

# Identifying Sequence Effects on Chain Dimensions of Disordered Proteins by Integrating Experiments and Simulations

Andrea Holla, Erik W. Martin, Thomas Dannenhoffer-Lafage, Kiersten M. Ruff, Sebastian L. B. König, Mark F. Nüesch, Aritra Chowdhury, John M. Louis, Andrea Soranno, Daniel Nettels, Rohit V. Pappu,\* Robert B. Best,\* Tanja Mittag,\* and Benjamin Schuler\*



Cite This: *JACS Au* 2024, 4, 4729–4743



Read Online

ACCESS |



Metrics & More



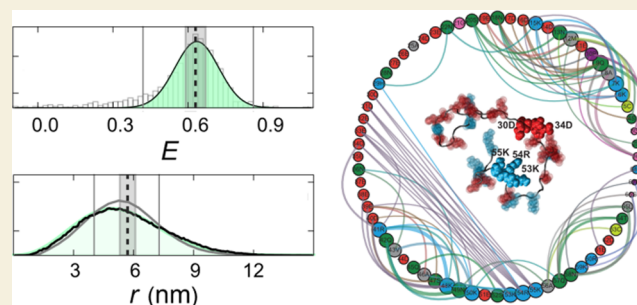
Article Recommendations



Supporting Information

**ABSTRACT:** It has become increasingly evident that the conformational distributions of intrinsically disordered proteins or regions are strongly dependent on their amino acid compositions and sequence. To facilitate a systematic investigation of these sequence-ensemble relationships, we selected a set of 16 naturally occurring intrinsically disordered regions of identical length but with large differences in amino acid composition, hydrophobicity, and charge patterning. We probed their conformational ensembles with single-molecule Förster resonance energy transfer (FRET), complemented by circular dichroism (CD) and nuclear magnetic resonance (NMR) spectroscopy as well as small-angle X-ray scattering (SAXS). The set of disordered proteins shows a strong dependence of the chain dimensions on sequence composition, with chain volumes differing by up to a factor of 6. The residue-specific intrachain interaction networks that underlie these pronounced differences were identified using atomistic simulations combined with ensemble reweighting, revealing the important role of charged, aromatic, and polar residues. To advance a transferable description of disordered protein regions, we further employed the experimental data to parametrize a coarse-grained model for disordered proteins that includes an explicit representation of the FRET fluorophores and successfully describes experiments with different dye pairs. Our findings demonstrate the value of integrating experiments and simulations for advancing our quantitative understanding of the sequence features that determine the conformational ensembles of intrinsically disordered proteins.

**KEYWORDS:** intrinsically disordered proteins, single-molecule spectroscopy, Förster resonance energy transfer (FRET), atomistic simulations, coarse-grained simulations, chain dimensions, local expansion and compaction



## INTRODUCTION

Large parts of the proteomes of higher eukaryotes consist of intrinsically disordered proteins (IDPs), which do not adopt a well-defined three-dimensional structure under physiological conditions.<sup>1</sup> For instance, ~58% of human proteins contain both folded domains and intrinsically disordered regions (IDRs).<sup>2</sup> IDRs occur in a variety of structural contexts, from tails and linkers between folded domains to fully disordered proteins, and they are particularly prevalent in regulation, such as in transcription and signaling,<sup>3</sup> as well as in cellular organization via phase separation.<sup>4</sup> Despite the lack of a well-defined tertiary structure, however, the conformational properties of IDPs are far from uniform: They range from compact states that can be rich in secondary structure to less compact ensembles all the way to highly expanded chains with no detectable secondary structure.<sup>5–16</sup>

For classifying and quantifying this continuous spectrum of disorder, concepts from polymer physics can be useful.<sup>3,14,17,18</sup> For instance, based on the combination of net charge per residue and fraction of charged residues, IDPs can be grouped

into strong and weak polyelectrolytes and polyampholytes,<sup>19</sup> and classified by their chain dimensions in terms of ensemble-averaged quantities, such as their hydrodynamic radius, radius of gyration, or end-to-end distance.<sup>6,12,18,20,21</sup> A helpful quantity for characterizing the dimensions of unfolded and disordered proteins independent of chain length is the scaling exponent,  $\nu$ ,<sup>9,22</sup> which relates the chain dimensions,  $R$ , to the number of residues or chain segments,  $N$ , as  $R \propto N^\nu$ . For infinitely long homopolymers,  $\nu$  can take values of 1/3 for compact globules (and globular folded proteins), 1/2 for Flory random coils, and ~0.588 for excluded volume chains.<sup>22</sup> However, intermediate and larger values are commonly

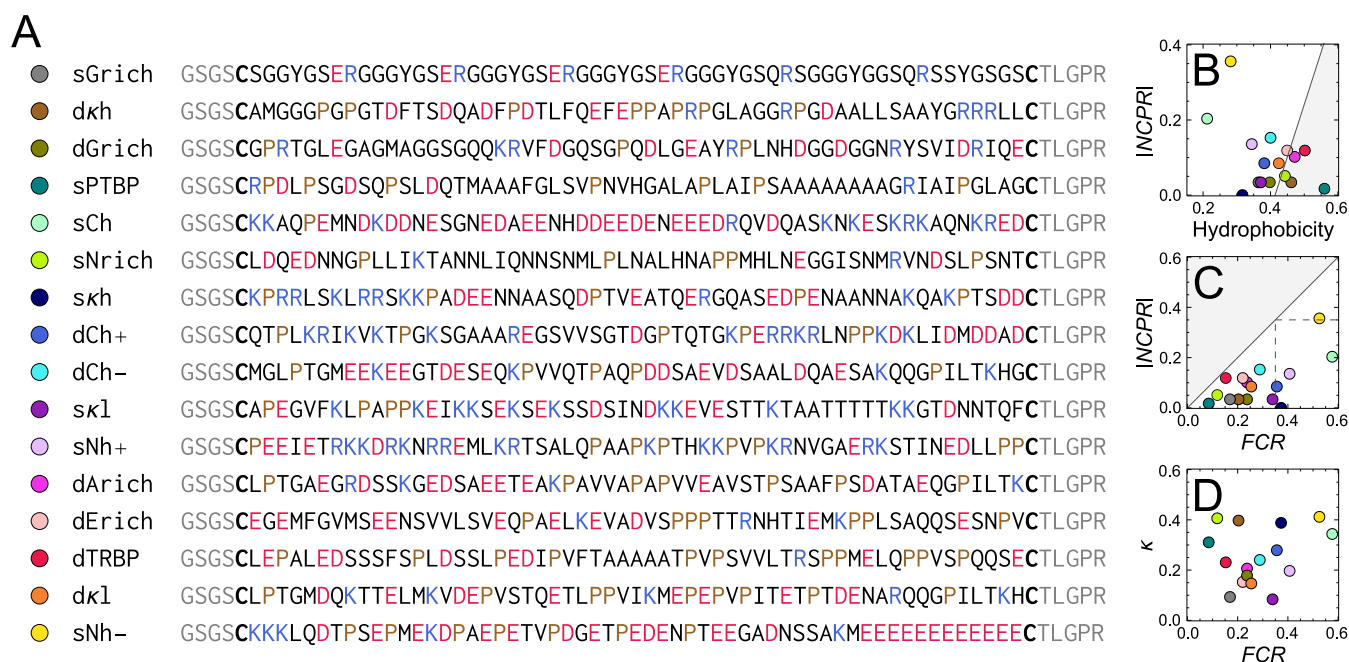
**Received:** July 25, 2024

**Revised:** September 13, 2024

**Accepted:** October 9, 2024

**Published:** November 14, 2024





**Figure 1.** Sequences and sequence properties of the selected IDRs. (A) Selected sequences of IDRs, with acidic amino acids shown in red, basic residues in blue, Pro residues in brown, and Cys residues used for labeling in bold. The residues terminal of Cys (gray) are a result of the cloning/expression strategy used. Those residues and Cys were not included for calculating sequence parameters. (B–D) Average sequence properties cover a wide range. Hydrophobicity, net charge per residue (NCPRI), and fraction of charged residues (FCR) are normalized by the total number of residues in each sequence. (B) The line indicates the separation of IDRs and folded proteins (gray region) suggested by Uversky et al.<sup>47</sup> (C) Most selected IDRs fall in the region of weak polyampholytes and polyelectrolytes with  $FCR \leq 0.35$  and  $INCPR \leq 0.35$ . IDRs with  $FCR > 0.35$  and  $INCPR \leq 0.35$  are considered strong polyampholytes, and those with  $FCR > 0.35$  and  $INCPR > 0.35$  strong polyelectrolytes<sup>19</sup> (dashed lines). The gray region cannot be populated. (D) The charge patterning metric  $\kappa$  describes the distribution of charged amino acids along the chain.<sup>10</sup>

observed in simulations and experiments.<sup>9,16,23–25</sup> Examples are highly charged sequences with pronounced electrostatic repulsion,<sup>9,14,26</sup> which can approach  $\nu \approx 1$  for rod-like conformations.<sup>27</sup> Other reasons for deviations from canonical scaling are finite-size effects<sup>28,29</sup> and heterogeneous patterns of intrachain interactions owing to the heteropolymeric nature of IDPs,<sup>12,30</sup> especially the contributions of high fractions of charged residues<sup>7–9</sup> and charge patterning.<sup>10</sup>

Considerable effort has been made to relate the dimensions of IDPs to their sequence properties and enable a predictive understanding of how intrachain interactions determine the sizes and shapes of the IDPs. Emerging consensus suggests that sequences rich in hydrophobic residues and certain polar tracts tend to favor compaction, whereas sequences rich in charged residues and proline tend to be more expanded.<sup>3,14–16,18,31–35</sup> Polyelectrolytes dominated by a single type of charge are most expanded,<sup>7,8,31</sup> whereas the attraction of opposite charges in polyampholytes can lead to compaction or long-range structural preferences depending on the patterning of oppositely charged residues.<sup>7,10,11,36</sup> The use of coarse-grained models parametrized based on experimental results<sup>15,37–40</sup> has enabled steps toward the analysis of conformational distributions across entire proteomes.<sup>16,25</sup> However, the systematic quantitative assessment of sequence contributions and the parametrization of IDP models is still complicated by the heterogeneity of molecular systems that have been studied experimentally, which usually vary both in length and sequence composition, and are investigated under disparate solution conditions. To furnish a data set that avoids such limitations, we thus selected naturally occurring IDRs of identical lengths but with very different sequence properties and probed their intrachain distances by single-molecule Förster resonance

energy transfer (FRET). In selected cases, we used complementary experimental methods, especially small-angle X-ray scattering (SAXS) for quantifying chain dimensions and NMR spectroscopy for identifying residue-specific intrachain interactions. We analyzed the results using atomistic simulations based on the ABSINTH model<sup>41</sup> to identify main determinants of chain dimensions and used the data to optimize a coarse-grained IDP model with an explicit representation of the fluorophores.

## RESULTS

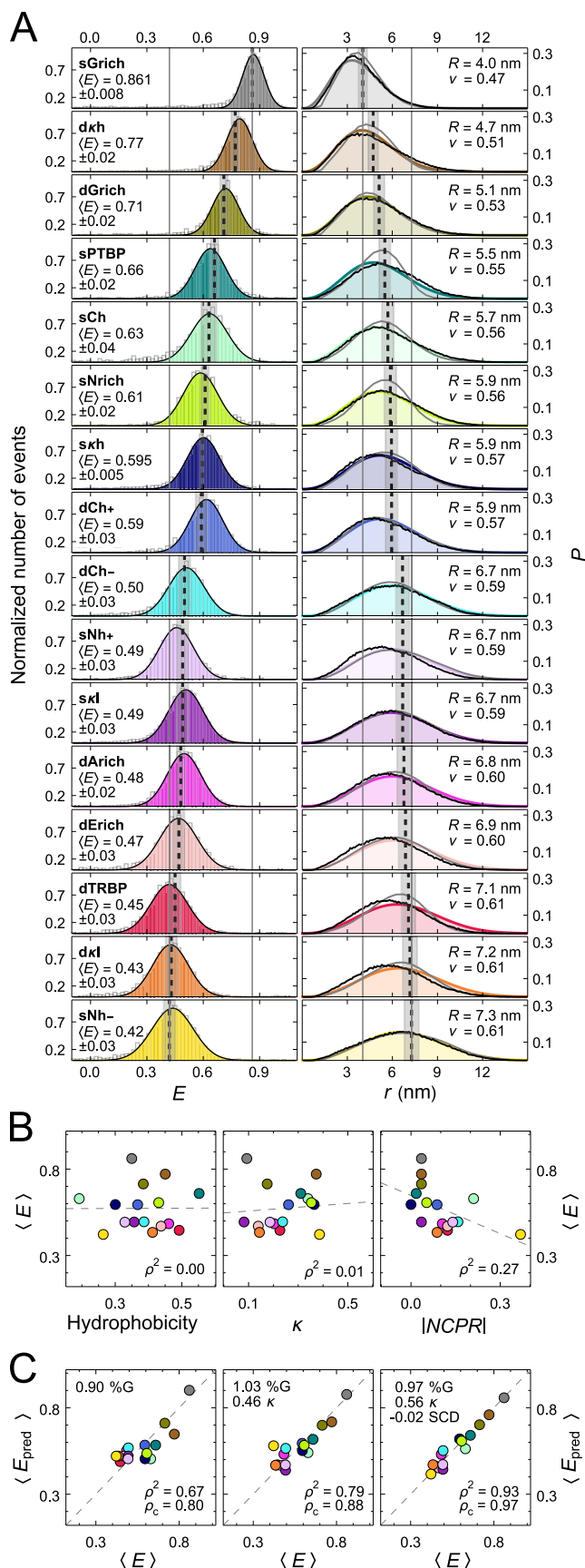
We selected 16 IDRs, each comprising 57 residues, from the linker regions connecting folded domains in RNA-binding proteins (Figure 1A, Figure S1). The large number of known RNA-binding proteins<sup>42</sup> allows for a wide variety of available sequences with very different physicochemical properties, while at the same time ensuring that the selected sequences are biologically relevant in terms of their amino acid composition. The sequence conservation of IDRs in RNA-binding proteins<sup>43</sup> attests to their functional importance beyond tethering of the folded domains. Their functions are possibly related to posttranslational modifications, interactions with the folded domains or RNA, or the optimal spacing of domains resulting from the sequence-encoded chain dimensions of the IDRs.<sup>44</sup> By scoring the corresponding linker sequences available in UniProtKB<sup>45</sup> for average hydrophobicity, net charge, fraction of charged residues, charge patterning, and amino acid composition (Figure S1), we identified examples that maximize the diversity of these properties (Figure 1B–D), from low to high hydrophobicity; from low to high net charge; from polyelectrolytes to

polyampholytes; from low to high charge segregation; and including examples enriched in individual amino acids, such as Gly, Glu, Ala, and Asn (see the [Materials and Methods](#) section for details). Only sequences with a disorder score in Metapredict<sup>46</sup> above 0.5 along the entire sequence were selected ([Figure S2](#)). Far-UV circular dichroism spectra of the recombinantly produced IDRs confirm the absence of pronounced secondary structure ([Figure S3](#)). The remaining differences between the spectra are suggestive of sequence-specific contributions to the conformational ensembles, but they are difficult to analyze quantitatively. Taken together, we have thus identified a set of naturally occurring IDRs of identical length that cover a broad spectrum of the key parameters commonly used to assess the properties of disordered proteins.

### Chain Dimensions Vary Widely among Sequences

The selected IDR sequences, bracketed with Cys residues for fluorophore labeling via maleimide chemistry, were expressed recombinantly, purified, and labeled with donor and acceptor dyes for single-molecule FRET. By working at picomolar protein concentrations in single-molecule measurements, we could avoid aggregation and phase separation, even for sequences with low solubility that are exceedingly difficult to investigate with ensemble experiments at high concentrations. Moreover, by resolving conformational subpopulations in single-molecule experiments, species such as small aggregates (which may go unnoticed in ensemble measurements) can be detected<sup>48</sup> and prevented by optimized sample handling or separated out in the analysis ([Figure S4](#)). We performed multiparameter confocal single-molecule measurements using pulsed interleaved excitation<sup>49</sup> with all 16 labeled IDRs and identified a single transfer efficiency peak under our experimental conditions for each sequence. The results reveal a remarkably broad range of intramolecular transfer efficiencies from  $\sim 0.4$  to  $\sim 0.9$  for the different IDRs despite their identical chain lengths ([Figure 2A](#)), reflecting the pronounced dependence of the chain dimensions on amino acid sequence.

We had previously observed that the charge of the fluorophores needs to be accounted for to quantitatively explain the dimensions of IDPs with simple polymer models.<sup>7,9</sup> We thus used two different FRET pairs with different net charges to assess such effects. One widely used pair comprises the dyes Alexa 488 and Alexa 594 (Förster radius  $R_0 = 5.4$  nm), each of which carries a net charge of  $-2$ ; the other pair comprises the dyes Cy3B and CF660R<sup>50</sup> ( $R_0 = 6.0$  nm), which carry a net charge of  $0$  and  $-1$ , respectively ([Figure S5](#)). We find that protein sequences rich in basic residues yield lower average intramolecular distances when labeled with the more negatively charged Alexa pair, whereas other sequences yield very similar results for both dye pairs ([Figure S6A](#)). Similarly, SAXS measurements of the Alexa 488-labeled IDR dCh $-$  showed the increase in  $R_g$  expected from the addition of the fluorophore compared to unlabeled dCh $-$ , but for sNh $+$  the increase was much smaller ([Figure S6B](#)). NMR spectroscopy confirmed the presence of more attractive interactions between positively charged residues and the Alexa fluorophores than with Cy3B and CF660R ([Figure S6C,D](#)). The large range of transfer efficiencies and chain dimensions we observe is robust with respect to the dye pair used, but to minimize the influence of the dyes on the FRET-based assessment of chain dimensions, we focus on the results obtained with Cy3B/CF660R.



**Figure 2.** Chain dimensions from single-molecule FRET and correlations with sequence parameters. (A, left) Transfer efficiency histograms of Cy3B/CF660R-labeled IDRs (gray) fit with Gaussian peak functions (color). The average of the mean transfer efficiency,

Figure 2. continued

$\langle E \rangle$ , from at least three experiments is indicated by a black dashed line, with the standard deviation shown as a gray band. Vertical gray lines indicate the lowest and highest average  $\langle E \rangle$  observed across the series. (A, right) Distance distributions based on the SAW- $\nu$  model<sup>26</sup> (colored line), reweighted ensembles from ABSINTH<sup>41</sup> simulations (gray line), and from an optimized coarse-grained model (black line). Vertical dashed lines indicate the average root-mean-squared distance ( $R$ ) from SAW- $\nu$ , with a gray error band based on a systematic uncertainty of  $\pm 7\%$  in the Förster radius.<sup>39,53</sup> Values for the average  $R$  and average  $\nu$  are shown in the top right corners. Vertical gray lines indicate the largest and smallest values of  $R$  across the series. (B) Correlations between different sequence parameters and average transfer efficiency. The dashed line and the coefficient of determination,  $\rho^2$ , were obtained from linear regression. (C) Multiple linear regression was used to identify combinations of sequence parameters that maximize the correlation with transfer efficiency. The regression coefficients for each of the sequence parameters used as regressors (%G,  $\kappa$ , and SCD) are indicated. The dashed line is the identity line, and  $\rho_c$  is the concordance correlation coefficient. Color code for the sequences is given in Figure 1.

Using the SAW- $\nu$  model, a semiempirical approximation with an adjustable length scaling exponent<sup>26</sup> to infer intramolecular distance distributions for the different sequences, we obtained root-mean-squared end-to-end distances,  $R$ , between 4.0 and 7.3 nm (Figure 2A), corresponding to almost a factor of 2 between the most compact (sGrich) and the most expanded IDR (sNh-), and a factor of  $\sim 6$  in chain volume. The inferred average scaling exponents,<sup>26</sup>  $\nu$ , are between 0.47 to 0.61, corresponding to the range from effective theta conditions to excluded volume chains.<sup>18,22</sup> Although a detailed interpretation of these scaling exponents is complicated by finite-size effects<sup>29</sup> and the contributions from heterogeneous interaction patterns within heteropolymers,<sup>12,30</sup> they imply that these IDRs are rather open chains and more expanded than a compact globule. However, it is worth noting that the two Gly-rich sequences sGrich and dGrich are among the most compact of the set, suggesting an important role for Gly in chain compaction.

It may not be surprising that highly charged sequences rank among the most expanded chains,<sup>7,8</sup> and correspondingly, the average net charge per residue (NCPR) shows a correlation with the observed transfer efficiency (Figure 2B). Average Kyte–Doolittle hydrophobicity<sup>51</sup> shows remarkably little correlation with transfer efficiency, which is likely to be connected to the requirement of alternative hydrophobicity scales to describe protein phase separation.<sup>38,40</sup> Similarly, the transfer efficiency shows little correlation with the charge patterning parameter  $\kappa$ , which only applies to sequences with high fractions of charged residues<sup>52</sup> (Figure 2B). However, sequence composition clearly influences the chain dimensions; examples of individual residues whose content in the sequences correlates with transfer efficiency with a coefficient of determination of  $\rho^2 \geq 0.36$  are Gly (favoring compaction), Arg (favoring compaction), Val (favoring compaction), Thr (favoring expansion), and Pro (favoring expansion) (Figure S6).

