Molecular details of protein condensates probed by microsecond-long atomistic simulations

Wenwei Zheng,[*,†,⊥] Gregory L. Dignon,[*,‡,⊥] Nina Jovic,[‡] Xichen Xu,[‡] Roshan M. Regy,[‡] Nicolas L. Fawzi,[¶] Young C. Kim,[*,§] Robert B. Best,[*,‖] and Jeetain Mittal[*,‡]

†*College of Integrative Sciences and Arts, Arizona State University, Mesa, Arizona, United States*

‡*Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, Pennsylvania, United States*

¶*Department of Molecular Pharmacology, Physiology and Biotechnology, Brown University, Providence, RI 02912, USA*

§*Center for Materials Physics and Technology, Naval Research Laboratory, Washington, DC 20375*

‖*Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland, United States*

⊥*These authors contributed equally to this work*

E-mail: wenweizheng@asu.edu; gregory.dignon@stonybrook.edu; youngchan.kim@nrl.navy.mil; robert.best2@nih.gov; jeetain@lehigh.edu

Abstract

The formation of membraneless organelles in cells commonly occurs via liquid-liquid phase separation (LLPS), and is in many cases driven by multivalent interactions between intrinsically disordered proteins (IDPs). Investigating the nature of these interactions, and their effect on dynamics within the condensed phase, is therefore of critical importance, but very challenging for either simulation or experiment. Here, we study these interactions and their dynamics by pairing a novel multiscale simulation strategy with microsecond all-atom MD simulations of a condensed, IDP-rich phase.We simulate two IDPs this way, the low complexity domain of FUS and the N-terminal disordered domain of LAF-1, and find good agreement with experimental information on average density, water content and residue-residue contacts. We go significantly beyond what is known from experiments by showing that ion partitioning within the condensed phase is largely driven by the charge distribution of the proteins and – in the cases considered – shows little evidence of preferential interactions of the ions with the proteins. Furthermore, we are able to probe the microscopic diffusive dynamics within the condensed phase, showing that water and ions are in dynamic equilibrium between dense and dilute phases, and their diffusion reduces by a factor of 2-3 in the dense phase. Despite their high concentration in the condensate, the proteins also remain mobile, explaining the observed liquid-like properties of this phase. We finally show that IDP self-association is driven by a combination of non-specific hydrophobic interactions, as well as hydrogen bonds, salt bridges, $\pi - \pi$ and cation-π interactions. The simulation approach presented here allows the structural and dynamical properties of biomolecular condensates to be studied in microscopic detail, and is generally applicable to single and multi-component systems of proteins and nucleic acids involved in LLPS.

Introduction

Biomolecular condensates are highly concentrated subcellular assemblies of biomolecules that occur naturally in biology, and may function as organelles, such as the nucleolus,[1–3]

ribonucleoprotein granules,[4,5] and many others.[6–10] The study of these bodies, often termed membraneless organelles (MLOs), has recently attracted tremendous research effort due to its novelty, and relevance to biological functions,[11–14] pathologies such as neurodegenerative diseases[15,16] and the design of biomimetic materials.[17–19] It is now accepted that many MLOs are formed through a process of phase separation, commonly liquid-liquid phase separation (LLPS) in which a dynamic liquid-like condensate organizes biomolecules including proteins and nucleic acids and allows them to diffuse freely within the condensate, and to exchange rapidly with the surrounding environment.[7] A physical understanding of the driving forces of biomolecular phase separation is essential for uncovering the mechanistic details of MLO formation and the pathology of relevant diseases.[20–25]

A frequent property of proteins involved in biomolecular phase separation is intrinsic disorder, which has been highlighted through estimates of enhanced disorder predicted within MLO-associated proteins.[26] Indeed, intrinsically disordered proteins (IDPs) have been shown to phase separate at relatively low concentrations compared to most folded proteins,[5,25,27] likely due to their polymeric nature, and consequent increased multivalent interactions.[28] Additionally, IDPs are generally more solvent exposed[29] than folded proteins, and thus more accessible to post-translational modifications, which provide an efficient mechanism of controlling the thermodynamic and dynamic properties of condensates.[23,30,31] Recent work has highlighted that single-molecule behavior of IDPs may yield information relevant to their phase behavior, since the intramolecular interactions driving single-chain collapse are related to the intermolecular interactions driving its homotypic phase separation.[32,33] This leads to the question of what exactly are the interactions that drive LLPS, how can they be determined, and how can they be manipulated to control phase behavior?[25]

Despite the advances in methodology for investigating structure formation inside LLPS droplets by experiment,[22] it is still challenging to obtain high resolution, sequence-resolved information on structure and dynamics from experiment alone. All-atom molecular dynamics (MD) simulation with explicit solvent is a promising technique for generating detailed

information on conformational ensembles of IDPs,[34,35] and the contacts occurring within a condensate composed of IDP molecules.[22,36,37] The approach has already been applied to simulating the condensed phase of disordered peptides and proteins.[38–40] However, the large system sizes and timescales required to observe equilibrium coexistence of two phases pose a major challenge for all-atom simulations. We have previously overcome this difficulty by developing coarse-grained (CG) simulation models for LLPS.[30,32,41–44] We have complemented CG simulations of LLPS with atomistic simulations of smaller fragments of the IDPs,[5,22,24,36,45] which yield detailed interactions occurring between the relevant proteins in dilute solution, and in principle, within the condensed phase.[32] In this work, we unify these two approaches, by using CG simulations to generate an initial, equilibrated configuration of phase-separated proteins, which is then mapped back to all-atom coordinates to investigate the details of atomic interactions occurring within a protein condensate.

Results and Discussion

As test systems, we have selected the low-complexity prion-like domain of FUS (hereafter, FUS LC),[4] and the disordered N-terminal domain of LAF-1 (LAF-1 RGG).[36,46]To set up the system, we initially equilibrated 40 chains of FUS LC or LAF-1 RGG in a planar slab geometry using our previously developed CG model[32,41] (Fig. 1a). The system size was chosen as it yields an atomic resolution system that is sufficiently small to run on Anton 2,[47] while being sufficiently large that finite size effects are small. We verify this by comparing with a larger system as we have used previously with 100 chains[32,41] and find similar coexistence densities (see supporting methods and Fig. S1). After setting up this system, we reconstructed all-atom coordinates from the C_α positions of the coarse-grained simulations by using a lookup table based on the protein structure database with the PULCHRA code[48] (Fig. 1b). Any conflicts between sidechains of different chains were resolved via a short simulation with the CAMPARI Monte Carlo engine and ABSINTH implicit solvent model

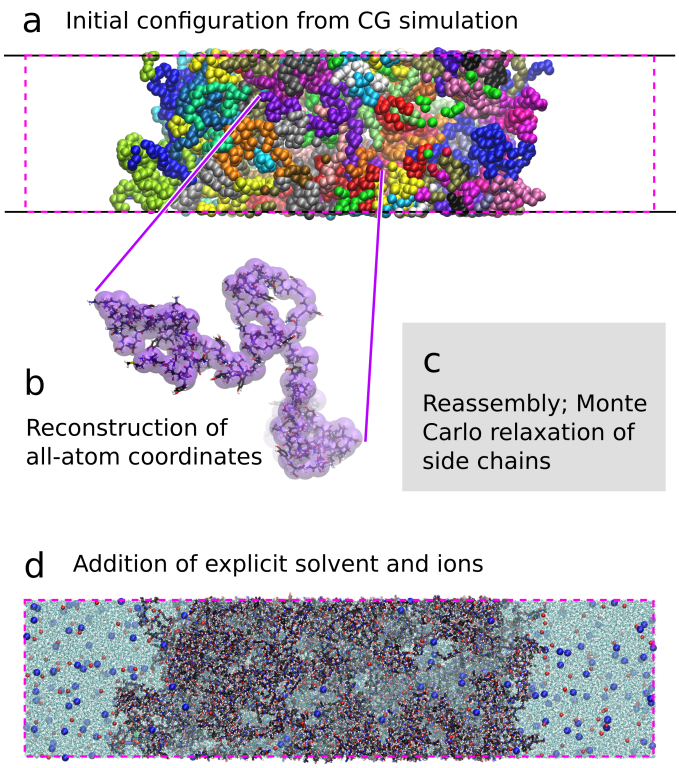


Figure 1: Procedure to set up all-atom simulations of dense phase. (a) An initial configuration is generated by CG simulations in a box with elongated z-dimension. (b) All-atom coordinates reconstructed from CG C_α coordinates templates for each chain using PDB database. (c) Reconstructed chains reassembled into condensed phase and sidechain clashes relieved using short Monte Carlo simulations with frozen backbone. (d) Addition of solvent and ions to generate complete system at all-atom resolution.

with fixed backbone[49] (Fig. 1c). Finally, the system was solvated and equilibrated with explicit solvent using the Amber ff03ws[50] force field, TIP4P/2005 water model,[51] and ∼100 mM NaCl[52] (Fig. 1d). By utilizing the specialized software and hardware from Anton 2 supercomputer developed by DE Shaw Research,[47] we equilibrated the system for 150 ns to relax it to its equilibrium density (Fig. S2) and collected a 2 μs trajectory in the NVT ensemble at 298 K for each sequence of interest (see supporting methods for details).

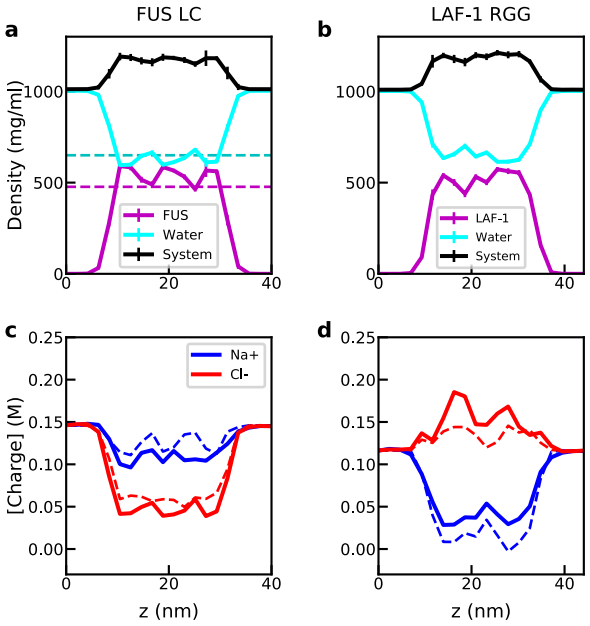


Figure 2: Density profiles from all-atom slab simulations of FUS LC (left) and LAF-1 RGG (right). Components are shown in the legend. Dashed lines in (a) indicate experimentally determined values[22] and dashed lines in (c) and (d) indicate the predicted ion concentration using concentrations of protein cationic and anionic residues.

All-atom simulations with explicit solvent can provide a great deal of information not accessible from CG models, most obviously how the solvent and ions partition into the dense phase, and how this depends on protein sequence. The initial protein concentration in the dense phase for both proteins was selected based on the NMR measurement of condensed phase FUS LC to be ∼477 mg/mL,[22] and typical for protein LLPS.[53] We note that there is also some indirect evidence for extremely low density condensates of LAF-1 RGG under certain conditions but it is not consistent with directly measured values for a human homolog

ddx4 with very similar sequence which forms very high density phases.[53,54] In both cases, the protein-rich phase has a higher total density than water (black lines), which agrees with the experimental observations that condensates of these proteins can be sedimented or separated using centrifugation.[4,46] In our system size with 40 chains, the expected number of chains in the dilute phase is close to zero. We therefore designed the simulation to have all the chains in the condensed slab. The density of the dilute phase cannot be estimated due to the fact that the escape of one or two chains from the slab happens at a much longer time scale than our simulation length. However, the diffusion of solvent and ions is very rapid and so their equilibrium partitioning can be readily probed from our all-atom simulations, as shown in Fig. 2 for FUS LC and LAF-1 RGG protein condensates. The water content inside both FUS LC and LAF-1 RGG protein-rich regions is on the order of ∼600 mg/mL (Figs. 2a and b), very similar despite significant differences in their sequence composition. The water content inside the FUS LC protein-rich phase from the simulation is consistent with the reported experimental estimate of 65% (by volume) by Murthy et al.[22]

Despite very similar protein and water density profiles for FUS LC and LAF-1 RGG, the partitioning of Na^+ and Cl^- ions differs considerably between the two systems (Fig. 2c and d). In the case of FUS LC, which only has two anionic residues (Asp), the concentration of Cl^- ions is greatly reduced inside the protein-rich phase, being preferentially excluded, while Na^+ ion concentration is only slightly reduced in the protein-rich phase (Fig. 2c). For LAF-1 RGG (Fig. 2d), which contains a more significant fraction of anionic and cationic charged residues (26% of charged amino acids), the Cl^- ions are preferentially incorporated into the protein-rich region, while the Na^+ ions are excluded. This likely has to do with the net +4 charge per protein chain for LAF-1 RGG. The equilibrium partition coefficient of ions reflects an interplay of direct charge-charge interactions between charged amino acids and ions and the free energy of transferring the ions from a solvent-rich to a protein-rich environment. Using a simple model, we can predict the local concentration of Na^+ and Cl^- from the local concentration of cationic and anionic residues (Fig. S3) and bulk concentrations of ions and

water. We set

$$[\mathrm{Na}^+] - [\mathrm{Cl}^-] + [\mathrm{protein}_{\mathrm{cation}}] - [\mathrm{protein}_{\mathrm{anion}}] = 0, \tag{1}$$

to represent electroneutrality, and

$$[\mathrm{Na}^+] + [\mathrm{Cl}^-] = \frac{[\mathrm{water}]}{[\mathrm{water}_{\mathrm{bulk}}]}([\mathrm{Na}^+_{\mathrm{bulk}}] + [\mathrm{Cl}^-_{\mathrm{bulk}}]), \tag{2}$$

which assumes negligible preferential interactions between ions and amino acids, as would be expected at these relatively low ion concentrations. The predicted Na^+ and Cl^- concentrations are plotted in Figs. 2c and d as dashed lines, and show good agreement with the concentrations obtained from the simulation. These results highlight the role of the charged amino acids in determining the density and composition of the protein condensates, which ultimately help to determine their function.

