SUPPLEMENTARY INFORMATION FOR:

DIRECT PREDICTION OF INTRINSICALLY DISORDERED PROTEIN CONFORMATIONAL PROPERTIES FROM SEQUENCE

Version 1.0 [2023-05-08]

Jeffrey M. Lotthammer[1,2,*], Garrett M. Ginell[1,2,*], Daniel Griffith[1,2,*], Ryan J. Emenecker[1,2], Alex S. Holehouse[1,2]

1 - Department of Biochemistry and Molecular Biophysics Washington University School of Medicine, St. Louis, MO, USA
2 – Center for Biomolecular Condensates, Washington University in St. Louis, St. Louis, MO, USA

* These authors contributed equally
Correspondence: alex.holehouse@wustl.edu

1. Extended methods

Mpipi fine-tuning

We first sought to fine-tune the Mpipi force field to address several minor discrepancies observed for IDR single-chain behavior between simulations and experiments. Mpipi is a one-bead-per-residue coarse-grained force field that was parameterized via a bottom-up, data-driven approach using statistics obtained from the PDB coupled with quantum mechanical calculations and all-atom simulations to derive parameters for a Wang-Frenkel (WF) potential **(Equation 1.1-1.2)** [1,2].

$$\phi(r) \;=\; \epsilon\,\alpha\,([\tfrac{\sigma}{r}]^{2\mu} - 1)\;([\tfrac{R}{r}]^{2\mu} \;-\; 1)^{(2\nu)} \qquad \textbf{Equation 1.1}$$

where

$$\alpha \;=\; 2\nu(\tfrac{R}{\sigma})^{2\mu} \times \left(\frac{1+2\nu}{2\nu\left[(\frac{R}{\sigma})^{2\mu}-1\right]}\right)^{(2\nu+1)} \qquad \textbf{Equation 1.2}$$

Where $R = 3\sigma$ and $\nu = 1$, as defined in the original Mpipi paper.

The WF potential provides a computationally convenient (efficient to compute, intercepts with 0 at long intermolecular distances) closed-form alternative to the more commonly used Lennard-Jones potential. Given the data-driven approach used for parameterization, Mpipi benefits from explicitly encoded inter-residue interaction values (i.e., $\varepsilon_{i,j}$ values) for all unique pairs of amino acids. This is in contrast to most other one-bead-per-residue force fields, where intrinsic residue-specific interaction strengths (i.e., ε_i or λ_i) are defined, and inter-residue interaction energies are then computed via so-called ‘mixing rules’. While Mpipi offers improved flexibility for capturing chemically complex interactions, the model also has many more parameters than most conventional force fields (i.e., $[n^2 + n]/2$ interaction parameters for a model with n amino acids). As such, despite the excellent accuracy of the original model, we sought to determine if Mpipi could be further improved.

We focused on four specific groups of pairwise interactions to fine-tune Mpipi. In doing so, we developed an augmented version we refer to as Mpipi-GG. In particular, we strengthened Gly:Gly and Gly:Ser interactions, weakened aromatic:charge interactions, increased the excluded volume of proline residues, and reparameterized aliphatic residues to have increased hydrophobicity.

The adjustment to proline was motivated by the observation that upon simulation with the original Mpipi parameters, many proline-rich IDRs were too compact compared with experiments (**Fig. S1A, B**). Proline predominantly drives IDR expansion via backbone restrictions and favorable solvation[3–6]. We reasoned that tuning the proline σ parameter (i.e., its excluded volume) would enhance its expansion-driving effects. To this end, after systematically titrating potential σ values and comparing the outcome of altering the parameters with all-atom simulations (**Fig. S1C**), we increased the proline σ by 33% for all pair-wise proline interactions, as shown in **Fig. S1B**. Applying this fix improved accuracy with respect to proline-rich IDRs with minimal loss of accuracy for other IDRs (**Fig. S1D**).

Following our adjustment of proline, we examined several polar-rich homo- or dipolymeric tracts for which experimental data have previously been obtained; poly-(GS), poly-(G), poly-(S), and poly-(Q). Previous work established that sufficiently long polyglutamine (poly-(Q)) tracts from compact globules, consistent with results from Mpipi[7]. However, we noticed that both poly-(GS) and poly-(G) scaled as a self-avoiding random walk (ν = ~0.60), despite the fact experimental work has suggested poly-(GS) behaves akin to a Gaussian chain (ν = ~0.5-0.55) and poly-(G) forms compact ensembles (ν = ~0.4) [8–11] (**Fig. S2A**). To address this discrepancy, we performed a titration series for poly-(G) chains. We tuned the G:G interactions by titrating the strength of the glycine-glycine attractive parameter in the WF potential ($\varepsilon_{G,G}$) for a poly-(G)$_{80}$ chain (**Fig. S2B**). Fitting these data to a coil-to-globule transition, we extracted the interaction strength (2.19x the original $\varepsilon_{G,G}$) that gave an apparent scaling exponent of 0.39, in line with previous experiments (**Fig. S2C**)[11,12]. Having established the correction factor for $\varepsilon_{G,G}$, we applied this same factor to the $\varepsilon_{G,S}$, such that poly-(GS) shows a slightly more compact scaling (ν = ~0.58) but is substantially more compact in terms of absolute dimensions, in better agreement with experiment (**Fig. 2D**). While this is not in perfect agreement with experimental work, it is an improvement on prior behavior. Without a reliable benchmark for poly-(S) we did not tune the $\varepsilon_{S,S}$ and instead focused on the $\varepsilon_{G,G}$ and $\varepsilon_{G,S}$ values. In summary, these changes provide a modest improvement in the expected polymer scaling behavior for glycine-rich sequences compared to the original Mpipi parameters.

To tune charge-aromatic interactions in Mpipi-GG we used aromatic-aromatic interactions as a benchmark. We compared the fraction of charged residues (FCR) and the fraction of aromatic residues with deviations in radii of gyration from experiment (ΔR_g). Our analysis revealed that Mpipi tended to over-compact sequences with greater aromatic and charge fractions (**Fig S3A**). We next plotted the pairwise WF-potentials, which suggested that the arginine, aspartic acid, and glutamic acid to aromatic interaction strengths were overestimated, likely driving this compaction (**Fig S3B**). Therefore, we tuned the $\varepsilon_{RED,FYW}$ values by systematically titrating to better fit the radii of gyration for these sequences (**Fig 2A, left**). The final $\varepsilon_{RED,FYW}$ value was 60% lower for the Mpipi-GG parameters. We confirmed our modifications better matched experimental radii of gyration by comparing the root mean squared error (RMSE) between simulated and experimental R_g values at different fractions of aromatic residues for the Mpipi-GG and original parameters shown in **Fig. 3A** (**Fig S3C**).

