# Improved Predictions of Phase Behaviour of Intrinsically Disordered Proteins by Tuning the Interaction Range

Giulio Tesei[1,*] and Kresten Lindorff-Larsen[1,*]

**1** Structural Biology and NMR Laboratory & the Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark

* giulio.tesei@bio.ku.dk, lindorff@bio.ku.dk

## Abstract

The formation and viscoelastic properties of condensates of intrinsically disordered proteins (IDPs) is dictated by amino acid sequence and solution conditions. Because of the involvement of biomolecular condensates in cell physiology and disease, furthering our understanding of the relationship between protein sequence and phase separation (PS) may have important implications in formulating new therapeutic hypotheses. Here, we present CALVADOS 2, a coarse-grained model of IDPs that accurately predicts conformational properties and propensity to undergo PS for diverse sequences and solution conditions. In particular, we systematically study the effect of varying the range of nonionic interactions and use our findings to improve the temperature scale of the model. We further optimize the residue-specific model parameters against experimental data on the chain compaction of 55 proteins and 70 hydrophobicity scales from the literature. Extensive testing shows that the model accurately predicts chain compaction and PS propensity for sequences of diverse length, charge patterning, and at disparate temperatures and salt concentrations.

## 1 Introduction

Biomolecular condensates may form via phase separation (PS) into coexisting solvent-rich and macromolecule-rich phases. PS is driven by multiple transient interactions which are in many cases engendered by intrinsically disordered proteins (IDPs) and low-complexity domains (LCDs) of multi-domain proteins [1–5]. The propensity of IDPs to phase separate and the viscoelastic properties of the condensates are dictated by the amino acid sequence of the constituent IDPs. Moreover, condensates of some IDPs reconstituted in vitro tend to undergo a transition to a dynamically-arrested state, wherein oligomeric species can nucleate and ultimately aggregate into fibrils [2, 3, 6–14]. As accumulating evidence suggests that these processes may be involved in neurodegeneration and cancer [15–18], understanding how PS and rheological properties of condensates depend on protein sequence is a current research focus. Due to the transient nature of the protein-protein interactions underpinning PS, quantitative characterization of biomolecular condensates via biophysical experimental methods is challenging, and hence molecular simulations have played an important role in aiding the interpretation of experimental data on condensates reconstituted in vitro [19]. Molecular simulations of the PS of IDPs require a minimal system size of $\sim$100 chains and long simulation times to sample the equilibrium properties of the two-phase system. Therefore, to enhance computational efficiency, it is useful to reduce the complexity of the system by modeling the solvent as a continuum while coarse-graining the atomistic representation of the protein to fewer interaction sites.

A widely used class of coarse-grained models of IDPs describes each residue as a single site centered at the C$\alpha$ atom. Charged residues interact via salt-screened electrostatic interactions whereas the remaining nonionic nonbonded interactions are incorporated in a single short-range potential characterized by a set of "stickiness" parameters. The "stickiness" parameters are specific to either the single amino acid or pairs of residues and were originally derived from a hydropho-

bicity scale [20]. Other models, based on the lattice-based simulation engines LaSSI [21] and PIMMS [22], classify the amino acids into a reduced number of residue types with distinct "stickiness", ranging from binary categorizations into stickers and spacers [4, 23, 24] to more detailed descriptions and parameterizations [25, 26]. Recently, the accuracy of the "stickiness" parameters has been considerably improved. This has been achieved by leveraging (i) experimental data on single-chain properties, (ii) statistical analyses of protein structures, and (iii) residue-residue free energy profiles calculated from all-atom simulations [26–31]. In particular, we and others have proposed an automated procedure to optimize the "stickiness" parameters so as to maximize the agreement with experimental data small-angle X-ray scattering (SAXS) and paramagnetic relaxation enhancement (PRE) NMR data for a large set of IDPs [26, 28, 29, 32]. To ensure the transferability of the model across sequence space, we employed a Bayesian regularization approach [32, 33]. As the regularization term, we defined the prior knowledge on the "stickiness" parameters in terms of 87 hydrophobicity scales reported in the literature. The resulting M1 parameters capture the relative propensities to phase separate of a wide range of IDP sequences. However, we also observed that a systematic increase in simulation temperature of about 30 °C was needed to quantitatively reproduce the experimental concentration of the dilute phase coexisting with the condensate on an absolute scale. Herein, we refer to this model as the first version of the CALVADOS (Coarse-graining Approach to Liquid-liquid phase separation Via an Automated Data-driven Optimisation Scheme) model (CALVADOS 1).

In this class of coarse-grained models of IDPs, nonionic interactions are modeled via a Lennard-Jones-like potential, which decays to zero only at infinite residue-residue distances. For computational efficiency, the potential is typically calculated up to a cutoff distance, $r_c$, and interactions between particles that are farther apart are ignored. Although this truncation may introduce severe artifacts, in the different implementations of the models, the value of $r_c$ has varied considerably between 1 and 4 nm [20, 29–32, 34, 35]. Here, we systematically investigate the effect of the cutoff of nonionic interactions on single-chain compaction and PS propensity. We find that decreasing the cutoff from 4 to 2 nm results in a small increase in the radius of gyration whereas the PS propensity significantly decreases. We exploit this effect to improve the temperature-dependence of the CALVADOS 1 model by tuning the cutoff of the nonionic potential. Further, we perform a Bayesian optimization of the "stickiness" parameters using a cutoff of 2.4 nm and an augmented training set. We show that the updated model (CALVADOS 2) has improved predictive accuracy.

## 2 Results and Discussion

When applying a cutoff scheme, we neglect the interactions of residues separated by a distance, $r$, larger than the cutoff, $r_c$. For the most strongly interacting residue pair (between two tryptophans), the nonionic potential of the CALVADOS 1 model at $r_c = 2$ nm takes the value of -5 J mol$^{-1}$, that is only a small fraction of the thermal energy (Fig. 1*A*). However, the Lennard-Jones potential falls off slowly whereas the number of interacting partners increases quadratically with increasing $r$. Therefore, in a simulation of a protein-rich phase, decreasing the cutoff from 4 to 2 nm (Fig. 1*A*) can imply ignoring a total interaction energy per protein of several times the thermal energy.

