Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range

Tesei, Giulio; Lindorff-Larsen, Kresten

Published in:
Open Research Europe

DOI:
10.12688/openreseurope.14967.2

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
RESEARCH ARTICLE

Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range [version 2; peer review: 2 approved]

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

Abstract

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

Keywords

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

This article is included in the Marie-Sklodowska-Curie Actions (MSCA) gateway.
Corresponding authors: Giulio Tesei (giulio.tesei@bio.ku.dk), Kresten Lindorff-Larsen (lindorff@bio.ku.dk)

Author roles: Tesei G: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Lindorff-Larsen K: Conceptualization, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101025063. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2023 Tesei G and Lindorff-Larsen K. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Tesei G and Lindorff-Larsen K. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range [version 2; peer review: 2 approved] Open Research Europe 2023, 2:94 https://doi.org/10.12688/openreseurope.14967.2

First published: 05 Aug 2022, 2:94 https://doi.org/10.12688/openreseurope.14967.1
1 Introduction

Biomolecular condensates may form via phase separation into coexisting solvent-rich and macromolecule-rich phases. Phase separation (PS) is driven by multiple, often transient, interactions which are in many cases engendered by intrinsically disordered proteins (IDPs) and low-complexity domains (LCDs) of multi-domain proteins\(^1\). The amino acid sequence dictates both the propensity of IDPs to phase separate and the viscoelastic properties of the condensates. Moreover, condensates of some IDPs reconstituted \textit{in vitro} tend to undergo a transition to a dynamically arrested state, in which oligomeric species can nucleate and ultimately aggregate into fibrils\(^2\)3,6\textendash}14. As accumulating evidence suggests that these processes may be involved in neurodegeneration and cancer\(^15\)\textendash}18, understanding how the amino acid sequence governs PS and rheological properties of condensates is a current research focus. Due to the transient nature of the protein-protein interactions that underpin PS, quantitative characterization of biomolecular condensates via biophysical experimental methods is challenging, and hence molecular simulations have played an important role in aiding the interpretation of experimental data on condensates reconstituted \textit{in vitro}\(^9\). However, molecular simulations of the PS of IDPs require a minimal system size of \(\sim100\) chains and long simulation times to sample the equilibrium properties of the two-phase system. To enhance the computational efficiency of these simulations, the atomistic representation of the phase-separating protein is typically coarse-grained to fewer interaction sites while the solvent is modelled as a continuum.

A widely used class of coarse-grained models of IDPs describes each residue as a single site centered at the C\(\alpha\) atom. Charged residues interact via salt-screened electrostatic interactions whereas the remaining nonionic nonbonded interactions are incorporated in a single short-range potential characterized by a set of “stickiness” parameters. The stickiness parameters are specific to either the single amino acid or pairs of residues and were originally derived by Dignon \textit{et al.} from a hydrophobicity scale\(^26\). Other models, based on the lattice simulation engines LaSSI\(^21\) and PIMMS\(^22\), classify the amino acids into a reduced number of residue types with distinct stickiness, ranging from binary categorizations into stickers and spacers\(^2,23,24\) to more detailed descriptions and parameterizations\(^23,26\). Recently, the accuracy of the stickiness parameters has been considerably improved. This has been achieved by leveraging (i) experimental data on single-chain properties, (ii) statistical analyses of protein structures, and (iii) residue-residue free energy profiles calculated from all-atom simulations\(^26\)\textendash}31. Notably, automated optimization procedures to improve of the stickiness parameters have been proposed by us and others\(26,28\)\textendash}32. In our previous work\(^32\), the procedure maximized the accuracy of the model with respect to experimental data reporting on conformational properties of IDPs, namely, small-angle X-ray scattering (SAXS) and paramagnetic relaxation enhancement (PRE) NMR data. To ensure the transferability of the model across sequence space, we trained the model on a large experimental data set and employed a Bayesian regularization approach\(^31\)\textendash}35. As the regularization term, we defined the prior knowledge on the stickiness parameters in terms of 87 hydrophobicity scales reported in the literature. The resulting optimal parameters (previously referred to as M1\(^1\)) were shown to capture the relative propensities to phase separate of a wide range of IDP sequences. However, we also observed that a systematic increase in simulation temperature of about 30 °C is needed to quantitatively reproduce the experimental concentration of the dilute phase coexisting with the condensate on an absolute scale. Herein, we refer to this model as the first version of the CALVADOS (Coarse-graining Approach to Liquid-liquid phase separation Via an Automated Data-driven Optimisation Scheme) model (CALVADOS 1).

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

2 Methods

2.1 Molecular simulations

Molecular dynamics simulations are conducted in the NVT ensemble using the Langevin integrator with a time step of 10 fs and friction coefficient of 0.01 ps\(^{-1}\). Non-bonded interactions between residues separated by one bond are excluded from the energy calculations. Functional forms and parameters for bonded and nonbonded interactions are reported in the “Bonded and nonbonded interactions” subsection. Single chains of \(N\) residues are simulated using HOOMD-blue v2.9.3 in a cubic box of side length \(0.38 \times (N - 1) + 4\) nm under periodic boundary conditions, starting from the fully extended linear conformation. Conformations are saved every \(\Delta t = 3 \times N^2\) fs if \(N > 100\) and \(\Delta t = 30\) ps otherwise. Each chain is simulated in ten replicas for a simulation time of \(600 \times \Delta t\). The initial 100 frames of each replica are discarded, so as to obtain 5,000 weakly correlated conformations for each protein (Figure 1). The functional form of \(\Delta t\) was inferred from calculations of the autocorrelation function of the radius of gyration, \(R_g\), for proteins of various \(N\). Direct-coexistence simulations are performed using openMM v7.5 in a cuboidal box of side lengths \([L_x, L_y, L_z] = [25, 25, 300], [17, 17, 300]\) and \([15, 15, 150]\) nm for Tau 2N4R, Ddx4 LCD, and for the remainder of the proteins, respectively. In the starting configuration, 100 chains are aligned along the \(z\)-axis and with their middle beads placed in the \(xy\)-plane at random \((x, y)\) positions which are more than 0.7 nm apart. Multi-chain simulations are carried out for at least 2 \(\mu s\), saving frames every 0.5 ns (Figure S1, S2, and S3). After discarding the initial 0.6 \(\mu s\), the slab is centered in the box at each frame as previously described and the equilibrium density profile, \(\rho(z)\), is calculated by averaging over the trajectory of the system at equilibrium. The densities of the dilute and protein-rich phases are estimated as the average densities in the regions \(|z| < z_{DS} - t/2\) and \(|z| > z_{DS} + 6t\) nm, where \(z_{DS}\) and \(t\) are the position of the dividing surface and the thickness of the interface, respectively. \(z_{DS}\) and \(t\) are obtained by fitting the semi-profiles in \(z > 0\) and \(z < 0\) to \(\rho(z) = (\rho_a + \rho_b)/2 + (\rho_b - \rho_a)/2 \times \tanh [(|z| - z_{DS})/t]\), where \(\rho_a\) and \(\rho_b\) are the densities of the protein-rich and dilute phases, respectively. The uncertainty of the density values is estimated as the standard error obtained from the blocking approach implemented in the BLOCKING software (github.com/fpesceKU/BLOCKING).