The final set of parameter modifications focuses on aliphatic residues. Aliphatic residues in the original Mpipi force field have very weak interaction strengths. To incorporate hydrophobicity, we made use of the Kyte-Doolittle hydropathy (KD_{hyro}) scale to reparameterize pairwise aliphatic interactions, such that ε_{ij} values are proportional to the sum of $KD_{hyro\ i+j}$ for a pairwise aliphatic interaction of i:j. **(Fig. S4A, B)**. Specifically, we modulated the aliphatic $\varepsilon_{AMLVI,AMLVI}$ values to strengthen aliphatic:aliphatic interactions **(Fig. S4B).**

In summary, small changes were made to parameters associated with twelve of the twenty natural amino acids: proline, glycine, arginine, aspartic acid, glutamic acid, phenylalanine, tyrosine, tryptophan, alanine, valine, isoleucine, leucine, and methionine.

IDR sequence library design

To construct *bona fide* disordered protein sequences, we leveraged the software package GOOSE, which enabled us to construct libraries of rationally designed disordered proteins with specific yet broad sequence chemistries. Therefore, we first designed disordered sequences with varying Fractions of Charged Residues (FCR), Net Charge per Residue (NCPR), and Kyte-Doolittle hydropathy scale values. In addition, we also generated disordered sequences with randomly assigned but specific amino acid fractions (where remaining amino acids were unrestrained) to sample across the sequence space that is accessible to disordered regions. Finally, we also ensured that we had broad coverage of charge distribution in our generated sequences by titrating across kappa, a charge asymmetry parameter where higher values mean greater charge asymmetry.

All-atom Excluded Volume (EV) simulations

Coarse-grained excluded volume (EV) simulations were performed in Mpipi-GG by adjusting the epsilon and sigma parameters such that the interaction potential overlaps for the repulsive component of the function but flattens to zero for distances greater than σ (**Fig. S10**). All-atom EV simulations for polyproline (**Fig. S1**) were performed using the CAMPARI simulation engine (V2) https://campari.sourceforge.net/) and the ABSINTH implicit solvent model[13]. EV simulations were performed as done previously[3,14]. Briefly, EV simulations involve scaling the attractive Lennard-Jones component, the solvation component, and the electrostatic component of the ABSINTH Hamiltonian to zero, such that the only determinant of the underlying ensemble reflects the excluded volume dictated by the repulsive component of the Lennard-Jones potential.

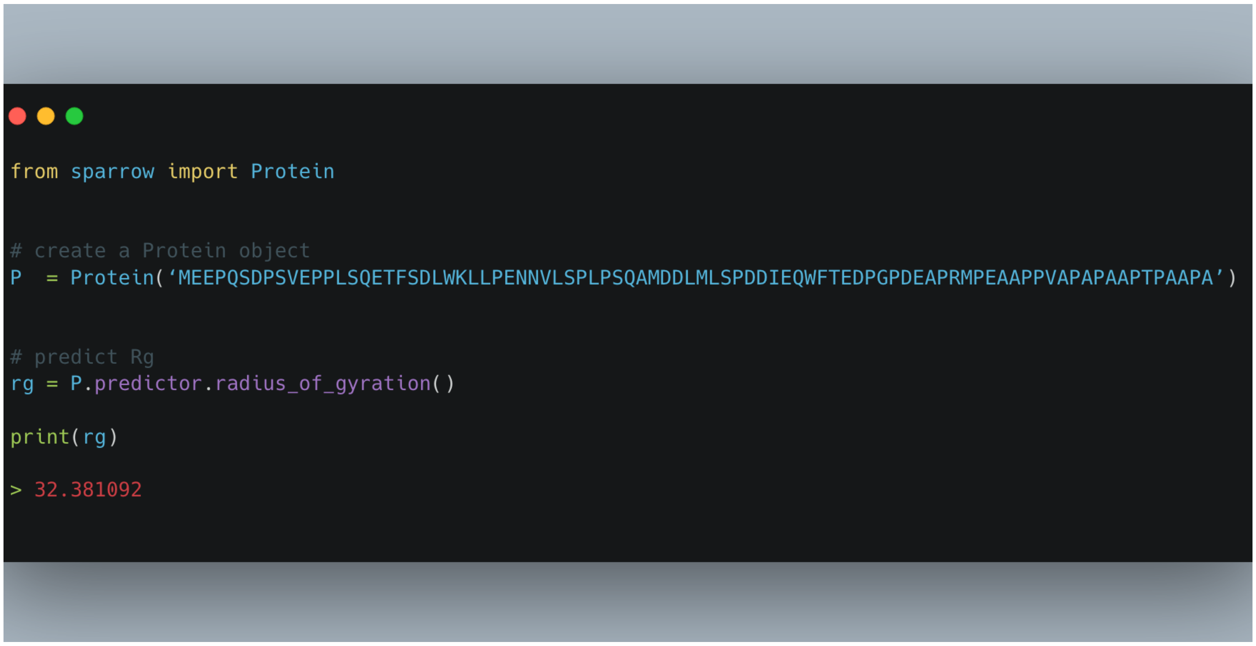
Scaled network training

For both the radius of gyration and the end-to-end distance networks, we trained BRNN-LSTM networks with and without sequence length normalization. For the normalized variations, we performed normalization by taking the respective metric and dividing it by the square root of the sequence length. The radius of gyration (or analogously the end-to-end distance) for a polymer can be defined as $R = A_0 \times N^{\nu}$ where R is either the radius of gyration or end-to-end distance, N is the length of the sequence, and $\boldsymbol{\nu}$ is the scaling exponent. A Gaussian chain is a chain that scales with $\boldsymbol{\nu}$ as 0.5. Therefore, to obtain the length-independent (i.e., sequence chemistry) contribution to the chain dimensions, one can normalize R by the root of the sequence length to derive the following relationship $\frac{R}{\sqrt{(N)}} = A_0 \times N^{(\nu-0.5)}$. This scaling normalizes the measure in polymer space and standardizes the ensemble dimension such that length is a less dominant factor of the learned network.

ALBATROSS distribution

In addition to providing a locally installable implementation of ALBATROSS via SPARROW, we also created a point-and-click style interface for ALBATROSS hosted on Google Colab to eliminate the software barrier of entry for users. By leveraging the cloud computing resources provided in Google Colab, we enable the accurate prediction of IDR conformational properties from sequence from anywhere in the world with an internet connection - even a smart device. Moreover, our Google Colab implementation enables users to specify either a single sequence or upload a fasta file of disordered protein sequences for ALBATROSS predictions. This means users can leverage the unique throughput of ALBATROSS predictions without even needing to write code to construct complex bioinformatic pipelines. Additionally, all predictions are filterable by numerical ranges for that property. For example, if one were trying to design a disordered sequence to serve as a synthetic IDR linker between protein domains, one could upload a FASTA file of proposed disordered proteins with different lengths and compositions, predict the global dimensions of each, and then filter for a specific set of dimensions. This innovation greatly expands the potential for well-designed and controlled synthetic biology experiments and applications, and we are actively implementing such a design protocol into GOOSE.