In view of these correlations, we thus asked how well the transfer efficiencies correlate with combined compositional biases. For instance, multiple linear regression combining the fraction of Gly and  $\kappa$  as regressors yields  $\rho^2 = 0.79$ , i.e., 79% of the variance in the observed transfer efficiencies can be

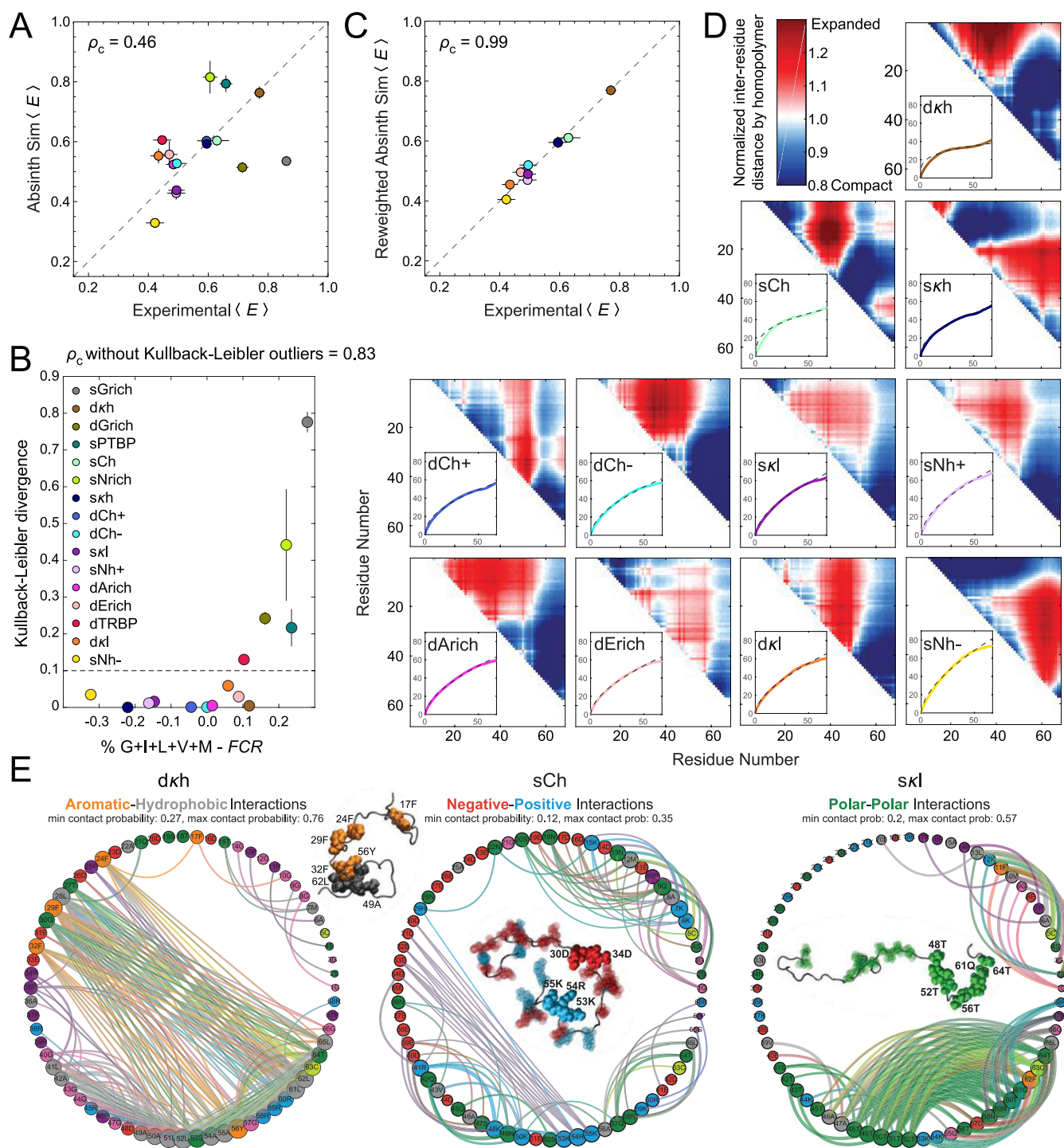
accounted for with this combination alone. Combining the fraction of Gly with  $\kappa$  and sequence charge decoration,<sup>11</sup> SCD, yields an even higher  $\rho^2$  value of 0.93 (Figure 2C). However, based on a leave-one-out analysis to identify the dominant contributions (Figure S7), we find that individual members of the data set have a large effect on the result; for instance, including sGrich greatly improves the  $\rho^2$  for linear regressions that account for the fractions of Gly or Tyr. These correlation analyses clearly show an important effect of sequence composition on chain dimensions and can thus provide interesting clues about which residues or sequence characteristics are relevant. However, the set of IDRs investigated here cannot provide sufficiently broad sampling of sequence space to uniquely define chain dimensions based on regression analysis alone. We thus turned to simulations for a more detailed analysis.

### Atomic-Level Characterization of Conformational Ensembles from Simulations

Key discoveries regarding sequence-ensemble relationships of IDPs have been made using atomistic simulations based on the ABSINTH implicit solvation model and force field paradigm.<sup>41</sup> In the ABSINTH model, polypeptides and solution ions are represented in atomistic detail, and the aqueous solvent is modeled implicitly. Measured free energies of solvation serve as a benchmark against which solvation inhomogeneities are calibrated. These inhomogeneities are gleaned by using solvent-accessible volumes, and the changes to solvation are balanced by changes to the effective charge, which is an efficient way of capturing dielectric inhomogeneities. To compare end-to-end distances from ABSINTH simulations to those inferred from FRET measurements, an atomistic representation of the fluorophores was included based on a rotamer library that takes into account dye configurations that are sterically allowed (see the Materials and Methods section). To compare FRET efficiencies from simulations and measurements (Figure 3A), we computed the concordance correlation coefficient<sup>34</sup> ( $\rho_c$ ), which combines correlation (precision) and deviation from perfect concordance (accuracy), yielding  $\rho_c = 0.46$  for the ABSINTH ensembles.

Overall, ABSINTH captures certain overall trends from the FRET measurements, but there are clear deviations from the experimental results. To better understand where ABSINTH fails and succeeds with the current data set, the conformational ensembles were reweighted, as described previously<sup>12</sup> (Figure 3C). We then compared the unweighted (prior) ensembles to the reweighted (posterior) ensembles. This analysis allowed us to assess whether there were specific sequence features that mandate substantial reweighting of the ABSINTH-derived ensembles when comparing the computed and measured FRET efficiencies. The results of this analysis are presented in terms of the Kullback–Leibler (KL) divergence between the unweighted and reweighted ensembles (Figure 3B), which is below 0.06 for 11 of the 16 sequences and greater than 0.1 for five of the sequences. The largest divergences result for the two Gly-rich sequences and the N-rich sequence. The two sequences with large fractions of aliphatic residues also show KL divergences above 0.1. Omitting the sequences with KL divergences above 0.1 from the analysis yields  $\rho_c = 0.83$  for the unweighted and  $\rho_c = 0.99$  for the reweighted ensembles (Figure 3C).

The most compact IDR observed experimentally, sGrich, contains several stretches rich in Gly. Water is a poor solvent



**Figure 3.** ABSINTH simulations provide atomic-level characterizations of conformational ensembles. (A) Correlation between the mean transfer efficiencies,  $\langle E \rangle$ , from experiment and unweighted ABSINTH simulations ( $\rho_c$ : concordance correlation coefficient). (B) Kullback–Leibler divergence quantifying the deviations between the unweighted (prior) and reweighted (posterior) ensembles obtained with ABSINTH, plotted as a function of the fraction of Gly and hydrophobic residues minus the fraction of charged residues ( $FCR$ ) of the sequences. Minimal deviations have a KL divergence of  $\leq 0.1$ . Based on this cutoff, five of the 16 sequences, characterized by high Gly or high aliphatic content and lower charge content, require substantial reweighting (sPTBP, dTRBP, sGrich, sNrich, and dGrich). (C) Correlation between  $\langle E \rangle$  from experiment and reweighted ABSINTH simulations for sequences with KL divergence  $< 0.1$ . (D) Average inter-residue distances from reweighted ABSINTH simulations (minimum spacing of 10 amino acids), relative to the value from the best fit of a homopolymer model (see color scale) determined by  $R_{ij} = \langle r_{ij}^2 \rangle^{1/2} = A_0 |i - j|^\nu$ , where  $r_{ij}$  is the spatial distance between residues  $i$  and  $j$ ,  $A_0$  is an adjustable prefactor that reports on the chain persistence length,  $|i - j|$  is the linear sequence separation between residues  $i$  and  $j$ , and  $\nu$  is the scaling exponent. Regions of local expansion relative to the equivalent homopolymer are shown in red, and areas of local compaction are in blue. The insets show ensemble-averaged inter-residue distances,  $\langle R_{ij} \rangle = \langle \langle r_{ij}^2 \rangle^{1/2} \rangle$  (in Å), versus  $|i - j|$  (colored line). The best homopolymer fit is shown as a dashed gray line. Here, the double average implies averaging over all pairs of residues  $i$  and  $j$  that are  $|i - j|$  apart in the sequence (the outer average), and the spatial separations,  $R$ , between specific pairs of residues across all conformations in the ensemble (the inner average). Only IDRs with a KL divergence  $\leq 0.1$  are shown. (E) Contact networks

Figure 3. continued

illustrate different interaction preferences for different IDRs. Residues are shown as nodes, with the circle size related to the mean contact probability between that residue and all other residues more than two residues away in sequence. Edges are drawn between two residues if their contact probability is at least 35% of the maximum contact probability observed for that IDR. Specifically, edges are shown for mean contact probabilities between 0.27 and 0.76, 0.12 and 0.35, and 0.2 and 0.57 for  $\alpha$ Ch, sCh, and sL, respectively. The width of an edge is 10 times the mean contact probability. Here, a contact distance of 10 Å is used, such that charge interactions can be observed. Gly is shown in pink, Ser, Thr, Asn, and Gln in green, Arg, Lys, and His in blue, Asp and Glu in red, Phe, Trp, and Tyr in orange, Met, Val, Ile, Leu, and Ala in gray, Pro in purple, and Cys in lime green. Edge colors are mixtures of interacting residue colors. Representative snapshots are visualized using VMD<sup>62</sup> and chosen by finding the frame that has the highest weight with a radius of gyration ( $R_g$ ) within 0.5 Å of the average  $R_g$  for the IDR. Contact networks for the remaining proteins are shown in Figure S10.

for polyglycine,<sup>20,55</sup> polyglutamine,<sup>6</sup> and other types of polar tracts.<sup>56</sup> The preference of Gly-<sup>57</sup> and Gln-rich sequences<sup>58</sup> for collapsed conformations and their low solubility has been predicted and computed using all-atom simulations<sup>59</sup> with different types of force fields,<sup>60</sup> and observed experimentally.<sup>20</sup> The challenges arise in ABSINTH for Gly- and Asn-rich sequences because of the delicate interplay between favorable hydration of the polar backbone and side chains and the favorable intrachain interactions between polar groups, even though ABSINTH does not have a challenge with Gln-rich or Gln- and Asn-rich sequences.<sup>61</sup>

Before analyzing the reweighted ensembles in detail, we computed three different parameters that quantify the ensemble-averaged global sizes and shapes for each of the IDRs using the ABSINTH-derived prior and posterior ensembles (Figure S8). First, we quantified the correspondence for the global radius of gyration ( $R_g$ ). Overall, the deviations are minimal, with the two largest outliers being the Gly-rich sequences (Figure S8). Next, we computed how the overall shape changes upon reweighting by computing the ensemble-averaged asphericity (Figure S8). Here, we observed a few more deviations compared to  $R_g$ , but the general consistency in compaction and expansion suggests that the sequence controls the local interactions and deviations from a homopolymer. Finally, we compared the root-mean-squared end-to-end distances,  $R$ , for unweighted versus reweighted ensembles. We find that the sequences with high KL divergence fall outside the 95% confidence interval. Across the three parameters, the deviation between unweighted and reweighted ensembles is the largest when we compare  $R$ .

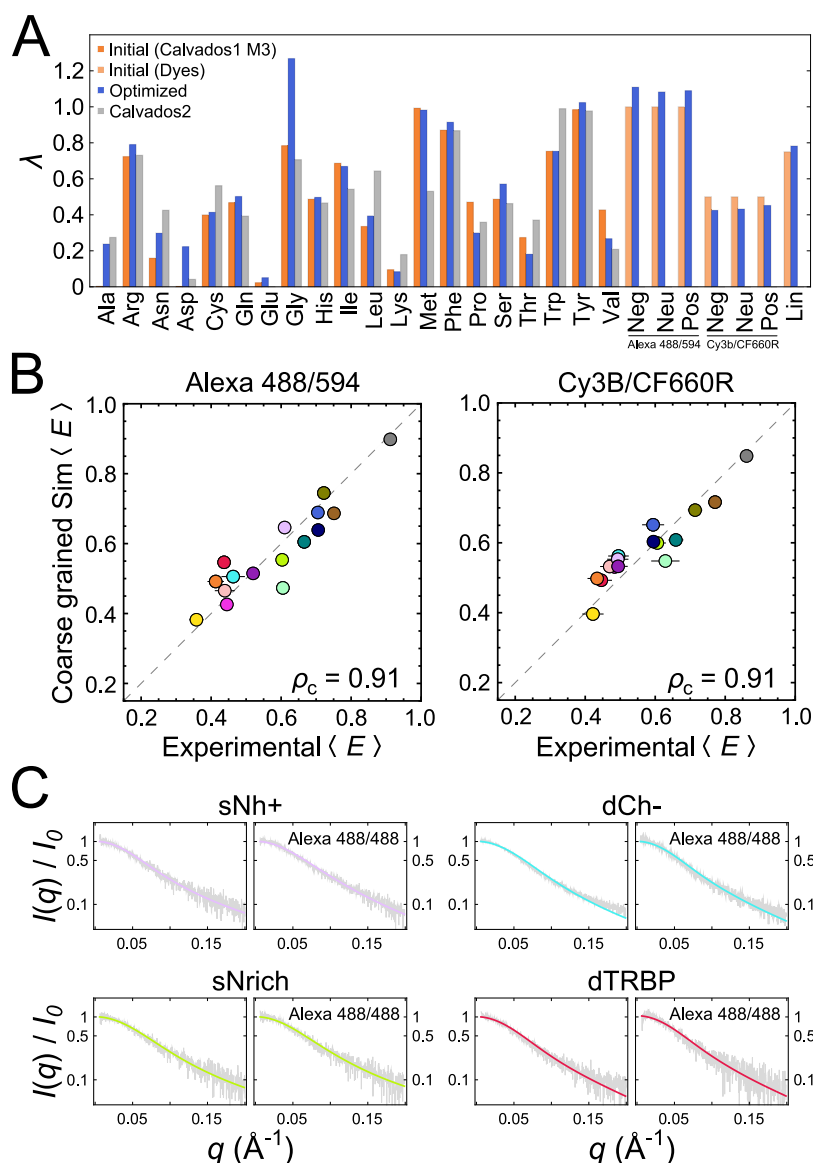
The preceding analysis suggests that robust inferences regarding conformational ensembles can be drawn from either the reweighted or unweighted ABSINTH simulations for 11 out of the 16 sequences, which we analyze in more detail (Figure 3D). We quantified the scaling of average inter-residue distances ( $\langle R_{ij} \rangle = \langle \langle r_{ij}^2 \rangle^{1/2} \rangle$ ) between residues  $i$  and  $j$  with an alternative approach to describing sequence separation ( $|i - j|$ ) (insets, Figure 3D). To determine how well the conformational ensembles can be described by homopolymer models, the standard polymer relationship  $\langle R_{ij} \rangle = A_0 |i - j|^\nu$  was fit to extract  $A_0$  and  $\nu$ , the apparent scaling exponent (insets in Figure 3D, dashed gray line). Although the value of  $\nu$  is not meaningful for internal scaling profiles that show plateauing behavior, this type of comparison can still be helpful to determine whether there are nonuniform interactions along a chain. Therefore, we examined the distance between residues normalized by the best-fit homopolymer models (Figure 3D). All IDRs show deviations from the uniform length scaling expected for homopolymers, as shown by regions along the sequence of compaction (blue) or expansion (red) compared to the IDR-specific homopolymer model. Additionally, the

degree of compaction or expansion relative to an equivalent homopolymer reference can vary along the sequence (Figure 3D, Figure S9), highlighting the heterogeneity of interactions within each sequence. Overall, these results suggest that even though global features, such as end-to-end distance, can be correlated with sequence composition, additional analyses from atomistic simulations can provide detailed insights into the heteropolymeric properties of the IDRs.

### Optimizing a Coarse-Grained IDP Model with Explicit Fluorophores

An alternative approach to describing sequence-ensemble relationships is to use the experimental data to variationally optimize the parameters of the simulation model itself.<sup>15,39</sup> Coarse-grained implicit solvent models, which have only a limited set of adjustable parameters, are naturally suited to this approach.<sup>15,16,37,39,40</sup> In this case, the amino acids are represented by beads with the appropriate volumes and charges, whose interactions are accounted for via a screened Coulomb potential and a residue-specific short-range potential that represents interactions such as hydrophobicity or hydrogen bonding in addition to the excluded volume. Force field parameters for such models have previously been optimized based on statistical potentials and/or by comparison with experimental data.<sup>37,38,40</sup> However, in general, the experimental data employed have been collected under different solution conditions (e.g., different temperatures, pHs, salt concentrations), complicating their coherent use for model refinement. Data sets comprising a large sequence diversity of IDRs measured under identical solution conditions are therefore expected to be useful for benchmarking and refining simulation models and parameters that are transferable to a wide range of IDPs.

We thus employed the experimental results from our 16 IDRs to identify the residue-specific short-range interaction parameters for a hydrophobicity scale (HPS) model<sup>38,63</sup> that best describe the entire data set. To this end, we use the values of Tesei et al.<sup>40</sup> as a starting point and employ the force balance approach,<sup>38,64,65</sup> where the short-range interaction parameters  $\lambda$  are iterated to optimize agreement between the simulated and experimental transfer efficiencies. A particular advantage of this method is that the fluorophores and their interactions with the rest of the sequence can be included in the model explicitly and parametrized with the same strategy. This approach thus enables interactions of the fluorophores to be taken into account that go beyond the excluded volume effects most commonly considered in accessible volume<sup>66,67</sup> and rotamer library<sup>50,68–71</sup> schemes. We chose a dye representation that reflects the size, structure, and charge distribution of the different fluorophores and consists of charged, uncharged, and dye linker beads (Figure S5).



**Figure 4.** Coarse-grained simulations consistently describe amino acid and dye interactions. (A) Initial (orange for amino acids and light orange for dyes), optimized (blue), CALVADOS 2<sup>72</sup> (gray) values of the short-range interaction parameter ( $\lambda$ ) for each amino acid, dye (Alexa 488/594 and Cy3B/CF660R; Neg: negatively charged, Neu: neutral, Pos: positively charged), and dye linker (Lin) beads. Note that the initial values for Ala and Asp are too small to be visible on this scale (Table S3). (B) Correlation between the mean transfer efficiencies from experimental data and from resulting coarse-grained simulations for both dye pairs used. The dashed line is the identity line, and  $\rho_c$  is the concordance correlation coefficient. (C) SAXS curves of selected IDRs (gray lines) compared to results based on the simulations (colored lines). For each IDR, results for unlabeled protein are shown in the left plot, and results for protein labeled with Alexa 488 at both Cys residues are shown in the right plot. The simulation value at a  $q$  of 0 was used for normalization of both experiments and simulations.

With the optimized parameters for the HPS model (Figure 4A), we obtained correlations with  $\rho_c = 0.91$  between experiment and simulation for the IDRs labeled with either dye pair (Figure 4B), compared with  $\rho_c = 0.70$  and  $0.64$  before optimization. The short-range interaction parameters for the amino acids after optimization are reasonably close to the starting values, which were obtained previously based on experimental data and hydrophobicity scales of amino acids;<sup>40</sup> indeed, the new parameters yield similar results when benchmarked against the original CALVADOS training set<sup>40</sup> (Figure S11). The increased value for Gly reflects the pronounced compaction of Gly-rich sequences we observe experimentally, and the increased values for Arg and Tyr are in line with previous results suggesting an important role of these

residues for chain compaction and phase separation.<sup>16,40,73–76</sup>

We note, however, that while the CALVADOS parameters slightly underestimate compaction for Gly-rich sequences in the present data set, our HPS parameters slightly overestimate compaction for the longer Gly-rich sequences in the CALVADOS training set, a conflict that likely points to limitations of the form of the HPS model itself. Importantly, by essentially treating the fluorophores as part of the sequence, both the results for the Alexa dye pair and Cy3B/CF660R can be described well. Although a detailed rationalization of the resulting short-range interaction parameters for the fluorophores based on their chemical structure is challenging at this level of coarse-graining, the larger values for the Alexa dyes are in accord with the stronger dye-peptide interactions observed

in the SAXS and NMR data compared to Cy3B/CF660R (Figure S6).

