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

Fan Cao[1], **Sören von Bülow**[1], **Giulio Tesei**[1], **Kresten Lindorff-Larsen**[1*]

***For correspondence:** lindorff@bio.ku.dk (KL-L)

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

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 centre 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 centre-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. 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.

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 signalling, catalysis and regulation (***Mackereth and Sattler, 2012***; ***Van Der Lee et al., 2014***; ***Delaforge et al., 2016***; ***Bondos et al., 2021***).

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 and Holehouse, 2021***) and the disordered regions may also affect the properties of the folded domains (***Yu and 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 et al., 2021b***). However, describing the conformational ensembles of MDPs in solution generally requires a combination of biophysical experiments and molecular dynamics (MD) simulations (***Thomasen and 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 (***Zheng et al., 2020***; ***Sekiyama et al., 2022***). 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 (***Neri et al., 2005***; ***Monticelli et al., 2008***; ***Bereau and Deserno, 2009***; ***Gopal et al., 2010***). The accuracy, transferability, and efficiency of such models, however, depend on the degree of coarse-graining and the parameterization strategy (***Heo and Feig, 2023***). 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***, ***2023***). 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 position 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 and 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 (***Dignon et al., 2018***; ***Joseph et al., 2021***; ***Regy et al., 2021***; ***Dannenhoffer-Lafage and Best, 2021***; ***Wessén et al., 2022***; ***Tesei and Lindorff-Larsen, 2023***; ***Valdes-Garcia et al., 2023***).

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 (***Souza et al., 2021***; ***Borges-Araújo et al., 2023***). 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. 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 (***Kim and Hummer, 2008***; ***Dignon et al., 2018***; ***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 (***Pappu et al., 1996***; ***Hyeon et al., 2006***; ***Maity et al., 2022***; ***Zhang et al., 2022***; ***Sieradzan et al., 2022***; ***Mugnai et al., 2023***; ***Zhang et al., 2023***; ***Yamada et al., 2023***). In some of these models, one bead is placed at C_α and the other one is at the centre of mass (COM) of side chain atoms. In this way, side chain interactions can be explicitly taken into account, improving the simulated dynamical behaviour 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) NMR data for a large set of IDPs (***Norgaard et al., 2008***; ***Tesei et al., 2021b***; ***Tesei and Lindorff-Larsen, 2023***). The most recent CALVADOS (Coarse-graining Approach to Liquid-liquid phase separation Via an Automated Data-driven Optimisation Scheme) model (CALVADOS 2) was further tuned to describe phase behaviour of multi-chain conformational ensembles of IDPs from simulations by reducing the range of non-ionic interactions (***Tesei and 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 3beta[1]). When we combine the COM representation of folded domains with harmonic restraints between residues in the folded domains and the CALVADOS 3beta 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.

[1] In a previous version of this preprint (available at https://www.biorxiv.org/content/10.1101/2024.02.03.578735v1) we had termed this model CALVADOS 3. We here rename this model to CALVADOS 3beta to indicate that it is slightly different from the final CALVADOS 3 model that will be published in a journal. All data and figures in this preprint refer to results with CALVADOS 3beta, and we have updated the naming in the text to indicate this fact.

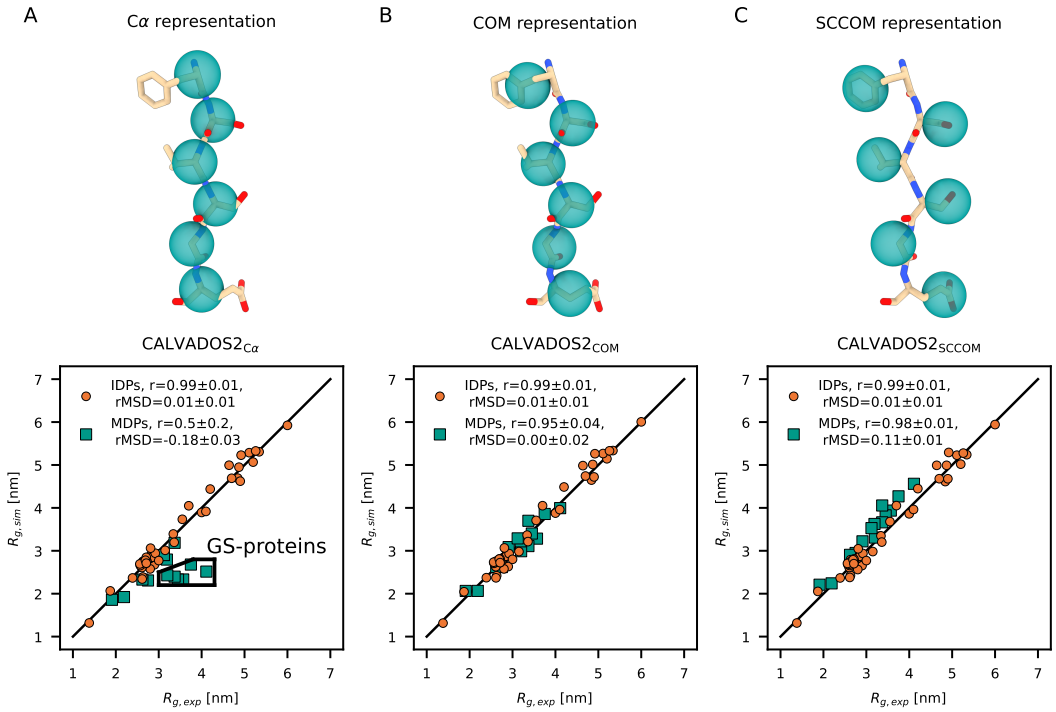

Figure 1. Simulations of MDPs and IDPs using a C_α representation, COM representation or side-chain centre-of-mass (SCCOM) representation. (A) Location of the interaction sites in a β-sheet when using a C_α representation (top), and comparison between simulated and experimental R_g values for IDPs (orange) and MDPs (green) using the CALVADOS2$_{C\alpha}$ model (CALVADOS 2 parameters and a C_α representation for both folded and disordered regions, bottom). (B) Location of the interaction sites in the same β-sheet when using a COM representation (top), and comparison between simulated and experimental R_g values for IDPs (orange) and MDPs (green) using the CALVADOS2$_{COM}$ model (CALVADOS 2 parameters and a COM representation for the interaction sites in the folded regions, bottom). (C) Location of the interaction sites in the same β-sheet when using a SCCOM representation (top), and comparison between simulated and experimental R_g values for IDPs (orange) and MDPs (green) using the CALVADOS2$_{SCCOM}$ model (CALVADOS 2 parameters and a SCCOM representation for the interaction sites in the folded regions, bottom). Pearson correlation coefficients (r) and relative mean signed deviation 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. The black lines indicate $y = x$.

Results

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 and Lindorff-Larsen, 2023***) and a C_α representation to run simulations of 56 IDPs and 14 MDPs (***Table S1***, ***Table S2***, ***Table 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 1***A, see Methods). 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 and 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 for IDPs (***Figure 1***A). In contrast, we found more substantial differences between experimental and calculated values of R_g for several MDPs (***Figure 1***A). In particular, we found that the R_g was underestimated for several MDPs including a series of two fluorescent proteins connected by Gly-Ser linkers of different lengths (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 1***A).

We 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 1***A); 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 1***B) 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 centred on the C$_\alpha$ 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 1***A and B). In contrast, simulations of the MDPs with CALVADOS2$_{COM}$ were in substantially better agreement with experiments than simulations with CALVADOS2$_{C\alpha}$ as evidenced e.g. by an increase in Pearson correlation coefficient from 0.5 to 0.95 and an increase in rMSD from -18% to 0% (***Figure 1***A and B). In addition to the COM representation, we also examined whether a side-chain centre-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 1***C). 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 1***B and C). Therefore, we proceeded by exploring 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 1***A). In the energy maps, we see evidence of substantial inter-domain interactions between residue 140–230 of one fluorescent protein and residue 340–440 of the other (***Figure 2***A). In contrast, these domain-domain interactions are not observed when simulating with COM representation (***Figure 2***B). 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 2***C).

Optimizing CALVADOS using a centre-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 et al., 2021b***) to optimize the λ stickiness parameters of the CALVADOS model targeting simultaneously SAXS and NMR data on 56 IDPs and 14 MDPs (***Table S1***, ***Table S2***, ***Table S3***). In these simulations we used the COM representation of the folded domains and we thus term the final model CALVADOS3beta$_{COM}$[2] to represent both the force field and the COM representation of the folded regions. The resulting λ values in CALVADOS3beta$_{COM}$ are similar to those in CALVADOS 2 (***Figure 3*** and ***Figure S1***). We found that simulations of IDPs with CALVADOS3beta$_{COM}$ and CALVADOS2$_{COM}$ gave similar agreement to SAXS experiments. Likewise, we find a similar agreement for the MDPs (***Figure 1***B, ***Figure 3***B and C).

Having optimized λ, we validated the CALVADOS3beta$_{COM}$ model on 25 IDPs and 9 MDPs (***Table S4***, ***Table 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 CALVADOS3beta$_{COM}$) with CALVADOS3beta$_{COM}$ performing comparably (***Figure 4***A–C). (Since the COM representation is only applied to the folded domains, these are all C$_\alpha$ representations and for IDPs CALVADOS2$_{C\alpha}$ CALVADOS2$_{COM}$ are therefore the same models). In contrast, for MDPs we found that CALVADOS3beta$_{COM}$ and CALVADOS2$_{COM}$ perform substantially better than CALVADOS2$_{C\alpha}$ (***Figure 4***A-C). Our validation results thus show that the CALVADOS3beta$_{COM}$ model gives improved agreement for simulations of MDPs while retaining the accuracy for IDPs with CALVADOS2$_{C\alpha}$. Across the 34 independent test proteins we find $\langle \chi^2_{R_g} \rangle$ values of 49, 22 and 15 for CALVADOS2$_{C\alpha}$, CALVADOS2$_{COM}$ and CALVADOS3beta$_{COM}$, respectively (***Figure 4***D), and both CALVADOS2$_{COM}$ and CALVADOS3beta$_{COM}$ have essentially no bias (rMSD≈0; ***Figure 4***B,C).

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 (***Dignon et al., 2018***; ***Tesei et al., 2021b***; ***Joseph et al., 2021***; ***Regy et al., 2021***; ***Dannenhoffer-Lafage and Best, 2021***; ***Wessén et al., 2022***; ***Tesei and Lindorff-Larsen, 2023***; ***Valdes-Garcia et al., 2023***). In some cases, these models have also been used to study phase separation of proteins that contain a mixture of folded and disordered regions (***Dignon et al., 2018***; ***Conicella et al., 2020***; ***Her et al., 2022***). We therefore examined whether the CALVADOS3beta$_{COM}$ model could be used to study phase separation of both

[2] In a previous version of this preprint (available at https://www.biorxiv.org/content/10.1101/2024.02.03.578735v1) we had termed this model CALVADOS3$_{COM}$. We here rename this model to CALVADOS3beta$_{COM}$ to indicate that it is slightly different from the final CALVADOS3$_{COM}$ model that will be published in a journal. All data and figures in this preprint refer to results with CALVADOS3beta$_{COM}$, and we have updated the naming in the text to indicate this fact.

