

Accurate predictions of conformational ensembles of disordered proteins with STARLING

Borna Novak^{*,1,2,3}, Jeffrey M. Lotthammer^{*,2,3}, Ryan J. Emenecker^{2,3}, Alex S. Holehouse^{2,3}✉

1. Medical Scientist Training Program, Washington University School of Medicine, St. Louis, MO, United States

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

3. Center for Biomolecular Condensates (CBC), Washington University in St. Louis, St. Louis, MO

* Authors contributed equally

✉ Corresponding author alex.holehouse@wustl.edu

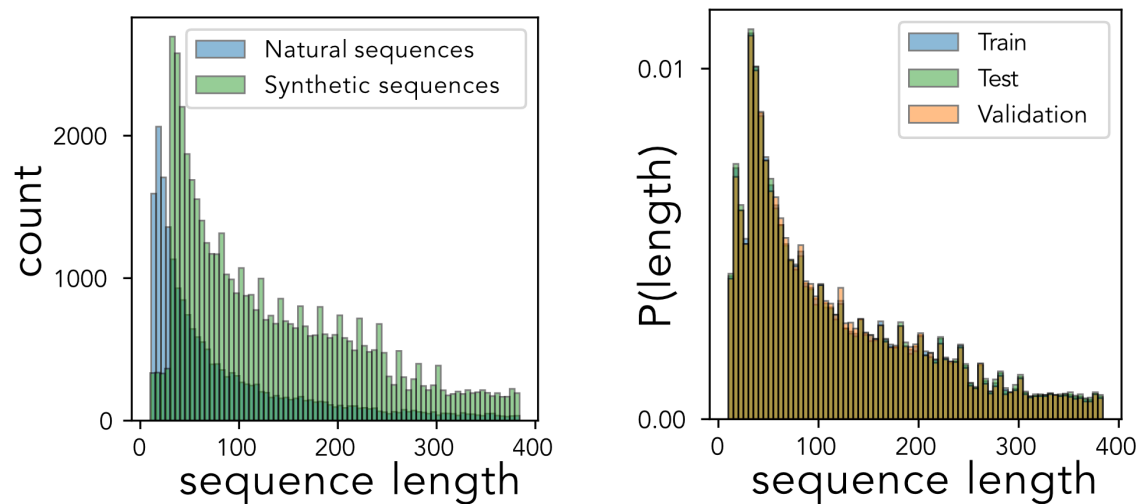


Figure S1: Our dataset of 70,563 consists of both natural sequences (N=20,349) and synthetic sequences designed by varying sequence features known to affect IDR ensembles (N=50,214) (left). The sequences were clustered and split into train (N=49,423), validation (N=10,703), and test (N=10,437) sets while maintaining the distribution of sequence lengths across the splits consistent (right).

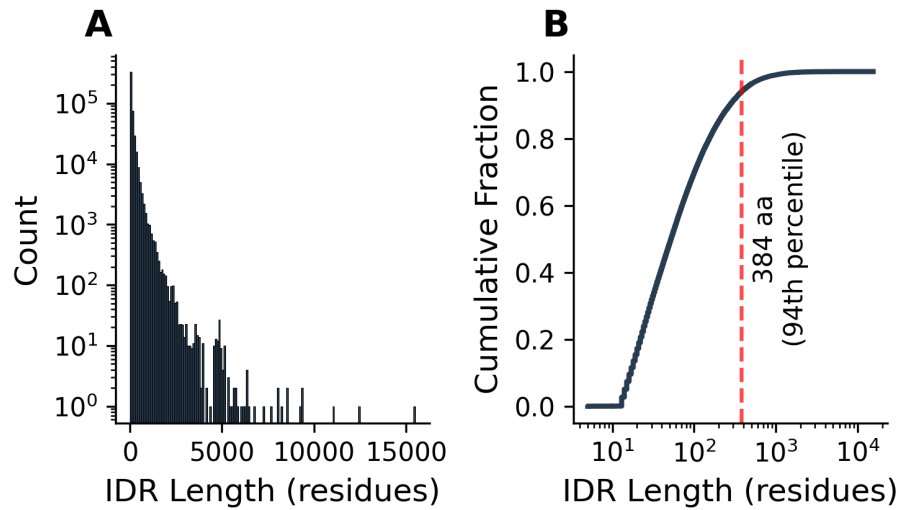
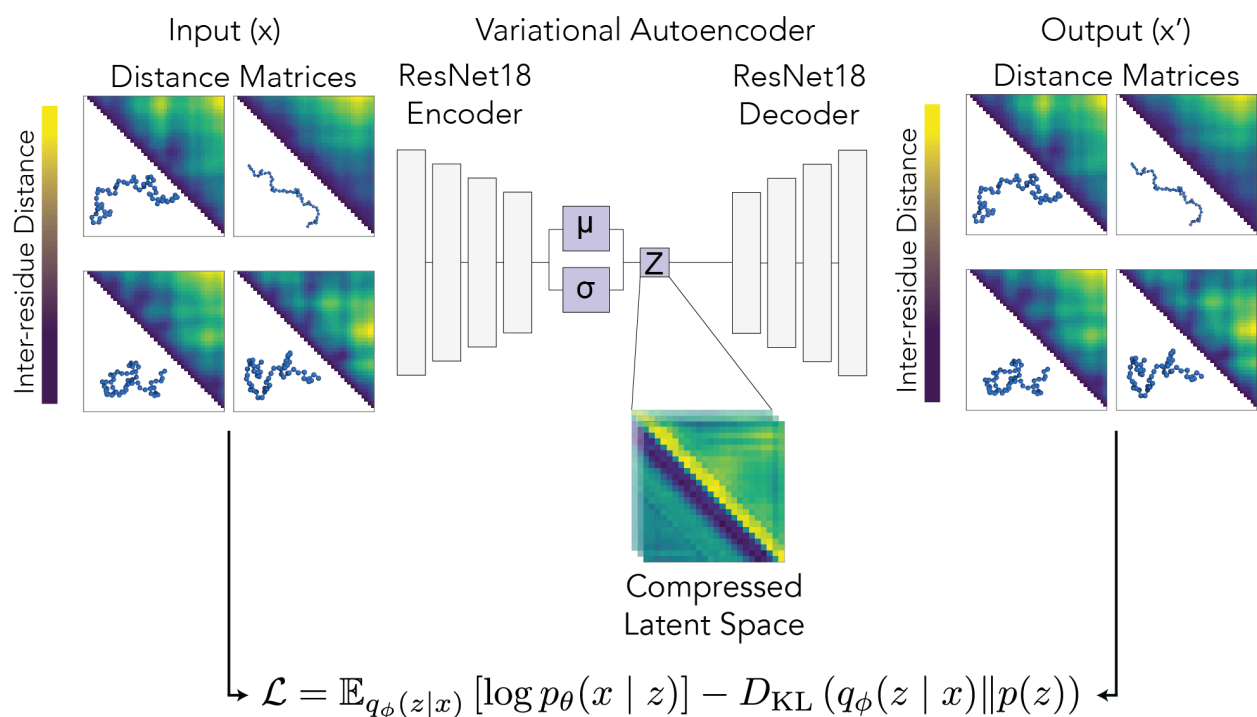


Figure S2: Disordered protein length distributions across a collection of model organism proteomes. **A.** Histogram of the protein sequence length distribution. **B.** A cumulative density function of the length distribution presented in A. 94% of the IDRs across common model organisms fall between 0 and 384 residues (the length cutoff of STARLING).



where $p_{\theta}(x | z) = \mathcal{N}(\mu_{\theta}(z), \sigma^2 I)$ and $x' \sim p_{\theta}(x | z) = \mathcal{N}(\mu_{\theta}(z), \sigma^2 I)$

Fig. S3: Schematic representation of the variational autoencoder (VAE). The input to our variational autoencoder is a collection of 2D inter-residue distance maps that undergo feature extraction and spatial compression through a ResNet18 encoder, which learns the μ and σ parameters of a normal distribution that parameterizes the latent features. The VAE is trained with an evidence-based lower bound (ELBO) loss comprising a reconstruction term and a Kullback–Leibler (KL) divergence term.

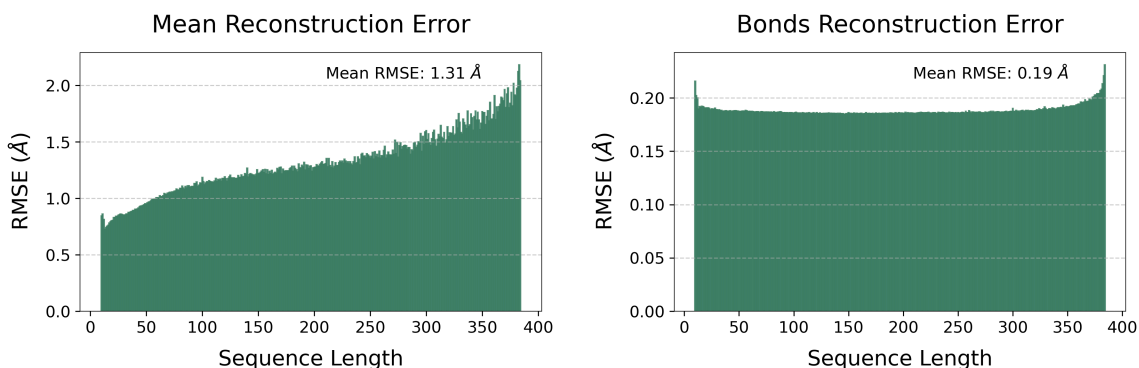


Figure S4: The variational autoencoder (VAE) is evaluated on a held-out test set of distance maps ($N = 1,425,849$). The VAE encoder spatially compresses these distance maps and is then reconstructed by the VAE decoder from their compressed representations. The root mean squared error (RMSE) is computed between the input and output distance maps. The trained VAE demonstrates a low mean RMSE, calculated over the upper triangle of each distance map (left). Accurate bond reconstruction is essential for generating physically realistic conformations, and our model achieves a very low mean RMSE (right).

