# Accurate model of liquid-liquid phase behaviour of intrinsically-disordered proteins from data-driven 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 & the Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark; [b]CSIC-Institute for Advanced Chemistry of Catalonia (IQAC), Barcelona, Spain

This manuscript was compiled on June 23, 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 compartmentalisation of intracellular biochemical reactions. The phase behaviour of IDPs is sequence-dependent, and its investigation through molecular simulations requires protein models that combine computational efficiency with an accurate description of intra- 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.**

biomolecular condensates | liquid–liquid phase separation | intrinsically disordered proteins

Many intrinsically disordered proteins (IDPs) and proteins with disordered regions can condense into liquid-like droplets, viz. a biomolecule-rich phase coexisting with a more dilute solution (1–5). This de-mixing process is known as liquid-liquid phase separation (LLPS) and is one of the ways cells compartmentalise 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 post-translational modifications (PTMs) on the propensity to phase separate, is key to understand these processes. The sequences of IDPs undergoing LLPS are generally characterized by stretches of polar residues (spacers) interspersed by aromatic residues (stickers), which are instrumental for the formation of reversible physical cross-links via $\pi$-$\pi$, cation-$\pi$ and sp$^2$-$\pi$ 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 accurately predicting the phase behaviour 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), previous models are unable to consistently provide accurate predictions of the phase behaviour 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. Our starting point is the hydrophobicity scale (HPS) model (21) (with minor modification; see SI Materials and Methods) wherein, besides steric repulsion and salt-screened charge-charge interactions, residue-residue interactions are determined by hydropathy parameters ($\lambda$) which were derived from the atomic partial charges of a classical all-atom force field (26). To address the current limitations of the HPS model, we reevaluate the $\lambda$ parameters based on the analysis of 87 hydrophobicity scales. This intermediate model is further improved by optimizing the $\lambda$ parameters through a Bayesian parameter-learning procedure (27–32). The training set comprises 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 (33). Second, we employ a Bayesian regularization approach to prevent over-fitting 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 comaparing 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). Four, we perform coexistence simulations to test the models against the phase behaviour of A2 LCD (22, 24), FUS LCD (34, 35), variants of hnRNP A1 LCD (A1 LCD) (15, 16) and

Author contributions: G.T., T.K.S, R.C., and K.L.-L. designed research; G.T. and T.K.S. performed research; G.T. and R.C. developed software; G.T. and T.K.S. performed molecular simulations.; G.T. and T.K.S. analyzed data; G.T. and K.L.-L. wrote the paper with input from T.K.S and R.C.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: giulio.tesei@bio.ku.dk (G.T.) & lindorff@bio.ku.dk (K.L.-L.)

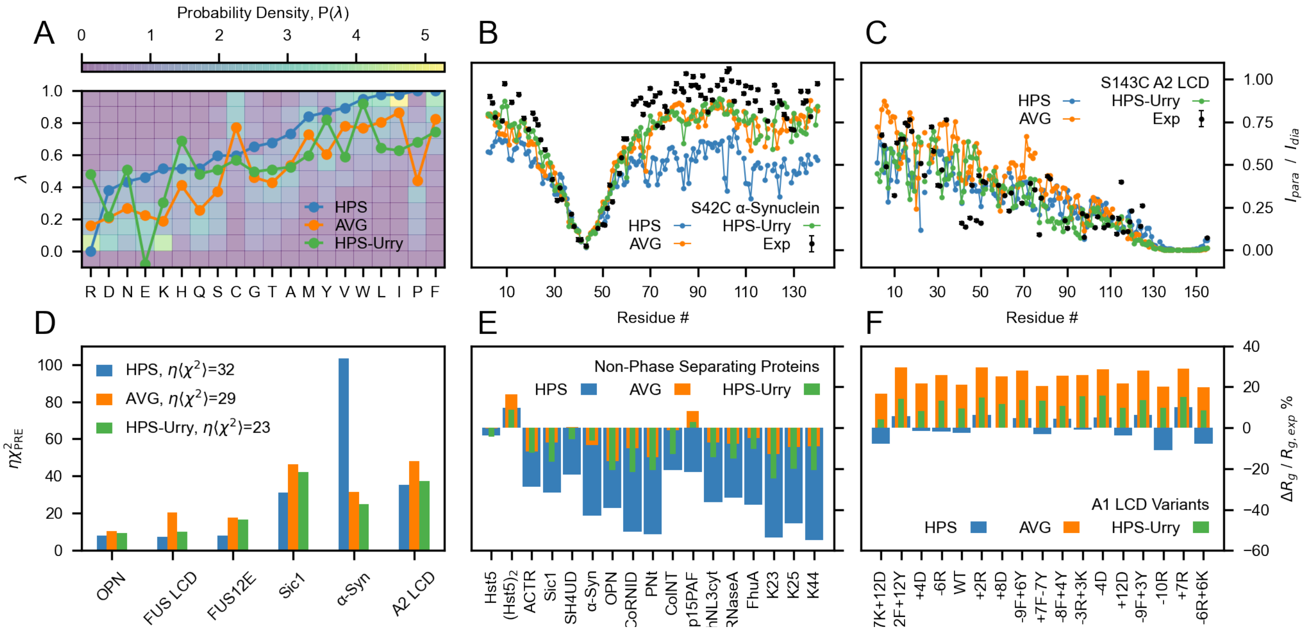

**Fig. 1.** (A) Probability distributions of the $\lambda$ parameters calculated from 87 min-max normalized hydrophobicity scales. Lines are the $\lambda$ parameters of the HPS model (blue), the average over the hydrophobicity scales (orange) and the HPS-Urry model (green) (36). Intramolecular PRE intensity ratios for the S43C mutant of $\alpha$-Synuclein (B) and the S243C mutant of A2 LCD (C) from simulations and experiments (22, 37) (black). (D) $\chi^2$ values quantifying the discrepancy between simulated and experimental intramolecular PRE data, scaled by the hyperparameter $\eta = 0.1$ (Materials and Methods). (E, F) Relative difference between simulated and experimental radii of gyration for non-phase separating sequences (E) and for variants of A1 LCD (F), with negative values corresponding to the simulated ensembles being more compact than in experiments.

the N-terminal region of the germ-granule protein Ddx4 (Ddx4 LCD) (8, 10, 13).

## Results and Discussion

**Analysis of Hydrophobicity Scales.** The $\lambda$ values of the original HPS model are based on a hydrophobicity scale derived from the atomic partial charges of the OPLS all-atom force field (26). 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 (38, 39). To put the HPS $\lambda$ values into context and find alternative models, we analyzed 98 hydrophobicity scales collected by Simm et al. (39). Each scale was 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. The resulting set of 87 scales was used to calculate the average scale (AVG) and the probability distribution of the $\lambda$ 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. 1A). Except for R and C residues, the HPS values are systematically larger than the AVG values. However, $\sum_{aa} P(\lambda_{aa}) = 37.2$ and 36.9 for the HPS and the AVG values, respectively, indicating that the two scales are comparably consistent with $P(\lambda)$.