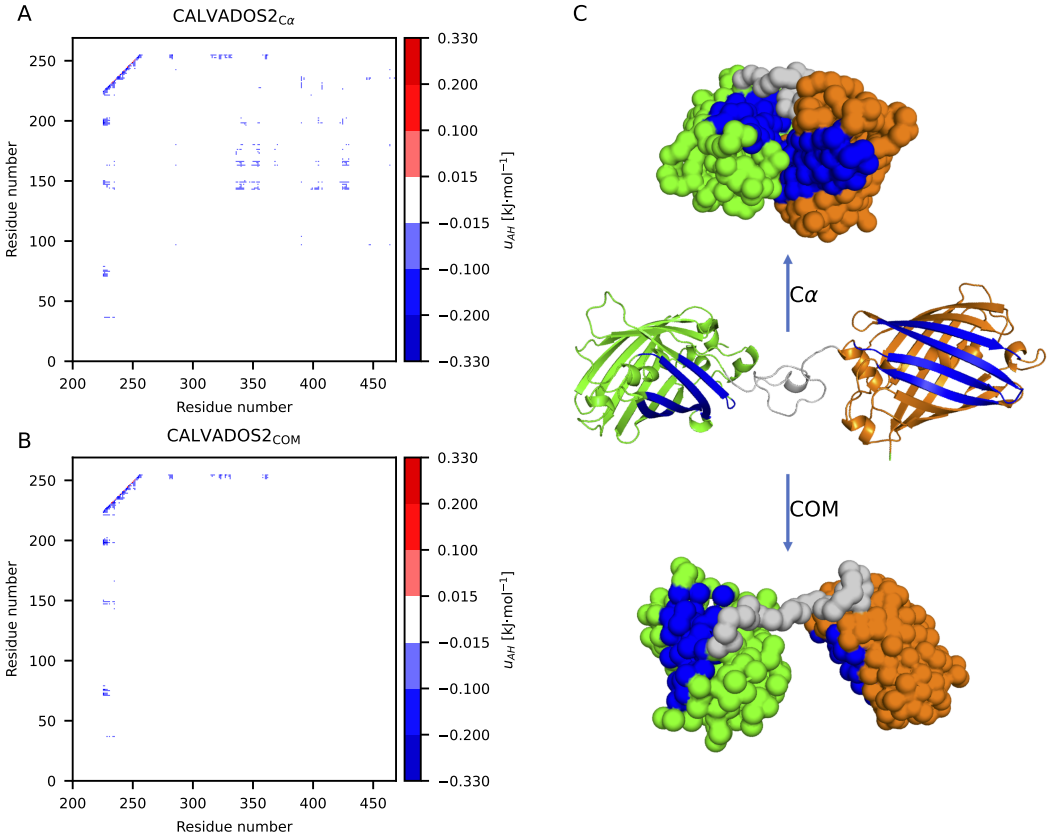

Figure 2. Energy calculations reveal substantial inter-domain interactions. We calculated an interaction energy map (of the Ashbaugh-Hatch term in the force field) from simulations using (A) the CALVADOS2$_{C\alpha}$ model and (B) the CALVADOS2$_{COM}$ model. We show only a subset of the map representing interactions between the first (residues 1–226) and second (residues 256–470) folded domains. (C) Examples of structures of GS0 with the same R_g as the average over simulations using CALVADOS2$_{C\alpha}$ (top) and CALVADOS2$_{COM}$ (bottom). 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 (grey). The regions that interact strongly in the CALVADOS2$_{C\alpha}$ simulations are coloured blue.

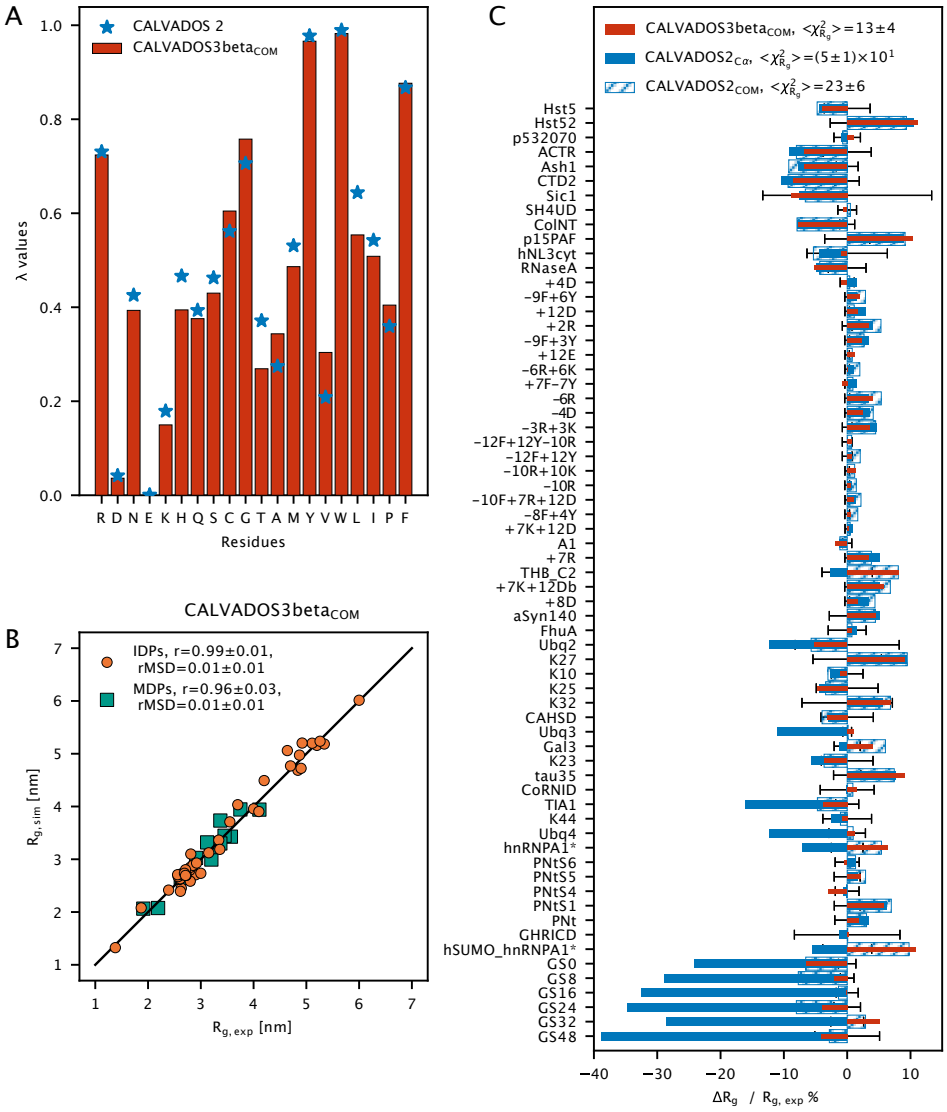

Figure 3. Optimizing the λ parameters using a COM representation for folded domains. (A) Comparison between λ values from CALVADOS 2 (blue stars) and CALVADOS3beta$_{COM}$ (red). (B) Comparison between simulated and experimental R_g values for IDPs (orange) and MDPs (green) using CALVADOS3beta$_{COM}$. Pearson correlation coefficients (r) and rMSD are reported in the legend. The black line indicates $y = x$. (C) Relative difference between experimental and simulated R_g values from CALVADOS3beta$_{COM}$ (red), CALVADOS2$_{C\alpha}$ (blue) and CALVADOS2$_{COM}$ (blue hatched). $\langle\chi^2_{R_g}\rangle$ values across IDPs and MDPs are reported in the legend. Error bars show the experimental error divided by $R_{g,exp}$.

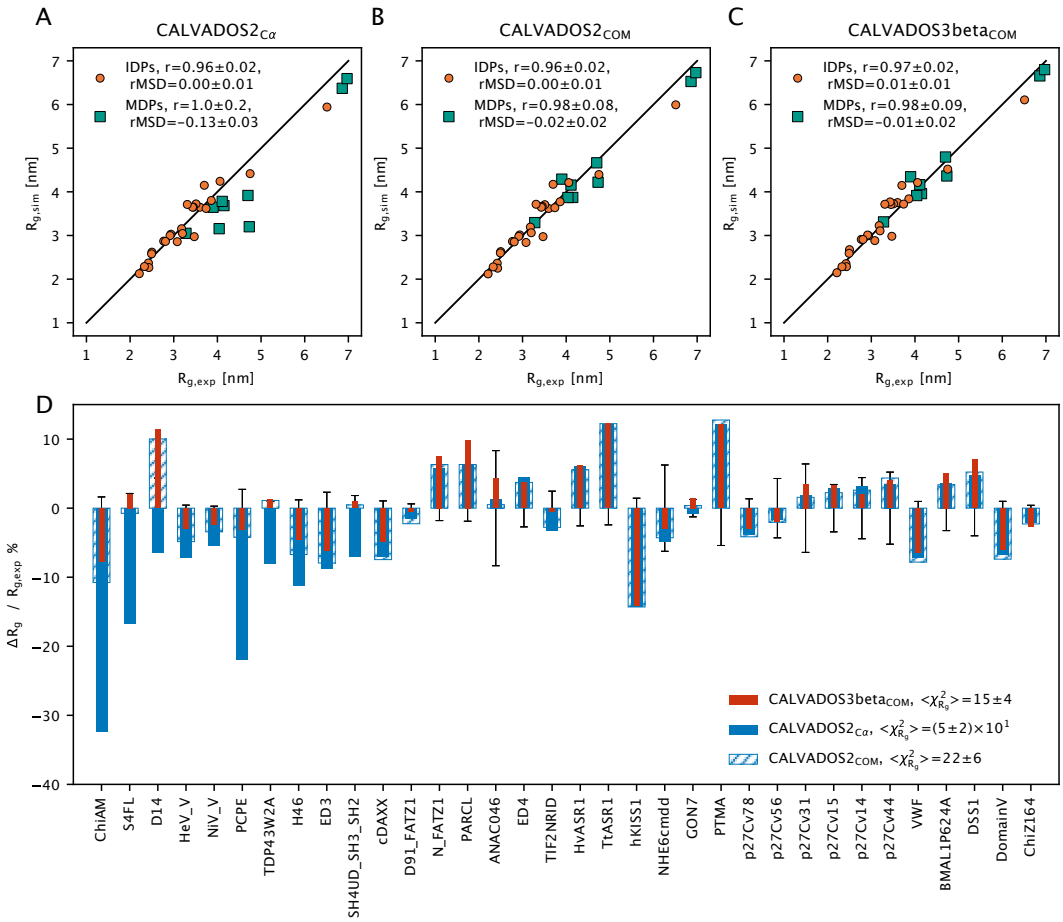

Figure 4. Validation of the CALVADOS3beta$_{COM}$ model using proteins that were not used during training. Comparison of simulated and experimental R_g values on a validation set using (A) CALVADOS2$_{C\alpha}$, (B) CALVADOS2$_{COM}$ and (C) CALVADOS3beta$_{COM}$. Pearson correlation coefficients (r) and rMSD are reported in the legend. The black lines indicate $y = x$. (D) Relative difference between experimental and simulated R_g values from CALVADOS3beta$_{COM}$ (red), CALVADOS2$_{C\alpha}$ (blue) and CALVADOS2$_{COM}$ (blue hatched). $\langle \chi^2_{R_g} \rangle$ values across IDPs and MDPs are reported in the legend. Error bars show the experimental error divided by $R_{g,exp}$.

