Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties

Giulio Tesei[a,1], Thea K. Schulze[a], Ramon Crehuet[a,b], and Kresten Lindorff-Larsen[a,1]

[a]Structural Biology and NMR Laboratory, Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark; and [b]CSIC-Institute for Advanced Chemistry of Catalonia (IQAC), E-08034 Barcelona, Spain

Edited by Susan Marqusee, University of California, Berkeley, CA, and approved September 15, 2021 (received for review June 24, 2021)

Many intrinsically disordered proteins (IDPs) may undergo liquid–liquid phase separation (LLPS) and participate in the formation of membraneless organelles in the cell, thereby contributing to the regulation and compartmentalization of intracellular biochemical reactions. The phase behavior of IDPs is sequence dependent, and its investigation through molecular simulations requires protein models that combine computational efficiency with an accurate description of intramolecular and intermolecular interactions. We developed a general coarse-grained model of IDPs, with residue-level detail, based on an extensive set of experimental data on single-chain properties. Ensemble-averaged experimental observables are predicted from molecular simulations, and a data-driven parameter-learning procedure is used to identify the residue-specific model parameters that minimize the discrepancy between predictions and experiments. The model accurately reproduces the experimentally observed conformational propensities of a set of IDPs. Through two-body as well as large-scale molecular simulations, we show that the optimization of the intramolecular interactions results in improved predictions of protein self-association and LLPS.

intrinsically disordered proteins | liquid–liquid phase separation | force field parameterization | biomolecular condensates | protein interactions

Many intrinsically disordered proteins (IDPs) and proteins with disordered regions can condense into liquid-like droplets, namely, a biomolecule-rich phase coexisting with a more dilute solution (1–5). This demixing process is known as liquid–liquid phase separation (LLPS) and is one of the ways cells compartmentalize proteins, often together with nucleic acids (6). While LLPS plays crucial biological roles in the cell, its dysregulation leads to maturation of biomolecular condensates into hydrogel-like assemblies, promoting the formation of neurotoxic oligomers and amyloid fibrils (5,7). A quantitative model for the "molecular grammar" of LLPS, including the influence of disease-associated mutations and posttranslational modifications (PTMs) on the propensity to phase separate, is key to understand these processes. The sequences of IDPs and intrinsically disordered regions that easily undergo LLPS are often characterized by stretches enriched in small polar residues (spacers) interspersed by, e.g., aromatic or arginine residues (stickers), which are instrumental for the formation of reversible physical cross-links via π–π, cation–π, and sp^2–π interactions (8–12). Y and R residues were shown to be necessary for the LLPS of a number of proteins including FUS, hnRNPA1, LAF-1, and Ddx4 (8, 10, 11, 13–17). While the propensity to undergo LLPS increases with the number of Y residues in the sequence, recent studies have revealed that the role of R residues is context dependent (16) and strongly affected by salt concentration (17), reflecting the unusual characteristics of the R side chain (18, 19).

Here we present the development of a coarse-grained (CG) model capable of predicting the phase behavior of IDPs based on amino acid sequence. CG models enable the combination of a sequence-dependent description with the computational efficiency necessary to explore the long time and large length scales involved in phase transitions (11, 20, 21). Although CG molecular simulations have been employed to explain the sequence dependence of the LLPS of a number of IDPs (11, 15, 17, 20–22) as well as the effect of phosphorylation on LLPS propensities (23, 24), such models have proven difficult to use to predict the phase behavior of very diverse sequences (25). Building on recent developments, including experimental phase diagrams of a number of IDPs (3, 4, 15, 16), we trained and tested a robust sequence-dependent model of the LLPS of IDPs. In particular, due to the similarity between intramolecular interactions within IDPs and intermolecular interactions between IDPs (12, 26), we reasoned that by optimizing a model to capture structural preferences for a broad set of monomeric IDPs, we could obtain a good model for interactions between IDPs.

The starting point for our analyses is the hydrophobicity scale (HPS) model (21) (with minor modification; *SI Appendix*) wherein, besides steric repulsion and salt-screened charge–charge interactions, residue–residue interactions are determined by hydropathy parameters (λ) which were derived from the

Significance

Cells may compartmentalize proteins via a demixing process known as liquid–liquid phase separation (LLPS), which is often driven by intrinsically disordered proteins (IDPs) and regions. Protein condensates arising from LLPS may develop into insoluble protein aggregates, as in neurodegenerative diseases and cancer. Understanding the process of formation, dissolution, and aging of protein condensates requires models that accurately capture the underpinning interactions at the residue level. In this work, we leverage data from biophysical experiments on IDPs in dilute solution to develop a sequence-dependent model which predicts conformational and phase behavior of diverse and unrelated protein sequences with good accuracy. Using the model, we gain insight into the coupling between chain compaction and LLPS propensity.

Author contributions: G.T., R.C., and K.L.-L. designed research; G.T., T.K.S., and R.C. performed research; G.T., T.K.S., and R.C. contributed new reagents/analytic tools; G.T., T.K.S., and K.L.-L. analyzed data; G.T. and K.L.-L. wrote the paper; and T.K.S. and R.C. edited the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: giulio.tesei@bio.ku.dk or lindorff@bio.ku.dk.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2111696118/-/DCSupplemental.

Published October 29, 2021.

atomic partial charges of a classical all-atom force field (27). Recently, the development of the HPS-Urry model (28) presented substantial improvements in accuracy over the original HPS model. These were achieved using a hydrophobicity scale derived from transition temperatures of elastin-like peptides (29) and further shifting the λ parameters by -0.08 to improve agreement with experimentally measured radii of gyration.

To address the current limitations, we improve upon these models by optimizing the λ parameters through a Bayesian parameter-learning procedure (30–33), leveraging as prior knowledge the probability distribution of the λ parameters evaluated from analyzing 87 hydrophobicity scales. The training set comprises small-angle X-ray scattering (SAXS) and paramagnetic relaxation enhancement (PRE) NMR data of 45 IDPs which we selected from the literature. First, we run Langevin dynamics simulations of single IDPs and estimate the experimental observables using state-of-the-art methods (34). Second, we employ a Bayesian regularization approach to prevent overfitting the training data and select three models which are equally accurate with respect to single-chain conformational properties. Third, through two-chain simulations, we validate the models by comparing predicted and experimental intermolecular PRE NMR data for the low-complexity domain (LCD) of the heterogeneous nuclear ribonucleoprotein (hnRNP) A2 (A2 LCD) (22) and the LCD of the RNA-binding protein fused in sarcoma (FUS LCD) (23). Fourth, we perform coexistence simulations to test the models against the phase behavior of A2 LCD (22, 24); FUS LCD (35, 36); variants of hnRNP A1 LCD (A1 LCD) (15, 16); the N-terminal region of the germ-granule protein Ddx4 (Ddx4 LCD) (8, 10, 13); and the N-terminal, R-/G-rich domain of the P granule protein LAF-1 (LAF-1 RGG domain). We use the final model to provide insight into the interactions between IDPs within condensates and to help elucidate the role of different amino acids to the driving force for LLPS.

Results and Discussion

Analysis of Hydrophobicity Scales. The λ values of the original HPS model are based on a hydrophobicity scale derived by Kapcha and Rossky from the atomic partial charges of an all-atom force field (27). Dozens of amino acid hydrophobicity scales have been derived from experimental as well as bioinformatics approaches such as the partitioning of amino acids between water and organic solvent, the partitioning of peptides to the lipid membrane interface, and the accessible surface area of residues in folded proteins (37, 38). To carry out the Bayesian optimization of the amino acid specific λ values, we sought to estimate the prior probability distribution of the hydropathy parameters from the analysis of 98 hydrophobicity scales collected by Simm et al. (38). Each scale was minimum–maximum (min–max) normalized, and after ranking in the ascending order of the HPS scale, we discarded all the scales yielding a linear fit with negative slope. This procedure allowed us to identify scales which were present in the set both in their original form and as the additive inverse of the hydropathy values (reversed scales). For most scales, the selection criterion resulted in discarding the reversed form. However, for scales where the most negative values of the hydropathy parameter correspond to the most hydrophobic amino acids–such as the scales by Bull and Breese (39), Guy (40), Bishop et al. (41). and Welling et al. (42)–we retained only the reversed form. The 87 scales that remained after this filtering were used to calculate the average scale (AVG) and the probability distribution of the λ values for the 20 amino acids, $P(\lambda)$, which is normalized so that $\sum_{aa} \int_{\lambda_{aa}=0}^{\lambda_{aa}=1} P(\lambda_{aa})\, \mathrm{d}\lambda_{aa} = 20$ (Fig. 1*A*). For the optimization described below we use the AVG scale as starting point, as well as an indication of the typical accuracy obtained from the prior knowledge encoded in $P(\lambda)$.

We assessed the HPS, HPS-Urry, and AVG parameter sets by running simulations of 45 IDPs ranging in length between 24 and 334 residues and compared the results against experiments.