We further tested the optimized HPS parameters by calculating scattering curves from coarse-grained simulations of several IDRs with and without fluorophores and comparing with SAXS measurements of labeled and unlabeled sNrh+, dCh-, sNrich, and dTRBP, and we found reasonable agreement (Figure 4C), further validating the model. As additional tests on independent sequences, we used previously published SAXS data on a series of IDPs<sup>12</sup> and a mutationally destabilized variant of the  $\beta$ -helix protein PNT labeled with zero, one, or two copies of Alexa 488;<sup>77</sup> overall, the data are well described by our model. In the case of PNT, the results indicate a very moderate decrease in radius of gyration upon attaching one or two dyes (although limitations in SAXS data quality yield a large uncertainty for the double-labeled variant, Figure S12A). For the same PNT variant with Alexa 488 and 594 attached (at the same sites as for the PNT variant doubly labeled with Alexa 488<sup>77</sup>), the measured transfer efficiency ( $\langle E \rangle = 0.6$ ; Figure S12B) is reasonably reproduced by the HPS model ( $\langle E \rangle = 0.52$ ). Altogether, the optimized coarse-grained model thus provides a residue-specific way of quantifying the dimensions of disordered proteins. Moreover, the fluorophores can be incorporated and their interactions parametrized within the same framework, treating them essentially as an additional set of residues, so that the effect of the FRET dyes on chain dimensions can be predicted and distances and distance distributions between the dyes can be obtained directly from the simulations and compared to experiment.

Finally, we compared end-to-end distance distributions resulting from the three different methods employed: the analytical SAW- $\nu$  polymer model, the optimized coarse-grained HPS model, and the reweighted atomistic ABSINTH simulations (Figure 2A). The distributions are similar in all cases, indicating the consistency of the different approaches at the level of the overall chain dimensions. For a more detailed comparison, we calculated distance maps from the HPS simulations and the reweighted ABSINTH simulations (Figure S13). This analysis reveals three groups of sequences. For nine of the 16 sequences, we find strong positive correlations between the normalized distance maps for the reweighted ABSINTH and HPS models (Pearson correlation coefficient  $\rho > 0.6$ ); for four of the 16 sequences, weak positive correlations ( $0.3 < \rho < 0.5$ ); and for three of the sequences—sGrich, sPTBP, and sNrich—we observe anticorrelation ( $\rho < 0$ ). These are also the IDRs for which extensive reweighting was required for the ABSINTH ensembles (Figure 3B), but for most of the IDRs, the intrachain interactions predicted by the models are similar. The discrepant cases highlight the challenges associated with the interplay between chain-solvent and intrachain interactions in arriving at a consistent description of ensembles for Gly- and Asn-rich sequences,<sup>61</sup> and for sequences where secondary structural preferences in the ABSINTH model cause deviations, such as sPTBP. Indeed, the circular dichroism spectra of sPTBP and some other sequences (Figure S3) show hints of residual secondary structure.

## DISCUSSION

We investigated a set of 16 IDRs selected from linker sequences of naturally occurring proteins with identical lengths but very different sequence compositions to probe the sequence dependence of the conformational ensembles of

disordered proteins. Notably, since all sequences investigated here originate from the linker regions between RNA-binding domains, their chain dimensions may have been an evolutionary factor contributing to the average distance between the domains<sup>44</sup> and their interaction with RNA. The experimental results, consisting of single-molecule FRET efficiencies measured with two different dye pairs and complemented with SAXS and NMR, serve as a benchmark and provide an opportunity for systematically refining simulation models and force field parameters. Here, we tested and compared three approaches at very different levels of coarse-graining for modeling the conformational ensembles of 16 IDRs.

Analytical homopolymer models can be useful for inferring overall distance distributions and effective length scaling exponents; although they cannot provide details on heterogeneities in local compaction or expansion along the sequence, they are a simple and useful way of interpreting experimental data in terms of distance distributions,<sup>17,18</sup> but their predictive power is limited. Simple correlations between chain dimensions and sequence composition also offer useful indications of the effect of individual residues on chain compaction, but simulations provide much more detail regarding the heteropolymer properties of disordered proteins. Our application of two different simulation approaches demonstrates the complementarity of atomistic and coarse-grained models. All-atom implicit solvent simulations using ABSINTH<sup>41,61,78</sup> in combination with ensemble reweighting<sup>12</sup> enable detailed analyses of residue-specific intrachain interaction networks that affect chain compaction and the resulting deviations from simple homopolymer models (Figure 3). Coarse-grained models facilitate the optimization of force field parameters to arrive at a transferable model, illustrated here with the HPS model<sup>38,40,63</sup> combined with the force balance approach<sup>38,64,65</sup> (Figure 4). We further show how FRET dyes can be incorporated and parametrized explicitly in the HPS model to achieve agreement between results using different fluorophores (Figure 4A,B). Overall, our work thus illustrates the mutual benefit of experiment and simulations: experimental data enable the testing and refinement of simulation models, and simulations enable a detailed structural interpretation of the experimental results.

## CONCLUSIONS

Our results demonstrate that the dimensions of IDRs exhibit a pronounced dependence on amino acid sequence. This result is consistent with a broad range of previous observations,<sup>7–10,13,14,16,18,19,24,31,34,79–81</sup> but a noteworthy feature of the current work is that we focused on sequences of identical length, thus ensuring that differences in sequence-ensemble relationships do not arise from additional structure or sequence context. The measurements were performed under identical solution conditions, such as buffer, salt concentration, and pH, which greatly simplifies their direct quantitative comparison. Furthermore, the sequences were selected from natural proteins to ensure the biological relevance of their sequence compositions and to represent a broad range of sequence characteristics, which allows many types of effects to be accounted for. The analysis of the results using polymer models and both all-atom and coarse-grained simulations consistently shows that the chain dimensions and conformational properties cover a very broad range. Particularly pronounced contributions to chain compaction come from the content in Gly and aromatic residues; contributions to



chain expansion come from charge repulsion and Pro residues. However, we did not identify a simple single descriptor that captures global chain dimensions, but compaction can be driven by different types of interactions in different sequences. From our data, we have derived a coarse-grained model that can be used to predict how such interactions affect the dimensions of other disordered proteins.

## MATERIALS AND METHODS

### Sequence Selection and Characterization

To identify disordered protein regions, UniProtKB was searched for proteins containing at least two double-stranded RNA-binding domains,<sup>82</sup> and sequences of interdomain linkers were identified that were 50–200 amino acids in length. In order to increase the compositional diversity of the sequences, a second pool of sequences was generated from proteins containing at least two RNA recognition motifs (RRMs).<sup>83</sup> All sequences were characterized in terms of the following sequence properties. Normalized hydrophobicity was calculated using the scale of Kyte and Doolittle, which assigns a relative hydrophobicity index,  $H_i$ , between  $-4.5$  and  $+4.5$  to each amino acid<sup>51</sup>

$$\text{hydrophobicity} = \frac{1}{N} \sum_{i=1}^{20} n_i \left( \frac{H_i}{9} + 0.5 \right)$$

Here,  $N$  is the total number of amino acids in the sequence and  $n_i$  is the number of each of the 20 amino acid types within the polypeptide chain. The normalized hydrophobicity may adopt values between 0 and 1. The fraction of charged residues,  $FCR$ , and the net charge per residue,  $NCPR$ , were calculated according to Das and Pappu<sup>10</sup> as

$$FCR = f_+ + f_-$$

$$NCPR = f_+ - f_-$$

where  $f_+$  and  $f_-$  denote the fractions of positively and negatively charged residues, respectively. We calculate the charge patterning factor  $\kappa$  and sequence charge decoration<sup>11</sup> (SCD) as described previously. A total of 16 linker sequences were selected to cover a large sequence parameter space. All sequences were shortened to 57 amino acids to rule out length-dependent effects in their comparison. In this process, care was taken to alter the average sequence properties as little as possible. Sequences containing Trp were excluded to minimize effects from dye quenching that can complicate quantitative analysis of FRET experiments.<sup>84</sup> Two natural Cys residues were replaced by Ser in dErich, and sKl contains a spontaneous Ser to Ile exchange due to the instability of the gene in *Escherichia coli*. The naming of the IDRs was chosen to be suggestive of characteristic sequence properties ('s': derived from ssRNA binding proteins, 'd': derived from dsRNA binding proteins, 'h': high, 'l': low, 'N': net charge, 'C': charge, '+' : positively charged, '-' : negatively charged, " $\kappa$ ": charge segregation, "Xrich": enriched in amino acid X). All sequences are shown in Figure 1, and the UniProt codes of the source proteins and the sequence parameters are listed in Table S1.

Multiple linear regression was performed for all single, double, and triple combinations of the following 26 compositional features: Fraction A, D, E, G, K, L, N, P, Q, R, S, T, V, K+R, D+E, Polar, Aliphatic, Aromatic, and Chain Expanding, as well as  $FCR$ ,  $INCPR$ , Hydrophobicity, Disorder Promoting, Isoelectric point,  $\kappa$ , and SCD. Compositional features were calculated using localCIDER.<sup>85</sup> Fractions of C, F, H, I, M, W, and Y were not considered, as they each account for less than 2.5% of all residues in the linker IDR sequences. For Figure S7A, the  $\rho^2$  values are shown for all 26 single compositional features, all double combinations of compositional features with a  $\rho^2 > 0.72$ , and all triple combinations of compositional features with a  $\rho^2 > 0.855$ . The boxplots in Figure S7A show the distributions of  $\rho^2$  values for all 16 leave-one-out analyses.

### Protein Expression, Purification, and Fluorescence Labeling

Codon-optimized DNA sequences encoding the IDRs with two terminal Cys residues for site-specific dye labeling were purchased from GeneArt (Regensburg, Germany). Linker IDR sequences were cloned into a pET-20b(+) based plasmid (EMD Millipore), which contained an N-terminal His<sub>6</sub>-tag, as well as a C-terminal GB1 domain for improved expression fused to a His<sub>6</sub>-tag, both of which were separated from the IDR of interest via a thrombin cleavage site.<sup>86</sup> For all constructs, thrombin cleavage resulted in a residual GSGSC overhang at the N-terminus and a CTLGPR overhang at the C-terminus of the protein.

The IDRs were expressed in *E. coli* Rosetta (DE3) cells (Merck Biosciences). Cultures were grown to an OD<sub>600</sub> of 0.8 in LB medium containing carbenicillin, induced with 1 mM IPTG, and incubated at 20 °C overnight. For the preparation of isotope-labeled proteins, M9 minimal medium containing <sup>15</sup>NH<sub>4</sub>Cl or <sup>13</sup>C<sub>6</sub>-glucose was used instead of LB medium. Cells were harvested, and pellets were resuspended in lysis buffer (100 mM NaH<sub>2</sub>PO<sub>4</sub>/Na<sub>2</sub>HPO<sub>4</sub>, 10 mM Tris-HCl, 6 M guanidinium chloride (GdmCl), 10 mM imidazole, 1 mM Tris(2-carboxyethyl)phosphine (TCEP), pH 8.0). Insoluble cell debris was removed by centrifugation. The soluble fraction was subjected to Nickel chelate affinity chromatography (Ni Sepharose excel, GE Healthcare Bio-Sciences). The lysis buffer was used for washing. For elution, the imidazole concentration was increased to 500 mM. Eluates were dialyzed against 50 mM Tris-HCl, 150 mM NaCl, and 10% (v/v) glycerol, pH 8.0, and the total protein concentration was quantified by measuring the absorbance at 280 nm (extinction coefficients: sGrich-GB1, 20,400 M<sup>-1</sup> cm<sup>-1</sup>; dGrich-GB1, 12,950 M<sup>-1</sup> cm<sup>-1</sup>; dkh-GB1, 11,460 M<sup>-1</sup> cm<sup>-1</sup>; all others: 9970 M<sup>-1</sup> cm<sup>-1</sup>). Subsequently, thrombin was added at 20 U/mg and the proteolytic digest was allowed to proceed for 1–3 h at room temperature. The reaction was quenched by adding 1 g/mL GdmCl. Protein solutions were then concentrated to a total volume of approximately 1 mL using Centrprep 3K centrifugal filter devices (EMD Millipore).

Protein samples were reduced by adding DTT at a final concentration of 10 mM and purified by reversed-phase high-performance liquid chromatography (RP-HPLC) on a C18 column (Reprosil Gold 200, Dr. Maisch GmbH) using 5% acetonitrile, 0.1% (v/v) trifluoroacetic acid (TFA) as buffer A and acetonitrile as buffer B. Eluates were lyophilized overnight and redissolved in 20 mM KH<sub>2</sub>PO<sub>4</sub>/K<sub>2</sub>HPO<sub>4</sub>, 6 M GdmCl, pH 7.3. Protein concentrations were quantified using a bicinchoninic acid (BCA) assay kit (Thermo Fisher Scientific Inc.) by measuring the absorbance at 562 nm. For fluorescence labeling, Alexa Fluor 488 C5 maleimide (Thermo Fisher Scientific Inc.) or maleimide-functionalized Cy3B (GE Healthcare AG) dissolved in anhydrous *N,N*-dimethylformamide (DMF) was added at a molar ratio of protein to dye of 1:0.7. The reaction was allowed to proceed at 4 °C overnight and quenched by the addition of DTT at a final concentration of 10 mM. Singly labeled protein was separated from unreacted and doubly labeled protein by RP-HPLC (see above), followed by lyophilization. Donor-labeled protein was redissolved in 20 mM KH<sub>2</sub>PO<sub>4</sub>/K<sub>2</sub>HPO<sub>4</sub>, 6 M GdmCl, pH 7.3. Alexa Fluor 594 C5 maleimide (Thermo Fisher Scientific, Inc.) or maleimide-functionalized CF660R (Biotium, Inc.) dissolved in anhydrous DMF was added in 3 times molar excess. The reaction was permitted to proceed as described above and quenched by adding DTT at a final concentration of 10 mM. Donor–acceptor labeled protein was purified by RP-HPLC (see above), lyophilized, and redissolved in 20 mM KH<sub>2</sub>PO<sub>4</sub>/K<sub>2</sub>HPO<sub>4</sub>, 6 M GdmCl, pH 7.3. Protein identity and site-specific labeling were confirmed by electrospray ionization mass spectrometry (ESI-MS). The reactivity of the two cysteine residues for the fluorophores is not identical, and some separation of labeling permutations was achieved during the purification of some IDRs, but in most cases, we used a mixture of permutants. In the case of skh, where the dye permutants could be separated, the difference in their transfer efficiency was 0.02,

indicating a minor effect on the results. Fluorescently labeled samples were stored at  $-80\text{ }^{\circ}\text{C}$  until further use.

A synthetic codon-optimized N-terminal 334-amino acid segment of pertactin<sup>77</sup> (PNt) with Cys residues in positions 29 and 117 was cloned into a pJ414 vector (ATUM) and transformed into *E. coli* BL21-DE3 cells (Agilent). Cells were grown at  $37\text{ }^{\circ}\text{C}$  in Luria–Bertani medium containing  $100\text{ }\mu\text{g mL}^{-1}$  carbenicillin and induced for expression at an  $\text{OD}_{600}$  of 0.7 for 3 h. The cell pellet derived from 0.5 L of culture was suspended in 70 mL of buffer A [ $50\text{ mM}$  Tris-HCl, pH 8,  $100\text{ mM}$  NaCl,  $1\text{ mM}$  ethylenediaminetetraacetic acid (EDTA),  $5\text{ mM}$  2-mercaptoethanol, and  $5\text{ mM}$  benzamidine], followed by the addition of lysozyme ( $100\text{ }\mu\text{g mL}^{-1}$ ) and sonicated at  $4\text{ }^{\circ}\text{C}$ . The insoluble recombinant protein was washed by resuspension in 70 mL of buffer A containing 1% Triton X-100 and subsequently in buffer A in the absence of Triton X-100. In all cases, the insoluble fraction was pelleted by centrifugation at  $20,000g$  for 30 min at  $4\text{ }^{\circ}\text{C}$ . The final pellet was solubilized in  $8\text{ M}$  urea,  $50\text{ mM}$  Tris-HCl, pH 8.0,  $5\text{ mM}$  EDTA, and  $5\text{ mM}$  tris(2-carboxyethyl)phosphine (TCEP) and 1/4th of the protein ( $\sim 15\text{ mg}$ ) was applied onto a Superdex-200 column ( $1.6 \times 60\text{ cm}$ , Cytiva) equilibrated in  $50\text{ mM}$  Tris-HCl, pH 8,  $4\text{ M}$  GdmCl,  $1\text{ mM}$  EDTA, and  $1\text{ mM}$  TCEP at a flow rate of  $1.4\text{ mL min}^{-1}$  at ambient temperature. Peak fractions with the highest purity were verified by mass spectrometry and used for the experiments. PNtCC was labeled with dyes as described for the other IDRs.

### Circular Dichroism Spectroscopy

Unlabeled constructs were dialyzed against  $20\text{ mM}$   $\text{KH}_2\text{PO}_4/\text{K}_2\text{HPO}_4$ ,  $1\text{ mM}$  DTT, pH 7.3, using Slide-A-Lyzer MINI Dialysis Devices, 3.5K MWCO (Thermo Fisher Scientific). Insoluble components were removed by centrifugation. Circular dichroism spectra from 190 to 250 nm were acquired on a spectropolarimeter (J-810, Jasco, or ChiraScan V100, Applied Photophysics) at  $22\text{ }^{\circ}\text{C}$  in quartz cells with a path length of 0.5 or 1 mm at concentrations of 0.1–0.5 mg/mL. Absorption data of those scans were used to determine the concentration of the peptides using their absorption at 214 nm.<sup>87</sup>

### Single-Molecule Spectroscopy

For single-molecule experiments, the donor–acceptor labeled IDRs were diluted to approximately  $100\text{ pM}$  in  $20\text{ mM}$   $\text{KH}_2\text{PO}_4/\text{K}_2\text{HPO}_4$ ,  $125\text{ mM}$  KCl, pH 7.3 with 0.001% Tween 20, and  $10\text{ mM}$  DTT for the Cy3B/CF660R-labeled or  $147\text{ mM}$  2-mercaptoethanol for the Alexa-labeled IDRs. The measurements were conducted at  $22\text{ }^{\circ}\text{C}$  using chambered cover slides ( $\mu$ -Slide, ibidi). Different light sources were used for excitation, depending on the fluorophores used. For Alexa 488 excitation, an LDH-D-C-485 diode laser (PicoQuant GmbH) was employed. Alexa 594 and Cy3B excitation was achieved using a supercontinuum fiber laser (SC-450-4, Fianium Ltd.) filtered by a z582/15 or HC543.5/2 band-pass, respectively (Chroma Technology). CF660R was excited with an LDH-D-C-640 diode laser (PicoQuant GmbH). Lasers were operated at a pulse repetition rate of  $20\text{ MHz}$  to achieve pulsed interleaved excitation of donor and acceptor.<sup>88</sup> Fluorescence photons were collected with a UplanApo 60x/1.20W objective (Olympus) and passed through a suitable multiband mirror and a  $100\text{-}\mu\text{m}$  confocal pinhole. Subsequently, photons were separated according to polarization by using a polarizing beam splitter and wavelength via suitable dichroic mirrors. Finally, photons were filtered by optical band-pass filters and detected by avalanche photodiodes. Photon arrival times were recorded with a HydraHarp 400 time-correlated single-photon counting system (PicoQuant) at a time resolution of 16 ps.

Photon bursts emitted by labeled IDRs diffusing through the confocal volume were identified as contiguous intervals of emission with interphoton times below  $150\text{ }\mu\text{s}$ .<sup>89</sup> FRET efficiency histograms shown in Figure 2A are based on a threshold of 50 photons per burst. Dual-channel-burst-search<sup>90</sup> was applied to avoid artifacts from bleaching and blinking. Subsequently, bursts were corrected for differences in chromophore quantum yields, differences in detection efficiency of the detectors and spectral crosstalk obtained from measurements of free dye solutions, and direct acceptor excitation and

background signal.<sup>91</sup> The stoichiometry ratio<sup>88,92</sup> of a photon burst was calculated according to

$$S = \frac{n_{\text{tot,Dex}}}{n_{\text{tot,Dex}} + n_{\text{tot,Aex}}}$$

where  $n_{\text{tot,Dex}}$  and  $n_{\text{tot,Aex}}$  denote the corrected total number of photons emitted after donor or acceptor excitation, respectively. Bursts with  $0.2 < S < 0.8$  were used to calculate the transfer efficiency

$$E = \frac{n_A}{n_A + n_D}$$

where  $n_D$  and  $n_A$  are the corrected donor and acceptor photon counts emitted upon donor excitation within a burst, respectively. Alternatively, the correction factors were inferred from the measurement of the IDRs with alternating excitation.<sup>39,92</sup> Figure 2 and Table S2 show the average of the mean transfer efficiencies of at least three independent measurements. The average standard deviations are  $\pm 0.014$  or  $\pm 0.024$  for the measurements using either the Alexa dye pair or Cy3B/CF660R, respectively, most of which were taken over the course of multiple years and on different instruments. Fluorescence polarization anisotropies were  $< 0.1$  for all samples and fluorophores, indicating that the orientational factor  $\kappa^2$  in Förster theory can be approximated by  $2/3$  due to rapid orientational averaging of donor and acceptor.<sup>53</sup> Data analysis was performed using the Mathematica (Wolfram Research) package Fretica (<https://github.com/SchulerLab>).