We assessed the HPS and AVG parameter sets by running simulations of 45 IDPs ranging in length between 24 and 334 residues and compared the results against experiments. Specifically, we compared the simulations with the radii of gyration, $R_g$, of 42 IDPs (Tab. S1) and intramolecular PRE data of six IDPs (Tab. S2) (16, 22, 23, 37, 40–53).

Compared to the AVG scale, the HPS model overestimates the compaction of $\alpha$-Synuclein whereas it closely reproduces the PRE data for A2 LCD (Fig. 1B 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. 1D and F), while the AVG scales is considerably more accurate at reproducing the $R_g$ of non-phase separating proteins (Fig. 1E). The recently proposed HPS-Urry model (36) is the most accurate at predicting the intramolecular PRE data while it shows intermediate accuracy for the $R_g$ values of both non-phase separating proteins and A1 LCD variants.

**Optimization of Amino-Acid Specific Hydrophobicity Values.** To obtain a model that accurately predicts the conformational properties of IDPs of diverse sequences and LLPS propensities, we trained the $\lambda$ values on a large set of experimental $R_g$

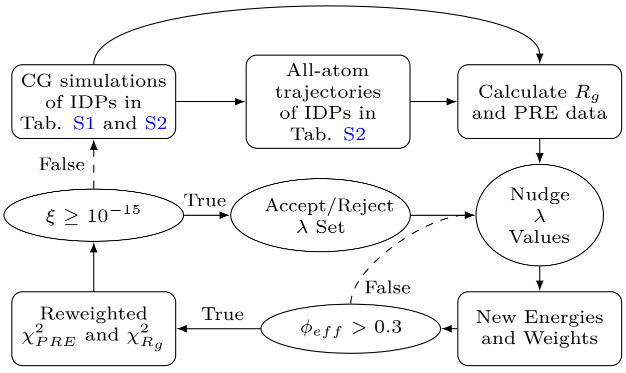
CG simulations of IDPs in Tab. S1 and S2
All-atom trajectories of IDPs in Tab. S2
Calculate $R_g$ and PRE data
False
$\xi \geq 10^{-15}$
True
Accept/Reject λ Set
Nudge λ Values
False
Reweighted $\chi^2_{PRE}$ and $\chi^2_{R_g}$
True
$\phi_{eff} > 0.3$
New Energies and Weights

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

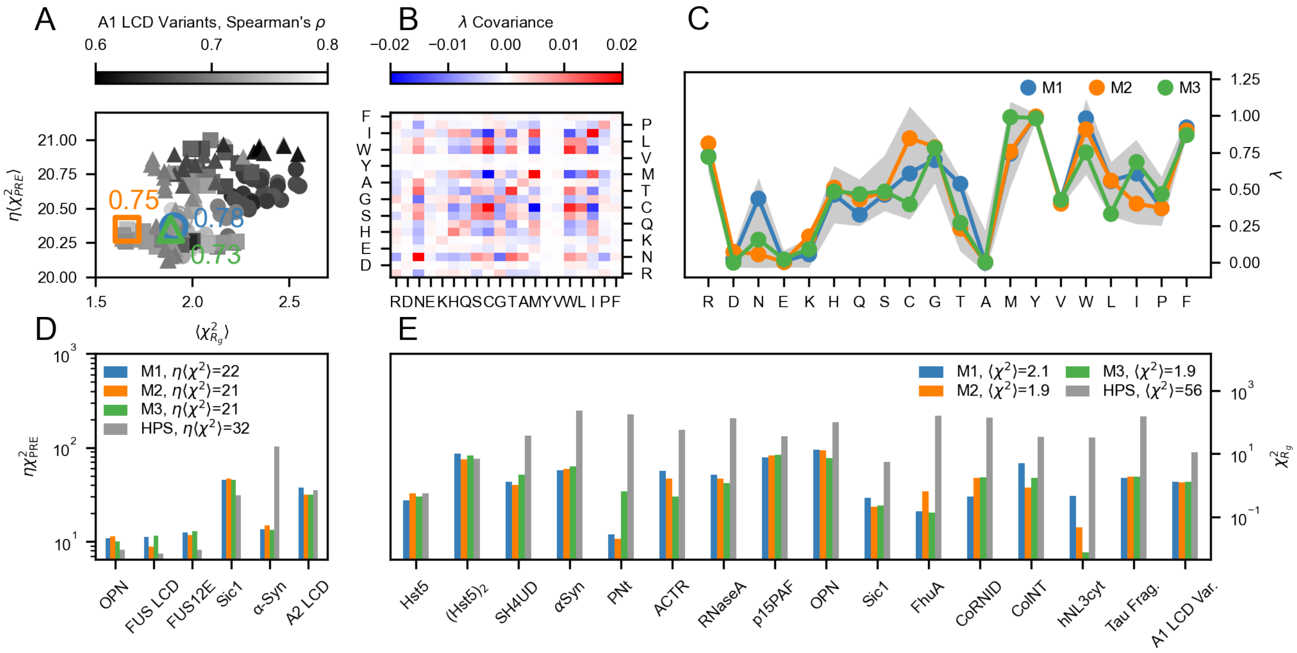

**Fig. 3.** (A) Overview of the optimal $\lambda$ sets with $\eta\chi^2_{PRE} < 21$ and $\chi^2_{R_g} < 3$ collected through the parameter learning procedures started from $\lambda_0$ =AVG (circles), M1 (squares) and M2 (triangles). The gray gradient shows the Spearman's correlation coefficient between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

and PRE data using a Bayesian parameter-learning procedure (27) shown schematically in Fig. 2 (Materials and Methods). We initially performed an optimization run starting from the AVG $\lambda$ values and setting the hyperparameters to $\theta = \eta = 0.1$ (Fig. S1A). We collected the optimized sets of $\lambda$ values which yielded $\eta\chi^2_{PRE} < 21$ and $\chi^2_{R_g} < 3$ (circles in Fig. 3A). The optimization was repeated starting from $\lambda = 0.5$ to assess that the parameter space sampled by our method is independent of the initial conditions (Fig. S2A and S1D). From the pool of optimized parameters, we selected the $\lambda$ 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 the selection of the optimal $\lambda$ set on the Spearman's correlation coefficient because we expect that capturing the experimental ranking in chain compaction will result in accurate predictions of the relative LLPS propensities (15, 16, 20, 54, 55). The selected model, referred to as M1 hereafter, is the starting point for two consecutive optimization cycles (Fig. S1C) which were performed with a lower weight for the prior ($\theta = 0.05$), yielding a new pool of optimized parameters (squares in Fig. 3A) 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 (Fig. S1D).

