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

ABSTRACT

Intrinsically disordered proteins and regions (collectively IDRs) are found across all kingdoms of life and play critical roles in virtually every eukaryotic cellular process. In contrast to folded proteins, IDRs lack a stable 3D structure and are instead described in terms of a conformational ensemble, a collection of energetically accessible interconverting structures. This unique structural plasticity facilitates diverse molecular recognition and function; thus, a convenient way to view IDRs is through their ensembles. Here, we combine advances in physics-based force fields for IDPs with the power of modern multi-scale generative modeling to develop STARLING, an approach for the rapid and accurate prediction of IDR ensembles directly from sequence. STARLING enables ensembles of hundreds of conformers to be generated in seconds and works on GPUs and CPUs. This, in turn, dramatically lowers the barrier to the computational interrogation of IDR function through the lens of emergent biophysical properties in addition to traditional bioinformatic approaches. We evaluate STARLING's accuracy against experimental data and offer a series of vignettes illustrating how STARLING can enable rapid hypothesis generation for IDR function or the interpretation of experimental data.

VERSION 1 (2025-02-14)

INTRODUCTION

Intrinsically disordered proteins and protein regions (IDRs) are structurally heterogeneous protein regions that are estimated to constitute ~30% of eukaryotic proteomes[1,2]. Despite lacking a fixed structure, IDRs play key roles in essential cellular processes such as transcription, translation, and cell signaling[1,3]. Due to their structural heterogeneity, IDRs are best described by a conformational ensemble - a dynamic set of rapidly interchanging conformations[1,4]. Just as folded domains follow a sequence-structure relationship, IDRs follow an analogous sequence-ensemble relationship, where the underlying sequence chemistry dictates the conformations that make up their ensembles[5,6]. Moreover, IDR ensembles can play essential roles in IDR function and may be perturbed in disease[7–9]. Consequently, just as structural biology has been instrumental in understanding the molecular basis for folded domain function, there is a growing appreciation that the characterization of IDR ensembles is important for understanding IDR function.

Experimental techniques commonly used to interrogate sequence-ensemble relationships include single-molecule Förster Resonance Energy Transfer (smFRET) spectroscopy[10], Nuclear Magnetic Resonance (NMR) spectroscopy[11], and Small-Angle X-ray Scattering (SAXS)[12]. While these report on specific aspects of IDR ensembles, they fall short of providing a holistic description of the distribution of conformers (referred to here as a 'full structural ensemble'). Integrating all-atom simulations with experimental data has proven one effective route to construct full structural ensembles[9,13–17]. However, these simulations require deep technical expertise to ensure reliable conclusions are drawn and are computationally expensive, often requiring significant time to achieve high quality data. While recent advances in coarse-grained (CG) simulations[18,19] offer a faster alternative, even CG simulations can still take hours to obtain sufficient sampling, making them inappropriate for large-scale investigation into sequence-ensemble relationships. Moreover, CG simulations still require a relatively high level of technical expertise to setup, run, and analyze. Recent deep learning predictors trained on CG simulations have enabled proteome-scale predictions for average values for some subset of observables (e.g. end-to-end distance) but are limited by which observables are predictable[20,21].

Deep learning approaches such as AlphaFold[22,23], RosettaFold[24,25], ESMFold[26,27], and others[28] have transformed the field of protein structure predictions, significantly reducing the barrier to the large-scale exploration of sequence-structure relationships (**Fig. 1A**). However, these methods are poorly suited to investigate IDRs[29] owing to a reduction in alignment-based conservation in disordered proteins, a paucity of appropriate experimental training data, and their training objective (predicting a single best structure from sequence, whereas IDRs should be described by large ensembles of structurally heterogeneous conformers) (**Fig. 1C)**. In short, while we now possess easy-to-use tools for accurately predicting the 3D structure of individual folded domains, equivalent tools for predicting full structural ensembles of IDRs are lacking.

Here, we address these challenges by developing a fast and accurate approach for predicting coarse-grained full structural ensembles from protein sequences. Our approach leverages advances in generative modeling, a deep learning technique capable of creating new and original data. However, developing a generative model poses a key challenge: the need for large training datasets. To address this, we performed large-scale coarse-grained simulations to generate full structural ensembles across tens of thousands of natural and synthetic IDRs. The resulting predictor, **conSTruction of intrinsicAlly disoRdered proteins ensembles efficientLy vIa multi-dimeNsional Generative models** (STARLING) furthers our descent into acronym madness but also allows us to generate full structural ensembles directly from sequence in seconds. A major goal in developing STARLING was to avoid hardware barriers. While STARLING is fast (~25 conformers/second) on GPUs, it can still generate ensembles in minutes on Intel/AMD CPUs and seconds on Apple CPUs.

STARLING-generated ensembles show good agreement with experimental data, enabling *de novo* exploration of uncharacterized IDRs or aid in the biophysical interpretation of experimental data. Moreover, here we show proof-of-concept that STARLING can be used to (i) investigate sequence-ensemble relationships for disordered proteins, (ii) explore bound-state conformational ensembles for binary disordered protein complexes, (iii) facilitate rational design of disordered proteins with specific ensembles. Taken together, we propose that STARLING's ease of use and speed make it a powerful tool for democratizing large-scale exploration of sequence-ensemble relationships in IDRs.

METHODS & RESULTS

Generative artificial intelligence (AI) has been transformative for text-to-image generation[30–35]. In text-to-image generative AI, an image is generated by passing a prompt (a short phrase describing the desired image) to a pre-trained deep-learning model. That model then generates an image consistent with the prompt, a process referred to as inference. Deep learning models capable of inference must first be trained. For models used in modern text-to-image generative AI, the training is not simply memorizing the images seen in training but learning the relationship between prompts and the features within the associated images. As a result, once a model is trained, if the same prompt is reused many times, it will generate many independent images entirely distinct from any of the individual instances observed in training. Notably, despite being different from one another, every image generated should be consistent with the prompt (**Fig. 1C**).

The mapping of a single text prompt to a collection of different images – each of which is consistent with the input prompt – is precisely the problem we wish to solve for IDR ensemble generation. In IDR ensemble generation, we want to take a text prompt (amino acid sequence) and generate a collection of many distinct and uncorrelated IDR conformations consistent with that prompt (**Fig. 1D**). Moreover, we want this generation process to be fast (seconds) and

possible on commodity hardware (laptops and desktops). To achieve this goal, we combine a variational autoencoder (VAE) with a discrete-time denoising-diffusion probabilistic model (DDPM) to create a latent diffusion model[31]. The resulting method (STARLING) enables the accurate and rapid prediction of coarse-grained conformational ensembles of IDRs.

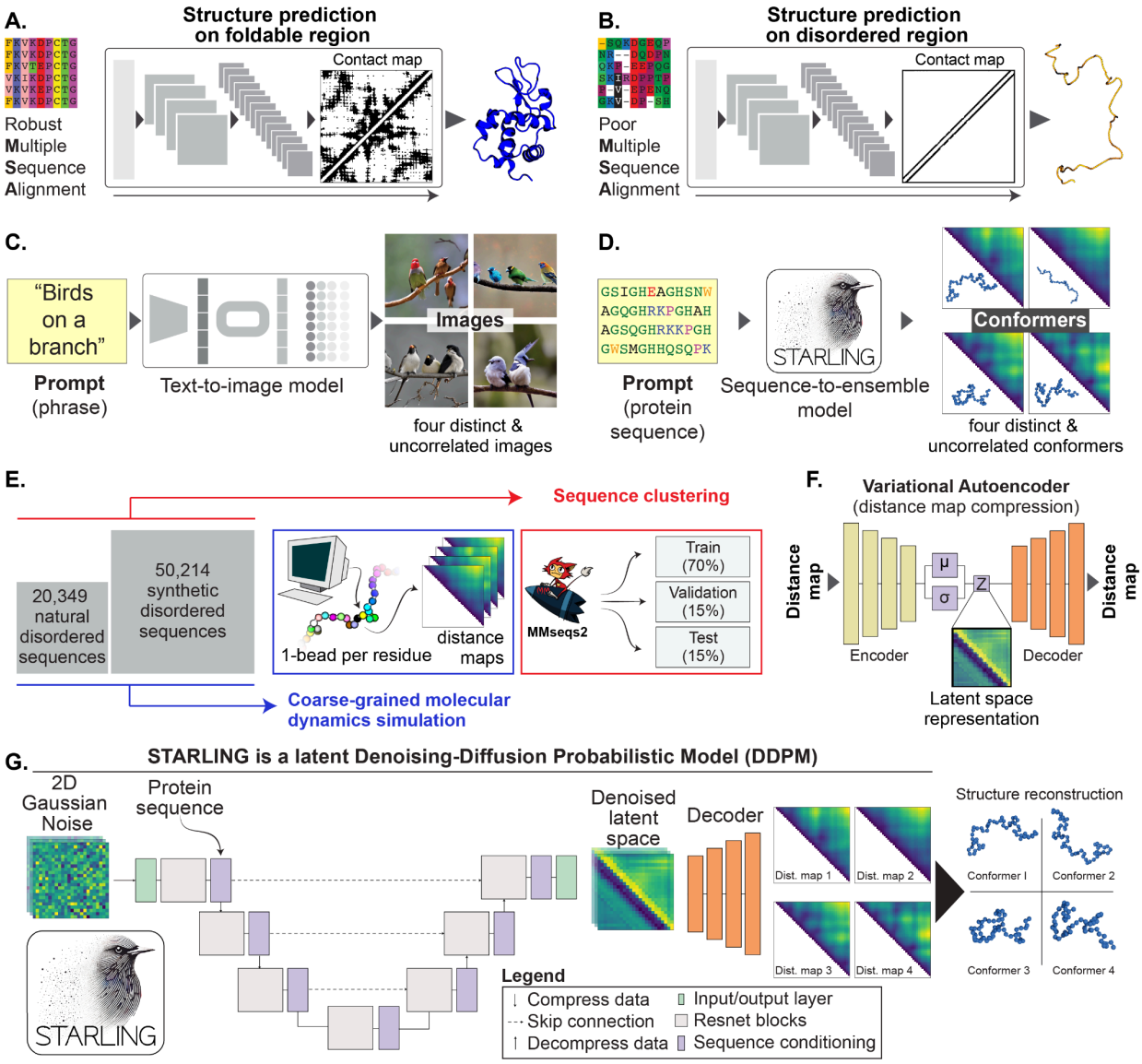


Figure 1. (A) Deep learning has revolutionized protein structure prediction, with major advances being facilitated by large-scale evolutionary information. **(B)** Structure prediction methods for folded domains are poorly suited for predicting the behavior of IDRs. These limitations come from the absence of a native-state structure and because evolutionary information is often poorly captured in MSAs. **(C)** Generative text-to-image models enable the creation of many unique and independent images consistent with a single input prompt. **(D)** IDR ensemble generation shares many similarities with text-to-image generation; we require many distinct and uncorrelated conformers, all of which are consistent with an input prompt (amino acid sequence). **(E)**. STARLING is trained on 70,563 amino acid sequences. For

each sequence, hundreds of distinct conformers are generated using coarse-grained molecular dynamics (MD) simulations, and each conformer is converted into a distance map (an image). Sequences are separated for training, testing, and validation. **(F)** STARLING makes use of a variational autoencoder (VAE) to compress distance maps to a latent space, allowing a denoising diffusion model to work in this latent space ('latent diffusion'). **(G)** The overall architecture of the STARLING model combines a latent-space probabilistic denoising diffusion model with a U-Net architecture using residual convolutions and transformer blocks. Latent space maps are decoded to real-space via the VAE decoder. Finally, if needed, distance maps are reconstructed into 3D coordinates via a parallelized multidimensional scaling approach.

STARLING is trained on coarse-grained simulations of 70,563 rationally designed and naturally occurring disordered protein sequences ranging between **10** and **384** residues in length (**Fig. 1E**, **Fig. S1**). Rationally designed sequences were designed using GOOSE[36]. While naturally occurring IDRs provide training data focused on the manifold of sequence features we may care about the most, the inclusion of a large number of rationally designed sequences allows us to construct training data that systematically titrates across sequence space to provide a well-rounded training set capturing the extremes of sequence composition and patterning. In short, synthetic sequences provide a well-rounded training set, capturing the extremes of sequence space which may not be well represented in natural sequences.

A necessary decision in generative models is defining the limits of your training data. We focused on sequences up to 384 amino acids in length for several reasons. First, almost 95% of naturally occurring IDRs across common model organisms are shorter than 384 amino acids (**Fig. S2**). Second, this length enables well-sampled coarse-grained simulations to be completed on a reasonable timescale. Third, the underlying architecture of the models requires us to have a fixed upper limit (see *Methods*).

Molecular Dynamics (MD) simulations were performed using the Mpipi-GG forcefield. Mpipi-GG[20] – a variant of the original Mpipi forcefield[18] – is a one-bead-per-residue coarse-grained model developed for disordered regions. Once run, conformations from simulations of a sequence of length *n* are converted into distance maps, *n* x *n* matrices where each element describes the distance between the ith and j^{th} residue for a specific conformation. This converts each IDR conformation into an "image", allowing us to leverage innovations developed for the conditional generation of images directly. Overall, our training dataset comprised nearly 10 million distance maps.

Two core limitations of DDPM models are their significant memory needs and a slow generative process at inference. These limitations come from performing the reverse denoising process in complex and high-dimensional data (e.g., images of size 384 x 384 pixels). To mitigate this, we developed a variational autoencoder (VAE) to compress each distance map into a lower-resolution latent space (24 x 24) (**Fig. 1F**). The denoising diffusion process can then occur

in this latent space (latent diffusion), significantly reducing memory requirements and inference time. Given this, STARLING was trained in two independent stages.

We first trained a highly accurate VAE that enables the compression of full-resolution distance maps into latent space (**Fig. S3**). Our VAE uses a ResNet18 architecture and learns parameters for an encoder (full-resolution to latent space) and a decoder (latent space to full resolution). We assessed the VAE's accuracy by encoding and decoding distance maps derived from sequences substantially different from those used in the training and validation sets. We evaluated our model on a held-out test set comprising 10,437 sequences, totaling 1,425,849 distance maps. The model achieved a root mean squared reconstruction error of 1.31 Å (**Fig. S4**). Furthermore, our model accurately reconstructs bond lengths with a root mean squared reconstruction error of 0.19 Å (**Fig. S4**), crucial for accurately modeling protein conformations.

In the second stage of training, we developed a discrete-time denoising-diffusion probabilistic model (DDPM) (**Fig 1D, Fig. S5**). Briefly, this model was trained to learn model parameters that map from random noise to individual latent-space conformer distance maps conditioned on the associated amino acid sequence. The training data came from 9,621,455 distance maps taken from 70,563 simulations. Each map was first compressed to latent space, and then the fixed forward diffusion process was used to add noise to each latent-space distance map. The model was then trained to learn parameters that reverse this forward diffusion process conditioned on the input amino acid sequence (**Fig. 1G**). Ultimately, ensemble generation is enabled by running many rounds of inference in parallel, leading to large numbers of independent distance maps.

While training requires the VAE encoder to generate latent space distance maps, only the VAE decoder is required for inference once the model is trained. Ultimately, the fully trained model (STARLING) combines two different models (the VAE decoder and the DDPM) that work together to enable rapid ensemble prediction. Using default settings, STARLING can generate 200 independent IDR conformations in approximately 9 seconds on a GPU (Nvidia A4500), 30 seconds on a Macbook Pro M3 CPU, and ~8 minutes on an Intel CPU (Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz) (**Figs. S6-8**).

We first checked that the STARLING-derived ensemble-averaged global dimensions agree with those obtained from the Mpipi-GG simulation of unseen sequences. In all cases, STARLING ensembles comprised 1000 conformers. Using a held-out test set of 10,437 sequences, we performed Mpipi-GG simulations and calculated radii of gyration (R_gs) and end-to-end distances (R_es). Our STARLING-derived R_gs (RMSE = 0.78 A, R^2 = 0.997) and R_es (RMSE = 3.21, R^2 = 0.991) are in excellent agreement with the simulations (**Fig. 2A,B**).