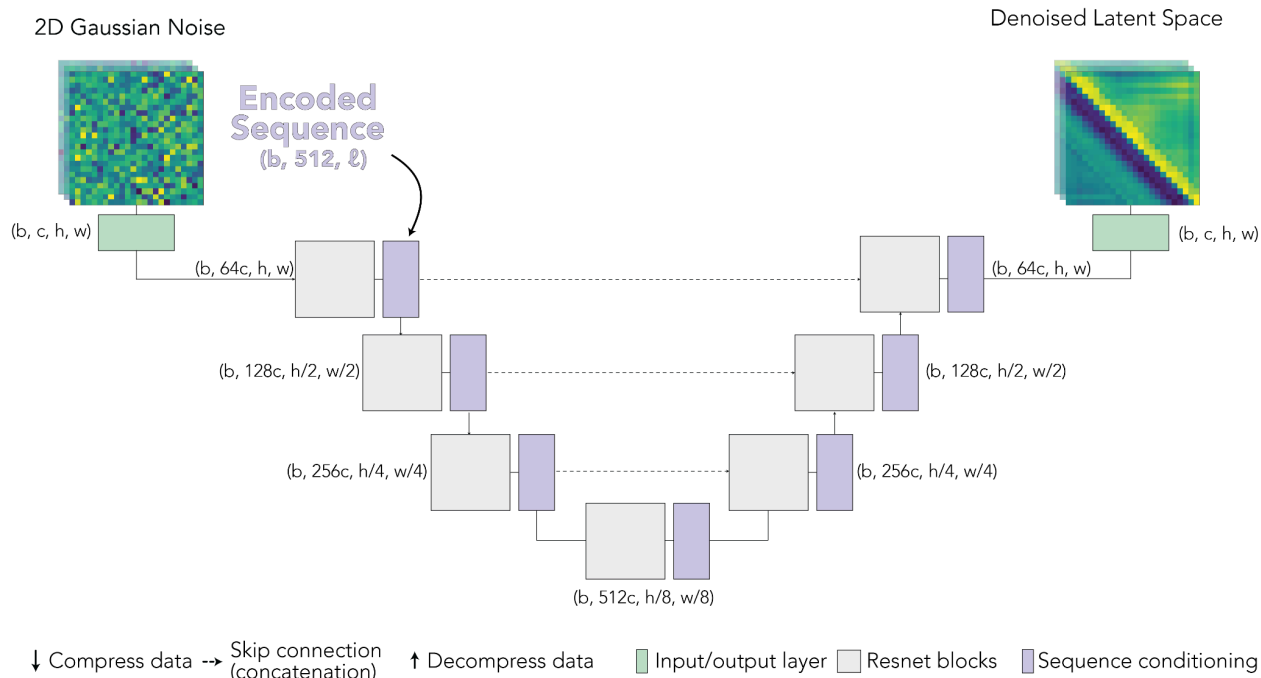


Fig. S5: Extended schematic representation of the Unet backbone of the STARLING discrete time denoising-diffusion probabilistic model (DDPM). The DDPM starts by taking a batch size (b) of 2D Gaussian noise with one channel (c) of features and spatial dimensions concordant with the VAE latent space (h, w). This data then undergoes three steps of spatial compression and feature extraction until we reach the base of the “U”, where we have 512x the features and 64x spatial compression ($8 \times 8 \times$). Following downsampling is a series of upsampling blocks that reverse the feature extraction and spatial compression, ultimately arriving at the original batch, channels, height, width dimensional vector (b, c, h, w) where the batch is the number of conformers, the channels are the noise estimate at each latent feature, the height and width are the compressed spatial dimensions which can subsequently be fed to the VAE decoder to project back to a $\ell \times \ell$ matrix of inter-residue distances. Notably, at each step of the downsampling and upsampling, we inject sequence information into the model by cross-attending to sequence embedding where each amino acid is encoded as a 512-dimensional vector comprising both learned amino acid token embeddings and positional encodings via the addition of sinusoidal embeddings.

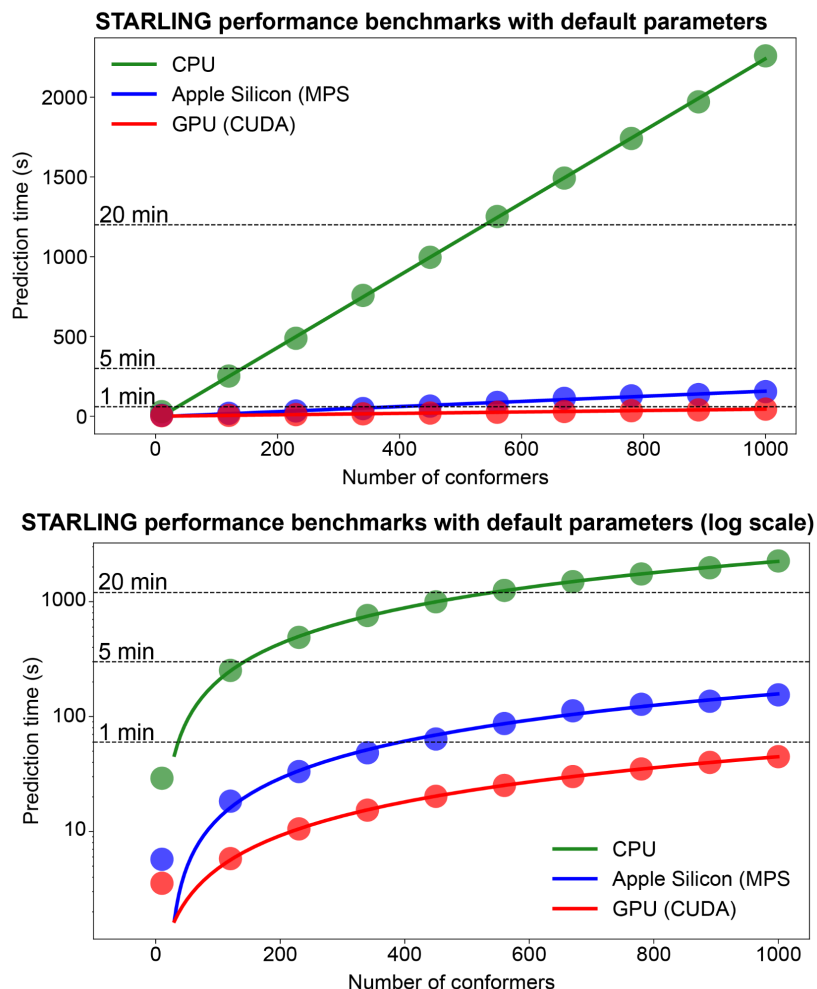


Fig. S6. STARLING benchmarking on different types of hardware. Runtime depends linearly on the number of conformers (although sequence length has no impact on runtime). The top and bottom panels show identical data in log vs. linear space. On GPUs and Apple Silicon (facilitated by CUDA and MPS, respectively), STARLING is highly performant, offering ensembles sufficient for detailed biophysical investigation in 10-15 seconds (GPU) or 30-60 seconds (MPS). Specific hardware tested here are Nvidia A4500 (GPU), Macbook Pro M3 Max CPU (MPS), and Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz (CPU).

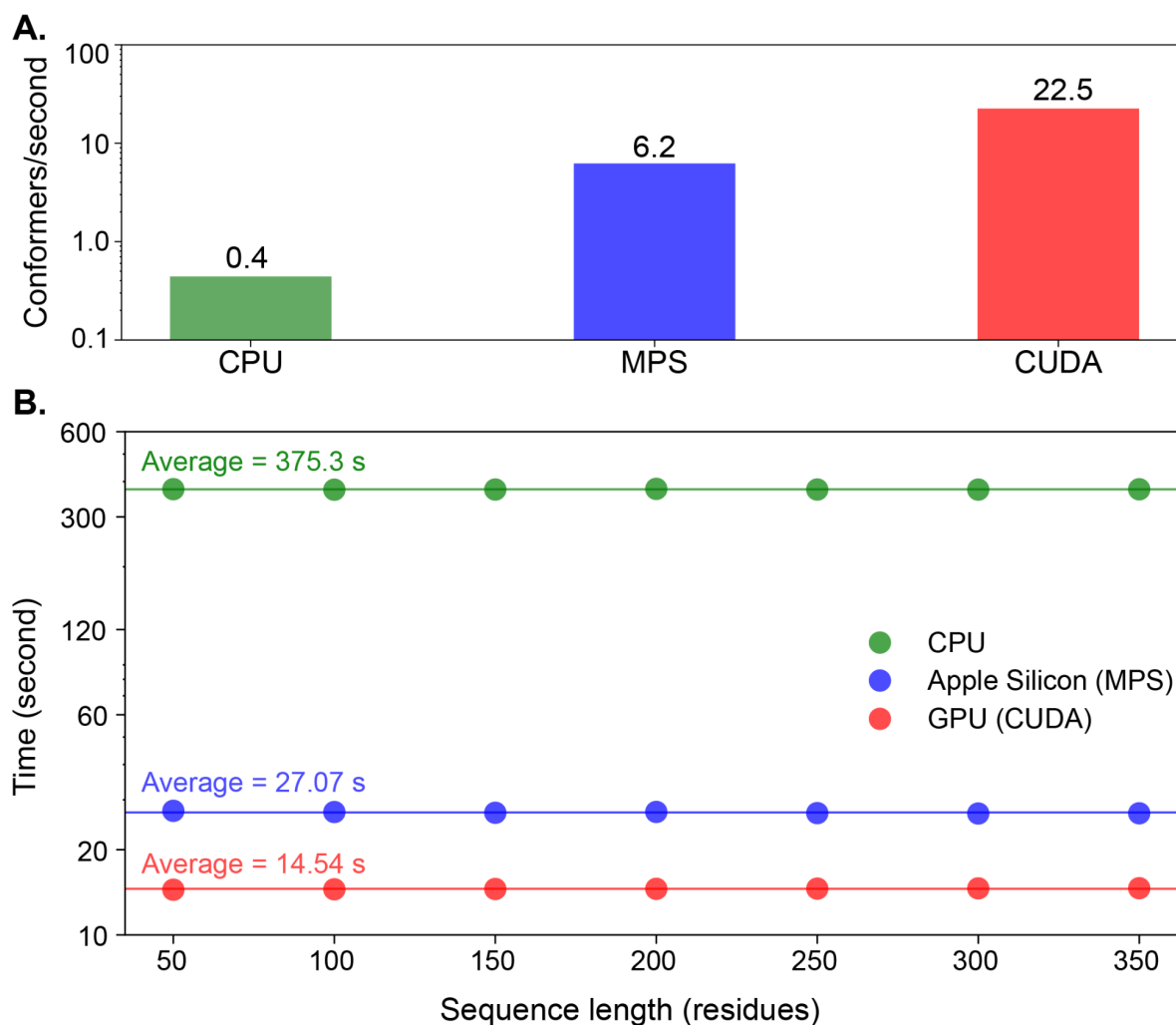


Fig. S7. (A) Number of conformers generated per second on different hardware types. Specific hardware tested here are Nvidia A4500 (GPU), Macbook Pro M3 Max CPU (MPS), and Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz (CPU). **(B)** STARLING ensemble predictions show neither runtime nor memory scaling based on sequence length. Note here the GPU used was an NVIDIA RTX A4000, hence the slight drop in performance compared to panel (A).

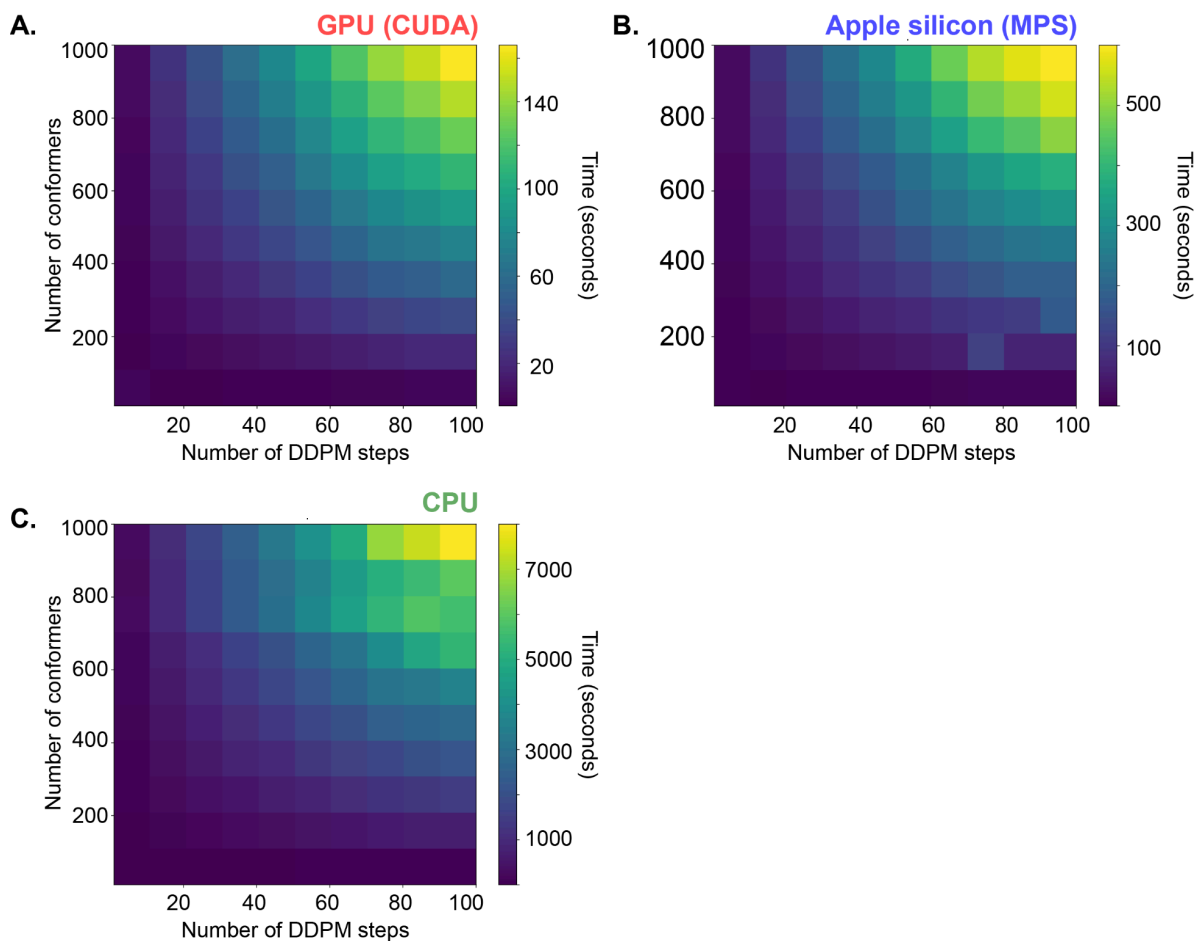


Fig. S8. Sweep of runtime vs. ensemble size and number of denoising steps on different hardware types. Specific hardware tested here are Nvidia A4500 (GPU), Macbook Pro M3 Max CPU (MPS), and Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz (CPU).

