviation of the aromaticity in the amino acid sequence (μ_A and σ_A). We find this is due to the fact that aromaticity is a binary in the sequence, and thus, the mean and standard deviation of aromaticity in a set composition are related to each other through the relation of random binary data, i.e., $\sigma = \sqrt{\mu(1-\mu)}$, since μ_A is constrained between 0 and 1. SAD will serve as a useful design parameter as it should be able to help design sequences with well-dispersed aromatic residues, and reduce the likelihood of blocks of aromatic residues that can lead to aggregation or more solid or gel-like behavior of condensates [22]. We are also curious about the redundancy of these three parameters in describing IDP dimensions since each is dependent on a summation over sequence pair separation $j-i$, and somewhat correlated physicochemical properties of each amino acid (i.e., hydrophobic residues are not charged). To evaluate this, we generated 10^5 random sequences with identical composition to 3 different IDPs used in this work, finding that the pairwise correlation of each pair of patterning parameters is quite minimal (Fig B in S1 Appendix). Indeed, the most significant Pearson correlation coefficient (r) is found between SHD and SAD in FUS LC with a value of 0.24.

Sequence Monte Carlo to design polypeptides with desirable properties

Contextualization and normalization of patterning parameters across all compositions and sizes offers great potential to accelerate design of sequences with multiple optimized physical properties. We have effectively created a coordinate space within which values of these patterning parameters are equally weighted regardless of differing sequence compositions and lengths. Toward this goal, we leverage a Monte Carlo sampling algorithm to design new sequences with desirable properties. This includes quickly formulating sequences with targeted values of patterning parameters, as well as creating identical compositions of existing sequences with a tuned propensity to phase separate.

We first define a loss function as a Monte Carlo energy function by inputting “target” values for a set of parameters, namely the aforementioned patterning parameters, and then measure the deviation of a designed sequence from each target value. Notably, this will favor sequence modifications that move each parameter (i) closer to the target value (i_{target}), and penalize modifications that move further away.

$$E_{seq} = \sum_{i \in [SCD, SHD, SAD, \dots]} W_i \times (|i - i_{target}|) \quad (12)$$

To ensure changes in each parameter are counted, and no one parameter is overly represented in the loss function, we set each parameter’s weight to $W_i = 1/\sigma_i$, where σ_i is the standard deviation of that parameter’s probability distribution calculated with respect to the composition and using [Equations 6, 8, and 11](#). This effectively makes the change in the MC energy equivalent to the sum of the change in SCD_{norm} , SHD_{norm} and SAD_{norm} . As a result of this weighting scheme, each parameter has roughly equivalent importance in the overall design and optimization of an IDP sequence. This approach can also incorporate other design parameters, such as aromatic patterning, and deviation from a set sequence composition ([Fig 4A–C](#)). We can also incorporate direct predictions of phase separation capability, such as the ΔG predicted from sequence using the algorithm developed by von Bülow et al. [\[39\]](#). For these terms, we are not able to produce similar weights as the previously described patterning terms, and thus opt to weigh them by their maximum possible value instead. We find that these weights keep term contributions within 2–3 orders of magnitude of changes to the patterning terms. We implement each of these and discuss further in the following sections. Notably, the process of sequence design is modular and can be made to incorporate even more design parameters than discussed in this work.

To carry out sequence design, we use a Metropolis Monte Carlo approach where perturbations are made to the sequence, the new sequence “energy” is calculated, and moves are accepted or rejected based on the Metropolis Criterion [\[44\]](#). To perturb the sequence, we introduce three types of “moves”, namely mutations, swaps and partial sequence shuffles ([Fig 4D](#)). The MC moves result in changes in the loss function (ΔE) which is then used to accept or reject the move based on the Metropolis criterion

$$P_{accept} = \min(e^{-\Delta E_{seq}/T}, 1) \quad (13)$$

where T is a scaling parameter that determines the likelihood of accepting uphill moves to higher energies, potentially aiding the escape of local minima in the energy function. For our study, we have set this to be 10^{-2} as an intermediate value allowing some exploration, but ensuring overall migration toward the optimized design parameters. This process will repeat until all parameters are within an interval of their desired values, or until the maximum set number of desired cycles has been reached.

As a trial run, we use LAF-1 RGG as an input and begin to design a target sequence with each design parameter set to several standard deviations away from the initial value, with the exception of composition, which is set to stay as close to the original sequence composition as possible—RMSD=0. SCD, SHD, and SAD optimization was particularly fast, reaching target values in around 200 moves ([Fig 5A,B](#)). Alternatively, the ΔG predictor was able to reach its target value within

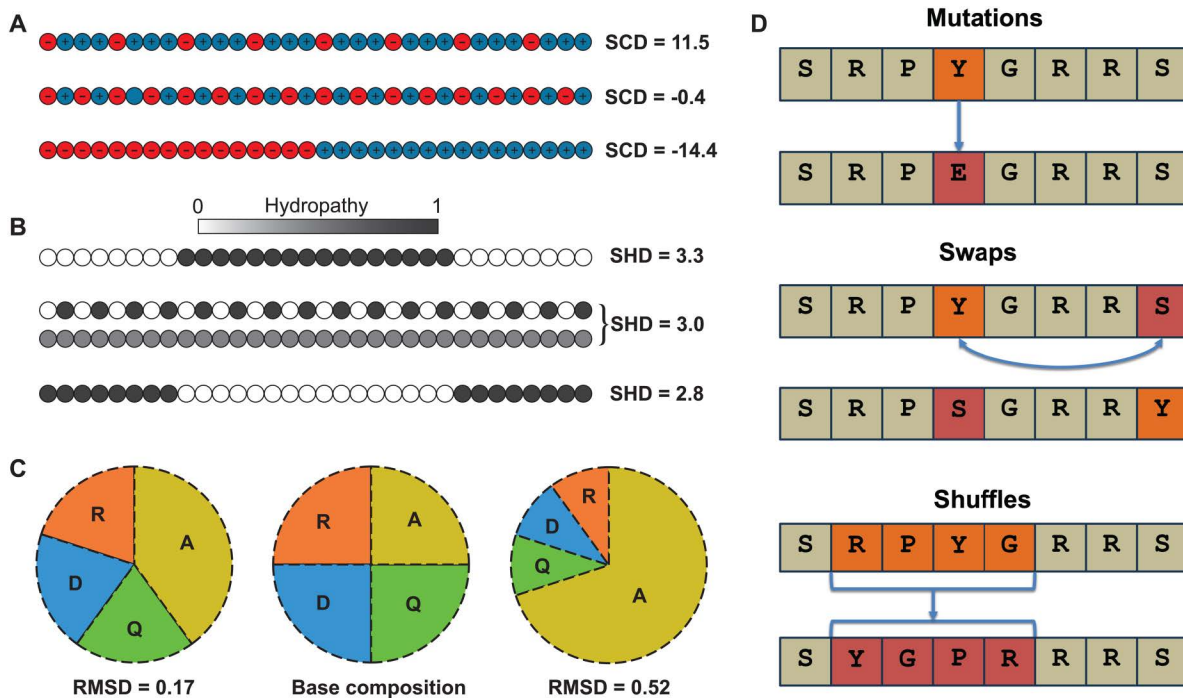


Fig 4. Design parameters and moveset for sequence Monte Carlo algorithm. **A)** SCD is related to the degree of segregation of charged residues, where well-mixed polyampholytes are close to zero, blocky polyampholytes have large negative SCD, and sequences with a significant net charge have a positive SCD. **B)** SHD tends to be higher for sequences having hydrophobic groups more localized toward the middle of the sequence. Typically, IDP sequences are not so extreme, so SHD functions to quantify blockiness of hydrophobic residues primarily, and is somewhat influenced by proximity of hydrophobic groups toward the edges of the sequence. **C)** Composition RMSD is a simple metric of how similar the current composition is from that of the base composition. **D)** Three types of moves are attempted for the MC code, mutation moves change a single residue's type, swap moves swap the positions of two random residues, and shuffle moves randomize a part of the sequence.

<https://doi.org/10.1371/journal.pcbi.1014462.g004>

an attempted 2000 Monte Carlo moves (Fig 5C,D). The composition of the sequence was allowed to fluctuate, but only did so moderately (Fig 5E), staying very close to the original sequence composition. Consequently, the total loss function reached a minimum at ~ 200 moves (Fig 5F). It is important to note that, while these trajectories converged successfully to target values in a low number of moves, this success is only possible because targets were reasonably attainable given the input sequence. For targets well outside the range of the optimized parameter's distribution for the given composition, the algorithm will be unsuccessful or take a significantly longer time to reach it through the gradual mutation of the sequence. It is also possible to optimize just a single parameter, such as SCD (Fig C in S1 Appendix), however we find that other parameters can widely vary and fluctuate throughout the design process (Fig I in S1 Appendix). Depending on one's design goals, this may be disadvantageous.

To evaluate the success of the normalization scheme for patterning parameters, we plotted the energy decomposition for additional design runs of LAF-1 RGG, breaking down the contributions of each parameter by what move type was attempted. In order to minimize drift throughout the course of the simulation, target values were set equal to the initial values of the input sequence (Fig B in S1 Appendix). Comparing between these parameters within a single move type, we can evaluate what each term's average energy contribution was. In this manner, we would expect a successful normalization scheme to keep all contributions within a reasonable range of one another. We find that for SCD, SHD, and SAD, each move type only exhibits variations within an order of magnitude of energy fluctuations (Fig B in S1 Appendix), indicating successful normalization of these parameters.

