Received: 1 February 2024 | Revised: 12 July 2024 | Accepted: 23 August 2024

DOI: 10.1002/pro.5172

METHODS AND APPLICATIONS

A coarse-grained model for disordered and multi-domain proteins

Fan Cao | **Sören von Bülow** | **Giulio Tesei** | **Kresten Lindorff-Larsen**

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

Correspondence
Kresten Lindorff-Larsen, Structural Biology and NMR Laboratory & the Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark.
Email: lindorff@bio.ku.dk

Funding information
European Molecular Biology Organization; China Scholarship Council; Novo Nordisk Fonden

Review Editor: Lynn Kamerlin

Abstract

Many proteins contain more than one folded domain, and such modular multi-domain proteins help expand the functional repertoire of proteins. Because of their larger size and often substantial dynamics, it may be difficult to characterize the conformational ensembles of multi-domain proteins by simulations. Here, we present a coarse-grained model for multi-domain proteins that is both fast and provides an accurate description of the global conformational properties in solution. We show that the accuracy of a one-bead-per-residue coarse-grained model depends on how the interaction sites in the folded domains are represented. Specifically, we find excessive domain–domain interactions if the interaction sites are located at the position of the C_α atoms. We also show that if the interaction sites are located at the center of mass of the residue, we obtain good agreement between simulations and experiments across a wide range of proteins. We then optimize our previously described CALVADOS model using this center-of-mass representation, and validate the resulting model using independent data. Finally, we use our revised model to simulate phase separation of both disordered and multi-domain proteins, and to examine how the stability of folded domains may differ between the dilute and dense phases. Our results provide a starting point for understanding interactions between folded and disordered regions in proteins, and how these regions affect the propensity of proteins to self-associate and undergo phase separation.

KEYWORDS

coarse graining, condensates, molecular dynamics, multi-domain proteins, protein dynamics

1 | INTRODUCTION

Multi-domain proteins (MDPs) consist of more than one folded domain that are often connected by linkers or longer intrinsically disordered regions (IDRs), and make up a large fraction (around 50%) of the proteomes in eukaryotic and prokaryotic organisms (Han et al., 2007; Van Der Lee et al., 2014). Like intrinsically disordered proteins (IDPs), MDPs can display large-amplitude motions that may play prominent roles in biomolecular

functions like signaling, catalysis and regulation (Bondos et al., 2021; Delaforge et al., 2016; Mackereth & Sattler, 2012; Van Der Lee et al., 2014).

The biological functions of MDPs depend both on the properties of the folded domains and the disordered regions, and so characterizing the conformational ensembles can be key to understanding how these proteins function. In many cases, the folded and disordered regions are studied separately, but the folded domains might affect the conformational properties of the disordered regions (Mittal et al., 2018; Taneja & Holehouse, 2021) and the disordered regions may also affect the properties of the folded domains (Yu & Sukenik, 2023). For example, there is a complex interplay between the folded and disordered regions in the RNA-binding protein hnRNPA1, that affects its conformational ensemble in solution and its propensity to undergo phase separation (Martin, Thomasen, et al., 2021). However, describing the conformational ensembles of MDPs in solution generally requires a combination of biophysical experiments and molecular dynamics (MD) simulations (Thomasen & Lindorff-Larsen, 2022).

All-atom MD simulations have been used to generate conformational ensembles of IDPs and MDPs and to study intra- and inter-domain interactions (Sekiyama et al., 2022; Zheng et al., 2020). Such simulations, however, are often limited by the large system sizes and long time scales which limit efficient sampling of these dynamic proteins. Coarse-grained (CG) models may increase the sampling efficiency by reducing the number of particles in the simulation systems (Bereau & Deserno, 2009; Gopal et al., 2010; Monticelli et al., 2008; Neri et al., 2005). The accuracy, transferability, and efficiency of such models, however, depend on the degree of coarse-graining and the parameterization strategy (Heo & Feig, 2024). One commonly used model is the Martini force field, which uses a four-to-one mapping scheme with explicit solvent (Souza et al., 2021). Different versions of Martini have been modified to produce improved ensembles of IDPs and MDPs (Benayad et al., 2020; Thomasen et al., 2022, 2024). For IDPs, there has in the last years been extensive work using even coarser models where each amino acid residue is represented by a single bead. The interaction sites are generally located at the C_α positions and separated by bonds that are 0.38 nm long, and we therefore here term these C_α models. Several related models rely on a similar functional form to the HPS model introduced by Dignon et al. (2018) and may include bonded terms, an Ashbaugh-Hatch potential (Ashbaugh & Hatch, 2008) for shorter-range interactions and a Debye-Hückel electrostatic screening potential. Such models have for example been used to study the conformational ensembles and interactions within and between IDPs (Dannenhoffer-Lafage & Best, 2021; Dignon et al., 2018; Joseph et al., 2021; Regy et al., 2021; Tesei & Lindorff-Larsen, 2023; Valdes-Garcia et al., 2023; Wessén et al., 2022).

Coarse-grained models developed for IDPs do not represent the stability of folded proteins well, because the finely balanced energy contributions from individual backbone and side-chain interactions are not captured by the reduced representation. As a consequence, additional (often harmonic) restraints are applied to maintain the folded configurations in folded proteins and MDPs (Borges-Araújo et al., 2023; Souza et al., 2021). Even when applying such restraints to models developed for IDPs, extra attention needs to be paid to interactions related to folded domains since it is still unclear whether the models are fully transferable to MDPs. In particular, C_α-based one-bead-per-residue mappings do not account for the specific orientations of side chains in folded proteins (Kolinski & Skolnick, 1998). For example, hydrophobic residues, whose side chains are "tucked away" in the hydrophobic core of the protein, may be exposed at the surface of the protein in a C_α based representation. One approach to help overcome this problem is to use a different or scaled set of force field parameters for interactions that involve folded regions (Dignon et al., 2018; Kim & Hummer, 2008; Krainer et al., 2021). Another possible solution is the introduction of more terms in the energy function to better describe long-range interactions (Li et al., 2012; Tan et al., 2023) or to introduce anisotropic interactions (Sieradzan et al., 2022).

As an alternative, other coarse-grained models represent a residue by more than one bead to represent backbone side chain orientations and interactions (Hyeon et al., 2006; Maity et al., 2022; Mugnai et al., 2023; Pappu et al., 1996; Sieradzan et al., 2022; Yamada et al., 2023; Zhang et al., 2022; Zhang et al., 2023). In some of these models, one bead is placed at C_α and the other one is at the center of mass (COM) of side chain atoms. In this way, side chain interactions can be explicitly taken into account, improving the simulated dynamical behavior of folded protein simulations and model transferability. In previous studies, this strategy has been used to study conformational ensembles of IDPs or unfolding pathways of proteins (Hyeon et al., 2006; Mugnai et al., 2023). While effective, using multi-bead-per-residue models increases the time to sample configurations in simulations, and requires the determination of a larger number of force field parameters.

We have previously developed and applied an automated procedure to optimize the "stickiness" parameters (λ) in a one-bead-per-residue model by improving the agreement with experimental small-angle X-ray scattering (SAXS) and paramagnetic relaxation enhancement

(PRE) nuclear magnetic resonance (NMR) data for a large set of IDPs (Norgaard et al., 2008; Tesei & Lindorff-Larsen, 2023; Tesei, Schulze, et al., 2021). The most recent CALVADOS (Coarse-graining Approach to Liquid–liquid phase separation Via an Automated Data-driven Optimization Scheme) model (CALVADOS 2) was further tuned to describe phase behavior of multi-chain conformational ensembles of IDPs from simulations by reducing the range of non-ionic interactions (Tesei & Lindorff-Larsen, 2023).

Here, we explore the use of the CALVADOS model for simulations of MDPs. We find that when the CALVADOS 2 parameters are used in simulations of MDPs with interaction sites at the C_α positions, the resulting structures in some cases show excessive interactions between the folded domains, leading to compact ensembles that do not agree with SAXS data. To remedy this problem, we describe a strategy where interaction sites in folded regions are located at the COM of the residue, and show that simulations with this model result in substantially improved agreement with experiments. We optimize the parameters in CALVADOS using the COM representation to derive a refined set of CALVADOS parameters (CALVADOS 3). When we combine the COM representation of folded domains with harmonic restraints between residues in the folded domains and the CALVADOS 3 parameters we obtain good agreement with experimental data on single-chain properties of MDPs and IDPs. Finally, we show how this model may be used to study the interactions between folded and disordered regions in proteins that undergo phase separation, and how the stability of folded domains might change during phase separation.