While equilibrium concentrations and compositions of MLOs are important to their function, another important factor is the dynamics within the dense phase, as it determines the rate at which components may pass through or rearrange within the condensate. Our MD simulations also provide detailed information on the dynamics within the condensed phase, and may be used to decouple the different components. The heterogeneous nature of our system with a distinct protein-rich environment, protein-poor bulk region, and an interfacial region poses some challenges to estimate diffusion coefficients unambiguously using the standard approach based on the mean square displacement. An alternative approach is to compute the probability distributions $P(\xi(t_0 + t) - \xi(t_0))$ for molecular displacement (i.e. propagators) in each direction $\xi = x, y, z$ as a function of the lag time t between observations (see supporting methods and Fig. S4). Since the simulation box is not cubic, we report diffusivity (D) values based on only the longer z-axis, in order to minimize the finite size effects.[55] We find it necessary to include more than one term (multiple D values) while fitting the propagator data from simulation to the expected distribution for one-dimensional diffusion. This behavior is consistent with the expected differences in the dynamics of solvent

molecules within the protein-rich and bulk phases. We find that the observed behavior of water and ions is best accounted for by three D values whereas one D value is sufficient for fitting protein propagator data (Fig. S4).

The fastest D value for water ($D_1 \approx 1.98$ nm^2/ns in FUS LC simulation) is consistent with the literature value (2.30 nm^2/ns[56]), and its relative contribution to the propagators is also compatible with the number of water molecules in the bulk region (Table S2). The second mode is significantly slower, by a factor of 5 from the bulk diffusion, very close to the 6-fold decreased diffusivity reported for buffer molecules within FUS condensates.[22] Based on this agreement, and its contribution to the propagator, we expect D_2 reflects slower water diffusion inside the protein-rich region (Table S2). The slowest mode (∼0.8% contribution) is difficult to pinpoint but is likely coming from a combination of factors, most importantly, water molecules directly interacting with protein atoms (Table S2). Similar to water diffusion, the dynamical behavior of ions reflects the presence of distinct populations. Most importantly, each mode's contribution and its relative difference from bulk diffusion appear to depend on the protein sequence (Table S2).

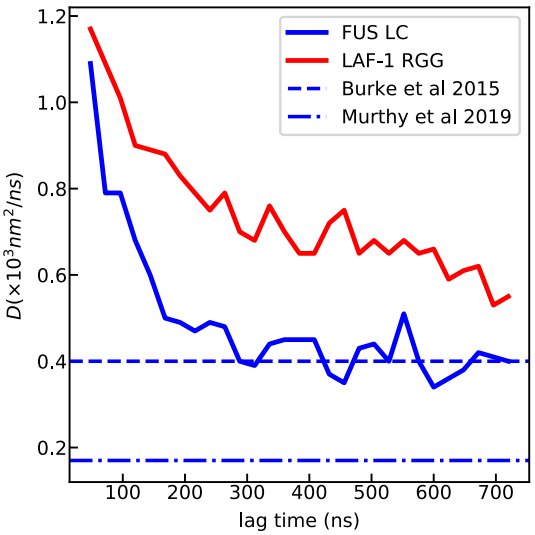


Figure 3: Protein self-diffusion coefficients within the slab along the z-axis, from all-atom slab simulations of FUS LC (blue) and LAF-1 RGG (red) are shown as a function of the lag time. Dashed horizontal line indicates experimentally determined diffusivity value for FUS LC.[4]

The protein dynamics inside the condensed phase is closely connected to its liquid-like properties, needed for maintaining the biological function of the biomolecular condensate.[57] To estimate the rate of relaxation of intramolecular protein degrees of freedom, we calculate the time autocorrelation function of the radius of gyration (Fig. S5) yielding average correlation times of 192 and 122 ns for FUS LC and LAF-1 RGG respectively. We note that relaxation timescales for LAF-1 RGG are somewhat shorter compared with those for FUS LC, and that they are comparable to experimental estimates for isolated IDPs of similar length.[58,59] This suggests that formation of the condensed phase has only a modest effect on intramolecular dynamics. The 2 μs long MD simulations are at least 10 times longer than this relaxation timescale, which gives reasonable confidence in our ability to directly compute the diffusivity values of these two proteins (Fig. 3). The diffusion coefficient obtained for FUS LC is in excellent agreement with a previously determined value from FRAP and NMR diffusion experiments by Fawzi and co-workers[4,22] (Fig. 3). Consistent with the faster chain relaxation time, the LAF-1 RGG diffusion coefficient is higher than that for the FUS LC. This may be explained because both the interchain and intrachain interactions governing frictional effects should have a similar dependence on the protein sequence.

Because of the potential significance of secondary structure elements in mediating interactions in condensed phases,[45] we examined the secondary structure populations of the proteins in the condensed phase using the DSSP algorithm[60] (Fig. S6). From this analysis, we find that the protein chains are largely disordered, with more than 50% of residues in a coil conformation, with local helices being the most common type of structured state (Fig. S6). This is consistent with experimental NMR studies showing a lack of structure within FUS condensates,[4,22,30] and condensates of a protein similar to LAF-1 IDR, Ddx4.[53]

The central goal of this work was to elucidate the atomic-resolution interactions stabilizing a condensed proteinaceous phase which cannot be accessed through lower-resolution CG simulations. This information is essential to gain a fundamental mechanistic understanding of molecular driving forces and developing theory for the sequence determinants of protein

assembly. Previous studies have highlighted the role of various interaction modes that may be responsible for driving the LLPS of different protein sequences, such as salt bridges,[28,53] cation-π interactions,[61,62] hydrophobic interactions,[22,39,63] sp^2/π interactions between several residue pairs including the protein backbone,[21,22] and hydrogen bonding interactions.[22,39,64] There is still a limited understanding of the relative importance of these different interaction modes in the context of a particular type of amino acid pair, or a protein sequence. We attempt to provide answers to some of these questions here.

To characterize the regions of each sequence most involved in molecular interactions, we start by computing the number of intermolecular van der Waals (vdW) contacts (see supporting methods for definitions) formed as a function of protein residue number (Figs. 4a and b) per frame, averaged over the entire trajectory. We find that contacts are relatively evenly distributed throughout the FUS LC sequence (Fig. 4a), which is consistent with NMR data.[22] One can observe intermittent peaks in the one-dimensional contact map data arising from the Tyr residues distributed throughout the FUS LC sequence (Fig. 4A, black dashed lines in the bottom panel). For LAF-1 RGG, the contacts are still distributed throughout the chain with a notable contact-prone region between residues 20-28 (Fig. 4b), which was identified previously from our CG model simulation and tested experimentally to be critical for promoting LLPS.[36]

To obtain a better understanding of how different amino acid types contribute to the formation of intermolecular contacts between protein chains, we combine the contact data for different pairs of the same kind. From this data in Figs. 4c and d, one can identify important residue pairs as well as residue types that are primarily responsible for interchain interactions and for stabilizing the condensed protein-rich phase. For both FUS LC and LAF-1 RGG, Tyr interactions with itself and other residues are highly abundant and likely essential drivers of LLPS.[21,22,28,37,62,64,65] Importantly, polar residues (Gln and Ser in the case of FUS LC and Asn in LAF-1 RGG) also participate in significant contacts, consistent with a recent mutagenesis study highlighting their role in LLPS.[22] Also, Gly residues appear to

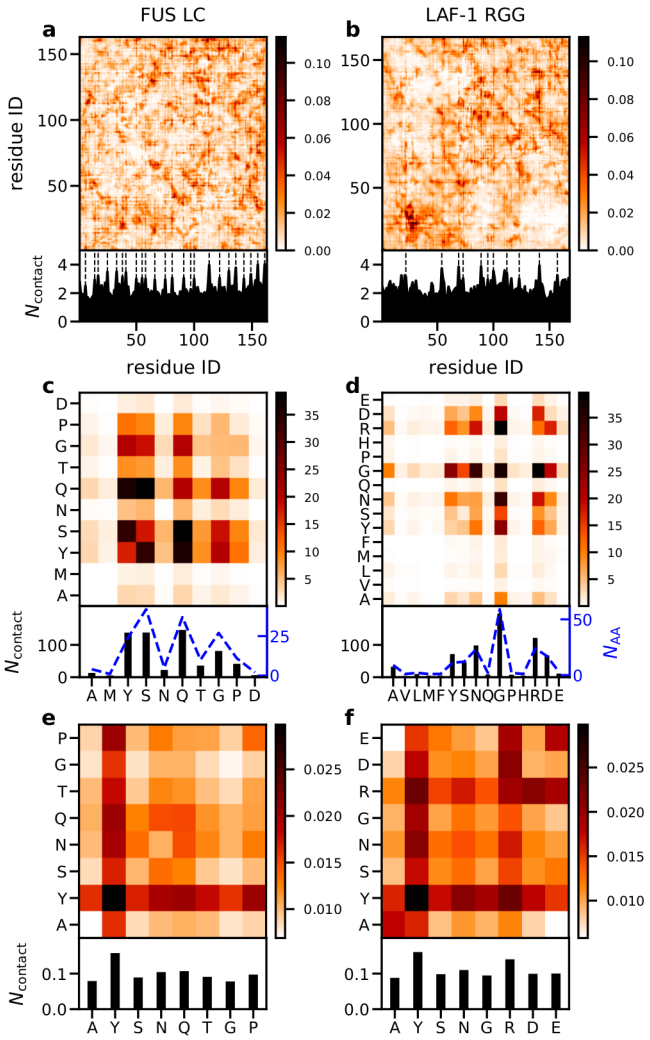


Figure 4: Intermolecular contacts within the condensed phase of FUS LC (left) and LAF-1 RGG (right), as a function of residue index (a and b) and amino acid types (c and d). The intermolecular contacts normalized by the relative abundance of each amino acid in the sequence are shown in e and f. In each of the figure, the bottom panel shows the one dimensional summation. The black dashed lines in a and b show the position of Tyr residues. The blues lines in c and d show the number of amino acids (N_{AA}) in the sequence.

be forming contacts with many other residue types in both proteins; this is highly visible in the LAF-1 RGG data for interactions with Arg, Gly, Tyr, and Asn. Lastly, LAF-1 RGG contact formation is enhanced by interactions between oppositely-charged residue pairs such as Arg-Asp pairs (Fig. 4d). To obtain the intrinsic propensity for each amino acid to form a contact, we also normalize the contacts by the relative abundance of each amino acid in the sequence (Figs. 4e and f). The overall values are largely consistent between the FUS LC and

LAF-1 RGG simulations, and supports the critical role of Tyr and Arg due to their intrinsic preference to form contacts while the Gly-involved contacts are present due to its abundance in both the sequences. Even though Gln and Ser contribute as many total contacts as Tyr, each individual Tyr contributes more. This is because there are more Ser and Gln than Tyr in FUS LC sequence. Additionally, Tyr is bigger than Ser and Gly, which may allow it to make more simultaneous contacts. We have also calculated the intramolecular interactions in the same way and obtained excellent agreement with the intermolecular interactions (Figs. S7 and S8), which supports our recent finding connecting the self-interaction properties of the single chain with LLPS behaviors.[32] Data for residues that are an insignificant fraction of the protein composition (appearing $\leq$2 times) have been excluded from the plot due to higher uncertainty associated with their contacts. In addition, we have included a comparison of the residue-specific contacts between the all-atom simulations and the initial configuration reconstructed from the CG simulation (Fig. S9 and S10). We find for different pairs of amino acids, the contact difference is not uniform, suggesting the variation does not solely come from compaction or expansion of the chain. The all-atom force field refines residue-level interactions through a finer description of atomic interactions and therefore is expected to provide a more accurate description of molecular interactions stabilizing the condensed protein-rich phase. Closer examination of the differences between coarse-grained and all-atom models indicates that the same residue pairs in each protein have similar shifts, suggesting that they represent small, but significant differences between the models.