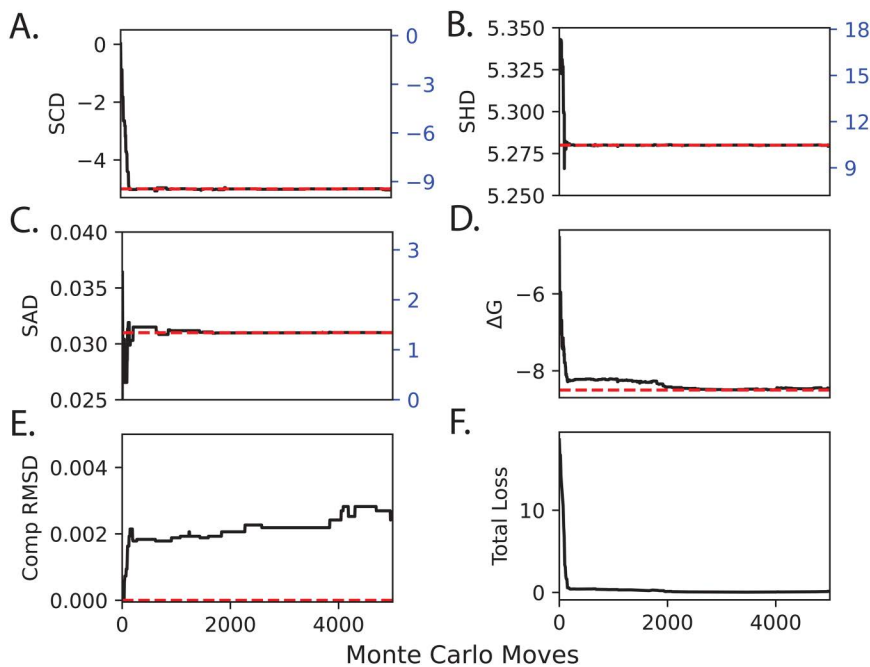


Fig 5. Multiple parameter optimization using Monte Carlo sequence design. (A-E) Trajectory of sequence MC for LAF-1 RGG converging to desirable properties of SCD, SHD, SAD, ΔG , and Composition RMSD, respectively. Blue axes represent the z-score of the constant-composition distribution for the patterning parameters. (F) Total loss progression over 5000 Monte Carlo moves.

<https://doi.org/10.1371/journal.pcbi.1014462.g005>

Efficient and guided design of extreme patterning in sequences

With the identification of significant changes in phase-behavior between protein sequences of fixed-composition, a commonly explored area of interest is the relation between patterning parameters and a protein's propensity for phase separation. In a previous study, we investigated the relationship between the phase-behavior of fixed-composition LAF-1 RGG variants at a wide-range of SCD values [35]. By performing 10 million random shuffles to the wild type sequence with an SCD of 0.6, the variant with the lowest attained SCD only reached -7.3. To recreate this effort using our Monte Carlo approach, we utilize LAF-1 RGG as a reference sequence to which further variants will be designed through shuffling. To keep composition from changing, we disabled mutation moves for this set of sequence designs. The shuffling moves attempted differ from the original study in that they are applied to only segments of the sequence, and their contribution to SCD is evaluated, such that changes that increase SCD are not accepted. This causes the sequence to walk a directed path toward the desired properties. We repeat this process 10 times for LAF-1 RGG variants with three different target SCDs of -5, -10, and -15, which averaged 180, 370, and 800 cycles, respectively, to complete (Fig 6A-C). Each design task, two of which extended significantly past the attainable sequences from the aforementioned paper, was attained in fewer than 1000 shuffle moves. Notably, the calculation of SCD and SHD for millions of sequences can be time consuming and scales poorly with increasing chain length, making it less feasible for longer IDP designs. Although the Monte Carlo algorithm relies on the same shuffling and calculation methodologies, its motivated search was able to reach more significant extremes of LAF-1 RGG's SCD sequence ensemble in a fraction of the iterations. As such, this tool provides an efficient means of designing polypeptide sequences that lie far outside the range of easily-accessible sequences.

We then tested these designed sequences using coarse-grained MD simulations in the HPS-Urry model with slab geometry [45]. The HPS-Urry model is well-suited for this problem as it has successfully been applied to capture single chain properties of a large library of IDPs, and also provide accurate phase diagram predictions for some IDPs with

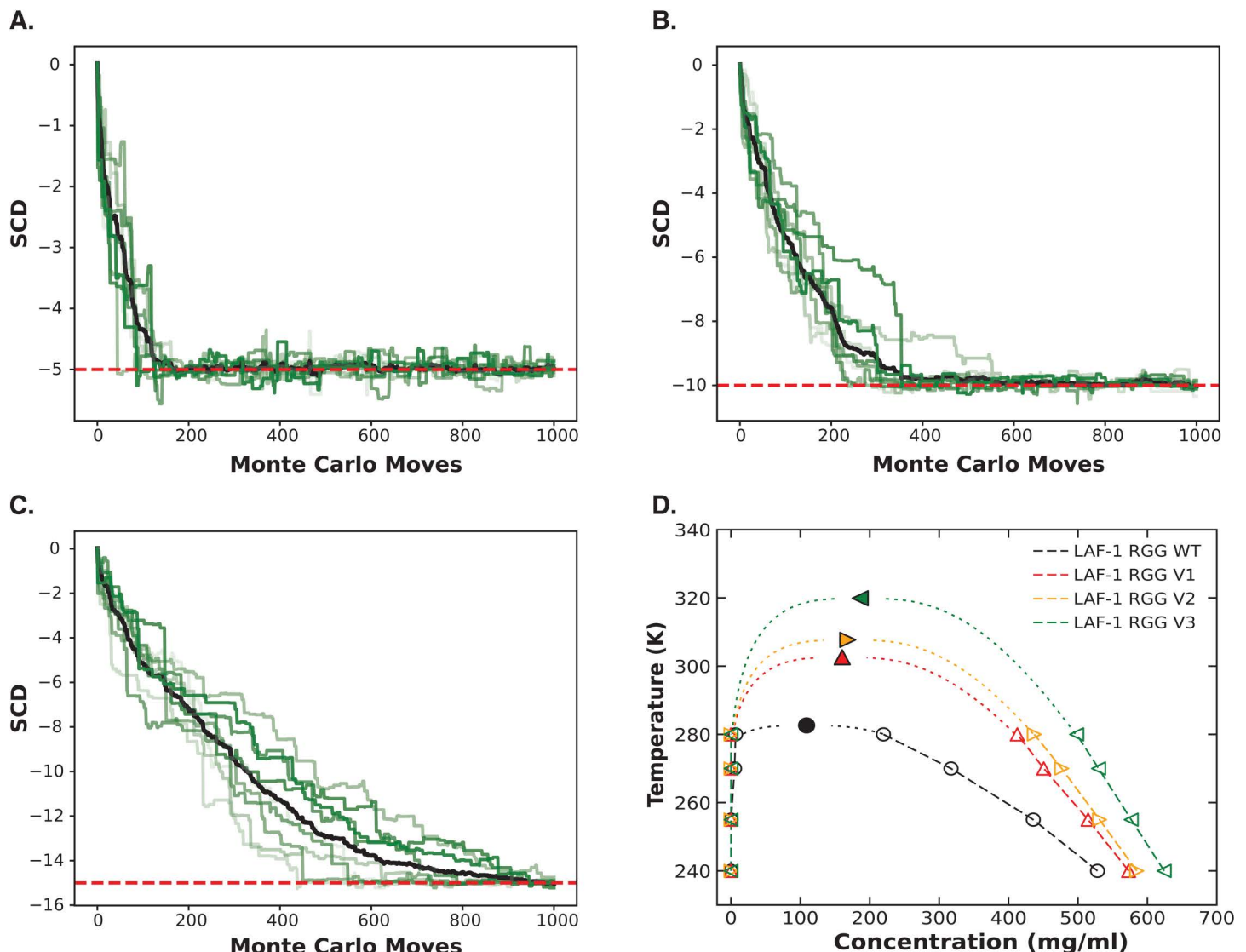


Fig 6. Design and validation of LAF-1 RGG-derived sequences. (A-C) 15 Monte Carlo runs starting with LAF-1 RGG and attempting to create variants with an SCD of -5 (A), -10 (B), and -15 (C). (D) Coexistence curve of LAF-1 RGG WT along with the 3 previously created variants.

<https://doi.org/10.1371/journal.pcbi.1014462.g006>

experimentally-measured phase diagrams [45]. We simulated the WT LAF-1 RGG as well as three designed variants with increasingly negative values of SCD, finding that more negative values of SCD resulted in upward shifts of the binodal phase envelope, indicating stronger self-association and greater phase separation propensity (Fig 6D). We also determine the critical temperature for these variants (Table 1) and find that the observed increase in critical temperature is comparable to that reported for shuffled sequences designed in previous work [35]. The designed sequences are listed in Table C in S1 Appendix.

Beyond critical temperature, we also investigated whether this method of LLPS optimization influences the intrinsic material properties of the simulated condensate. We computed the single-chain translational diffusivity and the R_g auto-correlation time (τ_{R_g}) of LAF-1 RGG WT and variants V1–V3 across the temperatures used to construct the phase diagrams (Fig 7). When plotted against temperature, all four sequences diffuse and re-orient faster at higher temperatures

Table 1. SCD, SHD, SAD, and T_c values for LAF-1 RGG wild type and variants.