### NMR Spectroscopy

All data were acquired on Bruker Avance 600 MHz spectrometers equipped with TCI triple-resonance cryogenic probes and pulsed-field gradient units. All spectra were referenced directly by using DSS for the  $^1\text{H}$  dimension;  $^{13}\text{C}$  and  $^{15}\text{N}$  frequencies were referenced indirectly. Samples were dissolved in a buffer identical to those used for smFRET measurements ( $20\text{ mM}$   $\text{KH}_2\text{PO}_4/\text{K}_2\text{HPO}_4$ ,  $0.125\text{ M}$  KCl,  $10\text{ mM}$  DTT, pH 7.3). For backbone assignment, the  $^{15}\text{N}$ ,  $^{13}\text{C}$  isotopically labeled peptides were prepared to an approximate concentration of  $0.5\text{ mM}$ . Standard 3D assignment experiments based on sensitivity-enhanced  $^1\text{H}$ ,  $^{15}\text{N}$  HSQC (8 scans,  $1024 \times 256$  complex data points) were collected. These included an HNCACB and CBCA(CO)NH (8 scans,  $1024 (^1\text{H}) \times 32 (^{15}\text{N}) \times 128 (^{13}\text{C})$  complex data points, with 11, 24, and 70 ppm as  $^1\text{H}$ ,  $^{15}\text{N}$  and  $^{13}\text{C}$  sweep width, respectively), an HN(CA)CO (8 scans,  $1024 (^1\text{H}) \times 32 (^{15}\text{N}) \times 75 (^{13}\text{C})$  complex data points, with 11, 24, and 18 ppm as  $^1\text{H}$ ,  $^{15}\text{N}$ , and  $^{13}\text{C}$  sweep widths), an HNCO (16 scans,  $1024 (^1\text{H}) \times 32 (^{15}\text{N}) \times 75 (^{13}\text{C})$  complex data points, with 11, 32, and 22 ppm as  $^1\text{H}$ ,  $^{15}\text{N}$ , and  $^{13}\text{C}$  sweep widths), and a HNCA (16 scans,  $1024 (^1\text{H}) \times 32 (^{15}\text{N}) \times 95 (^{13}\text{C})$  complex data points, with 16, 25, and 30 ppm as  $^1\text{H}$ ,  $^{15}\text{N}$ , and  $^{13}\text{C}$  sweep widths). Additionally, HN(CA)NNH (16 scans,  $1024 (^1\text{H}) \times 32 (^{15}\text{N} \text{ F1}) \times 60 (^{15}\text{N} \text{ F2})$  complex data points, with 11, 24, and 24 ppm as  $^1\text{H}$ ,  $^{15}\text{N} \text{ F1}$ , and  $^{15}\text{N} \text{ F2}$  sweep widths) spectra provided connectivity between  $i$  and  $i \pm 1$  amide nitrogen nuclei. Data were processed using BRUKER Topspin version 3.4, NMRPipe<sup>93</sup> (v.7.9) and analyzed using NMRfam SPARKY.<sup>94</sup> HSQC spectra were acquired for the unlabeled and Cy3B-, CF660R-, Alexa 488-, and Alexa 594-labeled sNh+ ( $16\text{ scans}$ ,  $1024 \times 256$  complex data points) at an approximate concentration of  $0.1\text{ mM}$ . Assignments were transferred to labeled sNh+ based on the assignments of the unlabeled protein. Broadening of the HSQC resonances was quantified using the ratio of the peak height of labeled protein to that of unlabeled. Chemical shift perturbation values were calculated as

$$\text{CSP} = \sqrt{\Delta\delta_{\text{H}}^2 + 0.15\Delta\delta_{\text{N}}^2}$$

### Small-Angle X-ray Scattering

SAXS experiments were performed on the BioCAT (18-ID-D) beamline at the Advanced Photon Source at Argonne National Laboratory.<sup>95</sup> Linker IDR samples were measured by coupling size-exclusion chromatography to a coflow X-ray sample chamber.<sup>96</sup> In short, a 5/150 Superdex 75 increase column (Cytiva) was equilibrated

in a buffer containing 20 mM  $\text{KH}_2\text{PO}_4/\text{K}_2\text{HPO}_4$ , 0.125 M KCl, pH 7.3, and 10 mM DTT. Elution of protein from the column was monitored by UV absorbance at 220 nm and integrated X-ray scattering intensity. Data reduction was performed at the beamline using the BioXTAS RAW software package.<sup>97</sup> Subsequent analysis and averaging of SEC-SAXS data was performed using custom Matlab routines<sup>98</sup> (Mathworks).

### Atomistic Simulations and Reweighting

Atomistic simulations of each of the IDRs were performed utilizing a homegrown adaptation of version 3 of the CAMPARI Monte Carlo simulation package (<http://campari.sourceforge.net>) and ABSINTH implicit solvation model and force field paradigm.<sup>41,78</sup> For each sequence, five independent simulations were performed. The simulations use spherical droplets with radii of 150 Å. Simulations utilize a modified `abs3.2_opls.prm` parameter with explicit representations of ions,<sup>99</sup> and the radii of sodium ions were set to 1.81 Å to avoid broken ergodicity due to ion chelation effects, especially around acidic groups. Neutralizing and excess  $\text{Na}^+$  and  $\text{Cl}^-$  ions were modeled explicitly, with an excess NaCl concentration of 20 mM. Simulations were performed at 340 K with  $6.15 \times 10^7$  steps, of which the first  $1 \times 10^7$  steps were taken as equilibration. The move set included translational, side chain rotation, concerted rotation, pivot, and proline puckering moves.<sup>78</sup>

For each replica, 1030 frames were saved and subjected to the addition of dyes using our in-house program COCOFRET. Briefly, for each frame 50 trials were attempted to attach Alexa 488 on the first Cys and attempts were discarded if the dye leads to steric clashes with the IDR. Additionally, 50 separate trials were attempted to attach Alexa 594 to the second Cys and attempts were discarded if steric clashes with the IDR exist. Attachment of dyes was performed by randomly selecting a rotamer from the HandyFRET rotamer library (<http://karri.anu.edu.au/handy/rl.html>) and making sure the  $\gamma$ -sulfur angles and bond lengths were ideal. Clashes were defined as any atoms within 5 Å of each other. Then, if at least 20 Alexa 488 and 20 Alexa 594 dyes were attached successfully, then all Alexa 488 and Alexa 594 dyes were attempted to be combined for the given frame conformation. If the dyes did not lead to steric clashes, then the distance between the dyes was saved. Transfer efficiencies per distance were determined using the Förster formula with  $R_0 = 6$  nm. For each frame, the mean transfer efficiency was calculated and used for the reweighting procedure.

The maximum entropy method COPER was utilized to reweight simulation ensembles to match experimental mean transfer efficiencies.<sup>100</sup> Briefly, the experimental mean transfer efficiencies as well as their associated errors listed in Table S2 were used as inputs to generate weights per frame that yield a global solution satisfying the inputs. The generated weights were then used to extract quantify conformational properties from the simulated ensembles.

### Analyses of ABSINTH Simulations

All analyses were performed using the Python-based simulation analysis package SOURSOP.<sup>101</sup> The weights extracted from COPER were used as inputs for the various analysis routines performed. Internal scaling profiles were calculated using the `get_internal_scaling_RMS()` analysis routine. The `get_scaling_exponent()` analysis routine was used to extract the best estimates of  $A_0$ , the prefactor which reports on the chain persistence length, and scaling exponent,  $\nu$ , from the standard homopolymer relationship,  $\langle R_{ij}^2 \rangle = \langle \langle r_{ij}^2 \rangle^{1/2} \rangle = A_0 |i - j|^\nu$  for each simulated ensemble. The  $A_0$  and  $\nu$  values extracted were then used as inputs into the analysis routine `get_polymer_scaled_distance_map()`. This routine determines how all residue distances compare to the best-fit standard homopolymer scaling behavior. The mode “scaled” was used which divides each weighted distance by the best-fit homopolymer model distance. Contact information was extracted using the analysis routine `get_contact_map()` with the mode “closest-heavy” and a contact distance threshold of 10 Å. The radius of gyration, asphericity, and secondary structure information per frame were calculated using the `get_radius_of_gyration()`, `get_asphericity()`, and `get_secondary_structure_DSSP()` analysis routines, respec-

tively. Contact networks were generated using the Python package NetworkX. When contacts (nodes in Figure 3E and Figure S10) and distance averages (Figure S10) were extracted per residue ( $i$ ), averages were taken only over residues greater than two residues away in linear sequence space, i.e.,  $j > i + 2$  and  $j < i - 2$ .

### Hydrophobicity Scale (HPS) Model Optimization and Simulations

We used the hydrophobicity scale model representation of disordered proteins,<sup>63</sup> in which each residue is represented by a single bead with size based on average residue volumes in crystal structures, linked by harmonic bonds with equilibrium length 0.38 nm and spring constant  $481.4 \text{ kJ nm}^{-2} \text{ mol}^{-1}$ . Interactions of each bead are determined by a scalar parameter  $\lambda$  characterizing “stickiness” with other beads. The value  $\lambda$  was based on hydrophobicity scales in the original model but should not be literally interpreted as hydrophobicity. Pairwise interactions between the beads are described by a modified Weeks–Chandler–Anderson potential in which the attractive part is determined by the arithmetic mean of the  $\lambda$ -values of the two beads. Further details are as described by Dannenhoffer-Lafage and Best.<sup>38</sup> The dyes were represented as shown in Figure S5, and force field parameters are given in Table S3. The dyes are linked by harmonic bonds with an equilibrium length of 0.38 nm and a spring constant of  $481.4 \text{ kJ nm}^{-2} \text{ mol}^{-1}$ . The shapes of the dyes are maintained by harmonic angle potential spring constants of  $48.14 \text{ kJ rad}^{-2} \text{ mol}^{-1}$ . The equilibrium angles for branch points were  $\pi/2$  and  $\pi$  elsewhere. Note that harmonic potentials for bond angles were only applied to the colored beads in Figure S5 and not to the dye linker beads. The mass of each dye bead was set to 100 atomic mass units.

Langevin dynamics was propagated at a temperature of 300 K with a time step of 10 fs and a friction coefficient of  $1.0 \text{ ps}^{-1}$  using the LAMMPS package, with typical run lengths of 600 ns for each protein, discarding the first 100 ns for equilibration. Run input files and final parameters are available on Zenodo at [10.5281/zenodo.11397637](https://doi.org/10.5281/zenodo.11397637). Force balance optimization was performed as described by Dannenhoffer-Lafage and Best,<sup>38</sup> optimizing the FRET efficiency and using a learning rate of 5.0. The initial parameters for the protein were obtained from the M1 parameter set of Tesei et al.,<sup>40</sup> while those for the dyes were obtained from a grid search over dye and linker  $\lambda$  values. Furthermore, L2 starting point regularization was employed with initial parameters used as the starting point and a regularization strength of 0.0001. Simulations were performed at an ionic strength of 185 mM, with CF660R or Alexa 594 attached to Cys5 and Cy3B or Alexa 488 attached to Cys64. Permuting the dyes led to an average difference in  $\langle E \rangle$  of 0.018, which was thus not taken into account in our comparison. Benchmark simulations of the CALVADOS data set<sup>40</sup> were performed with the CALVADOS M1 parameters and with our optimized “Linker-HPS” parameters (Figure S11).

Simulations of Pnt, ibb, n49, nul, and nus were propagated at 298 K with a time step of 10 fs, a friction coefficient of  $1.0 \text{ ps}^{-1}$ , and a Debye length corresponding to an ionic strength of 190 mM for Pnt and 165 mM for ibb, n49, nul, and nus. Pnt was simulated for 60  $\mu\text{s}$ , and the first 10  $\mu\text{s}$  was discarded for equilibration. The simulations of ibb, n49, nul, and nus were propagated for 10  $\mu\text{s}$ , and the first microsecond was discarded for equilibration.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/jacsau.4c00673>.

Selection scheme, disorder scores, and circular dichroism spectra for the 16 IDR sequences, single-molecule analysis of skh labeled with Cy3B/CF660R, coarse-grained dye structures, SAXS and NMR data, correlations between different sequence parameters and average transfer efficiency, and additional ABSINTH and HPS model analyses (Figures S1–S13); summary of UniProt11 identifiers and important physicochemical

parameters of the IDRs, original (CALVADOS 112 M3), optimized, and CALVADOS 213 short-range interaction parameters ( $\lambda$ ) (Tables S1–S3) (PDF)

## AUTHOR INFORMATION

### Corresponding Authors

**Rohit V. Pappu** – Department of Biomedical Engineering and Center for Biomolecular Condensates, Washington University in St. Louis, St. Louis, Missouri 63130, United States; [orcid.org/0000-0003-2568-1378](https://orcid.org/0000-0003-2568-1378); Email: [pappu@wustl.edu](mailto:pappu@wustl.edu)

**Robert B. Best** – Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, United States; [orcid.org/0000-0002-7893-3543](https://orcid.org/0000-0002-7893-3543); Email: [robert.best2@nih.gov](mailto:robert.best2@nih.gov)

**Tanja Mittag** – Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, United States; [orcid.org/0000-0002-1827-3811](https://orcid.org/0000-0002-1827-3811); Email: [Tanja.Mittag@stjude.org](mailto:Tanja.Mittag@stjude.org)

**Benjamin Schuler** – Department of Biochemistry, University of Zurich, 8057 Zurich, Switzerland; Department of Physics, University of Zurich, 8057 Zurich, Switzerland; [orcid.org/0000-0002-5970-4251](https://orcid.org/0000-0002-5970-4251); Email: [schuler@bioc.uzh.ch](mailto:schuler@bioc.uzh.ch)

### Authors

**Andrea Holla** – Department of Biochemistry, University of Zurich, 8057 Zurich, Switzerland

**Erik W. Martin** – Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, United States

**Thomas Dannenhoffer-Lafage** – Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, United States

**Kiersten M. Ruff** – Department of Biomedical Engineering and Center for Biomolecular Condensates, Washington University in St. Louis, St. Louis, Missouri 63130, United States; [orcid.org/0000-0003-3240-1856](https://orcid.org/0000-0003-3240-1856)

**Sebastian L. B. König** – Department of Biochemistry, University of Zurich, 8057 Zurich, Switzerland; Present Address: Federal Food Safety and Veterinary Office, 3003 Bern, Switzerland

**Mark F. Nüesch** – Department of Biochemistry, University of Zurich, 8057 Zurich, Switzerland

**Aritra Chowdhury** – Department of Biochemistry, University of Zurich, 8057 Zurich, Switzerland

**John M. Louis** – Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, United States; [orcid.org/0000-0002-0052-1899](https://orcid.org/0000-0002-0052-1899)

**Andrea Soranno** – Department of Biochemistry, University of Zurich, 8057 Zurich, Switzerland; Department of Biochemistry and Molecular Biophysics, Center for Biomolecular Condensates, Washington University in St. Louis, St. Louis, Missouri 63130, United States; [orcid.org/0000-0001-8394-7993](https://orcid.org/0000-0001-8394-7993)

**Daniel Nettels** – Department of Biochemistry, University of Zurich, 8057 Zurich, Switzerland

Complete contact information is available at: <https://pubs.acs.org/10.1021/jacsau.4c00673>

## Author Contributions

CRediT: **Andrea Holla** conceptualization, formal analysis, investigation, visualization, writing - original draft, writing - review & editing; **Erik Martin** formal analysis, investigation, visualization, writing - review & editing; **Thomas Dannenhoffer-Lafage** formal analysis, investigation, methodology, visualization, writing - review & editing; **Kiersten M. Ruff** formal analysis, investigation, visualization, writing - review & editing; **Sebastian L. B. König** formal analysis, investigation; **Mark F. Nüesch** formal analysis, investigation; **Aritra Chowdhury** formal analysis, investigation, writing - review & editing; **John M. Louis** investigation, writing - review & editing; **Andrea Soranno** conceptualization, formal analysis, investigation, writing - review & editing; **Daniel Nettels** formal analysis, software, writing - review & editing; **Rohit V Pappu** conceptualization, funding acquisition, resources, supervision, writing - review & editing; **Robert B. Best** conceptualization, formal analysis, investigation, methodology, resources, supervision, writing - review & editing; **Tanja Mittag** conceptualization, funding acquisition, resources, supervision, writing - review & editing; **Benjamin Schuler** conceptualization, formal analysis, funding acquisition, resources, supervision, writing - original draft, writing - review & editing.

## Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

We thank Lucia Franchini, Daniela Oswald, Gessica Priola, Moa Hasler, Vincent Maximilian Münch, Steffen Winkler, and Dina Heiligensetzer for their assistance in sample preparation, single-molecule measurements, and circular dichroism spectroscopy measurements, and Soundhar Gopi and Miloš Ivanović for helpful discussion. We thank the Functional Genomics Center Zurich for mass spectrometry analysis. This work was supported by the Swiss National Science Foundation (grant 310030\_197776, B.S.), the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement ID no. 898228 (A.C.), the US National Institutes of Health (grant R01NS121114, T.M. and R.V.P.), the US National Science Foundation (grant MCB-2227268, R.V.P.), the American Lebanese Syrian Associated Charities (to T.M.), and by the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (to T.D.-L, J.M.L., and R.B.B.). This research used resources of the Advanced Photon Source, which is a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. BioCAT was supported by grant P30 GM138395 from the National Institute of General Medical Sciences of the National Institutes of Health. In addition, the computational resources of the NIH HPC Biowulf cluster were utilized (<http://hpc.nih.gov>).

## REFERENCES

- (1) Holehouse, A. S.; Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat. Rev. Mol. Cell Biol.* **2024**, *25*, 187–211.
- (2) Tsang, B.; Pritisanac, I.; Scherer, S. W.; Moses, A. M.; Forman-Kay, J. D. Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations. *Cell* **2020**, *183*, 1742–1756.

- (3) van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; Kim, P. M.; Kriwacki, R. W.; Oldfield, C. J.; Pappu, R. V.; Tompa, P.; Uversky, V. N.; Wright, P. E.; Babu, M. M. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **2014**, *114*, 6589–6631.
- (4) Banani, S. F.; Lee, H. O.; Hyman, A. A.; Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **2017**, *18*, 285–298.
- (5) Uversky, V. N. What does it mean to be natively unfolded? *Eur. J. Biochem.* **2002**, *269*, 2–12.
- (6) Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V. Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 16764–16769.
- (7) Müller-Späh, S.; Soranno, A.; Hirschfeld, V.; Hofmann, H.; Rüegger, S.; Reymond, L.; Nettels, D.; Schuler, B. Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 14609–14614.
- (8) Mao, A. H.; Crick, S. L.; Vitalis, A.; Chicoine, C. L.; Pappu, R. V. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 8183–8188.
- (9) Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 16155–16160.
- (10) Das, R. K.; Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 13392.
- (11) Sawle, L.; Ghosh, K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J. Chem. Phys.* **2015**, *143*, No. 085101.
- (12) Fuetes, G.; Banterle, N.; Ruff, K. M.; Chowdhury, A.; Mercadante, D.; Koehler, C.; Kachala, M.; Estrada Girona, G.; Milles, S.; Mishra, A.; Onck, P. R.; Grater, F.; Esteban-Martin, S.; Pappu, R. V.; Svergun, D. I.; Lemke, E. A. Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E6342–E6351.
- (13) Best, R. B.; Zheng, W.; Borgia, A.; Buholzer, K.; Borgia, M. B.; Hofmann, H.; Soranno, A.; Nettels, D.; Gast, K.; Grishaev, A.; Schuler, B. Comment on "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **2018**, *361* (6405), No. eaar7101.
- (14) Holehouse, A. S.; Pappu, R. V. Collapse Transitions of Proteins and the Interplay Among Backbone, Sidechain, and Solvent Interactions. *Annu. Rev. Biophys.* **2018**, *47*, 19–39.
- (15) Shea, J. E.; Best, R. B.; Mittal, J. Physics-based computational and theoretical approaches to intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2021**, *67*, 219–225.
- (16) Tesei, G.; Trolle, A. I.; Jonsson, N.; Betz, J.; Knudsen, F. E.; Pesce, F.; Johansson, K. E.; Lindorff-Larsen, K. Conformational ensembles of the human intrinsically disordered proteome. *Nature* **2024**, *626*, 897–904.
- (17) O'Brien, E. P.; Morrison, G.; Brooks, B. R.; Thirumalai, D. How accurate are polymer models in the analysis of Förster resonance energy transfer experiments on proteins? *J. Chem. Phys.* **2009**, *130* (12), No. 124903.
- (18) Schuler, B.; Soranno, A.; Hofmann, H.; Nettels, D. Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins. *Annu. Rev. Biophys.* **2016**, *45*, 207–231.
- (19) Mao, A. H.; Lyle, N.; Pappu, R. V. Describing sequence-ensemble relationships for intrinsically disordered proteins. *Biochem. J.* **2013**, *449*, 307–318.
- (20) Teufel, D. P.; Johnson, C. M.; Lum, J. K.; Neuweiler, H. Backbone-Driven Collapse in Unfolded Protein Chains. *J. Mol. Biol.* **2011**, *409*, 250–262.
- (21) Borgia, A.; Zheng, W.; Buholzer, K.; Borgia, M. B.; Schuler, A.; Hofmann, H.; Soranno, A.; Nettels, D.; Gast, K.; Grishaev, A.; Best, R. B.; Schuler, B. Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *J. Am. Chem. Soc.* **2016**, *138*, 11714–11726.
- (22) Rubinstein, M.; Colby, R. H. *Polymer Physics*; Oxford University Press: Oxford, New York, 2003; p xi, 440 p.
- (23) Yu, M.; Heidari, M.; Mikhaleva, S.; Tan, P. S.; Mingu, S.; Ruan, H.; Reinkemeier, C. D.; Obarska-Kosinska, A.; Siggel, M.; Beck, M.; Hummer, G.; Lemke, E. A. Visualizing the disordered nuclear transport machinery in situ. *Nature* **2023**, *617*, 162–169.
- (24) Riback, J. A.; Bowman, M. A.; Zmyslowski, A. M.; Knoverek, C. R.; Jumper, J. M.; Hinshaw, J. R.; Kaye, E. B.; Freed, K. F.; Clark, P. L.; Sosnick, T. R. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **2017**, *358*, 238–241.
- (25) Lotthammer, J. M.; Ginell, G. M.; Griffith, D.; Emenecker, R. J.; Holehouse, A. S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat. Methods* **2024**, *21*, 465–476.
- (26) Zheng, W.; Zerze, G. H.; Borgia, A.; Mittal, J.; Schuler, B.; Best, R. B. Inferring properties of disordered chains from FRET transfer efficiencies. *J. Chem. Phys.* **2018**, *148*, No. 123329.
- (27) Valle, F.; Favre, M.; De Los Rios, P.; Rosa, A.; Dietler, G. Scaling Exponents and Probability Distributions of DNA End-to-End Distance. *Phys. Rev. Lett.* **2005**, *95*, No. 158105.
- (28) Steinhauser, M. O. A molecular dynamics study on universal properties of polymer chains in different solvent qualities. Part I. A review of linear chain properties. *J. Chem. Phys.* **2005**, *122*, No. 094901.
- (29) Sanchez, I. C. Phase-Transition Behavior of the Isolated Polymer-Chain. *Macromolecules* **1979**, *12*, 980–988.
- (30) Baul, U.; Chakraborty, D.; Mugnai, M. L.; Straub, J. E.; Thirumalai, D. Sequence Effects on Size, Shape, and Structural Heterogeneity in Intrinsically Disordered Proteins. *J. Phys. Chem. B* **2019**, *123*, 3462–3474.
- (31) Marsh, J. A.; Forman-Kay, J. D. Sequence Determinants of Compaction in Intrinsically Disordered Proteins. *Biophys. J.* **2010**, *98*, 2383–2390.
- (32) Das, R. K.; Ruff, K. M.; Pappu, R. V. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2015**, *32*, 102–112.
- (33) Martin, E. W.; Holehouse, A. S.; Grace, C. R.; Hughes, A.; Pappu, R. V.; Mittag, T. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J. Am. Chem. Soc.* **2016**, *138*, 15323–15335.
- (34) Best, R. B. Emerging consensus on the collapse of unfolded and intrinsically disordered proteins in water. *Curr. Opin. Struct. Biol.* **2020**, *60*, 27–38.
- (35) Zheng, W.; Dignon, G.; Brown, M.; Kim, Y. C.; Mittal, J. Hydrophobic Patterning Complements Charge Patterning to Describe Conformational Preferences of Disordered Proteins. *J. Phys. Chem. Lett.* **2020**, *11*, 3408–3415.
- (36) Vuzman, D.; Levy, Y. DNA search efficiency is modulated by charge composition and distribution in the intrinsically disordered tail. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 21004–21009.
- (37) Borgia, A.; Borgia, M. B.; Bugge, K.; Kissling, V. M.; Heidarsson, P. O.; Fernandes, C. B.; Sottini, A.; Soranno, A.; Buholzer, K. J.; Nettels, D.; Kragelund, B. B.; Best, R. B.; Schuler, B. Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **2018**, *555*, 61–66.
- (38) Dannenhoffer-Lafage, T.; Best, R. B. A Data-Driven Hydrophobicity Scale for Predicting Liquid-Liquid Phase Separation of Proteins. *J. Phys. Chem. B* **2021**, *125*, 4046–4056.