To dive deeper into the atomic interactions responsible for the observed role of the amino acids identified above, we determine the interaction modes present when two residues form a vdW contact. Based on the previous literature,[21,22,62] the most important modes are sp^2/π, hydrogen bonding, cation-π, and salt bridge. Here, we separate these interactions into contacts between backbone atoms (bb-bb), sidechain atoms (sc-sc), or backbone and sidechain (bb-sc) atoms (see supporting methods for definition of these interaction modes). The amino acid pairs are sorted by the number of vdW contacts formed (Fig. S11) and the

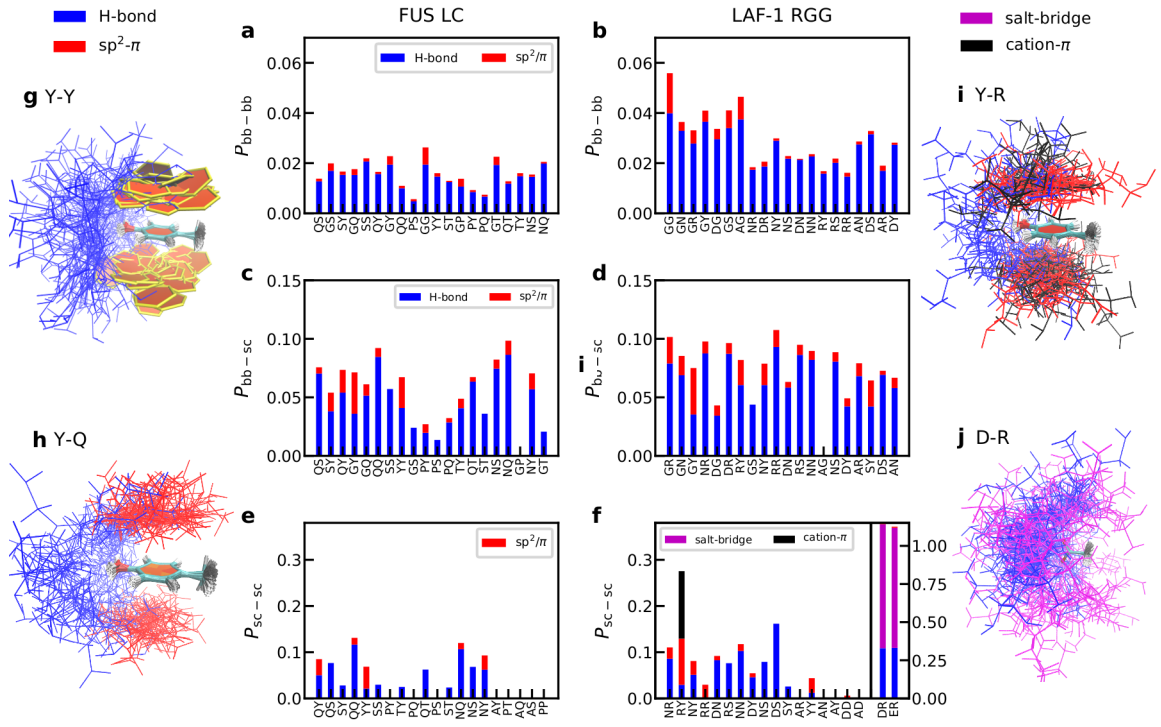


Figure 5: Interaction modes contributing to the intermolecular contacts separated into contributions between backbone atoms (bb-bb, a and b), between backbone and sidechain atoms (bb-sc, c and d) and between sidechain atoms (sc-sc, e and f). The amino acid pairs are sorted by the number of contacts formed between these pairs in each group as shown in Fig. S11. Some configurations of representative interacting amino acids are shown in g for FUS LC Tyr:Tyr, h for FUS LC Tyr:Gln, i for LAF-1 RGG Tyr:Arg and j for LAF-1 RGG Asp:Arg. The configurations are aligned to Tyr in g, h and i and to Asp in j. The color code is the same as the interaction mode shown in the legend. Rings are shown in paperchain representation.[66]

top 20 amino acid pair types for each group and each protein are shown in Fig. 5 with the full version in Fig. S12. The interaction modes from intramolecular interactions (Figs. S13 and S14) are highly similar to those from intermolecular interactions, so we only discuss the intermolecular version here.

For both FUS LC and LAF-1 RGG, we first note that most of the vdW contacts are non-specific and thus only a small fraction can be classified into any of the aforementioned specific interaction modes (Figs. 5a-f). This emphasizes the importance of non-specific interaction modes, including the hydrophobic interactions, in promoting LLPS. Within the context of

interaction modes in FUS LC, all pairs except for those involving Tyr are primarily stabilized by hydrogen bonds (Figs. 5a,c and e). Interactions involving sp^2/π groups are a relatively small fraction of the contacts, except for residue pairs involving Tyr, with contributions from the sp^2/π mode higher than from hydrogen bonds. The configurations of representative amino acid interactions are also shown in Figs. 5g and h. For both Tyr and Gln interactions, sp^2/π interaction modes tend to form on top or bottom of the sidechain whereas hydrogen bonds are around the side. This suggests for aromatic amino acids like Tyr, hydrogen bonds might not directly compete with forming sp^2/π interactions and can still be a major contribution to stabilizing the condensates.

For the LAF-1 RGG, we include two additional interaction modes, salt bridges and cation-π interactions, involving charged residues (Figs. 5b,d and f). We find that charged amino acids contribute heavily to LAF-1 sidechain interactions, with hydrogen bond and sp^2/π interactions from aromatic amino acids playing secondary roles (Fig. 5f). Previously, we have shown that certain pairings of residues can form contacts using different interaction modes, either switching between them, or forming multiple contacts cooperatively,[36] particularly cation-π and sp^2/π interactions between Arg and Tyr. Here we also find hydrogen bonds and salt bridges between cationic-aromatic pairings and oppositely-charged residues are among the strongest interactions occurring within the LAF-1 condensate (Fig. 5f), and different interaction modes can occur at the same time, e.g. cation-π and sp^2/π interactions between Arg and Tyr (Fig. 5i), and salt bridge and hydrogen bonds between Arg and Asp (Fig. 5j). This is also the reason the total probability of interaction modes might exceed 1 for interactions involving charged amino acids (Fig. 5f).

Conclusion

In this work, we present a general methodology for initializing, conducting, and analyzing all-atom explicit-solvent simulations of biomolecular condensates in coexistence with a sur-

rounding aqueous phase. We have optimized the procedure for systems with components of similar size to FUS LC and LAF-1 RGG so that similar simulations should be accessible using even general-purpose computing hardware (Table S1). We have leveraged our earlier work with CG simulations of IDP phase coexistence,[5,22,24,30,32,36,41,42,67] and atomistic studies of inter-protein interactions[5,22,24,36] to obtain important mechanistic details of the underlying molecular interactions of condensates. We note there are properties that cannot be adequately sampled by all-atom simulations, however we believe there are a number of key insights such as atomic level interactions and diffusion of protein, water and ions in the condensed phase that cannot be obtained in a CG simulation.

We find that the proteins are remarkably dynamic in the condensed phase, having intramolecular correlation times very comparable to those typical of isolated intrinsically disordered proteins. This flexibility is key to the liquid-like properties of the protein-rich phase. While the dense phase is highly viscous, we are also able to measure the protein diffusivity, finding excellent agreement with experimental results where available. Similarly, we show that water and ions are able to rapidly diffuse between phases, with diffusion coefficients within the dense phase reduced.

For both tested proteins, the equilibrium distribution of sodium and chloride ions within the condensed phase is essentially determined by the charge distribution and water content inside the phase-separated proteins. This implies that there is no strong preferential interaction of these ions with protein residues in these systems under the conditions we study. We note, however, that ions exhibiting stronger Hofmeister effects, or higher salt concentrations,[68–70] may alter this result, and would be interesting to consider in future work.

Finally, we find many types of residue-residue interactions are responsible for stabilizing the condensed phase, and contacts involving Gly are particularly abundant due to its frequency in the sequence. After normalizing for residue frequency, however, it appears that each Tyr contributes more interactions per residue than any other residue type, explaining its apparent importance in mutagenic approaches. For LAF-1 RGG, in addition to Tyr in-

teractions, we observe that both cation-π interactions (particularly involving Arg) and salt bridges contribute to the condensate's stability. The approach outlined here can be used to explore the generality of these findings in the context of other protein sequences.

Acknowledgement

We acknowledge useful discussions with Dr. Anastasia Murthy. This work was supported in part by the National Institutes of Health (NIH) grants R01GM120537 (J.M.), R01NS116176 (N.L.F. and J.M.), and R01GM118530 (N.L.F.), National Science Foundation grants DMR-2004796 (J.M.) and MCB-2015030 (W.Z.). R.B. was supported by the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases of the NIH and Y.C.K by the Office of Naval Research via the U.S. Naval Research Laboratory base program. This research used computational resources of Anton 2, XSEDE (supported by the NSF project no. TG-MCB120014), and the NIH HPC Biowulf cluster (http://hpc.nih.gov). The Anton 2 machine at PSC was generously made available by D.E. Shaw Research and the computer time was provided by the Pittsburgh Supercomputing Center (PSC) through NIH Grant R01GM116961.

Supporting Information Available

Supporting information includes: Detailed description of simulation setup, conversion to atomic resolution, minimization, and production simulation as well as description of modes of contact.

References

(1) Berry, J.; Weber, S. C.; Vaidya, N.; Haataja, M.; Brangwynne, C. P. RNA transcription modulates phase transition-driven nuclear body assembly. *Proc. Natl. Acad. Sci. U.S.A.*

2015, *112*, E5237–E5245.

(2) Feric, M.; Vaidya, N.; Harmon, T. S.; Mitrea, D. M.; Zhu, L.; Richardson, T. M.; Kriwacki, R. W.; Pappu, R. V.; Brangwynne, C. P. Coexisting liquid phases underlie nucleolar subcompartments. *Cell* **2016**, *165*, 1686–1697.

(3) Mitrea, D. M.; Cika, J. A.; Stanley, C. B.; Nourse, A.; Onuchic, P. L.; Banerjee, P. R.; Phillips, A. H.; Park, C.-G.; Deniz, A. A.; Kriwacki, R. W. Self-interaction of NPM1 modulates multiple mechanisms of liquid–liquid phase separation. *Nat. Commun.* **2018**, *9*, 1–13.

(4) Burke, K. A.; Janke, A. M.; Rhine, C. L.; Fawzi, N. L. Residue-by-residue view of in vitro FUS granules that bind the C-terminal domain of RNA polymerase II. *Mol. Cell* **2015**, *60*, 231–241.

(5) Ryan, V. H.; Dignon, G. L.; Zerze, G. H.; Chabata, C. V.; Silva, R.; Conicella, A. E.; Amaya, J.; Burke, K. A.; Mittal, J.; Fawzi, N. L. Mechanistic View of hnRNPA2 Low-Complexity Domain Structure, Interactions, and Phase Separation Altered by Mutation and Arginine Methylation. *Mol. Cell* **2018**,

(6) Riback, J. A.; Katanski, C. D.; Kear-Scott, J. L.; Pilipenko, E. V.; Rojek, A. E.; Sosnick, T. R.; Drummond, D. A. Stress-triggered phase separation is an adaptive, evolutionarily tuned response. *Cell* **2017**, *168*, 1028–1040.

(7) Brangwynne, C. P.; Eckmann, C. R.; Courson, D. S.; Rybarska, A.; Hoege, C.; Gharakhani, J.; Jülicher, F.; Hyman, A. A. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **2009**, *324*, 1729–1732.

(8) Marzahn, M. R.; Marada, S.; Lee, J.; Nourse, A.; Kenrick, S.; Zhao, H.; Ben-Nissan, G.; Kolaitis, R.-M.; Peters, J. L.; Pounds, S. et al. Higher-order oligomerization promotes localization of SPOP to liquid nuclear speckles. *EMBO J.* **2016**, e201593169.

(9) Sabari, B. R.; Dall'Agnese, A.; Boija, A.; Klein, I. A.; Coffey, E. L.; Shrinivas, K.; Abraham, B. J.; Hannett, N. M.; Zamudio, A. V.; Manteiga, J. C. et al. Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **2018**, *361*, eaar3958.

(10) Milovanovic, D.; Wu, Y.; Bian, X.; De Camilli, P. A liquid phase of synapsin and lipid vesicles. *Science* **2018**, *361*, 604–607.