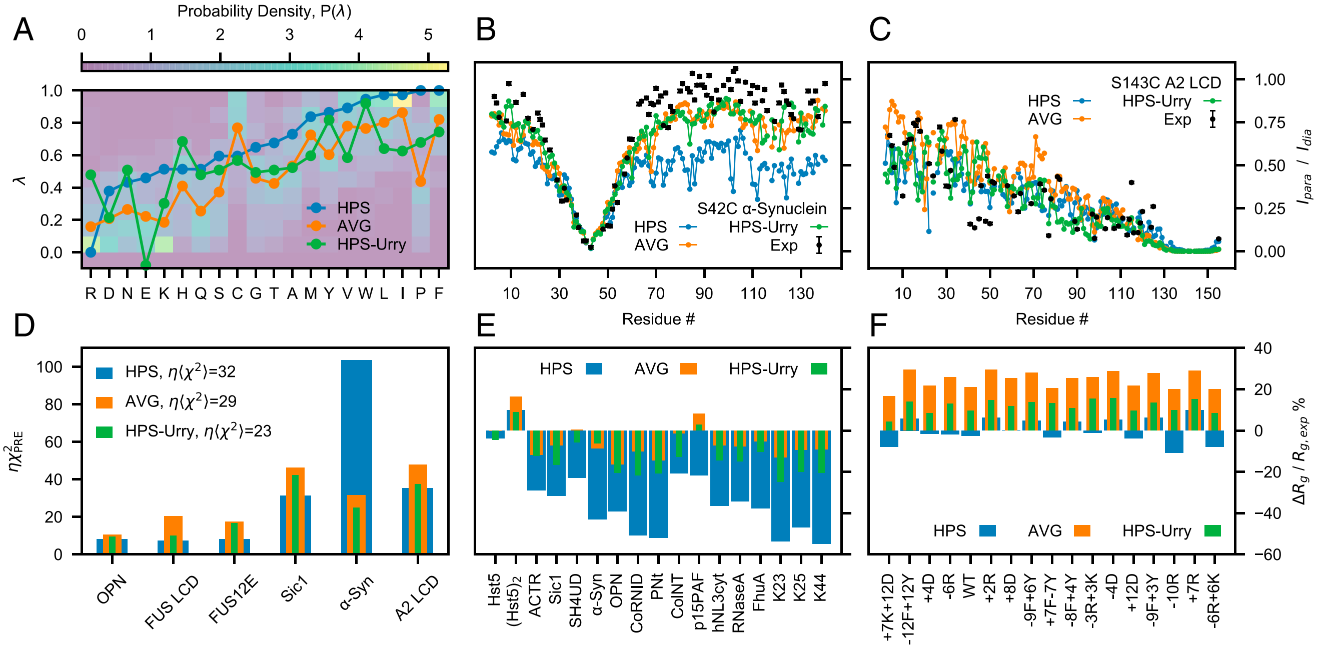

Fig. 1. Assessing the HPS, AVG, and HPS-Urry models using experimental data reporting on single-chain conformational properties. (*A*) Probability distributions of the λ parameters calculated from 87 min–max normalized hydrophobicity scales. Lines are the λ parameters of the HPS model (blue), the average over the hydrophobicity scales (orange) and the HPS-Urry model (green) (28). Intramolecular PRE intensity ratios for (*B*) the S42C mutant of α-Synuclein and (*C*) the S143C mutant of A2 LCD from simulations and experiments (22, 43) (black). (*D*) χ^2 values quantifying the discrepancy between simulated and experimental intramolecular PRE data, scaled by the hyperparameter $\eta = 0.1$ (*Materials and Methods*). Relative difference between simulated and experimental radii of gyration (*E*) for proteins that do not readily undergo phase separation alone and (*F*) for variants of A1 LCD, with negative values corresponding to the simulated ensembles being more compact than in experiments.

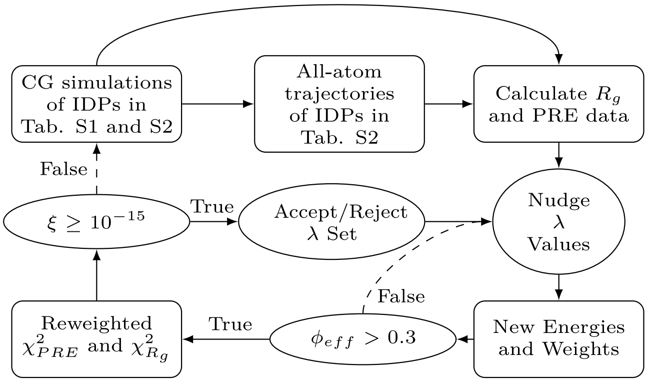
CG simulations of IDPs in Tab. S1 and S2
All-atom trajectories of IDPs in Tab. S2
Calculate R_g and PRE data
False
$\xi \geq 10^{-15}$
True
Accept/Reject λ Set
Nudge λ Values
False
Reweighted χ^2_{PRE} and $\chi^2_{R_g}$
True
$\phi_{eff} > 0.3$
New Energies and Weights

Fig. 2. Flowchart illustrating the Bayesian parameter-learning procedure (*Materials and Methods*).

Specifically, we compared the simulations with the radii of gyration, R_g, of 42 IDPs (*SI Appendix*, Table S1) and intramolecular PRE data of six IDPs (*SI Appendix*, Table S2) (16, 22, 23, 43–57). Compared to the AVG scale, the HPS model overestimates the compaction of α-Synuclein whereas it closely reproduces the PRE data for A2 LCD (Fig. 1 *B* and *C*). In general, the HPS model accurately predicts the conformational properties of sequences with high LLPS propensity, e.g., FUS LCD, A2 LCD, and A1 LCD (Fig. 1 *D* and *F*), while the AVG scale is considerably more accurate at reproducing the R_g of proteins that do not readily undergo phase separation alone (Fig. 1*E*). The recently proposed HPS-Urry model (28) is the most accurate at predicting the intramolecular PRE data while it shows intermediate accuracy for the R_g values of both proteins that do not readily undergo phase separation alone and A1 LCD variants. The HPS-Urry model in particular differs significantly from the HPS and AVG models for the λ parameters for R and E as well as the reversal of the order of hydrophobicity of Y and F (Fig. 1*A*).

Optimization of Amino Acid–Specific Hydrophobicity Values. To obtain a model that more accurately predicts the conformational properties of IDPs of diverse sequences and LLPS propensities, we trained the λ values on a large set of experimental R_g and PRE data using a Bayesian parameter-learning procedure (30) shown schematically in Fig. 2 (*Materials and Methods*). We initially performed an optimization run starting from the AVG λ values and setting the hyperparameters to $\theta = \eta = 0.1$ (*SI Appendix*, Fig. S1*A*). We collected the optimized sets of λ values which yielded $\eta\langle\chi^2_{PRE}\rangle < 21$ and $\langle\chi^2_{R_g}\rangle < 3$ (circles in Fig. 3*A*). The optimization was repeated starting from all $\lambda = 0.5$ to assess that the parameter space sampled by our method is independent of the initial conditions (*SI Appendix*, Figs. S1*D* and S2*A*). Thus, while we used the AVG model as starting point, our final parameters only depend on $P(\lambda)$ via its use as the prior in the Bayesian optimization.

From the pool of optimized parameters, we selected the λ set which resulted in the largest Spearman's correlation coefficient ($\rho = 0.78$) between simulated and experimental R_g values for the A1 LCD variants. We base this final selection of the optimal λ set on the Spearman's correlation coefficient of the A1 LCD variants because we expect that capturing the experimental ranking in chain compaction will result in accurate predictions of the relative LLPS propensities (15, 16, 20, 58, 59). Further, the systematic mutagenesis studies enable us to more clearly decouple the parameters for Y vs. F and R vs. K (15, 16). We note that while this selection uses only the A1 LCD variants, all three parameter sets result in good agreement with the full PRE and R_g dataset (Fig. 3*A*).

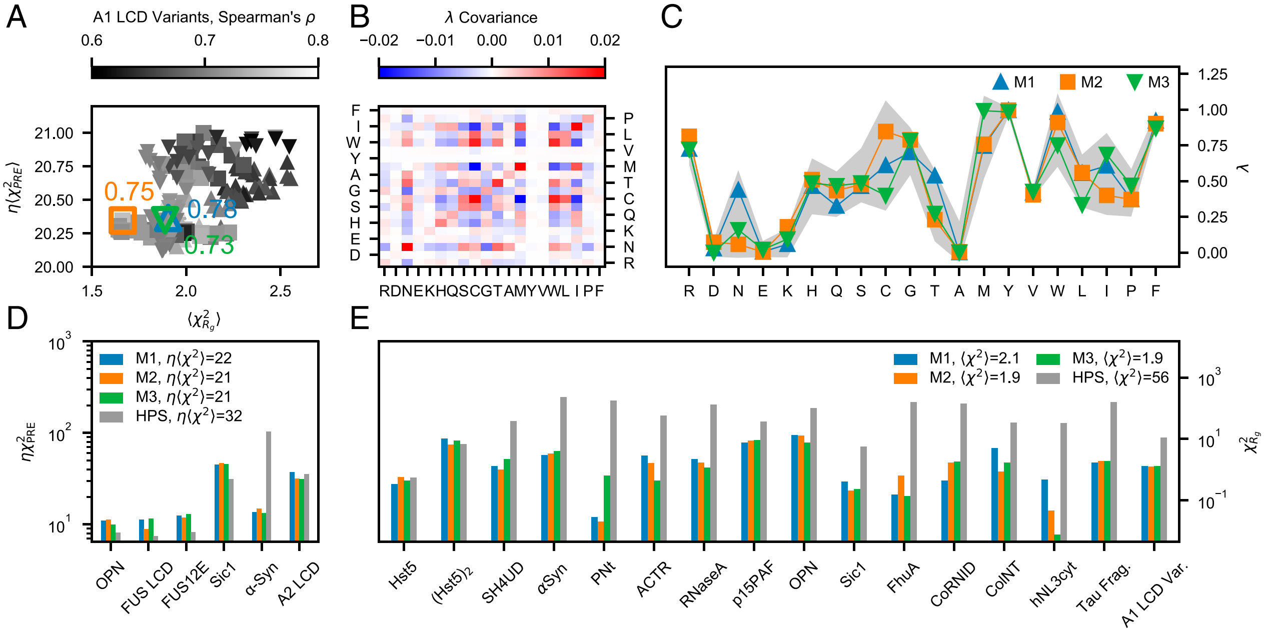

Fig. 3. Selection and performance of the M1–3 models with respect to the training data. (*A*) Overview of the optimal λ sets with $\eta\langle\chi^2_{PRE}\rangle < 21$ and $\langle\chi^2_{R_g}\rangle < 3$ collected through the parameter learning procedures started from $\lambda_0 =$ AVG (upward triangles), M1 (squares), and M2 (downward triangles). The gray gradient shows the Spearman's correlation coefficient between experimental and simulated R_g values for the A1 LCD variants in the training set. Colored open symbols indicate the M1 (blue upward triangle), M2 (orange square), and M3 (green downward triangle) scales, whereas the adjacent values are the respective Spearman's correlation coefficients. (*B*) Covariance matrix of the λ sets with $\eta\langle\chi^2_{PRE}\rangle < 21$ and $\langle\chi^2_{R_g}\rangle < 3$. (*C*) M1 (blue), M2 (orange), and M3 (green) scales. Solid lines are guides for the eye, whereas the gray shaded area shows the mean ±2 SD of the λ sets with $\eta\langle\chi^2_{PRE}\rangle < 21$ and $\langle\chi^2_{R_g}\rangle < 3$. Comparison between (*D*) $\eta\chi^2_{PRE}$ and (*E*) $\chi^2_{R_g}$ values for the HPS model (gray) and the optimized M1 (blue), M2 (orange), and M3 (green) models.