2 | RESULTS

2.1 | A modified representation improves accuracy for multi-domain proteins

We first evaluated the accuracy of the original CALVADOS 2 model for simulations of MDPs. We therefore used the CALVADOS 2 parameters (Tesei & Lindorff-Larsen, 2023) and a C_α representation to run simulations of 56 IDPs and 14 MDPs (Tables S1, S2, and S3). In all systems, the interaction sites are located at the C_α positions in both folded and disordered regions; for the MDPs, we applied an additional elastic network model to keep domains intact during simulations (Figure 1a, see Section 4). We term this combination of the force field parameters (CALVADOS 2) and the C_α representation of the interaction sites in the folded domains as CALVADOS2$_{C_\alpha}$. As expected and reported previously (Tesei & Lindorff-Larsen, 2023), we found that simulations of IDPs with CALVADOS2$_{C_\alpha}$ resulted in good agreement between experimental and calculated values of R_g (Figure 1b). In contrast, we found more substantial differences between experimental and calculated values of R_g for several MDPs (Figure 1b). In particular, we found that the R_g was underestimated for several MDPs including a series of two fluorescent proteins connected by Gly-Ser

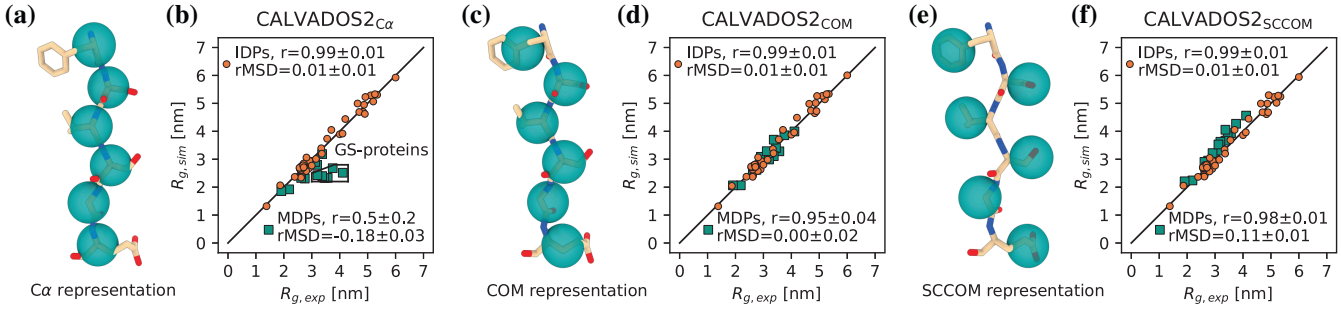

FIGURE 1 Simulations of MDPs and IDPs using a C_α representation, COM representation or side-chain center-of-mass (SCCOM) representation. Location of the interaction sites in a β-sheet when using (a) a C_α representation, (c) a COM representation, and (e) a SCCOM representation. Comparison between simulated and experimental R_g values for IDPs (orange) and MDPs (green) using (b) the CALVADOS2$_{C_\alpha}$ model (CALVADOS 2 parameters and a C_α representation for both folded and disordered regions), (d) the CALVADOS2$_{COM}$ model (CALVADOS 2 parameters and a COM representation for the interaction sites in the folded regions), and (f) the CALVADOS2$_{SCCOM}$ model (CALVADOS 2 parameters and a SCCOM representation for the interaction sites in the folded regions). The region labeled "GS-proteins" in panel B contains a number of proteins consisting of pairs of β-sheet-rich fluorescent protein connected by glycine-serine linkers (Moses et al., 2024). Pearson correlation coefficients (r) and relative mean signed deviation $\mathrm{rMSD} = \langle (R_{g,\mathrm{sim}} - R_{g,\mathrm{exp}})/R_{g,\mathrm{exp}} \rangle$ are reported in the legend, and errors represent standard errors of the mean calculated using bootstrapping. A negative rMSD value indicates that the calculated radii of gyration are systematically lower than the experimental values. The black diagonal lines in panel B, D and F indicate $y = x$.

linkers of different lengths (here termed GS-proteins; Moses et al., 2024). This observation was confirmed by calculations of the relative mean signed deviation, rMSD, between experimental and calculated values of R_g that shows that these are on average underestimated by 18% in the MDPs (Figure 1b).

As a first attempt at creating a model for both IDPs and MDPs, we used our previously described protocol (Norgaard et al., 2008; Tesei, Schulze, et al., 2021) to optimize the λ stickiness parameters of the CALVADOS model targeting simultaneously SAXS and NMR data on 56 IDPs and 14 MDPs. The resulting λ values were generally smaller than those in CALVADOS 2 (Figure S1a) in line with the finding that the MDPs were too compact using CALVADOS 2. Nevertheless, it was also clear that this new parameter set made the agreement worse for disordered proteins (Figure S1b–e) and did not result in a satisfactory model to describe both IDPs and MDPs.

We instead hypothesized that the compaction of several MDPs was a result of placing the interaction sites at the C_α positions in the folded domains. In particular for β-sheet-containing proteins, this geometry would mean that residues whose side chains are buried inside the folded domain are represented by interaction sites located closer to the protein surface (Figure 1a); thus buried hydrophobic residues might appear as solvent exposed. We therefore constructed a new model where the interaction sites within folded regions were placed at the COM of the residue (Figure 1c) and constrained by harmonic restraints; when used with the CALVADOS 2 parameters, we term this model CALVADOS2$_{COM}$. We stress that only the bead locations in the folded domains differ between the CALVADOS2$_{C_\alpha}$ and CALVADOS2$_{COM}$ models; residues in disordered regions are represented by one bead centered on the C_α positions in both models. In the absence of folded domains, CALVADOS2$_{COM}$ and CALVADOS2$_{C_\alpha}$ are thus identical and simulations with the two models gave comparable results (Figure 1b,d). In contrast, simulations of the MDPs with CALVADOS2$_{COM}$ were in substantially better agreement with experiments than simulations with CALVADOS2$_{C_\alpha}$ as evidenced, for example, by an increase in Pearson correlation coefficient from 0.5 to 0.95 and an increase in rMSD from −18% to 0% (Figure 1b,d). In addition to the COM representation, we also examined whether a side-chain center-of-mass (SCCOM) representation, shifting bead positions of buried residues further away from the surface, could yield even more accurate R_g predictions than the COM representation (Figure 1e). We performed single chain simulations with the CALVADOS 2 parameters and the SCCOM representation (CALVADOS2$_{SCCOM}$) and found that CALVADOS2$_{SCCOM}$ on average resulted in an overestimation of the R_g of MDPs of 11% (Figure 1d,f). As an alternative solution to decrease the too strong interactions between folded domains, it has previously been suggested to scale down interactions between pairs of folded domains (by a factor of 0.7) and between folded domains and disordered regions (by a factor of $0.84 = \sqrt{0.7}$) (Krainer et al., 2021). While applying this rescaling to CALVADOS 2 (termed CALVADOS2$_{C_\alpha}$ 70%) led to improved agreement with experiments, the improvement was smaller than when using the COM representation, and the simulations had a remaining bias towards underestimating the radii of gyration (Figure S2). Therefore, we proceeded by using the COM representation in this study.

To examine in more detail why the CALVADOS2$_{C_\alpha}$ model resulted in more compact conformations of MDPs than CALVADOS2$_{COM}$, we calculated the time-averaged non-ionic (Ashbaugh-Hatch) interaction energies between residues of different folded domains. For this analysis we selected GS0, a construct with two fluorescent proteins separated by a 29-residue-long linker (Moses et al., 2024), since the R_g value of GS0 deviates substantially from experiments in simulations with CALVADOS2$_{C_\alpha}$ (Figure 1b). In the energy maps, we see evidence of substantial inter-domain interactions between residues 140–230 of one fluorescent protein and residue 340–440 of the other (Figure 2a). In contrast, these domain–domain interactions are not observed when simulating with COM representation (Figure 2b). The comparison of the two energy maps thus supports the hypothesis that the too compact conformations of MDPs in simulations with CALVADOS2$_{C_\alpha}$ result from inter-domain attractions that are decreased in the COM representation (Figure 2c).

2.2 | Optimizing CALVADOS using a center-of-mass representation

Having shown that the COM representation gave an improved description of MDPs while preserving the accuracy when simulating IDPs, we proceeded to optimize the CALVADOS model further. We used our iterative Bayesian optimization scheme (Norgaard et al., 2008; Tesei, Schulze, et al., 2021) to optimize the λ stickiness parameters of the CALVADOS model targeting simultaneously SAXS and NMR data on 56 IDPs and 14 MDPs (Tables S1, S2, and S3). In these simulations we used the COM representation of the folded domains and we thus term the final model CALVADOS3$_{COM}$ to represent both the force field and the COM representation of the folded regions. The resulting λ values in CALVADOS3$_{COM}$ are similar to those in CALVADOS 2 (Figures 3 and S3). We found that simulations of IDPs with CALVADOS3$_{COM}$