2.2 Bonded and nonbonded interactions

In this study, we used the following truncated and shifted Ashbaugh-Hatch potential:

\[
S^{SP}_{u_{AH}}(r) = \begin{cases} 
\frac{u_L(r) - \lambda u_L(r_c)}{2} + \epsilon(1 - \lambda), & r \leq 2^{1/6}\sigma \\
\lambda[u_L(r) - u_L(r_c)], & 2^{1/6}\sigma < r \leq r_c \\
0, & r > r_c,
\end{cases}
\]

Figure 1. Values of the autocorrelation function (ACF) of the \(R_g\) for a lag time of one, two and three frames as a function of sequence length, \(N\). The autocorrelation is calculated for the sequences of Table 1 and Table 2 simulated using (A) CALVADOS 1 and (B) CALVADOS 2 for \(6 \times 0.3 \times N^2\) ps if \(N > 100\) and for 18 ns otherwise. 600 simulation frames are saved every \(0.003 \times N^2\) ps if \(N > 100\) and every 30 ps otherwise. The initial 100 frames are discarded from each simulation.
where \( \epsilon = 0.8368 \text{ kJ mol}^{-1}, r_c = 2 \text{ or } 4 \text{ nm}, \) and \( u_{LJ} \) is the Lennard-Jones potential:

\[
u_{LJ}(r) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right]. \tag{2}\]

\( \sigma \) and \( \lambda \) are arithmetic averages of amino acid specific parameters quantifying size and hydrophathy, respectively. For \( \sigma \), we use the values calculated from van der Waals volumes by Kim and Hummer\(^{40} \) whereas, for \( \lambda \), we use the recently proposed M1 parameters\(^{32} \) and the values optimized in this work.

Salt-screened electrostatic interactions are modeled via the Debye-Hückel potential,

\[
u_{DH}(r) = \frac{q_i q_j e^2 \exp(-r_c / D)}{4\pi \varepsilon_0 e r} \tag{3}\]

where \( q \) is the average amino acid charge number, \( e \) is the elementary charge, \( D = \sqrt{4\pi N_i F_c} \) is the Debye length of an electrolyte solution of ionic strength \( c_s \) and \( B(\varepsilon) \) is the Bjerrum length. Electrostatic interactions are truncated and shifted at the cutoff distance \( r_c = 4 \text{ nm} \), irrespective of the value of \( r_c \) used for Eq. 1. We use the following empirical relationship\(^{41} \)

\[
\epsilon_T(T) = \frac{5321}{T} + 233.76 - 0.9297 \times T + 1.417 \times 10^{-3} \times T^2 - 8.292 \times 10^{-7} \times T^3, \tag{4}\]

to model the temperature-dependent dielectric constant of the implicit aqueous solution. As previously observed\(^{31} \), accounting for the temperature-dependence of \( \epsilon \) has a small effect on the predictions of the model. Indeed, the relative change in \( D \) upon an increase in temperature from 4 to 50 °C is only \(-3\%\). Similarly, at \( c_s = 150 \text{ mM} \), the Debye-Hückel energy between like-charged residues at the cutoff distance, \( u_{DH}(r_c = 4 \text{ nm}) \), is 2.6 J mol\(^{-1} \) at 4 °C and 2.8 J mol\(^{-1} \) at 50 °C. The Henderson–Hasselbalch equation is used to estimate the average charge of the histidine residues, assuming a p\( K_a \) value of 6.42.

The amino acid beads are connected by harmonic potentials,

\[
u_{\text{bond}}(r) = \frac{1}{2} k (r - r_0)^2, \tag{5}\]

doing force constant \( k = 8033 \text{ kJ mol}^{-1} \text{ nm}^{-2} \) and equilibrium distance \( r_0 = 0.38 \text{ nm} \).

### 2.3 Optimization of the stickiness scale

The optimization of the stickiness parameters, \( \lambda \), is carried out to minimize the cost function

\[
\mathcal{L}(\lambda) = \langle x_{\text{PRE}}^2(\lambda) \rangle + \eta \langle x_{\text{PRE}}^2(\lambda) \rangle - \theta \ln[P(\lambda)] \tag{6}\]

using an algorithm which is analogous to the one we previously described\(^{31} \). \( x_{\text{PRE}} \) and \( x_{\text{PRE}}^2 \) quantify the discrepancy between model predictions and experimental data, and are defined as

\[
x_{\text{PRE}} = \sum_{j} \sum_{\text{labels}} N_{\text{labels}} \left[ \sum_{i} \left( y_{ij} - y_{ij}^{\text{cal}} \right)^2 / \sigma_{ij}^{\text{cal}}^2 \right],
\]

where \( \sigma_{ij}^{\text{cal}} \) is the error on the experimental values, \( Y \) is either PRE rates or intensity ratios and \( N_{\text{labels}} \) is the number of spin-labeled mutants used for the NMR PRE data. In the expression for the cost function, the coefficients are set to \( \eta = 0.1 \) and \( \theta = 0.05 \). The prior is the distribution of \( \lambda \), \( P(\lambda) \), derived from a subset of the hydrophobicity scales reported in Table 3 and 4 of Simm et al.\(^{41} \). Specifically, only the 70 scales that are unique after min-max normalization (Figure S4) are used for the calculation of \( P(\lambda) \), namely Wilmy, BULDG reverse, MANP780101, VHEG790101, JANIN, JANJ790102, WOLR790101, PONP800101–6, Wilson, FAUCH, ENGEL, ROSEM, JACWH, CowanWhittacker, ROSM880101 reverse, ROSM880102 reverse, COWR90101, BLAS910101, CASSI, CIDH920101, CIDH920105, CIDBB, CIDAM, CIDAB, CIDA+, CIDAB reverse, NADH010101–7, ZIMJ680101, NOZY710101, JONES, LEVIT, KYTJ820101, SWER830101, SWEET, EISEN, ROSEF, GUYFE, COHEN, NNEIG, MDK0, MDK1, JURD980101, SET1-3, CHOTA, CHOTH, Sweet & Eisenberg, KIDER, ROSEB, Welling reverse, Rao & Argos, GIBRA, and WOLR810101 reverse. \( P(\lambda) \) is obtained via multivariate kernel density estimation, as implemented in scikit-learn\(^{44} \), using a Gaussian kernel with bandwidth of 0.05. This prior is 20-dimensional and contains information on the \( \lambda \)-distribution of the single amino acid as well as on the covariance matrix inferred from our selection of 70 hydrophobicity scales (Figure 2).