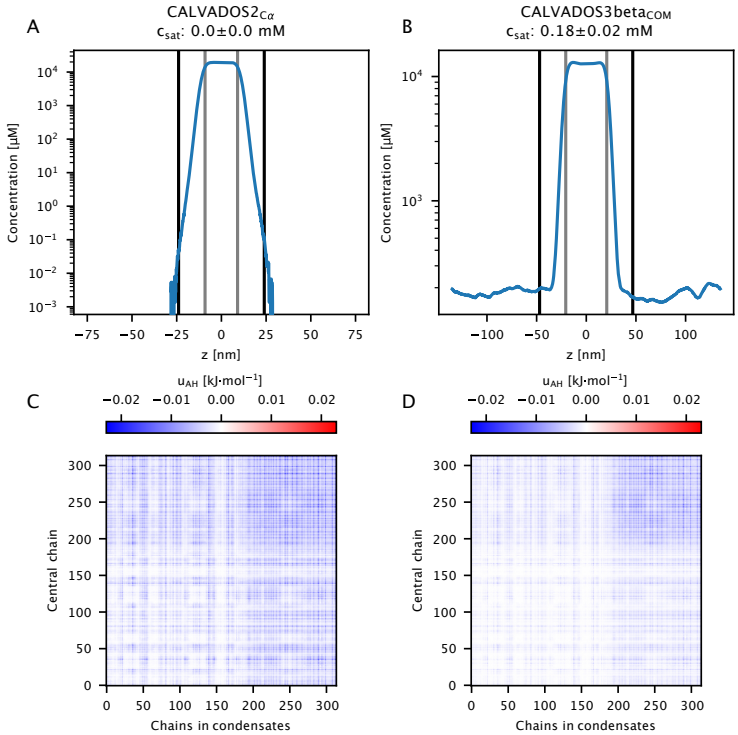

Figure 5. Phase coexistence simulations of hnRNPA1* using CALVADOS2$_{C\alpha}$ (left) and CALVADOS3beta$_{COM}$ (right). Simulations were performed at 293 K and an ionic strength of 0.15 M. Equilibrium density profile of hnRNPA1* using (A) CALVADOS2$_{C\alpha}$ and (B) CALVADOS3beta$_{COM}$. c_{sat} calculated from density profiles are 0 mM and 0.18 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) CALVADOS3beta$_{COM}$.

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 32 IDPs and found that simulations with CALVADOS3beta$_{COM}$ gave an agreement with experimental values of c_{sat} that is comparable to that of CALVADOS2$_{C\alpha}$ (***Table S6***, ***Figure S2***, ***Figure S3***).

We then proceeded to use CALVADOS3beta$_{COM}$ 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 multi-domain proteins with experimental estimates of c_{sat} (***Table S7***; ***Wang et al.*** (***2018***); ***Martin et al.*** (***2021b***)). 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 5***A). In contrast, simulations using CALVADOS3beta$_{COM}$ resulted in a lower propensity to phase separate and a calculated value of c_{sat}=0.18±0.02 mM that is closer to experiments (***Figure 5***B). To understand the origin of these differences we calculated an interaction energy map 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 (***Molliex et al., 2015***; ***Martin et al., 2021b***), and we indeed found evidence for substantial LCD-LCD interactions in the dense phases in simulations with both CALVADOS2$_{C\alpha}$ (***Figure 5***C) and CALVADOS3beta$_{COM}$ (***Figure 5***D). 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 CALVADOS3beta$_{COM}$ these interactions were much weaker, presumably explaining the increase of c_{sat} in these simulations.

Having demonstrated that CALVADOS3beta$_{COM}$ provides a more accurate description of the phase behaviour 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 (***Figure 6***, ***Figure S4***). 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 use of the COM representation in CALVADOS3beta$_{COM}$ decreases the protein-protein interactions, and thus substantially improves the agreement with experiments, though differences remain.

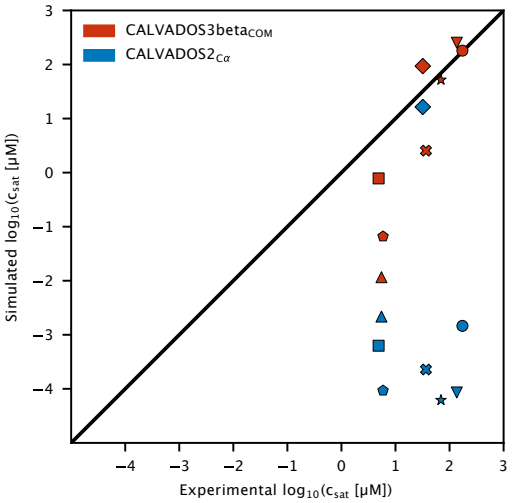

Figure 6. Comparison between simulated and experimental c_{sat} values for MDPs using the CALVADOS3beta$_{COM}$ model (red) and CALVADOS2$_{C\alpha}$ (blue). The simulated proteins are hnRNPA1* (circle), hSUMO_hnRNPA1* (downward triangle), FL_FUS (upward triangle), GFP_FUS (square), SNAP_FUS (pentagon), SNAP_FUS_PLDY2F_RBDR2K (star), SNAP_FUS_PLDY2F (x symbol) and FUS_PLDY2F_RBDR2K (diamond). The black line indicates $y = x$.

Discussion

In this work, we found that simulations with the CALVADOS2$_{C\alpha}$ model, previously shown to represent single-chain and multi-chain properties of IDPs, underestimated the radius of gyration of MDPs. Changing the CG mapping method from C$_\alpha$ to COM substantially improved the agreement with experimental data. We reoptimized the 'stickiness' parameters in the context of a COM-based model based on experimental data for both IDPs and MDPs. The resulting CALVADOS3beta$_{COM}$ 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$_\alpha$-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 (CALVADOS3beta$_{COM}$) 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. A summary of the training data and other properties of the CALVADOS models is shown in the supplement (***Table S8***). Together, our results show that the new parameter set and the centre-of-mass representation (CALVADOS3beta$_{COM}$) 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 3beta, with the implicit notion that this model is used with centre-of-mass representation of residues within folded regions.

When simulating MDPs with CALVADOS 3beta 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 and Sander, 1994***; ***Lau et al., 2023***) and might be combined with AlphaFold to set the harmonic restraints (***Jussupow and Kaila, 2023***).

Despite these current limitations, we envision that the CALVADOS 3beta model will enable detailed studies of the interactions within and between multi-domain proteins, 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., 2023***).

Methods

Description of the model

We modelled each amino acid by one bead. We generated C$_\alpha$-beads for IDPs and assigned C$_\alpha$ atom coordinates to bead positions for IDRs in multi-domain proteins according to their modelled or experimental structures (see below). For structured domains, we used the following rules for the different representations: we placed each bead position at the C$_\alpha$ atom (C$_\alpha$ representation), or the centre of mass calculated for all the atoms in a residue (COM representation), or the centre of mass calculated for only side chain atoms of residues (SCCOM represen-

tation). The MDP-CALVADOS 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,

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

with force constant $k = 8033$ kJ·mol^{-1}·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

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

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 $\epsilon = 0.8368$ kJ·mol^{-1} and $r_c = 2.2$ or 2 nm. Similar to previous work (***Tesei and Lindorff-Larsen, 2023***) we use $r_c = 2.2$ nm during the optimization of CALVADOS3beta$_{\text{COM}}$, and use 2 nm during validation and application (***Tesei and 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 (***Kim and Hummer, 2008***). λ values are treated as free parameters and optimized iteratively through a Bayesian parameter-learning procedure as described previously (***Tesei et al., 2021b***; ***Tesei and Lindorff-Larsen, 2023***) to minimize the differences in the simulated and experimental R_g and PRE data.

The Debye-Hückel potential is described by

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

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 and $B(\epsilon_r)$ is the Bjerrum length, ϵ_0 is 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 modelled by the following empirical relationship (***Akerlof and Oshry, 1950***):

$$\epsilon_r(T) = \frac{5321}{T} + 233.76 - 0.9297\times T + 1.417\times10^{-3}\times T^2 - 8.292\times10^{-7}\times T^3. \tag{5}$$

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

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

Here, the force constant k_d is 700 kJ·mol^{-1}·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 S9***). 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 would not fluctuate substantially (***Figure S5***). Non-bonded interactions (Ashbaugh-Hatch and Debye-Hückel potential) are excluded for the restrained pairs.

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., 2023***) or were modelled 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 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 $6.3\sim 77.7$ ns depending on the sequence length of the disordered regions (***Tesei and Lindorff-Larsen, 2023***; ***Tesei et al., 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. 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 1.47 nm interval. 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 centre 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 0.15 µs 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 and Lindorff-Larsen, 2023***).

Parameter optimization

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

$$\mathcal{L}(\lambda) = \langle \chi^2_{R_g} \rangle + \eta \langle \chi^2_{\mathrm{PRE}} \rangle - \theta \ln(P(\lambda)) \tag{7}$$

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

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

and

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

Here $P(\lambda)$ is a statistical prior of λ (***Tesei et al., 2021b***; ***Tesei and Lindorff-Larsen, 2023***), σ^{exp} is the error on the experimental values, Y is PRE data, either $I_{\mathrm{para}}/I_{\mathrm{dia}}$ or Γ_2 is calculated using the rotamer library approach implemented in DEER-PREdict (***Tesei et al., 2021a***), 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 to avoid overfitting. The coefficients are set to η = 0.1 and θ = 0.08. λ is not allowed to be negative but could be larger than 1.0 during optimization to explore the parameter space better.

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 and Lindorff-Larsen, 2023***; ***Thomasen et al., 2023***). 51 IDPs and 14 MDPs in this training set were used for fitting against experimental SAXS R_g data and 5 IDPs were used for fitting against experimental PRE data (***Table S1***, ***Table S2***, ***Table 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. 12 IDPs in this validation set were from our previous work and the rest (13 IDPs and 10 MDPs) were newly collected experimental R_g data in this work (***Table S4***, ***Table S5***). We also collected eight MDPs with measured values of c_{sat} to examine the accuracy of the phase behaviour simulated with the models presented in this work (***Table S7***).

The optimization procedure went through several cycles until convergence of the final total cost ($|\Delta\mathcal{L}| < 1$, $\Delta\mathcal{L}$ is the difference of final total cost between the current and previous cycle, ***Equation 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 all 0.5), 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 $w_i = \exp(-[U(r_i, \lambda_k) - U(r_i, \lambda_0)]/k_\mathrm{B}T)$, 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 $\phi_\mathrm{eff} = \exp[-\sum_i^{N_\mathrm{frames}} w_i \log(w_i \times N_\mathrm{frames})]$; if $\phi_\mathrm{eff} \geq 0.6$, trial λ_k acceptance probability is determined by the Metropolis criterion, $\min\{1, \exp\left(\frac{\mathcal{L}(\lambda_{k-1}) - \mathcal{L}(\lambda_k)}{\xi_k}\right)\}$, 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 4 cycles. Therefore, we used the resulting λ values from the fourth cycle as the final parameter set. We ran one additional optimization cycle to confirm the convergence of the training.

Data and software availability

Scripts and data to reproduce the work are available via https://github.com/KULL-Centre/_2024_Cao_CALVADOSCOM. We note that both the CALVADOS 3beta (referring to work described in this preprint) and CALVADOS 3 parameters are available from that link.

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).

References

Ahmed MC, Skaanning LK, Jussupow A, Newcombe EA, Kragelund BB, Camilloni C, Langkilde AE, Lindorff-Larsen K. Refinement of α-synuclein ensembles against SAXS data: Comparison of force fields and methods. Frontiers in molecular biosciences. 2021; 8:654333.

Akerlof G, Oshry H. The dielectric constant of water at high temperatures and in equilibrium with its vapor. Journal of the American Chemical Society. 1950; 72(7):2844–2847.

Alshareedah I, Borcherds W, Cohen S, Farag M, Singh A, Bremer A, Pappu R, Mittag T, Banerjee P. Sequence-specific interactions determine viscoelastic moduli and aging dynamics of protein condensates. . 2023; .

del Amo-Maestro L, Sagar A, Pompach P, Goulas T, Scavenius C, Ferrero DS, Castrillo-Briceño M, Taulés M, Enghild JJ, Bernadó P, et al. An integrative structural biology analysis of von Willebrand factor binding and processing by ADAMTS-13 in solution. Journal of Molecular Biology. 2021; 433(13):166954.

Arrondel C, Missoury S, Snoek R, Patat J, Menara G, Collinet B, Liger D, Durand D, Gribouval O, Boyer O, et al. Defects in t6A tRNA modification due to GON7 and YRDC mutations lead to Galloway-Mowat syndrome. Nature communications. 2019; 10(1):3967.