We first look into the effect of the choice of cutoff on the conformational ensembles of isolated proteins. We simulated single IDPs of different sequence length, $N = 71$–441, and average hydropathy, $\langle\lambda\rangle = 0.33$–0.63. The average radii of gyration, $\langle R_g\rangle$, calculated from simulation trajectories are systematically larger when we use $r_c = 2$ nm, compared to the values obtained using $r_c = 4$ nm. CALVADOS 1 was optimized using the longer $r_c$ and estimating the ensemble average $R_g$ values as the root-mean-square $R_g$, $\sqrt{\langle R_g^2\rangle}$. Since $\sqrt{\langle R_g^2\rangle}$ is systematically larger than $\langle R_g\rangle$, decreasing $r_c$ to 2 nm results in a slight improvement of the agreement between the calculated $\langle R_g\rangle$ values and the experimental data (Fig. 1*B*).

To gain further insight into the effect of the cutoff, we performed simulations of single chains of $\alpha$-Synuclein, hnRNPA1 LCD, PNt and Tau 2N4R (Tab. S1) using $r_c = 2$, 2.5, 3 and 4 nm. Irrespective of the sequence, $\langle R_g\rangle$ decreases monotonically with increasing $r_c$. However, the effect on compaction appears to be larger for long sequences and high

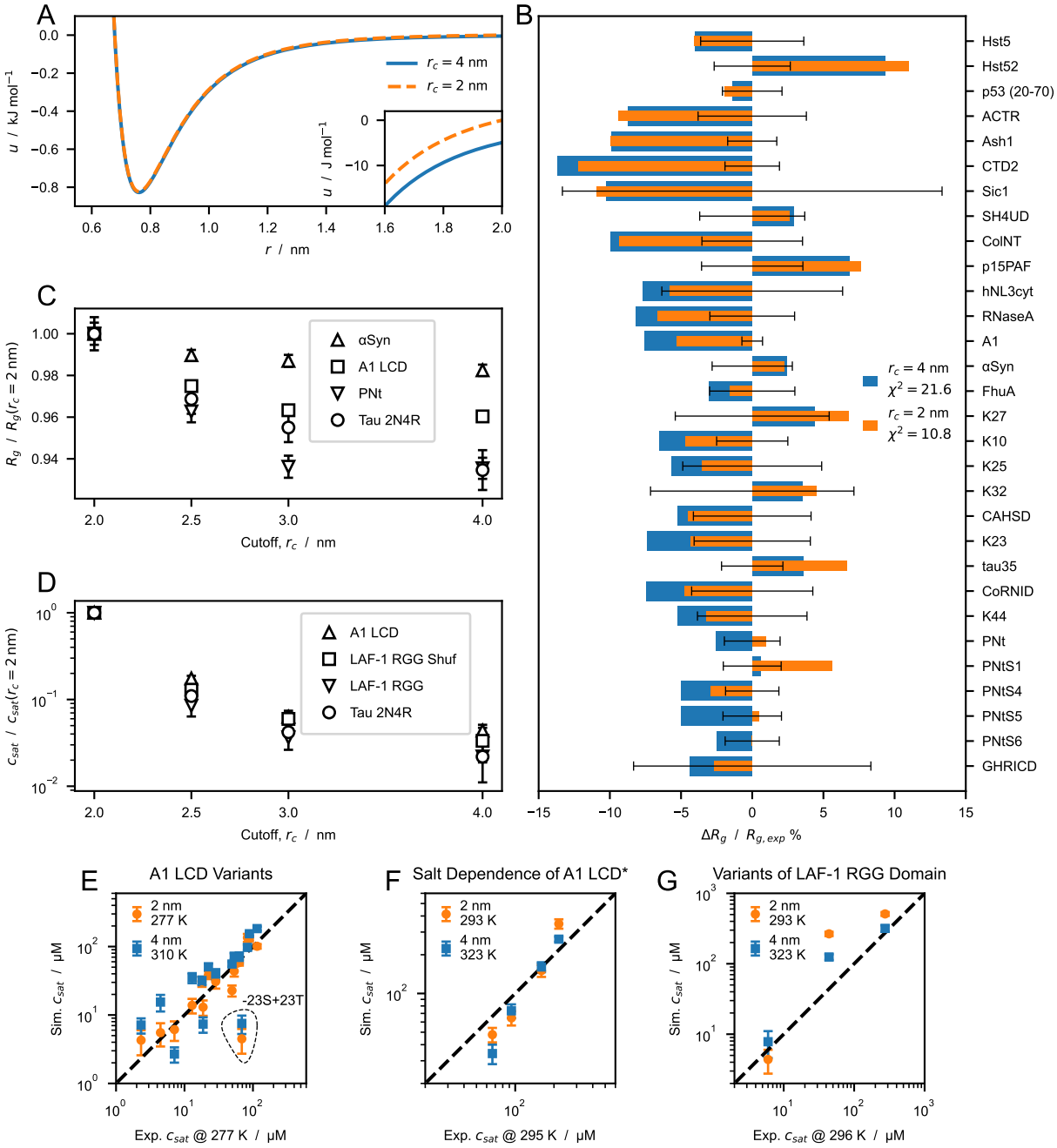


Figure 1: Effect of cutoff size on predictions of radii of gyration, $R_g$, and saturation concentration, $c_{sat}$, from simulations performed using the CALVADOS 1 parameters. (*A*) Nonionic Ashbaugh-Hatch potentials between two W residues with cutoff, $r_c$, of 4 (blue solid line) and 2 nm (orange dashed line). The inset highlights differences between the potentials for $r_s \leq r \leq r_c$. (*B*) Relative difference between experimental and predicted radii of gyration, $\langle R_g \rangle$, for $r_c = 4$ (blue) and 2 nm (orange). $\chi_r^2$ values reported in the legend are calculated for all the sequences in Table S1. (*C*) $\langle R_g \rangle$ of $\alpha$-Synuclein, hnRNPA1 LCD, PNt and human full-length tau (Table S1) from simulations performed with increasing cutoff size, $r_c$, and normalized by the value at $r_c = 2$ nm. (*D*) Saturation concentration, $c_{sat}$, for hnRNPA1 LCD, the randomly shuffled sequence of LAF-1 RGG domain, LAF-1 RGG domain and human full-length tau for increasing values of $r_c$ and normalized by the $c_{sat}$ at $r_c = 2$ nm. (*E–G*) Correlation between $c_{sat}$ from simulations and experiments for (*E*) A1 LCD variants, (*F*) A1 LCD* WT at $[\mathrm{NaCl}] = 0.15$, 0.2, 0.3 and 0.5 M and (*G*) variants of LAF-1 RGG domain (Table S2).

content of hydrophobic residues, both of which result in an increased number of shorter intramolecular distances. For example, upon increasing the $r_c$ from 2 to 4 nm, the $\langle R_g \rangle$ of $\alpha$-Synuclein ($N = 140$, $\langle \lambda \rangle = 0.33$) decreases by 2.3% whereas the effect is more pronounced for hnRNPA1 LCD ($N = 137$, $\langle \lambda \rangle = 0.61$) and tau 2N4R ($N = 441$, $\langle \lambda \rangle = 0.38$), with a decrease in $\langle R_g \rangle$ of 4.0% and 7.7%, respectively.