In the first step of the optimization procedure, the \( \lambda \) values for all the amino acids are set to 0.5, \( \lambda_0 = 0.5 \), and these parameters are used to simulate the proteins of the training set (Table 1 and Table 3). We
proceed with the first optimization cycle, wherein, at each $k$-th iteration, the $\lambda$ values of a random selection of five amino acids are nudged by random numbers picked from a normal distribution of standard deviation 0.05 to generate a trial $\lambda_k$ set. For each $i$th frame, we calculate the Boltzmann weight as $w_i = \exp \left\{ -\frac{U(r_i, \lambda_k) - U(r_i, \lambda_0)}{kT} \right\}$, where $U$ is the nonionic potential. The trial $\lambda_k$ is discarded if the effective fraction of frames, $\phi_{eff} = \exp \left\{ -\sum_i w_i \log (w_i \times N_{frames}) \right\}$, is lower than 60%. Otherwise, the acceptance probability follows the Metropolis criterion, $\min \left\{ 1, \exp \left[ \frac{L(\lambda_{k+1}) - L(\lambda_k)}{\xi_k} \right] \right\}$, where $\xi_k$ is a unitless control parameter. Each optimization cycle is divided into ten micro-cycles, wherein the control parameter, $\xi_k$, is initially set to $\xi_0 = 0.1$ and scaled down by 1% at each iteration until $\xi < 10^{-8}$. From the complete optimization cycle, we select the $\lambda$ set yielding the lowest estimate of $\mathcal{L}$. Consecutive optimization cycles are performed from simulations runs carried out with the intermediate optimal $\lambda$ set. To show that the procedure is reproducible and that the final $\lambda$ set is relatively independent of the initial conditions, we performed an additional optimization procedure starting from the M1 model, $\lambda_0 = \text{M1}$. The optimization performed in this work differs from our previous implementation also for the following details: (i) nine additional sequences have been included in the training set (Table 1 and Table 3); (ii) single chains are simulated as detailed in the “Molecular simulations” Subsection; (v) the average radius of gyration is calculated as $\langle R_g \rangle$ instead of $\sqrt{\langle R_g^2 \rangle}$.

Figure 2. (A–T) Probability distributions of the stickiness parameters, $P(\lambda)$, obtained from 70 min-max normalized hydrophobicity scales selected from the set by Simm et al. Blue bars are histograms with bin width of 0.1. Black lines are obtained as 1D projections of the multivariate kernel density estimation implemented in scikit-learn, using a Gaussian kernel with bandwidth of 0.05. (U) Covariance matrix of the 70 min-max normalized hydrophobicity scales selected from the set by Simm et al. The upper triangle of the matrix shows the covariance calculated directly from the 70 min-max normalized hydrophobicity scales whereas the lower triangle of the matrix shows the covariance calculated from the multivariate kernel density estimation averaging over the 70 min-max normalized hydrophobicity scales.
Table 1. Solution conditions and experimental radii of gyration of proteins included in the training set for the Bayesian parameter-learning procedure.

<table>
<thead>
<tr>
<th>Protein</th>
<th>N</th>
<th>$R_g$ (nm)</th>
<th>$T$ (K)</th>
<th>$c_s$ (M)</th>
<th>pH</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hst5</td>
<td>24</td>
<td>1.38 ± 0.05</td>
<td>293</td>
<td>0.15</td>
<td>7.5</td>
<td>45</td>
</tr>
<tr>
<td>(Hst5)$_2$</td>
<td>48</td>
<td>1.87 ± 0.05</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>46</td>
</tr>
<tr>
<td>p53 (20–70)</td>
<td>62</td>
<td>2.39 ± 0.05</td>
<td>277</td>
<td>0.1</td>
<td>7.0</td>
<td>47</td>
</tr>
<tr>
<td>ACTR</td>
<td>71</td>
<td>2.6 ± 0.1</td>
<td>278</td>
<td>0.2</td>
<td>7.4</td>
<td>48</td>
</tr>
<tr>
<td>Ash1</td>
<td>81</td>
<td>2.9 ± 0.05</td>
<td>293</td>
<td>0.15</td>
<td>7.5</td>
<td>49,50</td>
</tr>
<tr>
<td>CTD2</td>
<td>83</td>
<td>2.61 ± 0.05</td>
<td>293</td>
<td>0.12</td>
<td>7.5</td>
<td>50,51</td>
</tr>
<tr>
<td>Sic1</td>
<td>92</td>
<td>3.0 ± 0.4</td>
<td>293</td>
<td>0.2</td>
<td>7.5</td>
<td>52</td>
</tr>
<tr>
<td>SH4UD</td>
<td>95</td>
<td>2.7 ± 0.1</td>
<td>293</td>
<td>0.2</td>
<td>8.0</td>
<td>53</td>
</tr>
<tr>
<td>CoINt</td>
<td>98</td>
<td>2.8 ± 0.1</td>
<td>277</td>
<td>0.4</td>
<td>7.6</td>
<td>54</td>
</tr>
<tr>
<td>p15PAF</td>
<td>111</td>
<td>2.8 ± 0.1</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>55</td>
</tr>
<tr>
<td>hNL3cyt</td>
<td>119</td>
<td>3.2 ± 0.2</td>
<td>293</td>
<td>0.3</td>
<td>8.5</td>
<td>56</td>
</tr>
<tr>
<td>RNaseA</td>
<td>124</td>
<td>3.4 ± 0.1</td>
<td>298</td>
<td>0.15</td>
<td>7.5</td>
<td>57</td>
</tr>
<tr>
<td>A1</td>
<td>137</td>
<td>2.76 ± 0.02</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>-10R</td>
<td>137</td>
<td>2.67 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>-6R</td>
<td>137</td>
<td>2.57 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>+2R</td>
<td>137</td>
<td>2.62 ± 0.02</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>+7R</td>
<td>137</td>
<td>2.71 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>-3R+3K</td>
<td>137</td>
<td>2.63 ± 0.02</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>-6R+6K</td>
<td>137</td>
<td>2.79 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>-10R+10K</td>
<td>137</td>
<td>2.85 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>+12D</td>
<td>137</td>
<td>2.80 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>+4D</td>
<td>137</td>
<td>2.72 ± 0.03</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>+8D</td>
<td>137</td>
<td>2.69 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>-9F+3Y</td>
<td>137</td>
<td>2.68 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>+12E</td>
<td>137</td>
<td>2.85 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>+7K+12D</td>
<td>137</td>
<td>2.92 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>+7K+12D blocky</td>
<td>137</td>
<td>2.56 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>-4D</td>
<td>137</td>
<td>2.64 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>-8F+4Y</td>
<td>137</td>
<td>2.71 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>-10F+7R+12D</td>
<td>137</td>
<td>2.86 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>+7F-7Y</td>
<td>137</td>
<td>2.72 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>-12F+12Y</td>
<td>137</td>
<td>2.60 ± 0.02</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>-12F+12Y-10R</td>
<td>137</td>
<td>2.61 ± 0.02</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>-9F+6Y</td>
<td>137</td>
<td>2.65 ± 0.01</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
<td>58</td>
</tr>
<tr>
<td>αSyn</td>
<td>140</td>
<td>3.55 ± 0.1</td>
<td>293</td>
<td>0.2</td>
<td>7.4</td>
<td>59</td>
</tr>
</tbody>
</table>
3 Results and discussion

When applying a cutoff scheme, we neglect the interactions of residues separated by a distance, \(r\), larger than the cutoff, \(r_c\). For the most strongly interacting residue pair (between two tryptophan), the nonionic potential of the CALVADOS 1 model at \(r_c = 2\) nm takes the value of -5 J mol\(^{-1}\), that is, only a small fraction of the thermal energy (Figure 3A). However, the Lennard-Jones potential falls off slowly whereas the number of interacting partners increases quadratically with increasing \(r\). Therefore, in a simulation of a protein-rich phase, decreasing the cutoff from 4 to 2 nm (Figure 3A) can imply ignoring a total interaction energy per protein of several times the thermal energy. We first look into the effect of the choice of cutoff on the conformational ensembles of isolated proteins. We simulated single IDPs of different sequence length, \(N = 71–441\), and average hydropathy, \(\langle \lambda \rangle = 0.33–0.63\). The average radii of gyration, \(\langle R_g \rangle\), calculated from simulation trajectories are systematically larger when we use \(r_c = 2\) nm, compared to the values obtained using \(r_c = 4\) nm. CALVADOS 1 was optimized using the longer \(r_c\) and estimating the ensemble average \(R_g\) values as the root-mean-square \(\langle R_g^2 \rangle \). Since \(\langle R_g^2 \rangle\) is systematically larger than \(\langle R_g \rangle\), decreasing \(r_c\) to 2 nm results in a slight improvement of the agreement between the calculated \(\langle R_g \rangle\) values and the experimental data (Figure 3B).