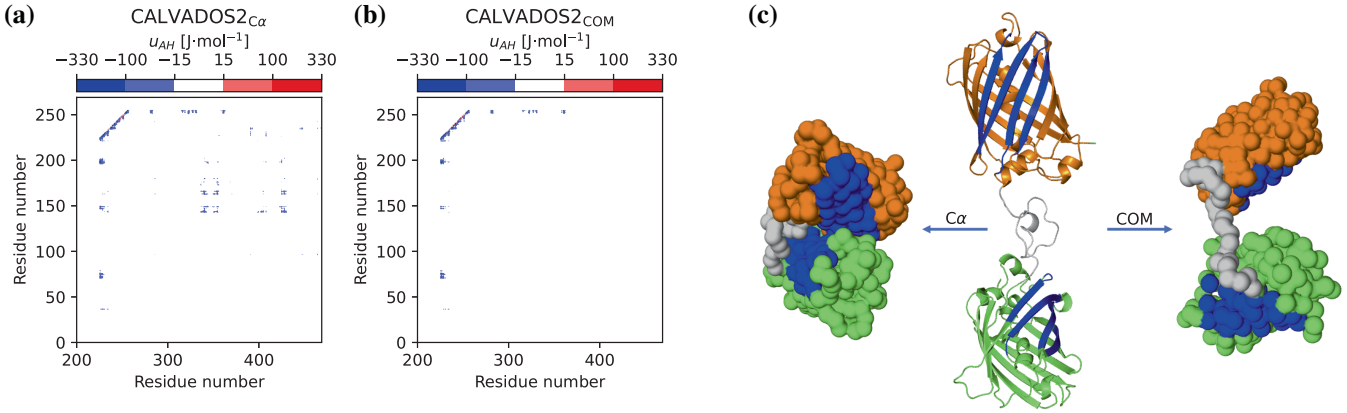

FIGURE 2 Energy calculations reveal substantial inter-domain interactions. We calculated interaction energy maps (of the Ashbaugh-Hatch term in the force field) from simulations using (a) the CALVADOS2$_{C_\alpha}$ model and (b) the CALVADOS2$_{COM}$ model. We show only a subset of the map representing interactions between the first (residues 1–226 on the *y*-axis) and second (residues 256–470 on the *x*-axis) folded domains. (c) Examples of structures of GS0 with the same R_g as the average over simulations using CALVADOS2$_{C_\alpha}$ (left) and CALVADOS2$_{COM}$ (right). The starting structure of the simulations is shown in the middle, where green and orange parts are the two fluorescent proteins connected by a flexible linker (gray). The regions that interact strongly in the CALVADOS2$_{C_\alpha}$ simulations are colored blue.

and CALVADOS2$_{COM}$ gave similar agreement to SAXS experiments. Likewise, we found a similar agreement for the MDPs (Figures 1d and 3b,c).

Having optimized λ, we validated the CALVADOS3$_{COM}$ model on 25 IDPs and 9 MDPs (Tables S4 and S5) that were not used in training for any of the models (Figure 4). For the 25 IDPs, we found good agreement for all three models (CALVADOS2$_{C_\alpha}$, CALVADOS2$_{COM}$, and CALVADOS3$_{COM}$) (Figure 4a–c). We note again that the COM representation is only applied to the folded domain. All IDPs have C_α representations, so CALVADOS2$_{C_\alpha}$ and CALVADOS2$_{COM}$ are the same models for IDPs. In contrast, for MDPs we found that CALVADOS3$_{COM}$ and CALVADOS2$_{COM}$ perform substantially better than CALVADOS2$_{C_\alpha}$ (Figure 4a–c). Our validation results thus show that the CALVADOS3$_{COM}$ model gives improved agreement for simulations of MDPs while retaining the accuracy of CALVADOS2$_{C_\alpha}$ for simulations of IDPs. Across the 34 independent test proteins we find $\left\langle \chi^2_{R_g} \right\rangle$ values of 50, 22, and 15 for CALVADOS2$_{C_\alpha}$, CALVADOS2$_{COM}$, and CALVADOS3$_{COM}$, respectively (Figure S4), and both CALVADOS2$_{COM}$ and CALVADOS3$_{COM}$ have essentially no bias (rMSD≈0; Figure 4b,c).

2.3 | Simulations of phase separation of disordered and multi-domain proteins

We and others have previously used one-bead-per-residue models such as CALVADOS to study the self-association and phase separation of IDPs (Dannenhoffer-Lafage & Best, 2021; Dignon et al., 2018; Joseph et al., 2021; Regy et al., 2021; Tesei & Lindorff-Larsen, 2023; Tesei, Schulze, et al., 2021; Valdes-Garcia et al., 2023; Wessén et al., 2022). In some cases, these models have also been used to study phase separation of proteins that contain a mixture of folded and disordered regions (Conicella et al., 2020; Dignon et al., 2018; Her et al., 2022). We therefore examined whether the CALVADOS3$_{COM}$ model could be used to study phase separation of both IDPs and MDPs. We used multi-chain simulations in a slab geometry (Dignon et al., 2018) to simulate the partitioning of proteins between a dilute and dense phase, and calculated the dilute phase concentration (the saturation concentration; c_{sat}) as a sensitive measure of the accuracy of the model. We first simulated 33 IDPs and found that simulations with CALVADOS3$_{COM}$ gave an agreement with experimental values of c_{sat} that is comparable to that of CALVADOS2$_{C_\alpha}$ (Table S6, and Figures S5, S6, and S7).

We then proceeded to use CALVADOS3$_{COM}$ to study the phase separation of MDPs including hnRNPA1* (where * denotes that residues 259–264 have been deleted from full-length hnRNPA1), full-length FUS (FL_FUS) and other MDPs with experimental estimates of c_{sat} (Table S7; Wang et al., 2018; Martin, Thomasen, et al., 2021). Simulations of hnRNPA1* with CALVADOS2$_{C_\alpha}$, under conditions where the experimental dilute phase concentration is 0.17 mM, resulted in essentially all proteins in the dense phase ($c_{sat} = 0$ mM; Figure 5a). In contrast, simulations using CALVADOS3$_{COM}$ resulted in a lower propensity to phase separate and a calculated value of $c_{sat} = 0.14 \pm 0.01$ mM that is comparable to experiments (Figure 5b).

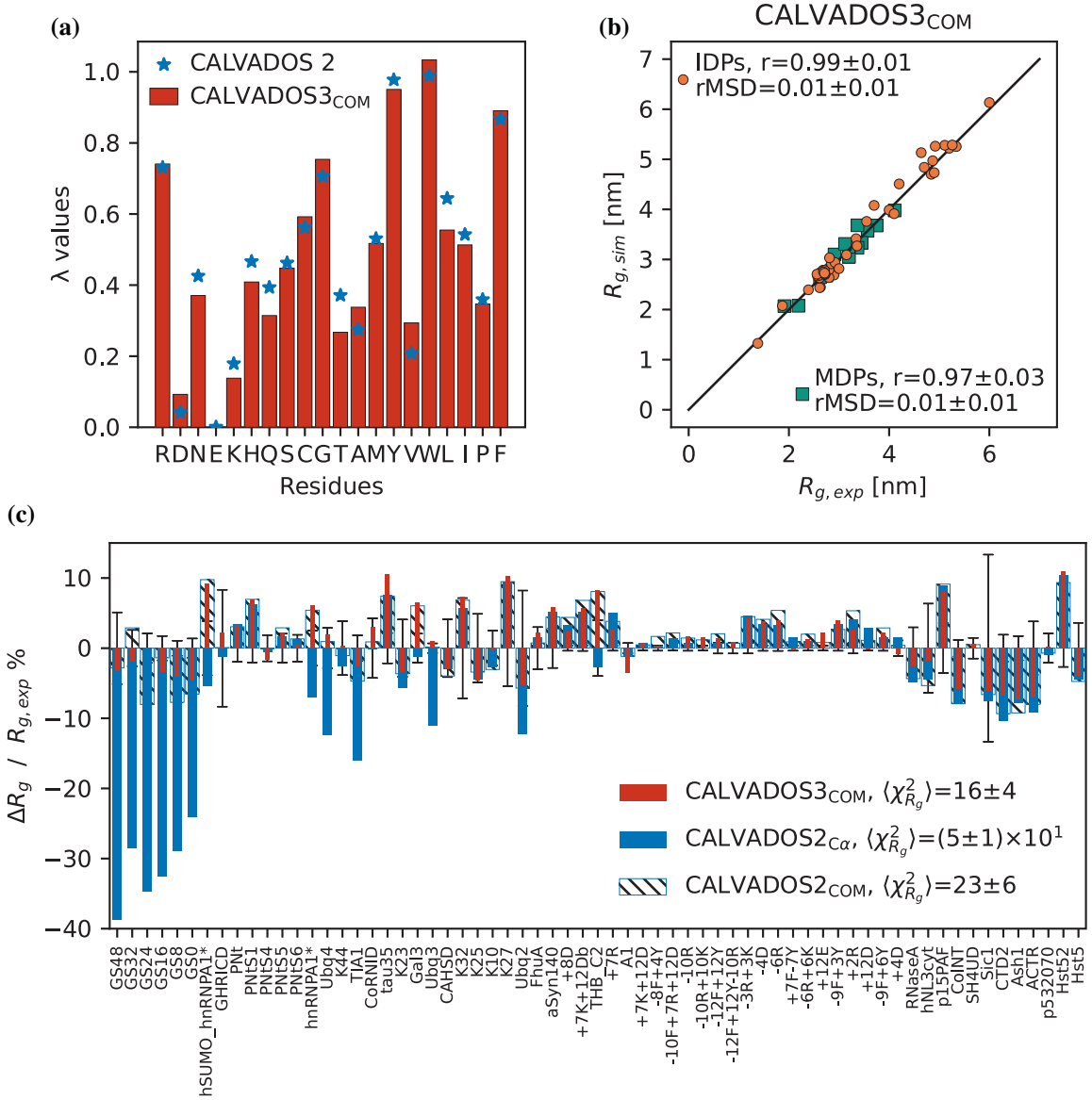

FIGURE 3 Optimizing the λ parameters using a COM representation for folded domains. (a) Comparison between λ values from CALVADOS 2 (blue) and CALVADOS3$_{\mathrm{COM}}$ (red). (b) Comparison between simulated and experimental R_g values for IDPs (orange) and MDPs (green) using CALVADOS3$_{\mathrm{COM}}$. Pearson correlation coefficients (r) and rMSD are reported in the legend. The black diagonal line indicates $y = x$. (c) Relative difference between experimental and simulated R_g values from CALVADOS3$_{\mathrm{COM}}$ (red), CALVADOS2$_{C_\alpha}$ (blue) and CALVADOS2$_{\mathrm{COM}}$ (blue hatched). $\left\langle \chi^2_{R_g} \right\rangle$ values across IDPs and MDPs in training set are reported in the legend. Error bars show the experimental error divided by $R_{\mathrm{g,exp}}$.