In addition to the distribution through Google Colab, the ALBATROSS networks are also integrated within the SPARROW sequence analysis package under the "predictors" object operator. In this context, proteome-scale predictions can be achieved in a few lines of code on commodity hardware, e.g.:



```
from sparrow import Protein


# create a Protein object
P  = Protein('MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPA')


# predict Rg
rg = P.predictor.radius_of_gyration()

print(rg)

> 32.381092
```

If GPUs are available, users can optionally perform batch predictions on GPUs, obtaining 1000s of sequence predictions per second. ALBATROSS predictions in SPARROW are, by default, memoized such that computations are not repeated after each call to the predictor operator. An optional override is provided to recompute predictions if desired. The lightweight and object-oriented nature of SPARROW makes it possible to build complex bioinformatic pipelines integrating both bioinformatic sequence properties as well as the emergent biophysical properties of a sequence.

G*ene ontology enrichment*

Gene ontology (GO) enrichment was performed using PANTHER[15]. We calculated enrichment using all IDR-containing proteins as our background (using PANTHER Overrepresentation Test - Released 20221013). For all reported GO terms, we focused on terms where (1) there were over 100 proteins with the term of interest, (2) fold-enrichment was 2x or higher.

2. Supplementary Figures

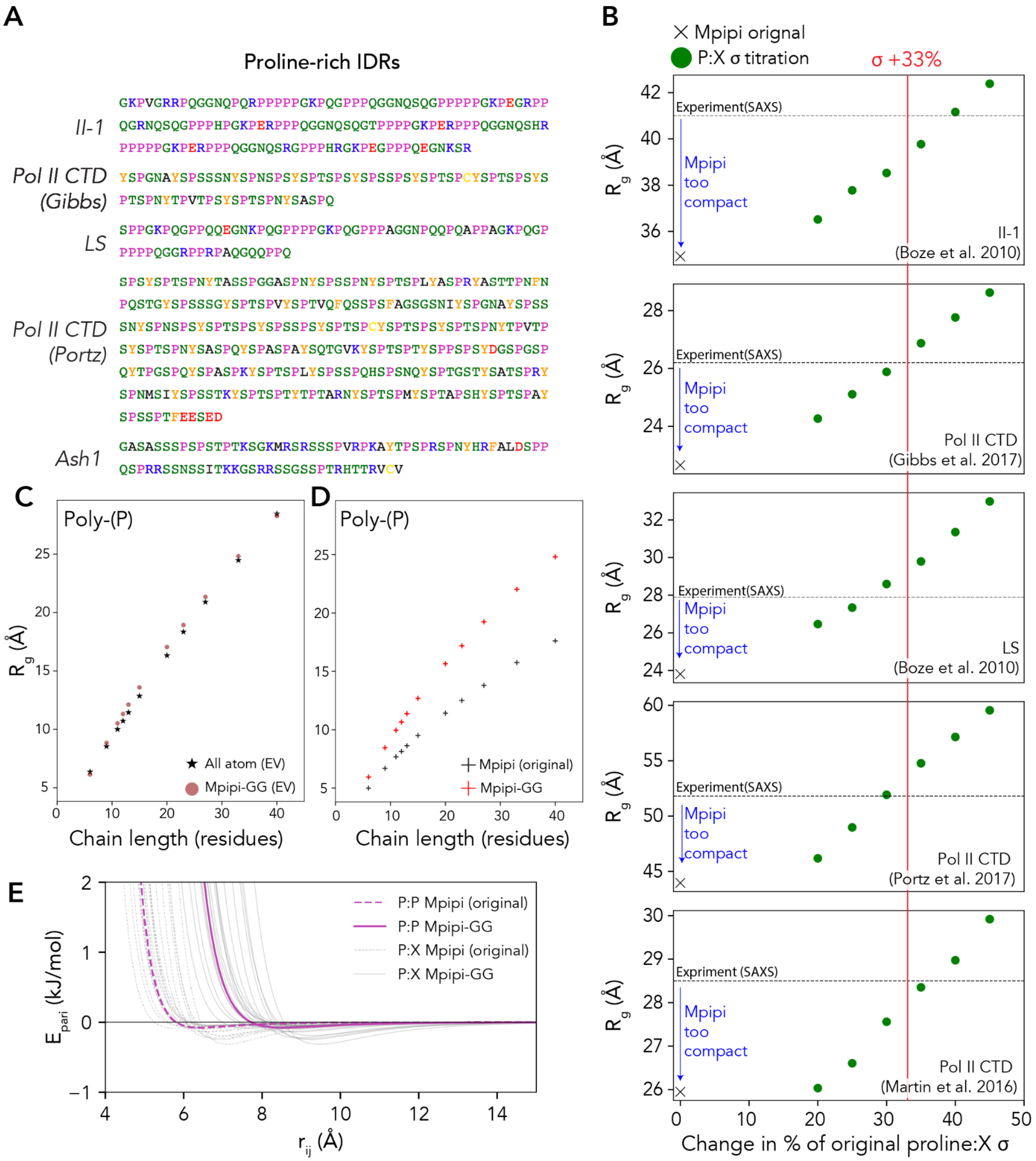
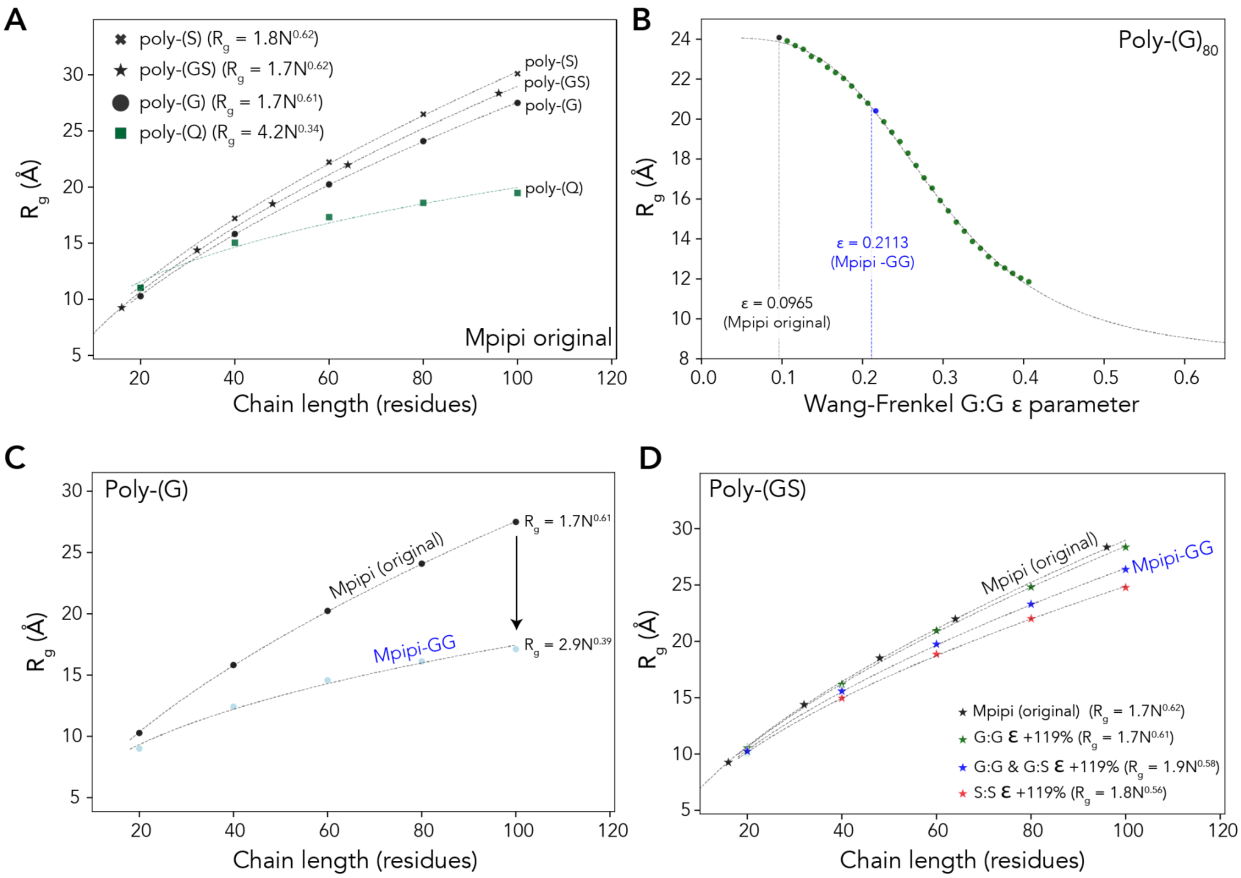
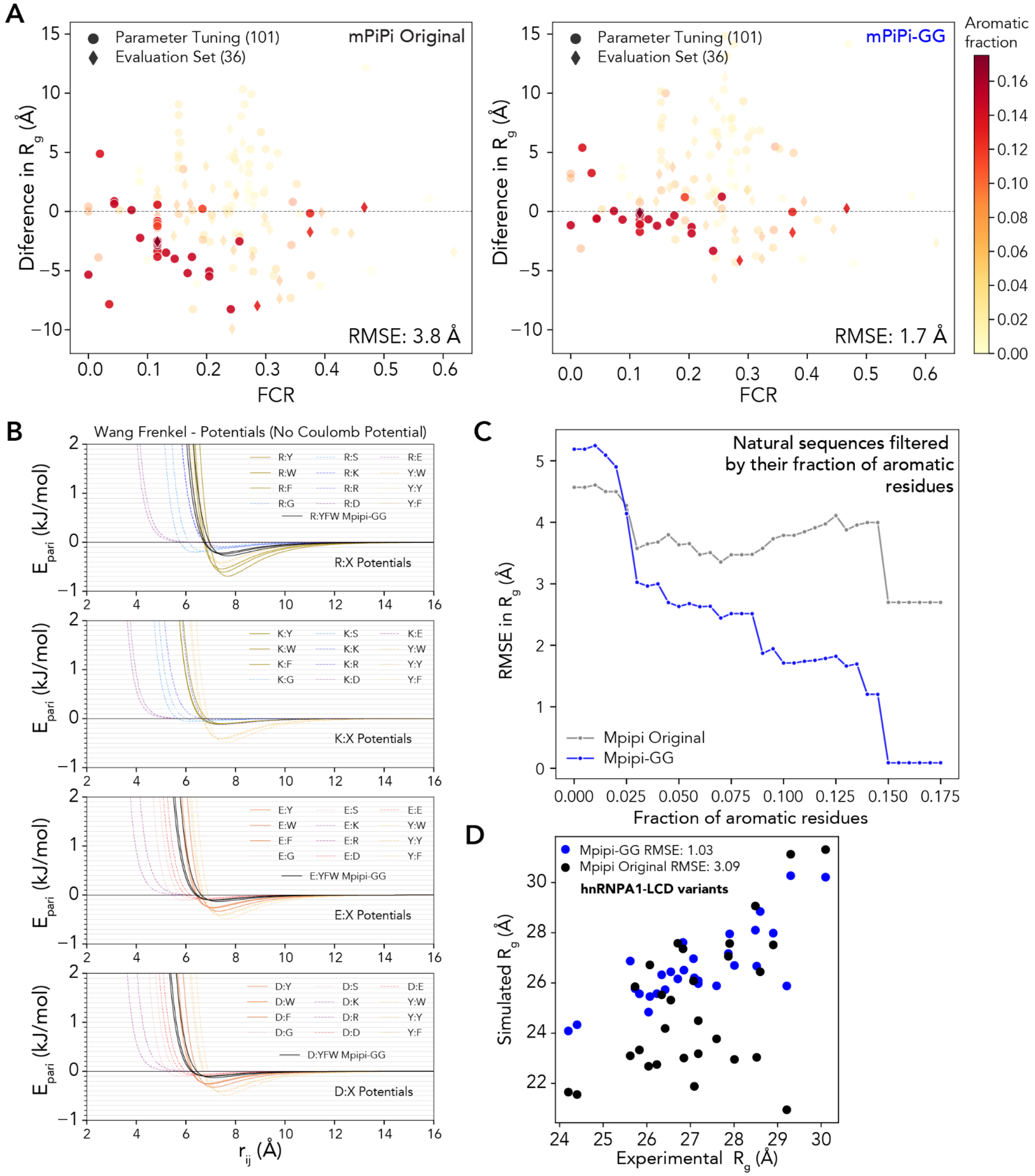