(11) Larson, A. G.; Elnatan, D.; Keenen, M. M.; Trnka, M. J.; Johnston, J. B.; Burlingame, A. L.; Agard, D. A.; Redding, S.; Narlikar, G. J. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature* **2017**, *547*, 236–240.

(12) Shin, Y.; Chang, Y.-C.; Lee, D. S. W.; Sanders, D. W.; Ronceray, P.; Wingreen, N. S.; Haataja, M. P.; Brangwynne, C. P. Liquid nuclear condensates mechanically sense and restructure the genome. *Cell* **2018**, *175*, 1481–1491.

(13) Sanulli, S.; Trnka, M. J.; Dharmarajan, V.; Tibble, R. W.; Bascal, B. D.; Burlingame, A. L.; Griffin, P. R.; Gross, J. D.; Narlikar, G. J. HP1 reshapes nucleosome core to promote phase separation of heterochromatin. *Nature* **2019**, *575*, 390–394.

(14) Gibson, B. A.; Doolittle, L. K.; Schneider, M. W. G.; Jensen, L. E.; Gamarra, N.; Henry, L.; Gerlich, D. W.; Redding, S.; Rosen, M. K. Organization of chromatin by intrinsic and regulated phase separation. *Cell* **2019**, *179*, 470–484.

(15) Patel, A.; Lee, H. O.; Jawerth, L.; Maharana, S.; Jahnel, M.; Hein, M. Y.; Stoynov, S.; Mahamid, J.; Saha, S.; Franzmann, T. M. et al. A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* **2015**, *162*, 1066–1077.

(16) Gomes, E.; Shorter, J. The molecular language of membraneless organelles. *J. Biol. Chem.* **2019**, *294*, 7115–7127.

(17) Simon, J. R.; Carroll, N. J.; Rubinstein, M.; Chilkoti, A.; López, G. P. Programming molecular self-assembly of intrinsically disordered proteins containing sequences of low complexity. *Nat. Chem.* **2017**, *9*, 509.

(18) Lau, H. K.; Paul, A.; Sidhu, I.; Li, L.; Sabanayagam, C. R.; Parekh, S. H.; Kiick, K. L. Microstructured Elastomer-PEG Hydrogels via Kinetic Capture of Aqueous Liquid–Liquid Phase Separation. *Adv. Sci.* **2018**, *5*, 1701010.

(19) Roberts, S.; Harmon, T. S.; Schaal, J. L.; Miao, V.; Li, K. J.; Hunt, A.; Wen, Y.; Oas, T. G.; Collier, J. H.; Pappu, R. V. et al. Injectable tissue integrating networks from recombinant polypeptides with tunable order. *Nat. Mater.* **2018**, *17*, 1154–1163.

(20) Nott, T. J.; Craggs, T. D.; Baldwin, A. J. Membraneless organelles can melt nucleic acid duplexes and act as biomolecular filters. *Nat. Chem.* **2016**, *8*, 569–575.

(21) Vernon, R. M.; Chong, P. A.; Tsang, B.; Kim, T. H.; Bah, A.; Farber, P.; Lin, H.; Forman-Kay, J. D. Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife* **2018**, *7*, e31486.

(22) Murthy, A. C.; Dignon, G. L.; Kan, Y.; Zerze, G. H.; Parekh, S. H.; Mittal, J.; Fawzi, N. L. Molecular interactions underlying liquid- liquid phase separation of the FUS low-complexity domain. *Nat. Struct. Mol. Biol.* **2019**, *26*, 637.

(23) Kim, T. H.; Tsang, B.; Vernon, R. M.; Sonenberg, N.; Kay, L. E.; Forman-Kay, J. D. Phospho-dependent phase separation of FMRP and CAPRIN1 recapitulates regulation of translation and deadenylation. *Science* **2019**, *365*, 825–829.

(24) Conicella, A. E.; Dignon, G. L.; Zerze, G. H.; Schmidt, H. B.; Alexandra, M.; Kim, Y. C.; Rohatgi, R.; Ayala, Y. M.; Mittal, J.; Fawzi, N. L. TDP-43 α-helical structure tunes liquid–liquid phase separation and function. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 5883–5894.

(25) Dignon, G. L.; Best, R. B.; Mittal, J. Biomolecular Phase Separation: From Molecular Driving Forces to Macroscopic Properties. *Annu. Rev. Phys. Chem.* **2020**, *71*, 53–75.

(26) Darling, A. L.; Liu, Y.; Oldfield, C. J.; Uversky, V. N. Intrinsically Disordered Proteome of Human Membrane-Less Organelles. *Proteomics* **2018**, *18*, 1700193.

(27) Asherie, N. Protein crystallization and phase diagrams. *Methods* **2004**, *34*, 266–272.

(28) Pak, C. W.; Kosno, M.; Holehouse, A. S.; Padrick, S. B.; Mittal, A.; Ali, R.; Yunus, A. A.; Liu, D. R.; Pappu, R. V.; Rosen, M. K. Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein. *Mol. Cell* **2016**, *63*, 72–85.

(29) Zerze, G. H.; Zheng, W.; Best, R. B.; Mittal, J. Evolution of All-atom Protein Force Fields to Improve Local and Global Properties. *J. Phys. Chem. Lett.* **2019**, *10*, 2227.

(30) Monahan, Z.; Ryan, V. H.; Janke, A. M.; Burke, K. A.; Zerze, G. H.; O'Meally, R.; Dignon, G. L.; Conicella, A. E.; Zheng, W.; Best, R. B. et al. Phosphorylation of FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. *EMBO J.* **2017**, *doi*, 10.15252/embj.201696394.

(31) Hofweber, M.; Hutten, S.; Bourgeois, B.; Spreitzer, E.; Niedner-Boblenz, A.; Schifferer, M.; Ruepp, M.-D.; Simons, M.; Niessing, D.; Madl, T. et al. Phase separation of FUS is suppressed by its nuclear import receptor and arginine methylation. *Cell* **2018**, *173*, 706–719.

(32) Dignon, G. L.; Zheng, W.; Best, R. B.; Kim, Y. C.; Mittal, J. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, 9929–9934.

(33) Lin, Y.-H.; Chan, H. S. Phase Separation and Single-Chain Compactness of Charged Disordered Proteins Are Strongly Correlated. *Biophys. J.* **2017**, *112*, 2043–2046.

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

(35) Best, R. B. Emerging consensus on the collapse of unfolded and intrinsically disordered proteins in water. *Curr. Opin. Struct. Biol.* **2020**, *60*, 27–38.

(36) Schuster, B. S.; Dignon, G. L.; Tang, W. S.; Kelley, F. M.; Ranganath, A. K.; Jahnke, C. N.; Simpkins, A. G.; Regy, R. M.; Hammer, D. A.; Good, M. C. 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*, 11421–11431.

(37) Martin, E. W.; Holehouse, A. S.; Peran, I.; Farag, M.; Incicco, J. J.; Bremer, A.; Grace, C. R.; Soranno, A.; Pappu, R. V.; Mittag, T. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **2020**, *367*, 694–699.

(38) Karandur, D.; Wong, K.-Y.; Pettitt, B. M. Solubility and aggregation of Gly5 in water. *J. Phys. Chem. B* **2014**, *118*, 9565–9572.

(39) Rauscher, S.; Pomès, R. The liquid structure of elastin. *Elife* **2017**, *6*, e26526.

(40) Paloni, M.; Bailly, R.; Ciandrini, L.; Barducci, A. Unraveling molecular interactions in a phase-separating protein by atomistic simulations. *bioRxiv* **2020**, *https://doi.org/10.1101/2020.05.16.099051*.

(41) Dignon, G. L.; Zheng, W.; Kim, Y. C.; Best, R. B.; Mittal, J. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* **2018**, *14*, e1005941.

(42) Dignon, G. L.; Zheng, W.; Kim, Y. C.; Mittal, J. Temperature-Controlled Liquid–Liquid Phase Separation of Disordered Proteins. *ACS Cent. Sci.* **2019**, *5*, 821.

(43) Perdikari, T. M.; Jovic, N.; Dignon, G. L.; Kim, Y. C.; Fawzi, N. L.; Mittal, J. A predictive coarse-grained model for position-specific effects of post-translational modifications on disordered protein phase separation. *bioRxiv* **2020**, *https://doi.org/10.1101/2020.06.12.148650*.

(44) Regy, R. M.; Dignon, G. L.; Zheng, W.; Kim, Y. C.; Mittal, J. Sequence dependent co-phase separation of RNA-protein mixtures elucidated using molecular simulations. *bioRxiv* **2020**, *https://doi.org/10.1101/2020.07.07.192047*.

(45) Conicella, A. E.; Zerze, G. H.; Mittal, J.; Fawzi, N. L. ALS mutations disrupt phase separation mediated by α-helical structure in the TDP-43 low-complexity C-terminal domain. *Structure* **2016**, *24*, 1537–1549.

(46) Elbaum-Garfinkle, S.; Kim, Y.; Szczepaniak, K.; Chen, C. C.-H.; Eckmann, C. R.; Myong, S.; Brangwynne, C. P. The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 7189–7194.

(47) Shaw, D. E.; Grossman, J.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H. et al. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. Proceedings of the international conference for high performance computing, networking, storage and analysis. 2014; pp 41–53.

(48) Rotkiewicz, P.; Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **2008**, *29*, 1460–1465.

(49) Vitalis, A.; Pappu, R. V. ABSINTH: A new continuum solvent model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* **2008**, *30*, 673–699.

(50) Best, R. B.; Zheng, W.; Mittal, J. Balanced protein–water interactions improve prop-

erties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* **2014**, *10*, 5113–5124.

(51) Abascal, J. L. F.; Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123*, 234505.

(52) Luo, Y.; Roux, B. Simulation of osmotic pressure in concentrated aqueous salt solutions. *J. Phys. Chem. Lett.* **2010**, *1*, 183–189.

(53) Brady, J. P.; Farber, P. J.; Sekhar, A.; Lin, Y.-H.; Huang, R.; Bah, A.; Nott, T. J.; Chan, H. S.; Baldwin, A. J.; Forman-Kay, J. D. 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*, E8194–E8203.

(54) Wei, M.-T.; Elbaum-Garfinkle, S.; Holehouse, A. S.; Chen, C. C.-H.; Feric, M.; Arnold, C. B.; Priestley, R. D.; Pappu, R. V.; Brangwynne, C. P. Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. *Nat. Chem.* **2017**,

(55) Bullerjahn, J. T.; von Bülow, S.; Hummer, G. Optimal estimates of self-diffusion coefficients from molecular dynamics simulations. *J. Chem. Phys.* **2020**, *153*, 024116.

(56) Mills, R. Self-diffusion in normal and heavy water in the range 1-45. deg. *J. Phys. Chem.* **1973**, *77*, 685–688.

(57) Forman-Kay, J. D.; Kriwacki, R. W.; Seydoux, G. Phase separation in biology and disease. *J. Mol. Biol.* **2018**, *430*, 4603.

(58) Soranno, A.; Buchli, B.; Nettels, D.; Cheng, R. R.; Müller-Späth, S.; Pfeil, S. H.; Hoffmann, A.; Lipman, E. A.; Makarov, D. E.; Schuler, B. Quantifying Internal Friction in Unfolded and Intrinsically Disordered Proteins with Single-molecule Spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 17800–17806.

(59) Borgia, A.; Borgia, M. B.; Bugge, K.; Kissling, V. M.; Heidarsson, P. O.; Fernandes, C. B.; Sottini, A.; Buholzer, K. J.; Nettels, D.; Kragelund, B. B. et al. Extreme disorder in an ultra-high-affinity protein complex. *Nature* **2018**, *555*, 61–66.

(60) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.

(61) Song, J.; Ng, S. C.; Tompa, P.; Lee, K. A. W.; Chan, H. S. Polycation-π Interactions Are a Driving Force for Molecular Recognition by an Intrinsically Disordered Oncoprotein Family. *PLoS Comput. Biol.* **2013**, *9*, e1003239.

(62) Wang, J.; Choi, J.-M.; Holehouse, A. S.; Lee, H. O.; Zhang, X.; Jahnel, M.; Maharana, S.; Lemaitre, R.; Pozniakovsky, A.; Drechsel, D. et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell* **2018**, *174*, 688–699.

(63) Fromm, S. A.; Kamenz, J.; Nöldeke, E. R.; Neu, A.; Zocher, G.; Sprangers, R. In Vitro Reconstitution of a Cellular Phase-Transition Process that Involves the mRNA Decapping Machinery. *Angew. Chem. Int. Edit.* **2014**, *53*, 7354–7359.

(64) Gabryelczyk, B.; Cai, H.; Shi, X.; Sun, Y.; Swinkels, P. J.; Salentinig, S.; Pervushin, K.; Miserez, A. Hydrogen bond guidance and aromatic stacking drive liquid-liquid phase separation of intrinsically disordered histidine-rich peptides. *Nat. Commun.* **2019**, *10*, 1–12.