- (39) Holmstrom, E. D.; Holla, A.; Zheng, W.; Nettels, D.; Best, R. B.; Schuler, B. Accurate Transfer Efficiencies, Distance Distributions, and Ensembles of Unfolded and Intrinsically Disordered Proteins From Single-Molecule FRET. In *Methods in Enzymology*; Elsevier, 2018; Vol. 611, pp 287–325.
- (40) Tesei, G.; Schulze, T. K.; Crehuet, R.; Lindorff-Larsen, K. Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118* (44), No. e2111696118.
- (41) Vitalis, A.; Pappu, R. V. ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* **2009**, *30*, 673–699.
- (42) Liao, J. Y.; Yang, B.; Zhang, Y. C.; Wang, X. J.; Ye, Y.; Peng, J. W.; Yang, Z. Z.; He, J. H.; Zhang, Y.; Hu, K.; Lin, D. C.; Yin, D. EuRBPDB: a comprehensive resource for annotation, functional and oncological investigation of eukaryotic RNA binding proteins (RBPs). *Nucleic Acids Res.* **2020**, *48*, D307–D313.
- (43) Varadi, M.; Zsolymoi, F.; Guharoy, M.; Tompa, P. Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. *PLoS One* **2015**, *10* (10), No. e0139731.
- (44) González-Foutel, N. S.; Glavina, J.; Borchers, W. M.; Safranchik, M.; Barrera-Vilarmau, S.; Sagar, A.; Estaña, A.; Barozet, A.; Garrone, N. A.; Fernandez-Ballester, G.; Blanes-Mira, C.; Sánchez, I. E.; de Prat-Gay, G.; Cortés, J.; Bernadó, P.; Pappu, R. V.; Holehouse, A. S.; Daughdrill, G. W.; Chemes, L. B. Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nat. Struct. Mol. Biol.* **2022**, *29*, 781–790.
- (45) UniProt-Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2022**, *51* (D1), D523–D531.
- (46) Emenecker, R. J.; Griffith, D.; Holehouse, A. S. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys. J.* **2021**, *120*, 4312–4319.
- (47) Uversky, V. N.; Gillespie, J. R.; Fink, A. L. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* **2000**, *41*, 415–427.
- (48) Hillger, F.; Nettels, D.; Dorsch, S.; Schuler, B. Detection and analysis of protein aggregation with confocal single molecule fluorescence spectroscopy. *J. Fluoresc.* **2007**, *17*, 759–765.
- (49) Kudryavtsev, V.; Sikor, M.; Kalinin, S.; Mokranjac, D.; Seidel, C. A. M.; Lamb, D. C. Combining MFD and PIE for Accurate Single-Pair Forster Resonance Energy Transfer Measurements. *ChemPhysChem* **2012**, *13*, 1060–1078.
- (50) Klose, D.; Holla, A.; Gmeiner, C.; Nettels, D.; Ritsch, I.; Bross, N.; Yulikov, M.; Allain, F. H.; Schuler, B.; Jeschke, G. Resolving distance variations by single-molecule FRET and EPR spectroscopy using rotamer libraries. *Biophys. J.* **2021**, *120*, 4842–4858.
- (51) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132.
- (52) Ruff, K. M. Predicting Conformational Properties of Intrinsically Disordered Proteins from Sequence. In *Intrinsically Disordered Proteins*, Methods in Molecular Biology; Springer US, 2020; Vol. 2141, pp 347–389.
- (53) Hellenkamp, B.; Schmid, S.; Doroshenko, O.; Opanasyuk, O.; Kuhnemuth, R.; Rezaei Adariani, S.; Ambrose, B.; Aznauryan, M.; Barth, A.; Birkedal, V.; Bowen, M. E.; Chen, H.; Cordes, T.; Eilert, T.; Fijen, C.; Gebhardt, C.; Gotz, M.; Gouridis, G.; Gratton, E.; Ha, T.; Hao, P.; Hanke, C. A.; Hartmann, A.; Hendrix, J.; Hildebrandt, L. L.; Hirschfeld, V.; Hohlbein, J.; Hua, B.; Hubner, C. G.; Kallis, E.; Kapanidis, A. N.; Kim, J. Y.; Krainer, G.; Lamb, D. C.; Lee, N. K.; Lemke, E. A.; Levesque, B.; Levitus, M.; McCann, J. J.; Naredi-Rainer, N.; Nettels, D.; Ngo, T.; Qiu, R.; Robb, N. C.; Rocker, C.; Sanabria, H.; Schlierf, M.; Schroder, T.; Schuler, B.; Seidel, H.; Streit, L.; Thurn, J.; Tinnefeld, P.; Tyagi, S.; Vandenberk, N.; Vera, A. M.; Weninger, K. R.; Wunsch, B.; Yanez-Orozco, I. S.; Michaelis, J.; Seidel, C. A. M.; Craggs, T. D.; Hugel, T. Precision and accuracy of single-molecule FRET measurements—a multi-laboratory benchmark study. *Nat. Methods* **2018**, *15*, 669–676.
- (54) Lin, L.-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **1989**, *45*, 255–268.
- (55) Holehouse, A. S.; Garai, K.; Lyle, N.; Vitalis, A.; Pappu, R. V. Quantitative Assessments of the Distinct Contributions of Polypeptide Backbone Amides versus Side Chain Groups to Chain Expansion via Chemical Denaturation. *J. Am. Chem. Soc.* **2015**, *137*, 2984–2995.
- (56) Mukhopadhyay, S.; Krishnan, R.; Lemke, E. A.; Lindquist, S.; Deniz, A. A. A natively unfolded yeast prion monomer adopts an ensemble of collapsed and rapidly fluctuating structures. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2649–2654.
- (57) Tran, H. T.; Mao, A.; Pappu, R. V. Role of Backbone–Solvent Interactions in Determining Conformational Equilibria of Intrinsically Disordered Proteins. *J. Am. Chem. Soc.* **2008**, *130*, 7380–7392.
- (58) Walters, R. H.; Murphy, R. M. Examining Polyglutamine Peptide Length: A Connection between Collapsed Conformations and Increased Aggregation. *J. Mol. Biol.* **2009**, *393*, 978–992.
- (59) Karandur, D.; Wong, K.-Y.; Pettitt, B. M. Solubility and Aggregation of Gly5 in Water. *J. Phys. Chem. B* **2014**, *118*, 9565–9572.
- (60) Karandur, D.; Harris, R. C.; Pettitt, B. M. Protein collapse driven against solvation free energy without H-bonds. *Protein Sci.* **2016**, *25*, 103–110.
- (61) Choi, J.-M.; Pappu, R. V. Improvements to the ABSINTH Force Field for Proteins Based on Experimentally Derived Amino Acid Specific Backbone Conformational Statistics. *J. Chem. Theory Comput.* **2019**, *15*, 1367–1382.
- (62) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (63) Dignon, G. L.; Zheng, W.; Kim, Y. C.; Best, R. B.; Mittal, J. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* **2018**, *14*, No. e1005941.
- (64) Norgaard, A. B.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Experimental Parameterization of an Energy Function for the Simulation of Unfolded Proteins. *Biophys. J.* **2008**, *94*, 182–192.
- (65) Wang, L. P.; Chen, J.; Van Voorhis, T. Systematic Parametrization of Polarizable Force Fields from Quantum Chemistry Data. *J. Chem. Theory Comput.* **2013**, *9*, 452–460.
- (66) Muschiello, A.; Andrecka, J.; Jawhari, A.; Brückner, F.; Cramer, P.; Michaelis, J. A nano-positioning system for macromolecular structural analysis. *Nat. Methods* **2008**, *5*, 965–971.
- (67) Sindbert, S.; Kalinin, S.; Nguyen, H.; Kienzler, A.; Clima, L.; Bannwarth, W.; Appel, B.; Müller, S.; Seidel, C. A. Accurate distance determination of nucleic acids via Forster resonance energy transfer: implications of dye linker length and rigidity. *J. Am. Chem. Soc.* **2011**, *133*, 2463–2480.
- (68) Klose, D.; Klare, J. P.; Grohmann, D.; Kay, C. W. M.; Werner, F.; Steinhoff, H. J. Simulation vs. Reality: A Comparison of In Silico Distance Predictions with DEER and FRET Measurements. *PLoS One* **2012**, *7* (6), No. e39492.
- (69) Warner, J. B.; Ruff, K. M.; Tang, P. S.; Lemke, E. A.; Pappu, R. V.; Lashuel, H. A. Monomeric Huntingtin Exon 1 Has Similar Overall Structural Features for Wild-Type and Pathological Polyglutamine Lengths. *J. Am. Chem. Soc.* **2017**, *139*, 14456–14469.
- (70) Grotz, K. K.; Nueesch, M. F.; Holmstrom, E. D.; Heinz, M.; Stelzl, L. S.; Schuler, B.; Hummer, G. Dispersion Correction Alleviates Dye Stacking of Single-Stranded DNA and RNA in Simulations of Single-Molecule Fluorescence Experiments. *J. Phys. Chem. B* **2018**, *122*, 11626–11639.
- (71) Montepietra, D.; Tesei, G.; Martins, J. M.; Kunze, M. B. A.; Best, R. B.; Lindorff-Larsen, K. FRETpredict: a Python package for FRET efficiency predictions using rotamer libraries. *Commun. Biol.* **2024**, *7*, No. 298.
- (72) Tesei, G.; Lindorff-Larsen, K. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Res. Eur.* **2022**, *2*, 94.
- (73) Pak, C. W.; Kosno, M.; Holehouse, A. S.; Padrick, S. B.; Mittal, A.; Ali, R.; Yunus, A. A.; Liu, D. R.; Pappu, R. V.; Rosen, M. K. Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein. *Mol. Cell* **2016**, *63*, 72–85.
- (74) Vernon, R. M.; Chong, P. A.; Tsang, B.; Kim, T. H.; Bah, A.; Farber, P.; Lin, H.; Forman-Kay, J. D. Pi-Pi contacts are an

- overlooked protein feature relevant to phase separation. *eLife* **2018**, *7*, No. e31486.
- (75) Martin, E. W.; Holehouse, A. S.; Peran, I.; Farag, M.; Incicco, J. J.; Bremer, A.; Grace, C. R.; Soranno, A.; Pappu, R. V.; Mittag, T. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **2020**, *367*, 694–699.
- (76) Bremer, A.; Farag, M.; Borchers, W. M.; Peran, I.; Martin, E. W.; Pappu, R. V.; Mittag, T. Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat. Chem.* **2022**, *14*, 196–207.
- (77) Riback, J. A.; Bowman, M. A.; Zmyslowski, A. M.; Plaxco, K. W.; Clark, P. L.; Sosnick, T. R. Commonly used FRET fluorophores promote collapse of an otherwise disordered protein. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 8889–8894.
- (78) Radhakrishnan, A.; Vitalis, A.; Mao, A. H.; Steffen, A. T.; Pappu, R. V. Improved Atomistic Monte Carlo Simulations Demonstrate That Poly-L-Proline Adopts Heterogeneous Ensembles of Conformations of Semi-Rigid Segments Interrupted by Kinks. *J. Phys. Chem. B* **2012**, *116*, 6862–6871.
- (79) Haran, G. How, when and why proteins collapse: the relation to folding. *Curr. Opin. Struct. Biol.* **2012**, *22*, 14–20.
- (80) Vancraenenbroeck, R.; Harel, Y. S.; Zheng, W.; Hofmann, H. Polymer effects modulate binding affinities in disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 19506–19512.
- (81) Sorensen, C. S.; Kjaergaard, M. Effective concentrations enforced by intrinsically disordered linkers are governed by polymer physics. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 23124–23131.
- (82) Masliah, G.; Barraud, P.; Allain, F. H. T. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cell. Mol. Life Sci.* **2013**, *70*, 1875–1895.
- (83) Maris, C.; Dominguez, C.; Allain, F. H. T. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J.* **2005**, *272*, 2118–2131.
- (84) Haenni, D.; Zosel, F.; Reymond, L.; Nettels, D.; Schuler, B. Intramolecular distances and dynamics from the combined photon statistics of single-molecule FRET and photoinduced electron transfer. *J. Phys. Chem. B* **2013**, *117*, 13015–13028.
- (85) Holehouse, A. S.; Das, R. K.; Ahad, J. N.; Richardson, M. O.; Pappu, R. V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys. J.* **2017**, *112*, 16–21.
- (86) Bao, W. J.; Gao, Y. G.; Chang, Y. G.; Zhang, T. Y.; Lin, X. J.; Yan, X. Z.; Hu, H. Y. Highly efficient expression and purification system of small-size protein domains in *Escherichia coli* for biochemical characterization. *Protein Expression Purif.* **2006**, *47*, 599–606.
- (87) Kuipers, B. J. H.; Gruppen, H. Prediction of molar extinction coefficients of proteins and peptides using UV absorption of the constituent amino acids at 214 nm to enable quantitative reverse phase high-performance liquid chromatography-mass spectrometry analysis. *J. Agric. Food Chem.* **2007**, *55*, 5445–5451.
- (88) Müller, B. K.; Zaychikov, E.; Bräuchle, C.; Lamb, D. C. Pulsed interleaved excitation. *Biophys. J.* **2005**, *89*, 3508–3522.
- (89) Eggeling, C.; Berger, S.; Brand, L.; Fries, J. R.; Schaffer, J.; Volkmer, A.; Seidel, C. A. M. Data registration and selective single-molecule analysis using multi-parameter fluorescence detection. *J. Biotechnol.* **2001**, *86*, 163–180.
- (90) Nir, E.; Michalet, X.; Hamadani, K. M.; Laurence, T. A.; Neuhauser, D.; Kovchegov, Y.; Weiss, S. Shot-Noise Limited Single-Molecule FRET Histograms: Comparison between Theory and Experiments. *J. Phys. Chem. B* **2006**, *110*, 22103–22124.
- (91) Schuler, B. Application of Single Molecule Förster Resonance Energy Transfer to Protein Folding. In *Protein Folding Protocols*, Methods in Molecular Biology; Humana Press, 2006; Vol. 350, pp 115–138.
- (92) Lee, N. K.; Kapanidis, A. N.; Wang, Y.; Michalet, X.; Mukhopadhyay, J.; Ebright, R. H.; Weiss, S. Accurate FRET Measurements within Single Diffusing Biomolecules Using Alternating-Laser Excitation. *Biophys. J.* **2005**, *88*, 2939–2953.
- (93) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **1995**, *6*, 277–293.
- (94) Lee, W.; Tonelli, M.; Markley, J. L. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **2015**, *31*, 1325–1327.
- (95) Fischetti, R.; Stepanov, S.; Rosenbaum, G.; Barrea, R.; Black, E.; Gore, D.; Heurich, R.; Kondrashkina, E.; Kropf, A. J.; Wang, S.; Zhang, K.; Irving, T. C.; Bunker, G. B. The BioCAT undulator beamline 18ID: a facility for biological non-crystalline diffraction and X-ray absorption spectroscopy at the Advanced Photon Source. *J. Synchrotron Radiat.* **2004**, *11*, 399–405.
- (96) Martin, E. W.; Hopkins, J. B.; Mittag, T. Chapter Seven - Small-angle X-ray scattering experiments of monodisperse intrinsically disordered protein samples close to the solubility limit. In *Methods in Enzymology*; Keating, C. D., Ed.; Academic Press, 2021; Vol. 646, pp 185–222.
- (97) Hopkins, J. B.; Gillilan, R. E.; Skou, S. BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. *J. Appl. Crystallogr.* **2017**, *50*, 1545–1553.
- (98) Martin, E. W.; Harmon, T. S.; Hopkins, J. B.; Chakravarthy, S.; Incicco, J. J.; Schuck, P.; Soranno, A.; Mittag, T. A multi-step nucleation process determines the kinetics of prion-like domain phase separation. *Nat. Commun.* **2021**, *12*, No. 4513.
- (99) Mao, A. H.; Pappu, R. V. Crystal lattice properties fully determine short-range interaction parameters for alkali and halide ions. *J. Chem. Phys.* **2012**, *137*, No. 064104.
- (100) Leung, H. T. A.; Bignucolo, O.; Aregger, R.; Dames, S. A.; Mazur, A.; Berneche, S.; Grzesiek, S. A Rigorous and Efficient Method To Reweight Very Large Conformational Ensembles Using Average Experimental Data and To Determine Their Relative Information Content. *J. Chem. Theory Comput.* **2016**, *12*, 383–394.
- (101) Lalmansingh, J. M.; Keeley, A. T.; Ruff, K. M.; Pappu, R. V.; Holehouse, A. S. SOURSOP: A Python Package for the Analysis of Simulations of Intrinsically Disordered Proteins. *J. Chem. Theory Comput.* **2023**, *19*, 5609–5620.