Figure S1. Tuning of proline sigma (σ) parameter in the Mpipi force field
A) Amino acid sequence of proline-rich IDRs previously studied by small angle X-ray scattering (SAXS)[3,5,16,17]. **B)** Comparison of experimental (dashed line) with predicted R_g obtained from Mpipi original ('X' tick, far left) demonstrates that proline-rich IDRs are overly compact in Mpipi. This compaction can be alleviated by increasing the σ parameter from the WF potential (see equation 1), in effect, making proline residues larger. Green points represent the result of a systematic titration of the σ value. **C)** The optimal change to the proline σ parameters was selected by

comparing excluded volume (EV) coarse-grained simulations from Mpipi with all-atom EV simulations and identifying the σ value that results in consistent R_g vs. N scaling. Shown here is a comparison of a +33% increase (used in Mpipi-GG) for Mpipi-GG EV simulations vs. all-atom EV simulations. **D)** Final comparison of polyproline dimensions for Mpipi-GG vs. original Mpipi. Mpipi-GG is more expanded than the original Mpipi, as also shown by the better agreement with experiment at a 33% increase, as shown in panel B. **E)** Wang-Frenkel (WF) potential for Pro:Pro interaction in the original Mpipi parameters (dashed purple line) vs. Mpipi-GG (solid purple line). Dashed and solid gray lines represent proline and each of the other twenty amino acids for Mpipi and Mpipi-GG, respectively.

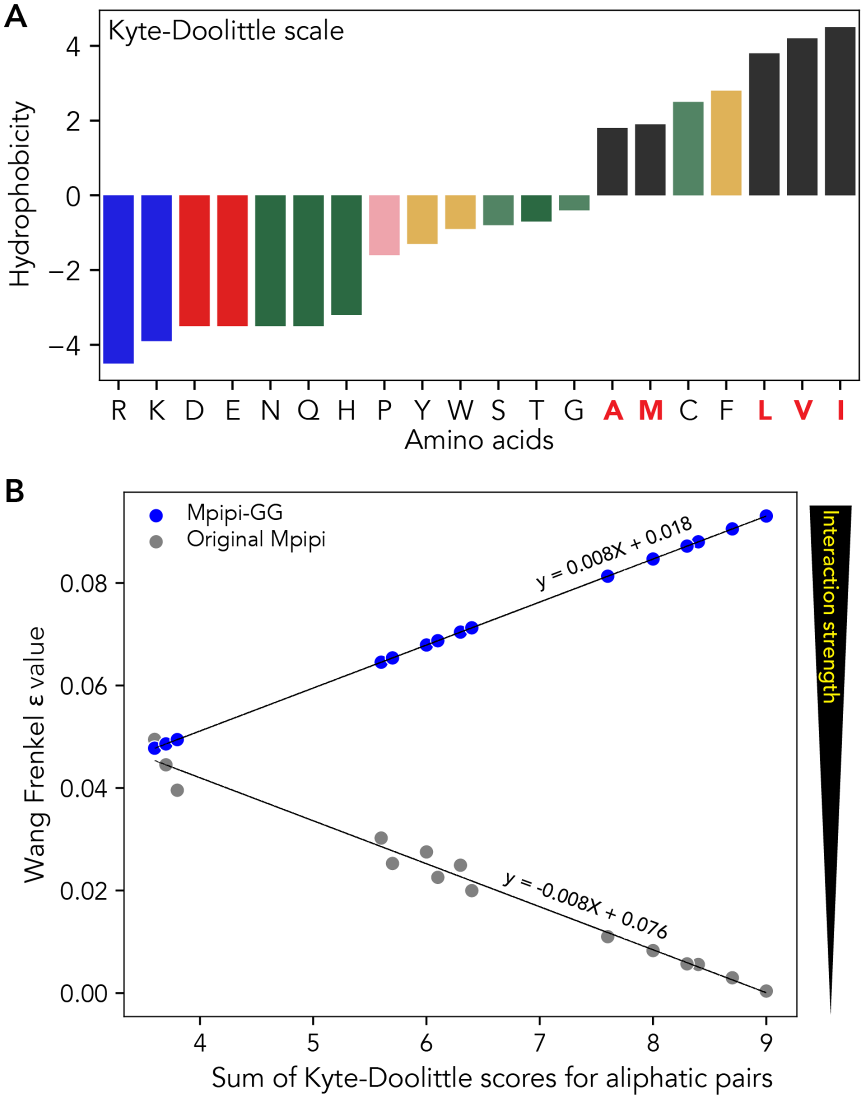


S2. Gly/Ser Mpipi-GG reparameterization A) Simulated scaling behavior for simple polymeric sequences performed using the original Mpipi model. Poly-(Q) compaction is consistent with experimental work[7]. However, poly-(G) scales as a self-avoiding random chain, despite prior work implicating a scaling exponent closer to 0.40[11]. Further, poly-(GS) scales as a self-avoiding random walk (ν = 0.6) against prior work from simulations and experiments which suggest poly-(GS) sequences behave closer to a Gaussian chain (ν = 0.5 - 0.55) [7–10]. These data suggest that G:G interactions are too weak. **B)** To reparameterize G:G strength we systematically titrated the glycine ε parameter, leading to a coil-to-globule titration from which we selected the ε value that best matches the expected scaling of 0.4. **C)** Comparison of Mpipi vs. Mpipi-GG, revealing the more compact scaling and smaller scaling exponent (0.39 vs. 0.61), in better agreement with experiment. **D)** Despite strengthening G:G interactions, poly-(GS) dipeptide repeat polymers are still relatively expanded (compare black and green data). To address this, we asked how changing the S:S: interaction (red) vs. G:G and G:S (blue) altered chain dimensions. Given the prevalence of serine residues in disordered regions, we released the Mpipi S:S interaction strength was likely already reasonable, such that we selected the same scaling for G:G and G:S to enhance cohesive interactions between glycine and serine.