To investigate the effect of the cutoff distance on PS propensity, we performed direct-coexistence simulations of

100 chains of hnRNPA1 LCD, LAF-1 RGG domain (WT and shuffled sequence with higher charge segregation) and Tau 2N4R (Tab. S2). From the simulation trajectories of the two-phase system at equilibrium, we calculate $c_{sat}$, i.e. the protein concentration in the dilute phase coexisting with the condensate. The higher the $c_{sat}$ value, the lower the propensity of the IDP to undergo PS. As expected from the increased contact density in the condensed phase, the choice of cutoff has a considerably larger impact on $c_{sat}$ than on chain compaction: decreasing $r_c$ from 4 to 2 nm results in an increase in $c_{sat}$ of over one order of magnitude. In contrast to what we observed for the $\langle R_g \rangle$, the decrease in $c_{sat}$ does not show a clear dependence on sequence length and average hydropathy. From the multi-chain trajectories of hnRNPA1 LCD, LAF-1 RGG domain (WT and shuffled sequence) and Tau 2N4R obtained using $r_c = 4$ nm, we estimate that the increase in nonionic energy per protein upon decreasing the cutoff from 4 to 2 nm is 13$\pm$1 kJ mol$^{-1}$ (mean$\pm$standard deviation), respectively (Fig. S1*A*). Assuming that the number of interactions neglected by the shorter cutoff is proportional to the sequence length and the amino acid concentration in the condensate, the small variance in the energy increase across the different IDPs finds explanation in the fact that the simulated systems display similar values of $N^2 \times c_{con}$ (Fig. S1*A*), where $c_{con}$ is the protein concentration in the condensate. The ratio $U(r_c = 2\ \mathrm{nm})/U(r_c = 4\ \mathrm{nm})$ of the nonionic energies for $r_c = 2$ and 4 nm is also largely system independent (Fig. S1*B*). Moreover, decreasing the temperature by $\sim$30 K in the range between 310 and 323 K has a rather small impact on the relative strength of the electrostatic interactions with respect to the thermal energy, due to the decrease in the dielectric constant of water with increasing temperature (Fig. S1*C*). Therefore, we speculate that the effect of decreasing $r_c$ can be compensated by simulating the system at a lower temperature (Fig. S1*B*).

With these considerations in mind, we use the CALVADOS 1 model with $r_c = 2$ nm to run direct-coexistence simulations of IDPs for which $c_{sat}$ has been measured experimentally (Tab. S3), i.e. variants of hnRNPA1 LCD, hnRNPA1 LCD$^*$ at various salt concentrations, and LAF-1 RGG domain variants. As we have shown that decreasing the range of the nonionic interactions disfavours PS, we perform these simulations at the experimental temperatures, which are lower by $\sim$30 K than those required to reproduce the experimental $c_{sat}$ values when the model is simulated with $r_c = 4$ nm (Fig. 1*E*–*G*). The two-fold decrease in $r_c$ enables the model to quantitatively recapitulate the experimental $c_{sat}$ data at the temperature at which the experiments were conducted. Notably, we show this for diverse sequences, across a wide range of ionic strengths, and for variants with different charge patterning and numbers of aromatic and charged residues. These results suggest that the range of interaction of the Lennard-Jones potential may be too large [36]. While the $r^{-6}$ dependence is strictly correct for dispersion interactions between atoms, the nonionic potential of our model incorporates a variety of effective nonbonded interactions between residues, and hence the Lennard-Jones potential is not expected to capture the correct interaction range [31].

Since CALVADOS 1 was developed using $r_c = 4$ nm, we examined whether reoptimizing the model with the shorter cutoff could result in a comparably accurate model. As detailed in the Methods Section, we performed a Bayesian parameter-learning procedure [32] using an improved algorithm, an expanded training set (Tab. S1), and $r_c = 2$ nm. Fig. S2 shows that the new model tends to underestimate the $c_{sat}$ values of the most PS-prone sequences. We hypothesize that during the optimization the reduction of attractive forces due to the shorter cutoff is overcompensated by an overall increase in $\lambda$. We tested this hypothesis by performing the optimization with increasing values of $r_c$, in the range between 2.0 and 2.5 nm, and found that the $c_{sat}$ values predicted from simulations performed with $r_c = 2.0$ nm increased monotonically with the $r_c$ used for the optimization (Fig. S3). Using $r_c = 2.4$ nm for the optimization resulted in a model with improved accuracy compared to CALVADOS 1 (Fig. 2), especially for the PS of LAF-1 RGG domain and the $-$23S+23T variant of A1 LCD. To test the robustness of the approach, the optimization was carried out starting from $\lambda_0 = 0.5$ for all the amino acids (Fig. 2) and from $\lambda_0$ =M1 (Fig. S4 and S5). The difference between the resulting sets of optimal $\lambda$ values (Fig. S5*A*) is lower than 0.08 for all the residues and exceeds 0.05 only for S, T and A. The model obtained