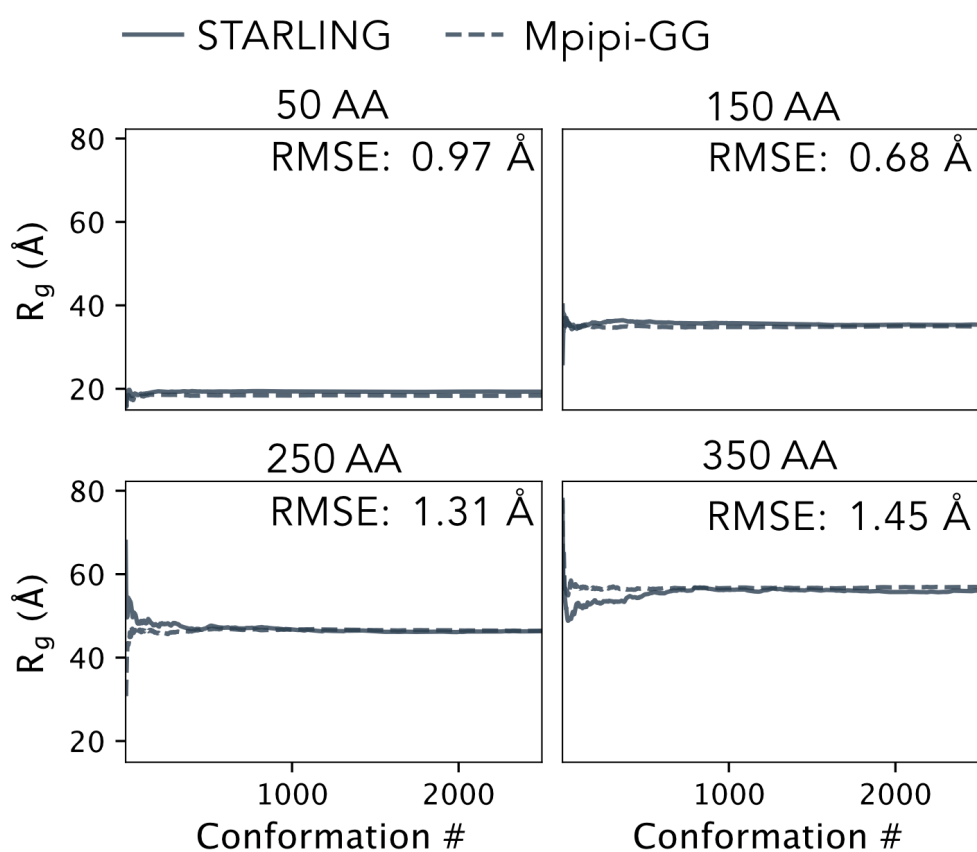


Fig. S9. Radius of gyration (R_g) convergence plot as a function of the size of the ensemble (number of conformations) for varied IDR lengths (50, 150, 250, and 350 residues in length). The dashed line is the ensemble-average R_g value taken from long Mpipi-GG simulations. STARLING-derived R_g values converge within the first 1000

conformers, exhibiting minimal error compared to Mpipi-GG ensembles. Based on this analysis, we selected 1000 STARLING conformers for R_g prediction comparisons in all benchmarking used here.

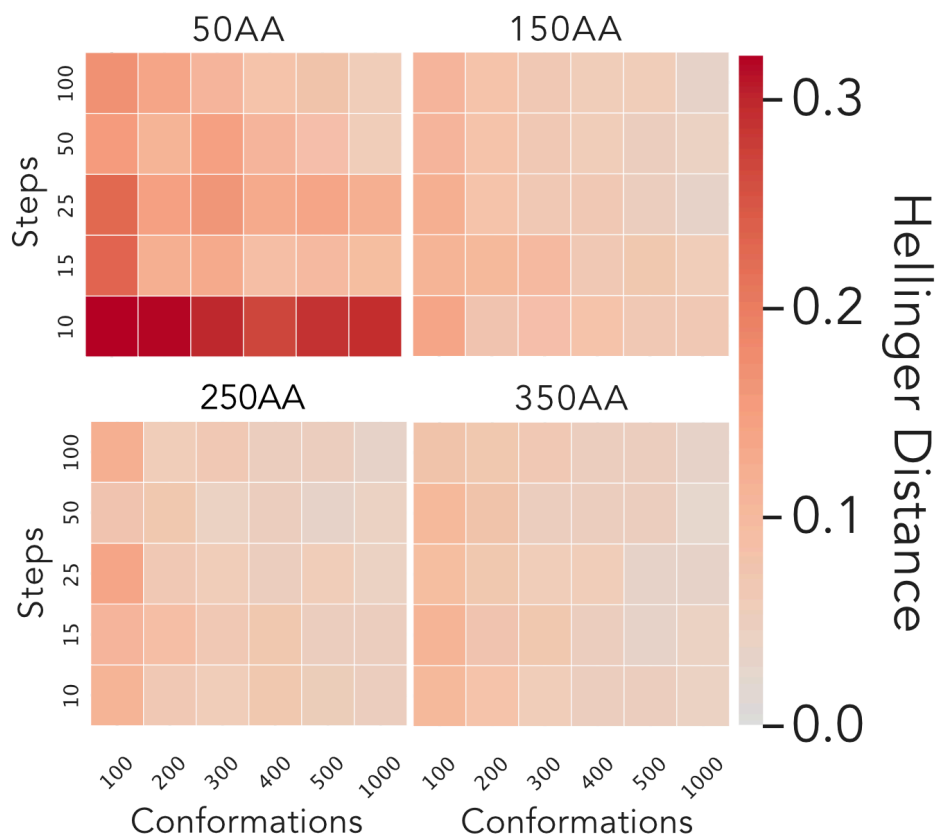


Fig. S10. Convergence plot depicting the relationship between Hellinger distance, the number of conformations sampled from STARLING, and the number of denoising steps for various IDR lengths (50, 150, 250, and 350 residues). The Hellinger distance quantifies the similarity between two distributions, where a value of zero indicates identical distributions and a value of one signifies no overlap (**Fig. 2F**). The Hellinger distance decreases as the number of conformations and denoising steps increase, with optimal STARLING performance observed at 1000 conformers and 25 denoising steps.

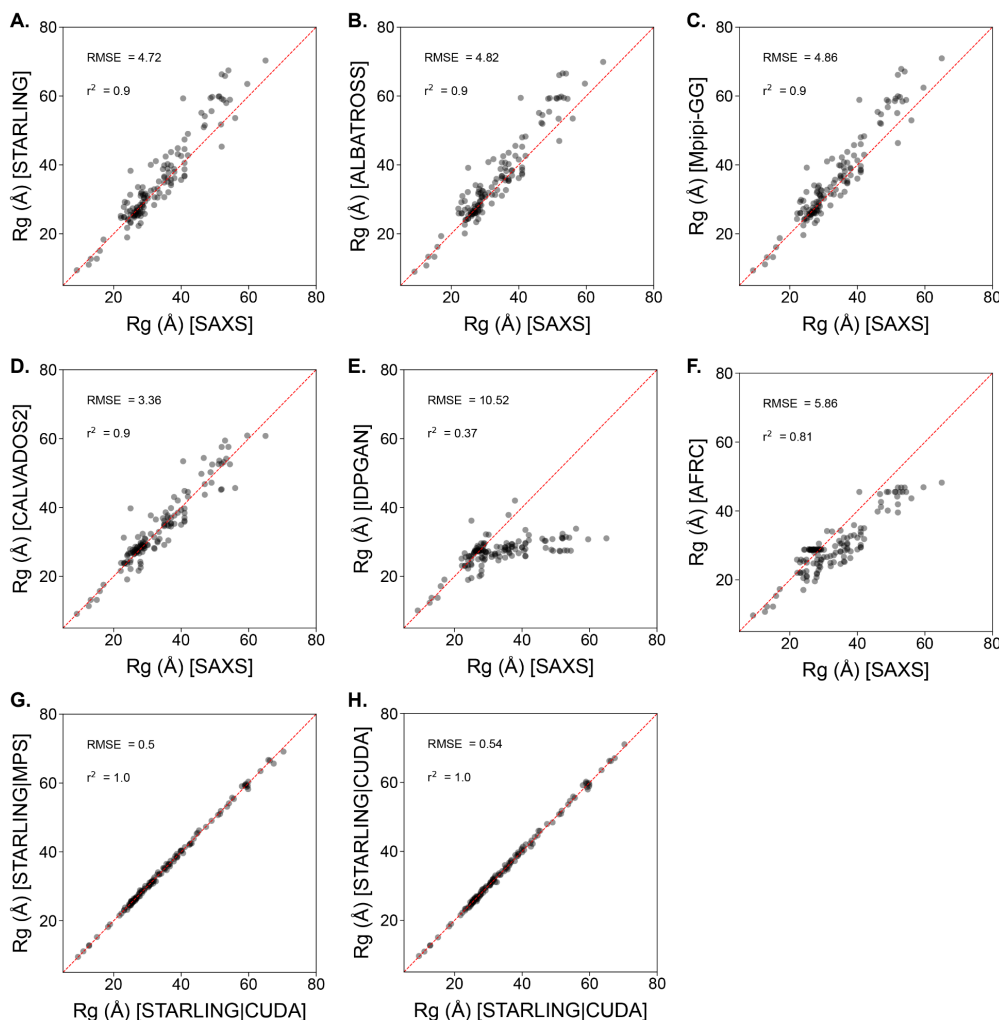


Fig. S11. Comparison of experimental R_g values obtained from 133 different experimental datasets with various state-of-the-art tools for ensemble prediction: **(A)** STARLING (this work), **(B)** ALBATROSS, a deep-learning-based predictor that can only predict average ensemble values, **(C)** Mpipi-GG, coarse-grained molecular dynamics simulations, **(D)** CALVADOS2, coarse-grained molecular dynamics simulations, **(E)** idpGAN, a Generative Adversarial Network for predicting ensembles from the sequence, **(F)** The Analytical Random Flory Coil [AFRC], a limiting null model that considers IDPs as Gaussian-like chains with no interactions between residues. We include this as this is the true “lower bound” for what a model should achieve in terms of accuracy and RMSE. **(G)** We also compared R_g predictions on GPU vs. MPS implementations for completeness. This revealed 1:1 agreement that matched the RMSE and correlation obtained when the same dataset was analyzed twice on the same hardware **(H)**.