From the collected optimal parameters (triangles in Fig. 3A), we selected M3 (largest $\rho = 0.73$). As shown in Fig. 3B, the optimal $\lambda$ values collected through the four independent optimization runs (Fig. S1A–D) are weakly intercorrelated. The covariance values range between -0.015 and 0.015 for most amino acids, with the exception of the standard deviations of N, C, T, M, W, and I. C, M, W, and I are among the least frequent amino acids in the training set (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. 3C shows that M1–3 fall within two standard deviations (SDs) above and below the mean of the $\lambda$ values yielding $\eta\chi^2_{PRE} < 21$ and $\chi^2_{R_g} < 3$ (gray shaded area). Despite the significant differences, M1–3 fit the training data equally accurately and result in an improvement in $\chi^2_{PRE}$ and $\chi^2_{R_g}$ of ∼30% and ∼95% with respect to the HPS model, respectively (Fig. 3D,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 and phase-separating IDPs (Fig. S4B–D and Fig. S5–S10) as well as in radii of gyration with relative errors

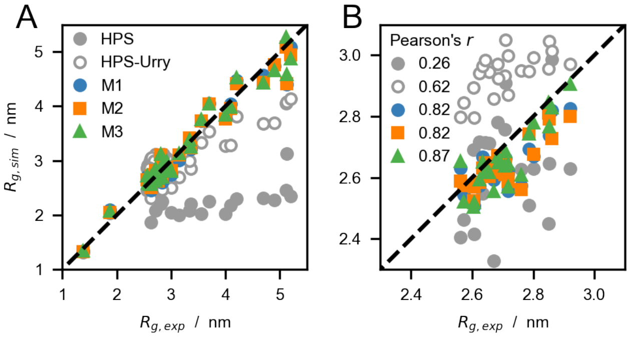

**Fig. 4.** (A) Comparison between experimental and predicted radii of gyration (Tab. S1), $R_g$, for the HPS, HPS-Urry, and M1–3 models. (B) Zoom-in on the $R_g$ values of the A1 LCD variants.

$-14\% < \Delta R_g \ / \ R_{g,exp} < 12\%$ (Fig. S4E,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, $\nu$, for the variants of A1 LCD (Fig. 4B and S11). The lower Pearson's correlation coefficients observed for $\nu$, compared to the corresponding $R_g$ data, may originate from the different forward models used to infer $\nu$ from SAXS experiments and simulation data, i.e., respectively, the molecular form factor method (16, 48) and least-squares fit to long intramolecular pairwise distances, $R_{ij}$, vs $|i-j| > 10$ (56) (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 $\lambda$ set and models M1–3 is the considerably smaller values for R and Y residues (Fig. S2C). Indeed, the large $\lambda$ 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, 57, 58).

**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, $\Gamma_2$, 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 $\Gamma_2$ 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 DEER-PREdict (33), assuming an effective correlation time of the spin label, $\tau_t$, of 100 ps and fitting the overall molecular correlation time, $\tau_c$, within the interval $1 \leq \tau_c \leq 20$ ns. In agreement with experiments, $\Gamma_2$ values predicted by the M1–3 models are characterized by no distinctive peaks along the protein sequence (Fig. 5A–E), which is consistent with transient and non-specific protein–protein interactions. Notably, while PRE rates for FUS LCD are of the same magnitude for all spin labeled sites, A2 LCD presents larger $\Gamma_2$ values for S99C than for S143C indicating that the tyrosine-rich aggregation-prone region (residues 84–107) is involved in more frequent intermolecular contacts with the entire sequence. The discrepancy between predicted and experimental intermolecular PRE data, $\chi^2_{PRE}$, varies significantly as a function of $\tau_c$ (Fig. 5F–G). For both FUS LCD and A2 LCD, the optimal $\tau_c$ is larger for M1 than for M3, which suggests that the latter has more attractive intermolecular interactions. While for M1 the minimum of $\chi^2_{PRE}$ is at $\tau_c = 17$ ns for both proteins, for M3 the optimal $\tau_c$ value is $\sim 8$ ns smaller for FUS LCD than A2 LCD. Although the accuracy of $\tau_c$ is difficult to assess in the case of transiently interacting IDPs, this large difference in $\tau_c$ (Fig. 5) suggests