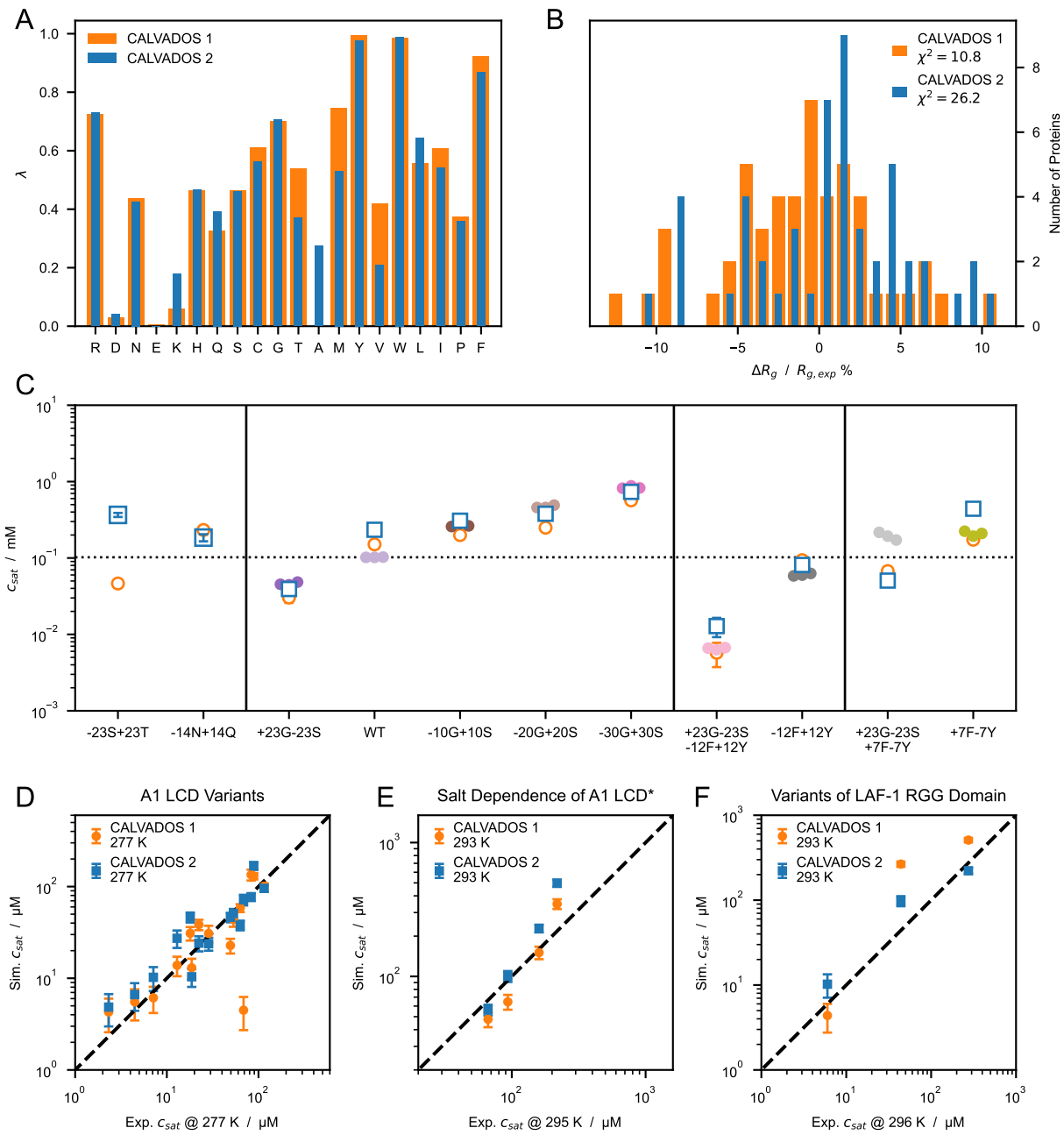


Figure 2: (*A*) Comparison between $\boldsymbol{\lambda}$ sets of CALVADOS 1 (orange) and CALVADOS 2 (blue). (*B*) Distribution of the relative difference between experimental (Table S1) and predicted radii of gyration, $\langle R_g \rangle$, for CALVADOS 1 (orange) and CALVADOS 2 (blue). (*C*) Comparison between saturation concentrations, $c_{sat}$, at 293 K of variants of hnRNPA1 LCD measured by Bremer, Farag, Borcherds et al. [37] (closed circles) and corresponding predictions of CALVADOS 1 (open orange circles) and CALVADOS 2 (open blue squares). (*D*–*F*) Correlation between $c_{sat}$ from simulations and experiments for (*D*) A1 LCD variants, (*E*) A1 LCD* WT at $[\mathrm{NaCl}] = 0.15$, 0.2, 0.3 and 0.5 M and (*F*) variants of LAF-1 RGG domain (Table S2).

starting from $\lambda_0 = 0.5$ is more accurate at predicting PS propensities and will be referred to as CALVADOS 2 hereafter. The $\lambda$ values of CALVADOS 1 and 2 differ mostly for K, T, A, M, and V, whereas the smaller deviations ($|\Delta\lambda| < 0.09$) observed for Q, L, I, and F (Fig. 2*A*) are within the range of reproducibility of the method (Fig. S5*A*). Although CALVADOS 1 was optimized using $r_c = 4$ nm, predictions of single-chain compaction from simulations performed using $r_c = 2$ nm are more accurate for CALVADOS 1 than for CALVADOS 2. This result can be explained by the opposing effects of decreasing the cutoff and calculating $R_g$ values as $\langle R_g \rangle$ insteead of $\sqrt{\langle R_g^2 \rangle}$. In fact, the $\sqrt{\langle R_g^2 \rangle}$ values predicted by CALVADOS 1 are strikingly similar to the $\langle R_g \rangle$ values predicted by CALVADOS 2 (Fig. S6).

The correlation between experimental and predicted $R_g$ values for the 67 proteins of Tab. S1 and S2 is excellent for both CALVADOS 1 and 2 (Fig. S7*A*). On the other hand, CALVADOS 2 is more accurate than CALVADOS 1 at predicting PS propensities, as evidenced by Pearson's correlation coefficients of 0.93 and 0.82, respectively, for the experimental and predicted $c_{sat}$ values of the 26 sequences of Tab. S3 (Fig. S7*B*).

Capturing the interplay between short-range nonionic and long-range ionic interactions is essential for accurately modeling the PS of IDPs [37–39]. Our results show that the decrease in the range of the nonionic potential reported in this work does not significantly perturb the balance between ionic and nonionic forces. In fact, CALVADOS 1 and 2 accurately predict the PS propensities of A1 LCD* at various salt concentrations, as well as the $c_{sat}$ of variants of A1 LCD and LAF-1 RGG domain with different charge patterning (Fig. 1*E–G* and 2*D–F*). Moreover, CALVADOS 1 and 2 recapitulate the effect of salt concentration and charge patterning on the chain compaction of A1 LCD* [40] and p27-C constructs [41], respectively (Fig. S8).

As additional test systems, we considered constructs of the 1–80 N-terminal fragment of yeast Lge1, which have been recently investigated using turbidity measurements [42]. CALVADOS 2 correctly predicts that the WT Lge1$_{1-80}$ construct undergoes PS at the experimental conditions, albeit with a hundred times larger $c_{sat}$ ($\sim$ 60 μM at $c_s = 100$ mM) compared to experiments ($<$ 1 μM). In agreement with experiments, CALVADOS 2 predicts that mutating all the 11 R residues to K increases $c_{sat}$ by over one order of magnitude whereas mutating the 14 Y residues of the 1–80 fragment to A abrogates PS.