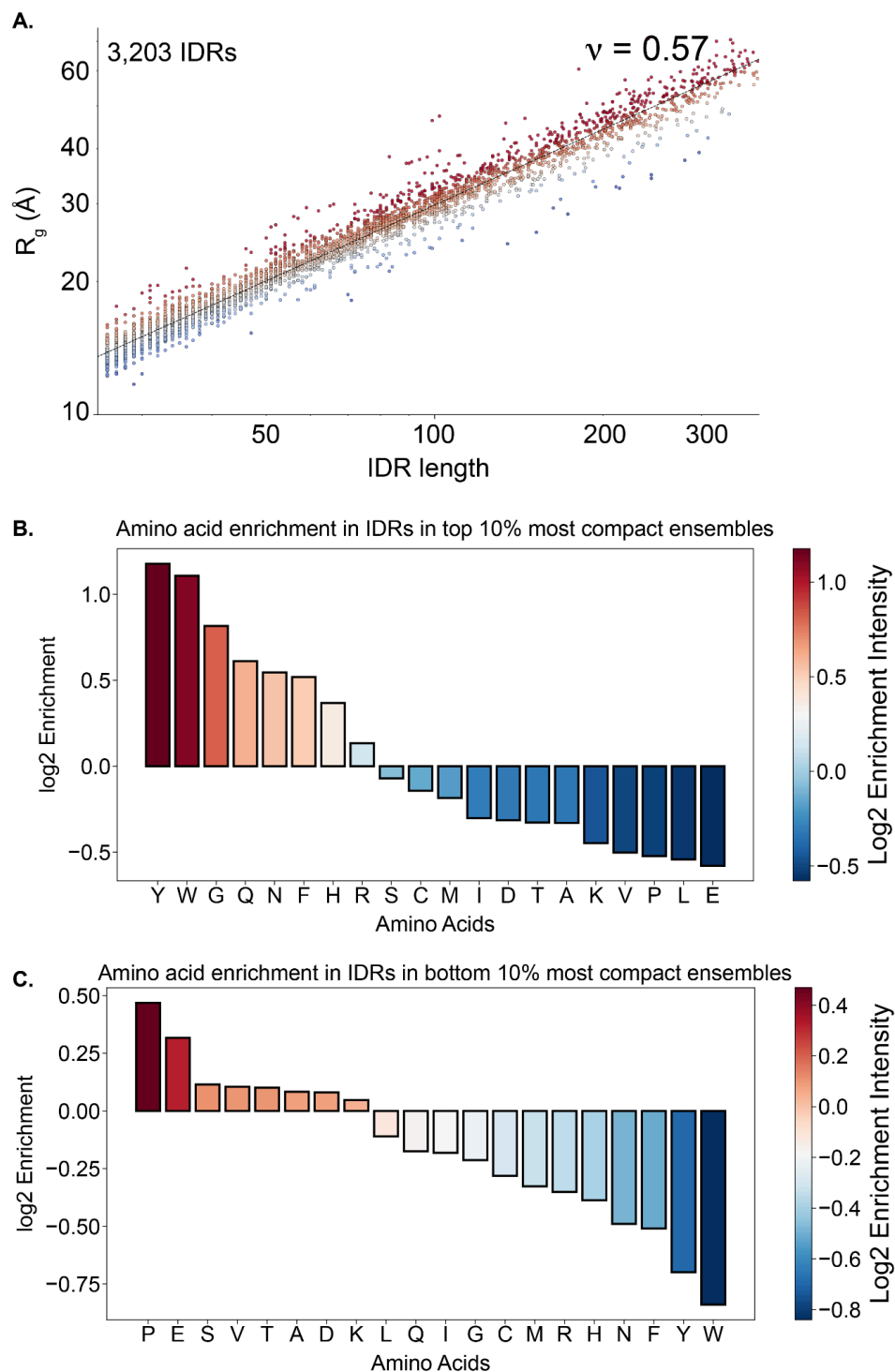


Figure S12. STARLING ensembles for DisProt. We predicted all ensembles in ~18 hours using a desktop computer with an NVIDIA A4500 GPU. **(A)** Fitting ensemble dimensions vs. length reveals an apparent scaling exponent of 0.57, indicating these IDRs are relatively expanded. However, we note an unavoidable acquisition bias in

experimentally-characterized IDRs, meaning highly soluble (and hence more expanded, fewer intra/intermolecular interactions) are preferred for biophysical characterization. We also note that we underestimate the contributions of aliphatic hydrophobes in driving intramolecular interactions and any effect from local helicity or the long-range consequences of intramolecular helix interaction. As such, we strongly caution against interpreting the value of 0.57 to mean all IDRs are highly expanded. **(B)** Analysis of sequences in the most compact 10% of IDRs identified aromatic residues as being over-represented here, in line with prior work ¹⁻³. **(C)** An analogous analysis for sequences in the top 10% of expanded IDRs also reveals negatively charged residues and proline as being enriched in sequences that drive expansion, again, in agreement with prior work^{1,4,5}.

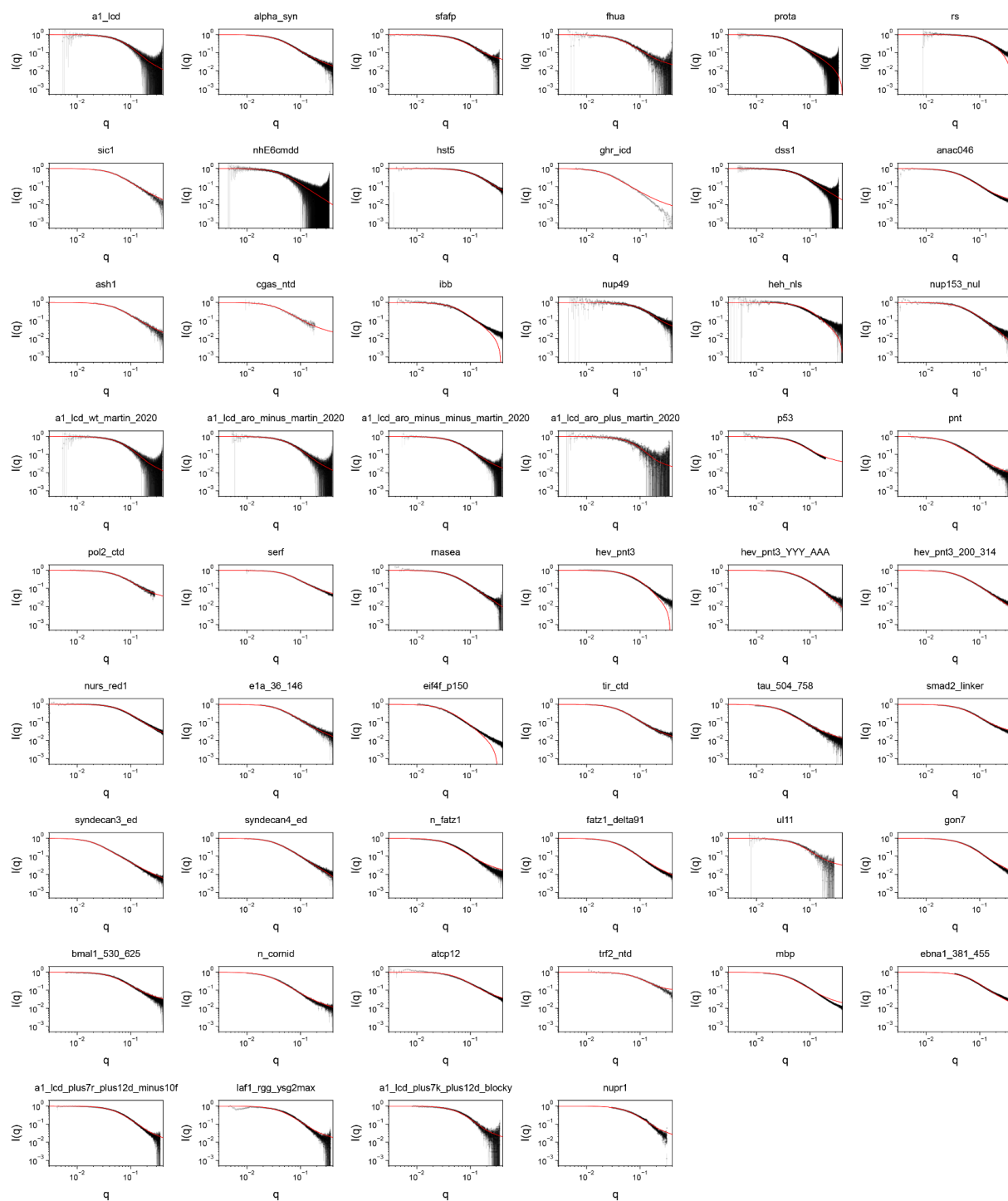


Figure S13. Comparison of STARLING-derived scattering data with experimental profiles from 52 sequences and datasets. Black is experimental and red in STARLING-derived scattering data. STARLING-derived curves were back-calculated

using FOXS. We emphasize that this agreement is not a direct fit to experimental data but rather a prediction.