To gain further insight into the effect of the cutoff, we performed simulations of single chains of \(\alpha\)-Synuclein, hnRNPA1 LCD, PNt and Tau 2N4R (Table 1 and Table 2) using \(r_c = 2, 2.5, 3\) and 4 nm. Irrespective of the sequence, \(\langle R_g \rangle\) decreases monotonically with increasing \(r_c\). However, the effect on compaction appears to be larger for long sequences and high content of hydrophobic residues, both of which result in an increased number of shorter intramolecular distances. For example, upon increasing the \(r_c\) from 2 to 4 nm, the \(\langle R_g \rangle\) of \(\alpha\)-Synuclein \((N = 140, \langle \lambda \rangle = 0.33)\) decreases by 2.3% whereas the effect is more pronounced for hnRNPA1 LCD \((N = 137, \langle \lambda \rangle = 0.61)\) and Tau 2N4R \((N = 441, \langle \lambda \rangle = 0.38)\), with a decrease in \(\langle R_g \rangle\) of 4.0% and 7.7%, respectively.

To investigate the effect of the cutoff distance on PS propensity, we performed direct-coexistence simulations of 100 chains of hnRNPA1 LCD, LAF-1 RGG domain (WT and shuffled sequence with higher charge segregation), and Tau 2N4R (Table 4). From the simulation trajectories of the two-phase system at equilibrium, we calculate \(c_{sat}\), i.e. the protein concentration in the dilute phase coexisting with the condensate. The higher the \(c_{sat}\) value, the lower the propensity of the IDP to undergo PS. As expected from the increased contact density in the condensed phase, the choice of cutoff has a considerably larger impact on \(c_{sat}\) than on chain compaction: decreasing \(r_c\) from 4 to 2 nm results in an increase in \(c_{sat}\) of over one order of magnitude. In contrast to what we observed for the \(\langle R_g \rangle\), the decrease in \(c_{sat}\) does not show a clear dependence on sequence length and average hydropathy.
Figure 3. Effect of cutoff size on predictions of radii of gyration, $R_p$, and saturation concentration, $c_{sat}$, from simulations performed using the CALVADOS 1 parameters. (A) Nonionic Ashbaugh-Hatch potentials between two W residues with cutoff, $r_c$, of 4 (blue solid line) and 2 nm (orange dashed line). The inset highlights differences between the potentials for $r_s \leq r \leq r_c$. (B) Relative difference between experimental and predicted radii of gyration, $\langle R_p \rangle$, for $r_c = 4$ (blue) and 2 nm (orange). $\chi^2$ values reported in the legend are calculated for all the sequences in Table 1. Error bars represent the experimental error relative to $expR_p$. (C) $\langle R_p \rangle$ of $\alpha$-Synuclein, hnRNPA1 LCD, Pnt and human full-length tau (Table 1 and Table 2) from simulations performed with increasing cutoff size, $r_c$, and normalized by the value at $r_c = 2$ nm. (D) Saturation concentration, $c_{sat}$, for hnRNPA1 LCD, the randomly shuffled sequence of LAF-1 RGG domain, LAF-1 RGG domain and human full-length tau for increasing values of $r_c$ and normalized by the $c_{sat}$ at $r_c = 2$ nm. (E–G) Correlation between $c_{sat}$ from simulations and experiments for (E) A1 LCD variants, (F) A1 LCD* WT at [NaCl] = 0.15, 0.2, 0.3 and 0.5 M and (G) variants of LAF-1 RGG domain (Table 4).

From the multi-chain trajectories of hnRNPA1 LCD, LAF-1 RGG domain (WT and shuffled sequence) and Tau 2N4R obtained using $r_c = 4$ nm, we estimate that the increase in nonionic energy per protein upon decreasing the cutoff from 4 to 2 nm is $13 \pm 1$ kJ mol$^{-1}$ (mean±standard deviation), respectively (Figure 4A). Assuming that the number of interactions neglected by the shorter cutoff is proportional to the sequence length and the
Table 2. Solution conditions and experimental radii of gyration of proteins simulated in this study but not included in the training set for the Bayesian parameter-learning procedure.

<table>
<thead>
<tr>
<th>Protein</th>
<th>N</th>
<th>$R_g$ (nm)</th>
<th>T (K)</th>
<th>$c_s$ (M)</th>
<th>pH</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSS1</td>
<td>71</td>
<td>2.5 ± 0.1</td>
<td>288</td>
<td>0.17</td>
<td>7.4</td>
<td>66</td>
</tr>
<tr>
<td>p27Cv14</td>
<td>107</td>
<td>2.936 ± 0.13</td>
<td>293</td>
<td>0.095</td>
<td>7.2</td>
<td>67</td>
</tr>
<tr>
<td>p27Cv15</td>
<td>107</td>
<td>2.915 ± 0.10</td>
<td>293</td>
<td>0.095</td>
<td>7.2</td>
<td>67</td>
</tr>
<tr>
<td>p27Cv31</td>
<td>107</td>
<td>2.81 ± 0.18</td>
<td>293</td>
<td>0.095</td>
<td>7.2</td>
<td>67</td>
</tr>
<tr>
<td>p27Cv44</td>
<td>107</td>
<td>2.492 ± 0.13</td>
<td>293</td>
<td>0.095</td>
<td>7.2</td>
<td>67</td>
</tr>
<tr>
<td>p27Cv56</td>
<td>107</td>
<td>2.328 ± 0.10</td>
<td>293</td>
<td>0.095</td>
<td>7.2</td>
<td>67</td>
</tr>
<tr>
<td>p27Cv78</td>
<td>107</td>
<td>2.211 ± 0.03</td>
<td>293</td>
<td>0.095</td>
<td>7.2</td>
<td>67</td>
</tr>
<tr>
<td>PTMA</td>
<td>111</td>
<td>3.7 ± 0.2</td>
<td>288</td>
<td>0.16</td>
<td>7.4</td>
<td>66</td>
</tr>
<tr>
<td>NHE6cmdd</td>
<td>116</td>
<td>3.2 ± 0.2</td>
<td>288</td>
<td>0.17</td>
<td>7.4</td>
<td>66</td>
</tr>
<tr>
<td>A1 LCD†</td>
<td>131</td>
<td>2.645 ± 0.02</td>
<td>293</td>
<td>0.05</td>
<td>7.5</td>
<td>68</td>
</tr>
<tr>
<td>A1 LCD*</td>
<td>131</td>
<td>2.65 ± 0.02</td>
<td>293</td>
<td>0.15</td>
<td>7.5</td>
<td>68</td>
</tr>
<tr>
<td>A1 LCD†</td>
<td>131</td>
<td>2.62 ± 0.02</td>
<td>293</td>
<td>0.3</td>
<td>7.5</td>
<td>68</td>
</tr>
<tr>
<td>A1 LCD*</td>
<td>131</td>
<td>2.528 ± 0.02</td>
<td>293</td>
<td>0.5</td>
<td>7.5</td>
<td>68</td>
</tr>
<tr>
<td>ANAC046</td>
<td>167</td>
<td>3.6 ± 0.3</td>
<td>298</td>
<td>0.14</td>
<td>7.0</td>
<td>66</td>
</tr>
<tr>
<td>Tau 2N3R</td>
<td>410</td>
<td>6.3 ± 0.3</td>
<td>298</td>
<td>0.15</td>
<td>7.4</td>
<td>62</td>
</tr>
<tr>
<td>Tau 2N4R</td>
<td>441</td>
<td>6.7 ± 0.3</td>
<td>298</td>
<td>0.15</td>
<td>7.4</td>
<td>62</td>
</tr>
</tbody>
</table>