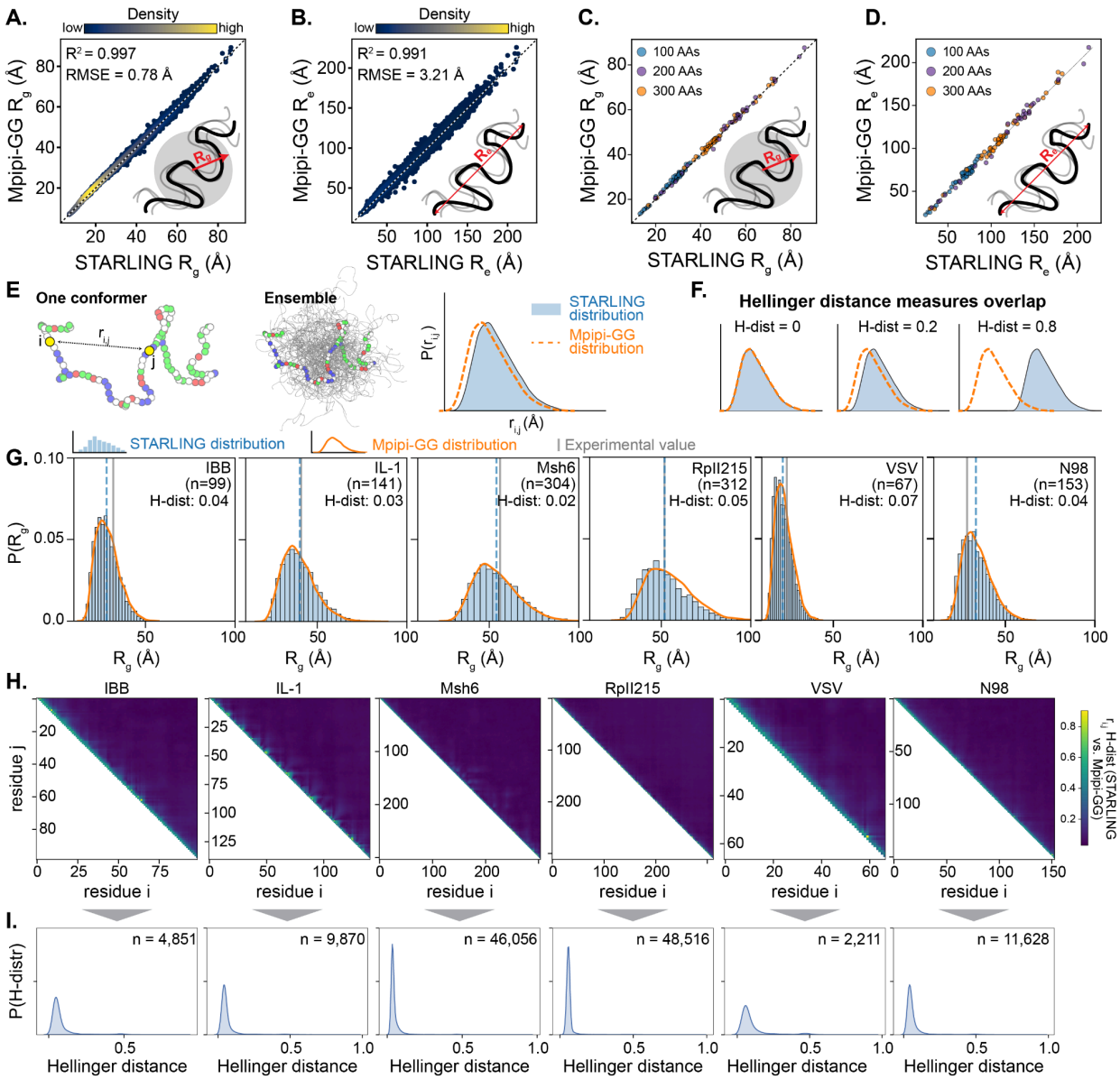


Figure 2. (A) Comparison of STARLING-derived average radii of gyration (R_gs) with those from Mpipi-GG (simulations) from a set of 10,437 unseen sequences. **(B)** Comparison of STARLING-derived average end-to-end distances (R_es) with those from Mpipi-GG (simulations) from 10,437 unseen sequences. **(C)** Comparison of STARLING-derived average radii of gyration (R_gs) for length-matched sequences with a broad range of chemistries for those from Mpipi-GG (simulations) from a set of 137 unseen sequences. This is done to verify STARLING has learned real sequence chemistry, not just that longer sequences are bigger. **(d)** Comparison of STARLING-derived average radii of gyration (R_es) for length-matched sequences with a broad range of chemistries for those from Mpipi-GG (simulations) from a set of 137 unseen sequences. This is done to verify STARLING has learned real sequence chemistry, not just that longer sequences are bigger. **(E)** STARLING generates full conformational ensembles, enabling ensemble distributions to be calculated (e.g., R_g, R_e, or $r_{i,j}$, inter-residue distance). For example, each (i,j) pair of residues in an ensemble yields a distribution of inter-residue distances, which can be quantified in STARLING ensembles and Mpipi-GG simulations. **(G)** Distribution similarity can be quantified by the

Hellinger distance (H-dist), where H-dist = 0 is perfect overlap and H-dist = 1 is no overlap. **(G)** Overlap of radii of gyration distributions obtained from STARLING (blue bars) and Mpipi-GG (orange line) with the experimental value shown for completeness. Overlap between the STARLING and Mpipi-GG distributions is quantified by the Hellinger distance (H-dist). **(H)** Hellinger distance values for all possible i,j inter-residue distance distributions across the same sequences examined in panel E. Almost all i,j distributions overlap very well, with deviations happening only at very short-range distances where the absolute magnitude of poor overlap is sub-angstrom. **(I)** Histogram of Hellinger distances in panels H.

An important determinant of model accuracy reflects ensemble size and the number of denoising steps. Using long timescale Mpipi-GG simulations to obtain an estimate of the true mean for the R_g, we systematically investigated how ensemble size influences the difference between the STARLING ensemble average value and the Mpipi-GG derived estimated mean (**Fig. S9**). We also examined the interplay between ensemble size and the number of denoising steps (**Fig. S10**). For the radius of gyration, the error with respect to our Mpipi-GG derived mean is flattened above 25 steps and 400 conformations. We, therefore, chose these values as default values for STARLING ensemble generation in our software implementation. However, we emphasize that larger ensembles may be necessary depending on the specific order parameter of interest.

Since R_e and R_g are highly correlated with sequence length, we next assessed the accuracy of our model on a set of length-matched sequences. This approach allowed us to determine whether the model had effectively captured the influence of sequence chemistry on the global dimensions of IDRs, or whether the model had simply learned to relate sequence length to global dimensions. Using all 100-, 200-, and 300-residue sequences with distinct sequence chemistries from the held-out test set, we found excellent agreement across a range of R_gs (**Fig. 2C,** RMSE = [0.64 Å, 0.79, 1.06] and R^2 =[0.99, 0.996, 0.996] for [100, 200, 300] residue lengths) and R_e (**Fig. 2D**, RMSEs = [2.7 Å, 3.5 Å, 4.4 Å], R^2 = 0.98, 0.99, 0.99] for [100, 200, 300] residue lengths). These results gave us confidence that STARLING had learned *bona fide* sequence-to-ensemble rules instead of simply learning polymer scaling theory.

In addition to ensemble-averaged observables, full structural ensembles enable us to calculate distributions of any observable of interest (**Fig. 2E**). We set out to assess how well STARLING-derived global dimension distributions match distributions from simulations. We used the Hellinger distance as a similarity measure to compare distributions. The Hellinger distance ranges from 0 to 1, where zero indicates that the distributions are identical, and one signifies that they are entirely disjoint (**Fig. 2F**). The STARLING-derived R_g distributions show excellent agreement with distributions from simulations (**Fig. 2G**). This overlap is quantified by a low Hellinger distance, confirming the high degree of similarity between the two distributions.

Finally, rather than focusing on distributions from global descriptors (e.g. R_g, R_e), we took the more stringent approach to assess distributions of all possible inter-residue distances between

simulations and STARLING ensembles. Here, we evaluated the Hellinger distance between every pair of inter-residue distances between Mpipi-GG and STARLING ensembles, again finding excellent agreement (**Fig. 2H, F**). In summary, all evidence supports that STARLING can directly predict meaningful conformational ensembles of disordered proteins from sequence.

Having established that STARLING can recapitulate simulated ensembles, we next investigated agreement with experimental data (**Fig. 3A**). We took a previously curated set of 133 sequences for which high-quality Small-angle X-ray Scattering (SAXS) data has been assembled, and average R_g values obtained. We found excellent agreement between average R_g values from STARLING with SAXS-derived R_g, obtaining values that are comparable to the state-of-the-art from simulations (RMSE = **4.72**, R^2 = **0.9**) (**Fig. 3B**, **Fig. S11**).

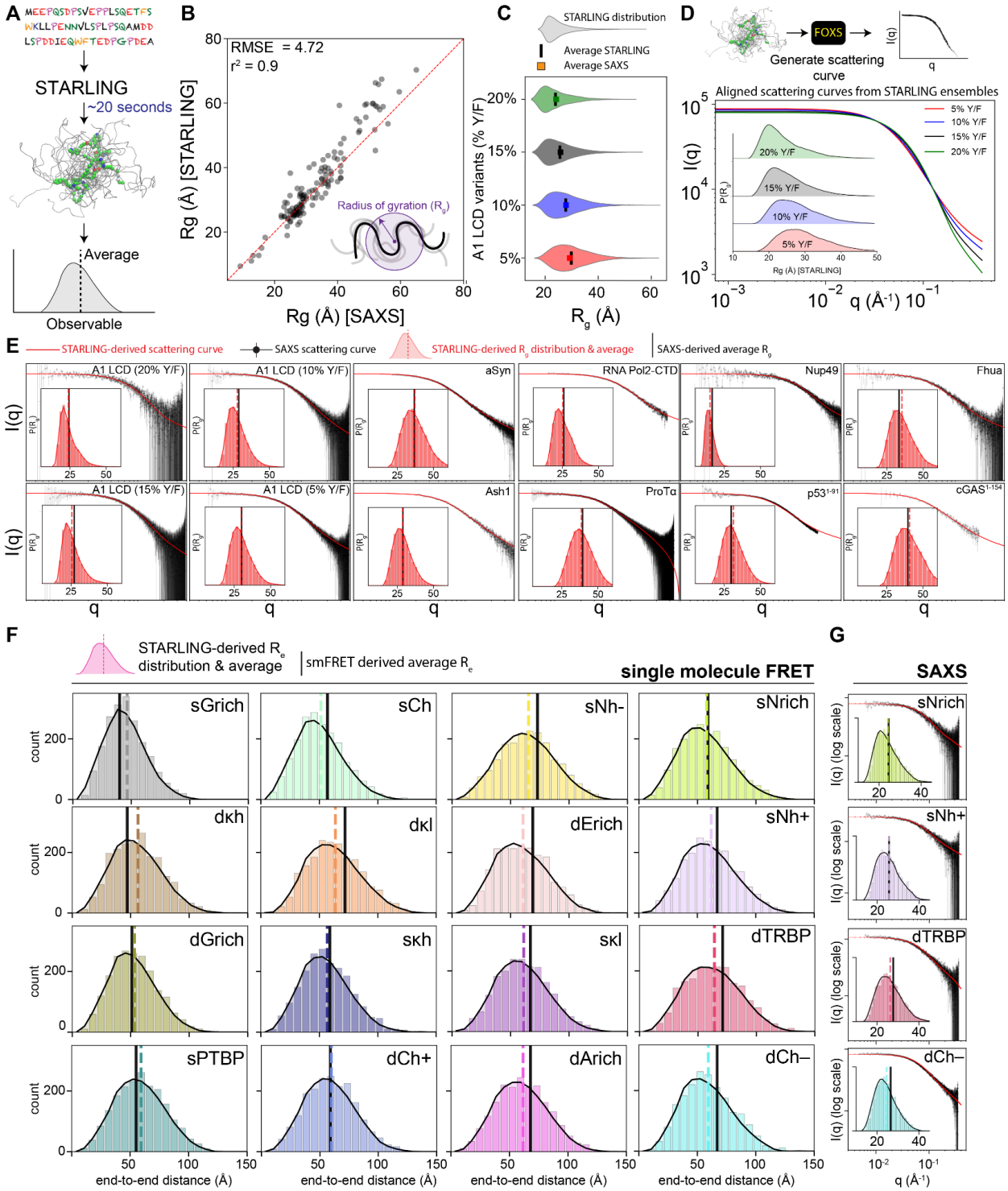


Figure 3. (A) STARLING enables rapid generation of full ensembles from sequence. **(B)** STARLING shows state-of-the-art accuracy in terms of predicting the average R_g from sequence. 133 IDR sequences comparing SAXS-derived and STARLING-derived average R_g values (see also Fig. ASHXX, ASHXX). **(C)** STARLING accurately captures the effect of small numbers of mutations, here illustrating how altering the number of aromatic residues changes the global dimensions of a low-complexity domain (LCD) taken

from the RNA binding protein hnRNPA1. **(D)** Full scattering curves can be back-calculated from STARLING ensembles using FOXS, illustrating how changes in conformational behavior alter both global dimensions (small angles, low q value) and the shape of the ensemble (intermediate and high q values). **(E)** Comparison of twelve different full small-angle X-ray scattering (SAXS) profiles with scattering profiles calculated from STARLING ensembles. Insets show STARLING Rg distributions with average R_g values from scattering data shown as a solid black line. See also **Fig. S13-14**. **(F)** STARLING-derived R_e distance distributions with experimental values are shown as black lines for sixteen different length-matched disordered protein sequences with different sequence chemistries. See also **Fig. S15. (G)** SAXS scattering curves for unlabeled proteins for a subset of those in pane F.

Next, we evaluated the sensitivity of our model to capture previously established trends between IDR sequence chemistry and ensemble dimensions. To do this, we predicted ensembles for all length-appropriate IDRs in DisProt[37], generating ensembles for 3,203 IDRs (**Fig. S12**). This revealed established sequence-ensemble relationships reported elsewhere (e.g., more compact IDRs are enriched in aromatic residues, whereas more expanded IDRs are enriched in proline and glutamic acid)[9,20,38,39]. To assess if STARLING can capture how small changes in sequence affect the ensemble, we generated ensembles for four variants of the hnRNPA1 low-complexity domain (A1-LCD) that vary the number of aromatic residues. These ensemble confirmed a clear relationship between aromatic content and dimensions, in quantitative agreement with the SAXS-derived experimental data (**Fig. 3C**)[9].

While we initially investigated SAXS data by comparing it against R_g values (**Fig. 3B, 3C**), SAXS experiments provide scattering data that reports on polymeric properties of IDRs across a range of lengths (down to 5-10 Å resolution). Given STARLING generates full ensembles, we can use existing tools for back-calculating scattering curves to reconstruct synthetic scattering data from our ensembles, enabling direct comparison between raw experimental data and STARLING-derived conformations. Using FOXS, we first showed that the subtle changes seen for the A1-LCD variants are clearly observable in the resulting scattering profiles (**Fig. 3D**)[9,40]. We, therefore, took twelve unrelated IDRs where high-resolution scattering data were available and compared scattering profiles derived from STARLING ensembles to experimental data, finding excellent agreement across a wide range of sequence chemistries and sequence lengths (**Fig. 3E**). We further examined an additional forty scattering profiles, finding the vast majority are in excellent agreement with STARLING-derived scattering profiles (**Fig. S13-14**). Taken together, our data suggests that our STARLING ensembles faithfully reproduce coarse-grained descriptions of IDR ensembles across a range of length scales and sequence chemistries.

Finally, we sought to compare against an alternative experimental modality. Recent elegant work illustrated how a series of sixteen length-matched sequences show a range of end-to-end distances depending on the underlying sequence chemistry, providing the most comprehensive experimental and computational investigation into sequence-ensemble relationships to date[17]. We compared predicted end-to-end distances (and, where available, SAXS-derived scattering

data) for those sixteen sequences, again finding good agreement between STARLING ensembles and experimental data (**Fig. 3F, Fig. S15**). Of note, our average error here across the sixteen sequences (RMSE 6 Å) is comparable to the differences obtained between different experimental groups in a recent smFRET benchmarking study[41]. Overall, our work suggests that STARLING can generate full coarse-grain ensembles of IDRs that accurately capture experimental data.

Finally, we applied STARLING to investigate a set of distinct IDRs in various contexts. We aim to illustrate the types of investigation STARLING enables rapidly and efficiently. The master transcriptional regulator cMyc (439 residues) underlies cell growth, metabolism, and proliferation and is frequently dysregulated in many cancers[42]. Besides a C-terminal HLH DNA binding domain and leucine zipper (both of which are likely partially or entirely disordered in an unbound state), cMyc is entirely disordered. The large size of its IDR (361 residues) coupled with its chemical composition has challenged biophysical characterization (**Fig. 4A**). Using STARLING, we generated conformational ensembles of $cMyc^{1-361}$ to investigate sequence-encoded conformational biases in the context of regions in Myc known to bind partners (so-called Myc boxes) (**Fig. 4B, Fig. S16**). Curiously, the STARLING-derived ensemble suggests that $cMyc^{1-361}$ can be divided into two halves (IDR1 and IDR2). IDR1 ($cMyc^{1-200}$) is more compact and engages in intramolecular interactions and some long-range interactions. IDR2 ($cMyc^{201-361}$) is more expanded with locally compact regions. Finally, conformationally distinct subregions align with established Myc-box boundaries, indicating local and long-range conformational properties can align with established functional annotations. Overall, these results predict $cMyc^{1-361}$ contains strong sequence-encoded biases that may be correlated with (or even underlie[43]) its function.

Although STARLING enables the investigation of hitherto unexplored IDRs, it can also provide new insight into existing systems. The C-terminal domain (CTD) of the DNA-directed RNA polymerase II subunit RPB1 consists of a series of heptad repeats with the consensus sequence YSPTSPS **(Fig. 4C)**[44]. This region stretches for 52 repeats in humans, and the number of repeats varies across organisms. Given the repetitive nature of the CTD, we reasoned principles from polymer physics would be well-suited to quantitatively describe the global dimensions as a function of CTD length. To this end, we generated STARLING ensembles across 26 different CTD lengths to generate a systematic series. We also performed all-atom Monte Carlo simulations across five different lengths[45,46]. The all-atom and STARLING derived ensembles show excellent agreement with one another as well as extant SAXS data **(Fig. 4D)**. Fitting the radius of gyration vs. number of amino acids to a polymer scaling law of the form $R_g = B_0N^v$ (where R_g is the radius of gyration, B_0 is a prefactor term that captures a combination of chain stiffness and monomer excluded volume, and v is the apparent Flory scaling exponent) revealed well-fit parameters of B_0 = 2.96 Å and v = 0.51. The convergence of STARLING ensembles, all-atom simulations, SAXS, and theory all suggest that – at least with respect to *in vitro* conditions – the global dimensions (including end-to-end distance, **Fig. S17**)

can be reasonably well-captured using a simple analytical description, enabling the 3D dimensions of CTDs from across the kingdoms of life to be estimated directly.