Protein name	SCD	SHD	SAD	T_c (K)
LAF-1 RGG WT	0.020	5.339	0.019	282.6
LAF-1 V1	-5.000	5.327	0.031	302.6
LAF-1 V2	-9.997	5.328	0.020	307.7
LAF-1 V3	-15.002	5.349	0.028	319.9

<https://doi.org/10.1371/journal.pcbi.1014462.t001>

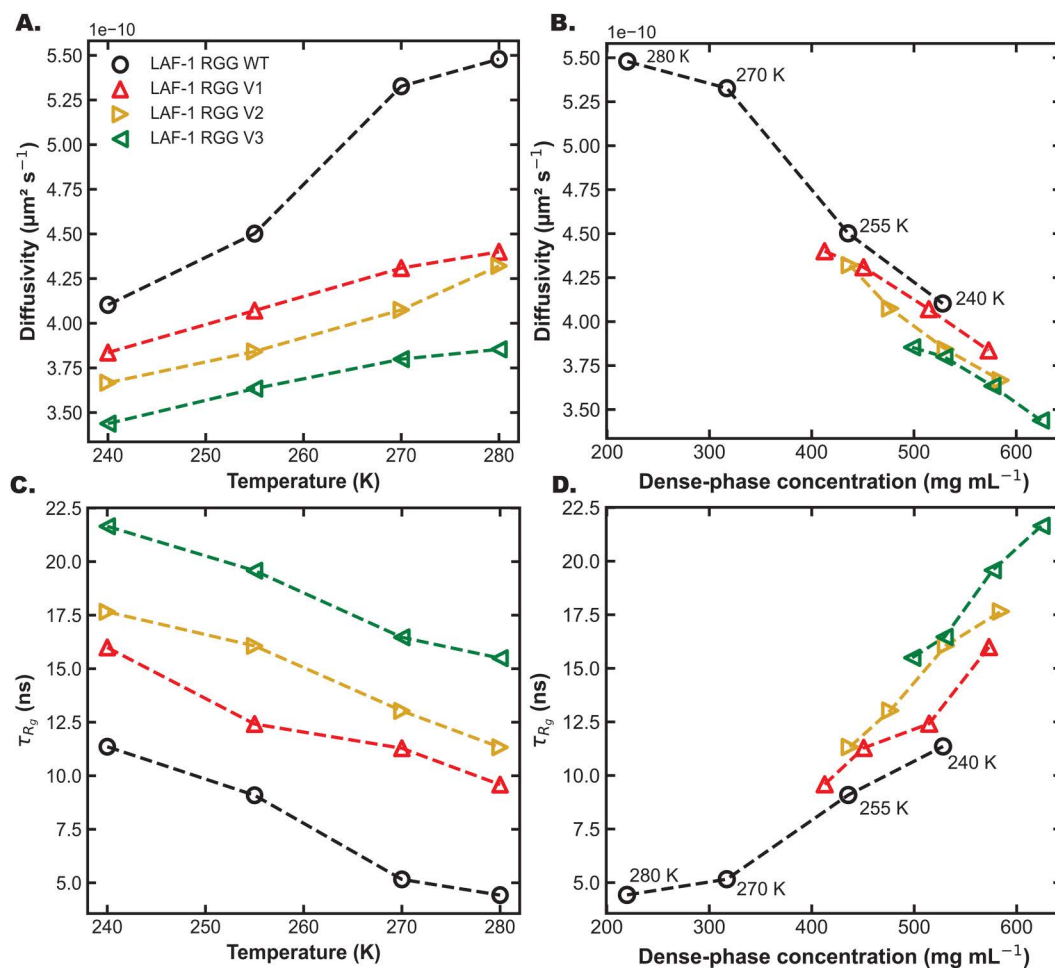


Fig 7. Effect of SCD-targeted optimization on intra-condensate dynamics of LAF-1 RGG. (A) Single-chain translational diffusivity in the dense phase versus temperature for LAF-1 RGG WT and the three designed variants V1, V2, V3 (Table 1). (B) The same diffusivity data plotted against the corresponding dense-phase concentration. (C) R_g autocorrelation time τ_{R_g} versus temperature for the same four sequences. (D) τ_{R_g} versus dense-phase concentration, with isotherms labeled.

<https://doi.org/10.1371/journal.pcbi.1014462.g007>

(Fig 7A,C), with each designed variant diffusing and reconfiguring more slowly than the WT at the same temperature, and showing progressively slower dynamics for sequences with a greater degree of charge segregation. One major contribution to this effect could be the higher density of the more charge segregated variants in the dense phase. Indeed, when re-plotted against the dense-phase concentration (Fig 7B,D), we find that the four sequences fall onto a much closer density-dependent trend, having lower diffusivity and longer τ_{R_g} at higher dense-phase concentration. The designed

sequences deviate only slightly from the WT, indicating that the density is indeed a major determinant of the materials properties of these dense phases. Importantly, slight variations between the different sequences at similar density cannot be explained by thermal effects because the designed variants are simulated at higher temperatures to achieve the same density as WT, but they exhibit slower dynamics and reconfiguration than the WT. Instead, this would likely be due to longer intrachain and interchain contact lifetimes [33].

We also investigated potential contributions of other parameters for influencing this increase in LLPS propensity for the designed variants. We find that T_c correlates most strongly with SCD ($r = -0.97$) and with ΔG ($r = -0.96$), but only weakly with SHD or SAD (Fig E(A) in [S1 Appendix](#)). All of these results agree with previous studies showing that more negative SCD values correspond to more strongly segregated charged residues and such sequences exhibit higher LLPS propensity [21,35,36].

While SCD is the major driver of increased T_c , we also wanted to determine if SHD could be designed to have a greater influence. Thus, we generated 15 additional fixed-composition LAF-1 RGG variants that nearly fully span the accessible range of SCD and SHD values (Fig 8A; Table D in [S1 Appendix](#)). For each of these new variants, we conducted simulations and calculated the binodal phase diagram, and T_c . Using this new dataset, we find that critical temperatures correlate strongly and monotonically with SCD (Fig 8B; $r = -0.895$), while showing essentially no dependence on SHD or SAD individually (Fig H(C,D) in [S1 Appendix](#)). However, when we combine the three parameters in a multilinear regression model of T_c , we find the agreement is moderately improved compared to predicting solely based on SCD (Fig H(A) in [S1 Appendix](#); $r = -0.911$). Our multilinear regression yields the standardized expression

$$T_c^* = -0.789 \cdot \text{SCD}^* - 0.117 \cdot \text{SHD}^* + 0.095 \cdot \text{SAD}^*,$$

where each variable is z-score standardized (mean-centered and scaled by its standard deviation) so that coefficients are directly comparable. The associated feature importance ranking attributes $\sim 79\%$ of the explained variance in T_c to SCD, with SHD and SAD together contributing the remaining $\sim 21\%$ (Fig H(B) in [S1 Appendix](#)). Taken together, these analyses indicate that, for sequences of fixed composition identical to that of LAF-1 RGG, the upward shift in T_c is predominantly controlled by enhanced charge segregation, and only minorly tuned by hydrophathy patterning or aromatic distribution. However, it is likely that in sequences having fewer charged residues, the phase behavior would be less predominantly described by SCD, such as we find with our designs in the following sections (Fig E in [S1 Appendix](#)).

These 15 new designs also allow us to revisit the material properties question on a larger and more diverse set. When colored by SCD, the diffusivity-versus-density (Fig 8C) and τ_{R_g} -versus-density (Fig 8D) curves overlap, indicating that dense-phase concentration, rather than SCD, is a significant determinant of the dynamics. The small differences that remain at a given density would not likely be due to temperature, since sequences with more negative SCD reach that density at a higher temperature, but generally exhibit slower dynamics as we saw previously with the smaller dataset (Fig 6).

The same framework can be used in the opposite direction to generate variants with reduced phase-separation propensity relative to the WT. As an example, sequence `protein14` in the extended LAF-1 RGG library has a T_c of 274.4 K (Table D in [S1 Appendix](#)), below the WT value of 282.6 K, and its coexistence curve lies below that of the WT (Fig G in [S1 Appendix](#)). This demonstrates that the Monte Carlo algorithm is not biased toward enhanced LLPS and can equally well design fixed-composition sequences that are less prone to phase separation. Combining all 18 LAF-1 RGG variants, we constructed an approximate 2D phase-behavior map in the (SCD, SHD) plane (Fig F in [S1 Appendix](#)), in which each sequence is annotated by its T_c . The map reproduces the trend quantified above, T_c varies most strongly with SCD and only weakly with SHD at fixed composition, and provides a practical visual reference for the SCD/SHD ranges that demarcate strongly versus weakly phase-separating LAF-1 RGG variants in our model. We caution that the location of these boundaries is composition and force-field dependent, so for sequences whose composition departs substantially from LAF-1 RGG, the same (SCD, SHD) coordinates may correspond to a different phase regime.