To understand the origin of these differences, we calculated interaction energy maps of the proteins in the dense phase. Experiments have shown that the LCD in hnRNPA1* (residues 186–320) plays a central role in driving phase separation (Martin, Thomasen, et al., 2021; Molliex et al., 2015), and we indeed found evidence for substantial LCD–LCD interactions in the dense phases in simulations with both CALVADOS2$_{C_\alpha}$ (Figure 5c) and CALVADOS3$_{\mathrm{COM}}$ (Figure 5d). In the simulations with CALVADOS2$_{C_\alpha}$ we, however, also observed more substantial interactions between the folded RRM (RNA recognition motif) domains (residues 14–97 and 105–185) and between the RRMs and the LCD. In simulations with CALVADOS3$_{\mathrm{COM}}$ these interactions were much weaker, presumably explaining the increase of c_{sat} in these simulations.

Having demonstrated that CALVADOS3$_{\mathrm{COM}}$ provides a more accurate description of the phase behavior of hnRNPA1* than CALVADOS2$_{C_\alpha}$, we proceeded to perform simulations of several other MDPs for which we found estimates of c_{sat} in the literature (Figures 6, S8, and S9). As for hnRNPA1*, we found that CALVADOS2$_{C_\alpha}$ substantially overestimates the tendency of these proteins to undergo phase separation (i.e., underestimate c_{sat}). The

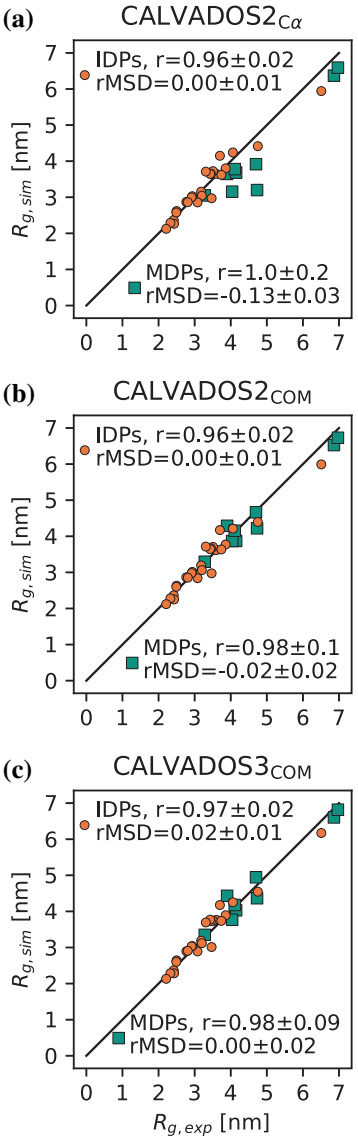

FIGURE 4 Validation of the CALVADOS3$_{COM}$ model using proteins that were not used during training. Comparison of simulated and experimental R_g values on a validation set using (a) CALVADOS2$_{C_\alpha}$, (b) CALVADOS2$_{COM}$ and (c) CALVADOS3$_{COM}$. Pearson correlation coefficients (r) and rMSD are reported in the legend. The black diagonal lines indicate $y = x$.

use of the COM representation in CALVADOS3$_{COM}$ decreases the protein–protein interactions, and thus substantially improves the agreement with experiments, though differences remain.

2.4 | Examining changes in folding stability in condensates

Experiments have shown that the protein-rich environment of condensates can modulate the stability of folded proteins or nucleic acids (Ahmed et al., 2024; Chen et al., 2024; Nott et al., 2015; Ruff et al., 2022). Inspired by these findings, we used the ability to simulate both folded and disordered regions with CALVADOS 3 to examine how partitioning into condensates may shift the folding equilibrium of a folded domain. As it is difficult to sample the folding-unfolding equilibrium by simulations, we studied it indirectly using a thermodynamic cycle that involves differences in partitioning of the folded and unfolded forms into a condensate (Nott et al., 2015).

To demonstrate how CALVADOS 3 enables such analyses, we simulated the isolated RRM1 and RRM2 from hnRNPA1* (Figure 7a) in the presence of a condensate of the LCD of hnRNPA1* and calculated the free energies of partitioning of the RRM domains in their native, folded state, ΔG^N_{part}. Using the same approach, we performed direct-coexistence simulations without applying harmonic networks to the RRMs to calculate the free energies of partitioning of the RRMs in their unfolded state, ΔG^U_{part}. A comparison of the concentration profiles from our direct-coexistence simulations shows that the unfolded RRMs accumulate in the condensate and are depleted from the dilute phase to a greater extent than the folded RRMs (Figure 7b–c); We quantify this via a more negative free energy of partitioning, $\Delta G^U_{\text{part}} < \Delta G^N_{\text{part}}$ (Figure 7d). The preference of the unfolded state for the condensate is particularly pronounced for RRM2, for which we estimate a two-fold decrease in the free energy of partitioning ($\Delta G^U_{\text{part}} - \Delta G^N_{\text{part}} = -0.7$ kcal/mol). From the thermodynamic cycle, this in turn means that the folding stability of RRM2 is 0.7 kcal mol^{-1} lower (less stable) in the condensate than in the dilute phase.

To put these changes into context, we used a recently developed machine learning approach (Cagiada et al., 2024) to predict the absolute protein folding stabilities of the isolated RRMs in the dilute phase, $\Delta G^{\text{dil}}_{N\to U}$, and obtained 6.6 kcal mol^{-1} for RRM1 and 4.4 kcal mol^{-1} for RRM2. Using these values and assuming a two-state model, we estimate that the partitioning into the condensate has a negligible effect on the amount of unfolded state for RRM1; in contrast we predict a four-fold increase in the population of the unfolded state of RRM2 from $\exp\left(-\Delta G^{\text{dil}}_{N\to U}/RT\right) \approx 1/2000$ to $\exp\left[-\left(\Delta G^{\text{dil}}_{N\to U} + \Delta G^U_{\text{part}} - \Delta G^N_{\text{part}}\right)/RT\right] \approx 1/500$. Although substantial additional work is needed to examine the accuracy of CALVADOS 3 for quantifying differences in partitioning of folded and unfolded proteins into condensates, these data show a promising use of our model for predicting unfolding in condensates.

3 | DISCUSSION

In this work, we found that simulations with the CALVADOS2$_{C_\alpha}$ model, previously shown to represent

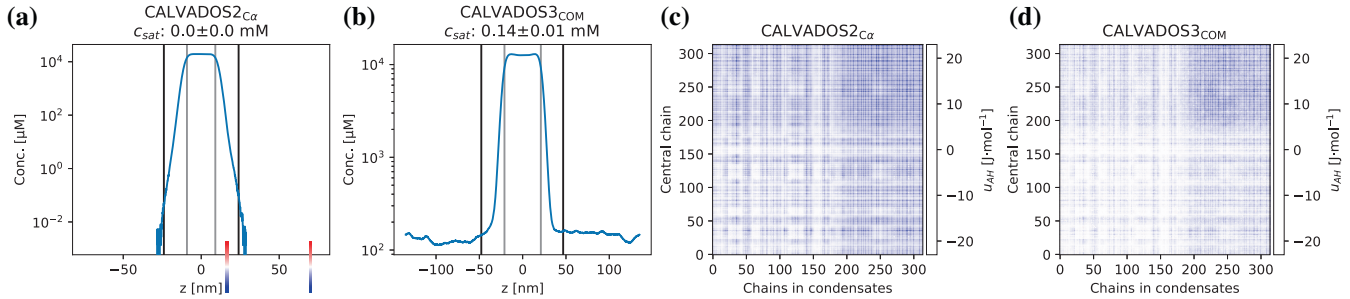

FIGURE 5 Phase coexistence simulations of hnRNPA1* using (a, c) CALVADOS2$_{C_\alpha}$ and (b, d) CALVADOS3$_{COM}$. Simulations were performed at 293 K and an ionic strength of 0.15 M. Equilibrium density profile of hnRNPA1* using (a) CALVADOS2$_{C_\alpha}$ and (b) CALVADOS3$_{COM}$. c_{sat} calculated from density profiles are 0 and 0.14 mM, respectively. Average residue–residue interaction energies (the Ashbaugh-Hatch term in the force field) between the most central chain and the rest of the condensate for (c) CALVADOS2$_{C_\alpha}$ and (d) CALVADOS3$_{COM}$.