The selected model, referred to as M1 hereafter, is the starting point for two consecutive optimization cycles (*SI Appendix*, Fig. S1*B*) which were performed with a lower weight for the prior ($\theta = 0.05$), yielding a new pool of optimized parameters (squares in Fig. 3*A*) and model M2 (largest $\rho = 0.75$). To generate a third model, we further decreased the confidence parameter to $\theta = 0.02$ and performed an additional optimization run starting from M2 (*SI Appendix*, Fig. S1*C*). From the collected optimal parameters (triangles in Fig. 3*A*), we selected M3 (largest $\rho = 0.73$). As shown in Fig. 3*B*, the optimal λ values collected through the four independent optimization runs (*SI Appendix*, Fig. S1 *A–D*) are weakly intercorrelated. The covariance values range between -0.015 and 0.015 for most amino acids, with the exception of the SDs of N, C, T, M, W, and I. C, M, W, and I are among the least frequent amino acids in the training set (*SI Appendix*, Fig. S3), and unsurprisingly, we observe the largest covariance values for C–W (0.017), C–M (-0.02) and C–I (–0.016). Fig. 3*C* shows that M1–3 fall within two SDs above and below the mean of the λ values yielding $\eta\langle\chi^2_{PRE}\rangle < 21$ and $\langle\chi^2_{R_g}\rangle < 3$ (gray shaded area). Despite their differences, M1–3 fit the training data equally accurately and result in an improvement in $\langle\chi^2_{PRE}\rangle$ and $\langle\chi^2_{R_g}\rangle$ of ~ 30 and $\sim 95\%$, respectively, with respect to the HPS model (Fig. 3 *D* and *E*).

Notably, the optimization procedure captures the sequence dependence of the chain dimensions (Fig. 4) and results in accurate predictions of intramolecular PRE data for both highly soluble IDPs and proteins that more readily phase separate (*SI Appendix*, Figs. S4 *B–D* and S5–S10), as well as in radii of gyration with relative errors $-14\% < \Delta R_g/R_{g,exp} < 12\%$ (*SI Appendix*, Fig. S4 *E* and *F*). Besides reproducing the experimental R_g values for the longer chains with high accuracy, the optimized models also capture the differences in R_g and scaling exponents, ν, for the variants of A1 LCD (Fig. 4*B* and *SI Appendix*, Fig. S11). The lower Pearson's correlation coefficients observed for ν, compared to the corresponding R_g data, may originate from the different models used to infer ν from SAXS experiments and simulation data, i.e., the molecular form factor method (16, 52) and least-squares fit to long intramolecular pairwise distances, R_{ij}, vs. $|i-j| > 10$ (60) (*SI Appendix*, Fig. S12).

To assess the impact of phase separating proteins on the optimized models, we perform an optimization run wherein the A1 LCD variants are removed from the training set. The major difference between the resulting optimal λ set and models M1–3 is the considerably smaller values for R and Y residues (*SI Appendix*, Fig. S2*C*). Indeed, the large λ values for R and Y residues in M1–3 relative to the HPS, AVG, and HPS-Urry models is a striking feature which resonates with previous experimental findings pointing to the important role of R and Y residues in driving LLPS (8, 14–16, 22, 61, 62).

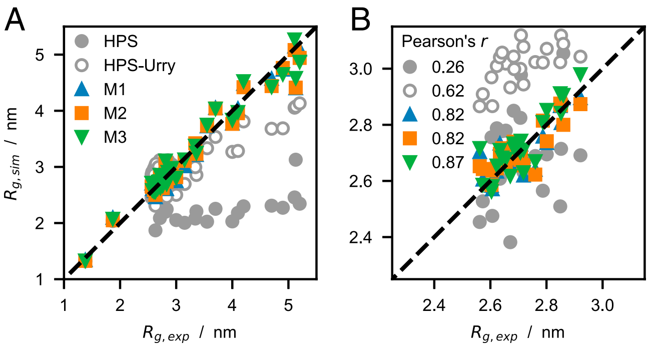

Fig. 4. (*A*) Comparison between experimental and predicted radii of gyration (*SI Appendix*, Table S1), R_g, for the HPS, HPS-Urry, and M1–3 models. (*B*) Zoom-in on the R_g values of the A1 LCD variants, with Pearson's *r* coefficients for this subset of the training data reported in the legend.

To identify the hydrophobicity scales which most closely resemble M1–3, we construct a dendrogram (*SI Appendix*, Fig. S13) complementing the 87 scales retained from the set by Simm et al. (38) with the Urry, Kapcha–Rossky, and M1–3 scales and using average linkage-based hierarchical clustering and Euclidean distances as the metric. This analysis reveals that the hydrophobicity scales by Urry et al. (29), Bishop et al. (41), Wimley and White (63), and the membrane protein surrounding hydrophobicity scale by Ponnuswamy and Gromiha (64) are those with greatest similarity to M1–3. These scales, which are characterized by a λ value for the R residue above the 80% quantile, are possibly the best of the unmodified scales for the properties that we optimized M1–3 to reproduce.

Testing Protein–Protein Interactions. To test whether the parameters trained on single-chain conformational properties are transferable to protein–protein interactions, we compared experimental intermolecular PRE rates, Γ_2, of FUS LCD and A2 LCD (22, 23) with predictions from two-chain simulations of the M1–3 models performed at the same conditions as the reference experiments. Intermolecular Γ_2 values were obtained from solutions of spin-labeled ^{14}N protein and ^{15}N protein without a spin label in equimolar amount and report on the transient interactions between a paramagnetic nitroxide probe attached to a cysteine residue of the spin-labeled chain and all the amide protons of the ^{15}N-labeled chain. We carried out the calculation of the PRE rates using the software DEER-PREdict (34), assuming an effective correlation time of the spin label, τ_t, of 100 ps and fitting an overall molecular correlation time, τ_c, within the interval $1 \leq \tau_c \leq 20$ ns. In agreement with experiments, Γ_2 values predicted by the M1–3 models are characterized by no distinctive peaks along the protein sequence (Fig. 5 *A–E*), which is consistent with transient and nonspecific protein–protein interactions. Notably, while PRE rates for FUS LCD are of the same magnitude for all spin-labeled sites, the A2 LCD presents larger Γ_2 values for S99C than for S143C indicating that the tyrosine-rich aggregation-prone region (residues 84 to 107) is involved in more frequent intermolecular contacts with the entire sequence. The discrepancy between predicted and experimental intermolecular PRE data, χ^2_{PRE}, varies significantly as a function of τ_c (Fig. 5 *F* and *G*). For both FUS LCD and A2 LCD, the optimal τ_c is larger for M1 than for M3, which suggests that the latter has more attractive intermolecular interactions. While for M1 the minimum of χ^2_{PRE} is at $\tau_c = 17$ ns for both proteins, for M3 the optimal τ_c value is ~ 8 ns smaller for FUS LCD than for A2 LCD. Although the accuracy of τ_c is difficult to assess in the case of transiently interacting IDPs, this large difference in τ_c (Fig. 5) suggests that the protein–protein interactions predicted for FUS LCD by M3 may be overly attractive.

To quantify protein–protein interactions with the optimized models, we calculated second virial coefficients, B_{22}, from two-chain simulations (*SI Appendix*). The net interactions are attractive for both the sequences ($B_{22} < 0$) and considerably stronger for A2 LCD than for FUS LCD. As expected from the λ values and amino acid compositions, M3 presents the most negative B_{22} values (large λ values for Q, G, and P), followed by M2 and M1 (Fig. 5*I*).

To test whether predictions of protein self-association by M1–3 are sequence dependent, we compared the probability of finding proteins in the bound dimeric state, p_B, in simulations of α-Synuclein, p15PAF, full-length tau (ht40), A2 LCD, and FUS LCD performed at the solution conditions of the reference experimental data (43, 50, 65) (*SI Appendix*). In agreement with experimental findings, we find that the highly soluble α-Synuclein, p15PAF, and ht40 proteins do not self-associate substantially in our simulations, whereas A2 LCD and FUS LCD have $p_B \sim 4$ and $\sim 1\%$, respectively. We further estimated the dissociation constants of A2 LCD and FUS LCD using