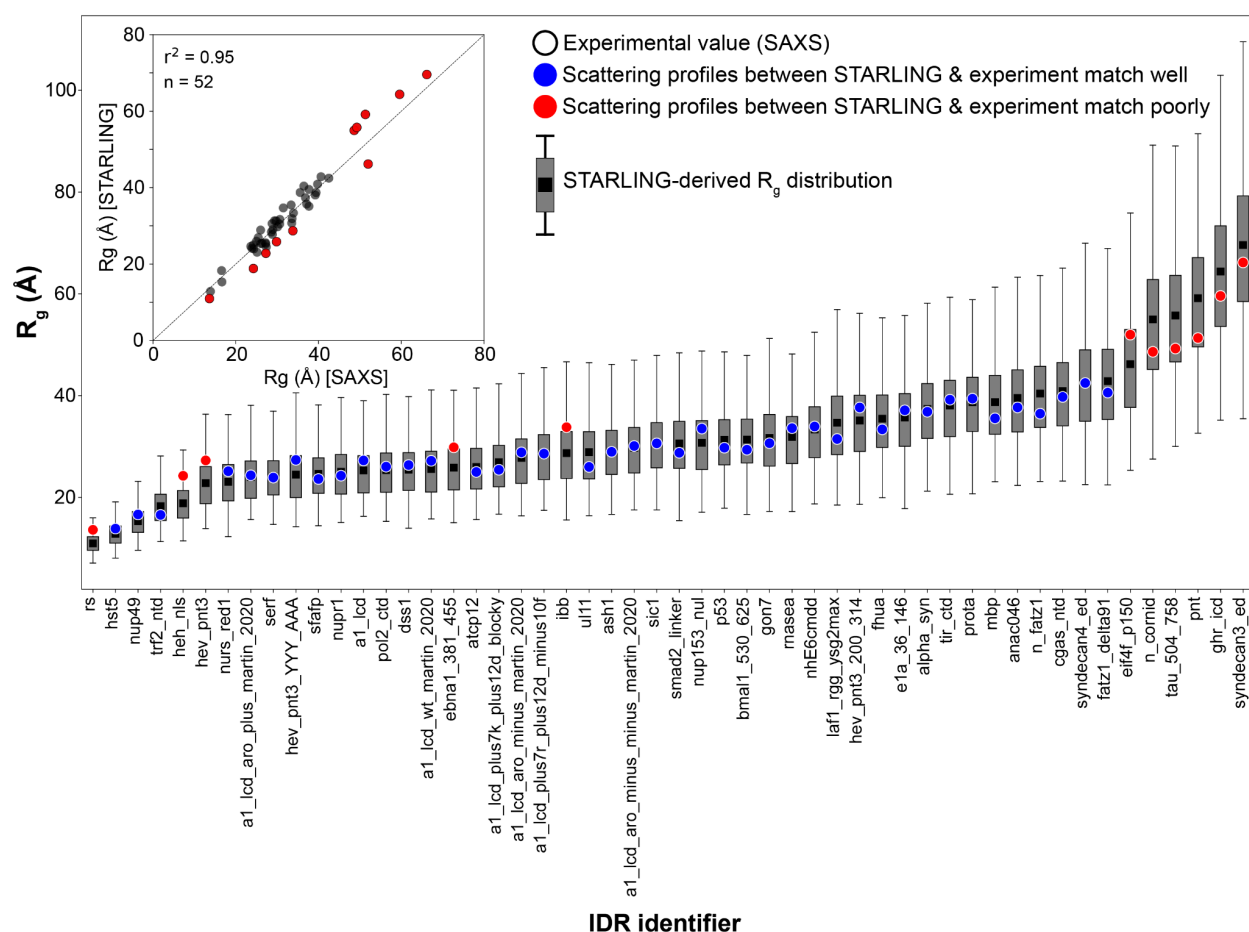


Figure S14. Comparison of SAXS-derived radii of gyration vs. STARLING-derived radii of gyration for the same 52 sequences as described in **Fig. S13**. Sequences in which experimental data is represented as a red dot are those where scattering profiles were in sub-optimal agreement with STARLING-derived scattering data—overall, focussing on this high-quality dataset (each of which we re-analyzed to obtain R_g values using the Molecular Form Factor (MFF) of Riback et al.⁶ we find even stronger agreement with STARLING-derived predictions (inset).

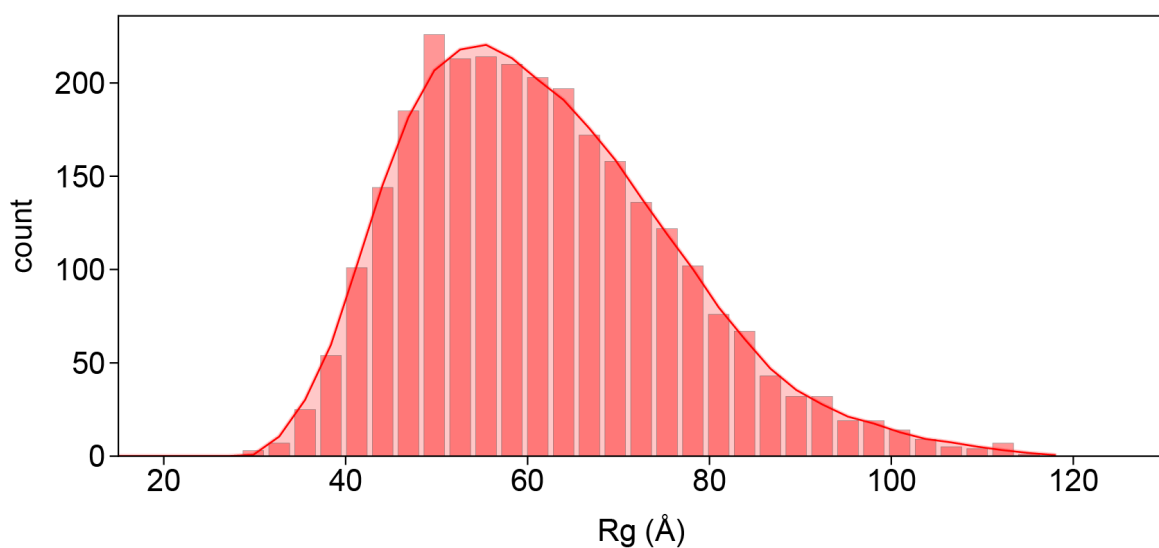


Fig. S16 R_g distribution for cMyc¹⁻³⁶¹ obtained from STARLING ensembles. We predict an average R_g of ~ 60 Å. We note this value may be an over-estimate given cMyc¹⁻³⁶¹ is predicted to contain several regions of transient secondary structure, which may reduce the R_g , such that we suggest a value closer to 55 Å may be a more accurate reflection on the expected dimensions for monomeric cMyc¹⁻³⁶¹.

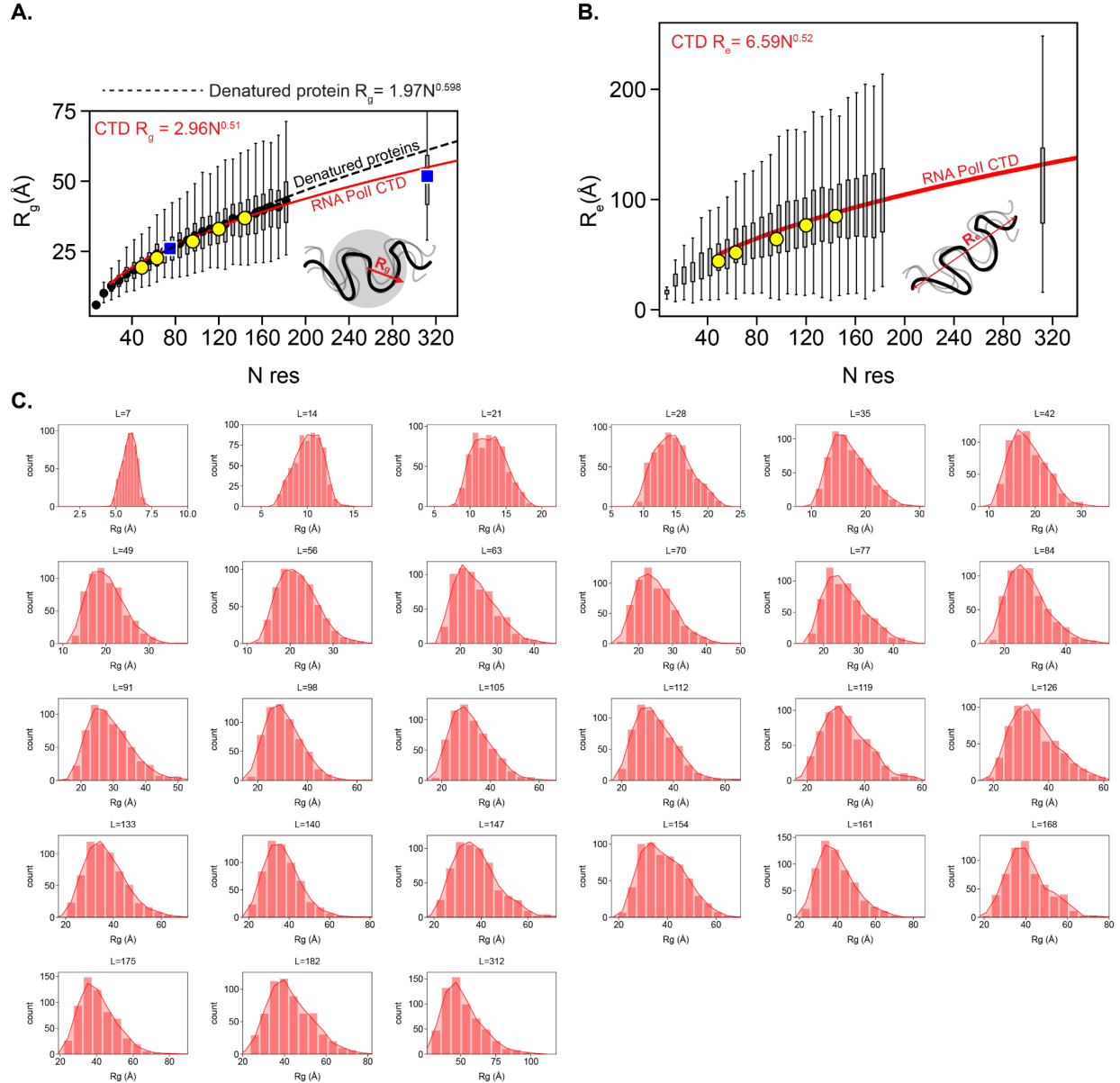


Fig. S17 Conformational behavior of RNA polymerase II CTD was examined using STARLING ensembles **(A)**. The same figure is shown for **Fig. 4D**, with an additional polymer scaling line here derived from fully denatured proteins measured by Kohn et al., also superimposed (black dashed line)⁸. This is shown because SAXS data on the 83-residue construct of the CTD measured by Gibbs *et al.* revealed a radius of gyration comparable to a fully denatured protein of the same number of amino acids (~28 Å in both cases)⁹. This observation is perhaps unexpected, given that fully denatured proteins are considered to engage in minimal intra (or inter) molecular interactions due to the “good solvent” provided by high denaturant concentrations. Yet, paradoxically, much work has shown that the RNA Pol2 CTD can phase separate and engage in

heterogeneous non-random intramolecular contacts^{10–13}. These results appear to resolve this apparent discrepancy; the larger dimensions of the 83-residue fragment originate primarily through the larger prefactor value in the scaling relationship ($B_0 = 2.96$ for CTD vs. 1.97 for denatured proteins), presumably driven by the stiffness of the proline residues and – perhaps to a lesser extent – the excluded volume of the tyrosine residue. As such, the CTD does indeed behave in a manner that involves extensive intramolecular interactions, but the overall global dimensions are dominated by the large prefactor (a convolution of chain persistence length and monomer excluded volume). We suggest the presence of proline residues – beyond serving as CKD recognition sites for phosphorylation – serves a biophysical role, enabling such a tyrosine-rich sequence to remain highly expanded and be primed for intermolecular interactions with a cost of cofactors during search, initiation, and elongation. **(B)** The analogous figure is shown in Fig. 4D, but the end-to-end distance is shown instead of the radius of gyration. **(C)** Histograms for R_g values across all CTD lengths.