(65) Lin, Y.; Currie, S. L.; Rosen, M. K. Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs. *J. Biol. Chem.* **2017**, *292*, 19110–19120.

(66) Cross, S.; Kuttel, M. M.; Stone, J. E.; Gain, J. E. Visualisation of cyclic and multi-branched molecules with VMD. *J. Mol. Graph. Model.* **2009**, *28*, 131–139.

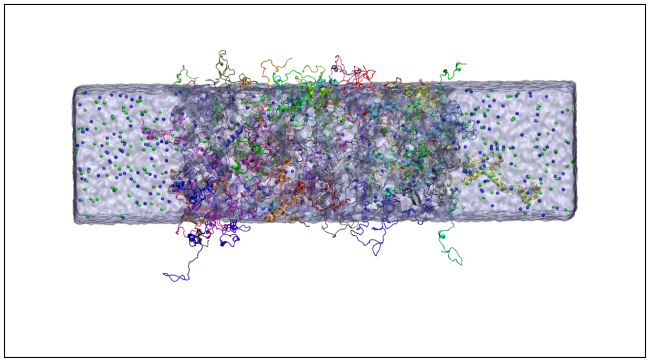
(67) Dignon, G. L.; Zheng, W.; Mittal, J. Simulation methods for liquid–liquid phase separation of disordered proteins. *Curr. Opin. Chem. Eng.* **2019**, *23*, 92–98.

(68) Baldwin, R. L. How Hofmeister ion interactions affect protein stability. *Biophys. J.* **1996**, *71*, 2056–2063.

(69) Vancraenenbroeck, R.; Harel, Y. S.; Zheng, W.; Hofmann, H. Polymer effects modulate binding affinities in disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 19506–19512.

(70) Welsh, T. J.; Krainer, G.; Espinosa, J. R.; Joseph, J. A.; Sridhar, A.; Jahnel, M.; Arter, W. E.; Saar, K. L.; Alberti, S.; Collepardo-Guevara, R. et al. Single particle zeta-potential measurements reveal the role of electrostatics in protein condensate stability. *bioRxiv* **2020**, *https://doi.org/10.1101/2020.04.20.047910*.

Graphical TOC Entry



Supporting Information:

Molecular details of protein condensates probed by microsecond-long atomistic simulations

Wenwei Zheng,[*,†,#] Gregory L. Dignon,[*,‡,#] Nina Jovic,[‡] Xichen Xu,[‡] Roshan M. Regy,[‡] Nicolas L. Fawzi,[§] Young C. Kim,[*,‖] Robert B. Best,[*,⊥] and Jeetain Mittal[*,‡]

†*College of Integrative Sciences and Arts, Arizona State University, Mesa, Arizona, United States*

‡*Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, Pennsylvania, United States*

¶*Current Address: Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York, United States,*

§*Department of Molecular Pharmacology, Physiology and Biotechnology, Brown University, Providence, RI 02912, USA*

‖*Center for Materials Physics and Technology, Naval Research Laboratory, Washington, DC 20375*

⊥*Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland, United States*

#*These authors contributed equally to this work*

Supporting Methods

Setup of all-atom slab simulations

The challenge in modeling the condensed phase of an IDP at atomic resolution is to equilibrate the density of the condensed phase. Even using the coarse-grained (CG) model, it takes about 1 μs simulation time to achieve the task.[S1] Here by using all-atom model, this is expected to be much slower considering the explicit solvent and a much greater number of degrees of freedom involved in the simulation. We therefore explore an alternative approach by initializing the all-atom simulation from a reasonable well-equilibrated CG conformation.

Using FUS LC as an example, we first generated an initial configuration of the FUS chains with our recently developed CG model–HPS model.[S1] We followed the same protocol used in the CG modeling by setting up a slab geometry, in which one box dimension is elongated, allowing for a semi-infinite condensed phase in two dimensions, and two flat interfaces along the elongated dimension (see Fig. 1A). This geometry has been widely used,[S2,S3] and demonstrated to be comparable with other accepted methods of sampling multiple phases, such as grand canonical Monte Carlo[S4] while reducing the effects of a finite-sized spherical droplet.[S5] The other advantage of the slab geometry in the case of explicit solvent is that it greatly reduces the box volume that is occupied largely by aqueous phase, compared to a brute-force droplet simulation. To reduce the overall box size, we have conducted CG simulations of a slab of FUS chains containing a different number of IDP chains, and at different box cross sections in order to obtain a system that has a sufficient fraction of the chains in the condensed phase region, rather than interfacial region (Fig. S1). The number of chains (40 here) included in the simulations have been carefully selected so that it is large enough to still reproduce exactly the density profile of our original CG simulations with 100 chains at the same temperature, whereas at the same time to be as small as possible to reduce the system size so that the simulation is still feasible.

With a reasonable slab configuration in CG representation as the starting point, we further prepared the all-atom configuration following the steps shown in the main text and Fig. 1. Finally, the system is solvated with explicit water molecules and ions, and further equilibrated using standard molecular dynamics engines (Fig. 1D). The equilibration is conducted using GROMACS 2018.[S6] Langevin dynamics was performed with a time step of 2 fs and a friction coefficient of 1 ps^{-1}. Berendsen pressure coupling was used for 10 ns followed by Parrinello-Rahman pressure coupling[S7] for 200 ns simulations. Lennard-Jones (LJ) pair interactions were cut off at 0.9 nm. Electrostatic energies were computed using particle-mesh Ewald[S8] with a grid spacing of 0.12 nm and a real-space cutoff of 0.9 nm. The protein force field was Amber ff03ws;[S9] the water model was TIP4P/2005[S10] and the ion parameters were from Luo et al.[S11] The final system we prepared for FUS contains 40 chains of the 163-residue protein, solvated in 98912 water molecules,[S10] 280 Na+ ions and 200 Cl- ions to neutralize the system net charge, and represent an ionic strength of about 100 mM, with a total of 485408 atoms (see Table S1).

Further 150 ns equilibration simulations using an NPT ensemble were conducted on Anton 2–a specialized supercomputer[S12] together with two production simulations, one for FUS LC and one for LAF-1 RGG for a total of 2 μs each using the same aforementioned force fields. We need to note that there have been a series of force fields developed for IDPs and can be used for similar simulations.[S9,S13–S15] The Amber ff03ws force fields have been demonstrated to capture the configurations and dynamics of disordered proteins.[S16,S17] It should be noted that performance on traditional resources, while one or two order of magnitude slower than on Anton 2, should be sufficiently fast to generate microseconds worth of data within a reasonable time frame. For instance we obtained the initial equilibration with a benchmark of about 45 ns/day with 896 Xeon E5-2680 CPUs. This makes such a 2 μs simulation feasible even using traditional computational resources. Please see Table S1 for simulation details and benchmarks.

Sequences of the proteins used in this work

N-terminal low complexity domain of the FUS protein (FUS LC)

MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG YSQSTDTSGY GQSSYSSYGQ
SQNTGYGTQS TPQGYGSTGG YGSSQSSQSS YGQQSSYPGY GQQPAPSSTS GSYGSSSQSS
SYGQPQSGSY SQQPSYGGQQ QSYGQQQSYN PPQGYGQQNQ YNS

N-terminal disordered domain of the LAF-1 protein (LAF-1 RGG)

MESNQSNNGG SGNAALNRGG RYVPPHLRGG DGGAAAAASA GGDDRRGGAG GGGYRRGGGN
SGGGGGGGYD RGYNDNRDDR DNRGGSGGYG RDRNYEDRGY NGGGGGGGNR GYNNNRGGGG
GGYNRQDRGD GGSSNFSRGG YNNRDEGSDN RGSGRSYNND RRDNGGDG

Calculation of diffusion coefficients

To estimate the diffusion coefficients in the simulation, we first compute the probability distributions (propagators) $P(\xi(t+t_0)-\xi(t_0))$ for molecular displacement in each direction $\xi = x, y, z$ as a function of the lag time t between observations to accurately describe the dynamics of protein, water, and ions (i.e. Na+ and Cl-). We then fit the probability distribution function at a specific lag time to the 1D diffusion equation with multiple diffusivity values as

$$P(x,t) = \Sigma_{i=1}^{n} p_i \frac{1}{\sqrt{4\pi D_{x,i} t}} \exp(\frac{-x^2}{4D_{x,i}t}), \tag{S1}$$

and

$$\Sigma_{i=1}^{n} p_i = 1. \tag{S2}$$

We increase the number (n) of diffusivity (D) values needed for fitting the probability distribution until the fitting is reasonably good as shown in Fig. S4. We find that for the protein, one D value ($n = 1$) is sufficient whereas for the water and ions, we need three D values ($n = 3$) for some of the lag time. We then check D as a function of lag time until the D values are plateaued at a range of the lag time. The average values of D at that window are reported (Table S2). The lag time window for calculating the protein diffusion

coefficients are from 408 to 720 ns and that for water and ions are from 3.6 to 6 ns.

Contact map calculation

We started with calculating the average number of contacts per frame of our simulation trajectories between all pairs of groups (i.e. backbone, sidechain of, or entire amino acids). A contact between two groups were considered if at least one atom from each of the two groups are within 6 Å distance.

The first contact map we would like to show is the residue contact map (163 residues for FUS LC and 168 residues for LAF-1 RGG), which suggests the contribution of specific residues in a protein sequence to stabilizing the condensates. Since there are 40 chains in the simulation, the residue contact map can be split into the intermolecular and intramolecular contributions. For the intramolecular residue contact map, we calculated the summation of the 40 intramolecular contact maps from 40 chains (Fig. S7a and b). We only considered the intramolecular contacts with a sequence separation of larger than 3, that is, $|i - j| > 3$ in which i and j represent the amino acid index in the sequence. For the intermolecular residue contact map, we calculated the summation from the $40 \times 39/2$ number of intramolecular contact maps from pairwise contributions of 40 chains. Permutation between two chains was excluded to avoid counting the same contact twice. With this consideration, the residue intermolecular contact map (Fig. 4a and b) is not symmetric. The contact between the i-th and j-th residues can come from two different molecules and therefore is different from that between the j-th and i-th residues. We normalized both intramolecular and intermolecular contact maps by the number of chains so as to consider the system size effect, since the total number of contacts of the system approximately scales with the number of chains in the simulation.

To further understand the contribution of specific residue types to stabilizing the condensates in a protein sequence, we would like to calculate residue-type contact map from the residue contact map. For the intramolecular version, we considered the

contribution between i-th and j-th residues if $i > 3$, due to the fact that the contact between i-th and j-th residues is the same as that between j-th and i-th residues (Fig. S7c and d). For the intermolecular residue-type contact map however, the contribution to the contact between the same pair of residue type could come from the contact both between the i-th and j-th residues and between the j-th and i-th residues since they represent different contacts in the simulation as discussed in the previous step. As we are more interested in the pair of residue types instead of the permutation of the two residue types, we added up the number of contacts from two terms: contacts between i-th and j-th residues and contacts between j-th and i-th residues (Fig. 4c and d). Such consideration guarantees that the total number of contacts from the intermolecular residue-type contact map still matches the average total number of contacts per frame of the trajectory.
At last we would like to ask the role of residue-type pairs without the contribution of the number of times these residue-type pairs appear in the protein sequence. We therefore normalized both our intramolecular and intermolecular contact maps with the largest number of contacts that can be formed between the two types of residues for a specific protein sequence. Due to limited statistics, we only show the contacts for residues that appear more than twice in the sequence. Practically for the normalized intramolecular residue-type contact map, the normalization factor is the product of the numbers of residues (Fig. S7e and f). For the normalized intermolecular residue-type contact map, the normalization factor is the square of the number of residues from the same types of residues, and two times the product of the numbers of residues from the two different residue types due to permutation (Fig. 4e and f).

Extracting molecular interactions from all-atom simulations

Hydrogen bond. MDAnalysis is used in calculating the hydrogen bond formation[S18] with a distance cutoff of 3.0 Å between the donor and acceptor and an angle cutoff of 120° between donor, hydrogen and acceptor.

sp^2/π interaction. We calculated the sp^2/π interactions based on the definition of a recent literature[S19] with small modification of the algorithm for efficiency. First we filtered all the pairs of sp^2/π groups by using a cutoff of 8 Å on the distance between the center of mass of the two groups. Second we calculated cosine angles between the normal vectors of the two plains and only kept the groups with absolute values of cosine angles larger than 0.8. Third both the plains defined by each group were raised by 1.5 Å and the distance between the center of mass of the two new plains were calculated. The pairs with the center of mass distances less than 4 Å were selected as forming the sp^2/π interactions.