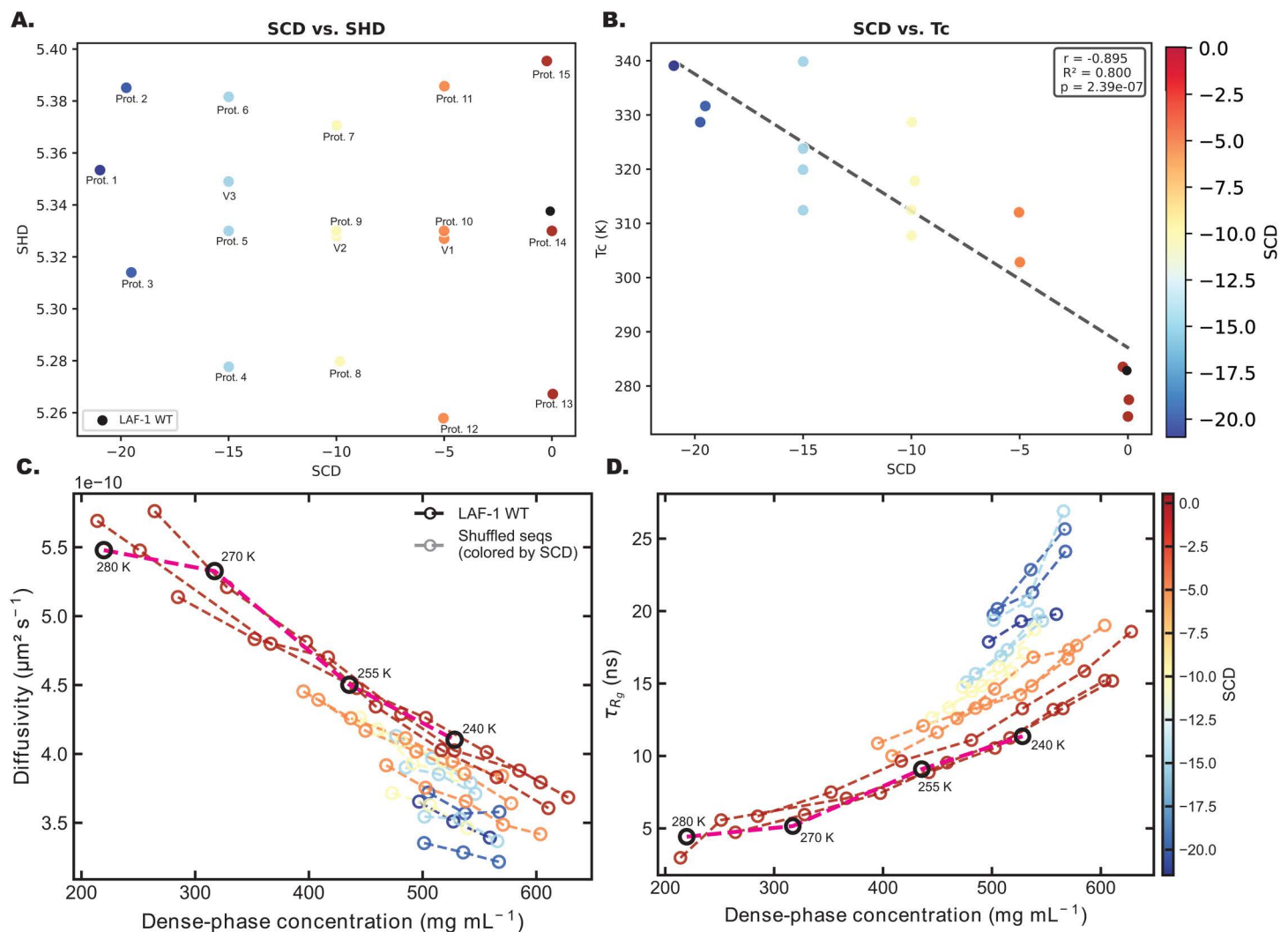


Fig 8. Fifteen additional fixed-composition LAF-1 RGG variants were designed to span a wide range of SCD and SHD values. (A) Coverage of the design library in the (SCD, SHD) plane, with WT shown as a black labeled marker. **(B)** Critical temperature T_c versus SCD. The strong, monotonic dependence of T_c on SCD, identifies SCD as the dominant determinant of LLPS stability across this composition-preserving library. **(C)** Single-chain diffusivity versus dense-phase concentration for the 15 designs at different temperatures, with points colored by SCD. **(D)** τ_{R_g} versus dense-phase concentration, colored by SCD.

<https://doi.org/10.1371/journal.pcbi.1014462.g008>

Adding new features to sequence optimization

Next, we incorporate an IDP's predicted transfer free energy (ΔG) that directly quantifies the degree to which a sequence will phase separate, being related to the log of the IDP's partition coefficient. To do this, we utilize the predictive algorithm from von Bülow et al. which predicts ΔG of phase separation directly from amino acid sequence [39]. This predictive algorithm was developed within an active-learning framework that iteratively combined coarse-grained molecular dynamics simulations with experimental data to train a neural network that maps amino acid sequence directly to the transfer free energy of phase separation (ΔG) and to the saturation concentration (c_{sat}). The training set spans a diverse collection of intrinsically disordered regions (IDRs), and the predictor has been validated against both simulated and experimental LLPS data, including a proteome-wide application to the human IDRome [39]. We note that simulations used in

training of this model were done using the CALVADOS 2 coarse-grained model [46] which employs a somewhat different hydropathy scale from the HPS-Urry scale used in this work. Thus our SHD parameter (and indeed the other patterning parameters) are likely to correlate somewhat with the ΔG predictions, as their influence on LLPS is likely to have been captured. Our intention in incorporating the ΔG predictions is to include a parameter that has some direct validation from experimental data to serve as a rough approximation of overall phase separation propensity, since none of our specific patterning parameters do a perfect job at predicting LLPS propensity in all cases (Fig 8B; Fig E in S1 Appendix; Fig H in S1 Appendix).

We also introduce composition root-mean-square deviation (compRMSD) as a parameter that quantifies how much a designed sequence deviates from a reference sequence composition while allowing for Mutation moves.

$$\text{compRMSD} = \sqrt{\frac{1}{20} \sum_{i=1}^{20} (f_i - f_{\text{target},i})^2} \quad (14)$$

where f_i is the fraction of residues in the sequence that are amino acid i . Unlike fixed-composition approaches, enabling mutation moves and restraining overall composition drift through compRMSD enables controlled variation, making it particularly valuable for designing IDPs that retain similar composition to the template sequences, but may differ somewhat in size or total length. The weight of the compRMSD component can also be adjusted to modify the importance of preserving sequence composition. In the absence of mutation moves, FUS LC cannot achieve significantly lower ΔG values than the WT sequence, even with our optimization. With mutations enabled, it rapidly reaches the target value (Fig I in S1 Appendix).

By integrating these two parameters into our Monte Carlo sequence design approach, we designed two variants of the FUS LC wild-type (WT) protein: Variant 1 (V1) and Variant 2 (V2). As shown in the phase diagrams (Fig 9), both variants exhibit a lower ΔG compared to WT, correlating with an increased propensity for phase separation. These results demonstrate how incorporating additional predictive metrics, such as transfer free energy, allows for the rational design of IDP sequences with enhanced phase behavior. We find that, unlike with the LAF-1 RGG like sequences, the T_c for these FUS LC-derived sequences are predicted well from SHD and SAD, likely owing to the lower content of charged residues.

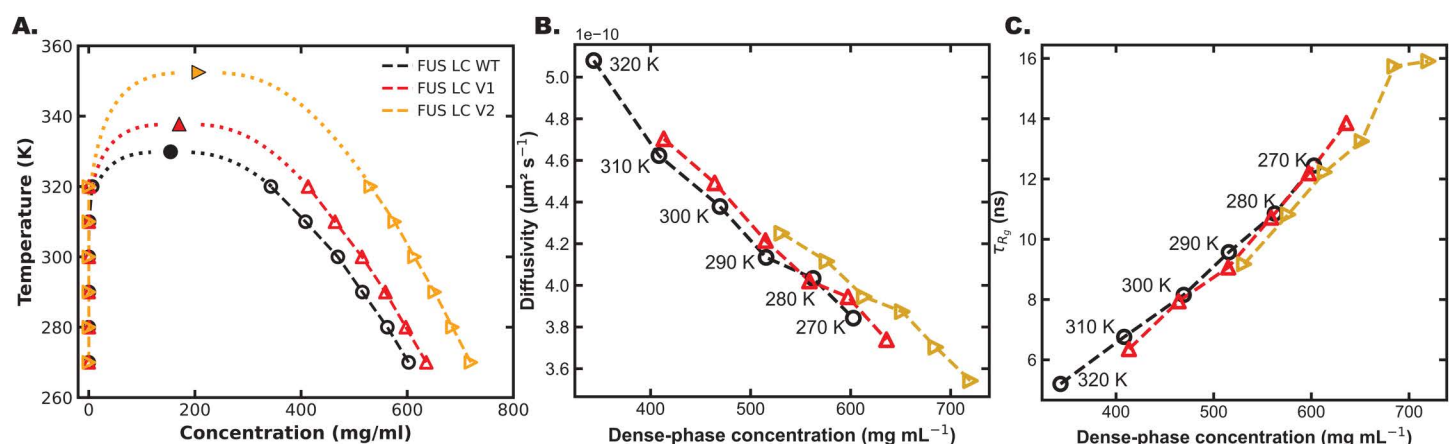


Fig 9. Phase behavior and intra-condensate dynamics of FUS LC WT and designed variants. (A) Temperature–concentration phase diagrams obtained from slab simulations for FUS LC WT and the two designed variants V1 and V2 (Table 2). Both variants exhibit upward-shifted binodals consistent with their more favorable predicted ΔG . (B) Single-chain translational diffusivity within the dense phase plotted against dense-phase concentration. (C) R_g autocorrelation time τ_{R_g} versus dense-phase concentration.