## Supporting Information

### Identifying sequence effects on chain dimensions of disordered proteins by integrating experiments and simulations

Andrea Holla<sup>1</sup>, Erik W. Martin<sup>2</sup>, Thomas Dannenhoffer-Lafage<sup>3</sup>, Kiersten M. Ruff<sup>4</sup>, Sebastian L. B. König<sup>1,5</sup>, Mark F. Nüesch<sup>1</sup>, Aritra Chowdhury<sup>1</sup>, John M. Louis<sup>3</sup>, Andrea Soranno<sup>1,6</sup>, Daniel Nettels<sup>1</sup>, Rohit V. Pappu<sup>4\*</sup>, Robert B. Best<sup>3\*</sup>, Tanja Mittag<sup>2\*</sup>, Benjamin Schuler<sup>1,7\*</sup>

<sup>1</sup>Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

<sup>2</sup>Department of Structural Biology, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, USA

<sup>3</sup>Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0520, USA

<sup>4</sup>Department of Biomedical Engineering and Center for Biomolecular Condensates, Washington University in St. Louis, St. Louis, MO 63130, USA

<sup>5</sup>Present Address: Federal Food Safety and Veterinary Office, 3003 Bern, Switzerland

<sup>6</sup>Department of Biochemistry and Molecular Biophysics, Center for Biomolecular Condensates, Washington University in St. Louis, St. Louis, USA

<sup>7</sup>Department of Physics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

\*Address correspondence to Rohit V. Pappu (pappu@wustl.edu), Robert B. Best (robert.best2@nih.gov), Tanja Mittag (Tanja.Mittag@stjude.org), or Benjamin Schuler (schuler@bioc.uzh.ch)



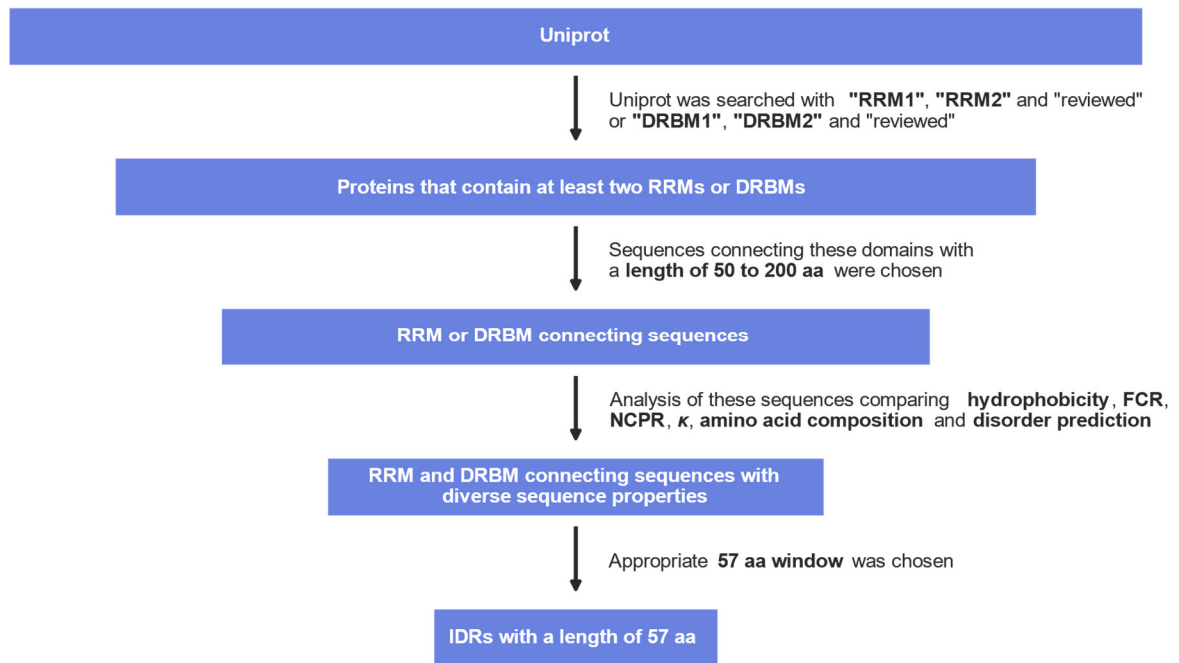
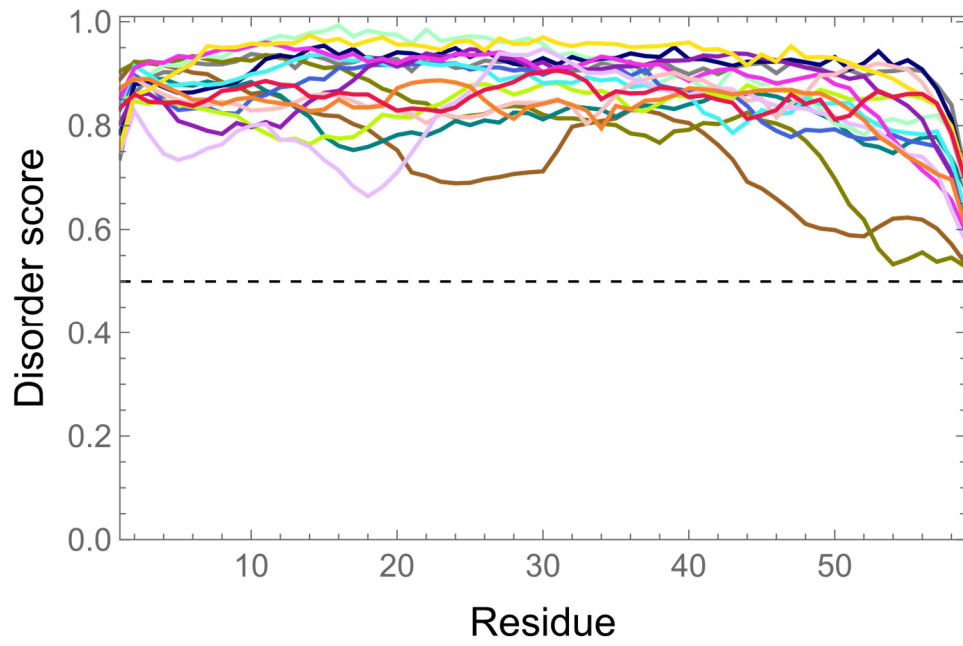
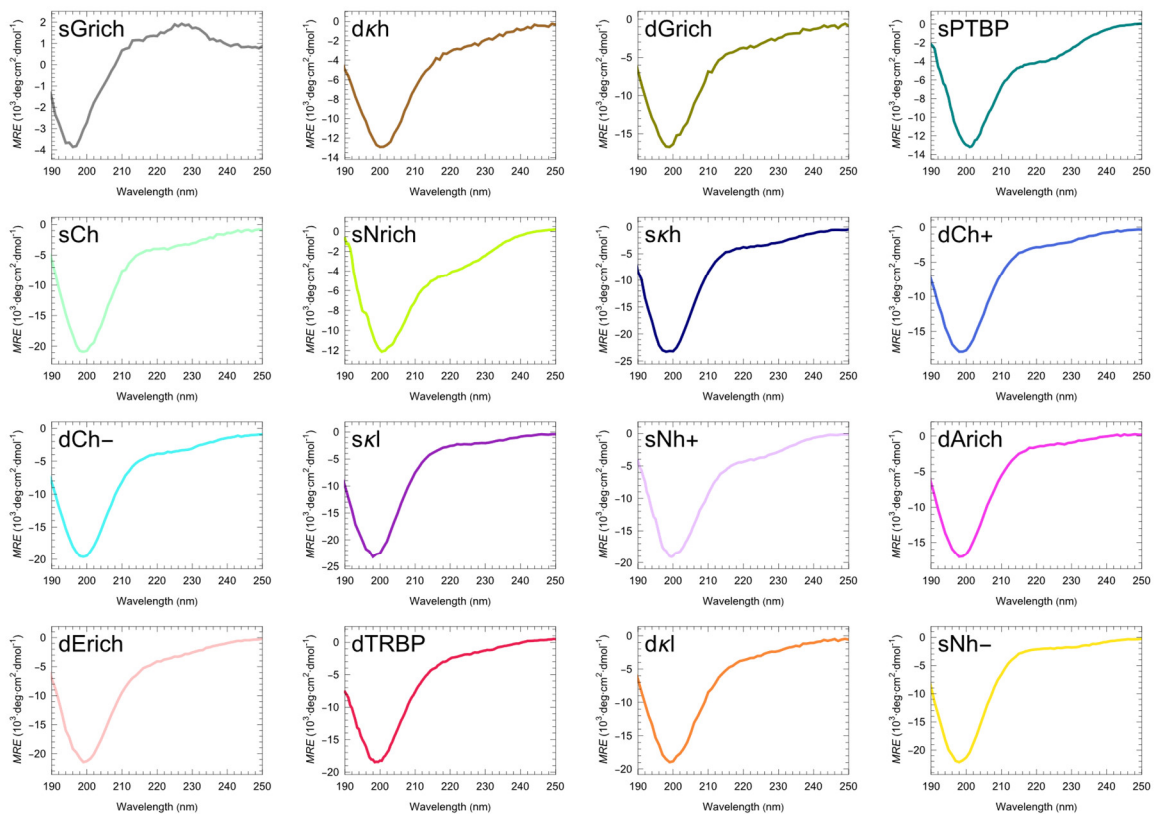


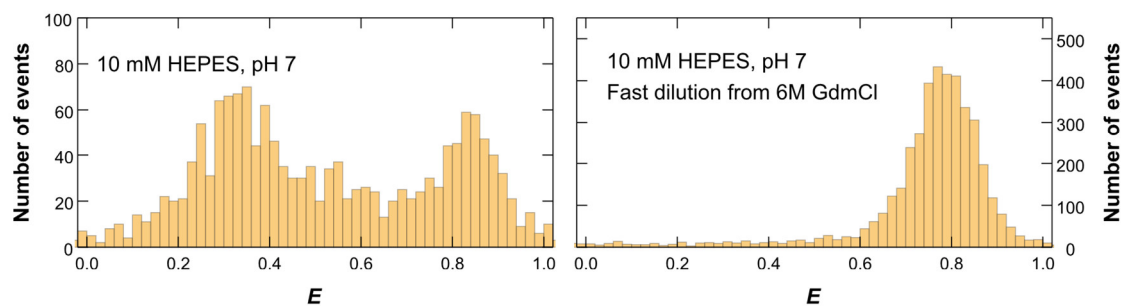
Figure S1. Selection scheme for the 16 IDR sequences.



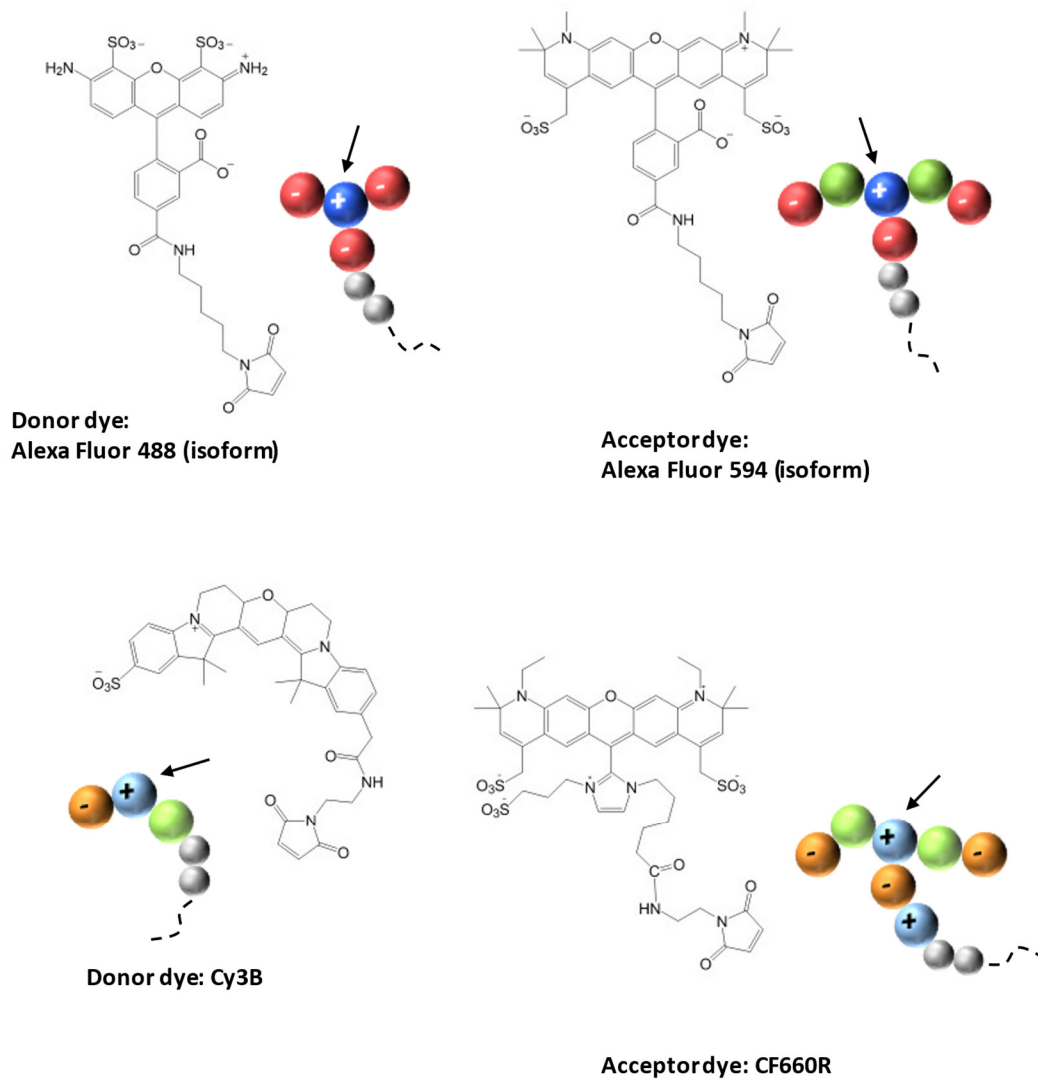
**Figure S2. Metapredict<sup>1</sup> disorder score for the 16 IDR sequences.** The disorder score is >0.5 (dashed line) for all sequences. Color code as in Fig. 1.



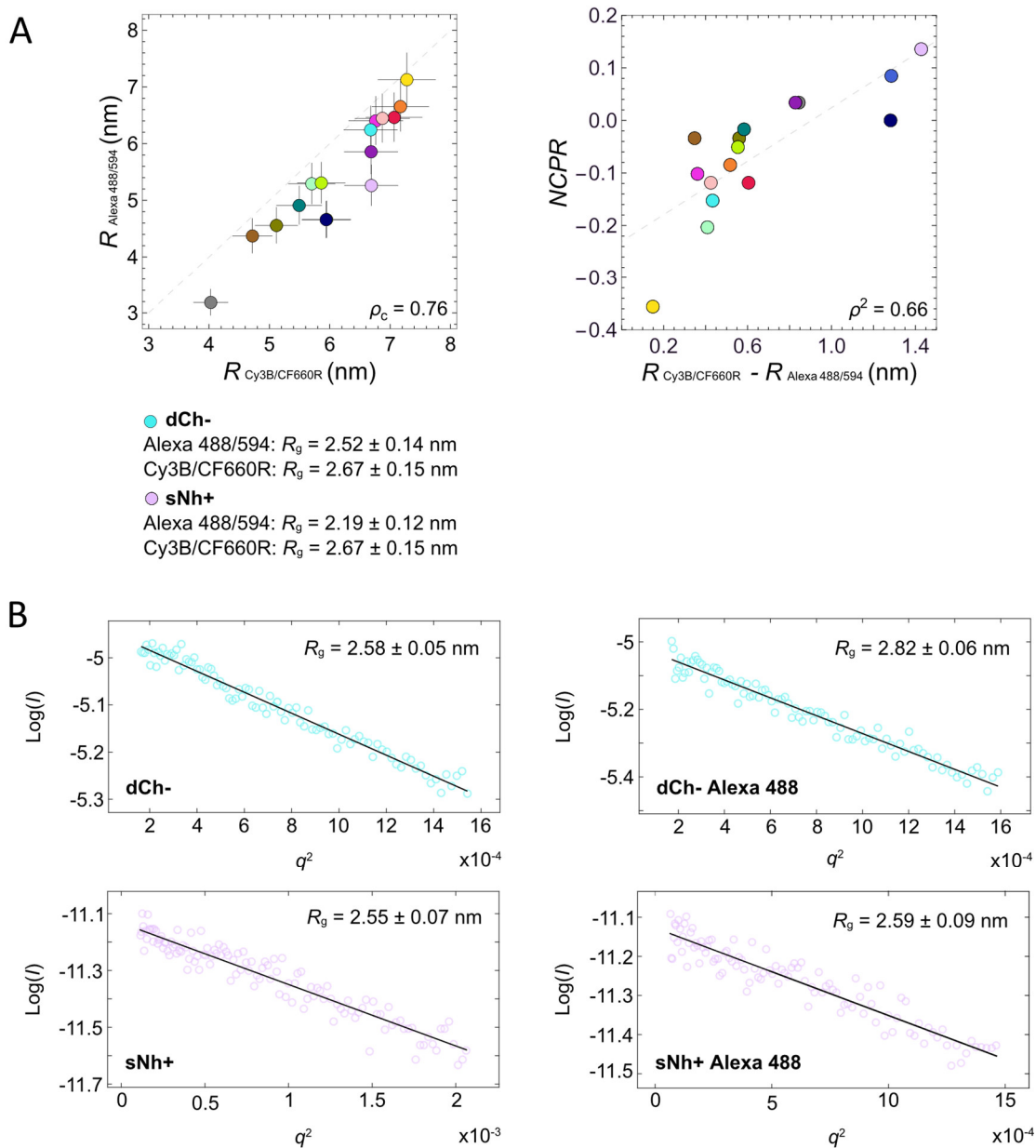
**Figure S3. Circular dichroism spectra of the IDRs indicate the absence of pronounced secondary structure.** The detectable sequence-specific differences between spectra are difficult to quantify reliably owing to the uncertainty in protein concentration measurements in sequences lacking aromatic amino acids (see Methods).



**Figure S4. Single-molecule analysis of *skh* labeled with Cy3B/CF660R identifies misfolding.** Initial analysis of *skh* in 10 mM 2-[4-(2-hydroxyethyl)piperazin-1-yl]ethane-1-sulfonic acid (HEPES), pH 7 (left histogram) revealed the presence of misfolded or aggregated species in addition to the monomeric form. However, by rapidly diluting *skh* to picomolar concentrations from the fully denatured state in 6 M guanidine chloride (GdmCl), the formation of these undesirable species could be prevented (right histogram). Single-molecule spectroscopy thus helps to identify issues with misfolding or aggregation that can then be eliminated by a suitable choice of solution conditions and the exceedingly low protein concentrations used.

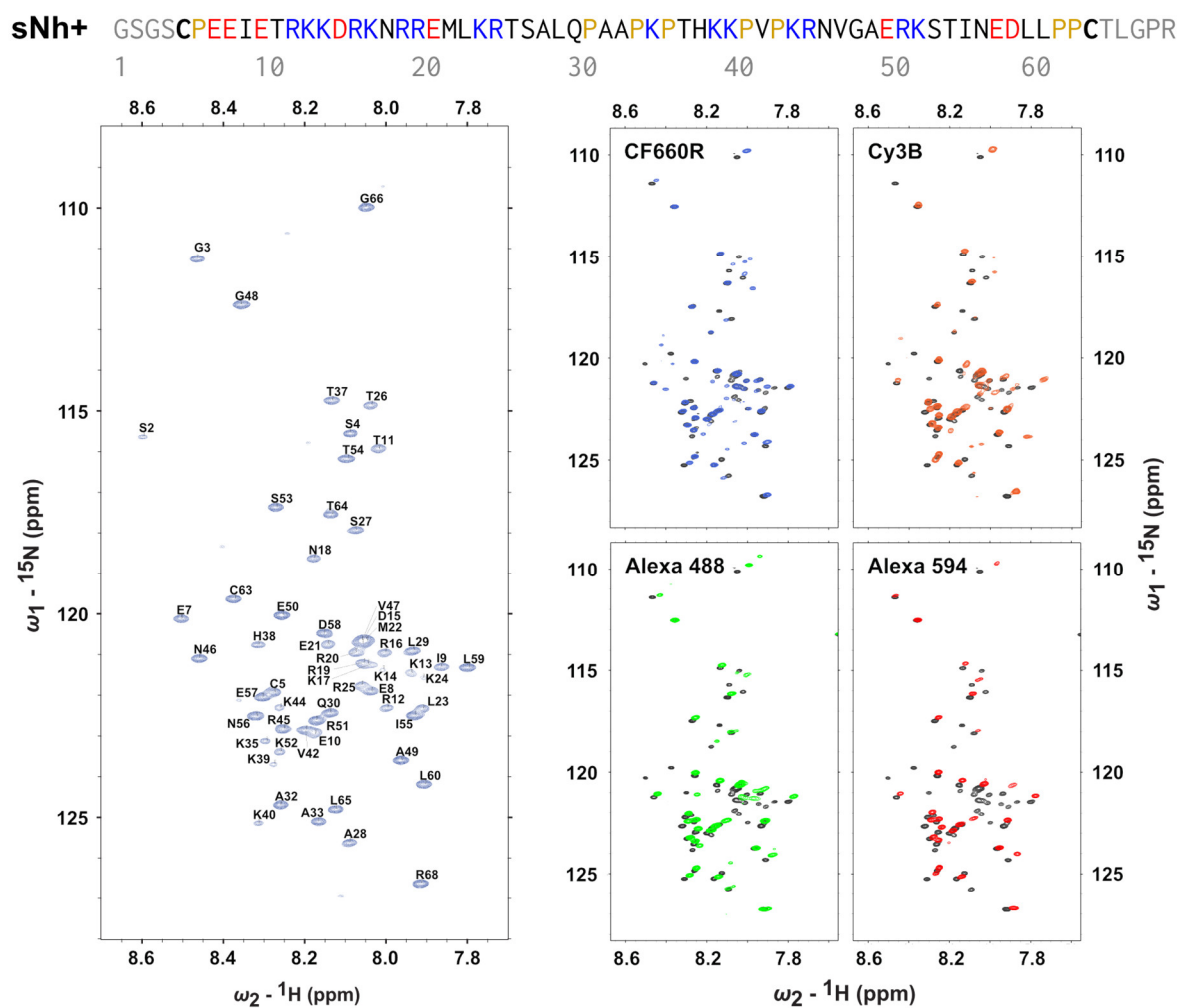


**Figure S5. Chemical structures of dyes Alexa 488/594 and Cy3B/CF660R, and their representation using beads for HPS model.** For both the Alexa 488/594 and Cy3B/CF660R dye pairs, we employ a representation with neutral, positively charged, and negatively charged beads. In the case of Alexa 488/594, these beads are color-coded as green (neutral), blue (positive), and red (negative), respectively; for Cy3B/CF660R, the corresponding beads are light green (neutral), light blue (positive) and orange (negative), with an arrow indicating the bead used for quantifying dye-dye distances. Additionally, dye linker beads (gray) are included in the representation.