## 3 Conclusions

In the context of a previously developed C$\alpha$-based IDP model (CALVADOS), we show that neglecting the long range of attractive Lennard-Jones interactions has a small impact on the compaction of a single chain while strongly disfavouring PS. The effect can be explained by the smaller number of neglected pair interactions for a residues in an isolated chain compared to the dense environment of a condensate. Moreover, we find that the effect of reducing the range of interaction by a factor of two is relatively insensitive to sequence length and composition. Therefore, decreasing the cutoff of the Lennard-Jones potential of the C$\alpha$-based model engenders a similar generic effect on chain compaction and PS as a corresponding increase in temperature. We take advantage of this finding to solve the temperature mismatch of the CALVADOS model. Namely, we decrease the cutoff of the nonionic interactions from 4 to 2 nm and obtain accurate $c_{sat}$ predictions at the experimental conditions, whereas simulations at temperatures higher by 30 °C were required in the original implementation. Finally, we used the shorter cutoff to reoptimize the "stickiness" parameters of the model against experimental data reporting on single-chain compaction. The small expansion of the chain conformations is overcompensated by an overall increase in "stickiness" so that the resulting model tends to underestimate the experimental $c_{sat}$ values. By systematically increasing the cutoff used in the development of the "stickiness" scale, we find that performing the optimization using $r_c = 2.4$ nm results in a model (CALVADOS 2) which yields accurate predictions from simulations run using $r_c = 2$ nm at the experimental conditions. We present CALVADOS 2 as an improvement of our previous model by testing on sets of experimental $R_g$ and $c_{sat}$ data comprising 16 and 36 systems, respectively, which were not used in the parameterization of the model.

## 4 Methods

### 4.1 Molecular Simulations

Molecular dynamics simulations are conducted in the NVT ensemble using the Langevin integrator with a time step of 10 fs and friction coefficient of 0.01 ps$^{-1}$. Non-bonded interactions between residues separated by one bond are excluded from the energy calculations. Functional forms and parameters for bonded and nonbonded interactions are reported

in the Supporting Information. Single chains of $N$ residues are simulated using HOOMD-blue v2.9.3 [43] in a cubic box of side length $0.38 \times (N-1) + 4$ nm under periodic boundary conditions, starting from the fully extended linear conformation. Each chain is simulated in ten replicas for $\sim 6 \times 0.3 \times N^2$ ps if $N > 100$ and for 18 ns otherwise. The initial 100 frames of each replica are discarded, so as to sample 5,000 weakly correlated conformations for each protein (Fig. S10). Direct-coexistence simulations are performed using openMM v7.5 [44] in a cuboidal box of side lengths $[L_x, L_y, L_z] = [25, 25, 300]$, $[17, 17, 300]$ and $[15, 15, 150]$ nm for Tau 2N4R, Ddx4 LCD, and for the remainder of the proteins, respectively. In the starting configuration, 100 chains are aligned along the $z$-axis and with their middle beads placed in the $xy$-plane at random $(x, y)$ positions which are more than 0.7 nm apart. Multi-chain simulations are carried out for at least 2 µs, saving frames every 0.5 ns (Fig. S11, S12, and S13). After discarding the initial 0.6 µs, the slab is centered in the box at each frame as previously described [32] and the equilibrium density profile, $\rho(z)$, is calculated by averaging over the trajectory of the system at equilibrium. The densities of the dilute and protein-rich phases are estimated as the average densities in the regions $|z| < z_{DS} - t/2$ and $|z| > z_{DS} + 6t$ nm, where $z_{DS}$ and $t$ are the position of the dividing surface and the thickness of the interface, respectively. $z_{DS}$ and $t$ are obtained by fitting the semi-profiles in $z > 0$ and $z < 0$ to $\rho(z) = (\rho_a + \rho_b)/2 + (\rho_b - \rho_a)/2 \times \tanh[(|z| - z_{DS})/t]$, where $\rho_a$ and $\rho_b$ are the densities of the protein-rich and dilute phases, respectively. The uncertainty of the density values is estimated as the standard error obtained from the blocking approach [45] implemented in the BLOCKING software github.com/fpesceKU/BLOCKING.

## 4.2 Optimization of the "Stickiness" Scale

The optimization of the "stickiness" parameters, $\boldsymbol{\lambda}$, is carried out to minimize the cost function $\mathcal{L}(\boldsymbol{\lambda}) = \langle \chi^2_{R_g}(\boldsymbol{\lambda}) \rangle + \eta \langle \chi^2_{PRE}(\boldsymbol{\lambda}) \rangle - \theta \ln[P(\boldsymbol{\lambda})]$ using an algorithm which is analogous to the one we previously described [32]. $\chi^2_{R_g}$ and $\chi^2_{PRE}$ quantify the discrepancy between model predictions and experimental data, and are defined as $\chi^2_{R_g} = [(R_g^{exp} - R_g^{calc})/\sigma^{exp}]^2$ and $\chi^2_{PRE} = \frac{1}{N_{labels} N} \sum_j^{N_{labels}} \sum_i^{N_{res}} [(Y_{ij}^{exp} - Y_{ij}^{calc})/\sigma_{ij}^{exp}]^2$, respectively, where $\sigma^{exp}$ is the error on the experimental values, $Y$ is either PRE rates or intensity ratios and $N_{labels}$ is the number of spin-labeled mutants used for the NMR PRE data. In the expression for the cost function, the coefficients are set to $\eta = 0.1$ and $\theta = 0.05$. The prior is the distribution of $\lambda$, $P(\boldsymbol{\lambda})$, derived from a subset of the hydrophobicity scales reported in Table 3 and 4 of Simm et al. [46]. Specifically, only the 70 scales that are unique after min-max normalization (Fig. S14) are used for the calculation of $P(\lambda)$, namely Wimley, BULDG reverse, MANP780101, VHEG790101, JANIN, JANJ790102, WOLR790101, PONP800101–6, Wilson, FAUCH, ENGEL, ROSEM, JACWH, CowanWhittacker, ROSM880101 reverse, ROSM880102 reverse, COWR900101, BLAS910101, CASSI, CIDH920101, CIDH920105, CIDBB, CIDA+, CIDAB, PONG1–3, WILM950101–2, WILM950104, Bishop reverse, NADH010101–7, ZIMJ680101, NOZY710101, JONES, LEVIT, KYTJ820101, SWER830101, SWEET, EISEN, ROSEF, GUYFE, COHEN, NNEIG, MDK0, MDK1, JURD980101, SET1–3, CHOTA, CHOTH, Sweet & Eisenberg, KIDER, ROSEB, Welling reverse, Rao & Argos, GIBRA, and WOLR810101 reverse. $P(\boldsymbol{\lambda})$ is obtained via multivariate kernel density estimation, as implemented in scikit-learn [47], using a Gaussian kernel with bandwidth of 0.05. This prior is 20-dimensional and contains information on the $\lambda$-distribution of the single amino acid (Fig. S15) as well as on the covariance matrix (Fig. S16) inferred from our selection of 70 hydrophobicity scales. Initially, the $\lambda$ values for all the amino acids are set to 0.5, $\lambda_0 = 0.5$, and simulated of all the proteins in the training set. We proceed with the first optimization cycle, wherein, at each $k$-th iteration, the $\lambda$ values of a random selection of five amino acids are nudged by random numbers picked from a normal distribution of standard deviation 0.05 to generate a trial $\boldsymbol{\lambda_k}$ set. For each $i$th frame, we calculate the Boltzmann weight as $w_i = \exp\{-[U(\boldsymbol{r_i}, \boldsymbol{\lambda_k}) - U(\boldsymbol{r_i}, \boldsymbol{\lambda_0})]/k_B T\}$, where $U$ is the nonionic potential. The trial $\boldsymbol{\lambda_k}$ is discarded if the effective fraction of frames, $\phi_{eff} = \exp\left[-\sum_i^{N_{frames}} w_i \log(w_i \times N_{frames})\right]$, is lower than 60%. Otherwise, the acceptance probability follows the Metropolis criterion, $\min\left\{1,\ \exp\left[\frac{\mathcal{L}(\boldsymbol{\lambda}_{k-1}) - \mathcal{L}(\boldsymbol{\lambda}_k)}{\xi_k}\right]\right\}$, where $\xi_k$ is a unitless control parameter.