<https://doi.org/10.1371/journal.pcbi.1014462.g009>

In summary, the designed FUS LC WT variants maintain compositions that are reasonably close to the wild type, with compRMSD values of 0.008 and 0.012 for variants V1 and V2, respectively (Table 2). By keeping compositional changes minimal, we ensure that the observed differences in predicted phase behavior primarily arise from altered residue patterning rather than shifts in overall amino acid content. The designed sequences are listed in Table C in S1 Appendix.

Beyond the critical temperature, we examined the intra-condensate dynamics of FUS LC WT and the two designed variants by computing single-chain translational diffusivity and the R_g autocorrelation time τ_{R_g} . When these observables are plotted against temperature (Fig J in S1 Appendix), the variants are systematically slower than the WT at matched temperature, diffusivity decreases and τ_{R_g} increases in the order WT $\rightarrow$ V1 $\rightarrow$ V2, mirroring their increasing phase-separation propensity. When the same data are instead plotted against dense-phase concentration (Fig 9B,C), however, the three sequences fall onto a single common trend diffusivity decreases and τ_{R_g} increases monotonically with dense-phase concentration. Notably, at matched dense-phase concentration the WT and the two variants are essentially indistinguishable unlike with the LAF-1 RGG derived variants. We note that because mutations were allowed (with compRMSD constraining but not eliminating compositional drift), the optimization raised the aromatic content from 24 residues in WT (14.7%) to 35 in V2 (21.5%), introducing 5 tryptophans and 2 phenylalanines absent from the WT and increased overall hydrophobic content from 24.5% to 31.3%, while the aromatic-decoration parameter SAD rose nearly threefold (0.065 $\rightarrow$ 0.178; Table 2). Thus, the alterations to the sequence are more distributed throughout the length of the sequence and varied properties are changing.

Additionally, since FUS LC contains essentially no charged residues (≤ 3 in any variant), changes to aromatic and hydrophobic content and clustering is what drives the variants to phase separate at higher temperatures and to reach higher dense-phase concentrations. The minor variations in SCD are not enough to explain variations in phase separation propensity in this case (Fig E(B) in S1 Appendix) providing a contrasting example to the previously designed LAF-1 RGG variants. In summary, we suggest that the total charge content is the major determinant of how much the phase behavior depends on SCD (Fig E in S1 Appendix).

Leveraging design strategy for short peptides to mimic larger IDPs

It is well-established that shorter chain lengths, particularly in molecules with lower valence, are less prone to phase separation [28,47–50]. However, recent work has shown that even short peptide sequences are capable of phase separation, usually owing to significant charge or aromatic content [27,51,52]. Short peptides have several advantages for experimental and computational research, including that simulations will converge more quickly and likely require fewer computing resources. We employed this method to design peptides that replicate the phase separation properties of FG Nucleoporin 153 (Nup153), a 1475-residue protein which is a component of the nuclear pore complex and implicated in binding to HIV-1 viral capsid proteins [53]. To demonstrate an extreme example of reducing protein sequence size, we have designed several miniature versions of Nup153, or mini-NUP sequences of length 30 (Table 3).

Since we have shortened the Nup153 sequence by a factor of 50, we expect that keeping identical composition will significantly reduce the phase separation propensity to a degree that it will not phase separate at accessible experimental conditions. Thus, we generated multiple variants of the peptide to have increasing likelihood of phase separation. To achieve the desired values of SHD, ΔG , and other sequence parameters, we increased the number of aromatic residues

Table 2. SCD, SHD, SAD, compRMSD, and ΔG values for FUS LC WT wild type and variants.

Protein name	SCD	SHD	SAD	compRMSD	ΔG (k _B T)
FUS LC WT	0.039	5.922	0.065	0	-5.9
FUS WT V1	-0.121	5.977	0.101	0.008	-7.9
FUS WT V2	0.000	6.187	0.178	0.012	-10.1

<https://doi.org/10.1371/journal.pcbi.1014462.t002>

Table 3. SCD, SHD, SAD, compRMSD, and ΔG values for designed mini-NUP sequences. Values in parentheses are normalized values.

Sequence	SCD	SHD	SAD	compRMSD	ΔG ($k_B T$)
mini-NUP seq1	-0.80(-8.18)	3.89(-1.98)	0.23(-2.45)	0.044	-6.39
mini-NUP seq2	-0.72(-7.39)	3.97(-2.05)	0.32(-2.12)	0.047	-7.66
mini-NUP seq3	-0.25(-4.71)	4.13(-3.66)	0.32(-2.11)	0.049	-7.84

<https://doi.org/10.1371/journal.pcbi.1014462.t003>

by introducing an additional 10% tyrosine and phenylalanine into the original FG Nucleoporin, Nup153. To quantify the compositional changes, we computed the compRMSD, which captures the difference in amino acid composition between the original sequence with added aromatic residues and the final designed sequences.

We conducted MD simulations of these peptide sequences to determine the phase separation propensity of each and find that we achieve a range of phase behaviors for each sequence (Fig 10). The sequences of designed Nup153-like peptides are included in Table C in S1 Appendix.

As with the LAF-1 RGG and FUS LC designs, we examined the intra-condensate dynamics of the three mini-NUP peptides by computing single-chain diffusivity and the R_g autocorrelation time τ_{R_g} as a function of temperature (Fig K in S1 Appendix) and dense-phase concentration (Fig 10B,C). One feature distinguishes these short peptides from the LAF-1 RGG case, T_c increases across the series seq1 $\rightarrow$ seq2 $\rightarrow$ seq3 (Table 3) while SCD becomes less negative ($-0.80 \rightarrow -0.72 \rightarrow -0.25$), the opposite of the trend established for LAF-1 RGG at fixed composition (Fig 8B and Fig E(A) in S1 Appendix), where T_c anticorrelates strongly with SCD. Inspection of the sequences (Table C in S1 Appendix) explains this reversal. In all three peptides the charges are strongly segregated—positively charged residues clustered near positions 3–6 and negatively charged residues near positions 22–27. The relevant difference between them is therefore not the patterning of charges but their total number, seq1 and seq2 each carry five charged residues, while seq3 carries only three. Because the design framework allowed mutations, the optimization satisfied the targeted increases in ΔG and SHD for seq3 by replacing two charged residues with aromatic/hydrophobic ones (aromatic fraction rises from 33.3% in seq1 to 36.7% in seq3, hydrophobic content from 50.0% to 60.0%, SAD rises from 0.23 to 0.32; Table 3). The less negative SCD in seq3 therefore does not indicate weaker charge patterning, the charges remain fully segregated,

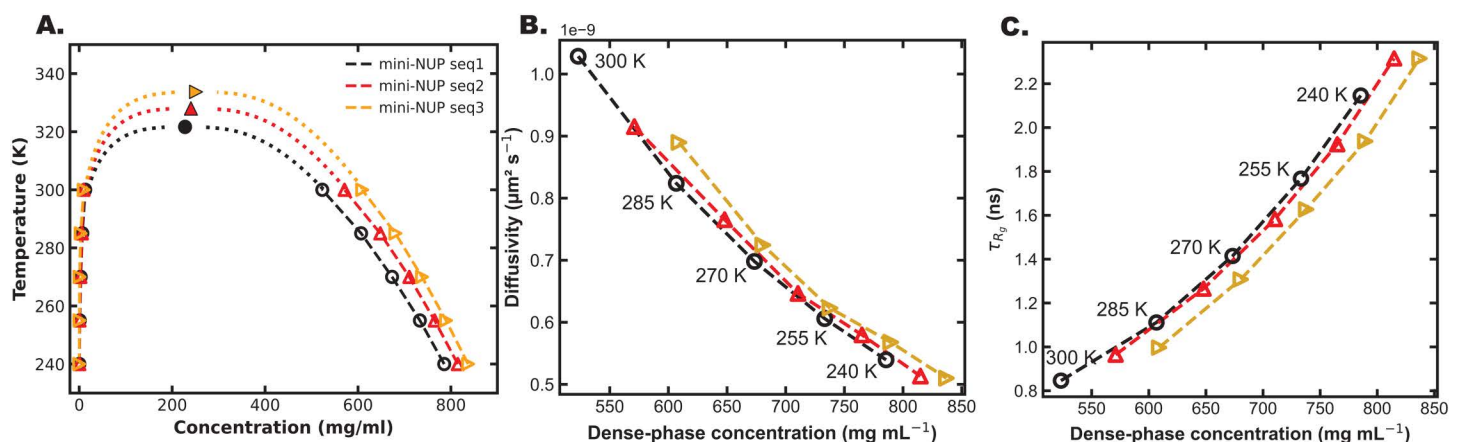


Fig 10. Phase behavior and intra-condensate dynamics of designed mini-NUP peptides. (A) Temperature–concentration phase diagrams obtained from slab simulations for the three 30-residue mini-NUP designs (seq1, seq2, seq3; Table 3). Despite all three peptides being only 30 residues long, the design framework yields variants that phase-separate across a broad range of conditions. (B) Single-chain translational diffusivity in the dense phase versus dense-phase concentration; isotherms (240, 255, 270, 285, 300 K) are labeled. (C) R_g autocorrelation time τ_{R_g} versus dense-phase concentration, with the same isotherms labeled.