Cation-π interaction. Similar to the sp2/π interactions, Cation-π interactions are also defined by using both a distance and an angle criterion. The distance between the charged nitrogen in the cationic side chain and the center of mass of the π group is first subjected to a cutoff of 6 Å. The absolute cosine angles between the normal vector of the π plain and the vector linking the charged nitrogen and the center of mass of the π group is further subjected to a cutoff of 0.8.

Salt bridge. We used a distance cutoff of 6 Å on the smallest distance between all charged nitrogen and oxygen atoms for every pair of charged amino acids to determine the formation of the salt bridge in our simulations.

Supporting Figures

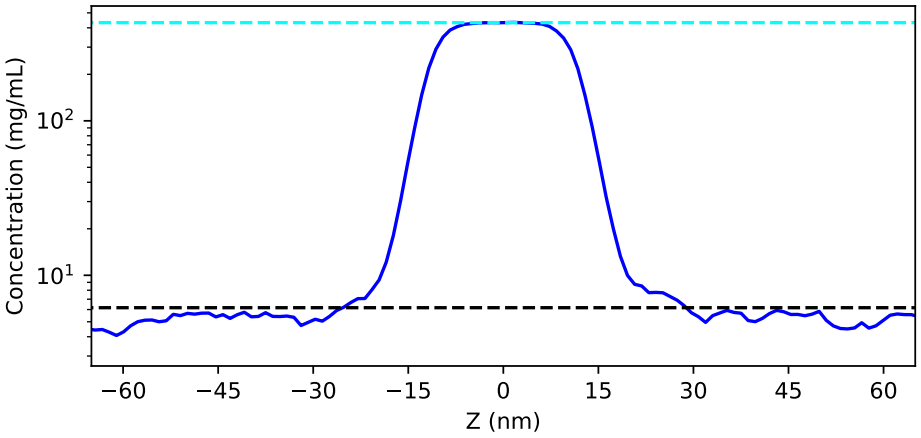


Figure S1: Test of CG simulations of FUS LC with 40 chains in slab geometry using a 10 nm box cross section at 340K compared to reference densities from a 100-chain slab simulation with a 15 nm box cross section as used in references.[S1,S5] The black dashed line indicates the low density phase of the reference and cyan dashed line indicates the high density phase of the reference.

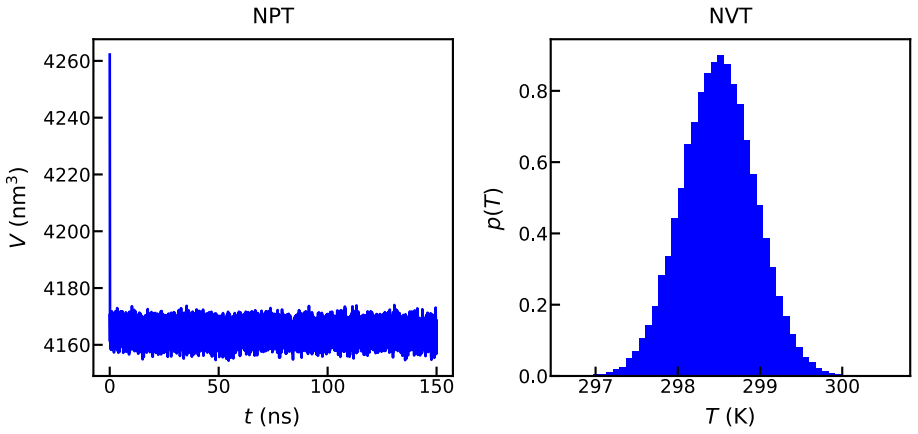


Figure S2: The volume fluctuation of NPT equilibration and temperature distribution of productive NVT simulation for LAF-1 RGG using Anton 2.

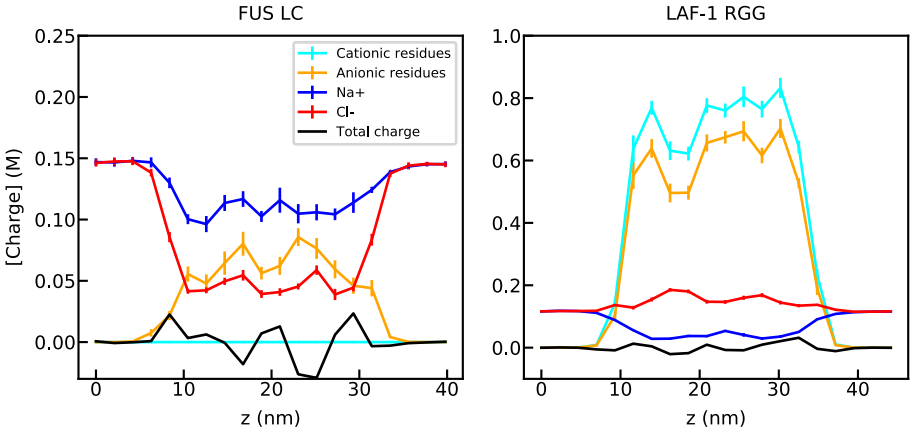


Figure S3: Density profiles of charges from all-atom slab simulations of FUS LC (left) and LAF-1 RGG (right). Components are shown in the legend.

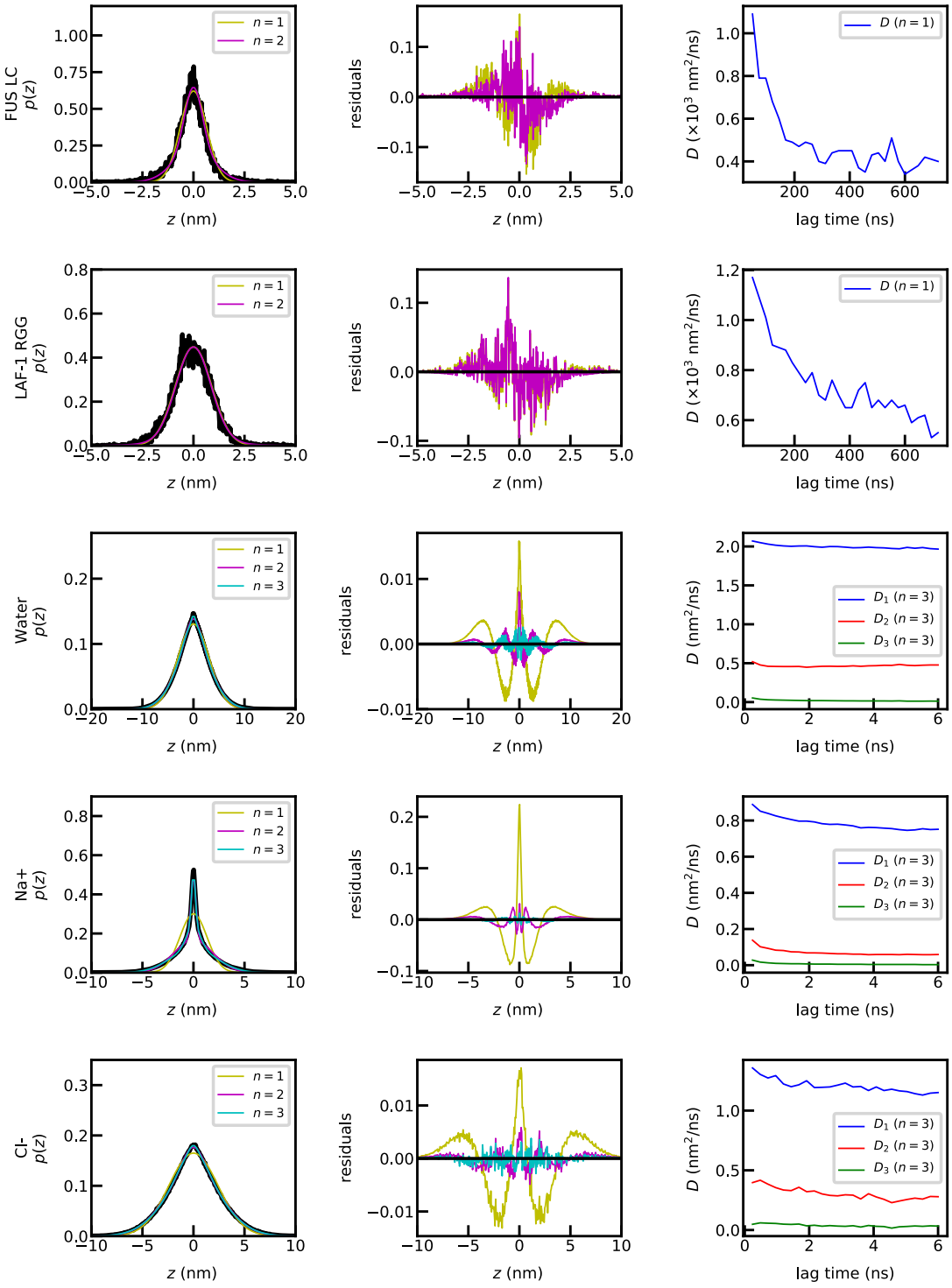


Figure S4: Representative distribution of mean square displacement at lag time of 600 ns for proteins and 3.6 ns for water and ions. The black lines show the results from simulations and the color lines show different strategies of fitting.

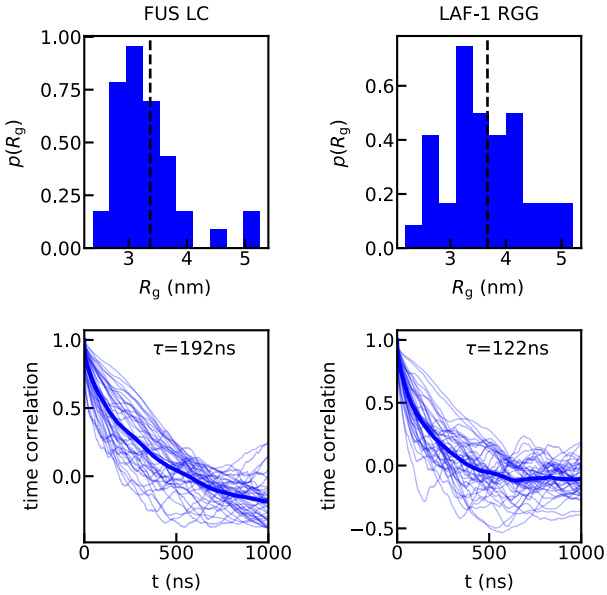


Figure S5: Radii of gyration (R_g) of FUS LC (left) and LAF-1 RGG (right). Top: The distribution of the average R_g of each of the 40 chains in the simulation. Bottom: Time correlation of R_g for each of the 40 chains (thin lines) and the average over the 40 chains (thick line). The relaxation time obtained by fitting the average time correlation to a single exponential function is shown in the legend.

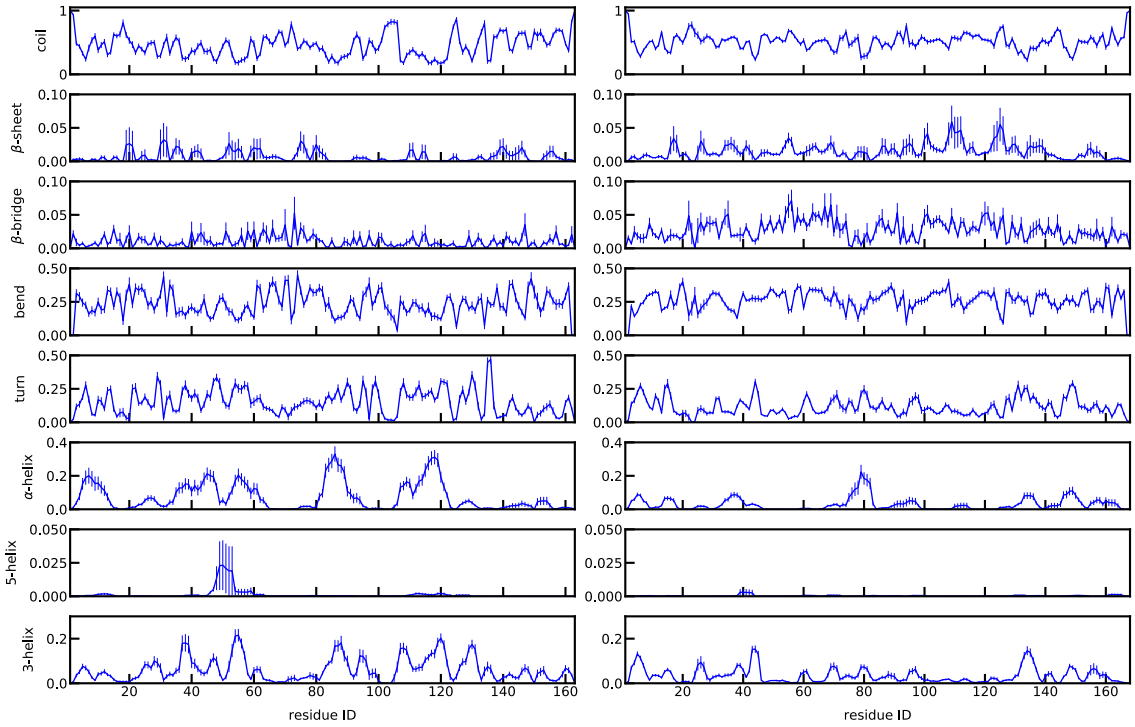


Figure S6: Secondary structures of FUS LC (left) and LAF-1 RGG (right) using DSSP.[S20]