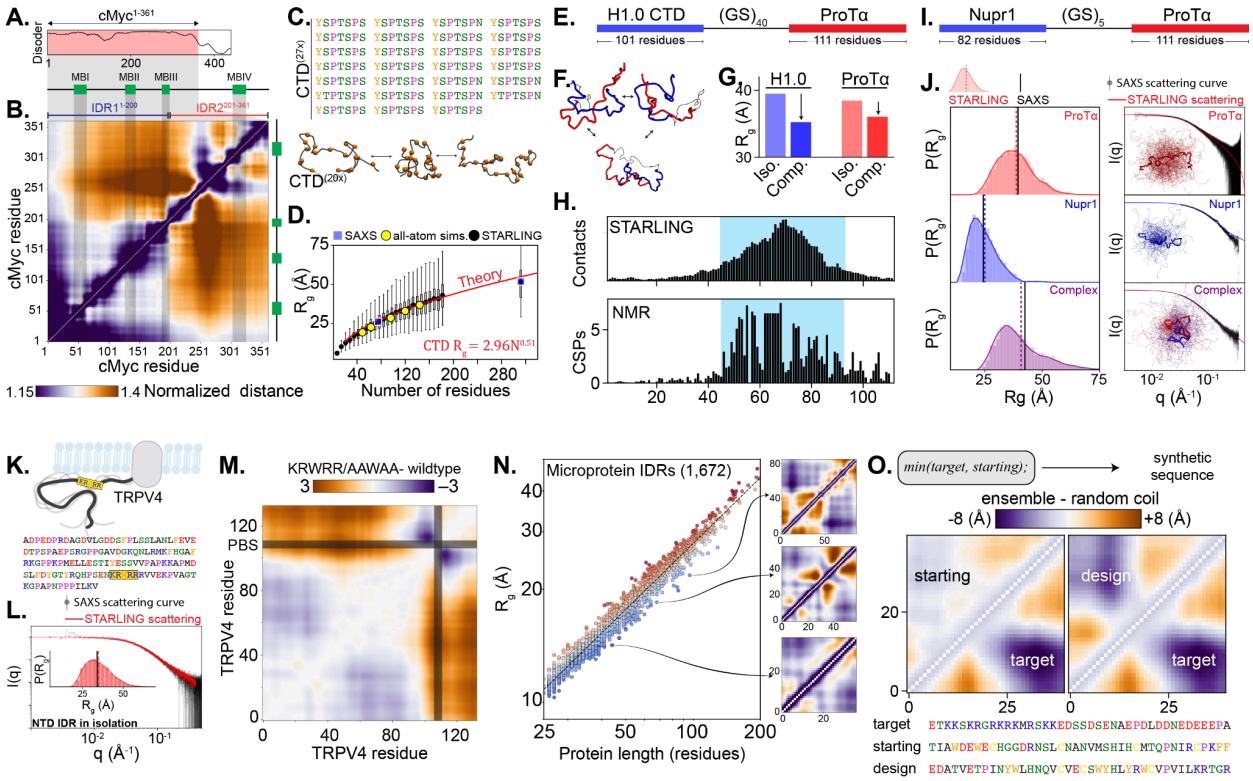


Figure 4. (A) The master transcriptional regulated cMyc is largely disordered. **(B)** STARLING provides a full description of its conformational ensemble, identifying distinct subregions and residues that drive attractive and repulsive intra-molecular interactions. Inter-residue distances here are normalized by limiting polymer model for a polypeptide as a Gaussian chain; darker colors are closer together and lighter colors are further away. **(C)** The RNA polymerase II CTD consists of a series of heptad repeats. **(D)** STARLING ensembles show excellent agreement with SAXS and all-atom simulations, enabling polymer theory to describe CTD dimensions with a simple analytical expression. **(E)** Histone H1.0 and ProTα form a high-affinity complex together. We generated ensembles for a construct in which the H1.0 CTD and ProTα are tethered by an inert $(GS)_{40}$ linker. **(F)** The resulting ensemble shows robust H1.0:ProTα interaction. **(G)** H1.0CTD and ProTα contract in the complex vs. in isolation. **(H)** Comparison of per-residue ProTα H1.0-CTD contacts and NMR-derived chemical shift perturbations (CSPs) show excellent agreement. **(I)** The proteins Nupr1 and ProTα have also been shown to form a complex, and we took a similar approach with a shorter $(GS)_5$ linker to enable reasonable characterize comparison to SAXS data. **(J)** STARLING-derived ensembles show excellent agreement with Nupr1, ProTα, and the Nupr1:ProTα complex. **(K)** The TRPV4 channel has a large N-terminal IDR with a basic PIP_2 binding site (PBS). **(L)** STARLING-derived ensembles show good agreement with small-angle X-ray scattering data. **(M)** The PBS engages in long-range interactions with a distal acidic region, and mutation of positive residues reduces these long-range interactions, in good agreement with extant experimental work. **(N)** STARLING enables systematic prediction of microprotein ensembles. **(O)** STARLING enables the rational design of sequences to match a desired distance map. Here, a random starting sequence is dissimilar to the target sequence, and after rounds of optimization, a new, entirely distinct sequence that recapitulates

the ensemble behavior of the target sequence is generated. Distances here are shown as absolute inter-residue distance with a random coil model subtracted.

We next wondered if we could use STARLING to characterize disordered bimolecular complexes. In the early days of AlphaFold2, biomolecular interactions were investigated by tethering two folded domains using a flexible glycine-serine (GS) linker. We reasoned a similar approach might be possible with STARLING **(Fig. 4E)**. Tethering the C-terminal disordered region from the Histone H1.0 ($H1.0^{97-198}$) to the histone chaperone Prothymosin α (ProTα) enabled the formation of an electrostatically-driven complex between the two proteins **(Fig. 4F)**, as reported previously[15]. Both proteins contract in the complex **(Fig. 4G)**, as observed and predicted by theory[47]. Moreover, the specific residues in ProTα identified by NMR-derived chemical shift perturbations (CSPs) match those residues from the STARLING complex ensemble that drive intermolecular interactions **(Fig. 4H)**. These results indicate that STARLING-derived binary complex ensembles can capture elements of IDR:IDR interaction ensembles.

While there is a limited number of biophysically-characterized disordered complexes, we also investigated the complex between ProTα and the fully disordered 82-residue protein Nupr1 (**Fig. 4I**). Prior work has characterized the proteins individually and in complex using SAXS[48]. We again find good agreement between STARLING ensembles and SAXS data for the individual proteins and – using a shorter GS linker (GS) – the disordered bi-molecular complex (**Fig. 4J**). Overall, our work suggests that STARLING has the potential for the prediction of fully disordered bi-molecular complexes, enabling the development of precise sequence-specific molecular hypotheses that can be tested experimentally.

Although our analyses thus far have primarily focused on ensemble differences between very different sequences, we wondered if STARLING could interpret relatively small amino acid changes. The N-terminal IDR from the Transient Receptor Potential Vanilloid 4 (TRPV4) ion channel has been thoroughly characterized through a battery of biophysical techniques to reveal a regulatory long-range interaction between a basic PIP_2 binding site and an acidic cluster (**Fig. 4K**)[49]. STARLING-derived ensembles of this region show good agreement with SAXS data (**Fig. 4L)**. Moreover, comparing the difference in intramolecular distances for the wildtype vs. a KRWRR-to-AAWAA mutant that was shown experimentally to reduce long-range interactions, we observe a commensurate reduction in long-range interactions in STARLING **(Fig. 4M)**. These results illustrate that STARLING enables the investigation of even relatively modest changes in sequence to examine sequence-ensemble-function relationships.

We were also interested in investigating STARLING's application to larger-scale ensemble prediction. Recent work has characterized a large set of microproteins – non-canonical ORFs

typically under 100 amino acids in size – using a combination of mass spectrometry and Ribo-Seq-based approaches[50]. While these microproteins are highly disordered, they possess distinct amino acid sequence biases from 'canonical' ORF-derived IDRs **(Fig. S18**). From 1785 microproteins in these data, we identified 1672 disordered regions and predicted ensembles for all of them. While many short IDRs are relatively expanded, we noticed many sequences with extensive intramolecular contacts driven primarily between aromatic and arginine residues (**Fig 4N**). These results suggest that even short-disordered proteins can possess sequence-encoded conformational biases.

Finally, given the speed with which STARLING enables ensemble prediction, we wondered if we could use STARLING to generate sequences that match the conformational ensemble of a target sequence. To this end, we developed an 'ensemble dissimilarity metric' based on STARLING-generated distance maps and used it as part of an objective function for the design of IDRs with desired ensemble properties (see *Methods*). As a proof-of-concept, we started with a randomly generated amino acid sequence and used STARLING to optimize the sequence towards matching the conformational ensemble of a short region in the C-terminal IDR of CTCF (Uniprot ID: P49711, residues 590-629). We generated a new sequence that matches the long and short-range behaviors of $CTCF^{590-629}$ **(Fig. 4O, Fig. S19)**. Alongside similar methods[36,51,52], the ability to rationally design IDRs with desired conformational properties opens the door to test the importance of conformational biases in disordered protein ensembles.

Overall, these vignettes demonstrate the versatile range of applications STARLING enables. Notably, the speed and ease with which STARLING generates ensembles means these investigations can be accomplished in hours or even minutes.

DISCUSSION & CONCLUSION

In this work, we present STARLING, a multiscale generative model capable of producing coarse-grained IDR ensembles directly from protein sequences in seconds. While there are many caveats and limitations (discussed below), the primary advantage we see with STARLING is that it drastically reduces the barrier to obtaining conformational predictions on sequence-dependent IDR behavior. While IDRs are often ignored or visualized in terms of AlphaFold "orange spaghetti," STARLING makes it straightforward for anyone to obtain realistic coarse-grained ensembles. Rather than "the end" of an investigation into IDR sequence-ensemble behavior, we view STARLING as a beginning, making it straightforward to develop hypotheses as to how an IDR's sequence may determine its conformational ensemble and/or how it may influence interactions with other IDRs.

STARLING does not provide information on IDR "dynamics". It is not an approach for generating dynamical trajectories from sequences but for rapidly generating large ensembles. While this

may be seen as a weakness, we see this as a strength in that the slow reconfiguration times of large chains often hampers simulations from obtaining a sufficient number of truly independent containers. In contrast, STARLING ensembles for long (200+ residue) IDRs are generated just as quickly as small ones, facilitating the computational investigation of IDRs that have been hitherto largely inaccessible.

A key feature of STARLING is the ability to generate full conformational ensembles instead of predicting just ensemble average values. While predicting average conformational behaviors (e.g., average R_g or average R_e) can be informative[20,21,53], it does not enable a 'structural' investigation into the determinants of a given behavior. With full conformational ensembles, the specific sets of residues responsible for an observed behavior can be directly identified (**Fig. S20-21**). Moreover, substantial effort has been placed into developing methods to back-calculate experimental observables from simulation trajectories to interpret experimental data. STARLING ensembles are amenable to these analyses, allowing comparison with a range of experimental results.

Our decision to focus on coarse-grained ensembles over all-atom ensembles was driven by several factors. Firstly, obtaining sufficient training data (especially to capture emergent long-range behaviors for larger IDRs) at all-atom resolution for larger IDRs would be prohibitively difficult with current approaches. While we could, in principle, back-map from coarse-grained to all-atom conformations, whether such back-mapping accurately reproduces local conformational biases at all-atom resolution is unclear, given the limitations of an isotropic amino acid bead. Moreover, it is not apparent whether sufficient experimental data exists to validate (or refute) such an approach.

While STARLING generates full coarse-grained conformational ensembles within seconds, it does have limitations. Firstly, the ensembles generated by STARLING lack any information about secondary structure. Since STARLING models are trained on data from one-bead-per-residue coarse-grained force fields, we are blind to putative secondary structural elements in the disordered protein sequence. Secondly, STARLING operates on single-chain, fully disordered protein sequences; however, disordered regions are generally found either adjacent to and/or between folded domains[1]. While STARLING is powerful at recapitulating the solution behavior of strictly disordered protein sequences, there is emerging biophysical evidence that folded domains (or membranes) can influence the biophysical ensembles of adjacent disordered protein sequences. Although a limitation of the raw predictive power of STARLING, we view this as a "feature" rather than a "bug" as deviations from conformational ensembles obtained in isolation could indicate biophysically meaningful interactions (and potentially function). Moreover, with enough compute, careful design, and consideration for sequence library design and structural coverage, future models could address this via massive large-scale simulations of folded domains tethered to physicochemically disparate disordered sequences.

Another limitation is that STARLING models are trained on data generated from simulations performed under “standard conditions”; however, an increasingly appreciated role for intrinsically disordered proteins is their involvement in environmental sensing in response to changes in the local chemical environment[54]. The increased solvent accessibility of disordered proteins and unique physicochemical sequence biases poises IDRs to be particularly responsive to changes in this local chemical environment (e.g., salt, pH, metabolites, nucleic acids, other proteins, etc.). Understanding how disordered proteins respond to these changes in the solution environment are thus an important area of future research. Since STARLING was not trained on these data types, its reliability in this field cannot be determined. Moreover, since the field lacks good experimental data to benchmark these models, it is unclear how useful a deep-learning-based surrogate model would be for these use cases, which starkly contrasts the typical solution behavior we rigorously compare against in this work. Nevertheless, similar models could be trained with more advanced simulation models as the field progresses.

Another limitation of STARLING is its inability to generate Boltzmann-distributed conformations. While scoring functions learned by DDPMs have been shown to approximate the underlying force field, these models are not sampling directly from the canonical ensemble. Nevertheless, it’s recently been shown that the scoring function learned by a DDPMs is a relatively strong approximation of the underlying force fields. Indeed, in some cases, this scoring function can actually be employed as a force field to directly sample conformers from the approximate force field[55].

Several other approaches for generating IDR ensembles from sequences have been released over the last year. Like STARLING, idpSAM is a latent-space probabilistic denoising diffusion model that enables C-alpha predictions[56]. However, unlike STARLING, idpSAM was trained on 3,259 ensembles up to 60 amino acids in length generated using all-atom simulations. For sequences over 60 amino acids, idpSAM’s predictive power is expected to be substantially reduced. Since many IDRs are substantially larger than 60 amino acids, STARLING addresses a niche that idpSAM does not currently serve, although we see STARLING as potentially synergistic with idpSAM. IDPFold is another approach for ensemble generation that combines protein language models with experimental and molecular dynamics data[57]. Specifically, IDPFold uses high-resolution crystal structures combined with NMR structures and coarse-grained conformers generated by the CALVADOS force field back mapped to all-atom resolution. Such an approach relies heavily on the accuracy of backmapping tools for local conformational behavior. Moreover, with prediction times of ~20 minutes per sequence, the requirement for GPUs, and correlations between experimental (SAXS) and IDPFold-derived radii of gyration of 0.80, the predictive power for at least global dimensions is relatively modest. Nevertheless, we again see IDPFold as a synergistic tool in an emerging ecosystem of ensemble prediction methods.

In conclusion, we hope STARLING will be useful to the disordered protein community and beyond. In contrast to many deep learning tools, we have made a substantial effort to ensure STARLING is easy to use and install (e.g., weights are managed automatically, installation is as easy as pip install, and prediction can be done using a single command-line command). The ease with which STARLING facilitates biophysical investigation into IDR conformational behavior is substantial. Full coarse-grained ensembles can be predicted on the fly on a whim, in meetings, or during seminars, while large-scale analysis can be performed on commodity hardware in hours. We continue to work on accuracy and performance and anticipate faster versions of STARLING in the coming months. While the validity of predictions should always be considered, and we do not see STARLING as a replacement for any existing technologies, we nevertheless see STARLING filling a currently unmet need in enabling biophysical investigation of IDR behavior to the broader community.

Author Contributions:

Conceptualization: BN, JML, ASH; **Methodology**: BN, JML; **Software**, **Validation**, **Formal analysis**, **Investigation**: BN, JML, RJE, ASH; **Resources & Funding Acquisition**: JML, ASH; **Data Curation**: BN, JML, ASH; **Writing & Visualization**: BN, JML, RJE, ASH;

Funding Sources:

This work was funded by The US National Science Foundation (NSF) through grant 2338129 (CAREER) to ASH and the US National Institutes of Health (NIH) through CA290639 (DP2) to ASH. JML was supported by the National Science Foundation via grant number DGE-2139839 and by the Frontera Computational Sciences Fellowship. We also acknowedlege generous support from

Acknowledgements:

We thank Jackie Pelham, Ramiz Somjee, and Bede Portz for their comments on the manuscript, and we thank Ute Hellmich for giving such an exceptional talk that inspired our investigation of the TRPV4 IDR. We especially thank our many colleagues for sharing and depositing raw experimental data (notably SAXS data on SASDB), which has greatly facilitated comparison with experiment. We also thank Jungsan Sohn, Scott Showalter, and Maria Grazia Ortore for sharing raw SAXS scattering data. We thank Birthe Kragelund and Freia Stryhn Buus for sharing chemical shift data for the ProTa:H1.0 complex.

Competing interest:

A.S.H. is on the scientific advisory board (SAB) for Prose Foods. All other authors declare no competing interests.

Version comment:

We note that this version of the manuscript lacks supplementary tables with all of the sequences used in this study. All these data will shortly be available at https://github.com/holehouse-lab/supportingdata/tree/master/2025/starling_2025.

METHODS

Sequence design and data splitting

We needed to construct a physicochemically diverse library of sequences to train our models for downstream simulations/model training. The sequence library presented in this work encompasses an expanded set of the constructs originally presented in Lotthammer *et al*.[20]. In addition to those sequences, we have also added 4,284 tri-block copolymer sequences (2994 in training, 654 in test, and 636 in validation). These tri-block sequences consist of three blocks: N and C terminal blocks of particular chemistry (consisting of a repetitive low complexity sequence) separated by a comparatively inert block. The N and C terminal blocks may be the same but generally differ. For example, one potential sequence would be positive charges, an inert GS linker, and negative charges.

After designing the sequence library and prior to model training, we clustered the sequences using MMseqs2[58] protocol (`mmseqs easy-cluster input.fasta output_directory tmp --min-seq-id 0.7 -c 0.8 --cov-mode 1`). Next, we stratified the cluster centers by length and assigned all sequences within each cluster into either train (49,423 sequences, 70%), validation (10,703 sequences, 15%), or test sets (10,437 sequences, 15%) (**Fig. S1**).

Molecular Dynamics Simulations and Analysis

Unless otherwise noted, we followed the same simulation protocol described previously[20]. Briefly, this process included performing Mpipi-GG simulations with the LAMMPS simulation engine[59]. We built initial configurations of all disordered proteins according to a random walk. For all simulations, we performed up to 1000 iterations of energy minimization until the force tolerance was below 1 x 10^{-8} (kcal/mol)/Å . Simulations were performed in the NVT ensemble at 150 mM NaCl. All simulations employed a weakly-coupled Langevin thermostat that was adjusted every 100 ps to maintain a target temperature of 300 K. The integration timestep was 20 femtoseconds for all simulations. Periodic boundary conditions were used in all simulations with a box length of 500 Å. All simulations were performed for 10 μs with trajectory coordinates saved every 2 ns.

All of our simulations were analyzed with SOURSOP and MDTraj[60,61]. For all simulations analyses, we chose to subset the data for (1) computational tractability and (2) to reduce the correlation between conformers by striding our analysis to only compute inter-residue distance maps every 30 frames, equating to one conformation for every 60 ns or 3 million integration steps. Collectively, we have a total of 9,621,465 inter-residue distance maps.

Generative Model Training

We coupled two deep learning methodologies, a variational autoencoder[62] (VAE) and a denoising-diffusion probabilistic model[30,63] (DDPM), for efficient sequence-to-ensemble predictions. We constructed training, validation, and test datasets using coarse-grained

simulations performed with the Mpipi-GG force field. Our dataset consists of sequences between 10 and 384 amino acids in length, and due to this variability in length, we padded all our distance maps to a consistent size of 384 by 384.

We trained a variational autoencoder to learn a compressed latent representation of distance maps. We chose to train a VAE to represent our distance maps in a 24 by 24 latent space, which amounts to a spatial compression ratio of 256. These design decisions enabled us to reduce the computational resources required for model training and inference while retaining strong performance. Following standard practices in VAE model training, our training objective consists of a reconstruction loss and a Kullback-Leibler (KL) divergence term. The reconstruction loss is modeled as negative log-likelihood under a Gaussian distribution, and KL loss regularizes the latent space by penalizing deviations from a unit Gaussian distribution (**Fig. S3**). We down-weighted the KLD loss by a factor of $1e^{-4}$, similar to previously reported approaches[31], to permit our model to balance accurate reconstruction of inter-residue distance maps with a regularized and structured latent space.