Each optimization cycle is divided into ten micro-cycles, wherein the control parameter, $\xi$, is initially set to $\xi_0 = 0.1$ and scaled down by 1% at each iteration until $\xi < 10^{-8}$. From the complete optimization cycle, we select the $\boldsymbol{\lambda}$ set yielding the lowest estimate of $\mathcal{L}$. Consecutive optimization cycles are performed from simulations runs carried out with the intermediate optimal $\boldsymbol{\lambda}$ set. To show that the procedure is reproducible and that the final $\boldsymbol{\lambda}$ set is relatively independent of the initial conditions, we performed an additional optimization procedure starting from the M1 model, $\lambda_0$ =M1 [32]. The optimization performed in this work differs from our previous implementation [32] also for the following details: (i) nine additional sequences have been included in the training set (Tab. S1 and S3); (ii) single chains are simulated as detailed in the Molecular Simulations Subsection; (v) the average radius of gyration is calculated as $\langle R_g \rangle$ instead of $\sqrt{\langle R_g^2 \rangle}$.

## Code and Data Availability

Scripts and parameters to perform single-chain and direct-coexistence simulations using CALVADOS 1 and 2 are available at github.com/KULL-Centre/CALVADOS. All code and data to reproduce the results presented in this work are available at github.com/KULL-Centre/papers/tree/main/2022/CG-cutoffs-Tesei-et-al.

## Acknowledgments

We thank Anna Ida Trolle for her help in setting up the protocol for single-chain simulations. We thank Rosana Collepardo, Jerelle A. Joseph, and Aleks Reinhardt for useful discussions that led us to explore the effects of cutoffs. We 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. 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.

## References


[1] J. Wang, J.-M. Choi, A. S. Holehouse, H. O. Lee, X. Zhang, M. Jahnel, S. Maharana, R. Lemaitre, A. Pozniakovsky, D. Drechsel, I. Poser, R. V. Pappu, S. Alberti, A. A. Hyman, "A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins", *Cell* **2018**, *174*, 688–699.e16, DOI 10.1016/j.cell.2018.06.006.

[2] Z. Monahan, V. H. Ryan, A. M. Janke, K. A. Burke, S. N. Rhoads, G. H. Zerze, R. O'Meally, G. L. Dignon, A. E. Conicella, W. Zheng, R. B. Best, R. N. Cole, J. Mittal, F. Shewmaker, N. L. Fawzi, "Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity", *The EMBO Journal* **2017**, *36*, 2951–2967, DOI 10.15252/embj.201696394.

[3] V. H. Ryan, G. L. Dignon, G. H. Zerze, C. V. Chabata, R. Silva, A. E. Conicella, J. Amaya, K. A. Burke, J. Mittal, N. L. Fawzi, "Mechanistic View of hnRNPA2 Low-Complexity Domain Structure, Interactions, and Phase Separation Altered by Mutation and Arginine Methylation", *Molecular Cell* **2018**, *69*, 465–479.e7, DOI 10.1016/j.molcel.2017.12.022.

[4] E. W. Martin, A. S. Holehouse, I. Peran, M. Farag, J. J. Incicco, A. Bremer, C. R. Grace, A. Soranno, R. V. Pappu, T. Mittag, "Valence and patterning of aromatic residues determine the phase behavior of prion-like domains", *Science* **2020**, *367*, 694–699, DOI 10.1126/science.aaw8653.

[5] T. Mittag, R. V. Pappu, "A conceptual framework for understanding phase separation and addressing open questions and challenges", *Molecular Cell* **2022**, *82*, 2201–2214, DOI 10.1016/j.molcel.2022.05.018.

[6] A. Patel, H. O. Lee, L. Jawerth, S. Maharana, M. Jahnel, M. Y. Hein, S. Stoynov, J. Mahamid, S. Saha, T. M. Franzmann, A. Pozniakovski, I. Poser, N. Maghelli, L. A. Royer, M. Weigert, E. W. Myers, S. Grill, D. Drechsel, A. A. Hyman, S. Alberti, "A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation", *Cell* **2015**, *162*, 1066–1077, DOI 10.1016/j.cell.2015.07.047.

[7] T. Murakami, S. Qamar, J. Q. Lin, G. S. K. Schierle, E. Rees, A. Miyashita, A. R. Costa, R. B. Dodd, F. T. Chan, C. H. Michel, D. Kronenberg-Versteeg, Y. Li, S.-P. Yang, Y. Wakutani, W. Meadows, R. R. Ferry, L. Dong, G. G. Tartaglia, G. Favrin, W.-L. Lin, D. W. Dickson, M. Zhen, D. Ron, G. Schmitt-Ulms, P. E. Fraser, N. A. Shneider, C. Holt, M. Vendruscolo, C. F. Kaminski, P. S. George-Hyslop, "ALS/FTD Mutation-Induced Phase Transition of FUS Liquid Droplets and Reversible Hydrogels into Irreversible Hydrogels Impairs RNP Granule Function", *Neuron* **2015**, *88*, 678–690, DOI 10.1016/j.neuron.2015.10.030.