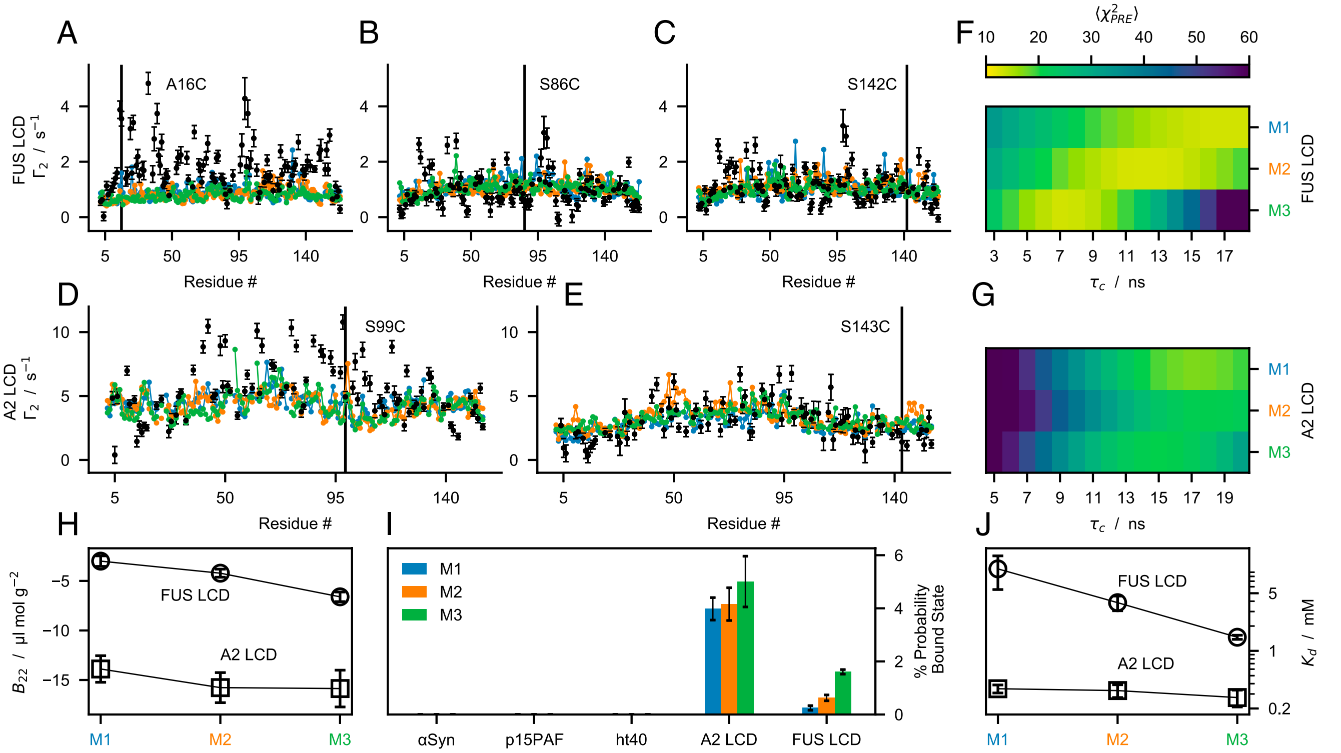

Fig. 5. Testing the M1–3 models using experimental findings on protein–protein interactions. Comparison between experimental (black) intermolecular PRE rates (*SI Appendix*, Table S3) and predictions from the M1 (blue), M2 (orange) and M3 (green) models for (*A*–*C*) FUS LCD and (*D* and *E*) A2 LCD calculated using the best-fit correlation time, τ_c. (*F* and *G*) Discrepancy between calculated and experimental intermolecular PRE rates χ^2_{PRE} as a function of τ_c. (*H*) Second virial coefficients, B_{22}, of FUS LCD (circles) and A2 LCD (squares) calculated from two-chain simulations of the M1–3 models. Error bars are SEMs estimated by bootstrapping 1,000 times 40 B_{22} values calculated from trajectory blocks of 875 ns. (*I*) Probability of the bound state estimated from protein-protein interaction energies in two-chain simulations of the M1–3 models. (*J*) Dissociation constants, K_d, of FUS LCD (circles) and A2 LCD (squares) calculated from two-chain simulations of the M1–3 models. For p_B and K_d, error bars are SDs of 10 simulation replicas. Lines in *H* and *J* are guides for the eye.

$K_d = (1-p_B)^2/(N_A p_B V)$ and $K_d = 1/(N_A p_B(V - B_{22}))$ self-consistently (66), where N_A is Avogadro's number (Fig. 5*J* and *SI Appendix*, Fig. S14).

Testing LLPS Propensities. To test the ability of the models to capture the sequence dependence of LLPS propensity, we performed multichain simulations in a slab geometry and calculated protein concentrations of the coexisting condensate, c_{con}, and dilute phase, c_{sat}. We compared our simulation results to an extensive set of sequences which have been shown to undergo LLPS below an upper critical solution temperature (UCST), namely, FUS LCD (23, 35, 36), A2 LCD (22, 24), the NtoS variant of A2 LCD (24), and LAF-1 RGG domain (11, 67–69), as well as variants of A1 LCD (15, 16) and Ddx4 LCD (8, 10, 13). From simulations of the optimized models at 37 °C, we observed that for a number of sequences in the test set, the predicted c_{sat} values are too low to allow for converged estimates from µs-timescale trajectories (*SI Appendix*, Fig. S15). Conversely, the least LLPS-prone variants of Ddx4 LCD yielded one-phase systems when simulated at 37 °C using HPS-Urry and M1–3 models. Thus, to be able to estimate converged c_{sat} values (*SI Appendix*, Figs. S16, S17, and S18), simulations were carried out at 50 °C, except for the HPS-Urry model which we simulated at 24 °C (*SI Appendix*, Table S4). The FtoA and RtoA variants of Ddx4 LCD were also simulated at 24 °C using the M1–3 models as in simulations of the same systems at 50 °C we only observed a single phase.

Simulations using M1 at 50 °C most closely recapitulate the experimental trend in c_{sat} across the diverse sequences (Fig. 6 *A*, *D*, and *G*) and reproduce the reference c_{con} and c_{sat} values measured at room temperature. Conversely, HPS overestimates the relative LLPS propensity of FUS LCD, whereas simulations using HPS-Urry at 24 °C show deviations of about an order of magnitude from the reference c_{sat} values for A2 LCD, Ddx4 LCD, A1 LCD, and FUS LCD. Regarding the LAF-1 RGG domain, all of the models overestimate by at least a factor of ∼ 5 the experimental c_{con} (68, 69), whereas M1 reproduces within a factor of ∼ 2 the experimental c_{sat} value from temperature-dependent turbidity measurements (11), both for the wild type (WT) and for variants with randomly shuffled sequence (LAF-1 shuf) and without residues 21 to 30 (LAF-1 Δ 21 to 30) (*SI Appendix*, Fig. S19). Although M1–3 fit the training data equally well, the prediction of LLPS propensities for the diverse sequences in Fig. 6 *A* and *D* differ considerably, with Pearson's correlation coefficients between simulation and experimental $\log_{10}(c_{sat})$ values ranging from 0.67 for M1 to 0.14 for M3 (Fig. 6*G*). The discrepancy is particularly evident for the Ddx4 LCD and FUS LCD which are rich in N and Q residues, respectively, i.e., the residues for which the M1 and M3 λ sets differ the most.

We further test our predictions against 15 variants of A1 LCD (Fig. 6 *B* and *E*). These include aromatic and charge variants, which were designed to decipher the role on the driving forces for phase separation of Y vs. F residues and of R, D, E, and K residues, respectively (16). The nomenclature, $\pm N_X \mathrm{X} \pm N_Z \mathrm{Z}$, denotes increase or decrease in the number of residues of type X and Z with respect to the WT, which is achieved by mutations to or from G and S residues while maintaining a constant G/S ratio. M1–3 are found to be equally accurate and present a considerable improvement over previous models with respect to their ability to

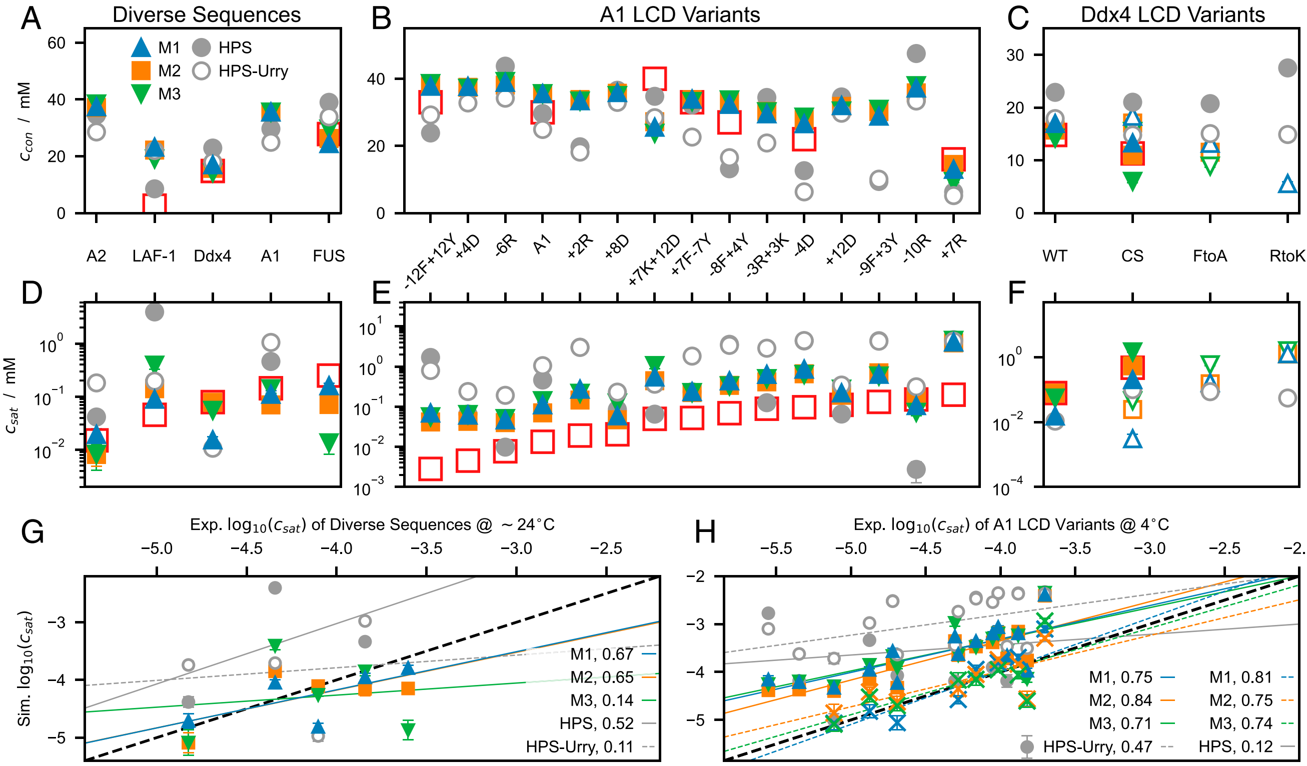

Fig. 6. Protein concentrations (*A*–*C*) in the condensate and (*D*–*F*) in the dilute phase from slab simulations of the M1–3, HPS, and HPS-Urry models performed at 50 °C (closed symbols), 37 °C (crosses in *H*), and 24 °C (open symbols). Red open squares show experimental measurements at ∼ 24 °C (*A*, *C*, *D*, and *F*) and ∼ 4 °C (*B* and *E*). Correlation between $\log_{10}(c_{sat}/\mathrm{M})$ from simulations and experiments for (*G*) diverse sequences and (*H*) A1 LCD variants. Solid lines show linear fits to the simulation data at 50 °C. Dashed lines show linear fits to the HPS-Urry data at 24 °C (*G* and *H*) and to the M1–3 data at 37 °C (*H*). Values reported in the legends are Pearson's correlation coefficients. Error bars are SEMs of averages over blocks of 0.3 μs. We note that the correlation coefficients reported in *G* are associated with a substantial uncertainty as they are calculated over only three (HPS), four (HPS-Urry), and five points (M1–3).

recapitulate the trends in LLPS propensity for the aromatic and charged variants of A1 LCD. Since M1–3 were selected based on their performance in predicting the experimental ranking for the R_g values of 21 A1 LCD variants (*SI Appendix*, Table S1), this result supports our model development strategy. For M1–3, Pearson's correlation coefficients exceed 0.7 between $\log_{10}(c_{sat})$ values measured at 4 °C (16) and simulation predictions at both 50 and 37 °C (Fig. 6*H*). Moreover, c_{sat} values from simulations at 37 °C are in agreement with the reference c_{sat} values at 4 °C (Fig. 6*H* and *SI Appendix*, Fig. S15). As we observed for the diverse sequences, quantitative agreement with the experimental c_{sat} values is achieved by carrying out simulations of the M1 model at a temperature systematically larger by ∼ 30 °C than the experimental conditions. In addition to the lack of temperature dependence of the hydropathy parameters (70), the inconsistency between the temperature dependence of chain compaction and phase separation might be attributed to the long range of the nonelectrostatic interactions, which we compute up to distances of 4 nm (*SI Appendix*). Moreover, the significant decrease in the number of interaction sites upon coarse-graining at the amino acid level, and the resulting reduction in configurational entropy (71, 72), may promote LLPS by lowering the entropic penalty associated with partitioning a chain from the dilute solution to the condensate.