<https://doi.org/10.1371/journal.pcbi.1014462.g010>

but rather fewer charges available to pattern, with the corresponding positions occupied by more cohesive residues. This compositional shift explains the reversed T_c -vs-SCD trend, as the gain in aromatic/hydrophobic cohesion outweighs the small reduction in charge content. The intra-condensate dynamics, by contrast, behave just as they do for the LAF-1 RGG and FUS LC systems: when diffusivity and τ_{R_g} are plotted against temperature (Fig K in [S1 Appendix](#)), seq3 is the slowest and seq1 the fastest, mirroring their order of phase-separation propensity, but when the same observables are plotted against dense-phase concentration ([Fig 10B,C](#)), the three peptides fall onto a single common trend and are essentially indistinguishable at matched dense-phase concentration and the small differences that remain may partly reflect temperature, since seq3 reaches a given dense-phase concentration at a higher temperature than seq1. As with the other systems, seq3 appears slower at matched temperature only because it reaches higher dense-phase concentrations, extending furthest along the common dynamics curve rather than occupying a distinct, more viscous material state. Taken together with the LAF-1 RGG and FUS LC results, this case illustrates two general points. First, intra-condensate dynamics are governed primarily by dense-phase concentration across all of our designs, so the optimization shifts the phase boundary without introducing a detectable change in material properties at matched density. Second, when mutations are enabled, SCD must be interpreted alongside composition, because it can change either through patterning at fixed composition (the LAF-1 RGG regime, where more negative SCD implies stronger LLPS) or through changes in charge content (the mini-NUP regime), only the former is captured by the fixed-composition heuristic.

Methods

Simulations of designed IDP sequences were conducted using OpenMM simulation software [\[54\]](#) using the HPS-Urry coarse-grained model [\[45\]](#). Each IDP was simulated at a range of temperatures where both a dense phase and a dilute phase were present to gather sufficient statistics to calculate the binodal coexistence curve. We ran simulations for at least 1 μ s per temperature case using Langevin dynamics. For all trajectories, the largest cluster was centered in the simulation box using an autocorrelation function of the instantaneous density profile of each simulation frame [\[55\]](#). The critical point of each sequence was determined using a fitting function, assuming the universality of the 3D Ising model as in previous work [\[49\]](#).

Force field and interaction potentials

The HPS-Urry coarse-grained model was employed, where each amino acid is represented by a single bead located at the position of its C_α atom. The total interaction energy (U_{total}) in this model is a sum of bonded and non-bonded terms:

$$U_{\text{total}} = \sum_{i=1}^{N-1} k_b (r_{i,i+1} - r_0)^2 + \sum_{i,j} \phi_{ij}^{\text{HPS}} + \sum_{i,j} \phi_{ij}^{\text{elec}}$$

The first term models the bonded interactions between adjacent residues as a harmonic potential with a spring constant $k_b = 20 \text{ kJ}/\text{\AA}^2$ and an equilibrium bond length of $r_0 = 3.82 \text{ \AA}$. The non-bonded interactions are composed of hydrophobic (ϕ_{ij}^{HPS}) and electrostatic (ϕ_{ij}^{elec}) terms.

Hydrophobic interactions. These interactions are described by a modified Ashbaugh-Hatch potential [\[56\]](#), which is based on the Lennard-Jones (LJ) potential:

$$\phi_{ij}(r) = \begin{cases} \phi_{ij}^{\text{LJ}}(r) + (1 - \lambda_{ij})\epsilon, & \text{for } r \leq 2^{1/6} \sigma_{ij} \\ \lambda_{ij} \phi_{ij}^{\text{LJ}}(r), & \text{for } r > 2^{1/6} \sigma_{ij} \end{cases}$$

where the standard LJ potential is defined as:

$$\phi_{ij}^{\text{LJ}}(r) = 4\epsilon \left[\left(\frac{\sigma_{ij}}{r} \right)^{12} - \left(\frac{\sigma_{ij}}{r} \right)^6 \right]$$

The energy scale is set by $\epsilon = 0.2$ kcal/mol. The residue-specific parameters for size (σ_i) and hydrophathy/stickiness (λ_i) are combined using an arithmetic mean: $\sigma_{ij} = (\sigma_i + \sigma_j)/2$ and $\lambda_{ij} = (\lambda_i + \lambda_j)/2 - \Delta$, where $\Delta = 0.08$.

Electrostatic Interactions. Electrostatic forces are modeled using a Debye-Hückel potential to account for charge screening in solution [57]:

$$\phi_{ij}^{\text{el}}(r) = \frac{q_i q_j}{4\pi D r} e^{-\kappa r}$$

Here, q_i and q_j are the charges of the interacting residues. The dielectric constant of the solvent (D) was set to 80 to represent water, and the inverse Debye screening length (κ) was set to 1 nm^{-1} , corresponding to a monovalent salt concentration of approximately 100 mM.

Analysis: translational diffusivity and R_g relaxation time. Single-chain translational diffusivity (D) and the radius-of-gyration autocorrelation time (τ_{R_g}) were computed from the same slab trajectories used to construct the binodal coexistence curves. All analysis was restricted to chains residing in the dense phase, identified as chains within the central high-density region of the slab. For the translational diffusivity, we tracked the position of the first $C\alpha$ bead of each chain and computed the mean-squared displacement (MSD) as a function of lag time t :

$$\text{MSD}(t) = \left\langle |\mathbf{r}(t_0 + t) - \mathbf{r}(t_0)|^2 \right\rangle \quad (15)$$

where the average is taken over all chains and all time origins t_0 , and displacements are evaluated under the minimum-image convention to account for periodic boundary conditions. Lag times up to 40 ns were computed (200 frames at a frame interval of 0.2 ns). The diffusion coefficient was extracted by fitting the MSD to the three-dimensional Einstein relation,

$$\text{MSD}(t) = 6Dt, \quad (16)$$

over the linear regime spanning 2–10 ns, giving D in units of $\mu\text{m}^2 \text{ s}^{-1}$. For the R_g autocorrelation time, the radius of gyration $R_g(t)$ was computed for each chain at every saved frame over the full production trajectory. The normalized autocorrelation function (ACF) was then calculated for each chain individually using a fast Fourier transform (FFT) algorithm:

$$C(t) = \frac{\langle \delta R_g(t_0) \delta R_g(t_0 + t) \rangle}{\langle \delta R_g^2 \rangle}, \quad \delta R_g(t) = R_g(t) - \langle R_g \rangle \quad (17)$$

and averaged over all chains in the simulation. Lag times spanning the full production trajectory (up to $\sim 1 \mu\text{s}$) were used. The relaxation time τ_{R_g} was extracted by fitting the chain-averaged ACF to a single exponential,

$$C(t) = \exp(-t/\tau_{R_g}), \quad (18)$$

restricted to the region where $C(t) \geq 0.05$ to avoid the noisy long-lag tail. All trajectory analysis was performed using MDAnalysis and MDTraj.

Conclusion

By defining trends between composition-specific terms and patterning parameters, we have successfully established a method of approximating the properties of a given IDP's constant-composition ensemble, and thus derived a normalization

scheme for the common patterning parameters, SCD and SHD. Using this normalization scheme, we then leverage this into a method of comparing changes between different patterning parameters, putting them on an equal playing field in the design space. We test this out with an IDP sequence design Monte Carlo algorithm which can simultaneously optimize multiple design criteria. Existing literature linking patterning parameters and LLPS further allows this Monte Carlo to represent a means of tuning the phase behavior of a given IDP by creating variants. We first demonstrate that our code can design extreme outliers of specific patterning parameters such as super charge-segregated LAF-1 RGG variants and, using an extended library of 15 additional fixed-composition LAF-1 RGG designs together with a multilinear regression of T_c on standardized (SCD, SHD, SAD), we show that for this charge-rich system the critical temperature is overwhelmingly controlled by charge patterning (Fig 8, Fig H in S1 Appendix). We apply this method to design variants of FUS that are more prone to LLPS than the wild-type, focusing primarily on predicted ΔG of phase separation. We also design mini-NUP variants that maintain similar composition to the base sequence, Nup153, while scaling the importance of perfect compositional similarity to enable design of more “sticky” sequences more capable of phase separation, offsetting the significantly reduced size of the sequence.

Beyond critical temperatures, we further examined intra-condensate dynamics for each designed system and find that the LAF-1 RGG variants preserve the WT mobility density relationship at matched dense-phase concentration, whereas the FUS LC and mini-NUP variants for which mutations were enabled and composition was allowed to drift exhibit lower diffusivity and longer τ_{R_g} at matched density. This indicates that the same framework can be used either to shift the phase boundary alone or to jointly tune the phase boundary and the material properties of the resulting condensate.

Looking forward, this work can be supplemented by inclusion of multiple other parameters for optimization, such as: secondary structure predictors, disorder predictors or any other sequence-based parameter. Furthermore, as the link between patterning parameters of an IDP and the properties of its resultant condensate continue to be elucidated, this work could incorporate condensate properties more directly into its loss function rather than only the general propensity for phase separation. For example, as SAD seems to indicate a more gel-like condensate, this work could be improved to give users better control over designed condensate properties by giving additional guidance into what parameters of SAD may correlate with liquid-like and gel-like condensates. The dynamics analyses introduced for the FUS LC and mini-NUP designs in this work are a first step in that direction, and we anticipate that pairing them with the experimental measurements outlined above will be the most effective route to validating and refining these material-property aware design criteria. Since the present study is entirely computational, we present the designed variants as testable hypotheses. Finally, we emphasize that design of fully shuffled or randomized IDP sequences could impose other unexpected changes to the sequence behavior, perhaps beyond what can be predicted by our methodology. Thus it is important to have a good understanding of the base protein sequence, its overall composition and patterning properties as well as any important motifs or domains that might be influential to its function before attempting to make drastic perturbations to the sequence.