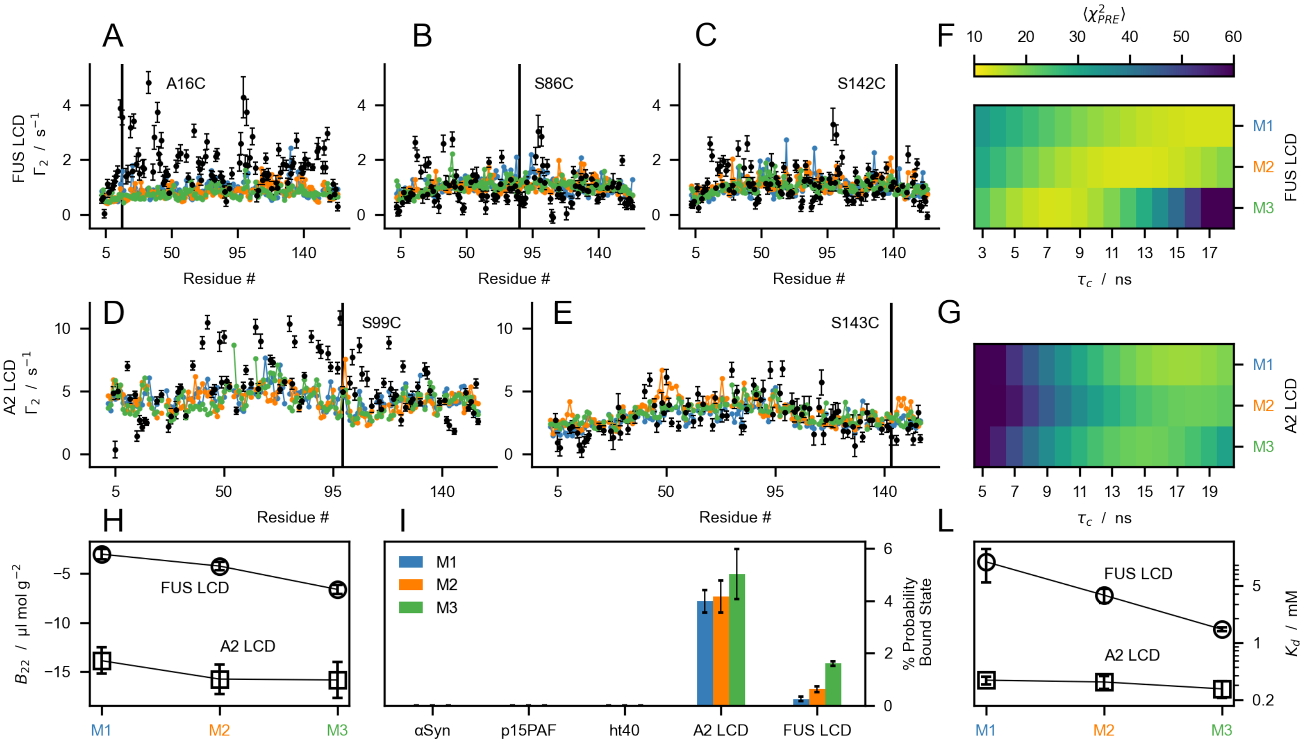

**Fig. 5.** (A–E) Comparison between experimental (black) intermolecular PRE rates (Tab. S3) and predictions from the M1 (blue), M2 (orange) and M3 (green) models for FUS LCD (A–C) and A2 LCD (D,E) calculated using the best-fit correlation time, $\tau_c$ . (F–G) Discrepancy between calculated and experimental intermolecular PRE rates $\chi^2_{PRE}$ as a function of $\tau_c$. (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. (L) 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 ten simulation replicas.

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 Materials and Methods). 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 $\lambda$ values and amino acid compositions, M3 presents the most negative $B_{22}$ values (large $\lambda$ values for Q, G and P), followed by M2 and M1 (Fig. 5I).

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 $\alpha$-Synuclein, p15PAF, full length tau (ht40), A2 LCD and FUS LCD performed at the solution conditions of the reference experimental data (37, 46, 59) (SI Materials and Methods). In agreement with experimental findings, we find that the highly soluble $\alpha$-Synuclein, p15PAF and ht40 proteins do not self-associate substantially in our simulations, whereas A2 LCD and FUS LCD have $p_B$ ~4% and ~1%, respectively. We further estimated the dissociation constants of A2 LCD and FUS LCD using $K_d = (1 - p_B)^2/(N_A p_B V)$ and $K_d = 1/(N_A p_B(V - B_{22}))$ self-consistently (60), where $N_A$ is Avogadro's number (SI Materials and Methods, Fig. S13 and Fig. 5L).

**Testing LLPS propensities.** To test the ability of the models to capture the sequence-dependence of LLPS propensity, we performed multi-chain simulations in a slab geometry and calculated protein concentrations of the coexisting condensate, $c_{con}$, and dilute phase, $c_{sat}$. Simulation results are tested against an extensive set of sequences which have been shown to undergo LLPS below an upper critical solution temperature (UCST), namely FUS LCD (23, 34, 35), A2 LCD (22, 24), the NtoS variant of A2 LCD (24) 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 (Fig. S14). 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 (Fig. S15 and S16) for all the proteins in Fig. 6, simulations were carried out at 50°C for most sequences, and at 24°C for the HPS-Urry model as well as for the Ddx4 LCD variants (Tab. S4).

Simulations of M1 and M2 at 50°C recapitulate the experimental trend in $c_{sat}$ across the diverse sequences (Fig. 6A,D) and also reproduce the reference $c_{con}$ and $c_{sat}$ values measured at room temperature. M3 and HPS overestimate the relative LLPS propensity of FUS LCD, whereas HPS-Urry underestimates the LLPS propensity of A1 LCD. We further test our predictions against 15 variants of A1 LCD (Fig. 6B,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, $\pm N_X$X$\pm N_Z$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 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 (Tab. S1), this result supports our model development strategy.

M1–3 and the recently proposed HPS-Urry model (36) reproduce the experimental ranking for LLPS propensity of the Ddx4 LCD variants, i.e. WT$\gg$CS>FtoA$\gtrsim$RtoK (Fig. 6C,F). Albeit smaller by an order of magnitude, the $c_{sat}$ values predicted by these models strongly correlate with the experimental values measured at the same temperature, with Pearson's correlation coefficients of 0.86 for the HPS-Urry model and

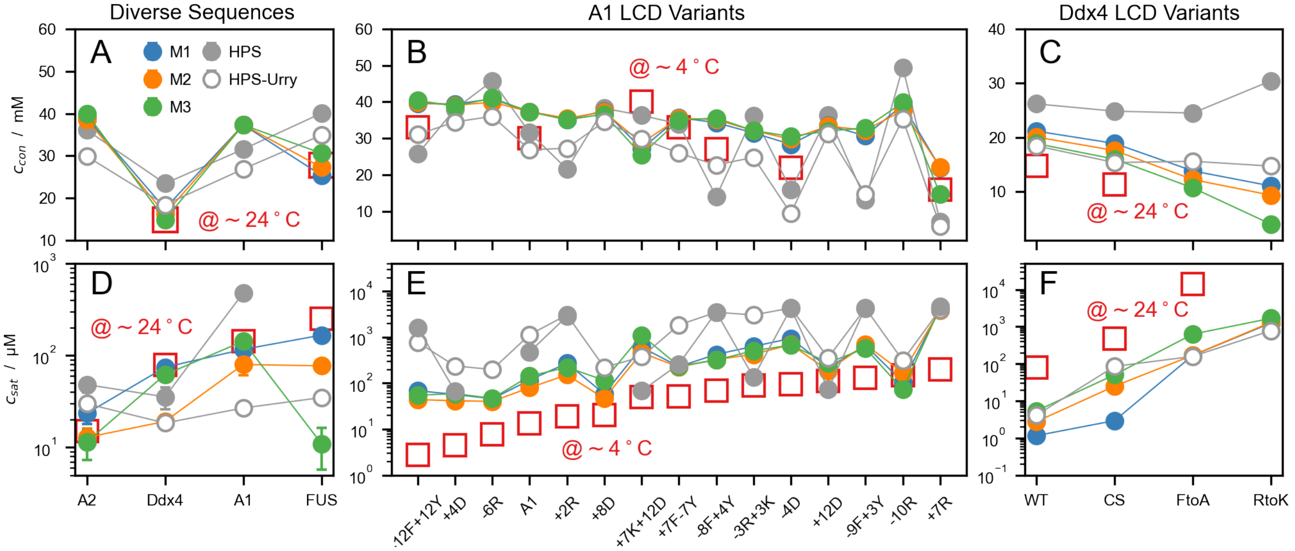

**Fig. 6.** Protein concentrations in the condensate (A–C) and in the dilute phase (D–E) from slab simulations of the M1–3, HPS and HPS-Urry models. Simulations of the M1–3 and HPS models were performed at 50°C, except for the Ddx4 LCD variants in panels D and F which were conducted at 24°C. The HPS-Urry model was simulated at 24°C for all the proteins. Red open squares indicate experimental measurements at ~24°C, except for the concentrations of A1 LCD variants in panels B and E which were measured at ~4°C. Error bars are SEMs of averages over blocks of 0.3 µs.

exceeding 0.99 for M1–3. On the other hand, 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. 6C), in apparent contrast to experimental observations (10, 13). M3 displays the lowest LLPS propensities, especially for the FtoA variant (14 F residues mutated to A) whereas M1 and M2 yield comparable $c_{sat}$ values and significantly differ only for the charge scrambled (CS) variant, which has the same net charge and amino acid composition as the WT but a more uniform charge distribution along the linear sequence.