Next, we developed a latent diffusion model by training a discrete time Denoising-Diffusion Probabilistic Model (DDPM) within the latent space of a variational autoencoder (VAE). The DDPM is a generative model that learns the data distribution by gradually denoising samples starting from isotropic Gaussian noise over a series of timesteps. To achieve this, we first corrupted the data through a noising process (i.e., the forward process) where Gaussian noise is progressively added according to a noise schedule.

$$q(X_{1:T}|X_0) := \prod_{t=1}^{T} q(x_t|x_{t-1})$$

where T ranges from t=1 to t=1000 and $q(x_t|x_{t-1}) := N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$, where beta is defined by the variance scheduler described below.

To ensure robust learning of the noise added at each time step – randomly sampled from discrete values between t=1 (original data) and t=1000 (fully noised data) – we employed a variance schedule that controls the magnitude of the added noise throughout the diffusion process. The schedule begins by adding a small amount of noise (low variance) where the original data is "information dense" and gradually increases. In our case, we used a cosine scheduler to select variances between 0.0001 and 0.999 with an offset factor of 0.008. This offset has a minor effect and is primarily motivated by addressing pixel binning issues in image models, where data is constrained to a [0,255] dynamic range. However, this rationale does not directly translate to our spatial dataset, which is unnormalized and has different structural properties. While the impact of this offset in our case is expected to be minimal, we acknowledge its suitability for our data is unverified. We did not conduct ablations to assess its

effects, and alternative approaches tailored to our dataset could be worth exploring in future work. A neural network was then trained to reverse this process conditioned upon the current timestep and the noisy data at that time step.

To parameterize a generative model, we train a neural network to reverse the time-dependent addition of Gaussian noise described in the forward process above. The reverse process is defined by the following relation:

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$

where $p_\theta(x_{t-1}|x_t)$ is a Gaussian distribution, and $p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \beta_t I)$ and $\mu_\theta(x_t, t)$ is computed from the predicted noise $\epsilon_\theta(x_t, t)$ according to the following equation:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \prod_{t=0}^{t} \alpha_t}} \epsilon_\theta(x_t, t)\right)$$

The parameter β_t is fixed according to the time-dependent variance noise scheduler described below and $\alpha_t := 1 - \beta_t$. The next denoised state, x_{t-1}, is given by $x_{t-1} = \mu_\theta(x_t, t) + \sigma_t \epsilon$ where

$$\sigma_t = \sqrt{\beta_t \frac{1 - \prod_{t=0}^{t-1} \alpha_t}{1 - \prod_{t=0}^{t} \alpha_t}} \text{ if } t > 1\text{; otherwise } x_{t-1} = \mu_\theta(x_t, t)\ .$$

We utilize a U-Net architecture with both residual convolutions and transformer blocks. The model processes batches of 2D isotropic Gaussian noise with a single feature channel and spatial dimensions concordant with the VAE latent space (**Fig. S5**). The input undergoes three steps of spatial compression and feature extraction, increasing the feature count by 512x and reducing the spatial dimensions by 64x at the base of the U-Net. This is followed by a series of upsampling blocks that gradually restore the spatial dimensions and reverse the feature extraction process, ultimately predicting the noise content within the original input. Prior to each downsampling and upsampling convolutional step, we augment our residual blocks with a transformer. Each transformer includes both self-attention, where each pixel in the latent space attends to all others, and cross-attention, which captures the relationship between the learned sequence embeddings and the latent representation of distance maps. The sequence embeddings are a (512, ℓ)-dimensional vector where ℓ is the 384 residues. If a given sequence

is less than 384 residues, we pad the input sequence to a length of 384. The 512-dimensional vector comprises both learned amino acid token embeddings and incorporated positional encodings via the addition of sinusoidal embeddings to those token embeddings. Our sinusoidal embeddings use the commonly used theta parameter of 10,000. Taken together, incorporating these cross-attention layers enables our model to conditionally generate sequence-specific distance maps.

Finally, the diffusion model is trained with a mean squared error loss function between the added and predicted noise according to the following equation:

$$L = E_{x_0,\epsilon,t}[\left|\left|\epsilon - \epsilon_\theta(x_t, t)\right|\right|^2]$$

During training, distance maps are encoded using a pre-trained VAE, with the output rescaled to achieve unit variance. Noise is then added to the latent representation, and our transformer-augmented U-Net predicts the amount of noise introduced.

Model inference

Putting it all together, during inference, 2D isotropic Gaussian noise alongside sequence information is iteratively input through the trained U-Net until all the noise is removed. The final output of the U-Net is then fed into the VAE decoder to project back to a ℓ x ℓ matrix of inter-residue distances consistent with the input sequence. Importantly, each instance of the 2D isotropic Gaussian noise is sampled independently, allowing us to generate decorrelated IDR conformations. The iterative process of removing noise can be time-consuming and typically requires thousands of iterations through the neural network. Custom samplers aim to reduce the number of passes through the neural network while maintaining high-quality samples by redefining the forward and reverse process of the DDPMs. Following advances in the image generation literature, we have implemented a denoising-diffusion implicit model[64] (DDIM) to speed up the generation of samples from our model. Since DDIMs share the same training objective as DDPMs, no additional training is required, and we can use DDIM samplers to improve the inference speed of any trained DDPM.

Projecting distance maps to structural ensembles.

To project from 2D inter-residue distance maps back to 3D coarse-grained structures with the same set of inter-residue distance distributions, we use a custom PyTorch implementation optimized for our use case of the Scaling by MAjorizing a COmplicated Function (SMACOF) algorithm - we note that we did not choose this approach for its apt name (although we endorse it, as it follows the spirit of our software package names), but rather for its improved rate of convergence on stress optimization. We note that in our preliminary tests our implementation offers a ~70x improvement over the default scipy implementation.

If PyTorch-compatible hardware (MPS or CUDA) acceleration is not available, we provide a fallback to the scipy implementation of the Multidimensional Scaling (MDS) algorithm available in sklearn (*sklearn.manifold.MDS*). Briefly, this algorithm works by projecting our 2D inter-residue distance matrices - where each entry d_{ij} in the input matrix represents the spatial distance between residues i and j - into 3D cartesian coordinate feature space. When performing this projection, we use these inter-residue distances as the precomputed dissimilarity measure. This projection aims to derive a set of 3D coordinates that best capture the spatial relationships between residues, as reflected in the input distance map.

Mathematically, this projection optimizes a stress function that quantifies the discrepancy between the pairwise distances in the 2D matrix and those in the projected 3D space. The stress function is defined as:

$$S = \sqrt{\frac{\sum_{i<j}(d_{ij}-\hat{d}_{ij})^2}{\sum_{i<j} d^2_{ij}}} \qquad Eq.\ 1$$

Where d_{ij} is the distance between residues *i* and *j* in the original 2D distance matrix. $\hat{d}_{ij}$ is the corresponding Euclidean distance between residues *i* and *j* computed from the projected 3D space.

Intrinsically Disordered Protein Ensemble Design

To enable IDR sequence design with target biophysical ensemble properties, we use the sequence optimization functionality from the Python package GOOSE (https://github.com/idptools/goose)[36]. Briefly, this sequence design approach iteratively modifies a starting sequence using random amino acid substitutions to minimize the mean squared error between a property or set of properties computed (or predicted) from a sequence. For the sequences designed in this manuscript, we leverage STARLING and this optimization approach to design novel sequences by computing an ensemble dissimilarity score. To compute this ensemble dissimilarity score, we use STARLING to generate distance maps for the target ensemble. Next, we take the sequence at the current iteration of the optimization process and compute the ensemble of distance maps for this sequence. Then, we compute the mean and standard deviations of the distances for each residue pair in the distance maps for both sequences. To account for variations in distance due to inter-residue proximity, these values are normalized by dividing by the distance from the diagonal. The sum of the absolute difference between the target sequence and the second sequence is then calculated for the normalized mean and standard deviation maps. The function then calculates the absolute difference between the target sequence and the current sequence for both the normalized mean and standard deviation maps. Finally, it combines these differences into a single scalar value that represents the dissimilarity between the two sequences.

The EnsembleDesign property comes with many tunable parameters that can be initialized and adaptively set during sequence optimization for improved performance. This primarily entails a progressive increase in the number of denoising steps and the number of conformations used as the initial sequence progresses closer and closer to the target ensemble. We also provide a patience functionality that controls how long optimization can continue without improving before increasing the number of steps or conformers.

For sequence optimization, we started with a randomly generated sequence. For our EnsembleDesign property, we set the conformations to always be 200 and the minimum and maximum number of steps to be 5 and 25, respectively. We also set the number of rounds of optimization without sequence improvement before increasing the number of steps to be 500 iterations. At each iteration, the starting sequence was randomly changed by substituting a random number of a randomly chosen amino acid for some other amino acid in the starting sequence. Next, the difference between the new sequence and the target sequence was quantified as described above. This evolutionary algorithm continues until convergence or the stopping criteria have been met.

Distribution overlap with Hellinger distance

To compute the Hillinger distance between two distributions, we use the implementation provided in the SOURSOP `sssampling` module, as described previously[65]. Briefly, this involves generating a density-normalized histogram for the two distributions of interest with the same bin values, and then measuring the overlap between those histogram densities (P and Q) as described below:

$$H(P,Q) = \frac{1}{\sqrt{2}} \times \sqrt{\sum_{i=0}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}$$

To ensure parity, all distributions are generated using identical numbers of datapoints, although in principle, this is not a pre-requisite for this approach to work.

Comparison with SAXS profiles

SAXS profiles are back-calculated using FOXS with default parameters and using 400 conformers (water_layer_c2 = 2.0, excluded_volume=1.0, max_q=0.4)[40]. The profiles were aligned by selecting a short region of experimental and computational profiles and identifying scalars that would best align the two profiles. Many examples of scattering data were obtained from SASBDB [66].

Large-scale ensemble predictions and bioinformatics

For DisProt prediction, we used DisProt obtained in the fall of 2024 (`DisProt_release_2024_06_with_ambiguous_evidences.fasta`) obtained from DisProt[37]. We the

filtered for IDRs between 25 and 380 amino acids in length. This resulted in 3,204 distinct IDR sequences.

For microprotein analysis, microprotein IDRs were predicted using metapredict V3. The reference human proteome was obtained from UniProt (`UP000005640_AND_revi_2025_02_07`) and all IDRs predicted using metapredict V3[2,67]. This analysis identified 33,473 IDRs in the human proteome. Compositional enrichment was calculated over all residues in distinct datasets.

All-atom simulations

All-atom simulations for CTD repeats of length 49, 63, 96, 120, and 144 amino acids were performed using the CAMPARI simulation engine and ABSINTH implicit solvent model using the ion parameters of Mao *et al.* as described previously [9,68–70]. Simulations were performed at 380 K to mitigate challenges in sampling (as described previously for aromatic-rich low-complexity sequences[9]), using 10 mM NaCl, with five independent replicas performed for 6 x 10^7 Monte Carlo steps each, with the first 1 x 10^7 steps discarded for equilibration. Conformers were saved every 2 x 10^4 steps, meaning each independent simulation generated 2500 conformations. Overall, for each CTD length, we obtained atomistic ensembles of 12,500 conformers. Atomistic ensembles were analyzed using SOURSOP[61].

Null model normalization

To enhance the visualization of inter-residue distances, we normalized inter-residue distances using the Analytical Flory Random Coil (AFRC) model[71]. This model represents the expected dimensions for a sequence-specific version of a polypeptide if the protein behaved as a true Gaussian (phantom) chain; that is, no excluded volume and no inter-residue attraction or repulsion. This becomes useful as a reference point to normalize the effects of sequence separation when comparing ensembles.

STARLING software implementation and installation

The STARLING software package is built using Python 3.11. STARLING makes heavy use of SciPy[72], NumPy[73], and PyTorch[74]. Documentation is built using Sphinx and hosted on ReadTheDocs. Testing is enabled by PyTest. Structure manipulation is facilitated by MDTraj [60].

Model weights

Initial weights are tagged as v1.0.0 and available at https://github.com/idptools/starling/releases/tag/v1.0.0. STARLING handles weights automatically on the backend, such that the first time a new install of STARLING is run it will automatically download the model weights for the VAE and DDPM. This happens once and requires an internet connection. After the weights have been downloaded once, STARLING defaults to using the cached versions found in `~/.cache/torch/hub/`. From the user perspective, this provides a seamless install experience.

Installation

To install, we strongly recommend using conda to set up an isolated environment, although we note STARLING has no “esoteric” dependencies. To install, simply run;

```
# create a conda environment
conda create -n starling  python=3.11 -y

# activate the environmennt
conda activate starling

# install starling
pip install git+https://github.com/idptools/starling.git

# check starling is working by running
starling -help
```

Note that the first time any starling command is run it may take 30-90 seconds to initialize various things in your new Python environment and download the starling weights in the background. This is all a one-time thing.

Command-line tools

To generate an ensemble with starling, simply run

```
# predict an ensemble
starling MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEA
```

To define the number of conformers, use the `-c` (or `--conformations`) flag. By default, the name of the output file generated if a single sequence is passed will be `sequence_1.starling`. To set a custom name for the output file, use the `--outname` argument. The **`starling`** command can take a single sequence or a FASTA file with multiple sequences. In this situation, the names of the output files will be defined by their FASTA headers. For a complete set of documentation run

```
# predict an ensemble
starling --help
```

By default, STARLING generates `.starling` files as output. These files contain all the data and metadata associated with the ensemble prediction. To visualize the ensemble as a set of molecular structures, you can convert `.starling` files into PDB trajectories (or XTC/PDB pairs) using our conversion tools:

```
# convert a .starling file to a PDB trajectory
# This generates a single PDB file with each confomer as
# a different MODEL in the PDB file.
```

```
starling2pdb <starlingfile name>

# generate .pdb file for topology and .xtc for trajectory
starling2xtc <starlingfile>
```

Python library

When analyzing STARLING ensembles we highly recommend using the STARLING Python library instead of converting ensembles to structural ensembles and then analyzing them in this context. This is because converting to structural ensembles requires structure reconstructing, which can introduce (minor) errors. In contrast, analysis of the predicted distance maps mitigates those errors.

Working with STARLING ensembles in Python is straightforward. STARLING has just two user-facing functions: **`generate()`** and **`load_ensemble()`**.

```
from starling import generate, load_ensemble

# load in a pre-existing Ensemble object
E = load_ensemble('sequence_1.starling')
```

The Ensemble object has a range of functionality enabling the investigation of ensemble properties. New Ensemble objects can also be generated using the **`generate()`** function.

```
# generate a new ensemble object
E = generate(MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEA,
return_single_ensemble=True)
```

The documentation and development for this is ongoing; for feature requests or errors, please raise an issue at https://github.com/idptools/starling/issues.

Compute Hardware

All model prototyping was performed on an in-house deep learning workstation comprising 4 NVIDIA A4500s. All production training runs were performed on NVIDIA GH200 GPUs provided via generous support courtesy of the Texas Advanced Computing Center and the Frontera Computational Sciences Fellowship.

Code distribution

STARLING is fully open source and provided here:
https://github.com/idptools/starling/

STARLING's documentation is available here:

https://idptools-starling.readthedocs.io/en/latest/

STARLING is provided as a Colab notebook here:
https://github.com/idptools/idpcolab/blob/main/STARLING/STARLING_demo.ipynb

The code and data associated with the figures in this manuscript will (shortly) be available here (we promise!):
https://github.com/holehouse-lab/supportingdata/tree/master/2025/starling_2025

REFERENCES

1. Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat. Rev. Mol. Cell Biol.* **25,** 187–211 (2024).
2. Lotthammer, J. M., Hernández-García, 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
3. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6,** 197–208 (2005).
4. Mittag, T. & Forman-Kay, J. D. Atomic-level characterization of disordered protein ensembles. *Curr. Opin. Struct. Biol.* **17,** 3–14 (2007).
5. Mao, A. H., Lyle, N. & Pappu, R. V. Describing sequence–ensemble relationships for intrinsically disordered proteins. *Biochem. J* **449,** 307–318 (2013).
6. 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.* **32,** 102–112 (2015).
7. González-Foutel, N. S., Glavina, J., Borcherds, 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.* **29,** 781–790 (2022).
8. 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 ERα disorder. *Nature* 1–9 (2025).
9. 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).
10. Cubuk, J., Stuchell-Brereton, M. D. & Soranno, A. The biophysics of disordered proteins from the point of view of single-molecule fluorescence spectroscopy. **66,** 875–890 (2022).
11. Camacho-Zarco, A. R., Schnapka, V., Guseva, S., Abyzov, A., Adamski, W., Milles, S., Jensen, M. R., Zidek, L., Salvi, N. & Blackledge, M. NMR Provides Unique Insight into the Functional Dynamics and Interactions of Intrinsically Disordered Proteins. *Chem. Rev.* **122,** 9331–9356 (2022).
12. Martin, E. W., Hopkins, J. B. & Mittag, T. Small-angle X-ray scattering experiments of monodisperse intrinsically disordered protein samples close to the solubility limit. *Methods Enzymol.* **646,** 185–222 (2021).
13. 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).
14. Gomes, G.-N. W., Krzeminski, M., Namini, A., Martin, E. W., Mittag, T.,

Head-Gordon, T., Forman-Kay, J. D. & Gradinaru, C. C. Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET. *J. Am. Chem. Soc.* **142,** 15697–15710 (2020).
15. 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* **555,** 61–66 (2018).
16. Ruff, K. M., Choi, Y. H., Cox, D., Ormsby, A. R., Myung, Y., Ascher, D. B., Radford, S. E., Pappu, R. V. & Hatters, D. M. Sequence grammar underlying the unfolding and phase separation of globular proteins. *Mol. Cell* **82,** 3193–3208.e8 (2022).
17. 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
18. Joseph, J. A., Reinhardt, A., Aguirre, A., Chew, P. Y., Russell, K. O., Espinosa, J. R., Garaizar, A. & Collepardo-Guevara, R. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat Comput Sci* **1,** 732–743 (2021).
19. 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.* **118,** (2021).
20. 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).
21. 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). doi:10.1038/s41586-023-07004-5
22. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).
23. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D. & Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630,** 493–500 (2024).
24. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R.,

Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J. & Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373,** 871–876 (2021).