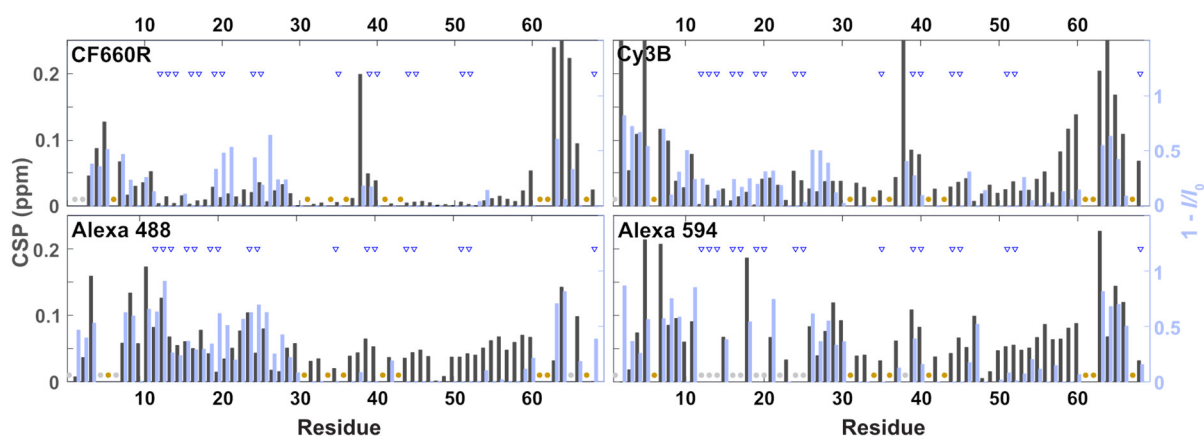


**Figure S6. Comparison of the IDRs labeled with the dye pairs Alexa 488/594 and Cy3B/CF660R by FRET, SAXS, and NMR.** (A) Correlation between the root-mean squared distance ( $R$ ) of the IDRs labeled with Alexa 488/594 and Cy3B/CF660R, respectively, from single-molecule FRET.  $R$  was inferred from the mean transfer efficiency assuming a SAW- $v$  distance distribution<sup>2</sup>, with error bars based on a systematic uncertainty of  $\pm 7\%$  in the Förster radius<sup>3</sup> used to calculate  $R$ ; radii of gyration ( $R_g$ ) for comparison with (B) were estimated from  $R$  and  $v$ .<sup>2</sup> Color code for the sequences as in Fig. 1. (B) SAXS measurements of dCh- and sNh+ unlabeled (left) and double-labeled with Alexa 488 (right).  $R_g$  was obtained from Guinier fits to the linear region of the scattering curve. [continued on the next page]

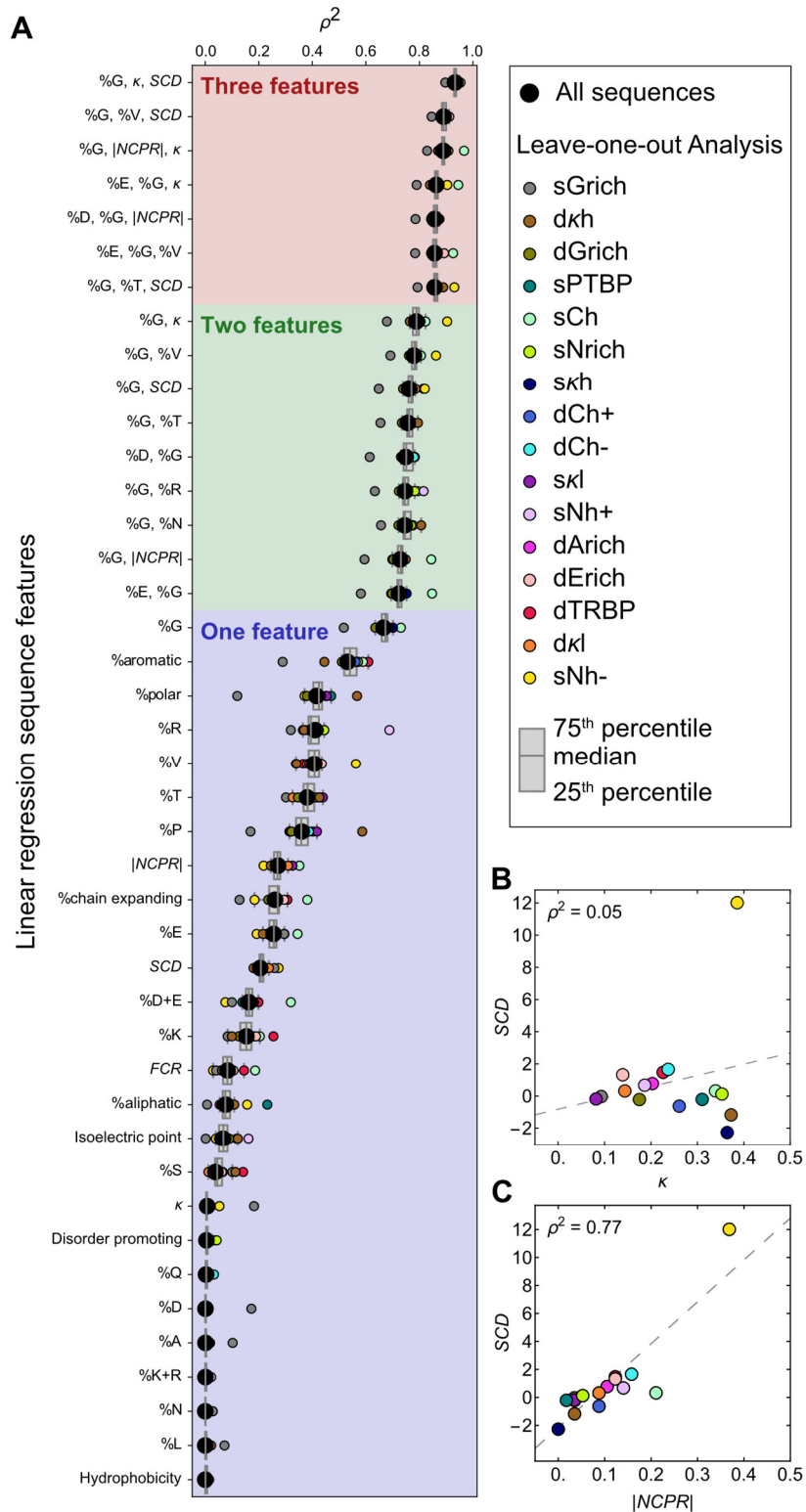
C



D

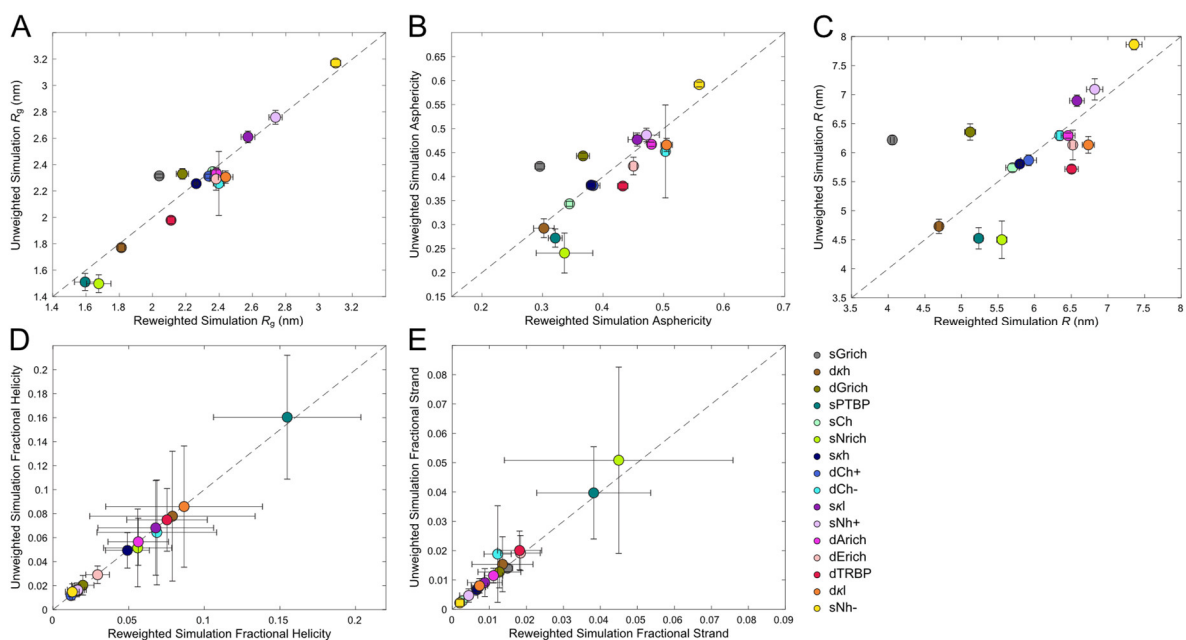


**Figure S6 [continued from previous page].** Comparison of the IDRs labeled with the dye pairs Alexa 488/594 and Cy3B/CF660R by FRET, SAXS, and NMR. (C) Left: Assigned HSQC spectrum of unlabeled sNh+; right: HSQC spectra of sNh+ double-labeled with the fluorophores indicated (colored spectra) and of the unlabeled IDRs (black spectra). (D) Chemical shift perturbations (CSP, dark gray bars) and resonance intensity decrease (light blue bars) upon double-labeling of sNh+ with the fluorophores indicated; blue triangles: positively charged amino acids; brown circles: resonances that were assigned in the unlabeled peptide but could not be assigned in the labeled peptide; light gray circles: proline residues.

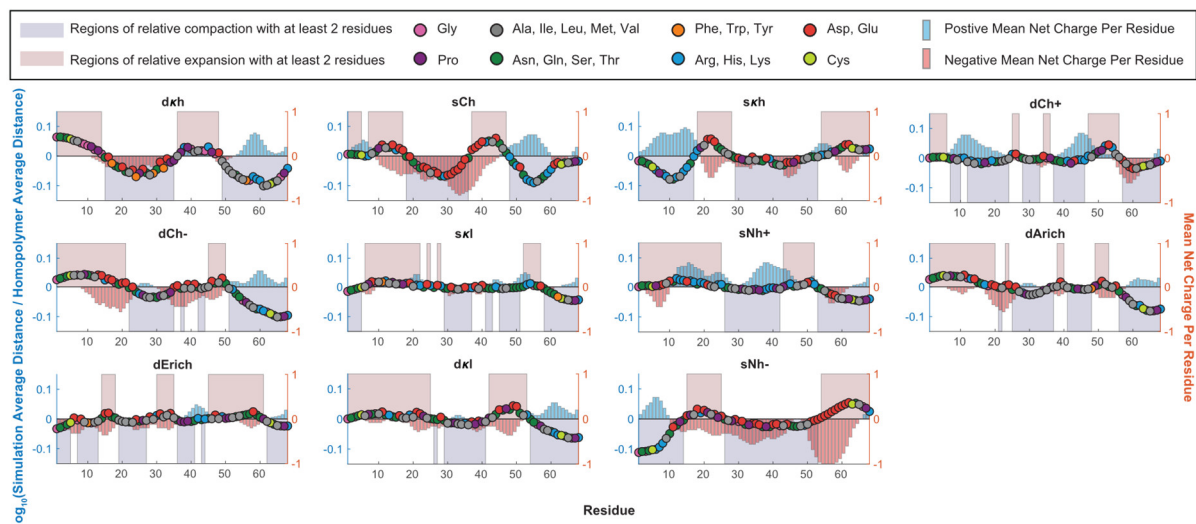


**Figure S7. Correlations between different sequence parameters and average transfer efficiency.** (A) Single-letter codes for amino acids are used; aro: aromatic residues; *SCD*: sequence charge decoration. Color code for the sequences as in Fig. 1.  $\rho^2$  analysis for linear regression of various compositional sequence features when including all sequences (black circle) or all sequences but one (colored circles). (B) and (C) Correlation analysis of  $\kappa$  and *SCD* and *SCD* and  $|NCPRI|$  showing correlation between *SCD* and  $|NCPRI|$ .

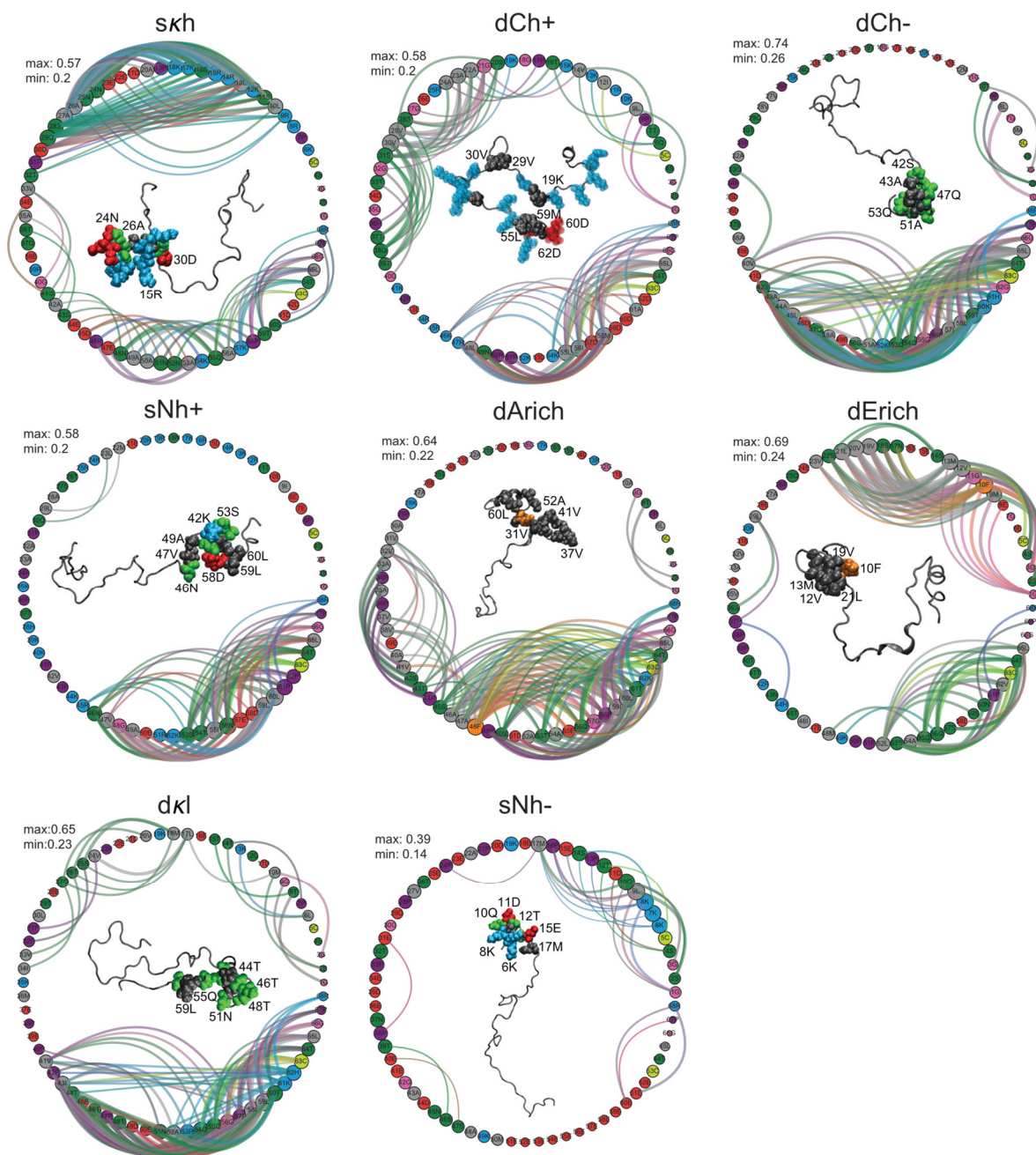




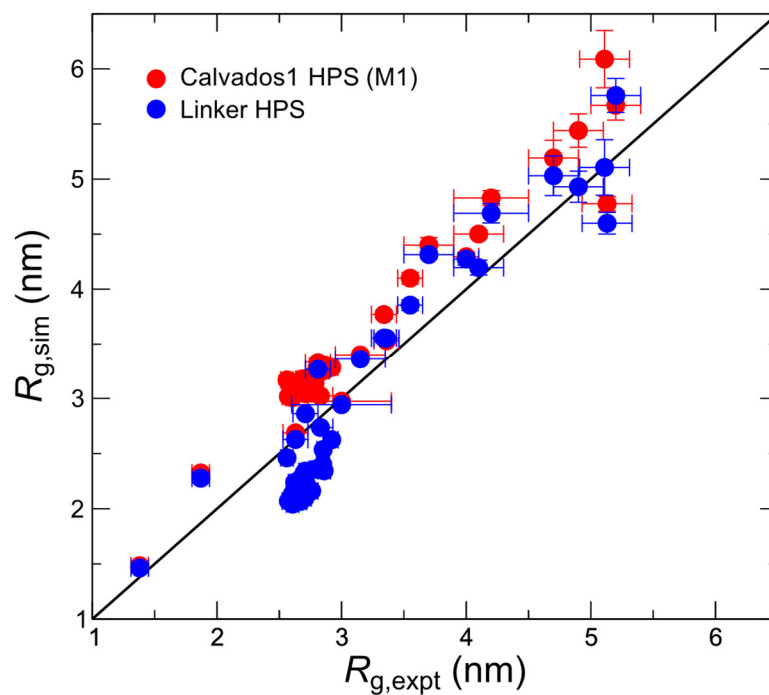
**Figure S8. Comparisons of conformational parameters of unweighted (prior) and reweighted (posterior) ABSINTH ensembles.** (A) The prior (ordinate) vs posterior values (abscissa) for (A) radius of gyration,  $R_g$ , (B) asphericity, (C) root-mean-squared end-to-end distance,  $R$ , (D) DSSP<sup>4</sup> fractional helical content, and (E) DSSP fractional strand content. For asphericity, values between 0.1 and 0.3 refer to roughly spherical envelopes for the ensembles. Values greater than 0.5 refer to prolate ellipsoids<sup>5</sup>. In all panels, the dashed lines define the lines of equality between prior and posterior values. Error bars represent the standard errors in the estimates of the mean values, which are shown as symbols.



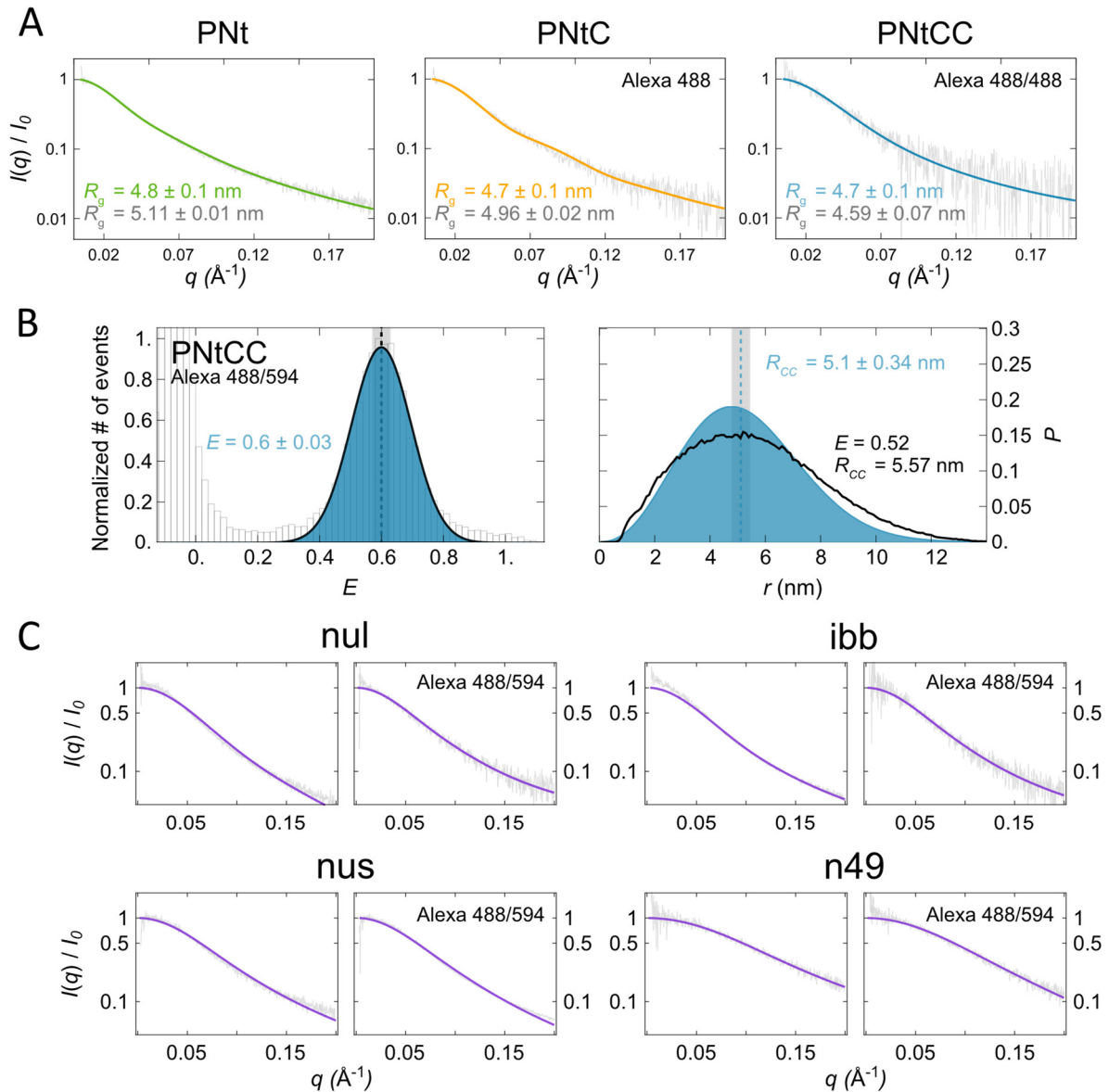
**Figure S9. Degree of local expansion and contraction relative to the best fit homopolymer model.** Circles are colored by their residue type. Consecutive regions of at least two residues that show compaction or expansion relative to the homopolymer model are shown by blue and red boxes, respectively. Mean net charge per residue profiles, averaged over five residue stretches, are shown as bar plots. Only the results for the sequences for which reweighting resulted in a Kullback-Leibler divergence below 0.1 are shown.