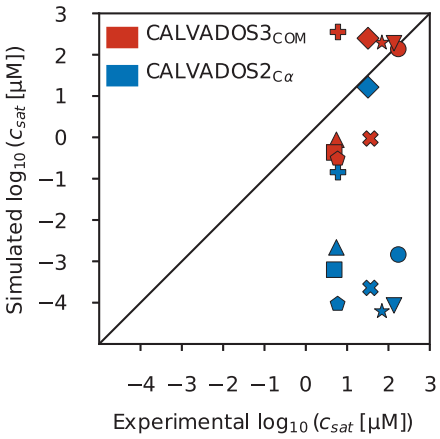

FIGURE 6 Comparison between simulated and experimental c_{sat} values for MDPs using the CALVADOS3$_{COM}$ model (red) and CALVADOS2$_{C_\alpha}$ (blue). The simulated proteins are hnRNPA1* (circle), hSUMO_hnRNPA1* (downward triangle), FL_FUS (upward triangle), GFP_FUS (square), SNAP_FUS (pentagon), SNAP_FUS_PLDY2F_RBDR2K (star), SNAP_FUS_PLDY2F (x symbol), FUS_PLDY2F_RBDR2K (diamond) and hnRNPA3 (plus symbol). The black diagonal line indicates $y = x$.

single-chain and multi-chain properties of IDPs, underestimated the radii of gyration of MDPs. Changing the CG mapping method from C_α to COM substantially improved the agreement with experimental data. This observation is in line with the finding that reconstruction of all-atom structures from a center-of-mass representation is more accurate than from a C_α representation (Heo & Feig, 2024). We reoptimized the "stickiness" parameters in the context of a COM-based model based on experimental data for both IDPs and MDPs. The resulting CALVADOS3$_{COM}$ model provides a good description of both single- and multi-chain simulations of both IDPs and MDPs.

The relatively low c_{sat} value calculated from slab simulations of hnRNPA1* with CALVADOS2$_{C_\alpha}$ further supported that interactions between the folded domains are overestimated by C_α-based models without any further modifications. Considering that the SCCOM-based model (CALVADOS2$_{SCCOM}$) overestimated R_g of MDPs, we suggest that the COM-based model (CALVADOS3$_{COM}$) appears to strike a good balance, leading to improved values of c_{sat} for MDPs. Nevertheless, some systematic differences remain even with this model, which resulted in underestimates of c_{sat} for different constructs of the protein FUS. Together, our results show that the new parameter set and the center-of-mass representation (CALVADOS3$_{COM}$) retain the accuracy of CALVADOS 2 for IDPs, but improve the description of proteins with both disordered and folded domains. We therefore term this new model CALVADOS 3, with the implicit notion that this model is used with center-of-mass representation of residues within folded regions. We note that a preprint describing our work (Cao et al., 2024) used a slightly different set of parameters, and we suggest to refer to that model as CALVADOS 3beta.

When simulating MDPs with CALVADOS 3 we need to restrain the folded domains using harmonic restraints. In the current work, we have manually determined the boundaries for which regions are considered to be folded, though automated methods will be needed for large-scale applications. Tools for automatic predictions of domain boundaries exist (Holm & Sander, 1994; Lau et al., 2023) and might be combined with AlphaFold to set the harmonic restraints (Jussupow & Kaila, 2023).

Despite these current limitations, we envision that the CALVADOS 3 model will enable detailed studies of the interactions within and between MDPs, and pave the way for proteome-wide simulation studies of full-length proteins similar to what has recently been achieved for IDRs (Tesei et al., 2024). We also envision that our approach to study changes in protein stability inside condensates can be used together with methods to predict

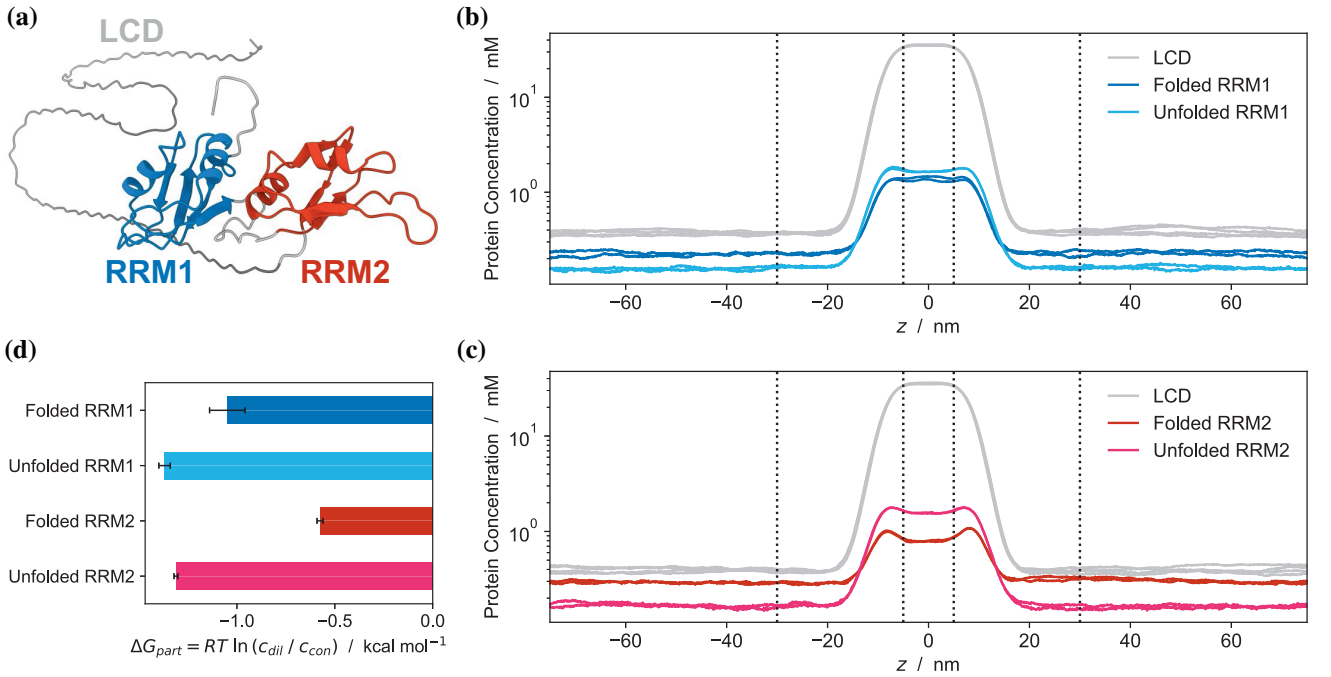

FIGURE 7 Predicting the effect of the protein-rich environment of a condensate on the stability of folded domains. (a) Structure of hnRNPA1* highlighting the low-complexity domain (gray) and RNA-recognition motifs 1 (blue) and 2 (red). (b) Concentration profiles of the LCD (gray) and RRM1 in the native (blue) and unfolded (cyan) state. (c) Concentration profiles of the LCD (gray) and RRM2 in the native (red) and unfolded (magenta) state. (d) Free energy of partitioning of RRM1 and RRM2 in native and unfolded states into condensates of the LCD. Data estimated from direct-coexistence simulations performed in two independent replicates. Error bars in (d) represent the differences between the replicates.

absolute protein stability (Cagiada et al., 2024) to learn and expand our knowledge on the rules that underlie phase separation and changes in stability of folded, globular proteins (Ruff et al., 2022).

4 | METHODS

4.1 | Description of the model

We modeled each amino acid by one bead. We generated C_α-beads for IDPs and assigned C_α atom coordinates to bead positions for IDRs in MDPs according to their modeled or experimental structures (Section 4.2). For structured domains, we used the following rules for the different representations: we placed each bead position at the C_α atom (C_α representation), or the center of mass calculated for all the atoms in a residue (COM representation), or the center of mass calculated for only side chain atoms of a residue (SCCOM representation). The CALVADOS 3 energy function consists of bonded interactions, non-bonded interactions and an elastic network model as described below.

Chain connectivity of the beads is described by a harmonic potential,

$$u_{\text{bond}}(r) = \frac{1}{2}k(r - r_0)^2, \tag{1}$$

with force constant $k = 8033\,\text{kJ}\cdot\text{mol}^{-1}\cdot\text{nm}^{-2}$. The equilibrium distance r_0 is set to 0.38 nm if two beads are both within IDRs, or the distance between two beads in the initial conformation if at least one bead is within a folded domain.