As we observe, M1 and M2 differ mainly for the $\lambda$ value of the N residues and perform equally well against the test set of Fig. 6. To assess which model is more accurate, we test the ability to predict the LLPS propensity of the NtoS variant of A2 LCD with respect to the wild type. Only the M1 model, which has $\lambda$ values for N and S of similar magnitude correctly predicts approximately the same LLPS propensity for variant and WT (Fig. S17), in agreement with experiments (24).

**Comparing intra- and inter-molecular interactions.** After establishing the ability of model M1 to accurately predict trends in LLPS propensity for diverse sequences, we analyze the non-electrostatic 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 intra- and intermolecular interaction patterns for both proteins (Fig. 7), consistent with a mostly uniform distribution of stickers along the linear sequence (Fig. 7G,H) (15, 61). 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 role of R as a sequence dependent sticker (16), whereby the large $\lambda$ value for R in models M1–3 presumably reflects the ability of the amphiphilic guanidinium moiety to engage in H-bonding, as well as $\pi$ stacking and charge-$\pi$ interactions (18). Further, in the dilute regime, the intra- 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 1D interaction energy projections. This result is consistent with the faster local motions of the terminal residues inferred from $^{15}$N NMR relaxation data e.g. for a number of phase separating IDPs (15, 22, 23). We also find that the aggregation-prone Y-rich region of A2 LCD (residues 84–107) interacts with the entire polypeptide chain (Fig. 7D–F) and thus likely drives chain compaction, self-association as well as LLPS. Finally, 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, and that differences in compaction between wild-type and charge variants of A1 LCD are greater in the dilute than in the condensed phase (Fig. S18)

**Correlating single-chain properties and phase separation.** Motivated by the above similarity, we perform a detailed analysis of the coupling between chain compaction and phase behaviour of the A1 LCD variants. The $\log_{10}(c_{sat})$ values predicted for A1 LCD variants by the M1–3 models at 50°C linearly correlate with the experimental values measured at

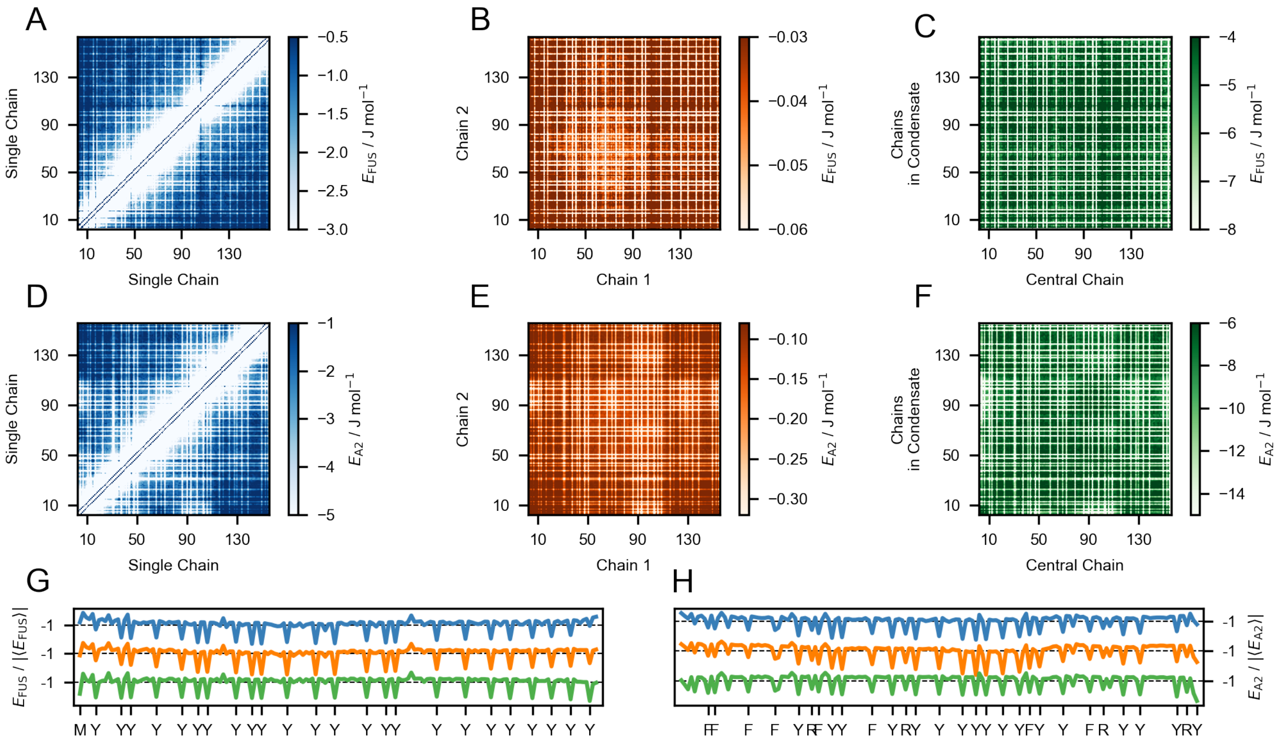

**Fig. 7.** Energy maps from simulations of the M1 model of FUS LCD (A–C) and A2 LCD (D–F) calculated using non-electrostatic interaction energies. (G–H) 1D projections of the energy maps for FUS LCD (G) and A2 LCD (H), normalized by the absolute average interaction energy $|\langle E\rangle|$ and shifted 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).

4°C (Pearson's correlation coefficients > 0.7) (Fig. 8A). Moreover, variants of intermediate LLPS propensity simulated at 37°C are in agreement with the predictions of the M1–3 models with the reference $c_{sat}$ values at 4°C (Fig. S14). Since $c_{sat}$ values from simulations at 37°C and 50 °C are highly correlated (Fig. 8B), we can use the linear correlations shown in Fig. 8A,C to infer the ranking in predictive performance of the various models with respect to the phase behaviour of the A1 LCD variants, i.e. M2 performs the best, followed by M1, M3, HPS-Urry and HPS.