[8] N. M. Kanaan, C. Hamel, T. Grabinski, B. Combs, "Liquid-liquid phase separation induces pathogenic tau conformations in vitro", *Nature Communications* **2020**, *11*, 2809, DOI 10.1038/s41467-020-16580-3.

[9] S. Wegmann, B. Eftekharzadeh, K. Tepper, K. M. Zoltowska, R. E. Bennett, S. Dujardin, P. R. Laskowski, D. MacKenzie, T. Kamath, C. Commins, C. Vanderburg, A. D. Roe, Z. Fan, A. M. Molliex, A. Hernandez-Vega, D. Muller, A. A. Hyman, E. Mandelkow, J. P. Taylor, B. T. Hyman, "Tau protein liquid–liquid phase separation can initiate tau aggregation", *The EMBO Journal* **2018**, *37*, e98049, DOI 10.15252/embj.201798049.

[10] T. R. Peskett, F. Rau, J. O'Driscoll, R. Patani, A. R. Lowe, H. R. Saibil, "A Liquid to Solid Phase Transition Underlying Pathological Huntingtin Exon1 Aggregation", *Molecular Cell* **2018**, *70*, 588–601.e6, DOI 10.1016/j.molcel.2018.04.007.

[11] S. Ray, N. Singh, R. Kumar, K. Patel, S. Pandey, D. Datta, J. Mahato, R. Panigrahi, A. Navalkar, S. Mehra, L. Gadhe, D. Chatterjee, A. S. Sawner, S. Maiti, S. Bhatia, J. A. Gerez, A. Chowdhury, A. Kumar, R. Padinhateeri, R. Riek, G. Krishnamoorthy, S. K. Maji, "$\alpha$-Synuclein aggregation nucleates through liquid–liquid phase separation", *Nature Chemistry* **2020**, *12*, 705–716, DOI 10.1038/s41557-020-0465-9.

[12] M. C. Hardenberg, T. Sinnige, S. Casford, S. Dada, C. Poudel, E. A. Robinson, M. Fuxreiter, C. Kaminksi, G. S. Kaminski-Schierle, E. A. A. Nollen, C. M. Dobson, M. Vendruscolo, "Observation of an $\alpha$-synuclein liquid droplet state and its maturation into Lewy body-like assemblies", *Journal of Molecular Cell Biology* **2021**, DOI 10.1093/jmcb/mjaa075.

[13] J. Wen, L. Hong, G. Krainer, Q.-Q. Yao, T. P. J. Knowles, S. Wu, S. Perrett, "Conformational Expansion of Tau in Condensates Promotes Irreversible Aggregation", *Journal of the American Chemical Society* **2021**, *143*, 13056–13064, DOI 10.1021/jacs.1c03078.

[14] S. T. Dada, M. C. Hardenberg, L. K. Mrugalla, M. O. McKeon, E. Klimont, T. C. T. Michaels, M. Vendruscolo, "Spontaneous nucleation and fast aggregate-dependent proliferation of $\alpha$-synuclein aggregates within liquid condensates at physiological pH", *bioRxiv* **2021**, DOI 10.1101/2021.09.26.461836.

[15] B. Wang, L. Zhang, T. Dai, Z. Qin, H. Lu, L. Zhang, F. Zhou, "Liquid-liquid phase separation in human health and diseases", *Signal Transduction and Targeted Therapy* **2021**, *6*, DOI 10.1038/s41392-021-00678-1.

[16] J. Lu, J. Qian, Z. Xu, S. Yin, L. Zhou, S. Zheng, W. Zhang, "Emerging Roles of Liquid-Liquid Phase Separation in Cancer: From Protein Aggregation to Immune-Associated Signaling", *Frontiers in Cell and Developmental Biology* **2021**, *9*, DOI 10.3389/fcell.2021.631486.

[17] J. H. Ahn, E. S. Davis, T. A. Daugird, S. Zhao, I. Y. Quiroga, H. Uryu, J. Li, A. J. Storey, Y.-H. Tsai, D. P. Keeley, S. G. Mackintosh, R. D. Edmondson, S. D. Byrum, L. Cai, A. J. Tackett, D. Zheng, W. R. Legant, D. H. Phanstiel, G. G. Wang, "Phase separation drives aberrant chromatin looping and cancer development", *Nature* **2021**, *595*, 591–595, DOI 10.1038/s41586-021-03662-5.

[18] S. F. Banani, L. K. Afeyan, S. W. Hawken, J. E. Henninger, A. Dall'Agnese, V. E. Clark, J. M. Platt, O. Oksuz, N. M. Hannett, I. Sagi, T. I. Lee, R. A. Young, "Genetic variation associated with condensate dysregulation in disease", *Developmental Cell* **2022**, DOI https://doi.org/10.1016/j.devcel.2022.06.010.

[19] N. L. Fawzi, S. H. Parekh, J. Mittal, "Biophysical studies of phase separation integrating experimental and computational methods", *Current Opinion in Structural Biology* **2021**, *70*, 78–86, DOI 10.1016/j.sbi.2021.04.004.

[20] 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 Computational Biology* **2018**, *14*, (Ed.: Y. Ofran), e1005941, DOI 10.1371/journal.pcbi.1005941.

[21] J.-M. Choi, F. Dar, R. V. Pappu, "LASSI: A lattice model for simulating phase transitions of multivalent proteins", *PLOS Computational Biology* **2019**, *15*, (Ed.: O. Keskin), e1007028, DOI 10.1371/journal.pcbi.1007028.

[22] A. S. Holehouse, R. V. Pappu, *PIMMS (0.24 pre-beta)*, version 0.24, **2019**, DOI 10.5281/zenodo.3588456.

[23] A. S. Holehouse, G. M. Ginell, D. Griffith, E. Böke, "Clustering of Aromatic Residues in Prion-like Domains Can Tune the Formation, State, and Organization of Biomolecular Condensates", *Biochemistry* **2021**, *60*, 3566–3581, DOI 10.1021/acs.biochem.1c00465.