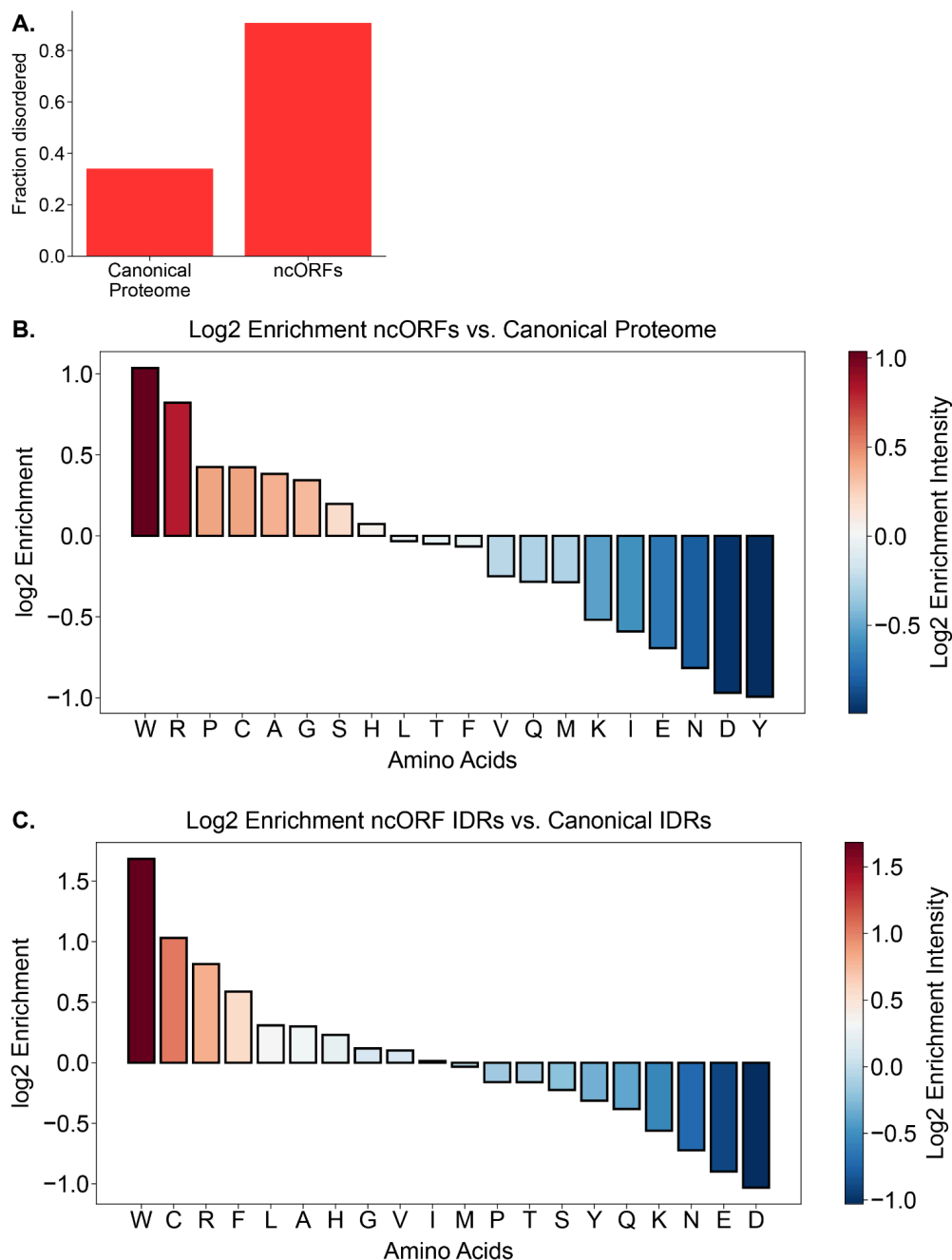


Fig. S18. (A) Microproteins are predicted to be highly disordered. The comparison here is the fraction of amino acids in the canonical human proteome found in disordered regions vs. the fraction of amino acids in the Deutsch *et al.* micro-proteome found in disordered regions¹⁴. The disorder was predicted using metapredict V3¹⁵. **(B)** Statistical enrichment for each of the 20 amino acids comparing frequency observed across all canonical proteins vs. frequency observed in microproteins. We excluded initiator methionine for both datasets, given the proportional impact of such a bias disproportionately influences shorter vs. longer sequences. As observed previously for

metazoan microproteins, we find enrichment for tryptophan and arginine and depletion acid residues^{16,17}. **(C)** To establish whether the enrichment observed here reflects something specific to microproteins or the enhanced fraction of disorder, we also compared enrichment in microprotein-derived IDRs vs. canonical-proteome-derived IDRs. This revealed an even stronger signature: microprotein IDRs are *highly* enriched for tryptophan, cysteine, arginine, and phenylalanine. In contrast, they are highly depleted for lysine, asparagine, glutamic acid, and aspartic acid.

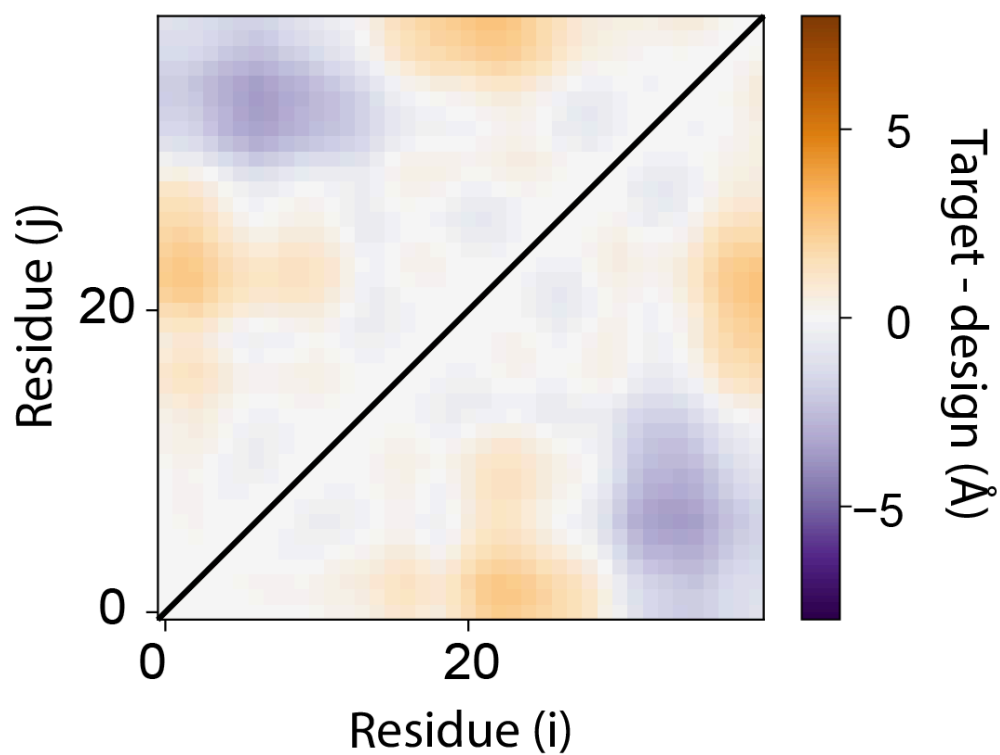


Fig. S19 Symmetric inter-residue distance matrix showing target designed sequence subtracted from the target sequence, yielding a difference map. All inter-residue errors are < 4 Å, with an RMSE error of 1.3 Å.

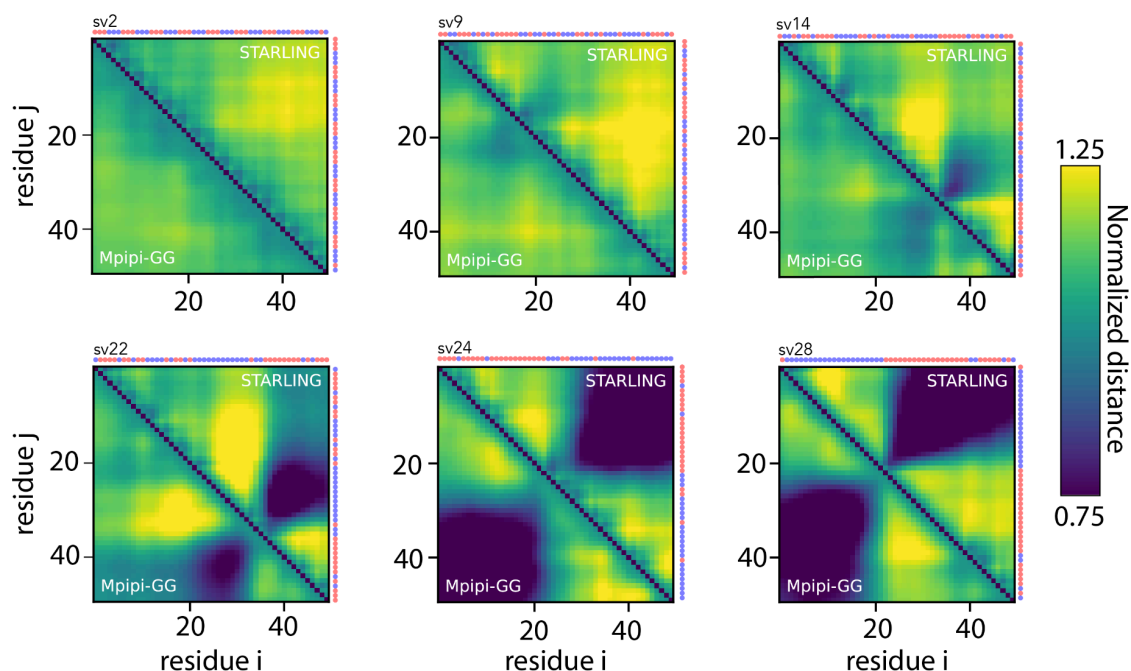


Fig. S20 Das-Kappa sequences in STARLING vs. Mpipi-GG. We computed normalized distance maps for several of the iconic “Das sequence”²⁰, revealing that STARLING ensembles offer good reproduction of the inter-residue distances.

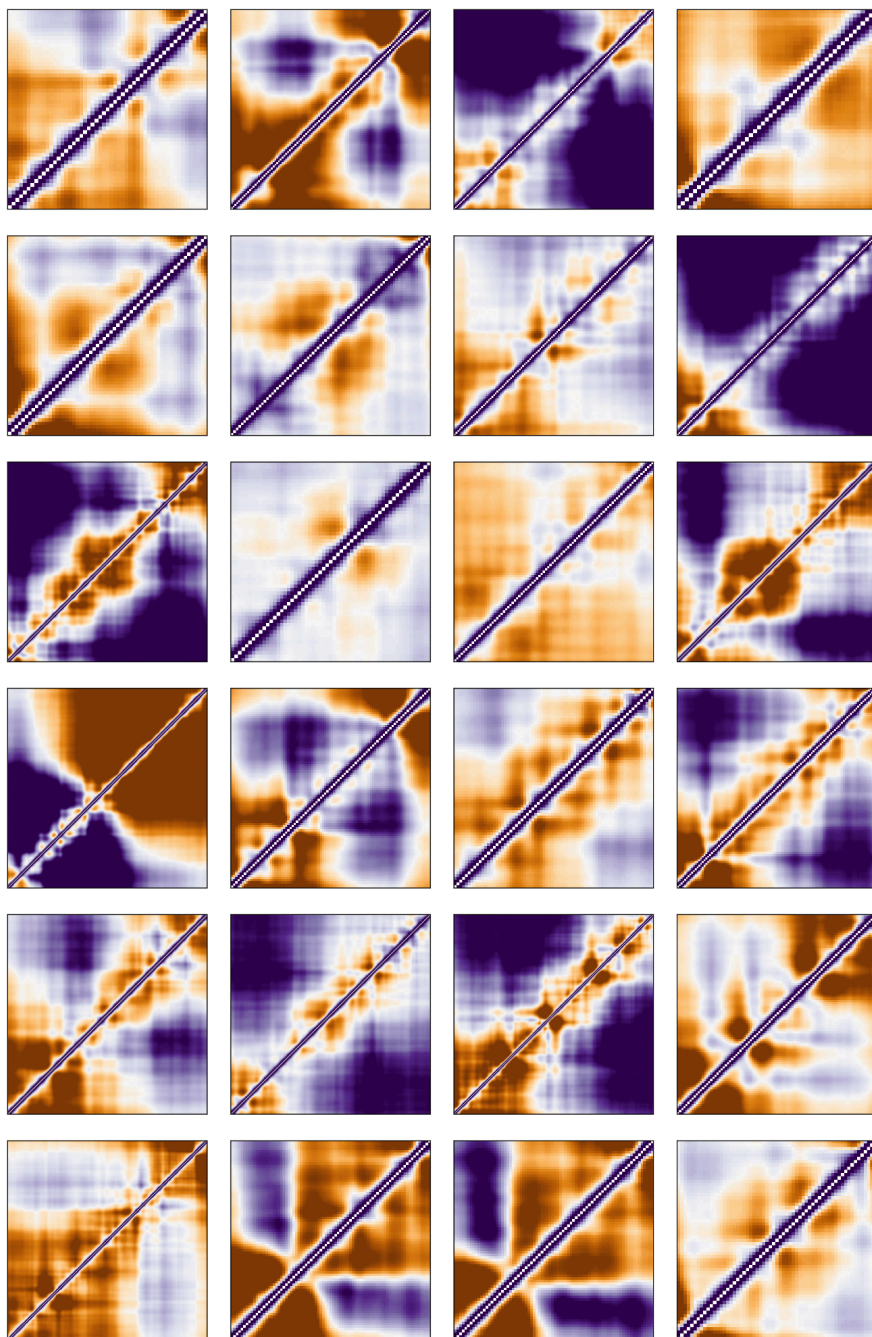


Fig. S21 Example of distance maps from DisProt

We examined normalized distance maps from a set of more compact ensembles identified from our DisProt sequences, revealing a wealth of sequence-encoded complexity in the underlying conformational behavior.