M1–3 reproduce the experimental ranking for LLPS propensity of the Ddx4 LCD variants, i.e., WT$\gg$ CS> FtoA$\gtrsim$ RtoK (Fig. 6 *C* and *F*), and for all the variants, M1 and M3 consistently display the highest and lowest LLPS propensities, respectively. Simulations at 50 °C using M2 are in quantitative agreement with the experimental c_{sat} values (13) for both WT and the CS variant, which has the same net charge and amino acid composition as the WT but a more uniform charge distribution along the sequence. Moreover, as observed experimentally (13), M1–3 predict a single phase for the RtoK variant at 24 °C. As previously shown by Das et al. (25), the HPS model predicts a considerable increase in LLPS propensity upon replacement of all 24 R residues in the Ddx4 LCD with K (RtoK variant; Fig. 6*C*), in apparent contrast to experimental observations (10, 13). Interestingly, augmenting the HPS model with stronger cation–π interactions for R-aromatic than for K-aromatic pairs (25) has been shown to be insufficient to capture the lower LLPS propensity of the RtoK variant compared to WT. On the other hand, our data for the M1–3 and HPS-Urry models indicate that making all the interactions involving R more favorable results in more accurate predictions. In fact, a large λ value for R may better mimic its relatively unfavorable free energy of hydration (19) as well as the occurrence of R-aromatic cation–π interactions, R-R π-stacking, and R-D/E bidentate H-bonding (10, 17, 18, 73). Compared to the Kapcha–Rossky scale, it is noteworthy that the increase in the λ values of R, Y, and G in M1–3 is accompanied by an overall decrease in the average λ value. Hence, the optimization procedure led to the enhancement of specific attractive forces while maintaining a balance between electrostatic and nonelectrostatic interactions (25), which reveals itself, for example, in the ability of M1–3 to recapitulate the lower LLPS propensity of the CS variant with respect to Ddx4 LCD WT.

The M1 and M2 parameter sets differ mainly for the λ value of the N residue (Fig. 3*C*) and perform equally well against the test set (Fig. 6). Therefore, we further test the ability of M1 and M2 to predict the LLPS propensity of the NtoS variant of A2 LCD with respect to the WT. Only the M1 model, which has

λ values for N and S of similar magnitude correctly predicts approximately the same LLPS propensity for variant and WT (*SI Appendix*, Fig. S20), in agreement with experiments (24).

Correlating Single-Chain Properties and Phase Separation. Motivated by recent experiments on the A1 LCD (15, 16), we perform a detailed analysis of the coupling between chain compaction and phase behavior of the A1 LCD variants. In agreement with previous observations (16), the $\log_{10}(c_{sat})$ values for the aromatic variants show a linear relationship with the scaling exponent, ν_{sim}, whereas changes in the number of charged residues (charge variants) result in significant deviations from the lines of best fit (Fig. 7 *A–C*). Following the approach of Bremer et al. (16), we plot the residuals for the charge variants with respect to the lines of best fit as a function of the net charge per residue (NCPR) (Fig. 7 *D–F*). The results for M1 and M2 show the V-shaped profile observed for the experimental data (16) and support the suggestion that mean-field electrostatic repulsion between the net charge of the proteins is responsible for breaking the coupling between chain compaction and LLPS propensity (16). In agreement with experimental data (16), we observe that for M1 and M2 the driving forces for LLPS are maximal for small positive values of NCPR (~ 0.02).

The dependence of LLPS on NCPR is clarified by comparing the residual nonelectrostatic energy maps of +8D (NCPR = 0), +4D (NCPR $\approx$ 0.03), and −4D (NCPR $\approx$ 0.09) with respect to the WT of A1 LCD (NCPR $\approx$ 0.06) (*SI Appendix*, Figs. S21 and S22). While in the case of NCPR = 0 the residual interaction patterns within the isolated chain and between chains in the condensate largely overlap, the energy baselines are clearly down- and up-shifted for NCPR $\approx$ 0.03 and NCPR $\approx$ 0.09, respectively (*SI Appendix*, Figs. S21 *G–I* and S22 *G–I*). Although the interaction patterns are still dominated by the stickers, deviations of the NCPR from ~ 0.02 result in electrostatic mean-field repulsive interactions that disfavor LLPS. The LLPS-promoting effect of small positive NCPR values finds explanation in the amphiphilic character of the R side chains (18) which compensates for the repulsion introduced by the excess positive charge by allowing for favorable interactions with both Y and negatively charged residues. As opposed to M1 and M2, the readily phase-separating M3 model shows a weaker dependence on NCPR, especially for variants of net negative charge. This suggests that the experimental observations regarding the coupling between conformational and phase behavior of A1 LCD stem from a well-defined balance between mean-field repulsion and sticker-driven LLPS which can be offset by an overall moderate increase of 3 to 4% in the λ values of the residues present in A1 LCD.

Comparing Intramolecular and Intermolecular Interactions. After establishing the ability of model M1 to accurately predict trends in LLPS propensity for diverse sequences, we analyze the nonelectrostatic residue–residue energies for FUS LCD and A2 LCD within a single chain, as well as between pairs of chains in the dilute regime and in condensates. We find a striking similarity between intramolecular and intermolecular interaction patterns for both proteins (Fig. 8), consistent with a mostly uniform distribution of stickers along the linear sequence (Fig. 8 *G* and *H*) (15, 74). Notably, besides the aromatic F and Y residues, the analysis also identifies an M residue and four R residues as stickers in FUS LCD and A2 LCD, respectively. Therefore, the parameter-learning procedure presented herein corroborates the important role of R as a sequence-dependent sticker (16), whereby the large λ value for R in models M1–3 presumably reflects the ability of the amphiphilic guanidinium moiety to engage in H-bonding, as well as π stacking and charge–π interactions (18). Further, in the dilute regime, the intramolecular and intermolecular interactions are weaker in the N- and C-terminal regions than for the rest of the chain, as evident from the upturning baselines of the one-dimensional (1D) interaction energy projections. This result is consistent with the faster local motions of the terminal residues inferred from ^{15}N NMR relaxation data for both unfolded proteins (75) and a number of phase separating IDPs (15, 22, 23). We also find that the aggregation-prone Y-rich region of A2 LCD (residues 84 to 107) interacts with the entire polypeptide chain (Fig. 8 *D–F*) and thus likely drives chain compaction and self-association as well as LLPS. Finally, in line with

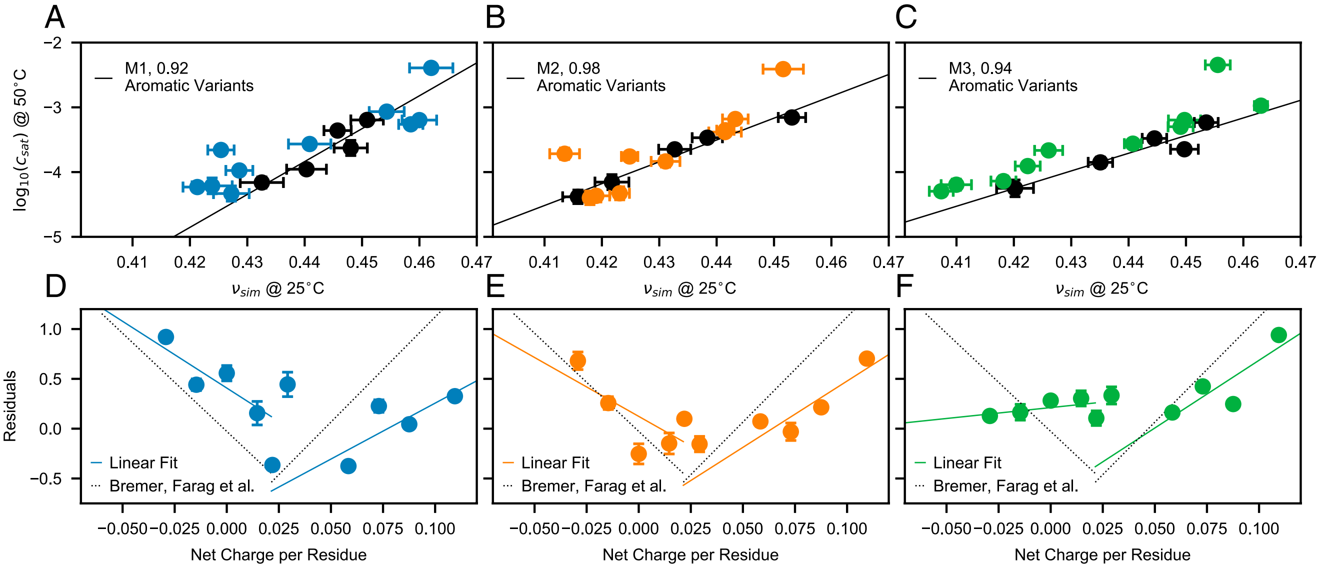

Fig. 7. Correlation between chain compaction and LLPS propensity for aromatic and charge variants of A1 LCD. $\log_{10}(c_{sat}/\mathrm{M})$ vs. ν_{sim} for A1 LCD variants from simulations performed using the (*A*) M1, (*B*) M2, and (*C*) M3 models. Black and colored circles indicate aromatic and charge variants, respectively. Black lines are linear fits to the aromatic variants. (*D–F*) Residuals from the linear fits of *A–C* for the charge variants of A1 LCD as a function of the NCPR. Values reported in the legends are Pearson's correlation coefficients. Error bars of $\log_{10}(c_{sat})$ values are SEMs of averages over blocks of 0.3 μs. Error bars of ν_{sim} are SDs from fits to $R_{ij} = R_0|i-j|^{\nu_{sim}}$ in the long-distance region, $|i-j| > 10$. Solid lines are linear fits to the data. Dotted lines in *D–F* are lines of best fit to the experimental data by Bremer et al. (16).