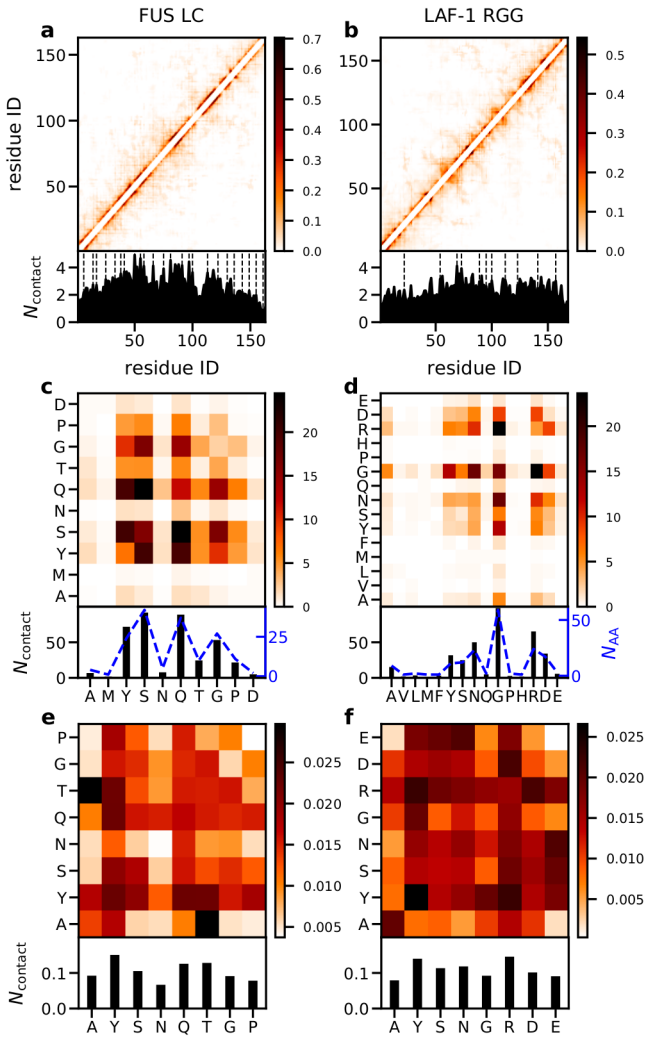


Figure S7: Intramolecular contacts within the condensed phase of FUS LC (left) and LAF-1 RGG (right), as a function of residue index (a and b) and amino acid types (c and d). The intramolecular contacts normalized by the relative abundance of each amino acid in the sequence are shown in e and f. In each of the figure, the bottom panel shows the one dimensional summation. The black dashed lines in a and b sho the position of Tyr residues. The blues lines in c and d show the number of amino acids (N_{AA}) in the sequence.

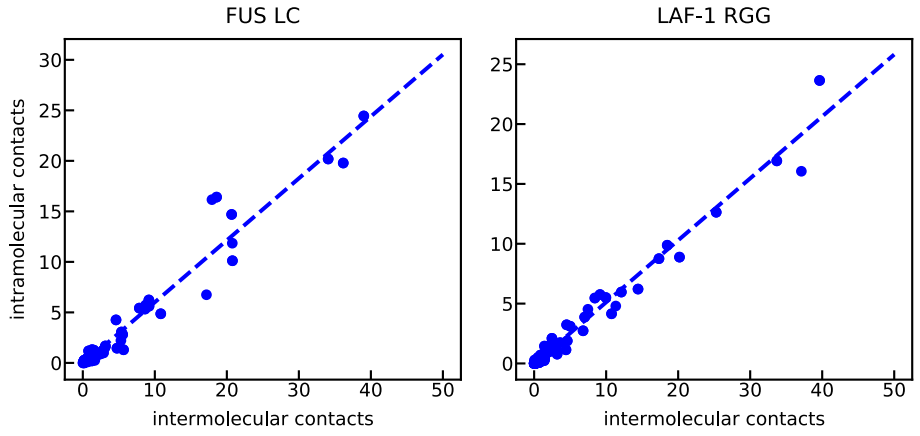


Figure S8: Correlation between intra- and inter-molecular contacts for each pair of amino acids within the condensed phase of FUS LC (left) and LAF-1 RGG (right).

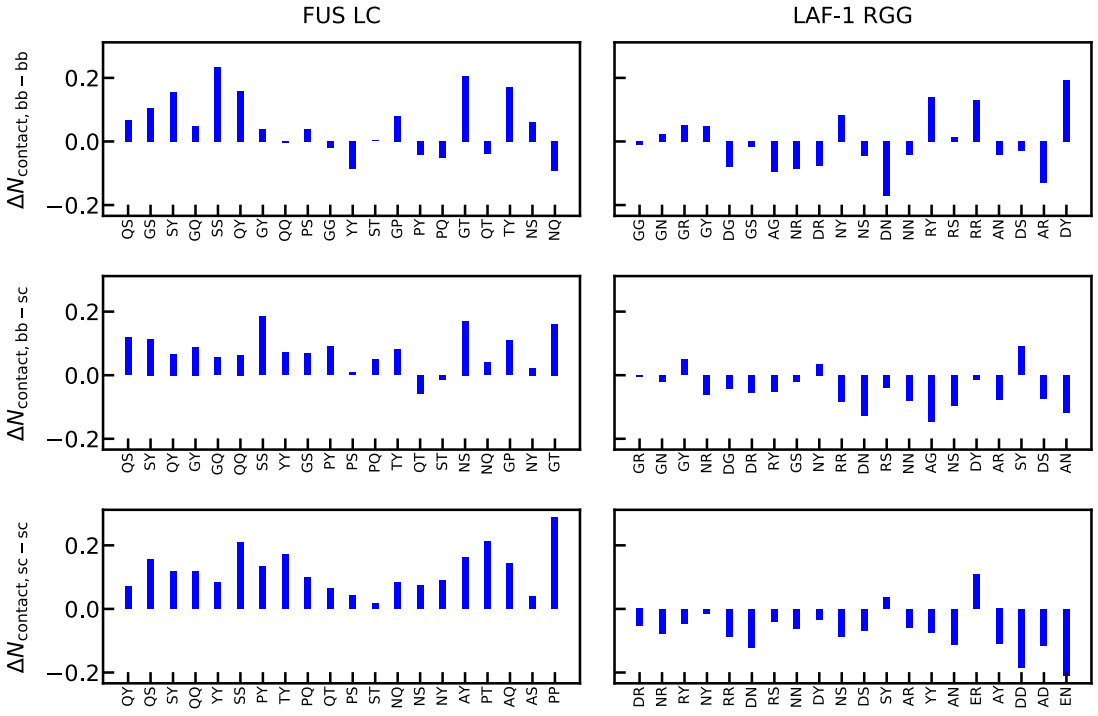


Figure S9: The relative difference ($\Delta N = (N_{\mathrm{all-atom}} - N_{\mathrm{initial}})/N_{\mathrm{initial}}$) between the number of intermolecular contacts shown in Fig. S11 averaged over the all-atom simulation and the initial configuration reconstructed from the coarse-grained simulation.

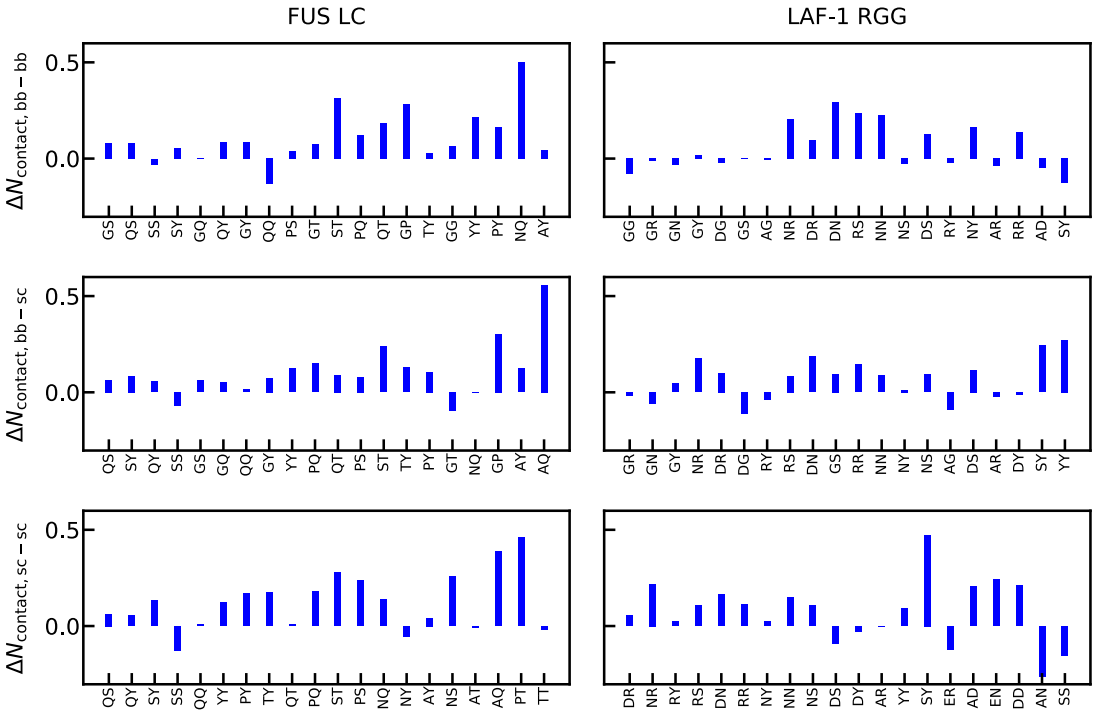


Figure S10: The relative difference ($\Delta N = (N_{\text{all-atom}} - N_{\text{initial}})/N_{\text{initial}}$) between the number of intramolecular contacts shown in Fig. S13 averaged over the all-atom simulation and the initial configuration reconstructed from the coarse-grained simulation.

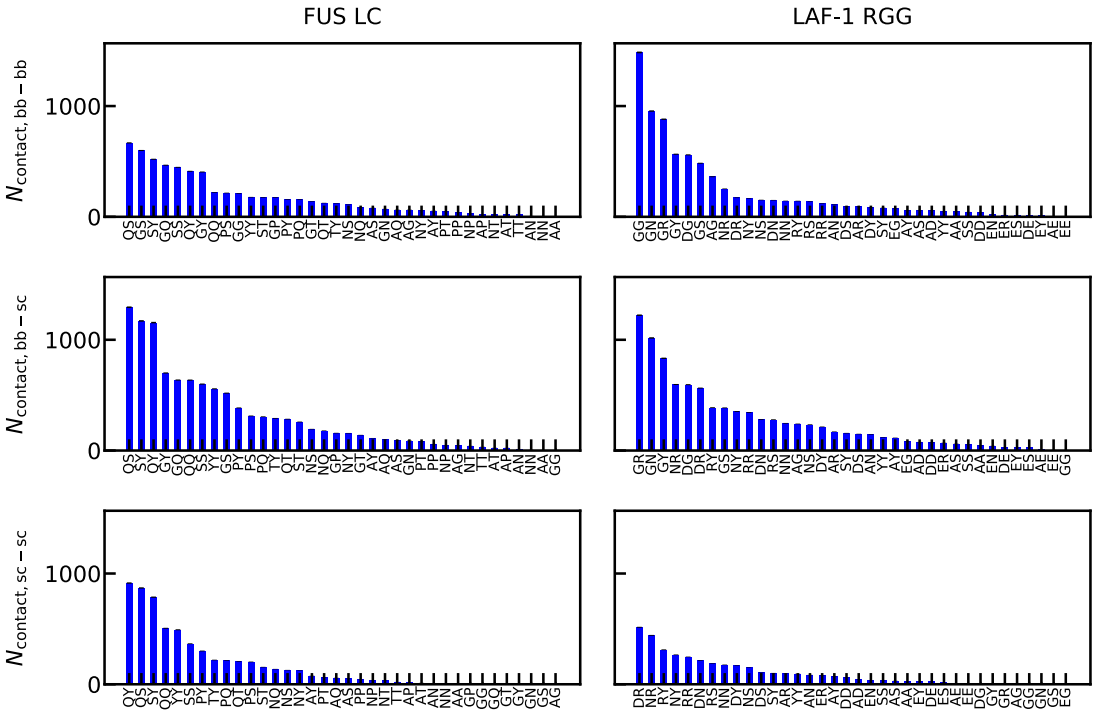


Figure S11: Number of intermolecular contacts for specific pairs of amino acids between backbone and backbone atoms (bb-bb), backbone and sidechain atoms (bb-sc), and sidechain and sidechain atoms (sc-sc).