Ashbaugh HS, Hatch HW. Natively unfolded protein stability as a coil-to-globule transition in charge/hydropathy space. Journal of the American Chemical Society. 2008; 130(29):9536–9542.

Benayad Z, von Bülow S, Stelzl LS, Hummer G. Simulation of FUS protein condensates with an adapted coarse-grained model. Journal of Chemical Theory and Computation. 2020; 17(1):525–537.

Bereau T, Deserno M. Generic coarse-grained model for protein folding and aggregation. The Journal of chemical physics. 2009; 130(23):06B621.

Bernocco S, Steiglitz BM, Svergun DI, Petoukhov MV, Ruggiero F, Ricard-Blum S, Ebel C, Geourjon C, Deléage G, Font B, et al. Low resolution structure determination shows procollagen C-proteinase enhancer to be an elongated multidomain glycoprotein. Journal of Biological Chemistry. 2003; 278(9):7199–7205.

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

Borges-Araújo L, Patmanidis I, Singh AP, Santos LH, Sieradzan AK, Vanni S, Czaplewski C, Pantano S, Shinoda W, Monticelli L, et al. Pragmatic Coarse-Graining of Proteins: Models and Applications. Journal of Chemical Theory and Computation. 2023; 19(20):7112–7135.

Bowman MA, Riback JA, Rodriguez A, Guo H, Li J, Sosnick TR, Clark PL. Properties of protein unfolded states suggest broad selection for expanded conformational ensembles. Proceedings of the National Academy of Sciences. 2020; 117(38):23356–23364.

Brady JP, Farber PJ, Sekhar A, Lin YH, Huang R, Bah A, Nott TJ, Chan HS, Baldwin AJ, Forman-Kay JD, et al. Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. Proceedings of the National Academy of Sciences. 2017; 114(39):E8194–E8203.

Bremer A, Farag M, Borcherds WM, Peran I, Martin EW, Pappu RV, Mittag T. Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry. 2022; 14(2):196–207.

Chan-Yao-Chong M, Deville C, Pinet L, Van Heijenoort C, Durand D, Ha-Duong T. Structural characterization of N-WASP domain V using MD simulations with NMR and SAXS data. Biophysical Journal. 2019; 116(7):1216–1227.

Conicella AE, Dignon GL, Zerze GH, Schmidt HB, D'Ordine AM, Kim YC, Rohatgi R, Ayala YM, Mittal J, Fawzi NL. TDP-43 α-helical structure tunes liquid–liquid phase separation and function. Proceedings of the National Academy of Sciences. 2020; 117(11):5883–5894.

Cordeiro TN, Sibille N, Germain P, Barthe P, Boulahtouf A, Allemand F, Bailly R, Vivat V, Ebel C, Barducci A, et al. Interplay of protein disorder in retinoic acid receptor heterodimer and its corepressor regulates gene expression. Structure. 2019; 27(8):1270–1285.

Dannenhoffer-Lafage T, Best RB. A data-driven hydrophobicity scale for predicting liquid–liquid phase separation of proteins. The Journal of Physical Chemistry B. 2021; 125(16):4046–4056.

Das RK, Huang Y, Phillips AH, Kriwacki RW, Pappu RV. Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. Proceedings of the National Academy of Sciences. 2016; 113(20):5616–5621.

De Biasio A, de Opakua AI, Cordeiro TN, Villate M, Merino N, Sibille N, Lelli M, Diercks T, Bernadó P, Blanco FJ. p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins. Biophysical journal. 2014; 106(4):865–874.

Dedmon MM, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson CM. Mapping long-range interactions in α-synuclein using spin-label NMR and ensemble molecular dynamics simulations. Journal of the American Chemical Society. 2005; 127(2):476–477.

Delaforge E, Milles S, Huang Jr, Bouvier D, Jensen MR, Sattler M, Hart DJ, Blackledge M. Investigating the role of large-scale domain dynamics in protein-protein interactions. Frontiers in Molecular Biosciences. 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 computational biology. 2018; 14(1):e1005941.

Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang LP, Simmonett AC, Harrigan MP, Stern CD, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS computational biology. 2017; 13(7):e1005659.

Elena-Real CA, Sagar A, Urbanek A, Popovic M, Morató A, Estaña A, Fournet A, Doucet C, Lund XL, Shi ZD, et al. The structure of pathogenic huntingtin exon 1 defines the bases of its aggregation propensity. Nature structural & molecular biology. 2023; 30(3):309–320.

Fagerberg E, Månsson LK, Lenton S, Skepö M. The effects of chain length on the structural properties of intrinsically disordered proteins in concentrated solutions. The Journal of Physical Chemistry B. 2020; 124(52):11843–11853.

Garg A, Orru R, Ye W, Distler U, Chojnacki JE, Köhn M, Tenzer S, Sönnichsen C, Wolf E. Structural and mechanistic insights into the interaction of the circadian transcription factor BMAL1 with the KIX domain of the CREB-binding protein. Journal of Biological Chemistry. 2019; 294(45):16604–16619.

Gibbs EB, Lu F, Portz B, Fisher MJ, Medellin BP, Laremore TN, Zhang YJ, Gilmour DS, Showalter SA. Phosphorylation induces sequence-specific conformational switches in the RNA polymerase II C-terminal domain. Nature communications. 2017; 8(1):15233.

Gomes GNW, Krzeminski M, Namini A, Martin EW, Mittag T, Head-Gordon T, Forman-Kay JD, Gradinaru CC. Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET. Journal of the American Chemical Society. 2020; 142(37):15697–15710.

Gomes T, Martin-Malpartida P, Ruiz L, Aragón E, Cordeiro TN, Macias MJ. Conformational landscape of multidomain SMAD proteins. Computational and Structural Biotechnology Journal. 2021; 19:5210–5224.

Gondelaud F, Bouakil M, Le Fèvre A, Miele AE, Chirot F, Duclos B, Liwo A, Ricard-Blum S. Extended disorder at the cell surface: The conformational landscape of the ectodomains of syndecans. Matrix Biology Plus. 2021; 12:100081.

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: Structure, Function, and Bioinformatics. 2010; 78(5):1266–1281.

Gurumoorthy V, Shrestha UR, Zhang Q, Pingali SV, Boder ET, Urban VS, Smith JC, Petridis L, O'Neill H. Disordered Domain Shifts the Conformational Ensemble of the Folded Regulatory Domain of the Multidomain Oncoprotein c-Src. Biomacromolecules. 2023; 24(2):714–723.

Hajizadeh NR, Pieprzyk J, Skopintsev P, Flayhan A, Svergun DI, Löw C. Probing the architecture of a multi-PDZ domain protein: Structure of PDZK1 in solution. Structure. 2018; 26(11):1522–1533.

Hamdi K, Salladini E, O'Brien DP, Brier S, Chenal A, Yacoubi I, Longhi S. Structural disorder and induced folding within two cereal, ABA stress and ripening (ASR) proteins. Scientific reports. 2017; 7(1):15544.

Han JH, Batey S, Nickson AA, Teichmann SA, Clarke J. The folding and evolution of multidomain proteins. Nature Reviews Molecular Cell Biology. 2007; 8(4):319–330.

Heo L, Feig M. One particle per residue is sufficient to describe all-atom protein structures. bioRxiv. 2023; p. 2023–05.

Her C, Phan TM, Jovic N, Kapoor U, Ackermann BE, Rizuan A, Kim YC, Mittal J, Debelouchina GT. Molecular interactions underlying the phase separation of HP1α: role of phosphorylation, ligand and nucleic acid binding. Nucleic Acids Research. 2022; 50(22):12702–12722.

Hesgrove CS, Nguyen KH, Biswas S, Childs CA, Shraddha K, Medina BX, Alvarado V, Yu F, Sukenik S, Malferrari M, et al. Tardigrade CAHS Proteins Act as Molecular Swiss Army Knives to Mediate Desiccation Tolerance Through Multiple Mechanisms. bioRxiv. 2021; p. 2021–08.

Hicks A, Escobar CA, Cross TA, Zhou HX. Sequence-dependent correlated segments in the intrinsically disordered region of ChiZ. Biomolecules. 2020; 10(6):946.

Holm L, Sander C. Parser for protein folding units. Proteins: Structure, Function, and Bioinformatics. 1994; 19(3):256–268.

Hyeon C, Dima RI, Thirumalai D. Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. Structure. 2006; 14(11):1633–1645.

Jephthah S, Staby L, Kragelund B, Skepö M. Temperature dependence of intrinsically disordered proteins in simulations: What are we missing? Journal of chemical theory and computation. 2019; 15(4):2672–2683.

Jin F, Gräter F. How multisite phosphorylation impacts the conformations of intrinsically disordered proteins. PLoS computational biology. 2021; 17(5):e1008939.

Johnson CL, Solovyova AS, Hecht O, Macdonald C, Waller H, Grossmann JG, Moore GR, Lakey JH. The two-state prehensile tail of the antibacterial toxin colicin N. Biophysical Journal. 2017; 113(8):1673–1684.

Joseph JA, Reinhardt A, Aguirre A, Chew PY, Russell KO, Espinosa JR, Garaizar A, Collepardo-Guevara R. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. Nature Computational Science. 2021; 1(11):732–743.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596(7873):583–589.

Jussupow A, Kaila VR. Effective Molecular Dynamics from Neural Network-Based Structure Prediction Models. Journal of Chemical Theory and Computation. 2023; 19(7):1965–1975.

Jussupow A, Messias AC, Stehle R, Geerlof A, Solbak SM, Paissoni C, Bach A, Sattler M, Camilloni C. The dynamics of linear polyubiquitin. Science advances. 2020; 6(42):eabc3786.

Kim YC, Hummer G. Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. Journal of molecular biology. 2008; 375(5):1416–1433.

Kjaergaard M, Nørholm AB, Hendus-Altenburger R, Pedersen SF, Poulsen FM, Kragelund BB. Temperature-dependent structural changes in intrinsically disordered proteins: Formation of α–helices or loss of polyproline II? Protein Science. 2010; 19(8):1555–1564.

Krainer G, Welsh TJ, Joseph JA, Espinosa JR, Wittmann S, de Csilléry E, Sridhar A, Toprakcioglu Z, Gudiškytė G, Czekalska MA, et al. Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions. Nature communications. 2021; 12(1):1085.

Kurzbach D, Vanas A, Flamm AG, Tarnoczi N, Kontaxis G, Maltar-Strmečki N, Widder K, Hinderberger D, Konrat R. Detection of correlated conformational fluctuations in intrinsically disordered proteins through paramagnetic relaxation interference. Physical Chemistry Chemical Physics. 2016; 18(8):5753–5758.

Lau AM, Kandathil SM, Jones DT. Merizo: a rapid and accurate protein domain segmentation method using invariant point attention. Nature Communications. 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. Proceedings of the National Academy of Sciences. 2012; 109(44):17789–17794.

Lin YH, Qiu DC, Chang WH, Yeh YQ, Jeng US, Liu FT, Huang Jr. The intrinsically disordered N-terminal domain of galectin-3 dynamically mediates multisite self-association of the protein through fuzzy interactions. Journal of Biological Chemistry. 2017; 292(43):17845–17856.

Lyu C, Da Vela S, Al-Hilaly Y, Marshall KE, Thorogate R, Svergun D, Serpell LC, Pastore A, Hanger DP. The disease associated tau35 fragment has an increased propensity to aggregate compared to full-length tau. Frontiers in Molecular Biosciences. 2021; 8:779240.

Mackereth CD, Sattler M. Dynamics in multi-domain protein recognition of RNA. Current opinion in structural biology. 2012; 22(3):287–296.

Maity H, Baidya L, Reddy G. Salt-induced transitions in the conformational ensembles of intrinsically disordered proteins. The Journal of Physical Chemistry B. 2022; 126(32):5959–5971.