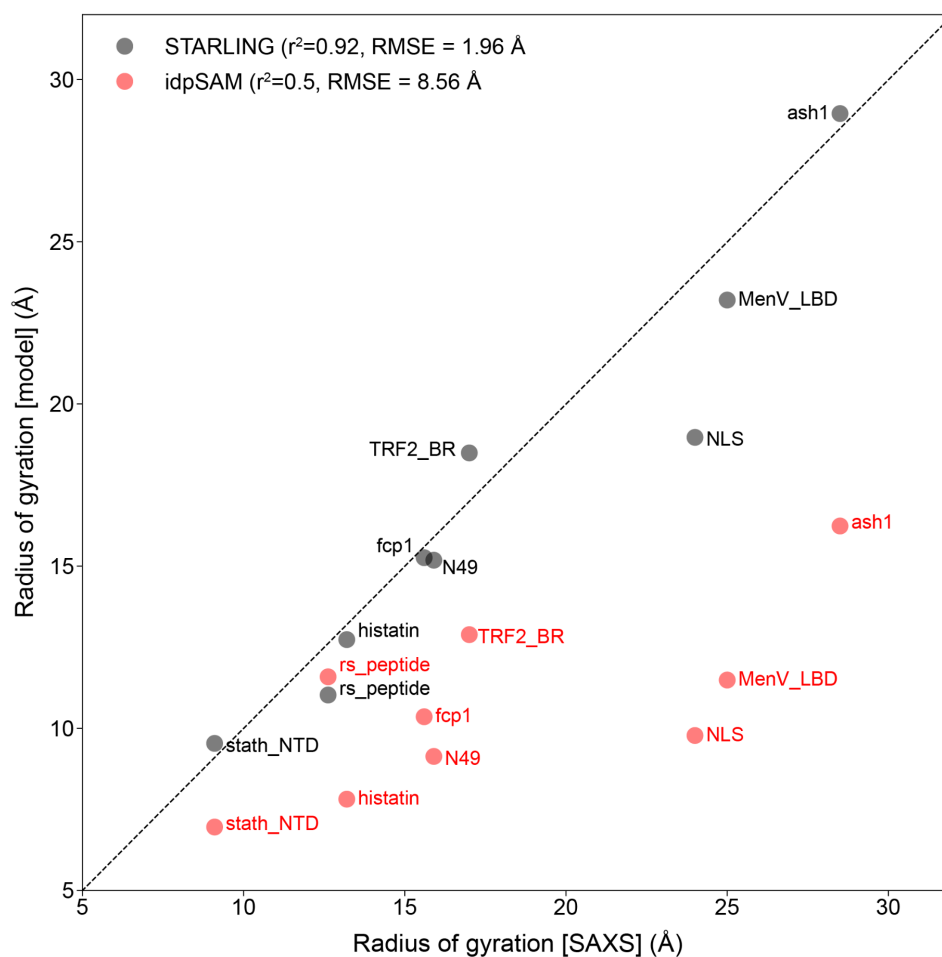


Fig. S22. Comparison of idpSAM-derived ensembles vs. STARLING compared to experimental data. Focussing on sequences less than 60 amino acids in length (with the exception of Ash1, discussed below), we compared R_g values obtained from idpSAM ensembles vs. R_g values from STARLING ensembles. In both cases, 1000 conformers were used (deferring to the idpSAM defaults). idpSAM offers limited predictive power. We note that the Analytical Flory Random Coil (AFRC) – a null model in which IDR sequence chemistry does not influence the overall dimensions and instead IDR dimensions depend solely on chain length and (weakly) the intrinsic dimensions of each amino acid – obtains an r^2 of 0.9 and an RMSE of 3.58 Å¹⁸. We included Ash1 (83 amino acids) because while it is larger than the ~60 amino acid cutoff used for training data in idpSAM, prior work has shown SAXS data for Ash1 is well reproduced by CAMPARI/ABSINTH simulations¹⁹. However, despite being trained on CAMPARI/ABSINTH simulations, idpSAM was unfortunately unable to recapitulate this behavior.

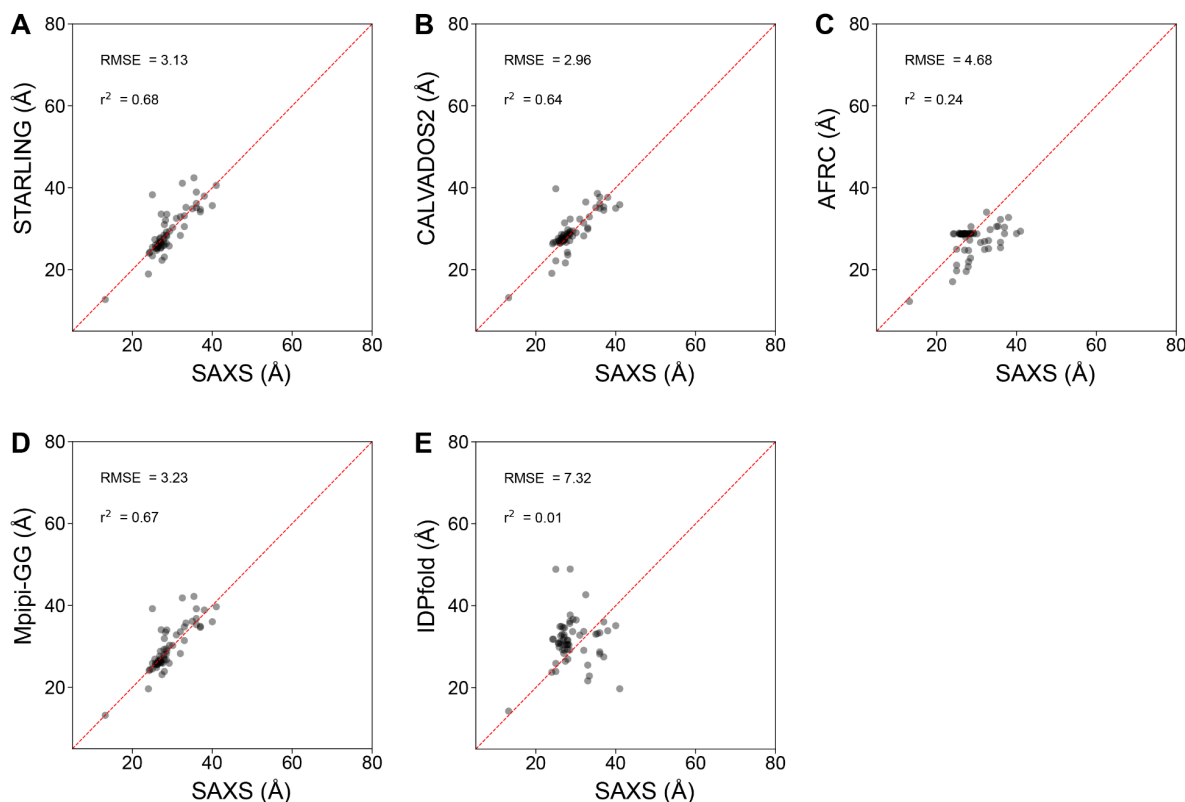


Fig. S23. Comparison of IDPFold-derived radii of gyration vs. other approaches compared to radii of gyration derived from experimental SAXS data. (A) Radii of gyration for STARLING vs. SAXS for 59 sequences where IDPFold ensembles were available at the time of submission. **(B)** Radii of gyration for CALVADOS2 simulations vs. SAXS for 59 sequences where IDPFold ensembles were available at the time of submission. **(C)** Radii of gyration for AFRC vs. SAXS for 59 sequences where IDPFold ensembles were available at the time of submission. **(D)** Radii of gyration for Mpipi-GG simulations vs. SAXS for 59 sequences where IDPFold ensembles were available at the time of submission. **(E)** Radii of gyration for IDPFold vs. SAXS for 59 sequences where IDPFold ensembles were available at the time of submission.

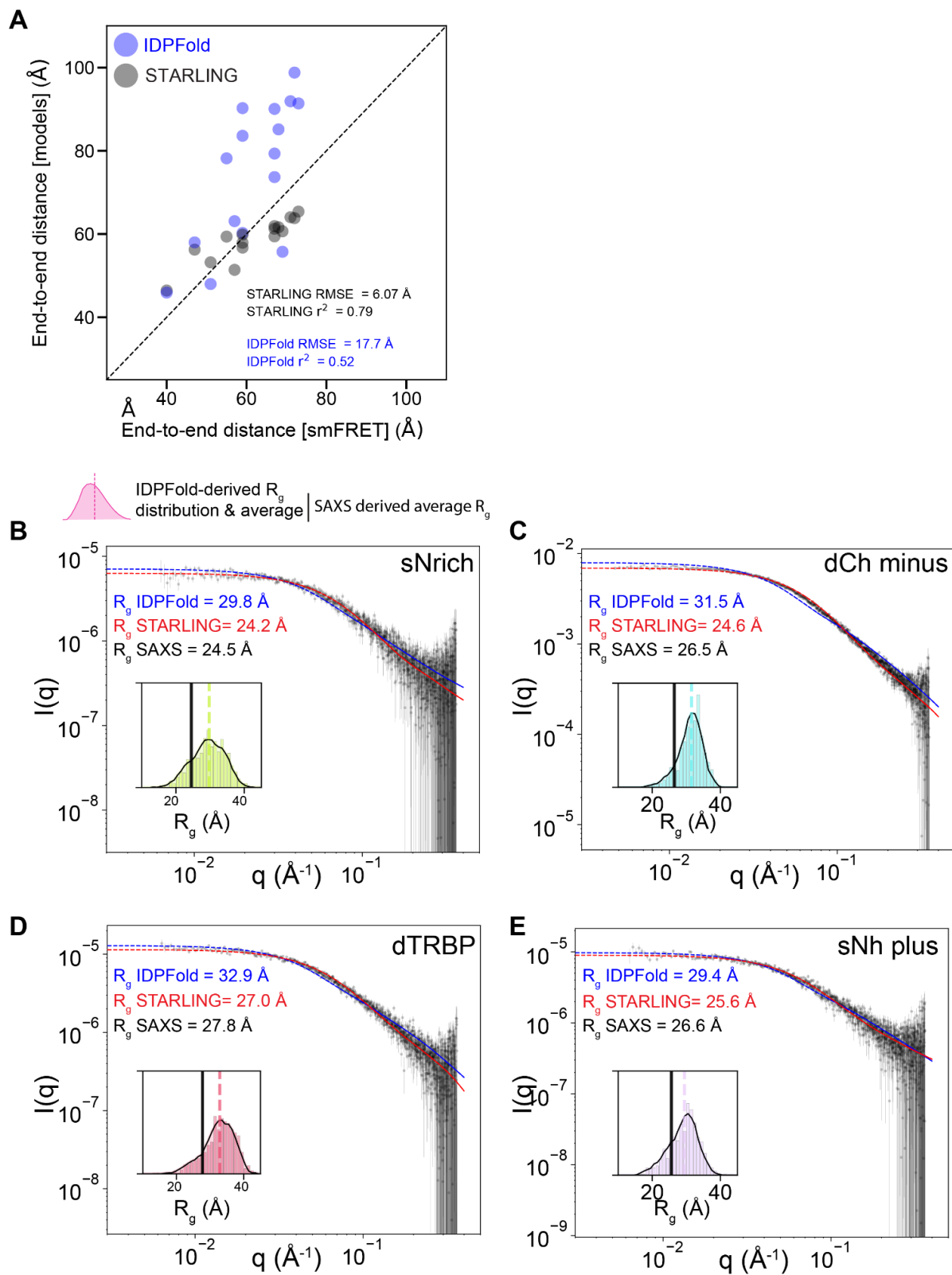


Fig. S24. Comparison of IDPFold-derived ensembles vs. STARLING compared to single-molecule FRET data. (A) Correlation between experimentally-derived