In agreement with previous observations (16), the $\log_{10}(c_{sat})$ values for the aromatic variants show a linear relationship with the scaling exponent, $\nu_{sim}$, whereas changes in the number of charged residues (charge variants) result in significant deviations from the lines of best fit (Fig. 8D–F). Following the approach of Bremer, Farag 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. 8G–I). The results for M1 and M2 show the V-shaped profile observed for the experimental data (16), which reveals that mean-field electrostatic repulsion between the net charge of the proteins is responsible for breaking the coupling between chain compaction and LLPS propensity. 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 non-electrostatic energy maps of +8D (NCPR=0), +4D (NCPR=0.3) and -4D (NCPR=0.9) with respect to the wild type of A1 LCD (Fig. S19 and S20). While in 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=0.3 and NCPR=0.9, respectively (Fig. S19G–I and S20G–I). Although the interaction patterns are still dominated by the stickers, deviations of the net charge 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 compensate for the repulsion introduced by the excess positive charge by allowing for favorable interactions with both Y and negatively charged residues. As opposed to M1–2, 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 behaviour 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 3–4% in the $\lambda$ values of the residues present in A1 LCD.

## 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 the method succeeds in agreement

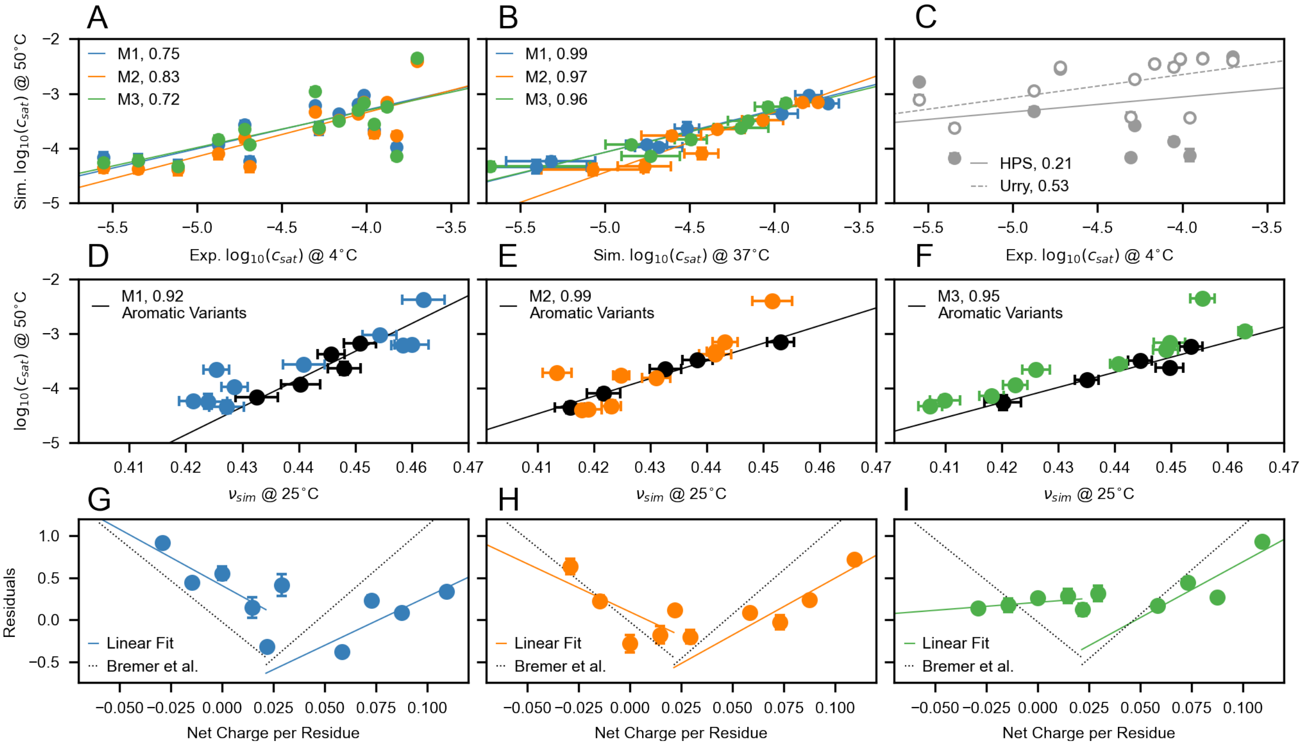

**Fig. 8.** (A) Correlation between $\log_{10}(c_{sat})$ from experiments at 4°C (16) and from simulations of the A1 LCD variants performed at 50°C using the M1–3 models. (B) Correlation between $\log_{10}(c_{sat})$ at 50 and 37°C calculated from simulations of the M1–3 models. (C) Correlation between $\log_{10}(c_{sat})$ from experiments at 4°C (16) and from simulations of the HPS and HPS-Urry models performed at 50 and 24°C, respectively. (D–F) $\log_{10}(c_{sat})$ vs $\nu_{sim}$ for A1-LCD variants from simulations performed using the M1 (D), M2 (E) and M3 (F) models. Black and colored circles indicate aromatic and charge variants, respectively. Black lines are linear fits to the aromatic variants. (G–I) Residuals from the linear fits of panels D–F for the charge variants of A1 LCD as a function of the net charge per residue. 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 $\nu_{sim}$ are SDs from the linear fit to $\log(R_{ij})$ vs $\log(|i-j|)$, with $|i-j| > 10$. Solid and dashed lines are linear fits to the data. Dotted lines in G–I are lines of best fit to the experimental data by Bremer et al. (16).

with the previously observed coupling between chain compaction and propensity for phase separation (15, 20, 54, 55), but also appears to recapitulate the recent discovery that charge effects may break this relationship (16). Our simulations reveal that, for sequences characterized by a uniform distribution of stickers, residue-residue interactions determining chain compaction also drive self-association and LLPS. Our model optimization with and without the A1 LCD variants indicates that the presence of phase-separating IDPs in the training set is necessary for the parameter-learning procedure to capture the role of Y and R residues as stickers. Finally, using the model optimized herein, we show that the experimentally observed dependency of LLPS on protein net charge appears to be captured by salt-screened electrostatic repulsion, 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 behaviour. Further, we suggest the model could be applied to study the influence of disease-associated mutations on the material properties of protein self-coacervates (62, 63), the LLPS of protein mixtures as a function of composition, and the partitioning of non-phase separating proteins into condensates (64). Finally, owing to the generalized parameter-learning approach, the model could be readily refined as new experimental data are collected and it should be possible to extend it to account for PTMs (65) and the temperature dependence of solvent mediated interactions (66).

## Materials and Methods

We use the C$\alpha$-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 Materials and Methods). Langevin dynamics simulations are conducted using HOOMD-blue v2.9 (67) in the $NVT$ ensemble using the Langevin thermostat with a time step of 5 fs and friction coefficient 0.01 ps$^{-1}$ (SI Materials and Methods).