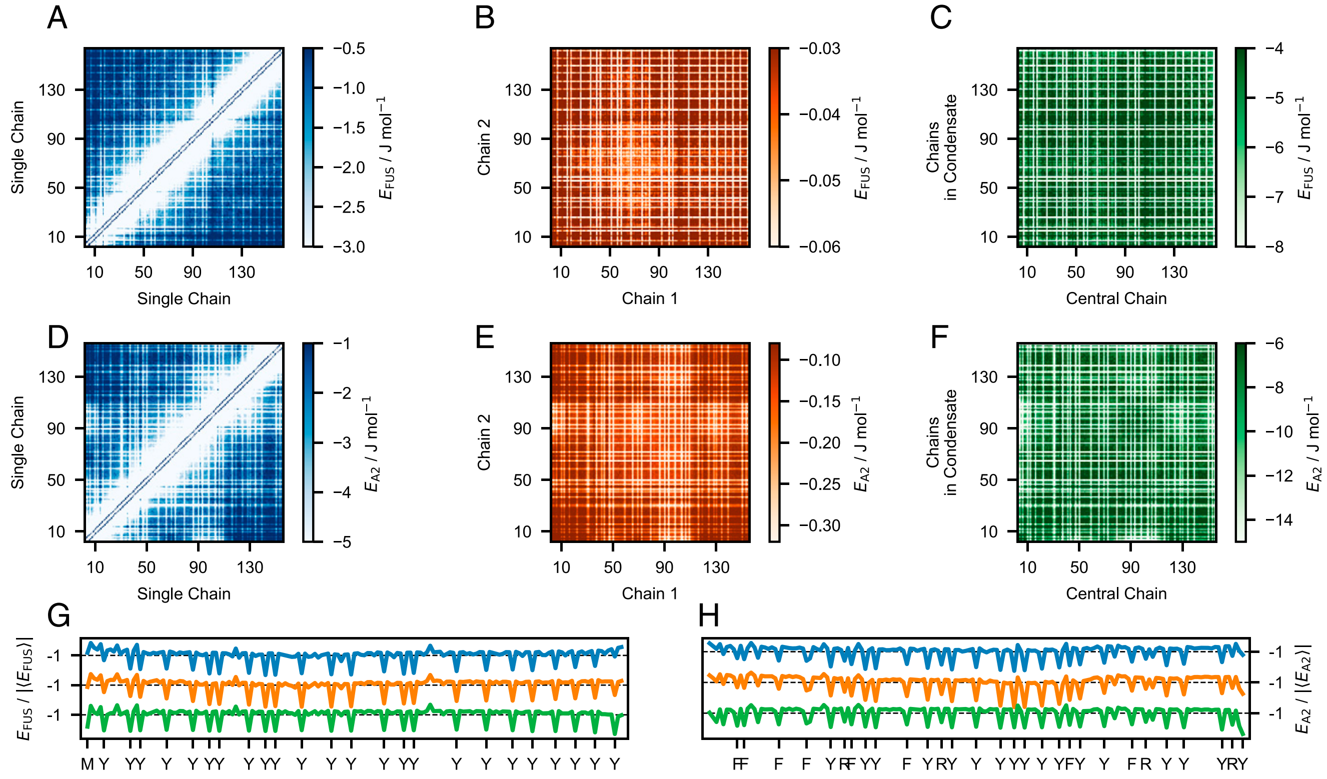

Fig. 8. Comparing residue–residue interactions in dilute solution and in the condensate. Energy maps from simulations of the M1 model of (*A*–*C*) FUS LCD and (*D*–*F*) A2 LCD calculated using nonelectrostatic interaction energies. The 1D projections of the energy maps for (*G*) FUS LCD and (*H*) A2 LCD, normalized by the absolute average interaction energy $|\langle E\rangle|$ and shifted vertically for clarity. Colors indicate that the energies were calculated within a single chain at infinite dilution (blue), between two chains in the dilute regime (orange), and between a chain located at the center of a condensate and the surrounding chains (green).

previous observations from theory, simulations, and experiments (16, 76, 77), we observe that the polypeptide chains of A1 LCD, A2 LCD, and FUS LCD are more expanded in the condensed phase than in the dilute phase (*SI Appendix*, Fig. S23). In particular, we find that the scaling exponents of the LCDs increase toward $\nu = 0.5$ in the condensed phase and that differences in compaction between WT and charge variants of A1 LCD are greater in the dilute than in the condensed phase (*SI Appendix*, Fig. S23).

Conclusions

In this work we implement and validate an automated procedure to develop an accurate model of the LLPS of IDPs based on experimental data reporting on single-chain conformational properties. We show that this strategy succeeds, in agreement with the previously observed coupling between chain compaction and propensity for phase separation (15, 20, 58, 59), but also appears to recapitulate the recent discovery that charge effects may break this relationship (16). Our work differs from related previous studies (28, 30, 33, 78) in several ways including the size of the dataset used for optimization, the use of both NMR PREs and R_g values, and the introduction of a prior for the λ values. Moreover, by carrying out model optimizations with and without the A1 LCD variants, we show that the presence of phase-separating IDPs in the training set helps the parameter-learning procedure to capture the role of Y and R residues as stickers. The accuracy and general applicability of our model can be tested further by future experiments on systems that were not used for training or testing. We also note that our automated, Bayesian optimization approach makes it relatively straightforward to continue to develop and improve the model as additional data become available.

Simulations performed using the model optimized herein reveal that at least for sequences characterized by a relatively uniform distribution of stickers, residue–residue interactions determining chain compaction also drive self-association and LLPS. Moreover, we show that the experimentally observed dependency of LLPS on protein net charge appears to be captured by salt-screened electrostatic repulsion, even when assuming a uniform dielectric constant throughout the two-phase system.

We have here shown how our model may be used to help elucidate the residues that are important for LLPS of IDPs with UCST behavior. Further, we suggest the model could be applied to study the influence of disease-associated mutations on the material properties of protein self-coacervates (79, 80), the LLPS of protein mixtures as a function of composition, and the partitioning of proteins that do not readily undergo phase separation alone into condensates formed by other proteins (81, 82). Finally, owing to the generalized parameter-learning approach, the model could readily be refined as new experimental data are collected, and it should be possible to extend it to account for specific pairwise interactions such as cation–π interactions (25), PTMs (83), the salting-out effect (84), and the temperature dependence of solvent-mediated interactions (70).

Materials and Methods

We use the Cα-based model proposed by Dignon et al. (21) augmented with extra charges for the termini and a temperature-dependent treatment for dielectric constant of water (*SI Appendix*). Langevin dynamics simulations are conducted using HOOMD-blue v2.9 (85) in the *NVT* ensemble using the Langevin thermostat with a time step of 5 fs and friction coefficient of 0.01 ps^{-1} (*SI Appendix*). Additionally, 100- and 300-chain simulations of LAF-1 RGG domain are also performed using openMM v7.5 (86) (*SI Appendix*, Fig. S20).

Bayesian Parameter-Learning Procedure. The λ values are optimized using a Bayesian parameter-learning procedure (30, 87, 88). The training set consists of the experimental R_g values of 42 IDPs (*SI Appendix*, Table S1) and the intramolecular PRE data of six proteins (*SI Appendix*, Table S2) (16, 22,23, 43–57). To guide the optimization within physically reasonable parameters and to avoid overfitting the training set, we introduce a regularization term which penalizes deviations of the λ values from the probability distribution, $P(\lambda)$, which is the prior knowledge obtained from the statistical analysis of 87 hydrophobicity scales. The optimization procedure consists of the following steps (Fig. 2):

1. Single-chain CG simulation of the proteins of the training set (*SI Appendix*, Table S1).
2. Conversion from CG to all-atom trajectories using the powerful chain restoration algorithm (PULCHRA v3.06) (89) for the proteins in *SI Appendix*, Table S2 for which we calculate the PRE data.
3. Calculation of per-frame radii of gyration and PRE data. The PRE rates, Γ_2, and intensity ratios, I_{para}/I_{dia}, are calculated using the rotamer library approach implemented in DEER-PREdict (34) with $\tau_t = 100$ ps and optimizing the correlation time, $\tau_c \in [1, 10]$ ns, against the experimental data.
4. Random selection of six λ values which are nudged by random numbers picked from a normal distribution of zero mean and SD 0.05. The prior probability distribution, $P(\lambda)$, sets the bounds of the parameter space: any λ_i for which $P(\lambda_i) = 0$ is further nudged until $P(\lambda_i) \neq 0$.
5. Calculation of the Boltzmann weights for the ith frame as $w_i = \exp\{-[U(\boldsymbol{r}_i, \boldsymbol{\lambda}_k) - U(\boldsymbol{r}_i, \boldsymbol{\lambda}_0)]/k_B T\}$, where $U(\boldsymbol{r}_i, \boldsymbol{\lambda}_k)$ and $U(\boldsymbol{r}_i, \boldsymbol{\lambda}_0)$ are the total Ashbaugh–Hatch energies of the ith frame for trial and initial λ values, respectively. If the effective fraction of frames,

$$\phi_{eff} = \exp\left[-\sum_{i}^{N_{frames}} w_i \log\left(w_i \times N_{frames}\right)\right], \qquad [1]$$