end-to-end distance (x-axis) and STARLING (black) or IDPFold (Blue) derived end-to-end distances. **(B)** Comparison of SAXS scattering profile for the sNrich ensemble derived from IDPFold ensembles (red) vs. experiment (black). Inferred radii of gyration from IDPFold, STARLING, and SAXS are provided. **(C)** Comparison of SAXS scattering profile for the dCh minus ensemble derived from IDPFold ensembles (red) vs. experiment (black). Inferred radii of gyration from IDPFold, STARLING, and SAXS are provided. **(D)** Comparison of SAXS scattering profile for the dTRBP ensemble derived from IDPFold ensembles (red) vs. experiment (black). Inferred radii of gyration from IDPFold, STARLING, and SAXS are provided. **(E)** Comparison of SAXS scattering profile for the sNh plus ensemble derived from IDPFold ensembles (red) vs. experiment (black). Inferred radii of gyration from IDPFold, STARLING, and SAXS are provided.

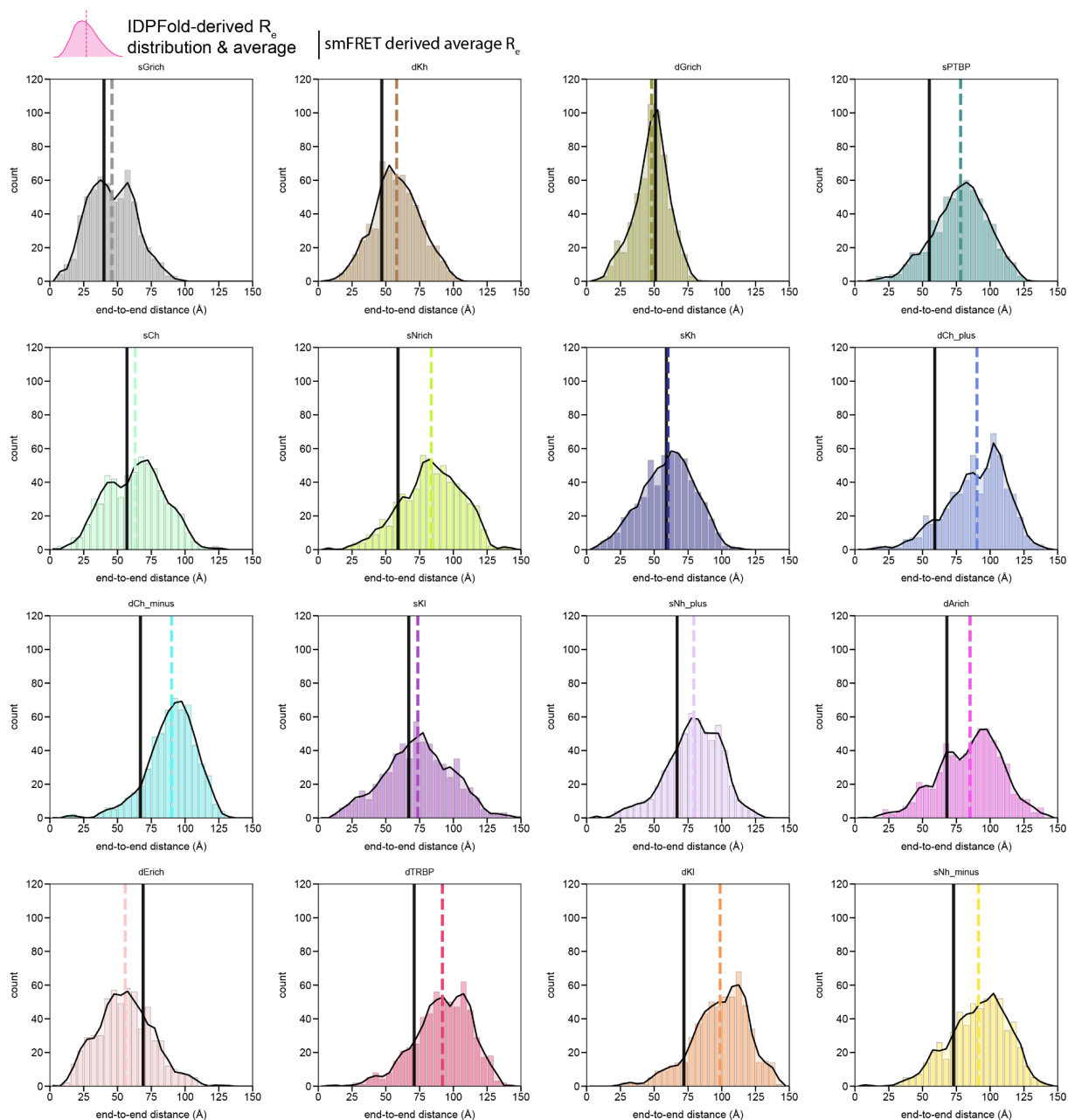


Fig. S25. Comparison of IDPFold-derived ensembles vs. STARLING compared to smFRET data. Distributions from IDPFold ensembles (colored histograms) with average values (colored dashed lines) for end-to-end distance vs. experimentally derived end-to-end distance (black line).

SUPPLEMENTARY REFERENCES

1. 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* **21**, 465–476 (2024).
2. Du, Z., Wang, H., Luo, S., Yun, Z., Wu, C., Yang, W., Buck, M., Zheng, W., Hansen, A. L., Kao, H.-Y. & Yang, S. The sequence–structure–function relationship of intrinsic ER α disorder. *Nature* 1–9 (2025).
3. 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* **367**, 694–699 (2020).
4. Marsh, J. A. & Forman-Kay, J. D. Sequence Determinants of Compaction in Intrinsically Disordered Proteins. *Biophys. J.* **98**, 2383–2390 (2010).
5. Müller-Späth, S., Soranno, A., Hirschfeld, V., Hofmann, H., Rügger, S., Reymond, L., Nettels, D. & Schuler, B. Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14609–14614 (2010).
6. 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* **358**, 238–241 (2017).
7. Holla, A., Martin, E. W., Dannenhoffer-Lafage, T., Ruff, K. M., König, S. L. B., Nüesch, M. F., Chowdhury, A., Louis, J. M., Soranno, A., Nettels, D., Pappu, R. V.,

- Best, R. B., Mittag, T. & Schuler, B. Identifying sequence effects on chain dimensions of disordered proteins by integrating experiments and simulations. *JACS Au* (2024). doi:10.1021/jacsau.4c00673
8. Kohn, J. E., Millett, I. S., Jacob, J., Zagrovic, B., Dillon, T. M., Cingel, N., Dothager, R. S., Seifert, S., Thiyagarajan, P., Sosnick, T. R., Hasan, M. Z., Pande, V. S., Ruczinski, I., Doniach, S. & Plaxco, K. W. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12491–12496 (2004).
 9. Gibbs, E. B., Lu, F., Portz, B., Fisher, M. J., Medellin, B. P., Laremore, T. N., Zhang, Y. J., Gilmour, D. S. & Showalter, S. A. Phosphorylation induces sequence-specific conformational switches in the RNA polymerase II C-terminal domain. *Nat. Commun.* **8**, 15233 (2017).
 10. Flores-Solis, D., Lushpinskaia, I. P., Polyansky, A. A., Changiarath, A., Boehning, M., Mirkovic, M., Walshe, J., Pietrek, L. M., Cramer, P., Stelzl, L. S., Zagrovic, B. & Zweckstetter, M. Driving forces behind phase separation of the carboxy-terminal domain of RNA polymerase II. *Nat. Commun.* **14**, 5979 (2023).
 11. Lu, H., Yu, D., Hansen, A. S., Ganguly, S., Liu, R., Heckert, A., Darzacq, X. & Zhou, Q. Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature* **558**, 318–323 (2018).
 12. Portz, B., Lu, F., Gibbs, E. B., Mayfield, J. E., Rachel Mehaffey, M., Zhang, Y. J., Brodbelt, J. S., Showalter, S. A. & Gilmour, D. S. Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. *Nat. Commun.* **8**, 15231 (2017).

13. Lu, F., Portz, B. & Gilmour, D. S. The C-Terminal Domain of RNA Polymerase II Is a Multivalent Targeting Sequence that Supports *Drosophila* Development with Only Consensus Heptads. *Mol. Cell* **73**, 1232–1242.e4 (2019).
14. Deutsch, E. W., Kok, L. W., Mudge, J. M., Ruiz-Orera, J., Fierro-Monti, I., Sun, Z., Abelin, J. G., Alba, M. M., Aspden, J. L., Bazzini, A. A., Bruford, E. A., Brunet, M. A., Calviello, L., Carr, S. A., Carvunis, A.-R., Chothani, S., Clauwaert, J., Dean, K., Faridi, P., Frankish, A., Hubner, N., Ingolia, N. T., Magrane, M., Martin, M. J., Martinez, T. F., Menschaert, G., Ohler, U., Orchard, S., Rackham, O., Roucou, X., Slavoff, S. A., Valen, E., Wacholder, A., Weissman, J. S., Wu, W., Xie, Z., Choudhary, J., Bassani-Sternberg, M., Vizcaíno, J. A., Ternet, N., Moritz, R. L., Prensner, J. R. & van Heesch, S. High-quality peptide evidence for annotating non-canonical open reading frames as human proteins. *bioRxiv* 2024.09.09.612016 (2024). doi:10.1101/2024.09.09.612016
15. Lotthammer, J. M., Hernandez-Garcia, J., Griffith, D., Weijers, D., Holehouse, A. S. & Emenecker, R. J. Metapredict enables accurate disorder prediction across the Tree of Life. *bioRxiv* 2024.11.05.622168 (2024). doi:10.1101/2024.11.05.622168
16. Martinez, T. F., Lyons-Abbott, S., Bookout, A. L., De Souza, E. V., Donaldson, C., Vaughan, J. M., Lau, C., Abramov, A., Baquero, A. F., Baquero, K., Friedrich, D., Huard, J., Davis, R., Kim, B., Koch, T., Mercer, A. J., Misquith, A., Murray, S. A., Perry, S., Pino, L. K., Sanford, C., Simon, A., Zhang, Y., Zipp, G., Bizarro, C. V., Shokhirev, M. N., Whittle, A. J., Searle, B. C., MacCoss, M. J., Saghatelian, A. & Barnes, C. A. Profiling mouse brown and white adipocytes to identify metabolically relevant small ORFs and functional microproteins. *Cell Metab.* **35**, 166–183.e11

(2023).

17. Martinez, T. F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M. N. & Saghatelian, A. Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* **16**, 458–468 (2020).
18. Alston, J. J., Ginell, G. M., Soranno, A. & Holehouse, A. S. The Analytical Flory Random Coil Is a Simple-to-Use Reference Model for Unfolded and Disordered Proteins. *J. Phys. Chem. B* **127**, 4746–4760 (2023).
19. 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.* **138**, 15323–15335 (2016).
20. 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.* **110**, 13392–13397 (2013).