Table 3. Protein and conditions related to the intramolecular PRE data included in the training set.

<table>
<thead>
<tr>
<th>Protein</th>
<th>N</th>
<th>$N_{labels}$</th>
<th>$\omega_I/2\pi$ (MHz)</th>
<th>T (K)</th>
<th>$c_s$ (M)</th>
<th>pH</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUS</td>
<td>163</td>
<td>3</td>
<td>850</td>
<td>298</td>
<td>0.15</td>
<td>5.5</td>
<td>2</td>
</tr>
<tr>
<td>FUS12E</td>
<td>164</td>
<td>3</td>
<td>850</td>
<td>298</td>
<td>0.15</td>
<td>5.5</td>
<td>2</td>
</tr>
<tr>
<td>OPN</td>
<td>220</td>
<td>10</td>
<td>800</td>
<td>298</td>
<td>0.15</td>
<td>6.5</td>
<td>69</td>
</tr>
<tr>
<td>αSyn</td>
<td>140</td>
<td>5</td>
<td>700</td>
<td>283</td>
<td>0.2</td>
<td>7.4</td>
<td>70</td>
</tr>
<tr>
<td>A2</td>
<td>155</td>
<td>2</td>
<td>850</td>
<td>298</td>
<td>0.005</td>
<td>5.5</td>
<td>3</td>
</tr>
</tbody>
</table>

With these considerations in mind, we use the CALVADOS 1 model with $r_c = 2$ nm to run direct-coexistence simulations of IDPs for which $c_{sat}$ has been measured experimentally (Table 4), i.e. variants of hnRNPA1 LCD, hnRNPA1 LCD* at various salt concentrations, and LAF-1 RGG domain variants. As we have shown amino acid concentration in the condensate, the small variance in the energy increase across the different IDPs finds explanation in the fact that the simulated systems display similar values of $N^2 \times c_{con}$ (Figure 4A), where $c_{con}$ is the protein concentration in the condensate. The ratio $U(r_c = 2 \text{ nm})/U(r_c = 4 \text{ nm})$ of the nonionic energies for $r_c = 2$ and 4 nm is also largely system independent (Figure 4B). Moreover, decreasing the temperature by $\sim 30$ K in the range between 310 and 323 K has a rather small impact on the relative strength of the electrostatic interactions with respect to the thermal energy, due to the decrease in the dielectric constant of water with increasing temperature (Figure 4C). Therefore, we speculate that the effect of decreasing $r_c$ can be compensated by simulating the system at a lower temperature (Figure 4B).
Table 4. Proteins and conditions used for the direct-coexistence simulations performed in this study and references to the experimental data. Shaded rows highlight systems which are not included in the correlation plot of Figure 7C.

<table>
<thead>
<tr>
<th>Protein</th>
<th>N</th>
<th>$c_i$ (mM)</th>
<th>pH</th>
<th>Ref.</th>
<th>$T$ (K)</th>
<th>4 nm</th>
<th>2 nm</th>
<th>Figure 1C</th>
</tr>
</thead>
<tbody>
<tr>
<td>6His-TEV-Lge11–80-StrepII WT</td>
<td>114</td>
<td>100</td>
<td>7.5</td>
<td>71</td>
<td>-</td>
<td>293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>6His-TEV-Lge11–80-StrepII -11R+11K</td>
<td>114</td>
<td>100</td>
<td>7.5</td>
<td>71</td>
<td>-</td>
<td>293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>6His-TEV-Lge11–80-StrepII -14Y+14A</td>
<td>114</td>
<td>100</td>
<td>7.5</td>
<td>71</td>
<td>-</td>
<td>293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD WT</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>4, 58</td>
<td>310 &amp; 323</td>
<td>277 &amp; 293</td>
<td>310</td>
<td></td>
</tr>
<tr>
<td>A1 LCD +7F-7Y</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310 &amp; 323</td>
<td>277 &amp; 293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD -12F+12Y</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310 &amp; 323</td>
<td>277 &amp; 293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD -23S+23T</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310 &amp; 323</td>
<td>277 &amp; 293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD -14N+14Q</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310 &amp; 323</td>
<td>277 &amp; 293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD -10G+10S</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310 &amp; 323</td>
<td>277 &amp; 293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD -20G+20S</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310 &amp; 323</td>
<td>277 &amp; 293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD -30G+30S</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>323</td>
<td>293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD +23G-23S</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>323</td>
<td>293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD +23G-23S-7F-7Y</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>323</td>
<td>293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD +23G-23S-12F+12Y</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>323</td>
<td>293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD -9F+3Y</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310</td>
<td>277</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD -8F+4Y</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310</td>
<td>277</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD -3R+3K</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310</td>
<td>277</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD -6R</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310</td>
<td>277</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD -4D</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310</td>
<td>277</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD +4D</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310</td>
<td>277</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD +8D</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310</td>
<td>277</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD +2R</td>
<td>137</td>
<td>150</td>
<td>7.0</td>
<td>58</td>
<td>310</td>
<td>277</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD+ WT</td>
<td>131</td>
<td>150</td>
<td>7.0</td>
<td>72</td>
<td>323</td>
<td>293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD+ WT</td>
<td>131</td>
<td>200</td>
<td>7.0</td>
<td>72</td>
<td>323</td>
<td>293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD+ WT</td>
<td>131</td>
<td>300</td>
<td>7.0</td>
<td>72</td>
<td>323</td>
<td>293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1 LCD+ WT</td>
<td>131</td>
<td>500</td>
<td>7.0</td>
<td>72</td>
<td>323</td>
<td>293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>LAF-1 RGG Domain</td>
<td>176</td>
<td>150</td>
<td>7.5</td>
<td>73</td>
<td>323</td>
<td>293</td>
<td>293</td>
<td></td>
</tr>
<tr>
<td>LAF-1 RGG Domain Shuffled</td>
<td>176</td>
<td>150</td>
<td>7.5</td>
<td>73</td>
<td>323</td>
<td>293</td>
<td>323</td>
<td></td>
</tr>
<tr>
<td>LAF-1 RGG Domain ∆21–30</td>
<td>166</td>
<td>150</td>
<td>7.5</td>
<td>73</td>
<td>323</td>
<td>293</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A2 LCD</td>
<td>155</td>
<td>10</td>
<td>5.5</td>
<td>74</td>
<td>-</td>
<td>297</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>FUS LCD</td>
<td>163</td>
<td>150</td>
<td>7.4</td>
<td>75</td>
<td>-</td>
<td>297</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Ddx4 LCD</td>
<td>236</td>
<td>130</td>
<td>6.5</td>
<td>76</td>
<td>-</td>
<td>297</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Human Full-Length Tau (2N4R)</td>
<td>441</td>
<td>70</td>
<td>7.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>277</td>
<td></td>
</tr>
</tbody>
</table>
that decreasing the range of the nonionic interactions disfavours PS, we perform these simulations at the experimental temperatures, which are lower by ∼30 K than those required to reproduce the experimental csat values when the model is simulated with \( r_c = 4 \) nm (Figure 3E–G). The two-fold decrease in \( r_c \) enables the model to quantitatively recapitulate the experimental csat data at the temperature at which the experiments were conducted. Notably, we show this for diverse sequences, across a wide range of ionic strengths, and for variants with different charge patterning and numbers of aromatic and charged residues. These results suggest that the range of interaction of the Lennard-Jones potential may be too large. While the \( r^{-6} \) dependence is strictly correct for dispersion interactions between atoms, the nonionic potential of our model incorporates a variety of effective nonbonded interactions between residues, and hence the Lennard-Jones potential is not expected to capture the correct interaction range.