Martin EW, Harmon TS, Hopkins JB, Chakravarthy S, Incicco JJ, Schuck P, Soranno A, Mittag T. A multi-step nucleation process determines the kinetics of prion-like domain phase separation. Nature communications. 2021; 12(1):4513.

Martin EW, Holehouse AS, Grace CR, Hughes A, Pappu RV, Mittag T. Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation. Journal of the American Chemical Society. 2016; 138(47):15323–15335.

Martin EW, Thomasen FE, Milkovic NM, Cuneo MJ, Grace CR, Nourse A, Lindorff-Larsen K, Mittag T. Interplay of folded domains and the disordered low-complexity domain in mediating hnRNPA1 phase separation. Nucleic acids research. 2021; 49(5):2931–2945.

Mazurkewich S, Helland R, Mackenzie A, Eijsink VG, Pope PB, Brändén G, Larsbrink J. Structural insights of the enzymes from the chitin utilization locus of Flavobacterium johnsoniae. Scientific Reports. 2020; 10(1):13775.

Michie KA, Kwan AH, Tung CS, Guss JM, Trewhella J. A highly conserved yet flexible linker is part of a polymorphic protein-binding domain in myosin-binding protein C. Structure. 2016; 24(11):2000–2007.

Mittal A, Holehouse AS, Cohan MC, Pappu RV. Sequence-to-conformation relationships of disordered regions tethered to folded domains of proteins. Journal of molecular biology. 2018; 430(16):2403–2421.

Molliex A, Temirov J, Lee J, Coughlin M, Kanagaraj AP, Kim HJ, Mittag T, Taylor JP. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. Cell. 2015; 163(1):123–133.

Monahan Z, Ryan VH, Janke AM, Burke KA, Rhoads SN, Zerze GH, O'Meally R, Dignon GL, Conicella AE, Zheng W, et al. Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. The EMBO journal. 2017; 36(20):2951–2967.

Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink SJ. The MARTINI coarse-grained force field: extension to proteins. Journal of chemical theory and computation. 2008; 4(5):819–834.

Moses D, Guadalupe K, Yu F, Flores E, Perez AR, McAnelly R, Shamoon NM, Kaur G, Cuevas-Zepeda E, Merg AD, et al. Structural biases in disordered proteins are prevalent in the cell. Nature Structural & Molecular Biology. 2024; p. 1–10.

Mugnai ML, Chakraborty D, Kumar A, Nguyen HT, Zeno W, Stachowiak JC, Straub JE, Thirumalai D. Sizes, conformational fluctuations, and SAXS profiles for Intrinsically Disordered Proteins. bioRxiv. 2023; p. 2023–04.

Murthy AC, Dignon GL, Kan Y, Zerze GH, Parekh SH, Mittal J, Fawzi NL. Molecular interactions underlying liquid- liquid phase separation of the FUS low-complexity domain. Nature structural & molecular biology. 2019; 26(7):637–648.

Mylonas E, Hascher A, Bernado P, Blackledge M, Mandelkow E, Svergun DI. Domain conformation of tau protein studied by solution small-angle X-ray scattering. Biochemistry. 2008; 47(39):10345–10353.

Nagai H, Kuwabara K, Carta G. Temperature dependence of the dissociation constants of several amino acids. Journal of Chemical & Engineering Data. 2008; 53(3):619–627.

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

Norgaard AB, Ferkinghoff-Borg J, Lindorff-Larsen K. Experimental parameterization of an energy function for the simulation of unfolded proteins. Biophysical journal. 2008; 94(1):182–192.

Ibáñez de Opakua A, Merino N, Villate M, Cordeiro TN, Ormaza G, Sánchez-Carbayo M, Diercks T, Bernadó P, Blanco FJ. The metastasis suppressor KISS1 is an intrinsically disordered protein slightly more extended than a random coil. PLoS One. 2017; 12(2):e0172507.

Ostendorp A, Ostendorp S, Zhou Y, Chaudron Z, Wolffram L, Rombi K, von Pein L, Falke S, Jeffries CM, Svergun DI, et al. Intrinsically disordered plant protein PARCL colocalizes with RNA in phase-separated condensates whose formation can be regulated by mutating the PLD. Journal of Biological Chemistry. 2022; 298(12).

Pappu RV, Schneller WJ, Weaver DL. Electrostatic multipole representation of a polypeptide chain: an algorithm for simulation of polypeptide properties. Journal of computational chemistry. 1996; 17(8):1033–1055.

Paz A, Zeev-Ben-Mordehai T, Lundqvist M, Sherman E, Mylonas E, Weiner L, Haran G, Svergun DI, Mulder FA, Sussman JL, et al. Biophysical characterization of the unstructured cytoplasmic domain of the human neuronal adhesion protein neuroligin 3. Biophysical Journal. 2008; 95(4):1928–1944.

Pesce F, Newcombe EA, Seiffert P, Tranchant EE, Olsen JG, Grace CR, Kragelund BB, Lindorff-Larsen K. Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins. Biophysical Journal. 2023; 122(2):310–321.

Regy RM, Thompson J, Kim YC, Mittal J. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. Protein Science. 2021; 30(7):1371–1379.

Riback JA, Bowman MA, Zmyslowski AM, Knoverek CR, Jumper JM, Hinshaw JR, Kaye EB, Freed KF, Clark PL, Sosnick TR. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. Science. 2017; 358(6360):238–241.

Ryan VH, Dignon GL, Zerze GH, Chabata CV, Silva R, Conicella AE, Amaya J, Burke KA, Mittal J, Fawzi NL. Mechanistic view of hnRNPA2 low-complexity domain structure, interactions, and phase separation altered by mutation and arginine methylation. Molecular cell. 2018; 69(3):465–479.

Ryan VH, Perdikari TM, Naik MT, Saueressig CF, Lins J, Dignon GL, Mittal J, Hart AC, Fawzi NL. Tyrosine phosphorylation regulates hnRNPA2 granule protein partitioning and reduces neurodegeneration. The EMBO Journal. 2021; 40(3):e105001.

Salladini E, Delauzun V, Longhi S. The Henipavirus V protein is a prevalently unfolded protein with a zinc-finger domain involved in binding to DDB1. Molecular BioSystems. 2017; 13(11):2254–2267.

Schmit JD, Bouchard JJ, Martin EW, Mittag T. Protein network structure enables switching between liquid and gel states. Journal of the American Chemical Society. 2019; 142(2):874–883.

Schuster BS, Dignon GL, Tang WS, Kelley FM, Ranganath AK, Jahnke CN, Simpkins AG, Regy RM, Hammer DA, Good MC, et al. Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. Proceedings of the National Academy of Sciences. 2020; 117(21):11421–11431.

Seiffert P, Bugge K, Nygaard M, Haxholm GW, Martinsen JH, Pedersen MN, Arleth L, Boomsma W, Kragelund BB. Orchestration of signaling by structural disorder in class 1 cytokine receptors. Cell Communication and Signaling. 2020; 18:1–30.

Sekiyama N, Takaba K, Maki-Yonekura S, Akagi Ki, Ohtani Y, Imamura K, Terakawa T, Yamashita K, Inaoka D, Yonekura K, et al. ALS mutations in the TIA-1 prion-like domain trigger highly condensed pathogenic structures. Proceedings of the National Academy of Sciences. 2022; 119(38):e2122523119.

Senicourt L, Le Maire A, Allemand F, Carvalho JE, Guee L, Germain P, Schubert M, Bernadó P, Bourguet W, Sibille N. Structural insights into the interaction of the intrinsically disordered co-activator TIF2 with retinoic acid receptor heterodimer (RXR/RAR). Journal of molecular biology. 2021; 433(9):166899.

Shrestha UR, Juneja P, Zhang Q, Gurumoorthy V, Borreguero JM, Urban V, Cheng X, Pingali SV, Smith JC, O'Neill HM, et al. Generation of the configurational ensemble of an intrinsically disordered protein from unbiased molecular dynamics simulation. Proceedings of the National Academy of Sciences. 2019; 116(41):20446–20452.

Sieradzan AK, Czaplewski C, Krupa P, Mozolewska MA, Karczyńska AS, Lipska AG, Lubecka EA, Gołaś E, Wirecki T, Makowski M, et al. Modeling the structure, dynamics, and transformations of proteins with the UNRES force field. Protein folding: Methods and protocols. 2022; p. 399–416.

Sonntag M, Jagtap PKA, Simon B, Appavou MS, Geerlof A, Stehle R, Gabel F, Hennig J, Sattler M. Segmental, Domain-Selective Perdeuteration and Small-Angle Neutron Scattering for Structural Analysis of Multi-Domain Proteins. Angewandte Chemie. 2017; 129(32):9450–9453.

Souza PC, Alessandri R, Barnoud J, Thallmair S, Faustino I, Grünewald F, Patmanidis I, Abdizadeh H, Bruininks BM, Wassenaar TA, et al. Martini 3: a general purpose force field for coarse-grained molecular dynamics. Nature methods. 2021; 18(4):382–388.

Sponga A, Arolas JL, Schwarz TC, Jeffries CM, Rodriguez Chamorro A, Kostan J, Ghisleni A, Drepper F, Polyansky A, De Almeida Ribeiro E, et al. Order from disorder in the sarcomere: FATZ forms a fuzzy but tight complex and phase-separated condensates with α-actinin. Science advances. 2021; 7(22):eabg7653.

Tan C, Niitsu A, Sugita Y. Highly Charged Proteins and Their Repulsive Interactions Antagonize Biomolecular Condensation. JACS Au. 2023; 3(3):834–848.

Taneja I, Holehouse AS. Folded domain charge properties influence the conformational behavior of disordered tails. Current Research in Structural Biology. 2021; 3:216–228.

Tesei G, Lindorff-Larsen K. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. Open Research 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 Computational Biology. 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. Proceedings of the National Academy of Sciences. 2021; 118(44):e2111696118.

Tesei G, Trolle AI, Jonsson N, Betz J, Pesce F, Johansson KE, Lindorff-Larsen K. Conformational ensembles of the human intrinsically disordered proteome: Bridging chain compaction with function and sequence conservation. bioRxiv. 2023; p. 2023–05.

Thomasen FE, Lindorff-Larsen K. Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. Biochemical Society Transactions. 2022; 50(1):541–554.

Thomasen FE, Pesce F, Roesgaard MA, Tesei G, Lindorff-Larsen K. Improving Martini 3 for disordered and multidomain proteins. Journal of Chemical Theory and Computation. 2022; 18(4):2033–2041.

Thomasen FE, Skaalum T, Kumar A, Srinivasan S, Vanni S, Lindorff-Larsen K. Recalibration of protein interactions in Martini 3. bioRxiv. 2023; p. 2023–05.

Valdes-Garcia G, Heo L, Lapidus LJ, Feig M. Modeling concentration-dependent phase separation processes involving peptides and RNA via residue-based coarse-graining. Journal of Chemical Theory and Computation. 2023; 19(2):669–678.

Van Der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, et al. Classification of intrinsically disordered regions and proteins. Chemical reviews. 2014; 114(13):6589–6631.

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic acids research. 2022; 50(D1):D439–D444.

Wang J, Choi JM, Holehouse AS, Lee HO, Zhang X, Jahnel M, Maharana S, Lemaitre R, Pozniakovsky A, Drechsel D, et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. Cell. 2018; 174(3):688–699.

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. The Journal of Physical Chemistry B. 2022; 126(45):9222–9245.

Wright GS, Watanabe TF, Amporndanai K, Plotkin SS, Cashman NR, Antonyuk SV, Hasnain SS. Purification and structural characterization of aggregation-prone human TDP-43 involved in neurodegenerative diseases. Iscience. 2020; 23(6):101159.

Yamada T, Miyazaki Y, Harada S, Kumar A, Vanni S, Shinoda W. Improved protein model in SPICA force field. Journal of Chemical Theory and Computation. 2023; 19(23):8967–8977.