25. Baek, M., Anishchenko, I., Humphreys, I. R., Cong, Q., Baker, D. & DiMaio, F. Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv* 2023.05.24.542179 (2023). doi:10.1101/2023.05.24.542179
26. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S. & Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379,** 1123–1130 (2023).
27. Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y. A., Mishra, C., Kim, C., Bartie, L. J., Nemeth, M., Hsu, P. D., Sercu, T., Candido, S. & Rives, A. Simulating 500 million years of evolution with a language model. *Science* eads0018 (2025).
28. Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T. J., Berenberg, D., Fisk, I., Zanichelli, N., Zhang, B., Nowaczynski, A., Wang, B., Stepniewska-Dziubinska, M. M., Zhang, S., Ojewole, A., Guney, M. E., Biderman, S., Watkins, A. M., Ra, S., Lorenzo, P. R., Nivon, L., Weitzner, B., Ban, Y.-E. A., Chen, S., Zhang, M., Li, C., Song, S. L., He, Y., Sorger, P. K., Mostaque, E., Zhang, Z., Bonneau, R. & AlQuraishi, M. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nat. Methods* **21,** 1514–1524 (2024).
29. Ruff, K. M. & Pappu, R. V. AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol.* **433,** 167208 (2021).
30. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. *arXiv [cs.LG]* (2020). at <http://arxiv.org/abs/2006.11239>
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. *arXiv [cs.CV]* (2021). at <https://github.com/CompVis/latent-diffusion>
32. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J. & Rombach, R. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv [cs.CV]* (2023). at <http://arxiv.org/abs/2307.01952>
33. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Sara Mahdavi, S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J. & Norouzi, M. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv [cs.CV]* (2022). at <http://arxiv.org/abs/2205.11487>
34. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I. & Chen, M. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv [cs.CV]* (2021). at <http://arxiv.org/abs/2112.10741>
35. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical

text-conditional image generation with CLIP latents. *arXiv [cs.CV]* (2022). at <http://arxiv.org/abs/2204.06125>
36. Emenecker, R. J., Guadalupe, K., Shamoon, N. M., Sukenik, S. & Holehouse, A. S. Sequence-ensemble-function relationships for disordered proteins in live cells. *bioRxiv* 2023.10.29.564547 (2023).
37. Aspromonte, M. C., Nugnes, M. V., Quaglia, F., Bouharoua, A., DisProt Consortium, Tosatto, S. C. E. & Piovesan, D. DisProt in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Res.* **52,** D434–D441 (2024).
38. Marsh, J. A. & Forman-Kay, J. D. Sequence Determinants of Compaction in Intrinsically Disordered Proteins. *Biophys. J.* **98,** 2383–2390 (2010).
39. Müller-Späth, 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.* **107,** 14609–14614 (2010).
40. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **105,** 962–974 (2013).
41. Hellenkamp, B., Schmid, S., Doroshenko, O., Opanasyuk, O., Kühnemuth, 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., Götz, 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., Hübner, 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., Röcker, C., Sanabria, H., Schlierf, M., Schröder, T., Schuler, B., Seidel, H., Streit, L., Thurn, J., Tinnefeld, P., Tyagi, S., Vandenberk, N., Vera, A. M., Weninger, K. R., Wünsch, 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* **15,** 669–676 (2018).
42. Dhanasekaran, R., Deutzmann, A., Mahauad-Fernandez, W. D., Hansen, A. S., Gouw, A. M. & Felsher, D. W. The MYC oncogene - the grand orchestrator of cancer growth and immune evasion. *Nat. Rev. Clin. Oncol.* **19,** 23–36 (2022).
43. Flores, E., Camacho, A. R., Cuevas-Zepeda, E., McCoy, M. B., Yu, F., Staller, M. V. & Sukenik, S. Correlating disordered activation domain ensembles with gene expression levels. *Biophys. Rep. (N. Y.)* **5,** 100195 (2025).
44. Eick, D. & Geyer, M. The RNA polymerase II carboxy-terminal domain (CTD) code. *Chem. Rev.* **113,** 8456–8490 (2013).
45. 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).
46. 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).
47. Post, C. B. & Zimm, B. H. Theory of DNA condensation: collapse versus

aggregation. *Biopolymers* **21,** 2123–2137 (1982).
48. Neira, J. L., Palomino-Schätzlein, M., Ricci, C., Ortore, M. G., Rizzuti, B. & Iovanna, J. L. Dynamics of the intrinsically disordered protein NUPR1 in isolation and in its fuzzy complexes with DNA and prothymosin α. *Biochim. Biophys. Acta: Proteins Proteomics* **1867,** 140252 (2019).
49. Goretzki, B., Wiedemann, C., McCray, B. A., Schäfer, S. L., Jansen, J., Tebbe, F., Mitrovic, S.-A., Nöth, J., Cabezudo, A. C., Donohue, J. K., Jeffries, C. M., Steinchen, W., Stengel, F., Sumner, C. J., Hummer, G. & Hellmich, U. A. Crosstalk between regulatory elements in disordered TRPV4 N-terminus modulates lipid-dependent channel activity. *Nat. Commun.* **14,** 4165 (2023).
50. 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., Ternette, N., Moritz, R. L., Prensner, J. R. & van Heesch, S. High-quality peptide evidence for annotating non-canonical open reading frames as human proteins. *bioRxivorg* 2024.09.09.612016 (2024). doi:10.1101/2024.09.09.612016
51. Pesce, F., Bremer, A., Tesei, G., Hopkins, J. B., Grace, C. R., Mittag, T. & Lindorff-Larsen, K. Design of intrinsically disordered protein variants with diverse structural properties. *Sci. Adv.* **10,** eadm9926 (2024).
52. Krueger, R., Brenner, M. P. & Shrinivas, K. Generalized design of sequence-ensemble-function relationships for intrinsically disordered proteins. *bioRxiv* 2024.10.10.617695 (2024). doi:10.1101/2024.10.10.617695
53. Chao, T.-H., Rekhi, S., Mittal, J. & Tabor, D. P. Data-Driven Models for Predicting Intrinsically Disordered Protein Polymer Physics Directly from Composition or Sequence. *ChemRxiv* (2023). doi:10.26434/chemrxiv-2023-wrnq1
54. Moses, D., Ginell, G. M., Holehouse, A. S. & Sukenik, S. Intrinsically disordered regions are poised to act as sensors of cellular chemistry. *Trends Biochem. Sci.* **48,** 1019–1034 (2023).
55. Arts, M., Satorras, V. G., Huang, C.-W., Zuegner, D., Federici, M., Clementi, C., Noé, F., Pinsler, R. & van den Berg, R. Two for one: Diffusion models and force fields for coarse-grained molecular dynamics. *arXiv [cs.LG]* (2023). at <http://arxiv.org/abs/2302.00600>
56. Janson, G. & Feig, M. Transferable deep generative modeling of intrinsically disordered protein conformations. *PLoS Comput. Biol.* **20,** e1012144 (2024).
57. Zhu, J., Li, Z., Zhang, B., Zheng, Z., Zhong, B., Bai, J., Hong, X., Wang, T., Wei, T., Yang, J. & Chen, H.-F. Precise generation of conformational ensembles for intrinsically disordered proteins via fine-tuned diffusion models. *bioRxiv* 2024.05.05.592611 (2024). doi:10.1101/2024.05.05.592611
58. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35,** 1026–1028 (2017).
59. Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M.,

Crozier, P. S., in ’t Veld, P. J., Kohlmeyer, A., Moore, S. G., Nguyen, T. D., Shan, R., Stevens, M. J., Tranchida, J., Trott, C. & Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271,** 108171 (2022).
60. McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L.-P., Lane, T. J. & Pande, V. S. MDTraj: a modern, open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109,** 1528–1532 (2015).
61. 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.* **19,** 5609–5620 (2023).
62. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv [stat.ML]* (2013). at <http://arxiv.org/abs/1312.6114v11>
63. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. in *Proceedings of the 32nd International Conference on Machine Learning* (eds. Bach, F. & Blei, D.) **37,** 2256–2265 (PMLR, 07--09 Jul 2015).
64. Song, J., Meng, C. & Ermon, S. Denoising Diffusion Implicit Models. *arXiv [cs.LG]* (2020). at <http://arxiv.org/abs/2010.02502>
65. Lotthammer, J. M. & Holehouse, A. S. Disentangling folding from energetic traps in simulations of disordered proteins. *bioRxiv* 2024.11. 06.622270 (2024). doi:10.1101/2024.11.06.622270
66. Kikhney, A. G., Borges, C. R., Molodenskiy, D. S., Jeffries, C. M. & Svergun, D. I. SASBDB: Towards an automatically curated and validated repository for biological scattering data. *Protein Sci.* **29,** 66–75 (2020).
67. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51,** D523–D531 (2023).
68. Vitalis, A. & Pappu, R. V. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* **30,** 673–699 (2009).
69. Vitalis, A. & Pappu, R. V. in *Annual Reports in Computational Chemistry* (ed. Wheeler, R. A.) **5,** 49–76 (Elsevier, 2009).
70. Mao, A. H. & Pappu, R. V. Crystal lattice properties fully determine short-range interaction parameters for alkali and halide ions. *J. Chem. Phys.* **137,** 064104 (2012).
71. 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).
72. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17,** 261–272 (2020).

73. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. Array programming with NumPy. *Nature* **585,** 357–362 (2020).
74. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. & Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv [cs.LG]* (2019). at <http://arxiv.org/abs/1912.01703>

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

VERSION 1 (2025-02-14)

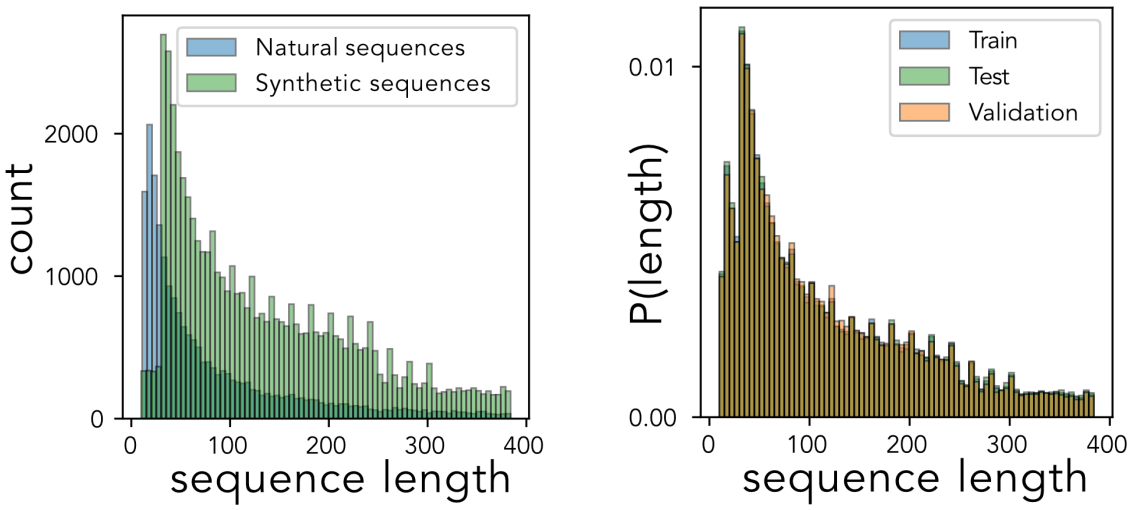


Figure S1: Our dataset of 70,563 consists of both natural sequences (N=20,349) and synthetic sequences designed by varying sequence features known to affect IDR ensembles (N=50,214) (left). The sequences were clustered and split into train (N=49,423), validation (N=10,703), and test (N=10,437) sets while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

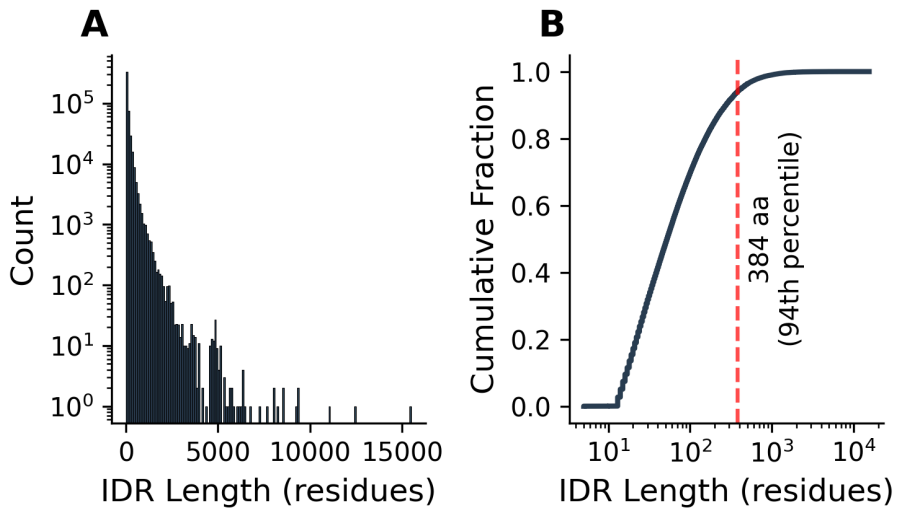
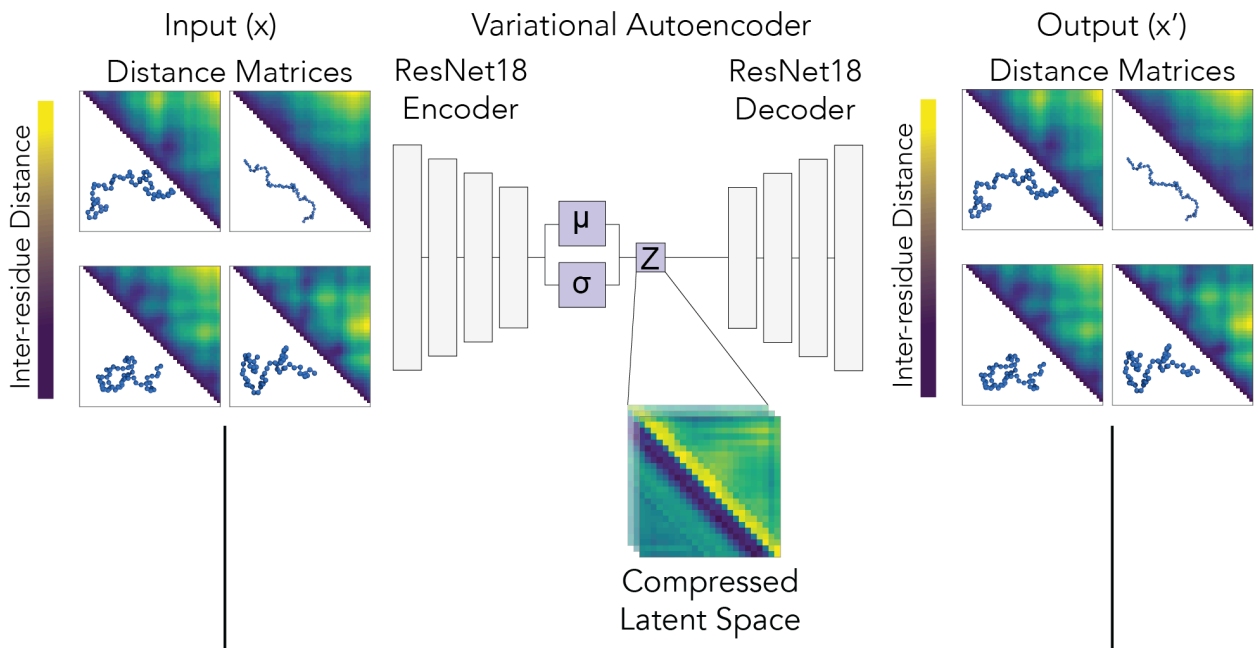


Figure S2: Disordered protein length distributions across a collection of model organism proteomes. **A.** Histogram of the protein sequence length distribution. **B.** A cumulative density function of the length distribution presented in A. 94% of the IDRs across common model organisms fall between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



$$\mathcal{L} = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - D_{\mathrm{KL}}\left(q_\phi(z \mid x) \| p(z)\right)$$

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

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.

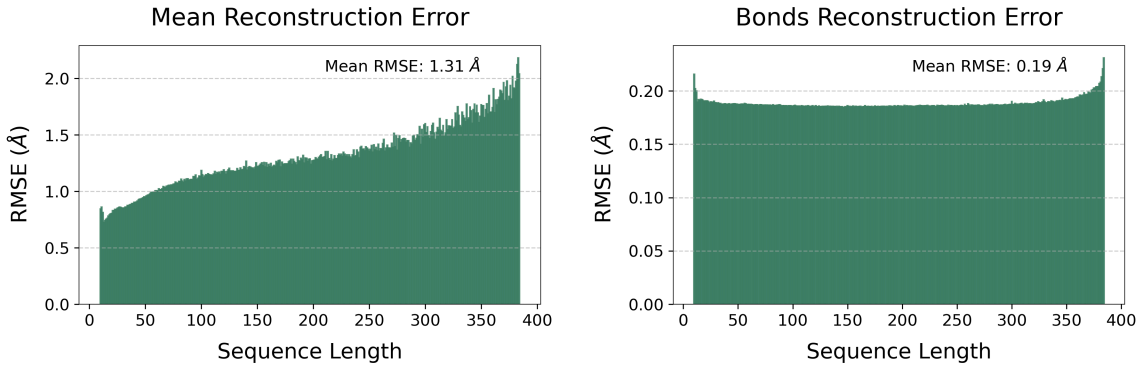


Figure S4: The variational autoencoder (VAE) is evaluated on a held-out test set of distance maps (N = 1,425,849). The VAE encoder spatially compresses these distance maps and is then reconstructed by the VAE decoder from their compressed representations. The root mean squared error (RMSE) is computed 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).