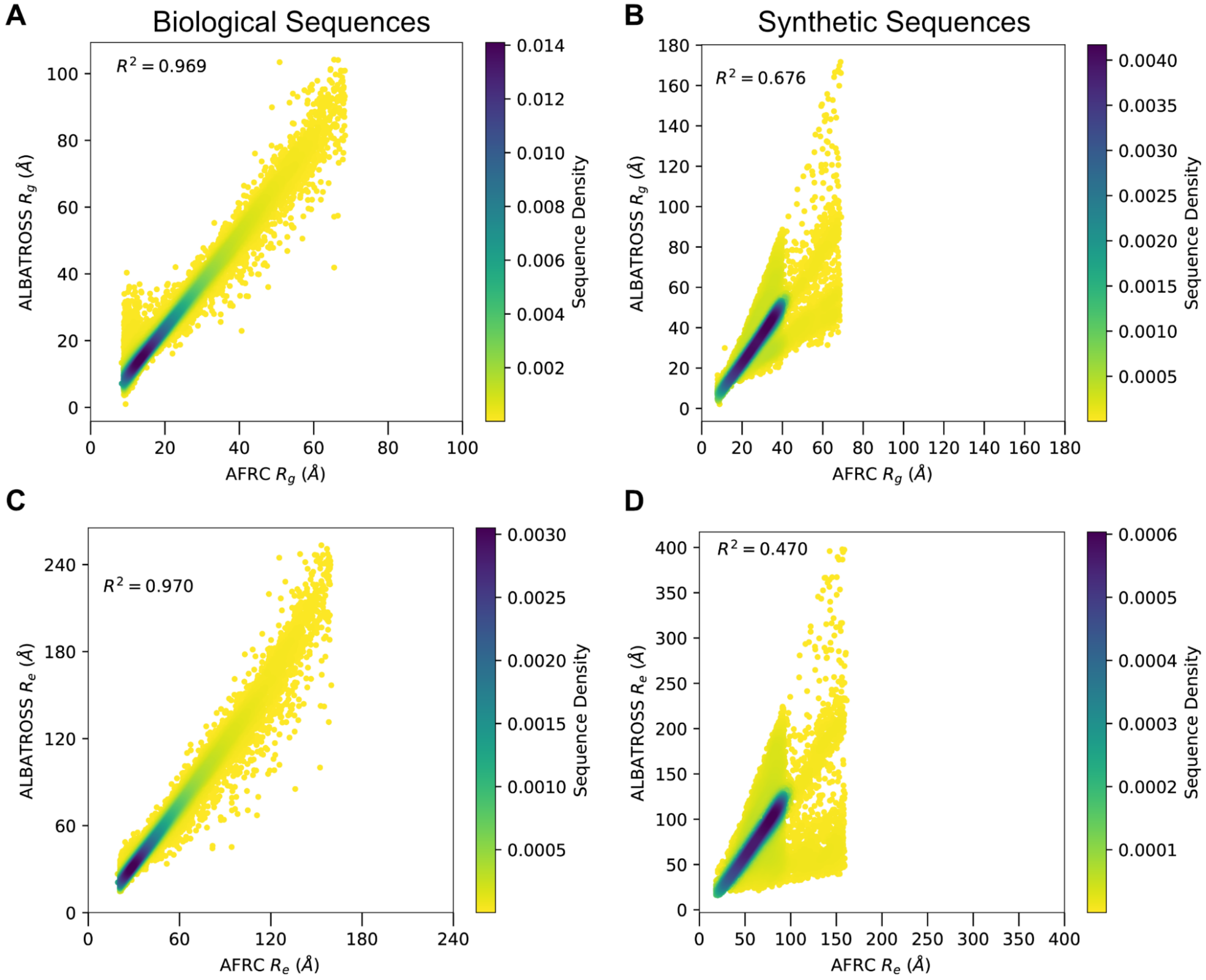


S3. Reparameterization of pairwise aromatic interactions for the Mpipi-GG force field. A) Residual plot for the deviations between experiment and Mpipi (left) and Mpipi-GG (right) as a function of the fraction of charged residues. When comparing all sequences in our curated dataset with >10% aromatic residues, the original Mpipi has an RMSE of 3.8 Å, whereas Mpipi-GG has an RMSE of 1.7 Å. We note specific improvement in sequences that are jointly aromatic and charge rich - i.e., >10% of charged residues by fraction. **B)** Pairwise Wang-Frenkel interaction potentials for arginine, lysine, aspartic acid, and glutamic acid relative to aromatic residues (solid lines) and benchmark residues (dotted lines). Updated Mpipi-GG potentials are drawn in black.

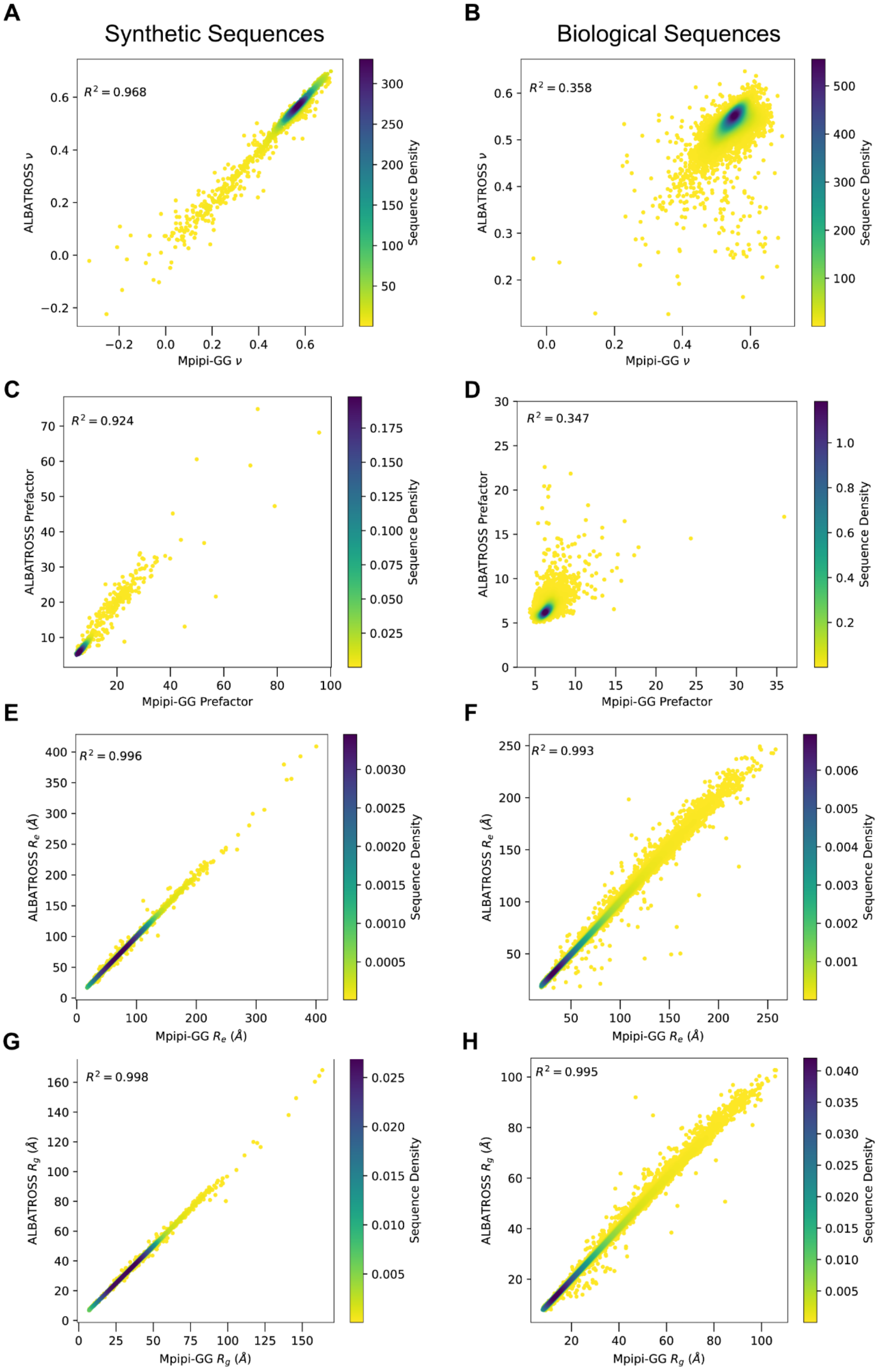
C) Panel **A** uses an aromatic threshold value of 0.10; however, this choice is somewhat arbitrary. Therefore, we chose to demonstrate generality by looking at many different potential thresholds. This plot looks at the root mean squared error as a function of aromatic thresholding. For each threshold value (x-axis), we took all sequences in the curated dataset with aromatic amino acid fractions equal to or exceeding the respective threshold value and computed the RMSE between the experiment and simulated results for each respective force field. As aromatics fractions are increased, Mpipi-GG consistently has modest improvements in recapitulating experimental SAXS radii of gyration. RMSEs near zero in Mpipi-GG are reflective of the fact that there are few sequences with greater than 15% aromatics in the curated library. Nevertheless, Mpipi-GG is highly accurate for these aromatic and charge-rich sequences. **D)** Comparison of simulated R_g vs. SAXS-derived R_g for hnRNPA1-LCD variants[18,19]. These sequences systematically vary charge and aromatic content, providing a convenient reference set for comparing Mpipi-GG vs. Mpipi in the context of aromatic/charge interactions.