   is below 30%, the trial $\boldsymbol{\lambda}_k$ is discarded.
6. The per-frame radii of gyration and PRE observables are reweighted, and the extent of agreement with the experimental data is estimated as

$$\chi^2_{R_g} = \left(\frac{R_g^{exp} - R_g^{calc}}{\sigma^{exp}}\right)^2 \qquad [2]$$

   and

$$\chi^2_{PRE} = \frac{1}{N_{labels} N_{res}} \sum_{j}^{N_{labels}} \sum_{i}^{N_{res}} \left(\frac{Y_{ij}^{exp} - Y_{ij}^{calc}}{\sigma_{ij}^{exp}}\right)^2, \qquad [3]$$

   where σ_{ij}^{exp} is the error on the experimental values, Y is either I_{para}/I_{dia} or Γ_2, N_{labels} is the number of spin-labeled mutants, and N_{res} is the number of measured residues.
7. Following the Metropolis criterion (90), the kth set of λ values is accepted with probability

$$A_{k-1\to k} = \begin{cases} \exp\left[\frac{\mathcal{L}(\boldsymbol{\lambda}_{k-1}) - \mathcal{L}(\boldsymbol{\lambda}_k)}{\xi_k}\right], & \mathcal{L}(\boldsymbol{\lambda}_k) > \mathcal{L}(\boldsymbol{\lambda}_{k-1}) \\ 1, & \mathcal{L}(\boldsymbol{\lambda}_k) \leq \mathcal{L}(\boldsymbol{\lambda}_{k-1}), \end{cases} \qquad [4]$$

   where the control parameter, ξ_k, scales with the number of iterations as $\xi = \xi_0 \times 0.99^k$. $\mathcal{L}$ is the cost function

$$\mathcal{L}(\boldsymbol{\lambda}) = \langle \chi^2_{R_g}(\boldsymbol{\lambda}) \rangle + \eta \langle \chi^2_{PRE}(\boldsymbol{\lambda}) \rangle - \theta \sum_i \ln\left[P(\lambda_i)\right], \qquad [5]$$

   where $\langle \chi^2_{R_g}(\boldsymbol{\lambda}) \rangle$ and $\langle \chi^2_{PRE}(\boldsymbol{\lambda}) \rangle$ are averages over the proteins in the training sets. θ and η are hyperparameters of the optimization procedure. θ determines the trade-off between overfitting and underfitting the training set, whereas η sets the relative weight of the PRE data with respect to the radii of gyration.

Steps 4 to 7 are iterated until $\xi < 10^{-15}$, when the reweighting cycle is interrupted and a new CG simulation is carried out with the trained λ values. A complete parameter-learning procedure consists of two reweighting cycles starting from $\xi_0 = 2$ followed by three cycles starting from $\xi_0 = 0.1$. The threshold on ϕ_{eff} results in average absolute differences between χ^2 values estimated from reweighting and calculated from trajectories performed with the corresponding parameters of $\sim$ 1.8 and $\sim$ 0.8 for $\eta\chi^2_{PRE}$ and $\chi^2_{R_g}$, respectively (*SI Appendix*, Fig. S24).

Data Availability. Datasets, amino acid sequences, code, and Jupyter Notebooks for reproducing our simulations and analyses have been deposited in publicly accessible repositories on GitHub (https://github.com/KULL-Centre/papers/tree/main/2021/CG-IDPs-Tesei-et-al) (91) and on Zenodo (DOI: 10.5281/zenodo.5005953) (92).

ACKNOWLEDGMENTS. We thank Veronica Ryan and Nicolas L. Fawzi for sharing the PRE data for FUS LCD, FUS12E LCD, and A2 LCD as well as Robert Konrat for sharing the intramolecular PRE data for Osteopontin. We thank Robert B. Best for sharing data on compaction of IDPs; Gregory L. Dignon and Jeetain Mittal for help setting up simulations with the HPS model; and Tanja Mittag, Massimiliano Bonomi, and Benjamin Schuler for helpful discussions. We acknowledge funding from the BRAINSTRUC structural biology initiative from the Lundbeck Foundation (Grant R155-2015-2666) and acknowledge access to computational resources from the Resource for Biomolecular Simulations (ROBUST) (supported by Novo Nordisk Foundation Grant NNF18OC0032608) and Biocomputing Core Facility at the Department of Biology, University of Copenhagen. T.K.S. acknowledges support from the Novo Scholarship Programme 2021. This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement 101025063.

1. A. Patel *et al.*, A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* **162**, 1066–1077 (2015).
2. S. Wegmann *et al.*, Tau protein liquid-liquid phase separation can initiate tau aggregation. *EMBO J.* **37**, e98049 (2018).
3. N. M. Kanaan, C. Hamel, T. Grabinski, B. Combs, Liquid-liquid phase separation induces pathogenic tau conformations in vitro. *Nat. Commun.* **11**, 2809 (2020).
4. S. Ray *et al.*, a-Synuclein aggregation nucleates through liquid-liquid phase separation. *Nat. Chem.* **12**, 705–716 (2020).
5. MC Hardenberg *et al.*, Observation of an α-synuclein liquid droplet state and its maturation into Lewy body-like assemblies. *J. Mol. Cell Biol.* **13**, 282–294 (2021).
6. Y. Shin, C. P. Brangwynne, Liquid phase condensation in cell physiology and disease. *Science* **357**, eaaf4382 (2017).
7. N. B. Nedelsky, J. P. Taylor, Bridging biophysics and neurology: Aberrant phase transitions in neurodegenerative disease. *Nat. Rev. Neurol.* **15**, 272–286 (2019).
8. T. J. Nott *et al.*, Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell* **57**, 936–947 (2015).
9. C. P. Brangwynne, P. Tompa, R. V. Pappu, Polymer physics of intracellular phase transitions. *Nat. Phys.* **11**, 899–904 (2015).
10. R. M. Vernon *et al.*, Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *eLife* **7**, e31486 (2018).
11. B. S. Schuster *et al.*, Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11421–11431 (2020).
12. G. L. Dignon, R. B. Best, J. Mittal, Biomolecular phase separation: From molecular driving forces to macroscopic properties. *Annu. Rev. Phys. Chem.* **71**, 53–75 (2020).
13. J. P. Brady *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.* **114**, E8194–E8203 (2017).
14. J. Wang *et al.*, A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell* **174**, 688–699.e16 (2018).
15. E. W. Martin *et al.*, Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367**, 694–699 (2020).
16. A. Bremer *et al.*, Deciphering how naturally occurring sequence features impact the phase behaviors of disordered prion-like domains. *bioRxiv* [Preprint] (2021). https://doi.org/10.1101/2021.01.01.425046 (Accessed 4 January 2021).
17. G. Krainer *et al.*, Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions. *Nat. Commun.* **12**, 1085 (2021).
18. M. Vazdar *et al.*, Arginine "magic": Guanidinium like-charge ion pairing from aqueous salts to cell penetrating peptides. *Acc. Chem. Res.* **51**, 1455–1464 (2018).
19. M. J. Fossat, X. Zeng, R. V. Pappu, Uncovering differences in hydration free energies and structures for model compound mimics of charged side chains of amino acids. *J. Phys. Chem. B* **125**, 4148–4161 (2021).
20. G. L. Dignon, W. Zheng, R. B. Best, Y. C. Kim, J. Mittal, Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9929–9934 (2018).
21. G. L. Dignon, W. Zheng, Y. C. Kim, R. B. Best, J. Mittal, Sequence determinants of protein phase behavior from a coarse-grained model. *PLOS Comput. Biol.* **14**, e1005941 (2018).
22. V. H. Ryan *et al.*, Mechanistic view of hnRNPA2 low-complexity domain structure, interactions, and phase separation altered by mutation and arginine methylation. *Mol. Cell* **69**, 465–479.e7 (2018).
23. Z. Monahan *et al.*, Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. *EMBO J.* **36**, 2951–2967 (2017).
24. V. H. Ryan *et al.*, Tyrosine phosphorylation regulates hnRNPA2 granule protein partitioning and reduces neurodegeneration. *EMBO J.* **40**, e105001 (2021).
25. S. Das, Y. H. Lin, R. M. Vernon, J. D. Forman-Kay, H. S. Chan, Comparative roles of charge, *p*, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 28795–28805 (2020).