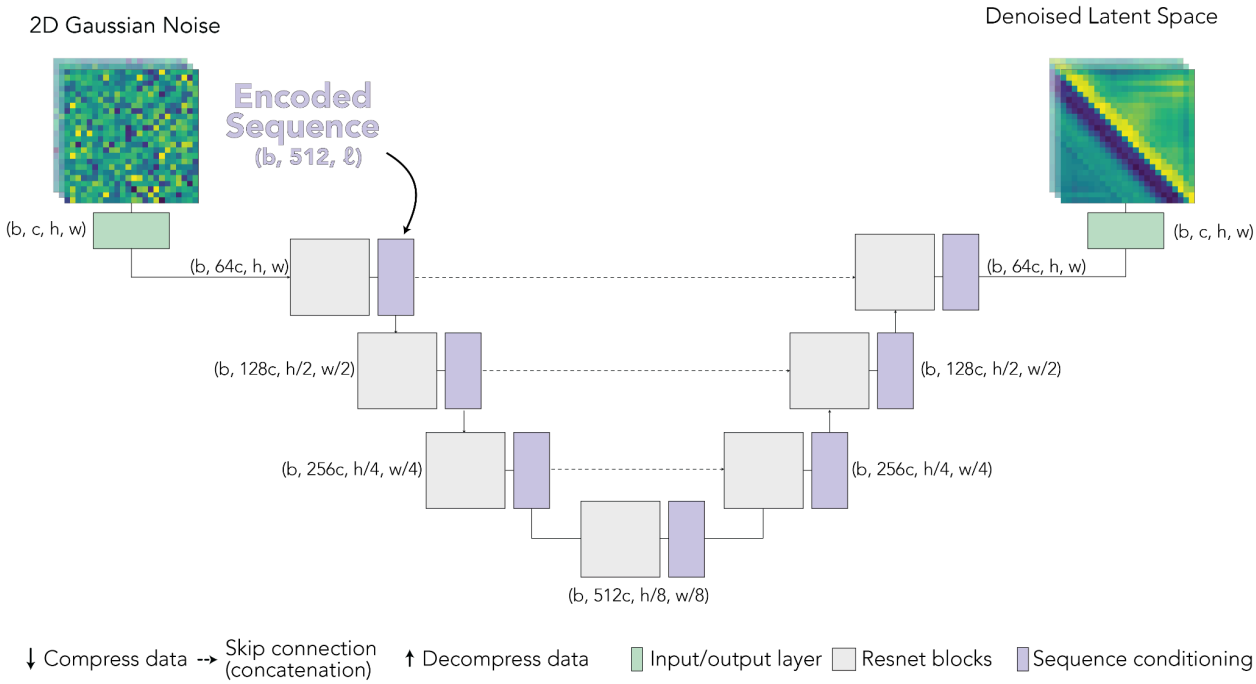


Fig. S5: Extended schematic representation of the Unet backbone of the STARLING discrete time denoising-diffusion probabilistic model (DDPM). The DDPM starts by taking a batch size (b) of 2D Gaussian noise with one channel (c) of features and spatial dimensions concordant with the VAE latent space (h,w). 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 (8x*8x). 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 ℓ x ℓ 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.

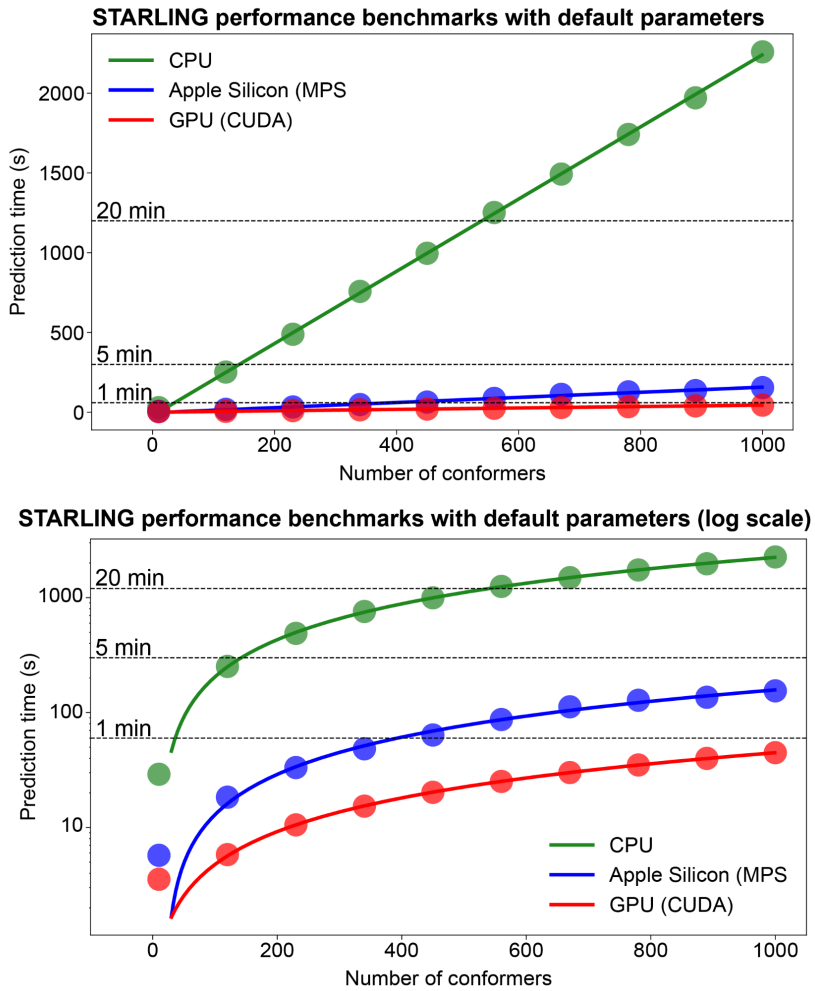


Fig. S6. STARLING benchmarking on different types of hardware. Runtime depends linearly on the number of conformers (although sequence length has no impact on runtime). The top and bottom panels show identical data in log vs. linear space. On GPUs and Apple Silicon (facilitated by CUDA and MPS, respectively), 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).

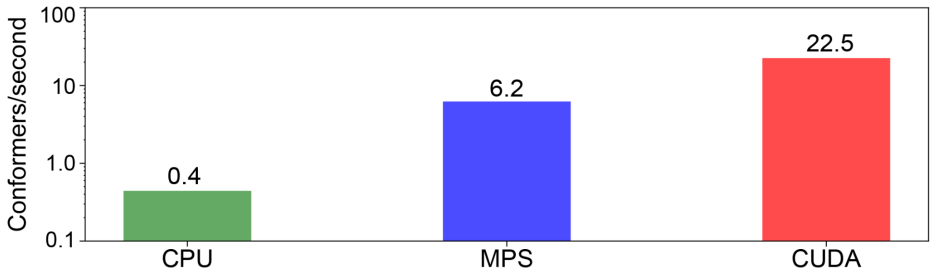


Fig. S7. Number of conformers generated per second on different hardware types. Specific hardware tested here are Nvidia A4500 (GPU), Macbook Pro M3 Max CPU (MPS), and Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz (CPU).

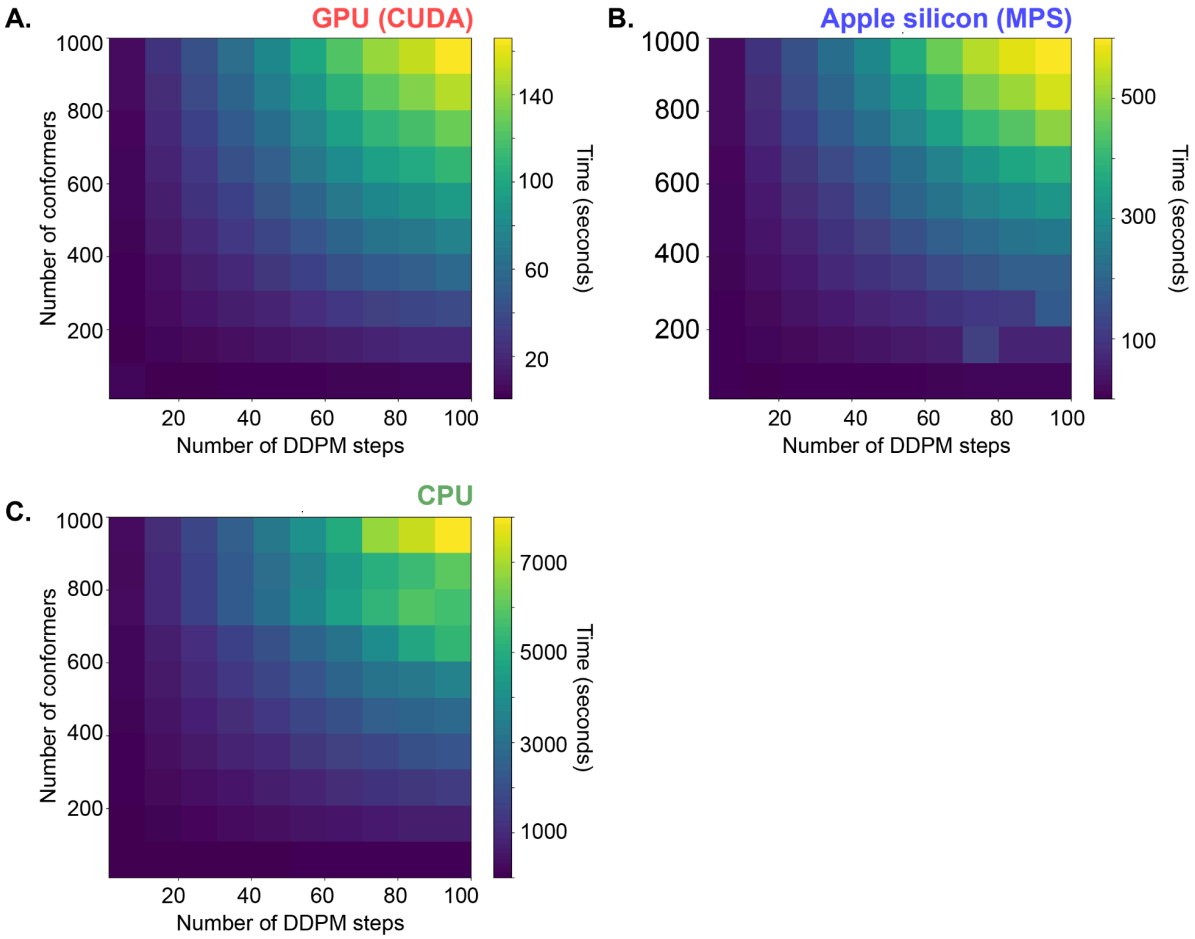


Fig. S8. Sweet of runtime vs. ensemble size and number of denoising steps on different hardware types. Specific hardware tested here are Nvidia A4500 (GPU), Macbook Pro M3 Max CPU (MPS), and Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz (CPU).

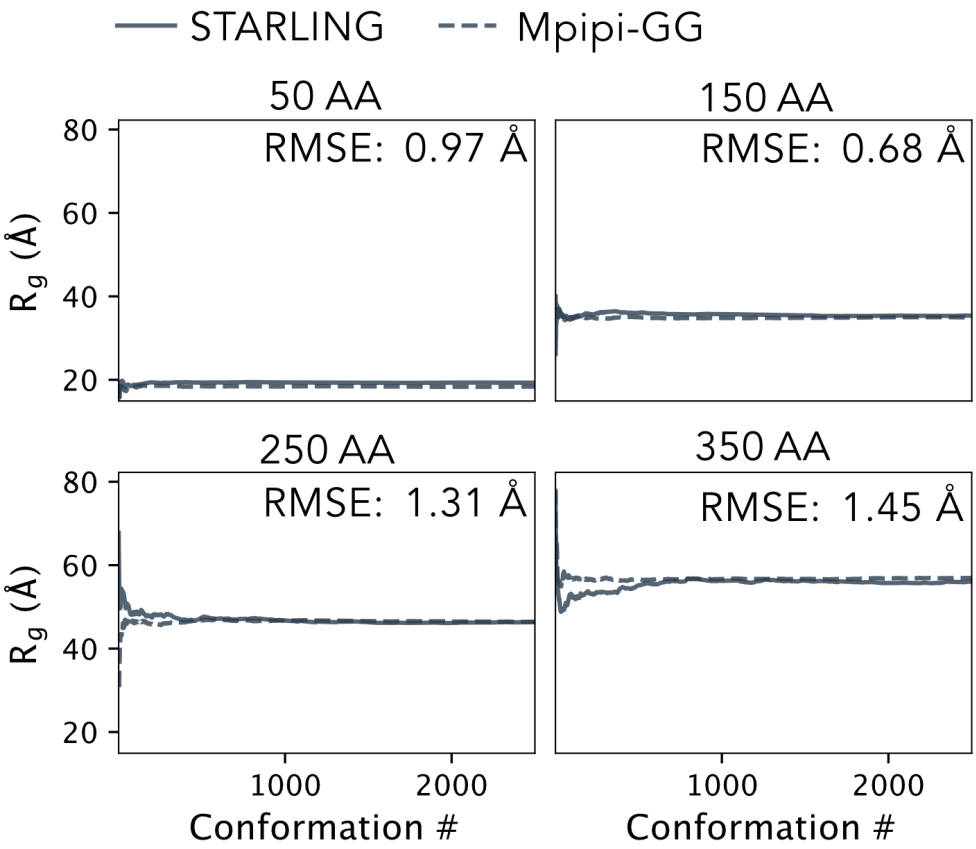


Fig. S9. Radius of gyration (R_g) convergence plot as a function of the size of the ensemble (number of conformations) for varied IDR lengths (50, 150, 250, and 350 residues in length). The dashed line is the ensemble-average R_g value taken from long Mpipi-GG simulations. STARLING-derived R_g values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

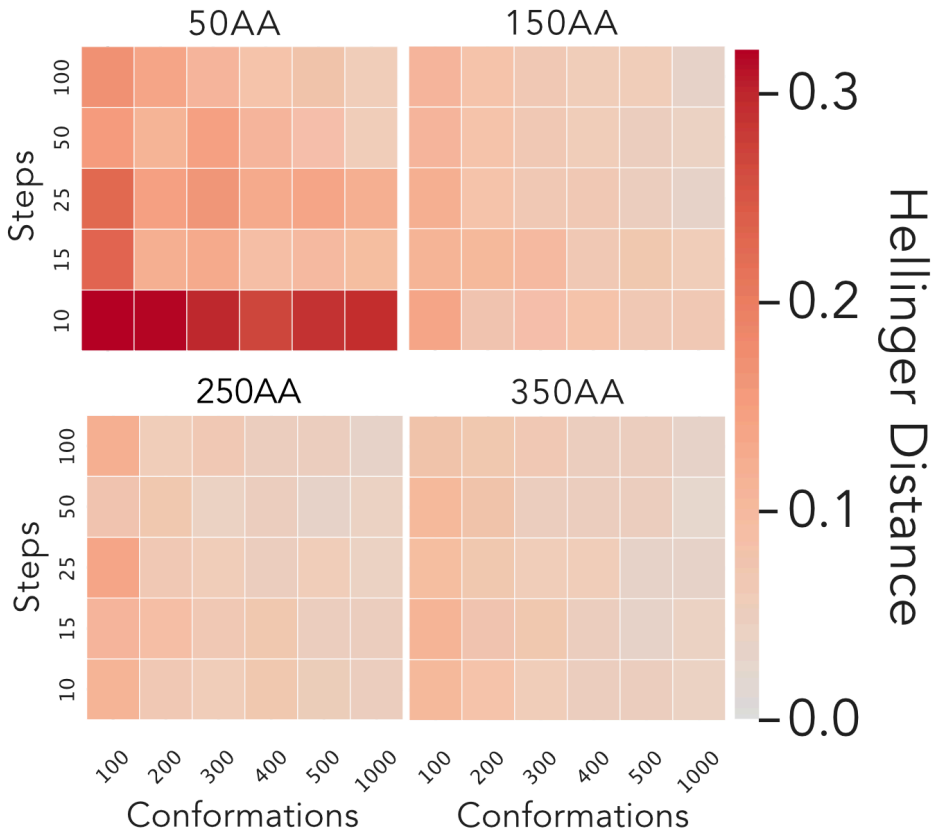


Fig. S10. Convergence plot depicting the relationship between Hellinger distance, the number of conformations sampled from STARLING, and the number of denoising steps for various IDR lengths (50, 150, 250, and 350 residues). The Hellinger distance quantifies the similarity between two distributions, 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.

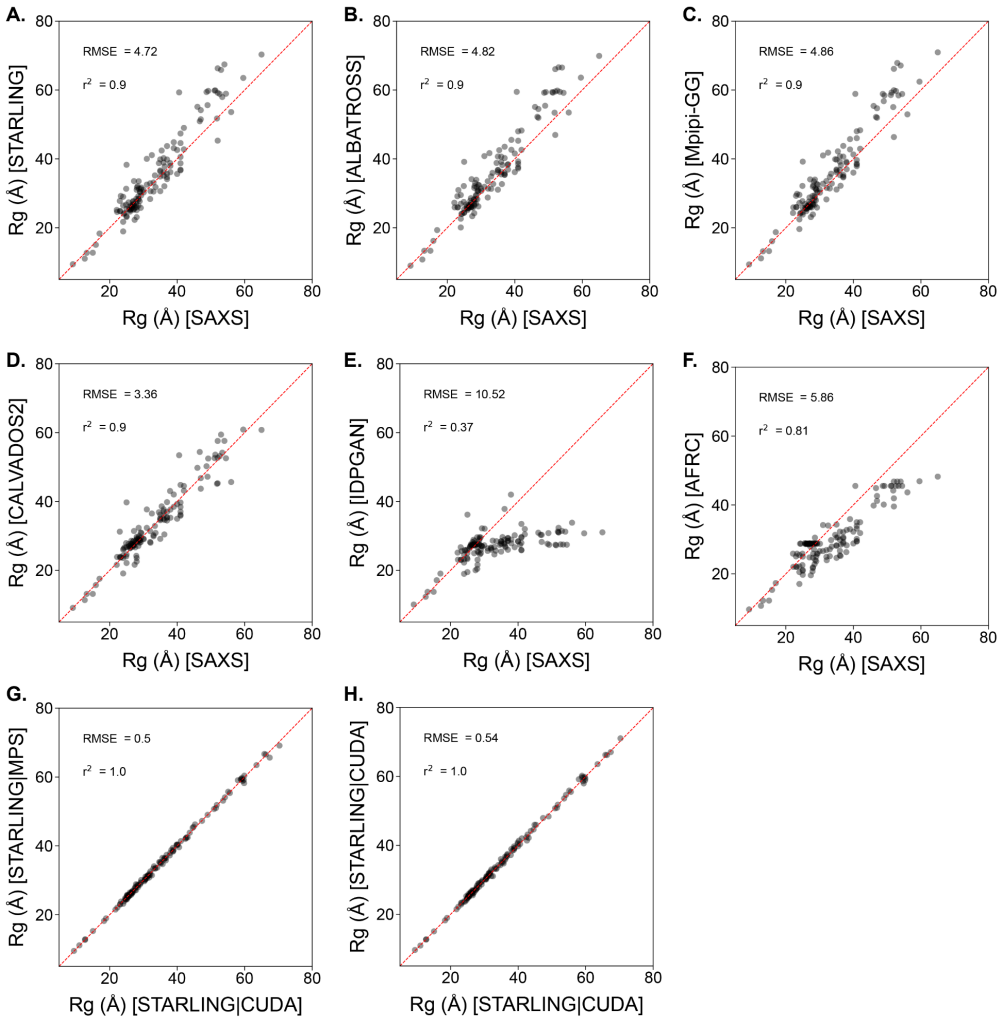


Fig. S11. Comparison of experimental R_g values obtained from 133 different experimental datasets with various state-of-the-art tools for ensemble prediction: **(A)** STARLING (this work), **(B)** ALBATROSS, a deep-learning-based predictor that can only predict average ensemble values, **(C)** Mpipi-GG, 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 **(F)**.