For non-bonded interactions, we use a truncated and shifted Ashbaugh-Hatch (AH) and Debye-Hückel (DH) potential to model van der Waals and salt-screened electrostatic interactions, respectively. The Ashbaugh-Hatch potential is described by

$$u_{\text{AH}}(r) = \begin{cases} u_{\text{LJ}}(r) - \lambda u_{\text{LJ}}(r_c) + \epsilon(1-\lambda), & r \le 2^{1/6}\sigma \\ \lambda[u_{\text{LJ}}(r) - u_{\text{LJ}}(r_c)], & 2^{1/6}\sigma < r \le r_c, \\ 0, & r > r_c \end{cases} \tag{2}$$

where $u_{\text{LJ}}(r)$ is the Lennard-Jones (LJ) potential,

$$u_{\text{LJ}}(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right], \tag{3}$$

and where $\epsilon = 0.8368\,\mathrm{kJ\cdot mol^{-1}}$ and $r_c = 2.2$ or 2 nm. Similar to previous work, we use $r_c = 2.2$ nm during the optimization of CALVADOS3$_{\mathrm{COM}}$, and use 2 nm during validation and application (Tesei & Lindorff-Larsen, 2023). Both σ and λ are calculated as the arithmetic averages of residue-specific bead size and stickiness, respectively. σ values are van der Waals volumes calculated by Kim and Hummer (2008)). λ values are treated as free parameters and optimized iteratively through a Bayesian parameter-learning procedure as described previously (Tesei & Lindorff-Larsen, 2023; Tesei, Schulze, et al., 2021) to minimize the differences in the simulated and experimental R_g and PRE data. In simulations where we scaled down interactions of folded domains (CALVADOS2$_{C_\alpha}$ 70%), we scaled down ϵ to 0.7ϵ for domain–domain interactions and to $\sqrt{0.7}\epsilon$ for domain-IDR interactions.

The Debye-Hückel potential is described by

$$u_{\mathrm{DH}}(r) = \frac{q_i q_j e^2}{4\pi\epsilon_0\epsilon_r}\frac{\exp(-r/D)}{r}, \tag{4}$$

where q is the average amino acid charge number, e is the elementary charge, $D = \sqrt{1/(8\pi B c_s)}$ is the Debye length of an electrolyte solution of ionic strength c_s, $B(\epsilon_r)$ is the Bjerrum length and ϵ_0 is the vacuum permittivity. Electrostatic interactions are truncated and shifted at the cutoff distance $r_c = 4$ nm. The temperature-dependent dielectric constant of the implicit aqueous solution is modeled by the following empirical relationship (Akerlof & Oshry, 1950):

$$\epsilon_r(T) = \frac{5321}{T} + 233.76 - 0.9297\times T + 1.417\times 10^{-3}\times T^2 - 8.292\times 10^{-7}\times T^3. \tag{5}$$

We use the Henderson-Hasselbalch equation to estimate the average charge of the histidine residues, assuming a pK_a value of 6 (Nagai et al., 2008).

We use an elastic network model (ENM) with a harmonic potential to restrain non-bonded pairs in the folded domains using

$$u_{\mathrm{ENM}}(r) = \frac{1}{2}k_d(r - r_0)^2. \tag{6}$$

Here, the force constant k_d is $700\,\mathrm{kJ\cdot mol^{-1}\cdot nm^{-2}}$, r is the distance between beads and equilibrium distances r_0 are directly taken from the reference structures. We only apply the ENM to residue pairs with an r_0 below a 0.9 nm cutoff. We determine the predefined boundary of each domain in MDPs by visual inspection of the three-dimensional structures (Table S8). Each domain has a starting amino acid and an ending amino acid indicating the range of the domain. Only residue pairs within the same domain are restrained by this harmonic potential except for bonded pairs, which are restrained by the aforementioned bonded potential. All boundaries of MDPs are consistent with definitions in their experimental or simulation articles. In some cases, one domain could be discontinuous because of long loops within the domain, so we exclude those regions when defining boundaries. Residues of α-helix, β-sheet and short loops in a structured domain are all restrained equally with the same force constant and cutoff distance. The application of ENM ensures that secondary structures within folded domains do not fluctuate substantially (Figure S10). Non-bonded interactions (Ashbaugh-Hatch and Debye-Hückel potential) are excluded for the restrained pairs.

4.2 | Simulations

We generated initial conformations of all IDPs as Archimedes' spirals with a distance of 0.38 nm between bonded beads. Atomistic structures of all MDPs used in optimization procedures, single-chain validation and slab simulations either came from our recent work (Thomasen et al., 2024) or were modeled by superposing experimental domain structures (if available) on AlphaFold predictions (Jumper et al., 2021; Varadi et al., 2022). We then mapped all of these MDPs to CG structures based on different CG representations (C_α, COM, SCCOM).

We conducted Langevin dynamics simulations using OpenMM 7.6.0 (Eastman et al., 2017) in the NVT ensemble with an integration time step of 10 fs and friction coefficient of $0.01\,\mathrm{ps^{-1}}$. Single chains of N residues were simulated in a cubic box with a $(N-1)\times 0.38 + 4$ nm box edge length under periodic boundary conditions. Each chain was simulated in 20 replicas for 6.3 ns to 77.7 ns depending on the sequence length of the disordered regions (Tesei et al., 2024; Tesei & Lindorff-Larsen, 2023). Final trajectories had 4000 frames for each protein, excluding the initial 10 frames in each replica.

We performed direct-coexistence simulations in a cuboidal box using $[L_x, L_y, L_z] = [17, 17, 300]$ and $[15, 15, 150]$ nm to simulate multi-chains of Ddx4WT and the other IDPs, respectively. For MDPs, box sizes are shown in Table S7. To keep the condensates thick enough and reduce finite-size surface effects, we chose 150 chains for hnRNPA1* and 100 chains for all the other IDPs and MDPs (see also below). We generated each IDP chain as an Archimedes' spiral with a distance of 0.38 nm between bonded beads in the *xy*-plane. Each spiral was

placed along the z-axis with a spacing of 1.47 nm. To avoid steric clashes of densely packed MDP input structures, we chose the most compact conformation sampled by single-chain simulations with CALVADOS 2 parameters and corresponding CG representation as the initial conformation for each MDP chain. Before production simulations, we performed equilibrium runs where we used an external force to push each chain towards the center of the box so that a condensate could be formed. We then continued to perform production simulations, saving frames every 0.125 ns and discarded the first 150 ns before analysis. The slab in each frame was centred in the box and the equilibrium density profile $\rho(z)$ was calculated by taking the averaged densities over the trajectories as previously described (Tesei & Lindorff-Larsen, 2023).

To examine finite-size effects of the direct-coexistence simulations we performed additional simulations of hnRNPA1* varying both the box dimensions (L_x, L_y, L_z) and the number of chains. We calculated both dense and dilute phase concentrations from each simulation and find that unless we use a very small patch ($L_x = L_y = 11$ nm), the results are consistent (Figures S11 and S12;Table S9), in line with previous analyses of such finite-size effects (Dignon et al., 2018; Joseph et al., 2021). Convergence of the IDP simulations was assessed as previously described (Tesei, Schulze, et al., 2021).

To indicate the computational performance of single- and multi-chain CALVADOS simulations, we show the performance for systems of different sizes run either on an Intel Xeon Gold 6130 CPU (for single-chain simulations) or on an NVIDIA Tesla V100 GPU (for multi-chain simulations) (Figure S13).

To estimate the free energy of partitioning of RRM1 (residues 11–89) and RRM2 (residues 105–179) into condensates of hnRNPA1* LCD (GS followed by residues 186–258 and 265–320), we performed direct-coexistence simulations at 298 K, pH 7.5, and 150 mM ionic strength, in a cuboidal box with sidelengths $[L_x, L_y, L_z] = [15, 15, 150]$ nm. The structures of the native states of RRM1 and RRM2 were based on the crystal structure (Shamoo et al., 1997) as previously described (Martin, Thomasen, et al., 2021). We performed two independent simulations, each 21 μs long, for each system and, after centering the LCD condensate in the middle of the box, calculated concentration profiles along the z-axis using the last 20 μs of each trajectory. We estimated the free energies of partitioning as $\Delta G_{\text{part}} = RT\ \ln(c_{\text{dil}} / c_{\text{con}})$ where R is the gas constant and c_{dil} and c_{con} are the average concentrations of the RRMs in the dilute phase and in the LCD condensate, respectively. The error on ΔG_{part} was estimated as the difference between the values from the two independent simulation replicas. Absolute folding stabilities of RRM1 and RRM2 were calculated using the Google Colab implementation of a recently described model for predicting absolute protein stability (Cagiada et al., 2024).