**Bayesian Parameter-Learning Procedure.** The $\lambda$ values are optimized using a Bayesian parameter-learning procedure (27, 30, 68). The training set consists of the experimental $R_g$ values of 42 IDPs (Tab. S1) and the intramolecular PRE data of six proteins (Tab. S2) (16, 22, 23, 37, 40–53). To guide the optimization within physically reasonable parameters and to avoid over-fitting the training set, we introduce a regularization term which penalizes deviations of the $\lambda$ 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 (Tab. S1);

2. Conversion of CG simulations into all-atom trajectories using PULCHRA (69);

3. Calculation of per-frame radii of gyration and PRE data. The PRE rates and intensity ratios are calculated using DEER-PREdict (33) 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 $\lambda$ values which are nudged by random numbers picked from a normal distribution of standard deviation 0.05. The prior probability distribution, $P(\lambda)$, sets the bounds of the parameter space: any $\lambda_i$ for which $P(\lambda_i) = 0$ is further nudged until $P(\lambda_i) \neq 0$.

5. Calculation of the Boltzmann weights for the $i^{th}$ frame as $w_i = \exp -[U(\boldsymbol{r_i}, \boldsymbol{\lambda_k}) - U(\boldsymbol{r_i}, \boldsymbol{\lambda_0})]/k_BT$, where $U(\boldsymbol{r_i}, \boldsymbol{\lambda_k})$ and $U(\boldsymbol{r_i}, \boldsymbol{\lambda_0})$ are the total Ashbaugh-Hatch energies of the $i^{th}$ frame for trial and initial $\lambda$ 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 $\sigma_{ij}^{exp}$ is the error on the experimental values, $Y$ is either $I_{para}/I_{dia}$ or $\Gamma_2$, $N_{labels}$ is the number of spin-labeled mutants and $N_{res}$ is the number of measured residues;

7. Following the Metropolis criterion (70), the $k^{th}$ set of $\lambda$ values is accepted with probability:
$$A_{k-1\rightarrow k} = \begin{cases} \exp\left[\frac{\mathcal{L}(\boldsymbol{\lambda_{k-1}})]-\mathcal{L}(\boldsymbol{\lambda_k})}{\xi_k}\right], & \mathcal{L}(\boldsymbol{\lambda_k}) > \mathcal{L}(\boldsymbol{\lambda_{k-1}}) \\ 1, & \mathcal{L}(\boldsymbol{\lambda_k}) \leq \mathcal{L}(\boldsymbol{\lambda_{k-1}}), \end{cases} \qquad [4]$$
where the control parameter, $\xi_k$, 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. $\theta$ and $\eta$ are hyperparameters of the optimization procedure. $\theta$ determines the trade-off between between over- and under-fitting the training set whereas $\eta$ sets the relative weight of the PRE data with respect to the radii of gyration.

Steps 4–7 are iterated until $\xi < 10^{-15}$, when the reweighting cycle is interrupted and new CG simulation carried out with the trained $\lambda$ 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 $\phi_{eff}$ results in average absolute differences between $\chi^2$ values estimated from reweighting and calculated from trajectories performed with the corresponding parameters of ~1.8 and ~0.8 for $\eta\chi^2_{PRE}$ and $\chi^2_{R_g}$, respectively (Fig. S21).

**Data deposition.** Datasets, code and Jupyter Notebooks for reproducing our analyses are available on GitHub at github.com/KULL-Centre/papers/tree/master/2021/CG-IDPs-Tesei-et-al and on Zenodo, DOI: 10.5281/zenodo.5005954.