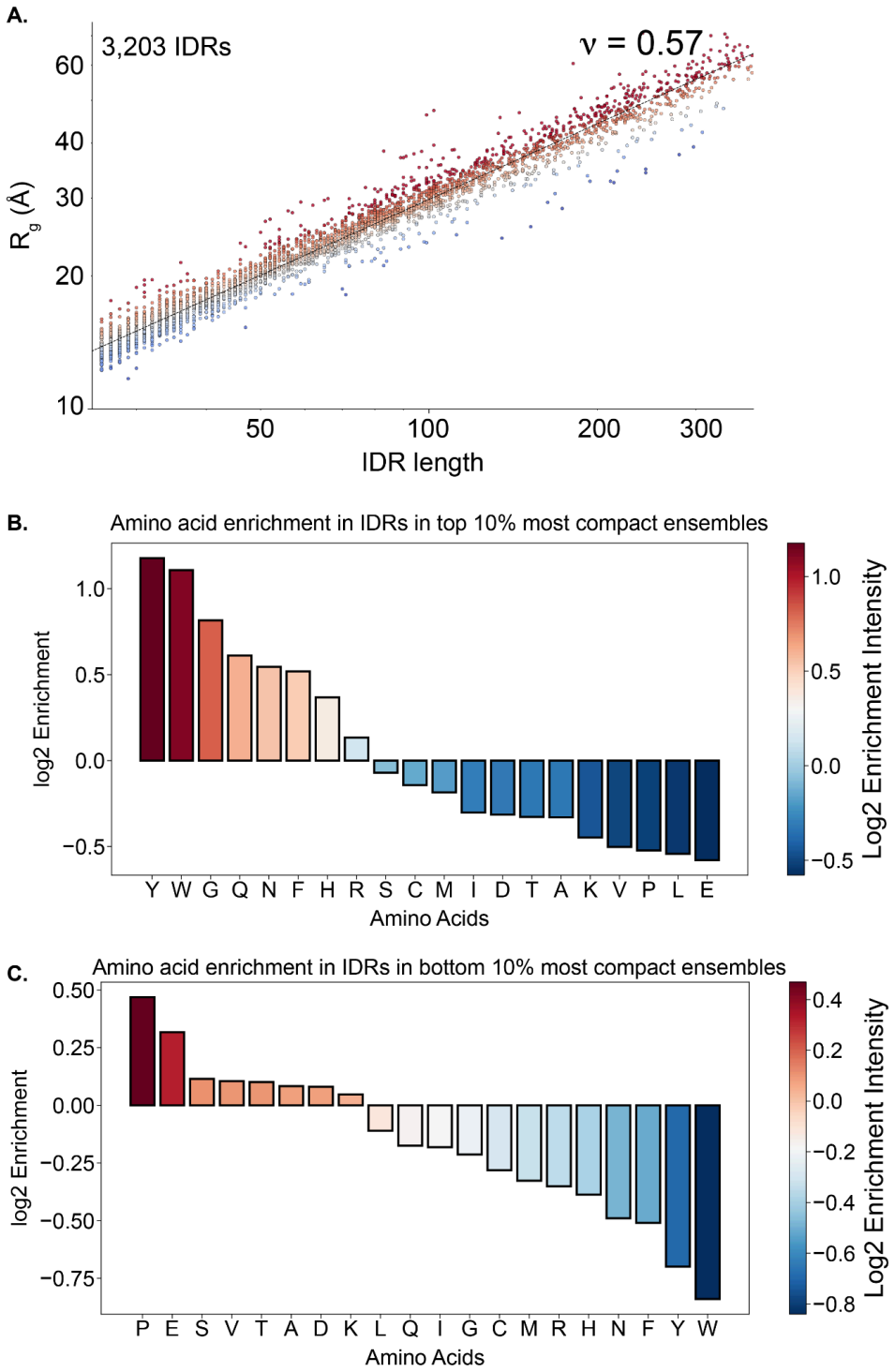


Figure S12. STARLING ensembles for DisProt. We predicted all ensembles in ~18 hours using a desktop computer with an NVIDIA A4500 GPU. **(A)** Fitting ensemble dimensions vs. length reveals an apparent scaling exponent of 0.57, indicating these IDRs are relatively expanded. However, we note an unavoidable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

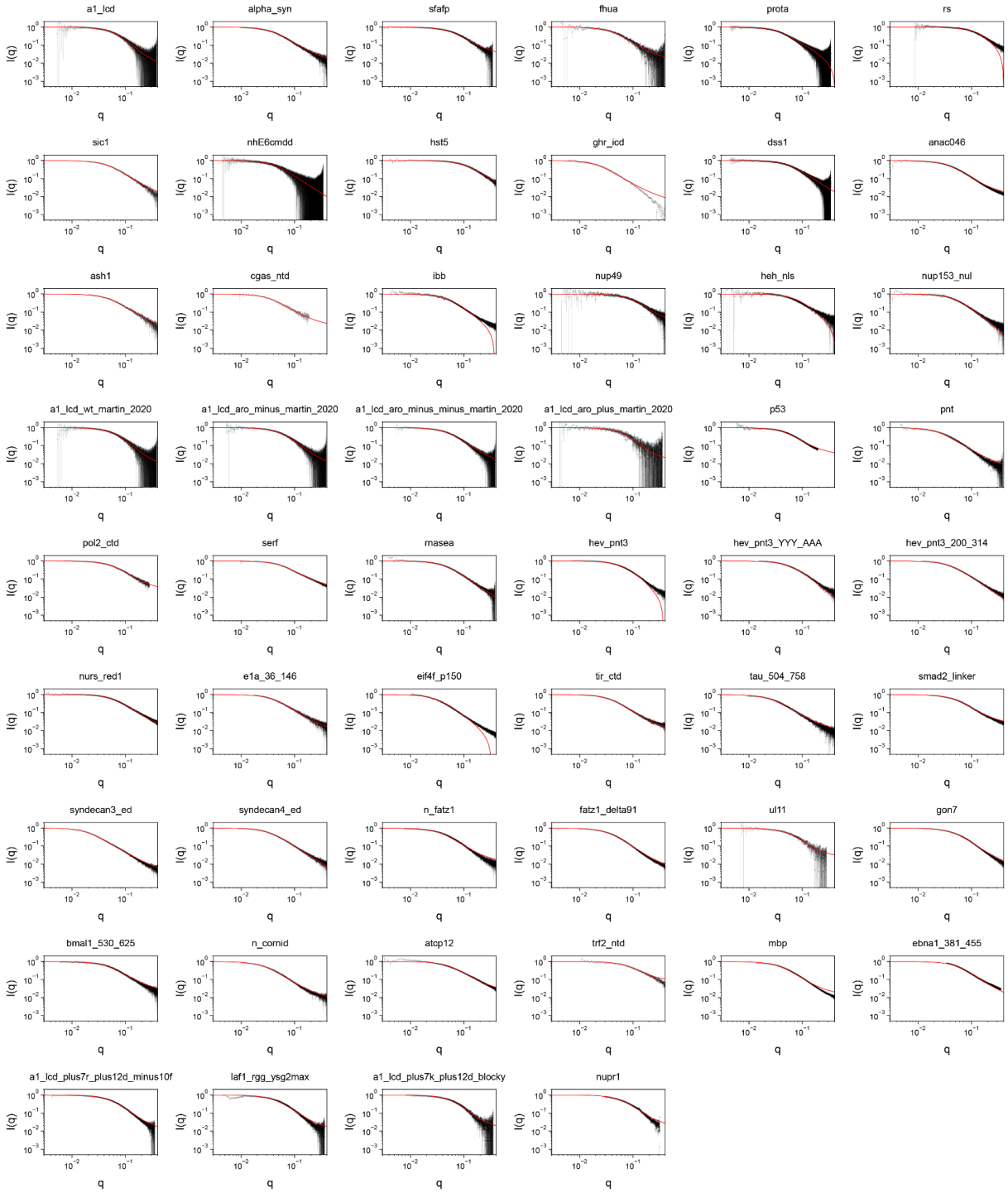


Figure S13. Comparison of STARLING-derived scattering data with experimental profiles from 52 sequences and datasets. Black is experimental and red in STARLING-derived scattering data. STARLING-derived curves were back-calculated using FOXS.

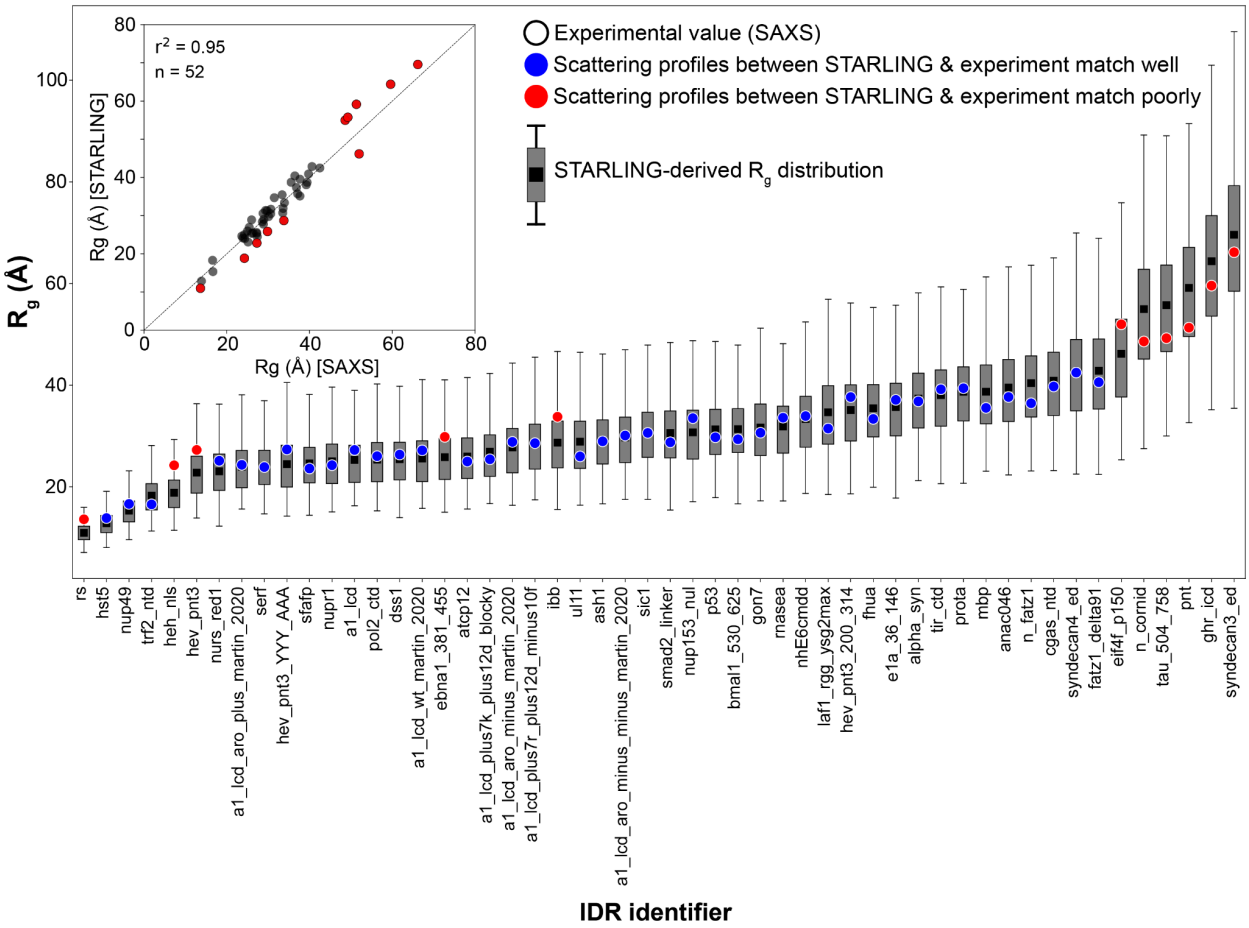


Figure S14. Comparison of SAXS-derived radii of gyration vs. STARLING-derived radii of gyration for the same 52 sequences as described in **Fig. S13**. Sequences in which experimental data is represented as a red dot are those where scattering profiles were in sub-optimal agreement with STARLING-derived 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).

A.

```
>sGrich
GSGSCSGGYGSERGGGYGSERGGGYGSERGGGYGSERGGGYGSQRSGGGYGGSQRSSYGSGSCTLGPR
>dKh
GSGSCAMGGGPGPGTDFTSDQADFPDTLFQEFEPPAPRPGLAGGRPGDAALLSAAYGRRRLLCTLGPR
>dGrich
GSGSCGPRTGLEGAGMAGGSGQQKRVFDGQSGPQDLGEAYRPLNHDGGDGGNRYSVIDRIQECTLGPR
>sPTBP
GSGSCRPDLPSGDSQPSLDQTMAAAFGLSVPNVHGALAPLAIPSAAAAAAAAGRIAIPGLAGCTLGPR
>sCh
GSGSCKKAQPEMNDKDDNESGNEDAEENHDDEEDENEEEDRQVDQASKNKESKRKAQNKREDCTLGPR
>sNrich
GSGSCLDQEDNNGPLLIKTANNLIQNNSNMLPLNALHNAPPMHLNEGGISNMRVNDSLPSNTCTLGPR
>sKh
GSGSCKPRRLSKLRRSKKPADEENNAASQDPTVEATQERGQASEDPENAANNAKQAKPTSDDCTLGPR
>dCh_plus
GSGSCQTPLKRIKVKTPGKSGAAAREGSVVSGTDGPTQTGKPERRKRLNPPKDKLIDMDDADCTLGPR
>dCh_minus
GSGSCMGLPTGMEEKEEGTDESEQKPVVQTPAQPDDSAEVDSAALDQAESAKQQGPILTKHGCTLGPR
>sKl
GSGSCAPEGVFKLPAPPKEIKKSEKSEKSSDSSNDKKEVESTTKTAATTTTTKKGTDNNTQFCTLGPR
>sNh_plus
GSGSCPEEIETRKKDRKNRREMLKRTSALQPAAPKPTHKKPVPKRNVGAERKSTINEDLLPPCTLGPR
>dArich
GSGSCLPTGAEGRDSSKGEDSAEETEAKPAVVAPAPVVEAVSTPSAAFPSDATAEQGPILTKCTLGPR
>dErich
GSGSCEGEMFGVMCEENSVVLSVEQPAELKEVADVSPPPTTRNHTIEMKPPLSAQQSECNPVCTLGPR
>dTRBP
GSGSCLEPALEDSSSFSPLDSSLPEDIPVFTAAAAATPVPSVVLTRSPPMELQPPVSPQQSECTLGPR
>dKl
GSGSCLPTGMDQKTTELMKVDEPVSTQETLPPVIKMEPEPVPITETPTDENARQQGPILTKHCTLGPR
>sNh_minus
GSGSCKKKLQDTPSEPMEKDPAEPETVPDGETPEDENPTEEGADNSSAKMEEEEEEEEEEEECTLGPR
```

B.

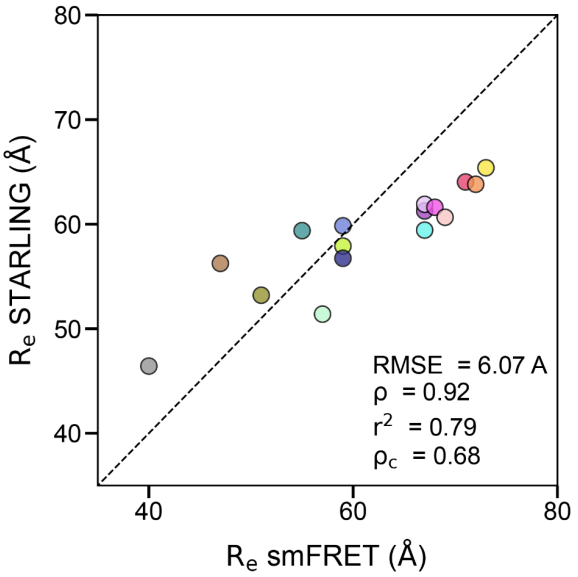


Fig. S15 Sequences and overall correlation obtained for sequences investigated by smFRET by [7]. **(A)** The amino acid sequence of the sequences investigated. **(B)** Correlation between STARLING-derived end-to-end distances and smFRET-derived end-to-end distances. Agreement between STARLING and experiment here is assessed in various ways, including root mean squared error (RMSE), Spearman correlation coefficient (ρ), the square of the Pearson correlation coefficient (r^2), and the concordance correlation coefficient ($ρ_c$). We note that STARLING appears to underestimate end-to-end distances for more expanded chains here, although we do not take dyes into explicit consideration, and it is possible that for the expanded negatively charged sequence, a small amount of (negatively-charged) dye:protein repulsions could account for the slight underestimation here.