Supporting information

S1 Appendix. Single supporting information document containing all supporting figures (Fig A–Fig K) and tables (Table A–Table D). Component captions are listed below. **Fig A.** Correlation of different predictive parameters with ensemble-calculated mean and standard deviation of SCD. **Fig B.** Correlations of the patterning terms SCD, SHD, and SAD for 10^5 shuffles of the IDPs FUS LC, LAF-1 RGG, and NUP-153. **Fig C.** FUS LC Monte Carlo runs with no mutations. Multiple parameters are tracked but the trajectories only optimize for targets of the maximum (blue) and minimum (green) possible SCD values. **Fig D.** Energy decomposition of five parameters with respect to mutations, swaps, and shuffles separately. These also include unaccepted Monte Carlo moves. The range of values observed for the three patterning parameters are within an a few-fold difference for each move type, indicating success in the normalization scheme for

application to MC sequence design. **Fig E.** Per-system correlations between critical temperature and sequence patterning metrics for the designed variants. (A) LAF-1 RGG variants T_c plotted against SCD, SHD, SAD, and the predicted ΔG . Pearson correlation coefficients are indicated in each panel. (B) FUS LC variants. (C) The three mini-NUP designs. **Fig F.** Approximate 2D phase-behavior map for fixed-composition LAF-1 RGG designs. All 18 LAF-1 RGG variants (WT plus V1–V3 from [Table 1](#) and the 15 additional designs in Table D) are placed at their (SCD, SHD) coordinates and annotated by their simulated critical temperature. **Fig G.** LAF-1 RGG variant with attenuated phase separation. Coexistence curves obtained from slab simulations for LAF-1 RGG WT ($T_c = 282.6$ K) and `protein14` (Table D; $T_c = 274.4$ K), a fixed-composition variant designed by the same Monte Carlo procedure. **Fig H.** Multilinear regression of critical temperature on patterning features for LAF-1 RGG designs. (A) Predicted versus observed T_c across the 18 LAF-1 RGG variants from a multilinear regression on the z-score standardized features. (B) Relative feature importance derived from the standardized coefficients. (C) T_c versus SHD. (D) T_c versus SAD. **Fig I.** FUS LC Monte Carlo runs optimizing for only ΔG with (A) no mutations and (B) mutations enabled. **Fig J.** Temperature dependence of intra-condensate dynamics for FUS LC variants. (A) Single-chain translational diffusivity in the dense phase versus temperature for FUS LC WT, V1, and V2. (B) R_g autocorrelation time τ_{R_g} versus temperature for the same three sequences. **Fig K.** Temperature dependence of intra-condensate dynamics for the mini-NUP designs. (A) Single-chain translational diffusivity within the dense phase versus temperature for mini-NUP `seq1`, `seq2`, and `seq3`. (B) R_g autocorrelation time τ_{R_g} versus temperature for the same three sequences. **Table A.** Urry normalized hydropathy values used for each amino acid. **Table B.** List of amino acid sequences for real IDPs used in initial shuffling study. **Table C.** List of designed sequences used in this work in MD simulations of phase separation. **Table D.** List of 15 additional designed LAF-1 RGG variants generated to span the (SCD, SHD) design space at fixed amino acid composition.

(PDF)

Acknowledgments

We thank José Villegas and Ben Schuster for useful discussions.

Author contributions

Conceptualization: Arjun Singh, Gregory L Dignon.

Data curation: Arjun Singh, Ali I Ukperaj, Gabriel F Porto, Gregory L Dignon.

Formal analysis: Arjun Singh, Ali I Ukperaj, Gabriel F Porto, Gregory L Dignon.

Funding acquisition: Gregory L Dignon.

Investigation: Arjun Singh, Ali I Ukperaj, Gabriel F Porto, Gregory L Dignon.

Methodology: Arjun Singh, Ali I Ukperaj.

Project administration: Gregory L Dignon.

Resources: Gregory L Dignon.

Software: Arjun Singh, Ali I Ukperaj, Gabriel F Porto.

Supervision: Gregory L Dignon.

Validation: Arjun Singh, Ali I Ukperaj, Gabriel F Porto.

Visualization: Arjun Singh, Gabriel F Porto.

Writing – original draft: Arjun Singh, Ali I Ukperaj.

Writing – review & editing: Arjun Singh, Ali I Ukperaj, Gabriel F Porto, Gregory L Dignon.

References

- Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol.* 2017;18(5):285–98. <https://doi.org/10.1038/nrm.2017.7> PMID: [28225081](#)
- Kelley FM, Ani A, Pinlac EG, Linders B, Favetta B, Barai M, et al. Controlled and orthogonal partitioning of large particles into biomolecular condensates. *Nat Commun.* 2025;16(1):3521. <https://doi.org/10.1038/s41467-025-58900-5> PMID: [40229261](#)
- Beutel O, Maraschini R, Pombo-García K, Martin-Lemaitre C, Honigsmann A. Phase Separation of Zonula Occludens Proteins Drives Formation of Tight Junctions. *Cell.* 2019;179(4):923–36.e11. <https://doi.org/10.1016/j.cell.2019.10.011> PMID: [31675499](#)
- Wang Y, Jiang J, Xiong Q, Li S, Shao J, Xie M, et al. Programmable solid-state condensates for spatiotemporal control of mammalian gene expression. *Nat Chem Biol.* 2025;21(9):1457–66. <https://doi.org/10.1038/s41589-025-01860-0> PMID: [40087540](#)
- Zamudio AV, Dall'Agnese A, Henninger JE, Manteiga JC, Afeyan LK, Hannett NM, et al. Mediator Condensates Localize Signaling Factors to Key Cell Identity Genes. *Mol Cell.* 2019;76(5):753–66.e6. <https://doi.org/10.1016/j.molcel.2019.08.016> PMID: [31563432](#)
- Strom AR, Emelyanov AV, Mir M, Fyodorov DV, Darzacq X, Karpen GH. Phase separation drives heterochromatin domain formation. *Nature.* 2017;547(7662):241–5. <https://doi.org/10.1038/nature22989> PMID: [28636597](#)
- Riback JA, Zhu L, Ferrolino MC, Tolbert M, Mitrea DM, Sanders DW, et al. Composition-dependent thermodynamics of intracellular phase separation. *Nature.* 2020;581(7807):209–14. <https://doi.org/10.1038/s41586-020-2256-2> PMID: [32405004](#)
- Lu Y, Wu T, Gutman O, Lu H, Zhou Q, Henis YI, et al. Phase separation of TAZ compartmentalizes the transcription machinery to promote gene expression. *Nat Cell Biol.* 2020;22(4):453–64. <https://doi.org/10.1038/s41556-020-0485-0> PMID: [32203417](#)
- Cai D, Feliciano D, Dong P, Flores E, Gruebele M, Porat-Shliom N, et al. Phase separation of YAP reorganizes genome topology for long-term YAP target gene expression. *Nat Cell Biol.* 2019;21(12):1578–89. <https://doi.org/10.1038/s41556-019-0433-z> PMID: [31792379](#)
- Soto C, Pritzkow S. Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases. *Nat Neurosci.* 2018;21(10):1332–40. <https://doi.org/10.1038/s41593-018-0235-9> PMID: [30250260](#)
- Sönmez A, Mustafa R, Ryll ST, Tuorto F, Wacheul L, Ponti D, et al. Nucleolar stress controls mutant Huntington toxicity and monitors Huntington's disease progression. *Cell Death Dis.* 2021;12(12):1139. <https://doi.org/10.1038/s41419-021-04432-x> PMID: [34880223](#)
- Lester E, Ooi FK, Bakkar N, Ayers J, Woerman AL, Wheeler J, et al. Tau aggregates are RNA-protein assemblies that mislocalize multiple nuclear speckle components. *Neuron.* 2021;109(10):1675–91.e9. <https://doi.org/10.1016/j.neuron.2021.03.026> PMID: [33848474](#)
- Jeon S, Jeon Y, Lim J-Y, Kim Y, Cha B, Kim W. Emerging regulatory mechanisms and functions of biomolecular condensates: implications for therapeutic targets. *Signal Transduct Target Ther.* 2025;10(1):4. <https://doi.org/10.1038/s41392-024-02070-1> PMID: [39757214](#)
- Banani SF, Afeyan LK, Hawken SW, Henninger JE, Dall'Agnese A, Clark VE, et al. Genetic variation associated with condensate dysregulation in disease. *Dev Cell.* 2022;57(14):1776–88.e8. <https://doi.org/10.1016/j.devcel.2022.06.010> PMID: [35809564](#)
- Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. *Curr Opin Struct Biol.* 2008;18(6):756–64. <https://doi.org/10.1016/j.sbi.2008.10.002> PMID: [18952168](#)
- Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu RV. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc Natl Acad Sci U S A.* 2010;107(18):8183–8. <https://doi.org/10.1073/pnas.0911107107> PMID: [20404210](#)
- Uversky VN. Natively unfolded proteins. *Unfolded Proteins From Denatured to Intrinsically Disordered.* 2008. p. 237–94.
- Uversky VN, Gillespie JR, Fink AL. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins: Struct Funct Bioinf.* 2000;41(3):415–27.
- Das RK, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci U S A.* 2013;110(33):13392–7. <https://doi.org/10.1073/pnas.1304749110> PMID: [23901099](#)
- Sawle L, Ghosh K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J Chem Phys.* 2015;143(8):085101. <https://doi.org/10.1063/1.4929391> PMID: [26328871](#)
- Lin Y-H, Chan HS. Phase Separation and Single-Chain Compactness of Charged Disordered Proteins Are Strongly Correlated. *Biophys J.* 2017;112(10):2043–6. <https://doi.org/10.1016/j.bpj.2017.04.021> PMID: [28483149](#)
- Martin EW, Holehouse AS, Peran I, Farag M, Incicco JJ, Bremer A, et al. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science.* 2020;367(6478):694–9. <https://doi.org/10.1126/science.aaw8653> PMID: [32029630](#)
- Kilgore HR, Chinn I, Mikhael PG, Mitnikov I, Van Dongen C, Zylberberg G, et al. Protein codes promote selective subcellular compartmentalization. *Science.* 2025;387(6738):1095–101. <https://doi.org/10.1126/science.adq2634> PMID: [39913643](#)
- Pesce F, Bremer A, Tesei G, Hopkins JB, Grace CR, Mittag T, et al. Design of intrinsically disordered protein variants with diverse structural properties. *Sci Adv.* 2024;10(35):eadm9926. <https://doi.org/10.1126/sciadv.adm9926> PMID: [39196930](#)
- Emenecker RJ, Guadalupe K, Shamoon NM, Sukenik S, Holehouse AS. Sequence-ensemble-function relationships for disordered proteins in live cells. *bioRxiv.* 2023.
- Krueger RK, Brenner MP, Shrinivas K. Generalized design of sequence-ensemble-function relationships for intrinsically disordered proteins. *Nat Comput Sci.* 2026;6(5):512–23. <https://doi.org/10.1038/s43588-025-00881-y> PMID: [41053403](#)