4.3 | Parameter optimization

Our Bayesian Parameter-Learning Procedure (Tesei & Lindorff-Larsen, 2023) of the "stickiness" parameters, λ, aimed to minimize the following cost function:

$$\mathscr{L}(\lambda) = \left\langle \chi^2_{R_g} \right\rangle + \eta \langle \chi^2_{\text{PRE}} \rangle - \theta \ln(P(\lambda)). \tag{7}$$

$\chi^2_{R_g}$ and χ^2_{PRE} denoting R_g and PRE differences between experiments and simulations are estimated as

$$\chi^2_{Rg} = \left(\frac{R_g^{\text{exp}} - R_g^{\text{calc}}}{\sigma^{\text{exp}}} \right)^2 \tag{8}$$

and

$$\chi^2_{\text{PRE}} = \frac{1}{N_{\text{labels}} N_{\text{res}}} \sum_j^{N_{\text{labels}}} \sum_i^{N_{\text{res}}} \left(\frac{Y_{ij}^{\text{exp}} - Y_{ij}^{\text{calc}}}{\sigma_{ij}^{\text{exp}}} \right)^2. \tag{9}$$

Here $P(\lambda)$ is a statistical prior of λ (Tesei & Lindorff-Larsen, 2023), σ^{exp} is the error on the experimental values, Y is PRE data, either $I_{\text{para}}/I_{\text{dia}}$ or Γ_2 is calculated using the rotamer library approach implemented in DEER-PREdict (Tesei, Martins, et al., 2021), N_{labels} is the number of spin-labeled mutants, and N_{res} is the number of measured residues. The prior loss, $\theta \ln(P(\lambda))$, quantifies the difference between prior distribution $P(\lambda)$ and current λ values (with min-max normalization at each step) to avoid overfitting. The coefficients are set to $\eta = 0.1$ and $\theta = 0.08$. λ is not allowed to be negative but can be greater than 1.0 during optimization.

We used a training set consisting of 56 IDPs and 14 MDPs to perform the optimization. All of those proteins were from our previous studies (Tesei & Lindorff-Larsen, 2023; Thomasen et al., 2023). A summary of the training data and other properties of different CALVADOS models is shown in the supporting material (Table S10). We used 51 IDPs and 14 MDPs as training set for fitting against experimental SAXS R_g data and 5 IDPs were used for fitting against experimental PRE data (Tables S1, S2, and S3). We then used a validation set to validate the performances of our new optimized models on reproducing experimental R_g. This validation set was composed of 25 IDPs and 9 MDPs. Twelve IDPs in this validation set were from our previous work and

the rest (13 IDPs and 9 MDPs) were newly collected experimental R_g data in this work (Tables S4 and S5). We also collected nine MDPs with measured values of c_{sat} to examine the accuracy of the phase behavior simulated with the models presented in this work (Table S7).

The optimization procedure went through several cycles until convergence of the final total cost ($|\Delta\mathscr{L}| < 1$, $\Delta\mathscr{L}$ is the difference between the lowest total cost of final total the current and previous cycle, Figure 7). Within each cycle, we use the optimized λ values from the previous cycle to perform new single-chain simulations (initial λ values for the first cycle are CALVADOS 2 parameters, (Tesei & Lindorff-Larsen, 2023)), calculate R_g and PRE for each frame and then nudge values in the λ set iteratively to minimize the cost function (five residues are randomly subjected to small perturbations sampled from a Gaussian distribution with $\mu = 0, \sigma = 0.05$). This trial λ set (λ_k) is used to calculate the Boltzmann weights of each frame by $w_i = \exp(-[U(r_i,\lambda_k) - U(r_i,\lambda_0)]/k_B T)$, where U is the AH potential, r_i are coordinates of a conformation, k_B is the Boltzmann constant and T is temperature. The resulting weights are then used to calculate the effective fraction of frames by $\phi_{eff} = \exp\left[-\sum_i^{N_{frames}} w_i \log(w_i \times N_{frames})\right]$; if $\phi_{eff} \geq 0.6$, trial λ_k acceptance probability is determined by the Metropolis criterion, $\min\left\{1, \exp\left(\frac{\mathscr{L}(\lambda_{k-1}) - \mathscr{L}(\lambda_k)}{\xi_k}\right)\right\}$, where ξ_k is a unitless control parameter, its initial value is set to 0.1 and scaled down by 1% at each iteration until $\xi < 10^{-8}$, which means a micro-cycle is complete. Within a cycle, a total of 10 micro-cycles are performed. In this work, the optimization procedure converged within three cycles. Therefore, we used the resulting λ values from the third cycle as the final parameter set. We ran one additional optimization cycle to confirm the convergence of the training.

AUTHOR CONTRIBUTIONS

Fan Cao: Investigation; writing – original draft; methodology; validation; visualization; writing – review and editing; software; formal analysis; data curation; conceptualization. **Sören von Bülow:** Conceptualization; investigation; writing – review and editing; methodology; validation; software; formal analysis; data curation. **Giulio Tesei:** Conceptualization; investigation; methodology; validation; writing – review and editing; software; formal analysis; data curation. **Kresten Lindorff-Larsen:** Conceptualization; investigation; writing – original draft; writing – review and editing; supervision; methodology.

ACKNOWLEDGMENTS

We acknowledge the use of computational resources from Computerome 2.0, the ROBUST Resource for Biomolecular Simulations (supported by the Novo Nordisk Foundation; NNF18OC0032608) and the Biocomputing Core Facility at the Department of Biology, University of Copenhagen. This research was supported by the PRISM (Protein Interactions and Stability in Medicine and Genomics) centre funded by the Novo Nordisk Foundation (NNF18OC0033950, to K.L.-L.), and CSC (China Scholarship Council, 202206340019). SB is a recipient of an EMBO postdoctoral fellowship (ALTF 810-2022).

DATA AVAILABILITY STATEMENT

Scripts and data to reproduce the work are available via https://github.com/KULL-Centre/_2024_Cao_CALVADOSCOM.

ORCID

Giulio Tesei https://orcid.org/0000-0003-4339-4460
Kresten Lindorff-Larsen https://orcid.org/0000-0002-4750-6039

REFERENCES

Ahmed R, Liang M, Hudson RP, Rangadurai AK, Huang SK, Forman-Kay JD, et al. Atomic resolution map of the solvent interactions driving SOD1 unfolding in CAPRIN1 condensates. Proc Natl Acad Sci. 2024; 121(35):e2408554121. https://doi.org/10.1101/2024.04.29.591724

Akerlof G, Oshry H. The dielectric constant of water at high temperatures and in equilibrium with its vapor. J Am Chem Soc. 1950;72(7):2844–7.

Ashbaugh HS, Hatch HW. Natively unfolded protein stability as a coil-to-globule transition in charge/hydropathy space. J Am Chem Soc. 2008;130(29):9536–42.

Benayad Z, von Bülow S, Stelzl LS, Hummer G. Simulation of FUS protein condensates with an adapted coarse-grained model. J Chem Theory Comput. 2020;17(1):525–37.

Bereau T, Deserno M. Generic coarse-grained model for protein folding and aggregation. J Chem Phys. 2009;130(23):6B621.

Bondos SE, Dunker AK, Uversky VN. On the roles of intrinsically disordered proteins and regions in cell communication and signaling. Cell Commun Signal. 2021;19(1):88.

Borges-Araújo L, Patmanidis I, Singh AP, Santos LH, Sieradzan AK, Vanni S, et al. Pragmatic coarse-graining of proteins: models and applications. Journal of Chemical Theory and Computation. 2023;19(20):7112–35.

Cagiada M, Ovchinnikov S, Lindorff-Larsen K. Predicting absolute protein folding stability using generative models. bioRxiv. 2024. https://doi.org/10.1101/2024.03.14.584940

Cao F, von Bülow S, Tesei G, Lindorff-Larsen K. A coarse-grained model for disordered and multi-domain proteins. bioRxiv. 2024. https://doi.org/10.1101/2024.02.03.578735

Chen R, Glauninger H, Kahan DN, Shangguan J, Sachleben JR, Riback JA, et al. HDX–MS finds that partial unfolding with sequential domain activation controls condensation of a cellular stress marker. Proc Natl Acad Sci U S A. 2024;121(13): e2321606121.

Conicella AE, Dignon GL, Zerze GH, Schmidt HB, D'Ordine AM, Kim YC, et al. TDP-43 α-helical structure tunes liquid–liquid phase separation and function. Proc Natl Acad Sci U S A. 2020; 117(11):5883–94.

Dannenhoffer-Lafage T, Best RB. A data-driven hydrophobicity scale for predicting liquid–liquid phase separation of proteins. J Phys Chem B. 2021;125(16):4046–56.

Delaforge E, Milles S, Jr H, Bouvier D, Jensen MR, Sattler M, et al. Investigating the role of large-scale domain dynamics in protein-protein interactions. Front Mol Biosci. 2016;3:54.

Dignon GL, Zheng W, Kim YC, Best RB, Mittal J. Sequence determinants of protein phase behavior from a coarse-grained model. PLoS Comput Biol. 2018;14(1):e1005941.

Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, et al. OpenMM 7: rapid development of high performance algorithms for molecular dynamics. PLoS Comput Biol. 2017;13(7):e1005659.

Gopal SM, Mukherjee S, Cheng YM, Feig M. PRIMO/PRIMONA: a coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. Proteins Struct Funct Bioinform. 2010;78(5):1266–81.

Han JH, Batey S, Nickson AA, Teichmann SA, Clarke J. The folding and evolution of multidomain proteins. Nat Rev Mol Cell Biol. 2007;8(4):319–30.