Yu F, Sukenik S. Structural preferences shape the entropic force of disordered protein ensembles. The Journal of Physical Chemistry B. 2023; .

Zhang Y, Li S, Gong X, Chen J. Toward Accurate Simulation of Coupling between Protein Secondary Structure and Phase Separation. Journal of the American Chemical Society. 2023; .

Zhang Y, Liu X, Chen J. Toward accurate coarse-grained simulations of disordered proteins and their dynamic interactions. Journal of Chemical Information and Modeling. 2022; 62(18):4523–4536.

Zhao J, Blayney A, Liu X, Gandy L, Jin W, Yan L, Ha JH, Canning AJ, Connelly M, Yang C, et al. EGCG binds intrinsically disordered N-terminal domain of p53 and disrupts p53-MDM2 interaction. Nature communications. 2021; 12(1):986.

Zheng W, Dignon GL, Jovic N, Xu X, Regy RM, Fawzi NL, Kim YC, Best RB, Mittal J. Molecular details of protein condensates probed by microsecond long atomistic simulations. The Journal of Physical Chemistry B. 2020; 124(51):11671–11679.

Table S1. Solution conditions and experimental radii of gyration of IDPs included in the training set for the Bayesian parameter-learning procedure.

Protein	N	R_g ± Err [nm]	T [K]	c_s [M]	pH	Ref.
Hst5	24	1.38 ± 0.05	293.00	0.150	7.5	(***Jephthah et al., 2019***)
Hst52	48	1.87 ± 0.05	298.00	0.150	7.0	(***Fagerberg et al., 2020***)
p532070	62	2.39 ± 0.05	277.00	0.100	7.0	(***Zhao et al., 2021***)
ACTR	71	2.63 ± 0.1	278.00	0.200	7.4	(***Kjaergaard et al., 2010***)
Ash1	81	2.9 ± 0.05	293.00	0.150	7.5	(***Martin et al., 2016***; ***Jin and Gräter, 2021***)
CTD2	83	2.614 ± 0.05	293.00	0.120	7.5	(***Jin and Gräter, 2021***)(***Gibbs et al., 2017***)]
Sic1	92	3.0 ± 0.4	293.00	0.200	7.5	(***Gomes et al., 2020***)
SH4UD	95	2.71 ± 0.04	293.15	0.216	8.0	(***Shrestha et al., 2019***)
ColNT	98	2.8 ± 0.033	277.00	0.433	7.6	(***Johnson et al., 2017***)
p15PAF	111	2.81 ± 0.1	298.00	0.150	7.0	(***De Biasio et al., 2014***)
hNL3cyt	119	3.15 ± 0.2	293.00	0.300	8.5	(***Paz et al., 2008***)
RNaseA	124	3.36 ± 0.1	298.00	0.150	7.5	(***Riback et al., 2017***)
+4D	137	2.72 ± 0.03	298.00	0.150	7.0	(***Bremer et al., 2022***)
-3R+3K	137	2.63 ± 0.02	298.00	0.150	7.0	(***Bremer et al., 2022***)
-6R+6K	137	2.79 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
-10R+10K	137	2.85 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
-4D	137	2.64 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
+2R	137	2.62 ± 0.02	298.00	0.150	7.0	(***Bremer et al., 2022***)
+12D	137	2.8 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
+12E	137	2.85 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
+7K+12D	137	2.92 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
+7R	137	2.71 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
-12F+12Y-10R	137	2.61 ± 0.02	298.00	0.150	7.0	(***Bremer et al., 2022***)
-10F+7R+12D	137	2.86 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
+8D	137	2.69 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
+7K+12Db	137	2.56 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
-9F+6Y	137	2.66 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
-10R	137	2.67 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
-9F+3Y	137	2.68 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
-8F+4Y	137	2.71 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
+7F-7Y	137	2.72 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
-12F+12Y	137	2.6 ± 0.02	298.00	0.150	7.0	(***Bremer et al., 2022***)
A1	137	2.76 ± 0.02	298.00	0.150	7.0	(***Bremer et al., 2022***)
-6R	137	2.57 ± 0.01	298.00	0.150	7.0	(***Bremer et al., 2022***)
aSyn140	140	3.55 ± 0.1	293.00	0.200	7.4	(***Ahmed et al., 2021***)
FhuA	144	3.34 ± 0.1	298.00	0.150	7.5	(***Riback et al., 2017***)
K27	167	3.7 ± 0.2	288.00	0.150	7.4	(***Mylonas et al., 2008***)
K10	168	4.0 ± 0.1	288.00	0.150	7.4	(***Mylonas et al., 2008***)
K25	185	4.1 ± 0.2	288.00	0.150	7.4	(***Mylonas et al., 2008***)
K32	198	4.2 ± 0.3	288.00	0.150	7.4	(***Mylonas et al., 2008***)
CAHSD	227	4.84 ± 0.2	293.00	0.070	7.0	(***Hesgrove et al., 2021***)
K23	254	4.9 ± 0.2	288.00	0.150	7.4	(***Mylonas et al., 2008***)
tau35	255	4.64 ± 0.1	293.20	0.150	7.4	(***Lyu et al., 2021***)
CoRNID	271	4.7 ± 0.2	293.15	0.192	7.5	(***Cordeiro et al., 2019***)
K44	283	5.2 ± 0.2	288.00	0.150	7.4	(***Mylonas et al., 2008***)
PNt	334	5.11 ± 0.1	298.00	0.150	7.5	(***Riback et al., 2017***; ***Bowman et al., 2020***)
PNtS1	334	4.92 ± 0.1	298.00	0.150	7.5	(***Bowman et al., 2020***)
PNtS4	334	5.34 ± 0.1	298.00	0.150	7.5	(***Bowman et al., 2020***)
PNtS5	334	4.87 ± 0.1	298.00	0.150	7.5	(***Bowman et al., 2020***)
PNtS6	334	5.26 ± 0.1	298.00	0.150	7.5	(***Bowman et al., 2020***)
GHRICD	351	6.0 ± 0.5	298.00	0.350	7.3	(***Seiffert et al., 2020***; ***Pesce et al., 2023***)

Table S2. Solution conditions and experimental PRE data included in the training set for the Bayesian parameter-learning procedure.

Proteins	N	N_{labels}	$\omega_I/2\pi[MHz]$	T [K]	c_s [M]	pH	Ref.
A2	155	2	850	298	0.005	5.5	(***Ryan et al., 2018***)
aSyn	140	5	700	283	0.200	7.4	(***Dedmon et al., 2005***)
OPN	220	10	800	298	0.150	6.5	(***Kurzbach et al., 2016***)
FUS	163	3	850	298	0.150	5.5	(***Monahan et al., 2017***)
FUS12E	164	3	850	298	0.150	5.5	(***Monahan et al., 2017***)

Table S3. Solution conditions and experimental radii of gyration of MDPs included in the training set for the Bayesian parameter-learning procedure.

Protein	N	R_g ± Err [nm]	T [K]	c_s [M]	pH	Ref.
THB_C2	137	1.91 ± 0.076	295.15	0.15	6.5	(***Michie et al., 2016***)
Ubq2	162	2.19 ± 0.18	293.00	0.33	8.0	(***Jussupow et al., 2020***)
Ubq3	228	2.62 ± 0.018	293.00	0.33	8.0	(***Jussupow et al., 2020***)
Gal3	250	2.91 ± 0.06	303.00	0.04	7.0	(***Lin et al., 2017***)
TIA1	275	2.75 ± 0.05	293.15	0.10	6.0	(***Sonntag et al., 2017***)
Ubq4	304	3.19 ± 0.092	293.00	0.33	8.0	(***Jussupow et al., 2020***)
hnRNPA1	314	3.12 ± 0.078	293.15	0.15	7.5	(***Martin et al., 2021b***)
hSUMO_hnRNPA1	433	3.37 ± 0.13	293.15	0.10	7.5	(***Martin et al., 2021b***)
GS0	470	3.2 ± 0.044	293.15	0.15	7.4	(***Moses et al., 2024***)
GS8	486	3.37 ± 0.036	293.15	0.15	7.4	(***Moses et al., 2024***)
GS16	502	3.45 ± 0.06	293.15	0.15	7.4	(***Moses et al., 2024***)
GS24	518	3.57 ± 0.075	293.15	0.15	7.4	(***Moses et al., 2024***)
GS32	534	3.75 ± 0.097	293.15	0.15	7.4	(***Moses et al., 2024***)
GS48	566	4.11 ± 0.21	293.15	0.15	7.4	(***Moses et al., 2024***)

Table S4. Solution conditions and experimental radii of gyration of IDPs included in the validation set.

Protein	N	R_g ± Err [nm]	T [K]	c_s [M]	pH	Ref.
ChiZ164	67	2.42 ± 0.01	293.00	0.0650	7.0	(***Hicks et al., 2020***)
DomainV	67	2.43 ± 0.024	288.15	0.1985	7.0	(***Chan-Yao-Chong et al., 2019***)
DSS1	71	2.5 ± 0.1	288.00	0.1700	7.4	(***Pesce et al., 2023***)
BMAL1P624A	98	2.77 ± 0.09	283.25	0.1540	7.2	(***Garg et al., 2019***)
VWF	103	3.08 ± 0.03	293.00	0.1530	7.4	(***del Amo-Maestro et al., 2021***)
p27Cv56	107	2.328 ± 0.1	293.00	0.0950	7.2	(***Das et al., 2016***)
p27Cv14	107	2.936 ± 0.13	293.00	0.0950	7.2	(***Das et al., 2016***)
p27Cv78	107	2.211 ± 0.03	293.00	0.0950	7.2	(***Das et al., 2016***)
p27Cv31	107	2.81 ± 0.18	293.00	0.0950	7.2	(***Das et al., 2016***)
p27Cv44	107	2.492 ± 0.13	293.00	0.0950	7.2	(***Das et al., 2016***)
p27Cv15	107	2.915 ± 0.1	293.00	0.0950	7.2	(***Das et al., 2016***)
PTMA	111	3.7 ± 0.2	288.00	0.1600	7.4	(***Pesce et al., 2023***)
GON7	114	3.18 ± 0.04	283.00	0.2110	6.5	(***Arrondel et al., 2019***)
NHE6cmdd	116	3.2 ± 0.2	288.00	0.1700	7.4	(***Pesce et al., 2023***)
hKISS1	120	3.47 ± 0.05	283.15	0.1590	7.0	(***Ibáñez de Opakua et al., 2017***)
TtASR1	141	3.31 ± 0.08	293.15	0.1500	7.3	(***Hamdi et al., 2017***)
HvASR1	143	3.51 ± 0.09	293.15	0.1500	7.3	(***Hamdi et al., 2017***)
TIF2NRID	150	3.74 ± 0.092	283.15	0.1750	6.8	(***Senicourt et al., 2021***)
ED4	163	4.06 ± 0.11	293.15	0.1530	7.4	(***Gondelaud et al., 2021***)
ANAC046	167	3.6 ± 0.3	298.00	0.1400	7.0	(***Pesce et al., 2023***)
PARCL	180	3.43 ± 0.065	293.15	0.1700	7.5	(***Ostendorp et al., 2022***)
N_FATZ1	191	3.45 ± 0.062	293.15	0.1920	7.5	(***Sponga et al., 2021***)
D91_FATZ1	209	4.0 ± 0.1	293.00	0.1800	7.5	(***Sponga et al., 2021***)
cDAXX	246	4.75 ± 0.05	293.00	0.1300	8.0	(***Schmit et al., 2019***)
ED3	373	6.51 ± 0.15	293.15	0.1530	7.4	(***Gondelaud et al., 2021***)

Table S5. Solution conditions and experimental radii of gyration of MDPs included in the validation set.