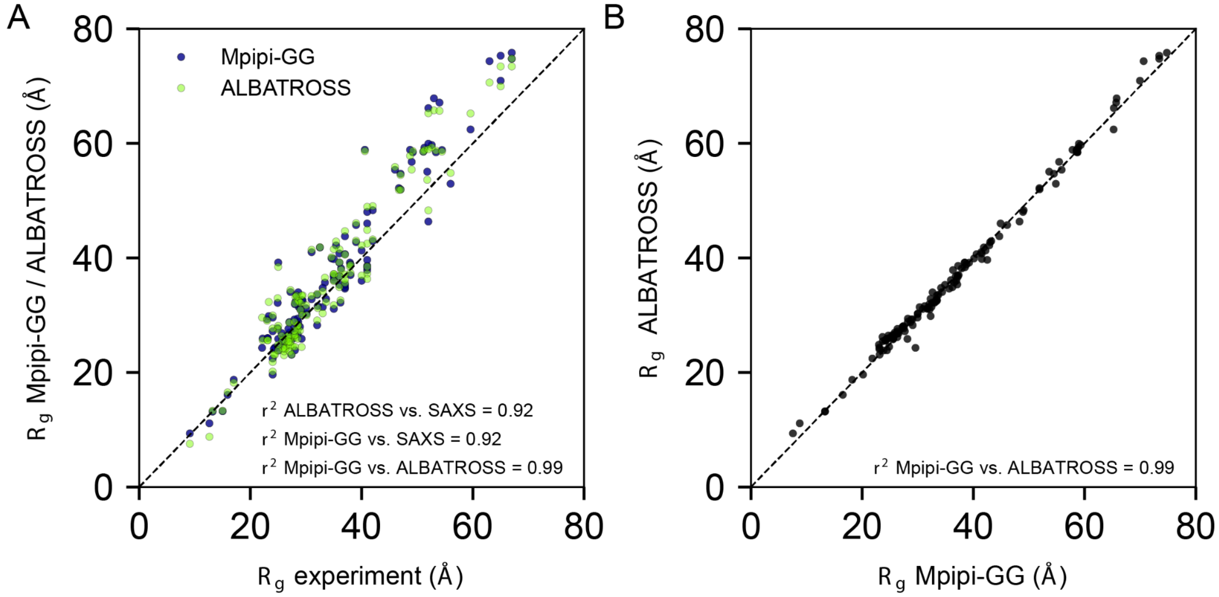
S4. Reparameterization of pairwise aromatic interactions for the Mpipi-GG force field. A) Kyte-Doolittle hydropathy scale for each of the twenty amino acids. Aliphatic residues are highlighted in bold and red along the bottom. **B)** Pairwise sum of Kyte Doolittle hydrophobicity (KD_{hyro}) values relative to Mpipi $\varepsilon_{AMLVI,AMLVI}$ values. Reparameterized $\varepsilon_{AMLVI,AMLVI}$ values in Mpipi-GG, $\varepsilon_{i,j}$ are equal to 0.0008*$KD_{hyro(i+j)}$ +0.018, where the slope of this line is inversely proportional to that of the $\varepsilon_{AMLVI,AMLVI}$ values in the original Mpipi force field, but scaled so the more hydrophobic pairs are stronger, as opposed to weaker.



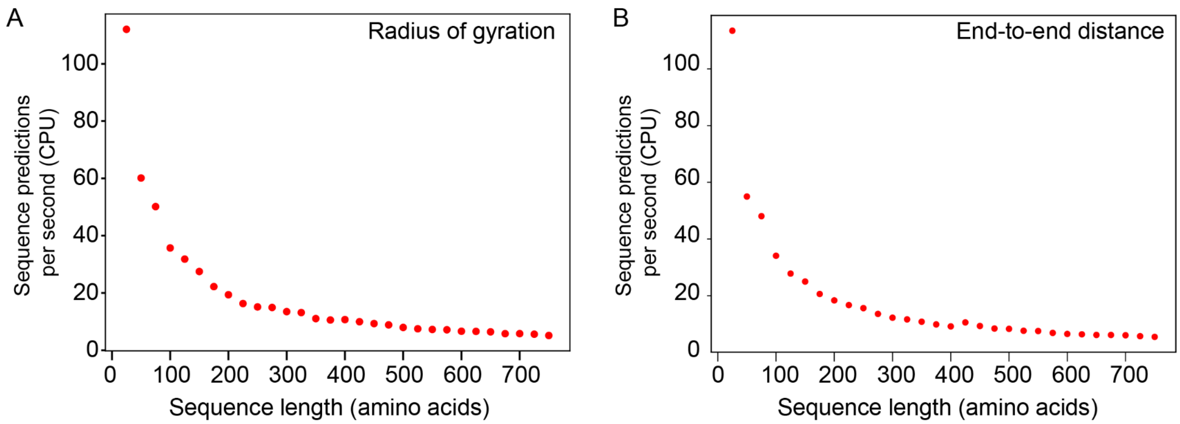
S5. Comparing the Analytical Flory Random Coil (AFRC) chain dimensions and the ALBATROSS predicted chain dimensions for the synthetic and biological sequence libraries. A-B) Correlations between modeled radii of gyration for both biological sequences (right) and the synthetic sequences (left). **C-D)** Correlations between modeled end-to-end distances for both biological sequences (right) and the synthetic sequences (left).