Heo L, Feig M. One bead per residue can describe all-atom protein structures. Structure. 2024;32(1):97–111.

Her C, Phan TM, Jovic N, Kapoor U, Ackermann BE, Rizuan A, et al. Molecular interactions underlying the phase separation of HP1α: role of phosphorylation, ligand and nucleic acid binding. Nucleic Acids Res. 2022;50(22):12702–22.

Holm L, Sander C. Parser for protein folding units. Proteins Struct Funct Bioinform. 1994;19(3):256–68.

Hyeon C, Dima RI, Thirumalai D. Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. Structure. 2006;14(11):1633–45.

Joseph JA, Reinhardt A, Aguirre A, Chew PY, Russell KO, Espinosa JR, et al. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. Nat Comput Sci. 2021;1(11):732–43.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.

Jussupow A, Kaila VR. Effective molecular dynamics from neural network-based structure prediction models. J Chem Theory Comput. 2023;19(7):1965–75.

Kim YC, Hummer G. Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. J Mol Biol. 2008;375(5):1416–33.

Kolinski A, Skolnick J. Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. Proteins Struct Funct Bioinform. 1998;32(4):475–94.

Krainer G, Welsh TJ, Joseph JA, Espinosa JR, Wittmann S, de Csilléry E, et al. Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions. Nat Commun. 2021;12(1):1085.

Lau AM, Kandathil SM, Jones DT. Merizo: a rapid and accurate protein domain segmentation method using invariant point attention. Nat Commun. 2023;14(1):8445.

Li W, Terakawa T, Wang W, Takada S. Energy landscape and multiroute folding of topologically complex proteins adenylate kinase and 2ouf-knot. Proc Natl Acad Sci U S A. 2012;109(44): 17789–94.

Mackereth CD, Sattler M. Dynamics in multi-domain protein recognition of RNA. Curr Opin Struct Biol. 2012;22(3):287–96.

Maity H, Baidya L, Reddy G. Salt-induced transitions in the conformational ensembles of intrinsically disordered proteins. J Phys Chem B. 2022;126(32):5959–71.

Martin EW, Thomasen FE, Milkovic NM, Cuneo MJ, Grace CR, Nourse A, et al. Interplay of folded domains and the disordered low-complexity domain in mediating hnRNPA1 phase separation. Nucleic Acids Res. 2021;49(5):2931–45.

Mittal A, Holehouse AS, Cohan MC, Pappu RV. Sequence-to-conformation relationships of disordered regions tethered to folded domains of proteins. J Mol Biol. 2018;430(16):2403–21.

Molliex A, Temirov J, Lee J, Coughlin M, Kanagaraj AP, Kim HJ, et al. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. Cell. 2015;163(1):123–33.

Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink SJ. The MARTINI coarse-grained force field: extension to proteins. J Chem Theory Comput. 2008;4(5):819–34.

Moses D, Guadalupe K, Yu F, Flores E, Perez AR, McAnelly R, et al. Structural biases in disordered proteins are prevalent in the cell. Nat Struct Mol Biol. 2024;31:1–10.

Mugnai ML, Chakraborty D, Kumar A, Nguyen HT, Zeno W, Stachowiak JC, et al. Sizes, conformational fluctuations, and SAXS profiles for intrinsically disordered proteins. bioRxiv. 2023. https://doi.org/10.1101/2023.04.24.538147

Nagai H, Kuwabara K, Carta G. Temperature dependence of the dissociation constants of several amino acids. J Chem Eng Data. 2008;53(3):619–27.

Neri M, Anselmi C, Cascella M, Maritan A, Carloni P. Coarse-grained model of proteins incorporating atomistic detail of the active site. Phys Rev Lett. 2005;95(21):218102.

Norgaard AB, Ferkinghoff-Borg J, Lindorff-Larsen K. Experimental parameterization of an energy function for the simulation of unfolded proteins. Biophys J. 2008;94(1):182–92.

Nott TJ, Petsalaki E, Farber P, Jervis D, Fussner E, Plochowietz A, et al. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. Mol Cell. 2015;57(5):936–47.

Pappu RV, Schneller WJ, Weaver DL. Electrostatic multipole representation of a polypeptide chain: an algorithm for simulation of polypeptide properties. J Comput Chem. 1996;17(8):1033–55.

Regy RM, Thompson J, Kim YC, Mittal J. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. Protein Sci. 2021;30(7):1371–9.

Ruff KM, Choi YH, Cox D, Ormsby AR, Myung Y, Ascher DB, et al. Sequence grammar underlying the unfolding and phase separation of globular proteins. Mol Cell. 2022;82(17):3193–208.

Sekiyama N, Takaba K, Maki-Yonekura S, Ki A, Ohtani Y, Imamura K, et al. ALS mutations in the TIA-1 prion-like domain trigger highly condensed pathogenic structures. Proc Natl Acad Sci. 2022;119(38):e2122523119.

Shamoo Y, Krueger U, Rice L, Williams K, Steitz T. Crystal structure of the two RNA binding domains of human hnRNP A1 at 1.75 Å resolution. Nat Struct Biol. 1997;4(3):215–22.

Sieradzan AK, Czaplewski C, Krupa P, Mozolewska MA, Karczyńska AS, Lipska AG, et al. Modeling the structure, dynamics, and transformations of proteins with the UNRES force field. Protein Fold Methods Protoc. 2022;2376:399–416.

Souza PC, Alessandri R, Barnoud J, Thallmair S, Faustino I, Grünewald F, et al. Martini 3: a general purpose force field for coarse-grained molecular dynamics. Nat Methods. 2021;18(4): 382–8.

Tan C, Niitsu A, Sugita Y. Highly charged proteins and their repulsive interactions antagonize biomolecular condensation. JACS Au. 2023;3(3):834–48.

Taneja I, Holehouse AS. Folded domain charge properties influence the conformational behavior of disordered tails. Curr Res Struct Biol. 2021;3:216–28.

Tesei G, Lindorff-Larsen K. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. Open Res Europe. 2023;2(94):94.

Tesei G, Martins JM, Kunze MB, Wang Y, Crehuet R, Lindorff-Larsen K. DEER-PREdict: software for efficient calculation of spin-labeling EPR and NMR data from conformational ensembles. PLoS Comput Biol. 2021;17(1):e1008551.

Tesei G, Schulze TK, Crehuet R, Lindorff-Larsen K. Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. Proc Natl Acad Sci U S A. 2021;118(44):e2111696118.

Tesei G, Trolle AI, Jonsson N, Betz J, Knudsen FE, Pesce F, et al. Conformational ensembles of the human intrinsically disordered proteome. Nature. 2024;626(8000):897–904.

Thomasen FE, Lindorff-Larsen K. Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. Biochem Soc Trans. 2022;50(1):541–54.

Thomasen FE, Pesce F, Roesgaard MA, Tesei G, Lindorff-Larsen K. Improving Martini 3 for disordered and multidomain proteins. J Chem Theory Comput. 2022;18(4):2033–41.

Thomasen FE, Skaalum T, Kumar A, Srinivasan S, Vanni S, Lindorff-Larsen K. Recalibration of protein interactions in Martini 3. Nat Commun. 2024;15(1):6645.

Valdes-Garcia G, Heo L, Lapidus LJ, Feig M. Modeling concentration-dependent phase separation processes involving peptides and RNA via residue-based coarse-graining. J Chem Theory Comput. 2023;19(2):669–78.

Van Der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. Chem Rev. 2014;114(13):6589–631.

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022; 50(D1):D439–44.

Wang J, Choi JM, Holehouse AS, Lee HO, Zhang X, Jahnel M, et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. Cell. 2018; 174(3):688–99.

Wessén J, Das S, Pal T, Chan HS. Analytical formulation and field-theoretic simulation of sequence-specific phase separation of protein-like Heteropolymers with short-and long-spatial-range interactions. J Phys Chem B. 2022;126(45):9222–45.

Yamada T, Miyazaki Y, Harada S, Kumar A, Vanni S, Shinoda W. Improved protein model in SPICA force field. J Chem Theory Comput. 2023;19(23):8967–77.

Yu F, Sukenik S. Structural preferences shape the entropic force of disordered protein ensembles. J Phys Chem B. 2023;127: 4235–44.

Zhang Y, Li S, Gong X, Chen J. Toward accurate simulation of coupling between protein secondary structure and phase separation. J Am Chem Soc. 2023;146:342-57.

Zhang Y, Liu X, Chen J. Toward accurate coarse-grained simulations of disordered proteins and their dynamic interactions. J Chem Inf Model. 2022;62(18):4523–36.

Zheng W, Dignon GL, Jovic N, Xu X, Regy RM, Fawzi NL, et al. Molecular details of protein condensates probed by microsecond long atomistic simulations. J Phys Chem B. 2020;124(51): 11671–9.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Cao F, von Bülow S, Tesei G, Lindorff-Larsen K. A coarse-grained model for disordered and multi-domain proteins. Protein Science. 2024;33(11):e5172. https://doi.org/10.1002/pro.5172