Protein	N	R_g ± Err [nm]	T [K]	c_s [M]	pH	Ref.
SH4UD_SH3_SH2	264	3.28 ± 0.06	293.15	0.216	8.0	(***Gurumoorthy et al., 2023***)
H46	381	4.15 ± 0.05	283.00	0.163	6.5	(***Elena-Real et al., 2023***)
TDP43W2A	415	4.11 ± 0.04	293.15	0.312	8.0	(***Wright et al., 2020***)
PCPE	424	4.04 ± 0.11	293.15	0.506	7.4	(***Bernocco et al., 2003***)
NiV_V	457	6.97 ± 0.02	293.15	0.232	8.0	(***Salladini et al., 2017***)
HeV_V	458	6.86 ± 0.03	293.15	0.232	8.0	(***Salladini et al., 2017***)
D14	483	3.9 ± 0.17	283.15	0.156	7.5	(***Hajizadeh et al., 2018***)
S4FL	552	4.7 ± 0.1	283.15	0.169	7.2	(***Gomes et al., 2021***)
ChiAM	682	4.73 ± 0.077	293.15	0.282	8.0	(***Mazurkewich et al., 2020***)

Table S6. IDPs and conditions used for slab simulations in this work.

Proteins	N	c_s [M]	pH	T [K]	$c_{\text{sat,exp}}$ [μM]	Ref.
LAF1	176	0.15	7.5	293	44.0	(***Schuster et al., 2020***)
LAF1D2130	166	0.15	7.5	293	275.0	(***Schuster et al., 2020***)
LAF1shuf	176	0.15	7.5	293	6.0	(***Schuster et al., 2020***)
A1S150	131	0.15	7.0	293	218.1	(***Martin et al., 2021a***)
A1S200	131	0.20	7.0	293	159.8	(***Martin et al., 2021a***)
A1S300	131	0.30	7.0	293	93.4	(***Martin et al., 2021a***)
A1S500	131	0.50	7.0	293	66.5	(***Martin et al., 2021a***)
-12F+12Y	137	0.15	7.0	293	60.3	(***Bremer et al., 2022***)
+4D	137	0.15	7.0	277	4.5	(***Bremer et al., 2022***)
-6R	137	0.15	7.0	277	7.1	(***Bremer et al., 2022***)
A1	137	0.15	7.0	293	102.2	(***Bremer et al., 2022***)
+2R	137	0.15	7.0	277	18.0	(***Bremer et al., 2022***)
+8D	137	0.15	7.0	277	18.7	(***Bremer et al., 2022***)
-14N+14Q	137	0.15	7.0	293	171.6	(***Bremer et al., 2022***)
-10G+10S	137	0.15	7.0	293	268.1	(***Bremer et al., 2022***)
+7F-7Y	137	0.15	7.0	293	209.0	(***Bremer et al., 2022***)
-20G+20S	137	0.15	7.0	293	469.4	(***Bremer et al., 2022***)
-23S+23T	137	0.15	7.0	293	342.2	(***Bremer et al., 2022***)
-8F+4Y	137	0.15	7.0	277	63.2	(***Bremer et al., 2022***)
-3R+3K	137	0.15	7.0	277	83.1	(***Bremer et al., 2022***)
-4D	137	0.15	7.0	277	88.8	(***Bremer et al., 2022***)
-9F+3Y	137	0.15	7.0	277	115.0	(***Bremer et al., 2022***)
+23G-23S	137	0.15	7.0	293	46.1	(***Bremer et al., 2022***)
+23G-23S+7F-7Y	137	0.15	7.0	293	194.0	(***Bremer et al., 2022***)
+23G-23S-12F+12Y	137	0.15	7.0	293	6.5	(***Bremer et al., 2022***)
-30G+30S	137	0.15	7.0	293	841.8	(***Bremer et al., 2022***)
FUS	163	0.15	5.5	297	105.0	(***Murthy et al., 2019***)
A2	155	0.01	5.5	297	15.0	(***Ryan et al., 2021***)
Ddx4WT	236	0.13	6.5	297	230.0	(***Brady et al., 2017***)
allF	137	0.15	7.0	293	250.0	(***Alshareedah et al., 2023***)
allY	137	0.15	7.0	293	85.0	(***Alshareedah et al., 2023***)
allW	137	0.15	7.0	293	1.0	(***Alshareedah et al., 2023***)

Table S7. Multi-domain proteins and conditions used for slab simulations in this work.

Proteins	N	c_s [M]	pH	T [K]	$c_{\text{sat,exp}}$ [μM]	Box [nm]	Ref.
hnRNPA1*	314	0.15	7.5	293	173.0	[20, 20, 270]	(***Martin et al., 2021b***)
hSUMO_hnRNPA1*	433	0.15	7.5	293	136.2	[25, 25, 190]	(***Martin et al., 2021b***)
FL_FUS	526	0.15	7.4	293	5.5	[20, 20, 270]	(***Wang et al., 2018***)
GFP_FUS	764	0.15	7.4	293	4.9	[25, 25, 300]	(***Wang et al., 2018***)
SNAP_FUS	708	0.15	7.4	293	5.9	[25, 25, 300]	(***Wang et al., 2018***)
SNAP_FUS_PLDY2F_RBDR2K	710	0.15	7.4	293	69.4	[29, 29, 340]	(***Wang et al., 2018***)
SNAP_FUS_PLDY2F	710	0.15	7.4	293	36.6	[25, 25, 300]	(***Wang et al., 2018***)
FUS_PLDY2F_RBDR2K	528	0.15	7.4	293	32.0	[25, 25, 340]	(***Wang et al., 2018***)

Table S8. Summary of CALVADOS models. The number of IDPs and MDPs, and cutoff distance for the AH potential used during optimization, cutoff distance of AH potential for validation (production) simulations, and references are shown.

models	IDPs	MDPs	cutoff_opt [nm]	cutoff_pro [nm]	Ref.
CALVADOS 1	48	0	4.0	4.0	(***Tesei et al., 2021b***)
CALVADOS 2	56	0	2.4	2.0	(***Tesei and Lindorff-Larsen, 2023***)
CALVADOS 3beta (CALVADOS3beta$_{COM}$)	56	14	2.2	2.0	this study

Table S9. Domain boundaries of MDPs used in this study for the Bayesian parameter-learning procedure and validation. Brackets indicate the first and last residue of the domain, respectively. Nested brackets indicate subdomains (restrained) separated by long linkers (unrestrained).

Protein	N	Domain 1	Domain 2	Domain 3	Domain 4	Domain 5	Domain 6	Domain 7
THB_C2	137	[6, 42]	[50, 137]					
Ubq2	162	[11, 82]	[87, 158]					
Ubq3	228	[1, 72]	[77, 148]	[153, 224]				
Gal3	250	[117, 250]						
SH4UD_SH3_SH2	264	[94, 150]	[166, 258]					
TIA1	275	[6, 82]	[95, 172]	[190, 275]				
Ubq4	304	[1, 72]	[77, 148]	[153, 224]	[229, 300]			
hnRNPA1*	314	[11, 89]	[105, 179]					
H46	381	[140, 355]						
TDP43W2A	415	[5, 77]	[107, 177]	[193, 260]	[321, 329]			
PCPE	424	[12, 125]	[134, 249]	[293, 412]				
hSUMO_hnRNPA1*	433	[44, 114]	[132, 209]	[224, 298]				
NiV_V	457	[406, 457]						
HeV_V	458	[404, 456]						
GS0	470	[1, 226]	[256, 470]					
D14	483	[31, 121]	[157, 246]	[265, 354]	[400, 479]			
GS8	486	[1, 226]	[272, 486]					
hSUMO_TIA1PrLD	492	[32, 102]	[114, 186]	[212, 287]	[321, 389]			
GS16	502	[1, 226]	[288, 502]					
GS24	518	[1, 226]	[304, 518]					
FL_FUS	526	[286, 368]	[423, 451]					
FUS_PLDY2F_RBDR2K	528	[288, 370]	[425, 453]					
GS32	534	[1, 226]	[320, 534]					
S4FL	552	[15, 138]	[[287, 294], [323, 466], [492, 542]]					
GS48	566	[1, 226]	[352, 566]					
ChiAM	682	[8, 89]	[92, 172]	[178, 257]	[266, 356]	[359, 462]	[471, 567]	[578, 668]
SNAP_FUS	708	[286, 368]	[423, 451]	[[537, 564], [586, 701]]				
SNAP_FUS_PLDY2F_RBDR2K	710	[288, 370]	[425, 453]	[[539, 566], [588, 703]]				
SNAP_FUS_PLDY2F	710	[288, 370]	[425, 453]	[[539, 566], [588, 703]]				
GFP_FUS	764	[286, 368]	[423, 451]	[529, 755]				

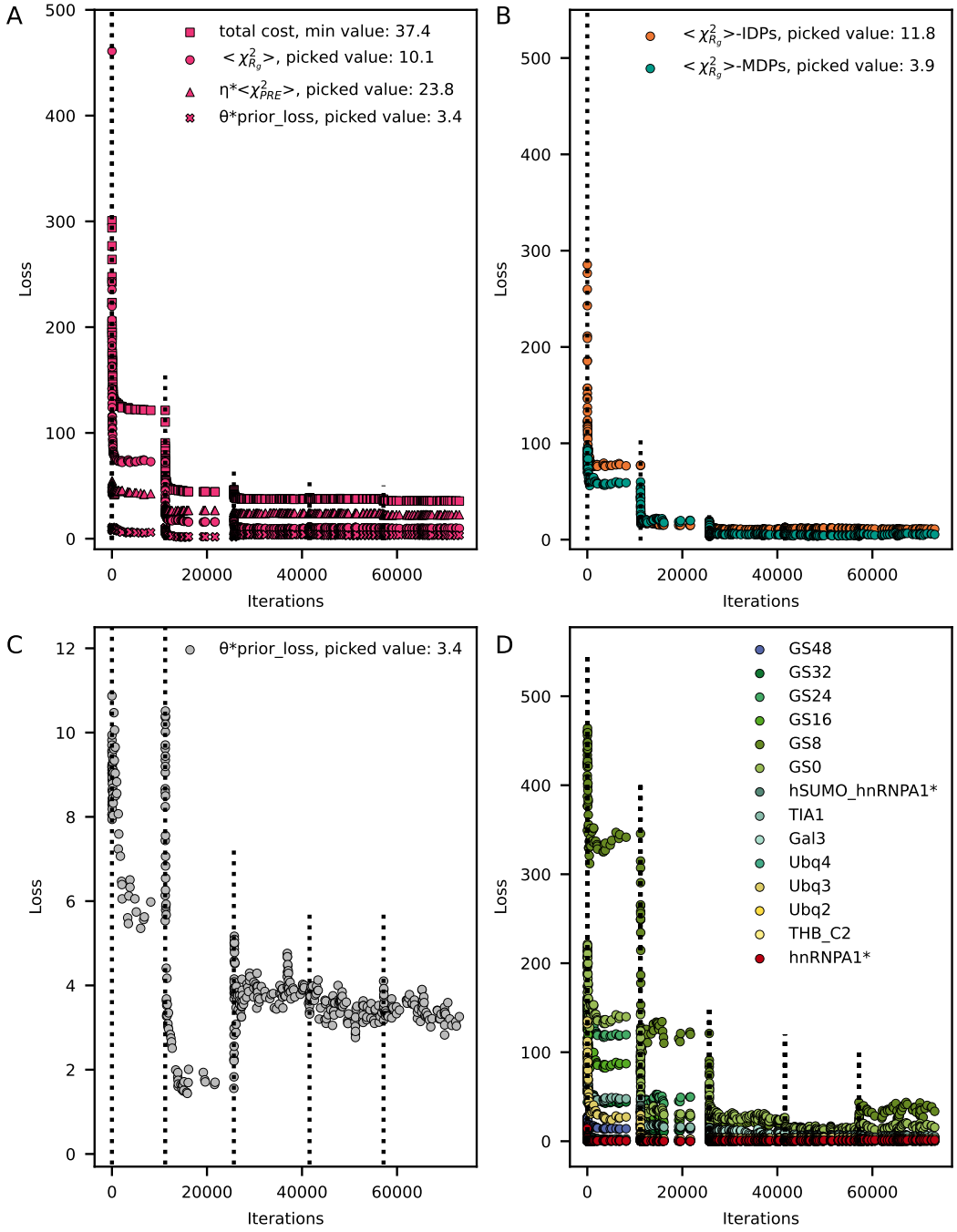

Figure S1. Optimization based on COM representation. The dotted lines indicate the starting points for each cycle. The minimal total cost was used to determine the final λ set and the corresponding values of each term are shown in the legends. (A) Evolution of the total cost (square), $\langle\chi^2_{R_g}\rangle$ (circle), $\eta\langle\chi^2_{\mathrm{PRE}}\rangle$ (triangle) and prior loss (cross) in the optimization, with $\theta = 0.08$ and $\eta = 0.1$. (B) Evolution of $\langle\chi^2_{R_g}\rangle$ contributed from IDPs (orange) and MDPs (green), respectively. (C) Evolution of prior loss. (D) Evolution of $\chi^2_{R_g}$ for each MDP during optimization.