Since CALVADOS 1 was developed using \( r_c = 4 \) nm, we examined whether reoptimizing the model with the shorter cutoff could result in a comparably accurate model. As detailed in the Methods Section, we performed a Bayesian parameter-learning procedure using an improved algorithm, an expanded training set (Table 1), and \( r_c = 2 \) nm. Figure S5 shows that the new model tends to underestimate the csat values of the most PS-prone sequences. We hypothesize that during the optimization the reduction of attractive forces due to the shorter cutoff is overcompensated by an overall increase in \( \lambda \). We tested this hypothesis by performing the optimization with increasing values of \( r_c \), in the range between 2.0 and 2.5 nm, and found that the csat values predicted from simulations performed with \( r_c = 2.0 \) nm tend to increase with the \( r_c \) used for the optimization (Figure 5).

Using \( r_c = 2.4 \) nm for the optimization resulted in a model with improved accuracy compared to CALVADOS 1 (Figure 6), especially for the PS of LAF-1 RGG domain and the −23S+23T variant of A1 LCD. To test the robustness of the approach, the optimization was carried out starting from \( \lambda_0 = 0.5 \) for all the amino acids (Figure 6) and from \( \lambda_0 = M1 \) (Figure S6 and S7). The difference between the resulting sets of optimal \( \lambda \) values (Figure S7A) is lower than 0.08 for all the residues and exceeds 0.05 only for S, T and A. The model obtained starting from \( \lambda_0 = 0.5 \) is more accurate at predicting PS propensities and will be referred to as CALVADOS 2 hereafter.

The \( \lambda \) values of CALVADOS 1 and 2 differ mostly for K, T, A, M, and V, whereas the smaller deviations (\(|\Delta \lambda| < 0.09\)) observed for Q, L, I, and F (Figure 6A) are within the range of reproducibility of the method (Figure S7A). Although CALVADOS 1 was optimized using \( r_c = 4 \) nm, predictions of single-chain compaction from simulations performed using \( r_c = 2 \) nm are more accurate for CALVADOS 1 than for CALVADOS 2.
Saturation concentrations, $c_{\text{sat}}$, as a function of the cutoff used to optimize the model. $c_{\text{sat}}$ values are calculated from simulations performed using $r_c = 2.0$ nm whereas the models are optimized using $r_c = 2.0$, 2.2, 2.4, and 2.5 nm. Horizontal dotted lines represent experimental $c_{\text{sat}}$ values from the references reported in Table 4.

This result can be explained by the opposing effects of decreasing the cutoff and calculating $\langle R_g \rangle$ instead of $\langle R_g^2 \rangle$. In fact, the $\langle R_g^2 \rangle$ values predicted by CALVADOS 1 are strikingly similar to the $\langle R_g \rangle$ values predicted by CALVADOS 2 (Figure 7A).

The correlation between experimental and predicted $R_g$ values for the 67 proteins of Table 1 and Table 2 is excellent for both CALVADOS 1 and 2 (Figure 7B). On the other hand, CALVADOS 2 is more accurate than CALVADOS 1 at predicting PS propensities, as evidenced by Pearson’s correlation coefficients of 0.93 and 0.82, respectively, for the experimental and predicted $c_{\text{sat}}$ values of the 26 sequences of Table 4 (Figure 7C).

Capturing the interplay between short-range nonionic and long-range ionic interactions is essential for accurately modeling the PS of IDPs. Our results show that the decrease in the range of the nonionic potential reported in this work does not significantly perturb the balance between ionic and nonionic forces. In fact, CALVADOS 1 and 2 accurately predict the PS propensities of A1 LCD at various salt concentrations, as well as the $c_{\text{sat}}$ of variants of A1 LCD and LAF-1 RGG domain with different charge patterning (Figure 3E–G and Figure 6D–F). Moreover, CALVADOS 1 and 2 recapitulate the effect of salt concentration and charge patterning on the chain compaction of A1 LCD and p27-C constructs (Figure 8A and 8B).

In the model, ionic interactions are also truncated and shifted. At the cutoff distance of 4 nm, the ionic energy decreases with increasing salt concentration and amounts to $\pm 2.7$ J mol$^{-1}$ at $c_s = 10$ mM and 20 °C. However, this energy is $\sim 43$ times larger at $c_s = 10$ mM, which suggests that the model may considerably underestimate the strength and range of charge-charge interactions at low salt concentrations. To investigate this aspect, we performed single-chain and direct-coexistence simulations using a longer cutoff of 6 nm for the ionic interactions. The change in cutoff has a small effect on both the $R_g$ (Figure S8A) and the $c_{\text{sat}}$ values predicted for systems at $c_s = 150$ mM (Figure S8C). Conversely, simulations at low salt concentration are considerably affected by the increase in cutoff. For the PRE data of A2 LCD at $c_s = 5$ mM, we observe an improvement in the agreement with experiments (Figure S8B). Instead, the accuracy of the phase behaviour predicted for A2 LCD at $c_s = 10$ mM decreases significantly as the $c_{\text{sat}}$ value shows a $\sim 100$-fold increase (Figure S8C). Since the vast majority of the available $R_g$ and $c_{\text{sat}}$ data in our training and test sets was measured at $c_s \approx 150$ mM, we are currently unable to further assess or improve the accuracy of the model at low salt concentrations.

As additional test systems, we considered constructs of the 1–80 N-terminal fragment of yeast Lge1, which have been recently investigated using turbidity measurements. CALVADOS 2 correctly predicts that the
WT Lge1-80 construct undergoes PS at the experimental conditions, albeit with a hundred times larger \( c_{sat} \) (50 ± 6 µM at \( c_s = 100 \text{mM} \)) compared to experiments (< 1 µM). In agreement with experiments, CALVADOS 2 predicts that mutating all the 11 R residues to K increases \( c_{sat} \) by over one order of magnitude whereas mutating the 14 Y residues of the 1–80 fragment to A abrogates PS (Figure 8C).