26. J. M. Choi, A. S. Holehouse, R. V. Pappu, Physical principles underlying the complex biology of intracellular phase transitions. *Annu. Rev. Biophys.* **49**, 107–133 (2020).
27. L. H. Kapcha, P. J. Rossky, A simple atomic-level hydrophobicity scale reveals protein interfacial structure. *J. Mol. Biol.* **426**, 484–498 (2014).
28. R. M. Regy, J. Thompson, Y. C. Kim, J. Mittal, Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. *Protein Sci.* **30**, 1371–1379 (2021).
29. D. W. Urry *et al.*, Hydrophobicity scale for proteins based on inverse temperature transitions. *Biopolymers* **32**, 1243–1250 (1992).
30. A. B. Norgaard, J. Ferkinghoff-Borg, K. Lindorff-Larsen, Experimental parameterization of an energy function for the simulation of unfolded proteins. *Biophys. J.* **94**, 182–192 (2008).
31. L. P. Wang, T. J. Martinez, V. S. Pande, Building force fields: An automatic, systematic, and reproducible approach. *J. Phys. Chem. Lett.* **5**, 1885–1891 (2014).
32. G. Tiana, L. Giorgetti, "Coarse graining of a giant molecular system: The chromatin fiber" in *Biomolecular Simulations: Methods in Molecular Biology*, M. Bonomi, C. Camilloni, Eds. (Springer, New York, 2019), vol. 2022, pp. 399–411.
33. T. Dannenhoffer-Lafage, R. B. Best, A data-driven hydrophobicity scale for predicting liquid–liquid phase separation of proteins. *J. Phys. Chem. B* **125**, 4046–4056 (2021).
34. G. Tesei *et al.*, DEER-PREdict: Software for efficient calculation of spin-labeling EPR and NMR data from conformational ensembles. *PLOS Comput. Biol.* **17**, e1008551 (2021).
35. K. A. Burke, A. M. Janke, C. L. Rhine, N. L. Fawzi, Residue-by-residue view of in vitro FUS granules that bind the C-terminal domain of RNA polymerase II. *Mol. Cell* **60**, 231–241 (2015).
36. A. C. Murthy *et al.*, Molecular interactions underlying liquid-liquid phase separation of the FUS low-complexity domain. *Nat. Struct. Mol. Biol.* **26**, 637–648 (2019).
37. H. S. Chan, *Amino Acid Side-Chain Hydrophobicity* (American Cancer Society, 2002).
38. S. Simm, J. Einloft, O. Mirus, E. Schleiff, 50 years of amino acid hydrophobicity scales: Revisiting the capacity for peptide classification. *Biol. Res.* **49**, 31 (2016).
39. H. B. Bull, K. Breese, Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues. *Arch. Biochem. Biophys.* **161**, 665–670 (1974).
40. H. R. Guy, Amino acid side-chain partition energies and distribution of residues in soluble proteins. *Biophys. J.* **47**, 61–70 (1985).
41. C. M. Bishop, W. F. Walkenhorst, W. C. Wimley, Folding of β-sheets in membranes: Specificity and promiscuity in peptide model systems. *J. Mol. Biol.* **309**, 975–988 (2001).
42. G. W. Welling, W. J. Weijer, R. van der Zee, S. Welling-Wester, Prediction of sequential antigenic regions in proteins. *FEBS Lett.* **188**, 215–218 (1985).
43. M. M. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo, C. M. Dobson, Mapping long-range interactions in a-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J. Am. Chem. Soc.* **127**, 476–477 (2005).
44. S. Jephthah, L. Staby, B. B. Kragelund, M. Skepö, Temperature dependence of intrinsically disordered proteins in simulations: What are we missing? *J. Chem. Theory Comput.* **15**, 2672–2683 (2019).
45. E. Fagerberg, L. K. Månsson, S. Lenton, M. Skepö, The effects of chain length on the structural properties of intrinsically disordered proteins in concentrated solutions. *J. Phys. Chem. B* **124**, 11843–11853 (2020).
46. M. Kjaergaard *et al.*, Temperature-dependent structural changes in intrinsically disordered proteins: Formation of a-helices or loss of polyproline II? *Protein Sci.* **19**, 1555–1564 (2010).
47. G. W. Gomes *et al.*, Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET. *J. Am. Chem. Soc.* **142**, 15697–15710 (2020).
48. U. R. Shrestha *et al.*, Generation of the configurational ensemble of an intrinsically disordered protein from unbiased molecular dynamics simulation. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 20446–20452 (2019).
49. C. L. Johnson *et al.*, The two-state prehensile tail of the antibacterial toxin colicin n. *Biophys. J.* **113**, 1673–1684 (2017).
50. A. De Biasio *et al.*, p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins. *Biophys. J.* **106**, 865–874 (2014).
51. A. Paz *et al.*, Biophysical characterization of the unstructured cytoplasmic domain of the human neuronal adhesion protein neuroligin 3. *Biophys. J.* **95**, 1928–1944 (2008).
52. J. A. Riback *et al.*, Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **358**, 238–241 (2017).
53. M. C. Ahmed *et al.*, Refinement of α-synuclein ensembles against SAXS data: Comparison of force fields and methods. *Front. Mol. Biosci.* **8**, 216 (2021).
54. E. Mylonas *et al.*, Domain conformation of tau protein studied by solution small-angle X-ray scattering. *Biochemistry* **47**, 10345–10353 (2008).
55. G. Platzer *et al.*, The metastasis-associated extracellular matrix protein osteopontin forms transient structure in ligand interaction sites. *Biochemistry* **50**, 6113–6124 (2011).
56. T. Mittag *et al.*, Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure* **18**, 494–506 (2010).
57. D. Kurzbach *et al.*, Detection of correlated conformational fluctuations in intrinsically disordered proteins through paramagnetic relaxation interference. *Phys. Chem. Chem. Phys.* **18**, 5753–5758 (2016).
58. A. Z. Panagiotopoulos, V. Wong, M. A. Floriano, Phase equilibria of lattice polymers from histogram reweighting Monte Carlo simulations. *Macromolecules* **31**, 912–918 (1998).
59. Y. H. Lin, H. S. Chan, Phase separation and single-chain compactness of charged disordered proteins are strongly correlated. *Biophys. J.* **112**, 2043–2046 (2017).
60. U. R. Shrestha, J. C. Smith, L. Petridis, Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations. *Commun. Biol.* **4**, 243 (2021).
61. J. A. Greig *et al.*, Arginine-enriched mixed-charge domains provide cohesion for nuclear speckle condensation. *Mol. Cell* **77**, 1237–1250.e4 (2020).
62. R. S. Fisher, S. Elbaum-Garfinkle, Tunable multiphase dynamics of arginine and lysine liquid condensates. *Nat. Commun.* **11**, 4628 (2020).
63. W. C. Wimley, S. H. White, Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* **3**, 842–848 (1996).
64. P. K. Ponnuswamy, M. M. Gromiha, Prediction of transmembrane helices from hydrophobic characteristics of proteins. *Int. J. Pept. Protein Res.* **42**, 326–341 (1993).
65. M. D. Mukrasch *et al.*, Structural polymorphism of 441-residue tau at single residue resolution. *PLoS Biol.* **7**, e34 (2009).
66. A. Jost Lopez, P. K. Quoika, M. Linke, G. Hummer, J. Köfinger, Quantifying protein–protein interactions in molecular simulations. *J. Phys. Chem. B* **124**, 4673–4685 (2020).
67. S. Elbaum-Garfinkle *et al.*, The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7189–7194 (2015).
68. M. T. Wei *et al.*, Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. *Nat. Chem.* **9**, 1118–1125 (2017).
69. N. O. Taylor, M. T. Wei, H. A. Stone, C. P. Brangwynne, Quantifying dynamics in phase-separated condensates using fluorescence recovery after photobleaching. *Biophys. J.* **117**, 1285–1300 (2019).
70. G. L. Dignon, W. Zheng, Y. C. Kim, J. Mittal, Temperature-controlled liquid-liquid phase separation of disordered proteins. *ACS Cent. Sci.* **5**, 821–830 (2019).
71. T. T. Foley, M. S. Shell, W. G. Noid, The impact of resolution upon entropy and information in coarse-grained models. *J. Chem. Phys.* **143**, 243104 (2015).
72. J. Jin, A. J. Pak, G. A. Voth, Understanding missing entropy in coarse-grained systems: Addressing issues of representability and transferability. *J. Phys. Chem. Lett.* **10**, 4549–4557 (2019).
73. G. Tesei *et al.*, Self-association of a highly charged arginine-rich cell-penetrating peptide. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 11428–11433 (2017).
74. X. Zeng, A. S. Holehouse, A. Chilkoti, T. Mittag, R. V. Pappu, Connecting coil-to-globule transitions to full phase diagrams for intrinsically disordered proteins. *Biophys. J.* **119**, 402–418 (2020).
75. J. Wirmer, W. Peti, H. Schwalbe, Motional properties of unfolded ubiquitin: A model for a random coil protein. *J. Biomol. NMR* **35**, 175–186 (2006).
76. G. Raos, G. Allegra, Chain collapse and phase separation in poor-solvent polymer solutions: A unified molecular description. *J. Chem. Phys.* **104**, 1626–1645 (1996).
77. J. Wen *et al.*, Conformational expansion of tau in condensates promotes irreversible aggregation. *J. Am. Chem. Soc.* **143**, 13056–13064 (2021).
78. A. P. Latham, B. Zhang, Maximum entropy optimized force field for intrinsically disordered proteins. *J. Chem. Theory Comput.* **16**, 773–781 (2020).
79. S. Elbaum-Garfinkle, Matter over mind: Liquid phase separation and neurodegeneration. *J. Biol. Chem.* **294**, 7160–7168 (2019).
80. D. G. Brown, J. Shorter, H. J. Wobst, Emerging small-molecule therapeutic approaches for amyotrophic lateral sclerosis and frontotemporal dementia. *Bioorg. Med. Chem. Lett.* **30**, 126942 (2020).
81. A Siegert *et al.*, Interplay between tau and α-synuclein liquid–liquid phase separation. *Protein Sci.* **30**, 1326–1336 (2021).
82. K. M. Ruff, F. Dar, R. V. Pappu, Polyphasic linkage and the impact of ligand binding on the regulation of biomolecular condensates. *Biophys. Rev.* **2**, 021302 (2021).
83. T. M. Perdikari *et al.*, A predictive coarse-grained model for position-specific effects of post-translational modifications. *Biophys. J.* **120**, 1187–1197 (2021).
84. S. Wohl, M. Jakubowski, W. Zheng, Salt-dependent conformational changes of intrinsically disordered proteins. *J. Phys. Chem. Lett.* **12**, 6684–6691 (2021).
85. J. A. Anderson, J. Glaser, S. C. Glotzer, HOOMD-blue: A python package for high-performance molecular dynamics and hard particle Monte Carlo simulations. *Comput. Mater. Sci.* **173**, 109363 (2020).
86. P. Eastman *et al.*, OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comput. Biol.* **13**, e1005659 (2017).
87. A. Cesari *et al.*, Fitting corrections to an RNA force field using experimental data. *J. Chem. Theory Comput.* **15**, 3425–3431 (2019).
88. S. Orioli, A. H. Larsen, S. Bottaro, K. Lindorff-Larsen, "How to learn from inconsistencies: Integrating molecular simulations with experimental data" in *Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly*, B. Strodel, B. Barz, Eds. (Elsevier, 2020), pp. 123–176.
89. P. Rotkiewicz, J. Skolnick, Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **29**, 1460–1465 (2008).
90. P. C. Schuur, Classification of acceptance criteria for the simulated annealing algorithm. *Math. Oper. Res.* **22**, 266–275 (1997).
91. G. Tesei, T. K. Schulze, R. Crehuet, K. Lindorff-Larsen, CG model of liquid-liquid phase behaviour of IDPs. GitHub. https://github.com/KULL-Centre/papers/tree/main/2021/CG-IDPs-Tesei-et-al. Accessed 30 September 2021.
92. G. Tesei, T. K. Schulze, R. Crehuet, K. Lindorff-Larsen, CG model of liquid-liquid phase behaviour of IDPs. Zenodo. https://doi.org/10.5281/zenodo.5005953. Accessed 10 September 2021.