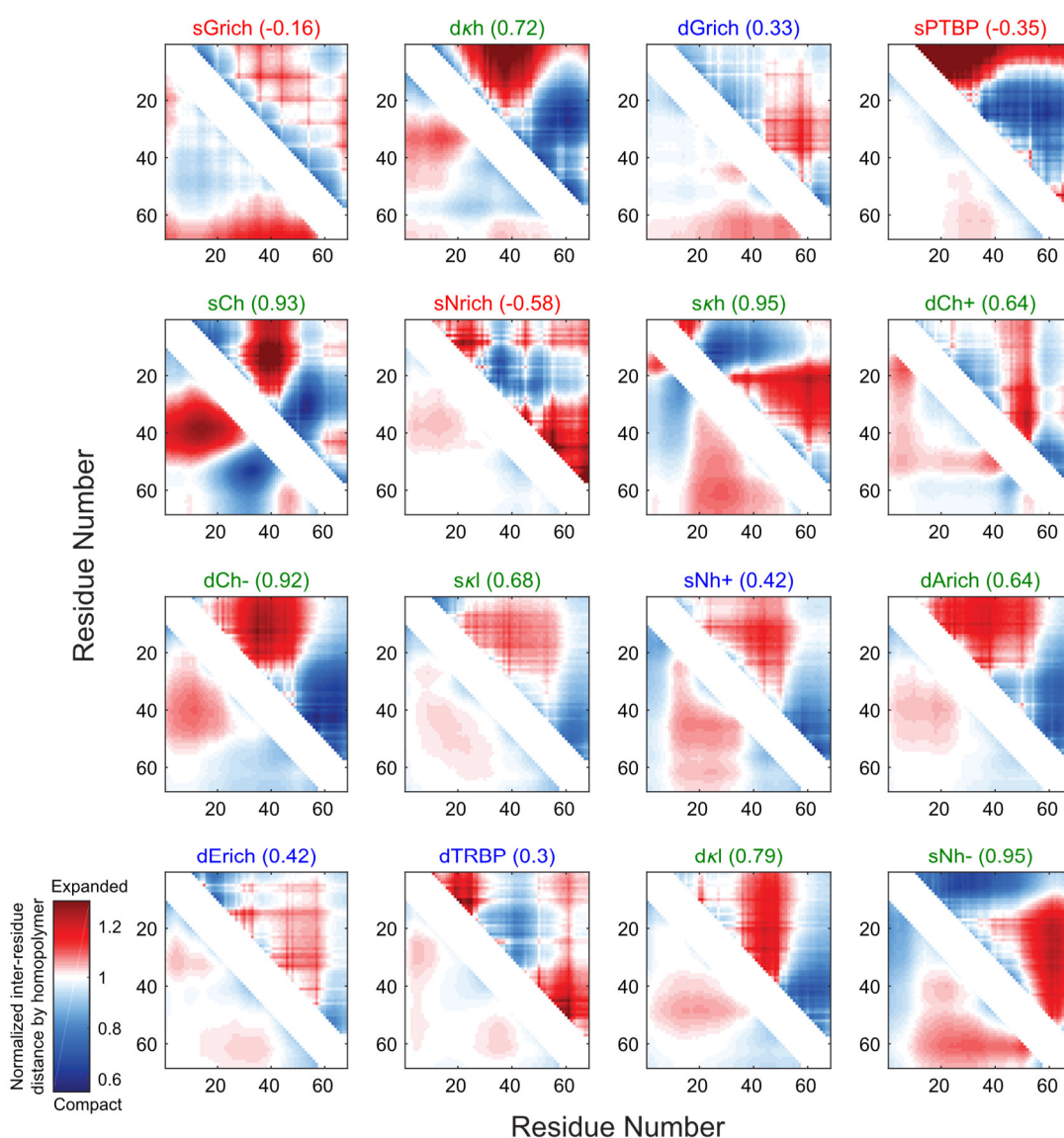
**Figure S10. Contact networks for the IDRs not shown in Fig. 3E.** Residues are shown as nodes with the circle size correlated to the mean contact probability between that residue and all other residues greater than two residues away in linear sequence space. Edges are drawn between two residues if they are at least 35 % of the maximum contact probability observed for that IDR. The maximum and minimum contact probabilities used to draw edges are listed in the figure as min and max. The width of the edge is 10 times the mean contact probability. Here, a contact distance of 10 Å is used such that charge interactions can be observed. Gly is shown in pink, Ser, Thr, Asn, and Gln in green, Arg, Lys, and His in blue, Asp and Glu in red, Phe, Trp, and Tyr in orange, Met, Val, Ile, Leu, and Ala in grey, Pro in purple, and Cys in lime green. Edge colors are the mixture of the interacting residue colors. Representative snapshots are visualized using VMD<sup>6</sup> and chosen by finding the frame that has the highest weight with a radius of gyration ( $R_g$ ) within 0.5 Å of the average  $R_g$  for the IDR. Only the results for the sequences for which reweighting resulted in a Kullback-Leibler divergence below 0.1 are shown.



















**Figure S11. Comparison of CALVADOS 1 HPS model<sup>7</sup> (M1) and “Linker HPS” model optimized here applied to the CALVADOS 1 training set.** The radii of gyration,  $R_{g,sim}$ , were computed from simulations of the 42 proteins in the CALVADOS training set<sup>7</sup> using the CALVADOS 1 (M1)  $\lambda$  values (red) or the  $\lambda$  values we are presenting here (“Linker HPS”, blue), and compared with experimental radii of gyration,  $R_{g,expt}$ . The cluster of A1 LCD-related sequences is located roughly between 2.5 and 3.0 nm in  $R_{g,expt}$ .



**Figure S12.** (A) SAXS curves of PNT variants<sup>8</sup>, left to right: unlabeled (PNT), labeled with one Alexa 488 dye (PNTc), and labeled with Alexa 488 at both Cys residues (PNTCC), compared with results based on the simulations with the optimized HPS model (colored lines). (B) Transfer efficiency histogram of PNTCC labeled with Alexa 488/594 measured in 10 mM Tris and 175 mM KCl, pH 7.4 (ionic strength 188 mM, accounting for residual GdmCl from dilution of labeled protein), fit with a Gaussian peak function (blue) to estimate the mean transfer efficiency. Distance distributions,  $P(r)$ , based on the SAW- $v$  model<sup>2</sup> (blue) and the optimized coarse-grained HPS model (black line). Vertical dashed line indicates the root-mean squared distance ( $R_{cc}$ ) from SAW- $v$ , with a gray error band based on a systematic uncertainty of  $\pm 7\%$  in the Förster radius<sup>9</sup>. The difference between experimental and simulation FRET efficiencies is likely due to the limitations in the HPS model in reproducing the sequence-specific dimensions of the 88-residue segment between the labels; the discrepancy is within the range of deviations obtained for the linker IDRs (Fig. 4B). Note that the  $R_g$  estimated from experimental or simulated FRET using the SAW- $v$  and extrapolated to the full sequence length is 4.51 and 5.07 nm respectively, lying on either side of the estimate from SAXS. (C) Comparison of experimental SAXS curves of the disordered proteins nul, ibb, nus and n49, unlabeled and labeled with Alexa 488/594 (gray lines) from Fuentes *et al.*<sup>10</sup> with SAXS curves calculated from simulations with the optimized HPS model including dyes (violet lines).



**Figure S13. Comparison of normalized intra-chain distances computed using the reweighted ABSINTH simulations (upper triangle) and the HPS simulations with Alexa 488/594 (lower triangle).** Results are shown with a minimum spacing of 10 amino acids and relative to the value from the best fit of a homopolymer model (see Fig. 3). Regions of local expansion relative to the equivalent homopolymer are shown in red, areas of local compaction in blue (see color scale). The title for each panel includes the name of the sequence, and the numbers in parentheses are Pearson correlation coefficients between  $O(10^3)$  pairs of normalized distances from the two models. Strong positive correlations corresponding to correlation coefficients greater than +0.6 are marked in green. Sequences marked in blue show weak positive correlations between +0.3 and +0.5. For three sequences, the correlation coefficients are negative, implying that the reweighted ABSINTH ensembles are inconsistent the HPS ensembles. For sGrich, the HPS ensemble shows uniform compaction through the middle of the chain with the ends avoiding one another. In contrast, the reweighted ABSINTH ensemble yields numerous local loops and very few long-range attractions. For sNrich, the regions of attractions and repulsions are inverted across the two models. Finally, for sPTBP, the reweighted ABSINTH ensemble shows some helicity and a preference for turn-like structures. The HPS ensemble shows weaker overall contact preferences, suggestive of interactions closer to the homopolymer model than those in ABSINTH.

	UniProt	Netcharge	FCR	NCPR	Hydrophobicity	$\kappa$	SCD
sNh-	 Q9Y4C8	-21	0.54	0.37	0.26	0.39	12.01
dkl	 Q6NXA4	-5	0.26	0.09	0.41	0.14	0.32
dTRBP	 Q15633	-7	0.16	0.12	0.49	0.23	1.47
dErich	 Q6GPZ1	-7	0.23	0.12	0.44	0.14	1.32
dArich	 Q12906	-6	0.25	0.11	0.46	0.20	0.77
sNh+	 074968	+8	0.42	0.14	0.33	0.19	0.67
skl	 Q54PB2	+2	0.35	0.04	0.36	0.08	-0.18
dCh-	 Q91550	-9	0.30	0.16	0.39	0.24	1.66
dCh+	 P25159	+5	0.37	0.09	0.37	0.26	-0.63
skh	 A7TQR2	0	0.39	0.00	0.30	0.36	-2.27
sNrich	 P32831	-3	0.12	0.05	0.43	0.35	0.13
sCh	 P37838	-12	0.60	0.21	0.19	0.34	0.32
sPTBP	 P26599	-1	0.09	0.02	0.55	0.31	-0.21
dGrich	 P24785	-2	0.25	0.04	0.38	0.17	-0.22
dkh	 Q9NS39	-2	0.21	0.04	0.45	0.37	-1.17
sGrich	 Q43349	+2	0.18	0.04	0.35	0.09	-0.03

**Table S1:** Summary of UniProt<sup>11</sup> identifiers and important physicochemical parameters of the IDRs

	$\langle E \rangle_{A488A594}$	$\langle E \rangle_{Cy3bCF660R}$	$R_{A488A594}$ [nm]	$R_{Cy3bCF660R}$ [nm]	$V_{A488A594}$	$V_{Cy3bCF660R}$
sNh-	$0.36 \pm 0.02$	$0.42 \pm 0.03$	$7.1^{+0.5}_{-0.5}$	$7.3^{+0.5}_{-0.5}$	$0.61^{+0.02}_{-0.02}$	$0.61^{+0.02}_{-0.02}$
dk1	$0.41 \pm 0.02$	$0.43 \pm 0.03$	$6.7^{+0.4}_{-0.4}$	$7.2^{+0.5}_{-0.5}$	$0.59^{+0.02}_{-0.02}$	$0.61^{+0.02}_{-0.02}$
dTRBP	$0.437 \pm 0.006$	$0.45 \pm 0.03$	$6.5^{+0.4}_{-0.4}$	$7.1^{+0.5}_{-0.5}$	$0.59^{+0.02}_{-0.02}$	$0.61^{+0.02}_{-0.02}$
dErich	$0.44 \pm 0.03$	$0.47 \pm 0.03$	$6.4^{+0.4}_{-0.4}$	$6.9^{+0.5}_{-0.5}$	$0.59^{+0.02}_{-0.02}$	$0.60^{+0.02}_{-0.02}$
dArich	$0.445 \pm 0.009$	$0.48 \pm 0.02$	$6.4^{+0.4}_{-0.4}$	$6.8^{+0.4}_{-0.4}$	$0.58^{+0.02}_{-0.02}$	$0.60^{+0.02}_{-0.02}$
sNh+	$0.610 \pm 0.007$	$0.49 \pm 0.03$	$5.3^{+0.4}_{-0.4}$	$6.7^{+0.4}_{-0.4}$	$0.54^{+0.02}_{-0.02}$	$0.59^{+0.02}_{-0.02}$
sk1	$0.52 \pm 0.02$	$0.49 \pm 0.03$	$5.9^{+0.4}_{-0.4}$	$6.7^{+0.4}_{-0.4}$	$0.56^{+0.02}_{-0.02}$	$0.59^{+0.02}_{-0.02}$
dCh-	$0.46 \pm 0.03$	$0.50 \pm 0.03$	$6.2^{+0.4}_{-0.4}$	$6.7^{+0.4}_{-0.4}$	$0.58^{+0.02}_{-0.02}$	$0.59^{+0.02}_{-0.02}$
dCh+	$0.70 \pm 0.02$	$0.59 \pm 0.03$	$4.7^{+0.3}_{-0.3}$	$5.9^{+0.4}_{-0.4}$	$0.51^{+0.02}_{-0.02}$	$0.57^{+0.02}_{-0.02}$
skh	$0.706 \pm 0.002$	$0.595 \pm 0.005$	$4.7^{+0.3}_{-0.3}$	$5.9^{+0.4}_{-0.4}$	$0.51^{+0.02}_{-0.02}$	$0.57^{+0.02}_{-0.02}$
sNrich	$0.60 \pm 0.02$	$0.61 \pm 0.02$	$5.3^{+0.4}_{-0.4}$	$5.9^{+0.4}_{-0.4}$	$0.54^{+0.02}_{-0.02}$	$0.56^{+0.02}_{-0.02}$
sCh	$0.605 \pm 0.003$	$0.63 \pm 0.04$	$5.3^{+0.4}_{-0.4}$	$5.7^{+0.4}_{-0.4}$	$0.54^{+0.02}_{-0.02}$	$0.56^{+0.02}_{-0.02}$
sPTBP	$0.666 \pm 0.006$	$0.66 \pm 0.02$	$4.9^{+0.3}_{-0.3}$	$5.5^{+0.4}_{-0.4}$	$0.52^{+0.02}_{-0.02}$	$0.55^{+0.02}_{-0.02}$
dGrich	$0.72 \pm 0.02$	$0.71 \pm 0.02$	$4.6^{+0.3}_{-0.3}$	$5.1^{+0.4}_{-0.4}$	$0.50^{+0.02}_{-0.02}$	$0.53^{+0.02}_{-0.02}$
dkh	$0.75 \pm 0.02$	$0.77 \pm 0.02$	$4.4^{+0.3}_{-0.3}$	$4.7^{+0.3}_{-0.3}$	$0.49^{+0.02}_{-0.02}$	$0.51^{+0.02}_{-0.02}$
sGrich	$0.912 \pm 0.004$	$0.861 \pm 0.008$	$3.2^{+0.2}_{-0.2}$	$4.0^{+0.3}_{-0.3}$	$0.42^{+0.02}_{-0.02}$	$0.47^{+0.02}_{-0.02}$

**Table S2:** Mean transfer efficiencies,  $\langle E \rangle$ , averaged from at least three independent measurements, the corresponding average root-mean squared distances,  $R$ , and scaling exponents,  $\nu$ , from SAW- $\nu$ . The uncertainties for  $R$  and  $\nu$  are based on a systematic uncertainty of  $\pm 7\%$  in the Förster radius<sup>9</sup>.



	Initial	Optimized	Calvados2
Ala	0.003	0.238	0.274
Arg	0.723	0.79	0.731
Asn	0.16	0.298	0.426
Asp	0.002	0.223	0.042
Cys	0.4	0.414	0.562
Gln	0.468	0.502	0.393
Glu	0.022	0.05	0.001
Gly	0.784	1.268	0.706
His	0.487	0.497	0.466
Ile	0.687	0.669	0.542
Leu	0.335	0.394	0.644
Lys	0.095	0.083	0.179
Met	0.993	0.982	0.531
Phe	0.871	0.915	0.867
Pro	0.471	0.299	0.359
Ser	0.487	0.572	0.463
Thr	0.274	0.181	0.371
Trp	0.753	0.753	0.989
Tyr	0.984	1.024	0.977
Val	0.428	0.267	0.208
Aneg	1	1.109	
Aneu	1	1.083	
Apos	1	1.089	
Cneg	0.5	0.424	
Cneu	0.5	0.432	
Cpos	0.5	0.452	
Lin	0.75	0.783	

**Table S3. Original (CALVADOS 1<sup>12</sup> M3), optimized, and CALVADOS 2<sup>13</sup> short-range interaction parameters ( $\lambda$ ).** Aneg, Aneu, and Apos are the parameters for the negatively charged, neutral, and positively charged Alexa 488/594 dye beads, respectively. Cneg, Cneu, and Cpos are the parameters for the negatively charged, neutral, and positively charged Cy3B/CF660R dye beads, and Lin for the linking dye beads, respectively (see Fig. S5). Values for the parameters  $\sigma$  and  $\varepsilon$  in the HPS model were taken from the previous version of the model<sup>14</sup>; for each dye bead, a value of 0.582 nm was used for  $\sigma$ .

## References

1. Emenecker, R. J.; Griffith, D.; Holehouse, A. S., Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys. J.* **2021**, *120* (20), 4312-4319.
2. Zheng, W.; Zerze, G. H.; Borgia, A.; Mittal, J.; Schuler, B.; Best, R. B., Inferring properties of disordered chains from FRET transfer efficiencies. *J. Chem. Phys.* **2018**, *148* (12), 123329.
3. Hellenkamp, B.; Schmid, S.; Doroshenko, O.; Opanasyuk, O.; Kuhnemuth, R.; Rezaei Adariani, S.; Ambrose, B.; Aznauryan, M.; Barth, A.; Birkedal, V.; Bowen, M. E.; Chen, H.; Cordes, T.; Eilert, T.; Fijen, C.; Gebhardt, C.; Gotz, M.; Gouridis, G.; Gratton, E.; Ha, T.; Hao, P.; Hanke, C. A.; Hartmann, A.; Hendrix, J.; Hildebrandt, L. L.; Hirschfeld, V.; Hohlbein, J.; Hua, B.; Hubner, C. G.; Kallis, E.; Kapanidis, A. N.; Kim, J. Y.; Krainer, G.; Lamb, D. C.; Lee, N. K.; Lemke, E. A.; Levesque, B.; Levitus, M.; McCann, J. J.; Naredi-Rainer, N.; Nettels, D.; Ngo, T.; Qiu, R.; Robb, N. C.; Rocker, C.; Sanabria, H.; Schlierf, M.; Schroder, T.; Schuler, B.; Seidel, H.; Streit, L.; Thurn, J.; Tinnefeld, P.; Tyagi, S.; Vandenberg, N.; Vera, A. M.; Weninger, K. R.; Wunsch, B.; Yanez-Orozco, I. S.; Michaelis, J.; Seidel, C. A. M.; Craggs, T. D.; Hugel, T., Precision and accuracy of single-molecule FRET measurements—a multi-laboratory benchmark study. *Nat. Methods* **2018**, *15* (9), 669-676.
4. Kabsch, W.; Sander, C., Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22* (12), 2577-2637.
5. Steinhauser, M. O., A molecular dynamics study on universal properties of polymer chains in different solvent qualities. Part I. A review of linear chain properties. *J. Chem. Phys.* **2005**, *122* (9), 094901.
6. Humphrey, W.; Dalke, A.; Schulten, K., VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33-38.
7. Tesei, G.; Schulze, T. K.; Crehuet, R.; Lindorff-Larsen, K., Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Natl. Acad. Sci. USA* **2021**, *118* (44).
8. Riback, J. A.; Bowman, M. A.; Zmyslowski, A. M.; Plaxco, K. W.; Clark, P. L.; Sosnick, T. R., Commonly used FRET fluorophores promote collapse of an otherwise disordered protein. *Proc. Natl. Acad. Sci. USA* **2019**, *116* (18), 8889-8894.
9. Holmstrom, E. D.; Holla, A.; Zheng, W.; Nettels, D.; Best, R. B.; Schuler, B., Accurate Transfer Efficiencies, Distance Distributions, and Ensembles of Unfolded and Intrinsically Disordered Proteins From Single-Molecule FRET. *Methods Enzymol.* **2018**, *611*, 287-325.
10. Fuertes, G.; Banterle, N.; Ruff, K. M.; Chowdhury, A.; Mercadante, D.; Koehler, C.; Kachala, M.; Estrada Girona, G.; Milles, S.; Mishra, A.; Onck, P. R.; Grater, F.; Esteban-Martin, S.; Pappu, R. V.; Svergun, D. I.; Lemke, E. A., Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proceedings of the National Academy of Sciences of the United States of America* **2017**, *114* (31), E6342-E6351.
11. UniProt, C., UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2022**.
12. Tesei, G.; Schulze, T. K.; 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).
13. Tesei, G.; Lindorff-Larsen, K., Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Res Eur* **2022**, *2*, 94.
14. Dannenhoffer-Lafage, T.; Best, R. B., A Data-Driven Hydrophobicity Scale for Predicting Liquid-Liquid Phase Separation of Proteins. *J. Phys. Chem. B* **2021**, *125* (16), 4046-4056.