4 Conclusions
In the context of a previously developed Cα-based IDP model (CALVADOS), we show that neglecting the long range of attractive Lennard-Jones interactions has a small impact on the compaction of a single chain while strongly disfavouring PS. The effect can be explained by the smaller number of neglected pair interactions
Figure 7. (A) Relative difference between experimental and predicted radii of gyration for CALVADOS 1 (orange) and CALVADOS 2 (blue). Full and hatched bars show \((R_g^{\text{calc}} - R_g^{\text{exp}}) / R_g^{\text{exp}}\) where \(R_g^{\text{calc}}\) is calculated as the mean \(2 \langle gR \rangle\) or the root mean square \(\langle gR^2 \rangle^{1/2}\), respectively. The vertical dashed line splits the plot into the 51 and 16 sequences or solution conditions of the training set (Table 1) and test set (Table 2), respectively. Error bars represent the experimental error relative to \(R_g^{\text{exp}}\). \(\chi^2\) values in the legend are averages over 67 different sequences or solution conditions (Table 1 and Table 2). (B and C) Comparison between experimental and predicted \(R_g^{\text{calc}}\) (Table 1 and Table 2) and \(c_{sat}\) values for CALVADOS 1 (orange) and CALVADOS 2 (blue). Pearson’s \(r\) coefficients are reported in the legend. Small squares in C show the same data as in Figure 6C–F whereas the large upward triangle, downward triangle, and circle show values for A2 LCD, FUS LCD, and Ddx4 LCD, respectively, at the conditions reported in Table 4.

Figure 8. Comparison between experimental \(R_g\) values and predictions of CALVADOS 1 (orange) and CALVADOS 2 (blue) for (A) A1 LCD at different salt concentrations (50 mM < \(c_s\) < 500 mM) and (B) p27-C constructs of different charge patterning (0.1 < \(\kappa\) < 0.8). Experimental conditions and references are reported in Table 2. (C) Predictions of CALVADOS 2 direct-coexistence simulations of the PS of constructs of the 1–80 N-terminal fragment of yeast Lge1 simulated at \(c_s = 100\) mM. Protein concentration profiles are shown as a function of the long side of the simulation cell for WT (blue), -11R+11K variant (orange), and -14Y+14A variant (green).
for a residues in an isolated chain compared to the dense environment of a condensate. Moreover, we find that the effect of reducing the range of interaction by a factor of two is relatively insensitive to sequence length and composition. Therefore, decreasing the cutoff of the Lennard-Jones potential of the Cα-based model engenders a similar generic effect on chain compaction and PS as a corresponding increase in temperature. We take advantage of this finding to solve the temperature mismatch of the CALVADOS model. Namely, we decrease the cutoff of the nonionic interactions from 4 to 2 nm and obtain accurate $c_{\text{sat}}$ predictions at the experimental conditions, whereas simulations at temperatures higher by 30 °C were required in the original implementation. Finally, we used the shorter cutoff to reoptimize the stickiness parameters of the model against experimental data reporting on single-chain compaction. The small expansion of the chain conformations is overcompensated by an overall increase in stickiness so that the resulting model tends to underestimate the experimental $c_{\text{sat}}$ values. By systematically increasing the cutoff used in the development of the stickiness scale, we find that performing the optimization using $r_c = 2.4$ nm results in a model (CALVADOS 2) which yields accurate predictions from simulations run using $r_c = 2$ nm at the experimental conditions. We present CALVADOS 2 as an improvement of our previous model by testing on sets of experimental $R_g$ and $c_{\text{sat}}$ data comprising 16 and 36 systems, respectively, which were not used in the parameterization of the model.

**Ethics and consent**

Ethical approval and consent were not required.

**Data availability**

Underlying data

Code and data to reproduce the results presented in this work are available at github.com/KULL-Centre/papers/tree/main/2022/CG-cutoffs-Tesei-et-al and archived on Zenodo at doi.org/10.5281/zenodo.7437501 under the terms of the Creative Commons Attribution 4.0 International license.

Extended data

Supplementary figures S1–S7 are deposited on Zenodo at doi.org/10.5281/zenodo.7437501 under the terms of the Creative Commons Attribution 4.0 International license.

**Software availability**

Scripts and parameters to perform single-chain and direct-coexistence simulations using CALVADOS 1 and 2 are available at github.com/KULL-Centre/CALVADOS and archived on Zenodo at doi.org/10.5281/zenodo.7437501 under the terms of the GNU Affero General Public License v3.0.

**Acknowledgments**

We thank Anna Ida Trolle for her help in setting up the protocol for single-chain simulations. We thank Rosana Collepardo, Jerelle A. Joseph, and Aleks Reinhardt for useful discussions that led us to explore the effects of cutoffs. We acknowledge access to computational resources from the ROBUST Resource for Biomolecular Simulations (supported by the Novo Nordisk Foundation; NNF18OC0032608) and Biocomputing Core Facility at the Department of Biology, University of Copenhagen. An earlier version of this article can be found on bioRxiv at doi.org/10.1101/2022.07.09.499434.

**References**


7. Murakami T, Qamar S, Lin Q, et al.: ALS/FTD mutation-induced...

PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


Publisher Full Text


Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


Publisher Full Text


Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


Reference Source


PubMed Abstract | Publisher Full Text | Free Full Text

Frauke Gräter
1 Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany
2 Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany

Camilo Aponte-Santamaria
1 Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany
2 Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany

Tesei and Lindorff-Larsen present CALVADOS 2 which is an improvement of CALVADOS 1, their original coarse-grained model to simulate the dynamics of intrinsically disordered protein chains or condensates of them. They systematically studied the effect of the cutoff of the short-range interactions on the compactness of single chains and the phase behavior propensity. They optimized the model using a large set of intrinsically disordered proteins of varying amino acid length and for which experimental data exist.

This is a very important study as it shows to what extent the truncation of the short range interactions affect the dynamics of single chains and condensates of IDPs and how this feature can be used to balance the excess thermal energy needed to calibrate the original CALVADOS implementation.

I have the following comments.

Similarly, as it was done for the short-range interactions, the electrostatic interactions are also truncated (at a cutoff distance of 4 nm which is about 4 fold the Debye length at 300 K and 150 mM ionic strength). I wonder what the influence of this cutoff is. The authors could comment on that.

From equations 3 and 4, I understand that a temperature-dependent Debye length was used. How much does this length change when changing the temperature (relative to the cutoff)? I think it is important to comment on that in the paper.
Minor points:
  ○ what are the black lines in Fig 3B?
  ○ Intro 2nd paragraph: What is M1? M1 has not been defined.
  ○ Methods: simulation length of 6*0.3*N^2 ps. What was the motivation to choose this particular simulation length?
  ○ Fig 4B, orange-yaxis: title confusing. It is not the normalized temperature ratio but the kinetic energy ratio.

**Is the work clearly and accurately presented and does it cite the current literature?**
Yes

**Is the study design appropriate and does the work have academic merit?**
Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**
Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**
Yes

**Are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions drawn adequately supported by the results?**
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Molecular Dynamics, IDPs, coarse-graining

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 14 Dec 2022

**Giulio Tesei**

We thank Frauke Gräter and Camilo Aponte-Santamaria for their comments and positive view on our work.