27. Baruch Leshem A, Sloan-Dennison S, Massarano T, Ben-David S, Graham D, Faulds K, et al. Biomolecular condensates formed by designer minimalist peptides. *Nat Commun.* 2023;14(1):421. <https://doi.org/10.1038/s41467-023-36060-8> PMID: [36702825](#)
28. De Gennes PG. *Scaling concepts in polymer physics.* Cornell University Press; 1979.
29. Panagiotopoulos AZ, Wong V, Floriano MA. Phase equilibria of lattice polymers from histogram reweighting Monte Carlo simulations. *Macromolecules.* 1998;31(3):912–8.
30. Dignon GL, Zheng W, Best RB, Kim YC, Mittal J. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc Natl Acad Sci U S A.* 2018;115(40):9929–34. <https://doi.org/10.1073/pnas.1804177115> PMID: [30217894](#)
31. Harmon TS, Holehouse AS, Rosen MK, Pappu RV. Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *Elife.* 2017;6:e30294. <https://doi.org/10.7554/eLife.30294> PMID: [29091028](#)
32. Zheng W, Dignon G, Brown M, Kim YC, Mittal J. Hydropathy Patterning Complements Charge Patterning to Describe Conformational Preferences of Disordered Proteins. *J Phys Chem Lett.* 2020;11(9):3408–15. <https://doi.org/10.1021/acs.jpclett.0c00288> PMID: [32227994](#)
33. Sundaravadivelu Devarajan D, Wang J, Szala-Mendyk B, Rekhi S, Nikoubashman A, Kim YC, et al. Sequence-dependent material properties of biomolecular condensates and their relation to dilute phase conformations. *Nat Commun.* 2024;15(1):1912. <https://doi.org/10.1038/s41467-024-46223-w> PMID: [38429263](#)
34. Martin EW, Holehouse AS, Grace CR, Hughes A, Pappu RV, Mittag T. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J Am Chem Soc.* 2016;138(47):15323–35. <https://doi.org/10.1021/jacs.6b10272> PMID: [27807972](#)
35. Schuster BS, Dignon GL, Tang WS, Kelley FM, Ranganath AK, Jahnke CN, et al. Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proc Natl Acad Sci U S A.* 2020;117(21):11421–31. <https://doi.org/10.1073/pnas.2000223117> PMID: [32393642](#)
36. Brady JP, Farber PJ, Sekhar A, Lin Y-H, Huang R, Bah A, et al. Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proc Natl Acad Sci U S A.* 2017;114(39):E8194–203. <https://doi.org/10.1073/pnas.1706197114> PMID: [28894006](#)
37. Cohan MC, Shinn MK, Lalmansingh JM, Pappu RV. Uncovering Non-random Binary Patterns Within Sequences of Intrinsically Disordered Proteins. *J Mol Biol.* 2022;434(2):167373. <https://doi.org/10.1016/j.jmb.2021.167373> PMID: [34863777](#)
38. Ruff KM, King MR, Ying AW, Liu V, Pant A, Lieberman WE, et al. Molecular grammars of predicted intrinsically disordered regions that span the human proteome. *Cell.* 2026;189(1):323–342.e17. <https://doi.org/10.1016/j.cell.2025.10.019> PMID: [41232529](#)
39. von Bülow S, Tesei G, Zaidi FK, Mittag T, Lindorff-Larsen K. Prediction of phase-separation propensities of disordered proteins from sequence. *Proc Natl Acad Sci U S A.* 2025;122(13):e2417920122. <https://doi.org/10.1073/pnas.2417920122> PMID: [40131954](#)
40. Lin YH, Brady JP, Forman-Kay JD, Chan HS. Charge pattern matching as a 'fuzzy' mode of molecular recognition for the functional phase separations of intrinsically disordered proteins. *New J Phys.* 2017;19(11):115003.
41. Wang R, Wang ZG. Theory of polymer chains in poor solvent: Single-chain structure, solution thermodynamics, and θ point. *Macromolecules.* 2014;47(12):4094–102.
42. Urry DW, Gowda DC, Parker TM, Luan CH, Reid MC, Harris CM, et al. Hydrophobicity scale for proteins based on inverse temperature transitions. *Biopolymers.* 1992;32(9):1243–50. <https://doi.org/10.1002/bip.360320913> PMID: [1420991](#)
43. Firman T, Ghosh K. Sequence charge decoration dictates coil-globule transition in intrinsically disordered proteins. *J Chem Phys.* 2018;148(12):123305. <https://doi.org/10.1063/1.5005821> PMID: [29604827](#)
44. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *J Chem Phys.* 1953;21(6):1087–92.
45. Regy RM, Thompson J, Kim YC, Mittal J. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. *Protein Sci.* 2021;30(7):1371–9. <https://doi.org/10.1002/pro.4094> PMID: [33934416](#)
46. Tesei G, Lindorff-Larsen K. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Res Eur.* 2023;2:94. <https://doi.org/10.12688/openreseurope.14967.2> PMID: [37645312](#)
47. Flory PJ. Thermodynamics of high polymer solutions. *J Chem Phys.* 1942;10(1):51–61.
48. Li P, Banjade S, Cheng H-C, Kim S, Chen B, Guo L, et al. Phase transitions in the assembly of multivalent signalling proteins. *Nature.* 2012;483(7389):336–40. <https://doi.org/10.1038/nature10879> PMID: [22398450](#)
49. Dignon GL, Zheng W, Kim YC, Best RB, Mittal J. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput Biol.* 2018;14(1):e1005941. <https://doi.org/10.1371/journal.pcbi.1005941> PMID: [29364893](#)
50. Zeng X, Pappu RV. Developments in describing equilibrium phase transitions of multivalent associative macromolecules. *Curr Opin Struct Biol.* 2023;79:102540. <https://doi.org/10.1016/j.sbi.2023.102540> PMID: [36804705](#)
51. Choi S, Meyer MO, Bevilacqua PC, Keating CD. Phase-specific RNA accumulation and duplex thermodynamics in multiphase coacervate models for membraneless organelles. *Nat Chem.* 2022;14(10):1110–7. <https://doi.org/10.1038/s41557-022-00980-7> PMID: [35773489](#)
52. Sun Y, Wu X, Li J, Radiom M, Mezzenga R, Verma CS, et al. Phase-separating peptide coacervates with programmable material properties for universal intracellular delivery of macromolecules. *Nat Commun.* 2024;15(1):10094. <https://doi.org/10.1038/s41467-024-54463-z> PMID: [39572548](#)

53. Li S, Patel JS, Yang J, Crabtree AM, Rubenstein BM, Lund-Andersen PK, et al. Defining the HIV Capsid Binding Site of Nucleoporin 153. *mSphere*. 2022;7(5):e00310-22. <https://doi.org/10.1128/msphere.00310-22> PMID: [36040047](https://pubmed.ncbi.nlm.nih.gov/36040047/)
54. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol*. 2017;13(7):e1005659. <https://doi.org/10.1371/journal.pcbi.1005659> PMID: [28746339](https://pubmed.ncbi.nlm.nih.gov/28746339/)
55. Jung H, Yethiraj A. A simulation method for the phase diagram of complex fluid mixtures. *J Chem Phys*. 2018;148(24):244903. <https://doi.org/10.1063/1.5033958> PMID: [29960369](https://pubmed.ncbi.nlm.nih.gov/29960369/)
56. Ashbaugh HS, Hatch HW. Natively unfolded protein stability as a coil-to-globule transition in charge/hydrophobicity space. *J Am Chem Soc*. 2008;130(29):9536–42. <https://doi.org/10.1021/ja802124e> PMID: [18576630](https://pubmed.ncbi.nlm.nih.gov/18576630/)
57. Debye P, Hückel E. De la theorie des electrolytes. I. abaissement du point de congelation et phenomenes associes. *Phys Z*. 1923;24(9):185–206.