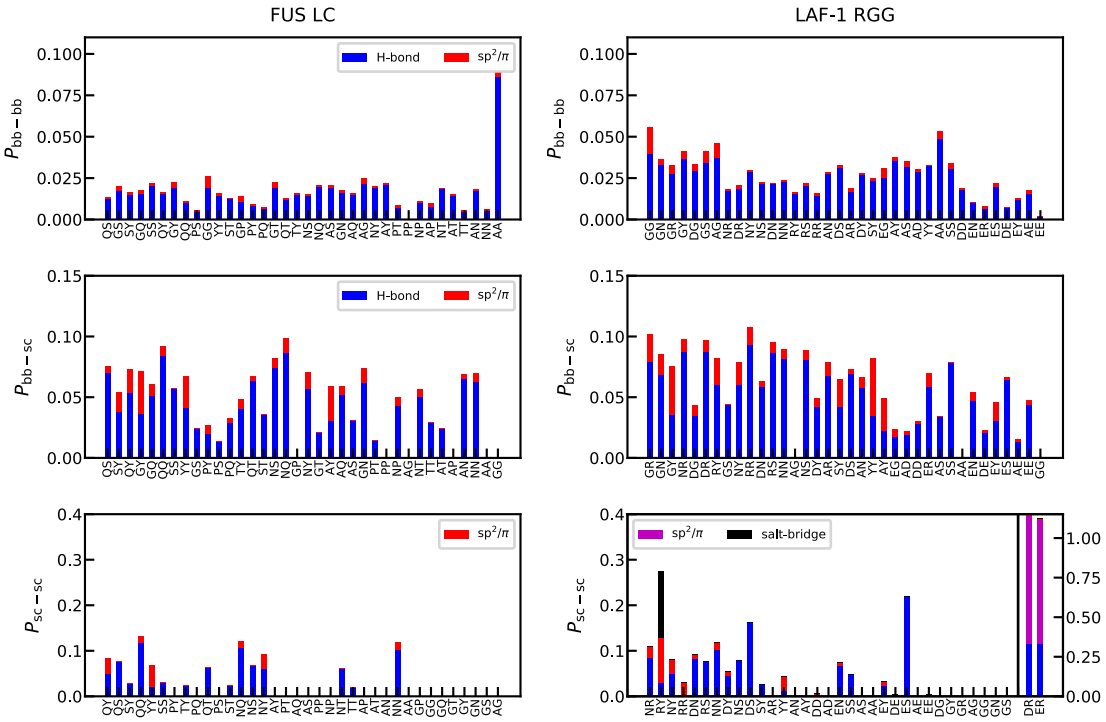


Figure S12: Interaction modes contributing to the intermolecular contacts.

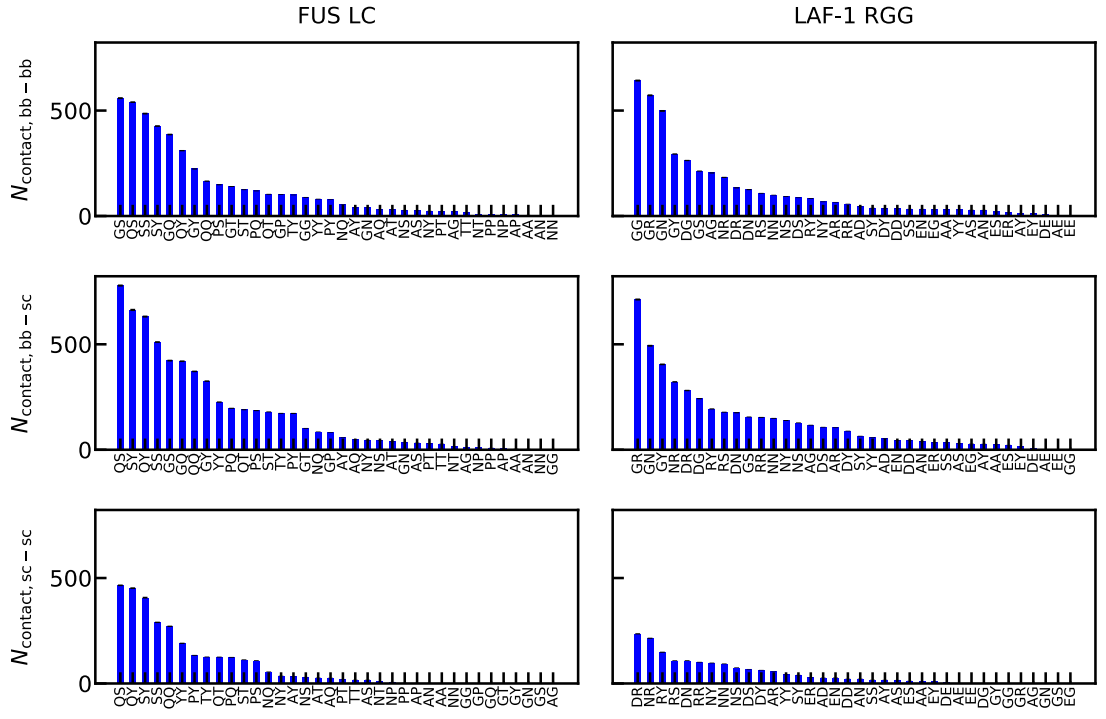


Figure S13: Number of intramolecular contacts for specific pairs of amino acids.

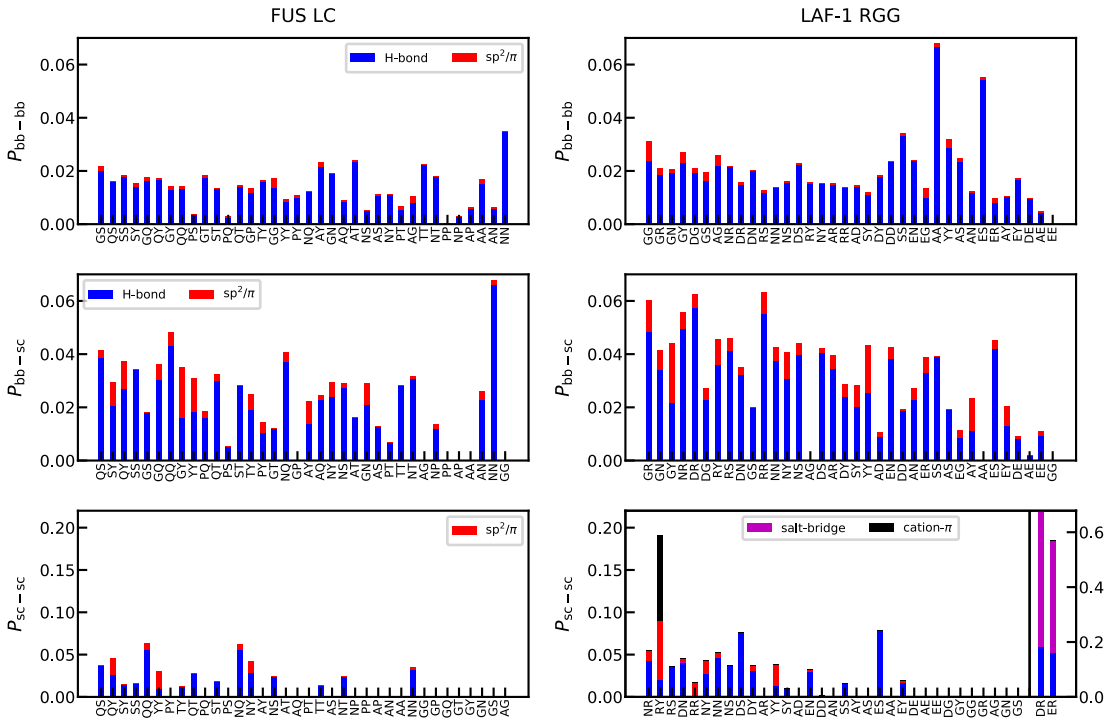


Figure S14: Interaction modes contributing to the intramolecular contacts.

Supporting Tables

Table S1: System sizes of all-atom simulations. *The benchmark of local simulations is estimated by using Gromacs 2018.3[S21] and 896 CPUs of Xeon E5-2680.

Name	FUS LC	LAF-1 RGG
number of chains	40	40
number of amino acids per chain	163	168
number of water molecules	98912	114727
number of Na+	280	188
number of Cl-	200	348
total number of atoms	485408	546884
equlibrated box dimension (nm)	$9.7 \times 9.7 \times 39.8$	$9.7 \times 9.7 \times 44.1$
ionic strength (mM)	106	107
local equilibration benchmark	46 ns/day*	44 ns/day*
local equilibration length	210 ns	210 ns
Anton 2[S12] benchmark	4.4 μs/machine day	4.2 μs/machine day
Anton 2 equilibration	0.15 μs NPT	0.15 μs NPT
Anton 2 simulation length	2 μs NVT	2 μs NVT

Table S2: Diffusion coefficients of proteins, water and ions along z-axis. See supporting methods for details.

$n = 1$	$D(\times 10^3 \mathrm{nm}^2/\mathrm{ns})$					
FUS LC	0.404					
LAF-1 RGG	0.642					
$n = 3$	$D_1(\mathrm{nm}^2/\mathrm{ns})$	p_1	$D_2(\mathrm{nm}^2/\mathrm{ns})$	p_2	$D_3(\mathrm{nm}^2/\mathrm{ns})$	p_3
Water (FUS LC)	1.980	0.713	0.472	0.279	0.014	0.008
Water (LAF-1 RGG)	2.123	0.710	0.505	0.286	0.012	0.005
Na+ (FUS LC)	0.754	0.645	0.059	0.228	0.004	0.127
Na+ (LAF-1 RGG)	0.854	0.777	0.063	0.159	0.003	0.064
Cl- (FUS LC)	1.164	0.763	0.265	0.217	0.030	0.020
Cl- (LAF-1 RGG)	1.143	0.559	0.202	0.390	0.020	0.051

References

(S1) Dignon, G. L.; Zheng, W.; Kim, Y. C.; Best, R. B.; Mittal, J. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* **2018**, *14*, e1005941.

(S2) Blas, F. J.; MacDowell, L. G.; de Miguel, E.; Jackson, G. Vapor-liquid interfacial properties of fully flexible Lennard-Jones chains. *J. Chem. Phys.* **2008**, *129*, 144703.

(S3) Silmore, K. S.; Howard, M. P.; Panagiotopoulos, A. Z. Vapour–liquid phase equilibrium and surface tension of fully flexible Lennard–Jones chains. *Mol. Phys.* **2017**, *115*, 320–327.

(S4) Jung, H.; Yethiraj, A. A simulation method for the phase diagram of complex fluid mixtures. *J. Chem. Phys.* **2018**, *148*, 244903.

(S5) Dignon, G. L.; Zheng, W.; Best, R. B.; Kim, Y. C.; Mittal, J. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, 9929–9934.

(S6) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(S7) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(S8) Darden, T.; York, D.; Pedersen, L. Particle mesh ewald: an $N - \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(S9) Best, R. B.; Zheng, W.; Mittal, J. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* **2014**, *10*, 5113–5124.

(S10) Abascal, J. L. F.; Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123*, 234505.

(S11) Luo, Y.; Roux, B. Simulation of osmotic pressure in concentrated aqueous salt solutions. *J. Phys. Chem. Lett.* **2009**, *1*, 183–189.

(S12) Shaw, D. E.; Grossman, J.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H., et al. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. Proceedings of the international conference for high performance computing, networking, storage and analysis. 2014; pp 41–53.

(S13) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B* **2015**, *119*, 5113–5123.

(S14) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71–73.

(S15) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E4758–E4766.

(S16) Gül Zerze,; Mittal, J.; Best, R. B. Diffusive dynamics of contact formation in disordered polypeptides. *Phys. Rev. Lett.* **2016**, 068102.

(S17) Zerze, G. H.; Zheng, W.; Best, R. B.; Mittal, J. Evolution of All-atom Protein Force Fields to Improve Local and Global Properties. *J. Phys. Chem. Lett.* **2019**, *10*, 2227.

(S18) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: a

toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **2011**, *32*, 2319–2327.

(S19) Vernon, R. M.; Chong, P. A.; Tsang, B.; Kim, T. H.; Bah, A.; Farber, P.; Lin, H.; Forman-Kay, J. D. Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife* **2018**, *7*, e31486.

(S20) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.

(S21) Kutzner, C.; Páll, S.; Fechner, M.; Esztermann, A.; de Groot, B. L.; Grubmüller, H. More bang for your buck: Improved use of GPU nodes for GROMACS 2018. *J. Comput. Chem.* **2019**, *40*, 2418–2431.