Below is our point-by-point response to the comments:
  ○ *Similarly, as it was done for the short-range interactions, the electrostatic interactions are also truncated (at a cutoff distance of 4 nm which is about 4 fold the Debye length at 300 K and 150 mM ionic strength). I wonder what the influence of this cutoff is. The authors*
could comment on that.

Response: We agree with the reviewers that the effect of the cutoff on the ionic interactions is an interesting aspect to examine. We first tested the effect on the single chain data. Although some proteins (A2 LCD and CAHSD) were simulated at low ionic strength \( (c_s = 5-70 \text{ mM}) \), we found that increasing the cutoff value from 4 to 6 nm leads to relative changes in the predicted \( R_g \) values below 2.5% (Figure S8A). However, the effect on the predicted PRE data is more pronounced. In particular, for A2 LCD we observe a ~30% decrease in the \( \chi^2_{\text{PRE}} \) value (Figure S8B).

We also tested the effect on the phase behavior of A2 LCD \( (c_s = 10 \text{ mM}) \), A1 LCD* \( (c_s = 150 \text{ mM}) \), FUS LCD \( (c_s = 150 \text{ mM}) \), LAF-1 RGG domain \( (c_s = 150 \text{ mM}) \), and its shuffled variant \( (c_s = 150 \text{ mM}) \). As expected, increasing the cutoff from 4 to 6 nm has a negligible effect on the phase separation at high \( c_s \) whereas a considerable increase in saturation concentration, \( c_{\text{sat}} \), is observed for simulations of A2 LCD at \( c_s = 10 \text{ mM} \) (Figure S8C). Coincidentally, we found that the \( c_{\text{sat}} \) predicted for A2 LCD using the shorter cutoff is in better agreement with the reference experimental value. We now discuss this point in the results section: In the model, ionic interactions are also truncated and shifted. At the cutoff distance of 4 nm, the ionic energy decreases with increasing salt concentration and amounts to \( \pm 2.7 \text{ J mol}^{-1} \) at \( c_s = 150 \text{ mM} \) and 20 °C. However, this energy is ~43 times larger at \( c_s = 10 \text{ mM} \), which suggests that the model may considerably underestimate the strength and range of charge-charge interactions at low salt concentrations. To investigate this aspect, we performed single-chain and direct-coexistence simulations using a longer cutoff of 6 nm for the ionic interactions. The change in cutoff has a small effect on both the \( R_g \) (Figure S8A) and the \( c_{\text{sat}} \) values predicted for systems at \( c_s = 150 \text{ mM} \) (Figure S8C). Conversely, simulations at low salt concentration are considerably affected by the increase in cutoff. For the PRE data of A2 LCD at \( c_s = 5 \text{ mM} \), we observe an improvement in the agreement with experiments (Figure S8B). Instead, the accuracy of the phase behaviour predicted for A2 LCD at \( c_s = 10 \text{ mM} \) decreases significantly as the \( c_{\text{sat}} \) value shows a ~100-fold increase (Figure S8C). Since the vast majority of the available \( R_g \) and \( c_{\text{sat}} \) data in our training and test sets was measured at \( c_s \approx 150 \text{ mM} \), we are currently unable to further assess or improve the accuracy of the model at low salt concentrations.

○ From equations 3 and 4, I understand that a temperature-dependent Debye length was used. How much does this length change when changing the temperature (relative to the cutoff)? I think it is important to comment on that in the paper.

Response: The Debye length, \( D \), used in the model shows a weak temperature dependence. For example, the relative change in \( D \) upon an increase in temperature from 4 to 50 °C is only -3%. Ionic interaction energies between like-charge residues at a cutoff distance of 4 nm, \( u_{\text{DH}}(r = 4 \text{ nm}) \), also show a weak temperature dependence. At \( c_s = 150 \text{ mM} \), \( u_{\text{DH}}(r = 4 \text{ nm}) \) is 2.6 J mol\(^{-1}\) at 4 °C and 2.8 J mol\(^{-1}\) at 50 °C. We now discuss this point in the methods section: As previously observed [doi:10.1038/s43588-021-00155-3], accounting for the temperature-dependence of \( \epsilon \) has a small
effect on the predictions of the model. Indeed, the relative change in $D$ upon an increase in temperature from 4 to 50 °C is only -3%. Similarly, at $c_s = 150$ mM, the Debye-Hückel energy between like-charged residues at the cutoff distance, $u_{DH}(r = 4 \text{ nm})$, is 2.6 J mol$^{-1}$ at 4 °C and 2.8 J mol$^{-1}$ at 50 °C.

- **What are the black lines in Fig 3B?**

  **Response:** We have clarified in the figure caption that the black error bars represent the relative error of the experimental measurement, i.e. $\sigma_{\text{exp}}^*/R_g^{\text{exp}}$.

- **Intro 2nd paragraph: What is M1? M1 has not been defined.**

  **Response:** We have changed the text to specify that M1 is the name used to refer to the CALVADOS 1 optimal stickiness parameters in our previous publication on the model (Tesei et al. DOI: 10.1073/pnas.2111696118).

- **Methods:** simulation length of $6 \times 0.3 \times N^2$ ps. What was the motivation to choose this particular simulation length?

  **Response:** We have changed the text to clarify how we chose sampling frequencies and simulation lengths based on sequence length, $N$. Briefly, we saved every $\Delta t \approx 3 \times N^2$ fs if $N > 100$, and $\Delta t = 30$ ps otherwise, to ensure that the radii of gyration for consecutive frames were weakly correlated irrespective of sequence length, $N$ (Figure 1). The quadratic $N$-dependence of $\Delta t$ was inferred from the lag time at which the autocorrelation function of the $R_g$ approximates 1/2, for a subset of proteins of different $N$. We simulated ten replicas per sequence, each for a simulation time of $600 \times \Delta t$. After discarding the initial 100 frames of each replica, we obtained 5,000 weakly correlated conformations for each protein. In the context of the optimization procedure, we observed that this number of frames is sufficient for accurately reweighting the trajectories, when the fractions of effective frames exceeds 60%.

- **Fig 4B, orange y-axis: title confusing. It is not the normalized temperature ratio but the kinetic energy ratio.**

  **Response:** We have clarified in the caption that the orange y-axis shows the ratio of thermal energies, $RT' / RT$, where $R$ is the gas constant.

**Competing Interests:** No competing interests were disclosed.
Alex Holehouse

1 Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO, USA
2 Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO, USA

This paper is a timely update to the recent CALVADOS forcefield developed by Tesei et al. and published late last year. The authors demonstrate how while tweaking aspects of CALVDOS do not substantially alter the single-chain radii of gyrations, the phase behavior can be altered substantially.

The updated version provides some interesting discussion, the data are all shared on GitHub and Zenodo, and all software parameters are made available in a convenient format. The paper is well-written, the methods well-described and logically motivated and the figures clear.

What more could one want from a paper? I strongly support indexing in its current form.

As a note, the conciseness of this review should not be seen as a lack of attention to detail. However, would any suggestions or comments I make materially affect the conclusions, clarity, or availability of data? I do not think so, and as such, it is in the author's best interest to use their time on the next set of questions than fine-tune what is already a very strong manuscript.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Coarse-grained simulations of disordered proteins

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
Giulio Tesei

We thank Alex Holehouse for the very kind comments and positive view on our work.

**Competing Interests:** No competing interests were disclosed.