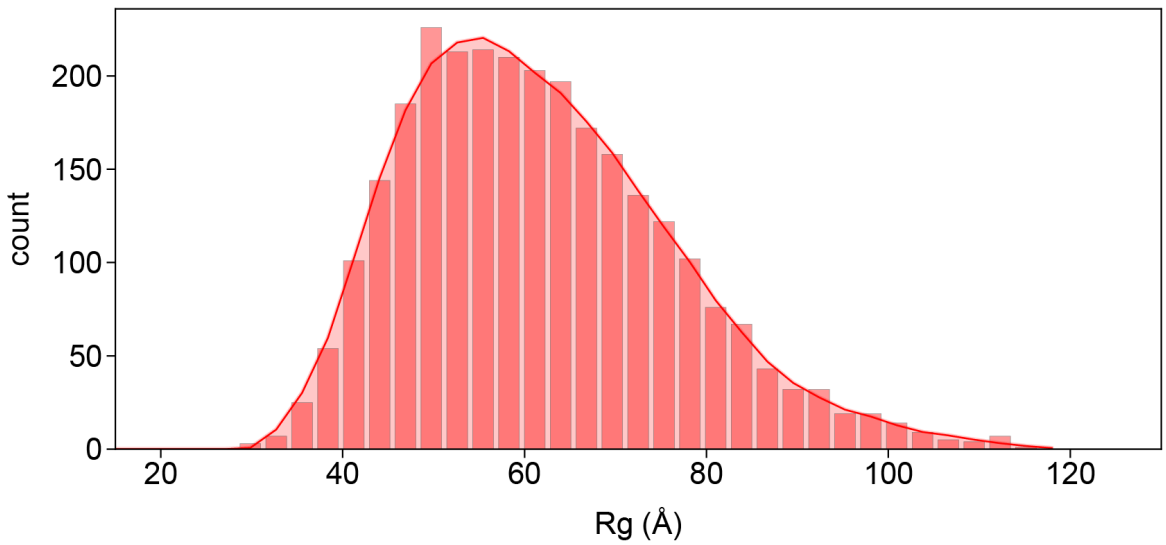


Fig. S16 R_g distribution for cMyc$^{1-361}$ obtained from STARLING ensembles. We predict an average Rg of ~60 Å. We note this value may be an over-estimate given cMyc$^{1-361}$ is predicted to contain several regions of transient secondary structure, which may reduce the R_g, such that we suggest a value closer to 55 Å may be a more accurate reflection on the expected dimensions for monomeric cMyc$^{1-361}$.

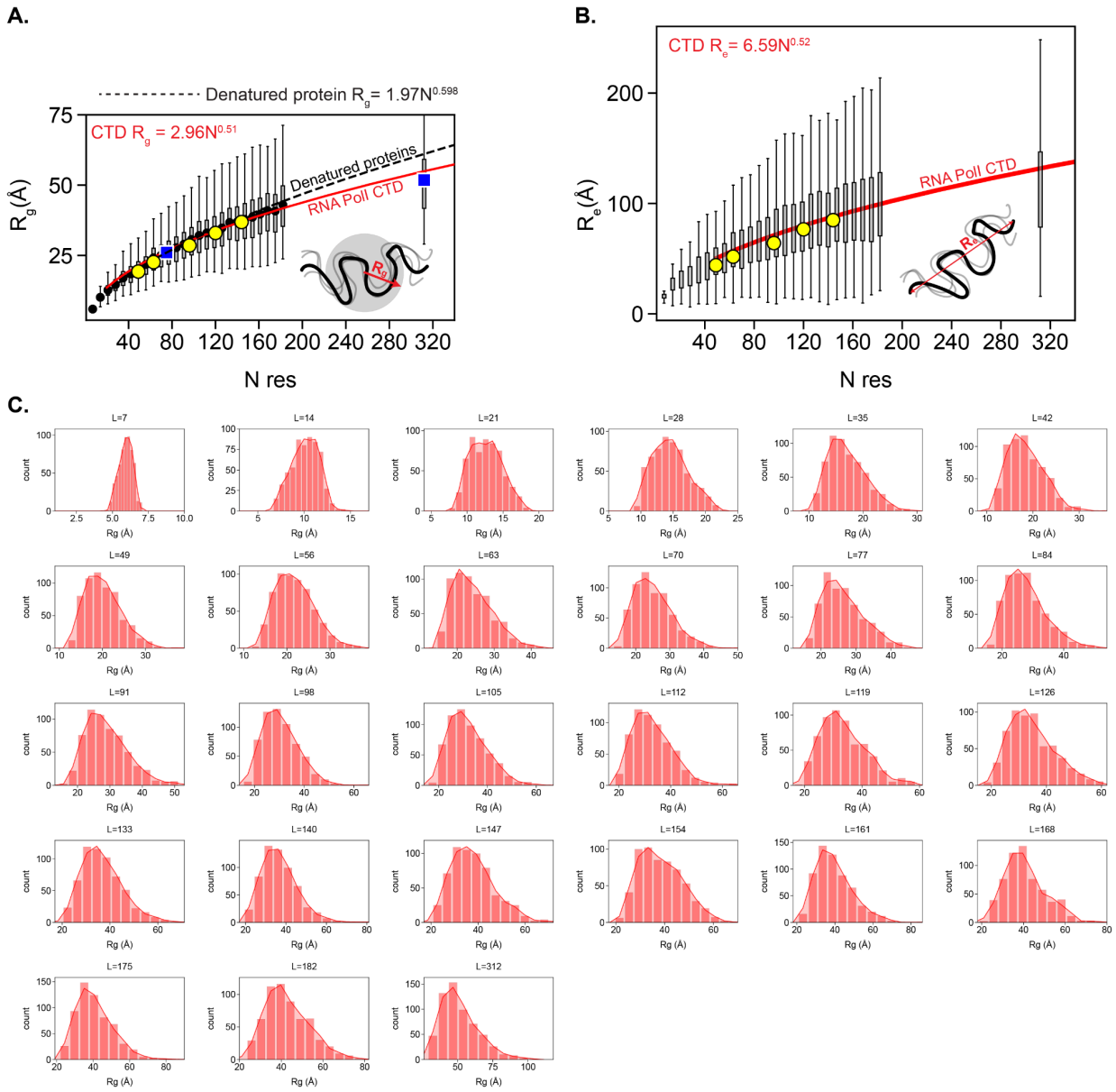


Fig. S17 Conformational behavior of RNA polymerase II CTD was examined using STARLING ensembles **(A)**. The same figure is shown for **Fig. 4D**, with an additional polymer scaling line here derived from fully denatured proteins measured by Kohn et al., also superimposed (black dashed line)[8]. This 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.

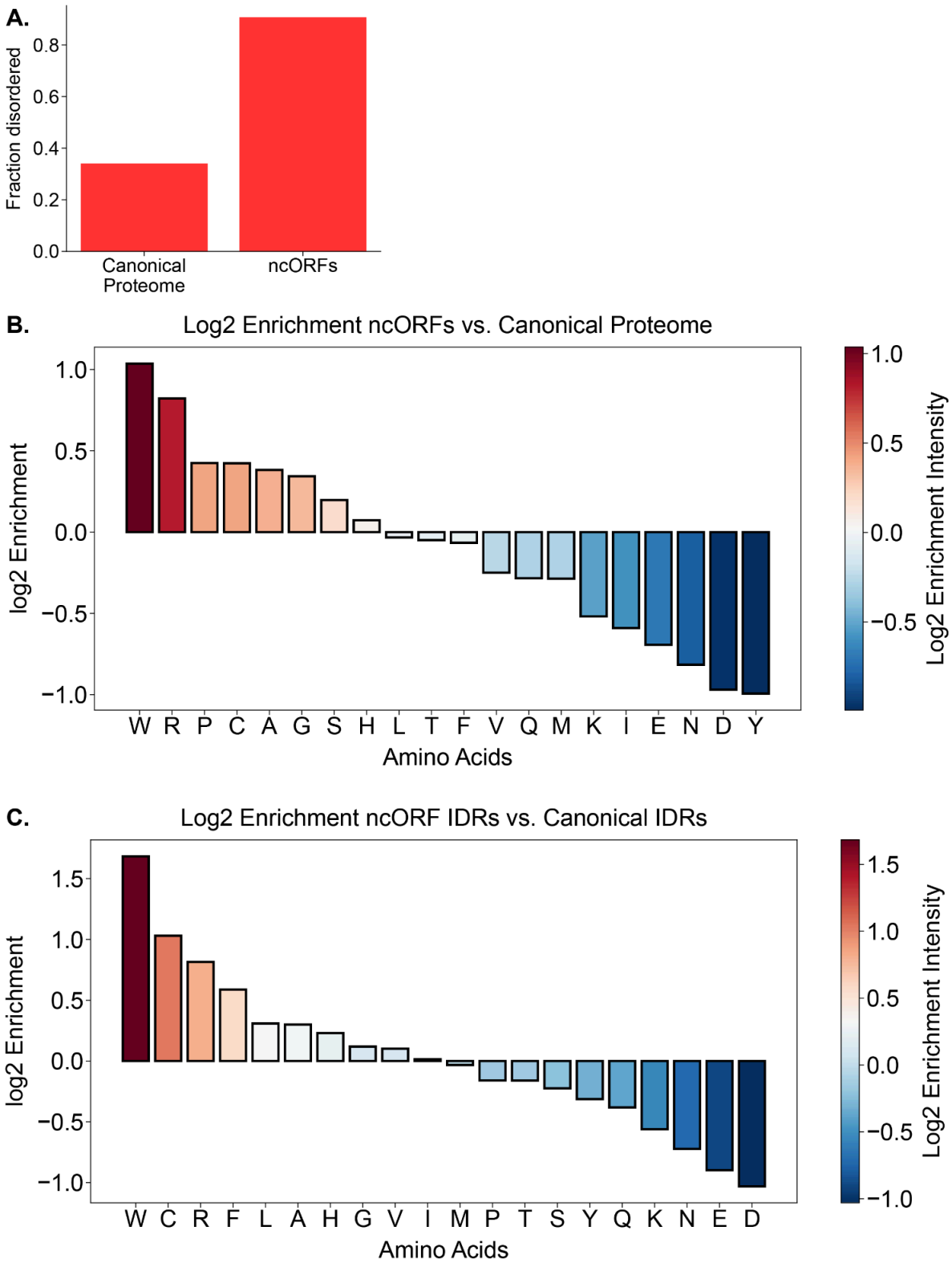

A.
Fraction disordered
0.0
0.2
0.4
0.6
0.8
Canonical Proteome
ncORFs
B.
Log2 Enrichment ncORFs vs. Canonical Proteome
log2 Enrichment
1.0
0.5
0.0
−0.5
−1.0
W R P C A G S H L T F V Q M K I E N D Y
Amino Acids
Log2 Enrichment Intensity
1.0
0.5
0.0
−0.5
C.
Log2 Enrichment ncORF IDRs vs. Canonical IDRs
log2 Enrichment
1.5
1.0
0.5
0.0
−0.5
−1.0
W C R F L A H G V I M P T S Y Q K N E D
Amino Acids
Log2 Enrichment Intensity
1.5
1.0
0.5
0.0
−0.5
−1.0

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 regions[14]. The disorder was predicted using metapredict V3[15]. **(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 disproportionally 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.

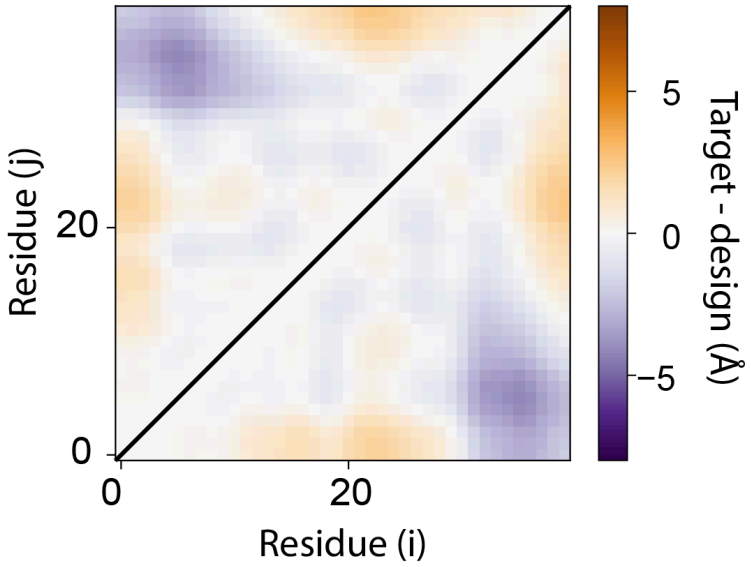


Fig. S19 Symmetrical matrix showing target designed sequence subtracted from the target sequence, yielding a difference map. All inter-residue errors are < 4 Å, with an RMSE error of 1.3 Å.

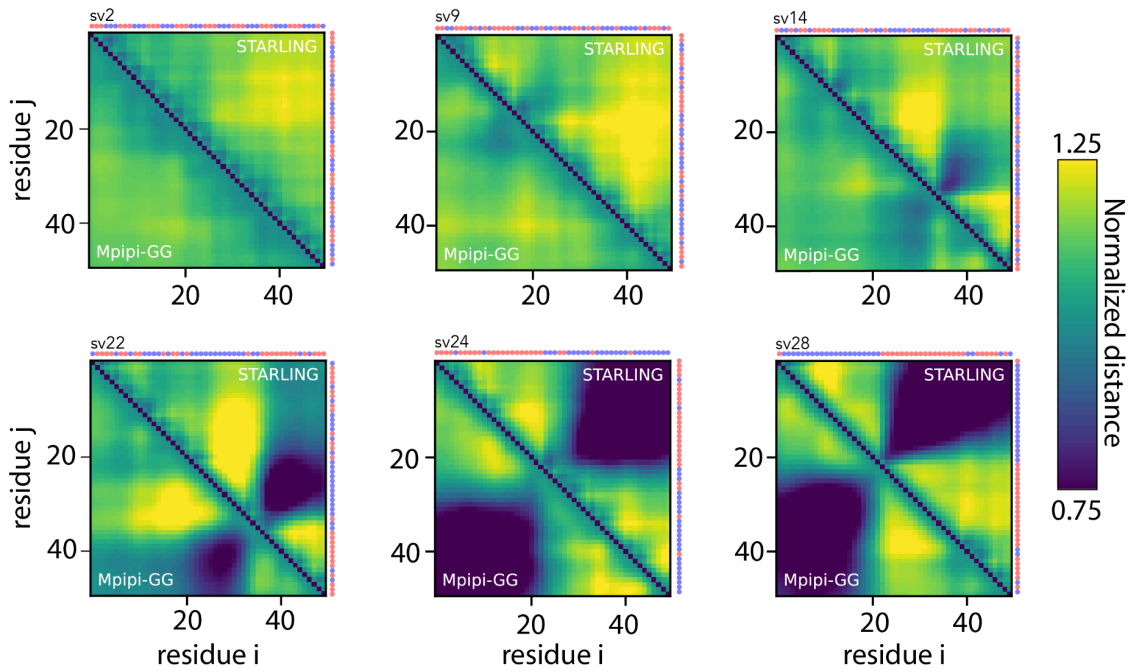


Fig. S20 Das-Kappa sequences in STARLING vs. Mpipi-GG. We computed normalized distance maps for several of the iconic "Das sequence"[18], revealing that STARLING ensembles offer good reproduction of the inter-residue distances.

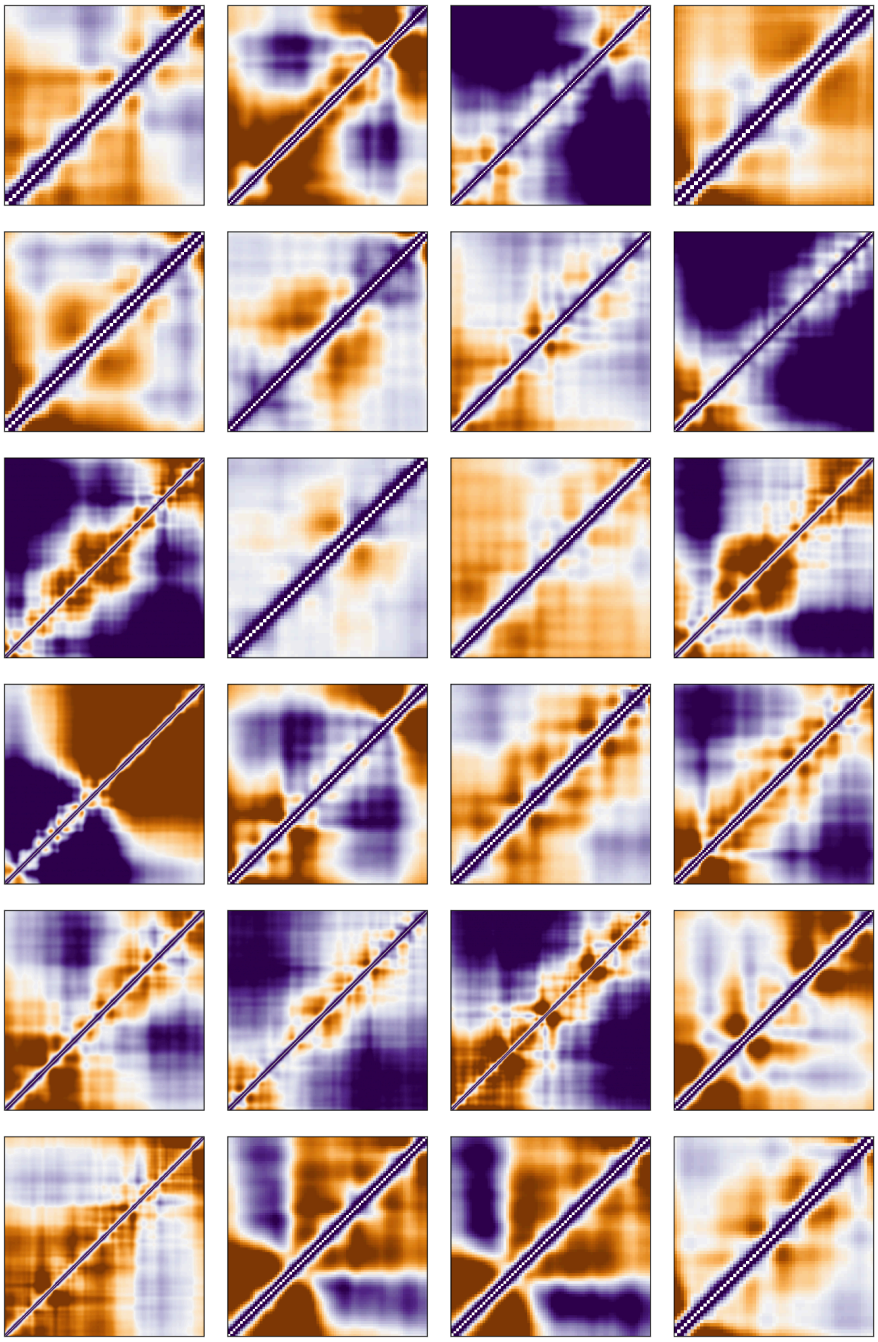


Fig. S21 Example of distance maps from DisProt

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.

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 ERα 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., 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.* **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., Ternette, N., Moritz, R. L., Prensner, J. R. & van Heesch, S. High-quality peptide evidence for annotating non-canonical open reading frames as human proteins. *bioRxivorg* 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. 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).