**ACKNOWLEDGMENTS.** We thank Veronica Ryan and Nicolas L. Fawzi for generously sharing the PRE data for FUS LCD, FUS12E LCD and A2 LCD as well as Robert Konrat for generously sharing the intramolecular PRE data for Osteopontin. We thank Robert B. Best for sharing data on compaction of IDPs, and Tanja Mittag for comments on the manuscript. We acknowledge funding from the BRAINSTRUC structural biology initiative from the Lundbeck Foundation (R155-2015-2666), and acknowledge access to computational resources from the ROBUST Resource for Biomolecular Simulations (supported by the Novo Nordisk Foundation; 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 programme under the Marie Skłodowska-Curie grant agreement No. 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. *The EMBO J.* **37**, e98049 (2018).
3. NM 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., $\alpha$-synuclein aggregation nucleates through liquid–liquid phase separation. *Nat. Chem.* **12**, 705–716 (2020).
5. MC Hardenberg, et al., Observation of an $\alpha$-synuclein liquid droplet state and its maturation into lewy body-like assemblies. *J. Mol. Cell Biol.* **n/a** (2021) mjaa075.
6. Y Shin, CP Brangwynne, Liquid phase condensation in cell physiology and disease. *Science* **357**, eaaf4382 (2017).
7. NB Nedelsky, JP Taylor, Bridging biophysics and neurology: aberrant phase transitions in neurodegenerative disease. *Nat. Rev. Neurol.* **15**, 272–286 (2019).
8. TJ Nott, et al., Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell* **57**, 936–947 (2015).
9. CP Brangwynne, P Tompa, RV Pappu, Polymer physics of intracellular phase transitions. *Nat. Phys.* **11**, 899–904 (2015).
10. RM Vernon, et al., Pi-pi contacts are an overlooked protein feature relevant to phase separation. *eLife* **7** (2018).
11. BS Schuster, et al., Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proc. Natl. Acad. Sci.* **117**, 11421–11431 (2020).
12. GL Dignon, RB Best, J Mittal, Biomolecular phase separation: From molecular driving forces to macroscopic properties. *Annu. Rev. Phys. Chem.* **71**, 53–75 (2020).
13. JP 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.* **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. EW 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* **n/a** (2021).
17. G Krainer, et al., Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions. *Nat. Commun.* **12** (2021).
18. M Vazdar, et al., Arginine “magic”: Guanidinium like-charge ion pairing from aqueous salts to cell penetrating peptides. *Accounts Chem. Res.* **51**, 1455–1464 (2018).
19. MJ Fossat, X Zeng, RV Pappu, Uncovering differences in hydration free energies and structures for model compound mimics of charged side chains of amino acids. *The J. Phys. Chem. B* **125**, 4148–4161 (2021).
20. GL Dignon, W Zheng, RB Best, YC Kim, J Mittal, Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl. Acad. Sci.* **115**, 9929–9934 (2018).
21. GL Dignon, W Zheng, YC Kim, RB Best, J Mittal, Sequence determinants of protein phase behavior from a coarse-grained model. *PLOS Comput. Biol.* **14**, e1005941 (2018).
22. VH 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. *The EMBO J.* **36**, 2951–2967 (2017).
24. VH Ryan, et al., Tyrosine phosphorylation regulates hnRNPA2 granule protein partitioning and reduces neurodegeneration. *The EMBO J.* **40** (2020).
25. S Das, YH Lin, RM Vernon, JD Forman-Kay, HS Chan, Comparative roles of charge, $\pi$, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proc. Natl. Acad. Sci.* **117**, 28795–28805 (2020).
26. LH Kapcha, PJ Rossky, A simple atomic-level hydrophobicity scale reveals protein interfacial structure. *J. Mol. Biol.* **426**, 484–498 (2014).
27. AB 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).
28. LP Wang, TJ Martinez, VS Pande, Building force fields: An automatic, systematic, and reproducible approach. *The J. Phys. Chem. Lett.* **5**, 1885–1891 (2014).
29. AP Latham, B Zhang, Maximum entropy optimized force field for intrinsically disordered proteins. *J. chemical theory computation* **16**, 773–781 (2019).
30. A Cesari, et al., Fitting corrections to an rna force field using experimental data. *J. chemical theory computation* **15**, 3425–3431 (2019).
31. G Tiana, L Giorgetti, *Coarse Graining of a Giant Molecular System: The Chromatin Fiber*, eds. M Bonomi, C Camilloni. (Springer New York, New York, NY), pp. 399–411 (2019).
32. T Dannenhoffer-Lafage, RB Best, A data-driven hydrophobicity scale for predicting liquid–liquid phase separation of proteins. *The J. Phys. Chem. B* **125**, 4046–4056 (2021).
33. 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).
34. KA Burke, AM Janke, CL Rhine, NL 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).
35. AC Murthy, et al., Molecular interactions underlying liquid-liquid phase separation of the FUS low-complexity domain. *Nat. Struct. & Mol. Biol.* **26**, 637–648 (2019).
36. RM Regy, J Thompson, YC Kim, J Mittal, Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. *Protein Sci.* **n/a** (2021).
37. MM Dedmon, K Lindorff-Larsen, J Christodoulou, M Vendruscolo, CM Dobson, Mapping long-range interactions in $\alpha$-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J. Am. Chem. Soc.* **127**, 476–477 (2005).
38. HS Chan, *Amino Acid Side-chain Hydrophobicity.* (American Cancer Society), (2002).
39. S Simm, J Einloft, O Mirus, E Schleiff, 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification. *Biol. Res.* **49** (2016).
40. S Jephthah, L Staby, BB Kragelund, M Skepö, Temperature dependence of intrinsically disordered proteins in simulations: What are we missing? *J. Chem. Theory Comput.* **15**, 2672–2683 (2019).
41. E Fagerberg, LK Månsson, S Lenton, M Skepö, The effects of chain length on the structural properties of intrinsically disordered proteins in concentrated solutions. *The J. Phys. Chem. B* **124**, 11843–11853 (2020).
42. M Kjaergaard, et al., Temperature-dependent structural changes in intrinsically disordered proteins: Formation of $\alpha$-helices or loss of polyproline ii? *Protein Sci.* **19**, 1555–1564 (2010).
43. GNW 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).
44. UR Shrestha, et al., Generation of the configurational ensemble of an intrinsically disordered protein from unbiased molecular dynamics simulation. *Proc. Natl. Acad. Sci.* **116**, 20446–20452 (2019).
45. CL Johnson, et al., The two-state prehensile tail of the antibacterial toxin colicin n. *Biophys. J.* **113**, 1673–1684 (2017).
46. AD 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).
47. 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).
48. JA Riback, et al., Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **358**, 238–241 (2017).
49. MC Ahmed, et al., Refinement of $\alpha$-synuclein ensembles against SAXS data: Comparison of force fields and methods. *Front. Mol. Biosci.* **8** (2021).
50. E Mylonas, et al., Domain conformation of tau protein studied by solution small-angle x-ray scattering. *Biochemistry* **47**, 10345–10353 (2008).
51. G Platzer, et al., The metastasis-associated extracellular matrix protein osteopontin forms transient structure in ligand interaction sites. *Biochemistry* **50**, 6113–6124 (2011).
52. 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).
53. 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).
54. AZ Panagiotopoulos, V Wong, MA Floriano, Phase equilibria of lattice polymers from histogram reweighting monte carlo simulations. *Macromolecules* **31**, 912–918 (1998).
55. YH Lin, HS Chan, Phase separation and single-chain compactness of charged disordered proteins are strongly correlated. *Biophys. J.* **112**, 2043–2046 (2017).
56. UR Shrestha, JC Smith, L Petridis, Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations. *Commun. Biol.* **4** (2021).
57. JA Greig, et al., Arginine-enriched mixed-charge domains provide cohesion for nuclear speckle condensation. *Mol. cell* **77**, 1237–1250 (2020).
58. RS Fisher, S Elbaum-Garfinkle, Tunable multiphase dynamics of arginine and lysine liquid condensates. *Nat. communications* **11**, 1–10 (2020).
59. MD Mukrasch, et al., Structural polymorphism of 441-residue tau at single residue resolution. *PLoS Biol.* **7**, e1000034 (2009).
60. AJ Lopez, PK Quoika, M Linke, G Hummer, J Köfinger, Quantifying protein–protein interactions in molecular simulations. *The J. Phys. Chem. B* **124**, 4673–4685 (2020).
61. X Zeng, AS Holehouse, A Chilkoti, T Mittag, RV Pappu, Connecting coil-to-globule transitions to full phase diagrams for intrinsically disordered proteins. *Biophys. J.* **119**, 402–418 (2020).
62. S Elbaum-Garfinkle, Matter over mind: Liquid phase separation and neurodegeneration. *J. Biol. Chem.* **294**, 7160–7168 (2019).
63. DG Brown, J Shorter, HJ Wobst, Emerging small-molecule therapeutic approaches for amyotrophic lateral sclerosis and frontotemporal dementia. *Bioorganic & Medicinal Chem. Lett.* **30**, 126942 (2020).
64. A Siegert, et al., Interplay between tau and $\alpha$-synuclein liquid–liquid phase separation. *Protein Sci.* **n/a** (2021).
65. TM Perdikari, et al., A predictive coarse-grained model for position-specific effects of post-translational modifications. *Biophys. J.* **120**, 1187–1197 (2021).
66. GL Dignon, W Zheng, YC Kim, J Mittal, Temperature-controlled liquid-liquid phase separation of disordered proteins. *ACS Cent. Sci.* **5**, 821–830 (2019).
67. JA Anderson, J Glaser, SC Glotzer, HOOMD-blue: A python package for high-performance molecular dynamics and hard particle monte carlo simulations. *Comput. Mater. Sci.* **173**, 109363 (2020).
68. S Orioli, AH 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.* (Elsevier), pp. 123–176 (2020).
69. P Rotkiewicz, J Skolnick, Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **29**, 1460–1465 (2008).
70. PC Schuur, Classification of acceptance criteria for the simulated annealing algorithm. *Math. Oper. Res.* **22**, 266–275 (1997).