[24] M. Kar, F. Dar, T. J. Welsh, L. T. Vogel, R. Kühnemuth, A. Majumdar, G. Krainer, T. M. Franzmann, S. Alberti, C. A. M. Seidel, T. P. J. Knowles, A. A. Hyman, R. V. Pappu, "Phase-separating RNA-binding proteins form heterogeneous distributions of clusters in subsaturated solutions", *Proceedings of the National Academy of Sciences* **2022**, *119*, DOI 10.1073/pnas.2202222119.

[25] A. Bremer, M. Farag, W. M. Borcherds, I. Peran, E. W. Martin, R. V. Pappu, T. Mittag, "Deciphering how naturally occurring sequence features impact the phase behaviors of disordered prion-like domains", *bioRxiv* **2021**, DOI 10.1101/2021.01.01.425046.

[26] M. Farag, S. R. Cohen, W. M. Borcherds, A. Bremer, T. Mittag, R. V. Pappu, "Condensates of disordered proteins have small-world network structures and interfaces defined by expanded conformations", *bioRxiv* **2022**, DOI 10.1101/2022.05.21.492916.

[27] S. Das, Y.-H. Lin, R. M. Vernon, J. D. Forman-Kay, H. S. Chan, "Comparative roles of charge, $\pi$, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins", *Proceedings of the National Academy of Sciences* **2020**, *117*, 28795–28805, DOI 10.1073/pnas.2008122117.

[28] A. P. Latham, B. Zhang, "Consistent Force Field Captures Homologue-Resolved HP1 Phase Separation", *Journal of Chemical Theory and Computation* **2021**, *17*, 3134–3144, DOI 10.1021/acs.jctc.0c01220.

[29] T. Dannenhoffer-Lafage, R. B. Best, "A Data-Driven Hydrophobicity Scale for Predicting Liquid–Liquid Phase Separation of Proteins", *The Journal of Physical Chemistry B* **2021**, *125*, 4046–4056, DOI 10.1021/acs.jpcb.0c11479.

[30] R. M. Regy, J. Thompson, Y. C. Kim, J. Mittal, "Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins", *Protein Science* **2021**, *n/a*, DOI 10.1002/pro.4094.

[31] J. A. Joseph, A. Reinhardt, A. Aguirre, P. Y. Chew, K. O. Russell, J. R. Espinosa, A. Garaizar, R. Collepardo-Guevara, "Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy", *Nature Computational Science* **2021**, *1*, 732–743, DOI 10.1038/s43588-021-00155-3.

[32] G. Tesei, T. K. Schulze, R. Crehuet, K. Lindorff-Larsen, "Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties", *Proceedings of the National Academy of Sciences* **2021**, *118*, DOI 10.1073/pnas.2111696118.

[33] A. B. Norgaard, J. Ferkinghoff-Borg, K. Lindorff-Larsen, "Experimental Parameterization of an Energy Function for the Simulation of Unfolded Proteins", *Biophysical Journal* **2008**, *94*, 182–192, DOI 10.1529/biophysj.107.108241.

[34] A. R. Tejedor, A. Garaizar, J. Ramírez, J. R. Espinosa, "'RNA modulation of transport properties and stability in phase-separated condensates", *Biophysical Journal* **2021**, *120*, 5169–5186, DOI https://doi.org/10.1016/j.bpj.2021.11.003.

[35] S. Das, A. N. Amin, Y.-H. Lin, H. S. Chan, "Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters", *Phys. Chem. Chem. Phys.* **2018**, *20*, 28558–28574, DOI 10.1039/C8CP05095C.

[36] X. Wang, S. Ramírez-Hinestrosa, J. Dobnikar, D. Frenkel, "The Lennard-Jones potential: when (not) to use it", *Physical Chemistry Chemical Physics* **2020**, *22*, 10624–10633, DOI 10.1039/c9cp05445f.

[37] A. Bremer, M. Farag, W. M. Borcherds, I. Peran, E. W. Martin, R. V. Pappu, T. Mittag, "Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains", *Nature Chemistry* **2021**, DOI 10.1038/s41557-021-00840-w.

[38] I. Alshareedah, T. Kaur, J. Ngo, H. Seppala, L.-A. D. Kounatse, W. Wang, M. M. Moosa, P. R. Banerjee, "Interplay between Short-Range Attraction and Long-Range Repulsion Controls Reentrant Liquid Condensation of Ribonucleoprotein–RNA Complexes", *Journal of the American Chemical Society* **2019**, *141*, 14593–14602, DOI 10.1021/jacs.9b03689.

[39] M. K. Hazra, Y. Levy, "Biophysics of Phase Separation of Disordered Proteins Is Governed by Balance between Short- And Long-Range Interactions", *The Journal of Physical Chemistry B* **2021**, *125*, 2202–2211, DOI 10.1021/acs.jpcb.0c09975.

[40] E. W. Martin, F. E. Thomasen, N. M. Milkovic, M. J. Cuneo, C. R. Grace, A. Nourse, K. Lindorff-Larsen, T. Mittag, "Interplay of folded domains and the disordered low-complexity domain in mediating hnRNPA1 phase separation", *Nucleic Acids Research* **2021**, *49*, 2931–2945, DOI 10.1093/nar/gkab063.

[41] R. K. Das, Y. Huang, A. H. Phillips, R. W. Kriwacki, R. V. Pappu, "Cryptic sequence features within the disordered protein $p27^{Kip1}$ regulate cell cycle signaling", *Proceedings of the National Academy of Sciences* **2016**, *113*, 5616–5621, DOI 10.1073/pnas.1516277113.

[42] A. A. Polyansky, L. D. Gallego, R. G. Efremov, A. Köhler, B. Zagrovic, "Protein compactness and interaction valency define the architecture of a biomolecular condensate across scales", *bioRxiv* **2022**, DOI 10.1101/2022.02.18.481017.

[43] J. A. Anderson, J. Glaser, S. C. Glotzer, "HOOMD-blue: A Python package for high-performance molecular dynamics and hard particle Monte Carlo simulations", *Computational Materials Science* **2020**, *173*, 109363, DOI 10.1016/j.commatsci.2019.109363.

[44] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, V. S. Pande, "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics", *PLOS Computational Biology* **2017**, *13*, (Ed.: R. Gentleman), e1005659, DOI 10.1371/journal.pcbi.1005659.

[45] H. Flyvbjerg, H. G. Petersen, "Error estimates on averages of correlated data", *The Journal of Chemical Physics* **1989**, *91*, 461–466, DOI 10.1063/1.457480.

[46] S. Simm, J. Einloft, O. Mirus, E. Schleiff, "50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification", *Biological Research* **2016**, *49*, DOI 10.1186/s40659-016-0092-5.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.