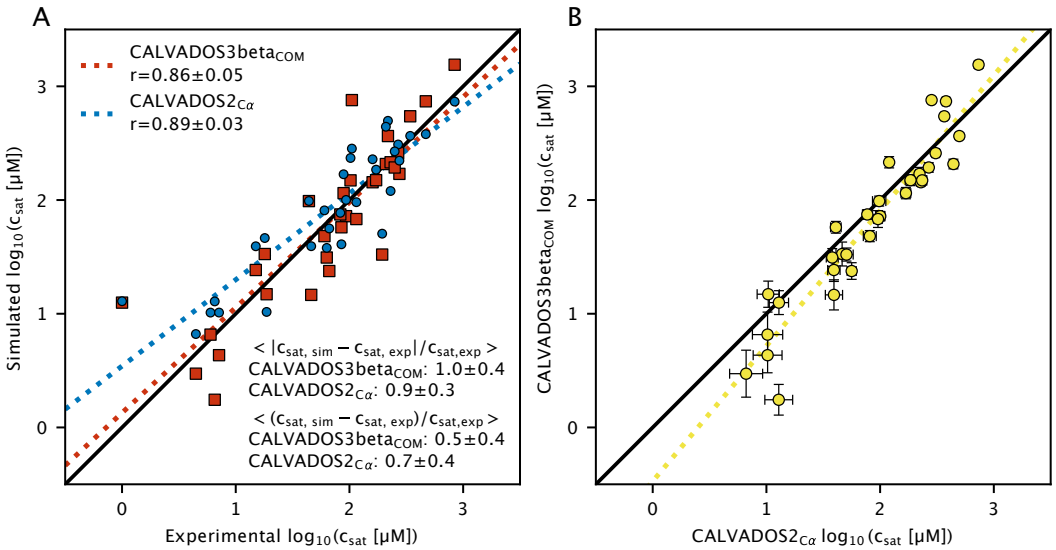

Figure S2. (A) Correlation between c_{sat} from simulations yielded by CALVADOS3beta$_{\text{COM}}$ (red) or CALVADOS2$_{C\alpha}$ (blue) and experiments for IDPs. (B) Correlation of simulated c_{sat} between CALVADOS3beta$_{\text{COM}}$ and CALVADOS2$_{C\alpha}$ for the same set of IDPs. Error bars show the error propagation with logarithm from simulated c_{sat}. The black lines indicate $y = x$. The dotted lines indicate the corresponding linear regression.

Figure S3. Time evolution of the protein concentration profiles from slab simulations of 32 IDPs using CALVADOS3beta$_{COM}$ parameters. A more intense colour intensity indicates higher protein concentration.

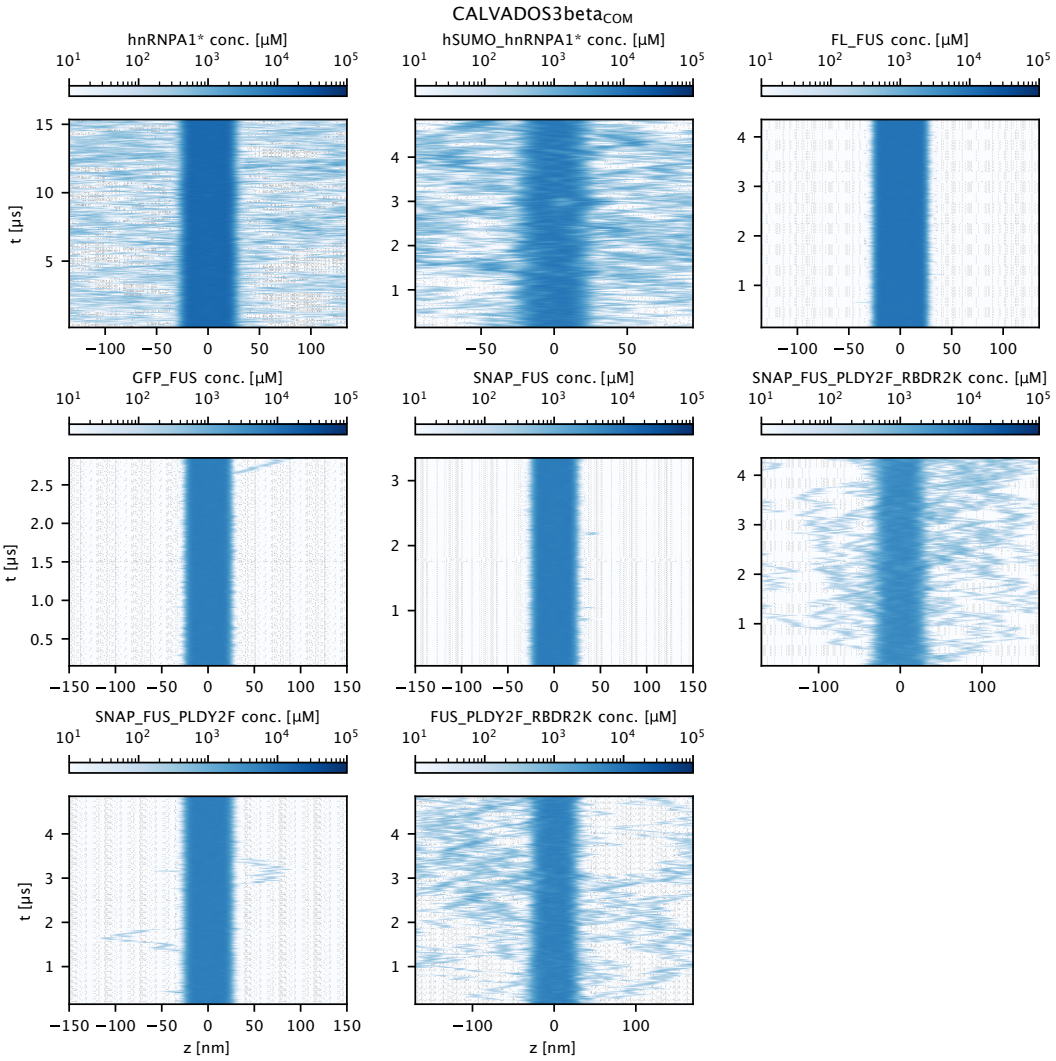

Figure S4. Time evolution of the protein concentration profiles from slab simulations of 8 MDPs using CALVADOS3beta$_{COM}$ parameters. A more intense colour intensity indicates higher protein concentration.

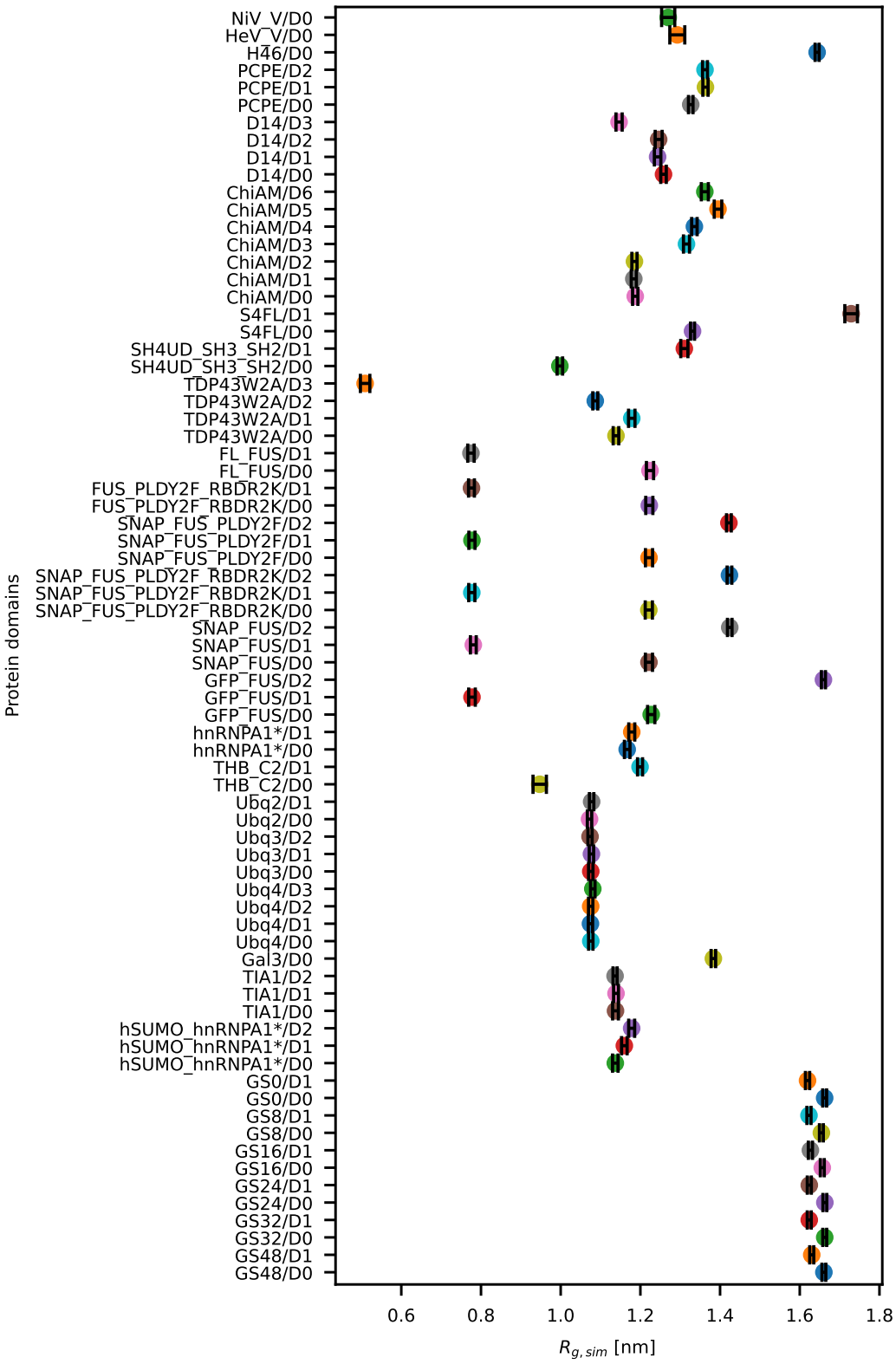

Figure S5. Simulated R_g of domains restrained by elastic network model. Each domain in a protein is indicated by D0, D1, D2, etc. Multi-domain proteins in the training set, validation set and slab set are shown.