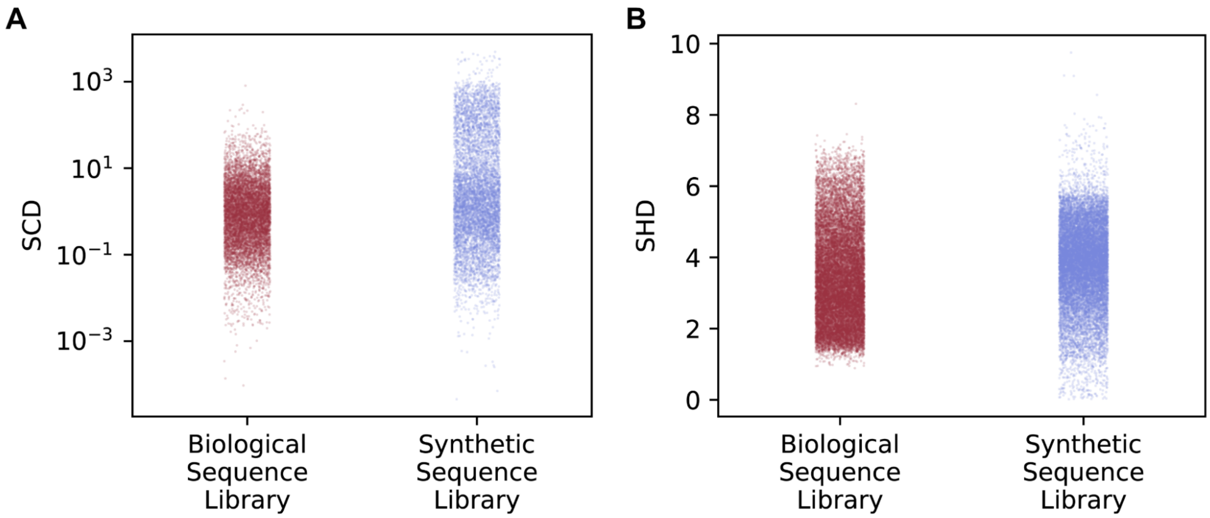
S6. Evaluating the performance of the ALBATROSS networks on held-out test sets for synthetic and biological sequences. A) Performance of the ALBATROSS scaling exponent network on a held-out test set of synthetic sequences unseen during training. **B)** Performance of the ALBATROSS scaling exponent network on a held-out test set of biological sequences unseen during training. **C)** Performance of the ALBATROSS polymer prefactor network on a held-out test set of synthetic sequences unseen during training. **D)** Performance of the ALBATROSS prefactor network on a held-out test set of biological sequences unseen during training. **E)** Performance of the ALBATROSS scaled end-to-end distance network on a held-out test set of synthetic sequences unseen during training. **F)** Performance of the ALBATROSS scaled end-to-end distance network on a held-out test set of biological sequences unseen during training. **G)** Performance of the ALBATROSS scaled radius of gyration network on a held-out test set of synthetic sequences unseen during training. **H)** Performance of the ALBATROSS scaled radius of gyration network on a held-out test set of biological sequences unseen during training.



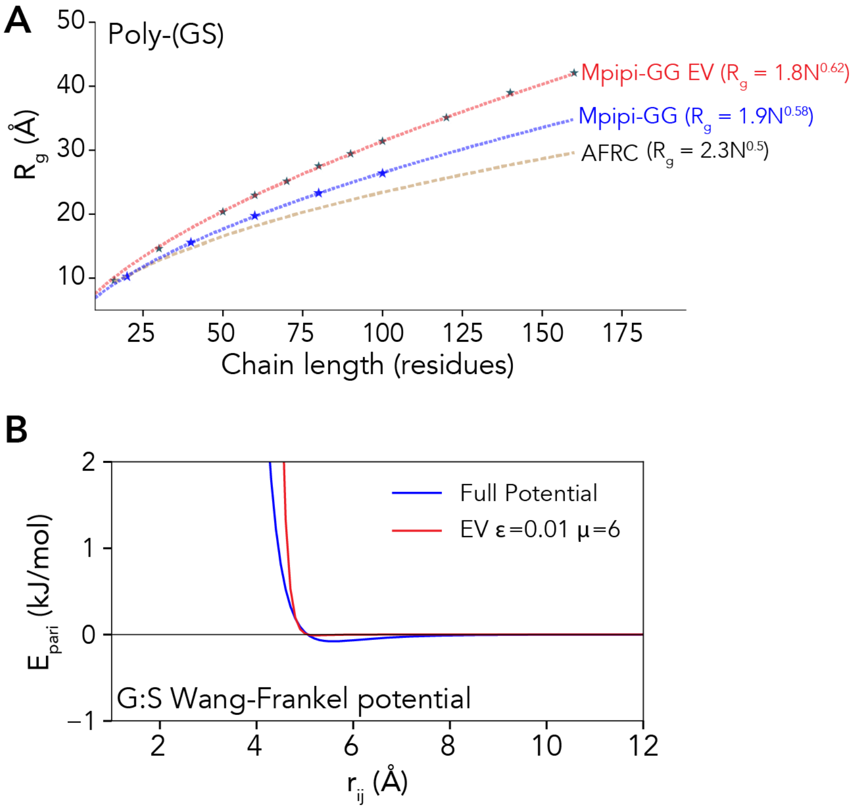
S7. Evaluating the performance of the ALBATROSS network at recapitulating experimental SAXS measurements. A) The ALBATROSS R_g network, consistent with the Mpipi-GG simulations, accurately reproduces experimental SAXS data ($R^2 = 0.92$) **B)** For the same set of sequences, ALBATROSS essentially maps 1-1 with Mpipi-GG radii of gyration, as shown also in Fig. S6 and Fig. 4.



S8. Network performance on standard commodity hardware. We measured predictive power on standard CPUs for the various networks as a function of sequence length. For 100-residue IDRs, performance sits around 30-40 sequences per second. We emphasize that on a Google Colab notebook using GPUs, the entire human proteome takes ~8 seconds, but we focus here on CPU performance given the broad availability of CPUs.

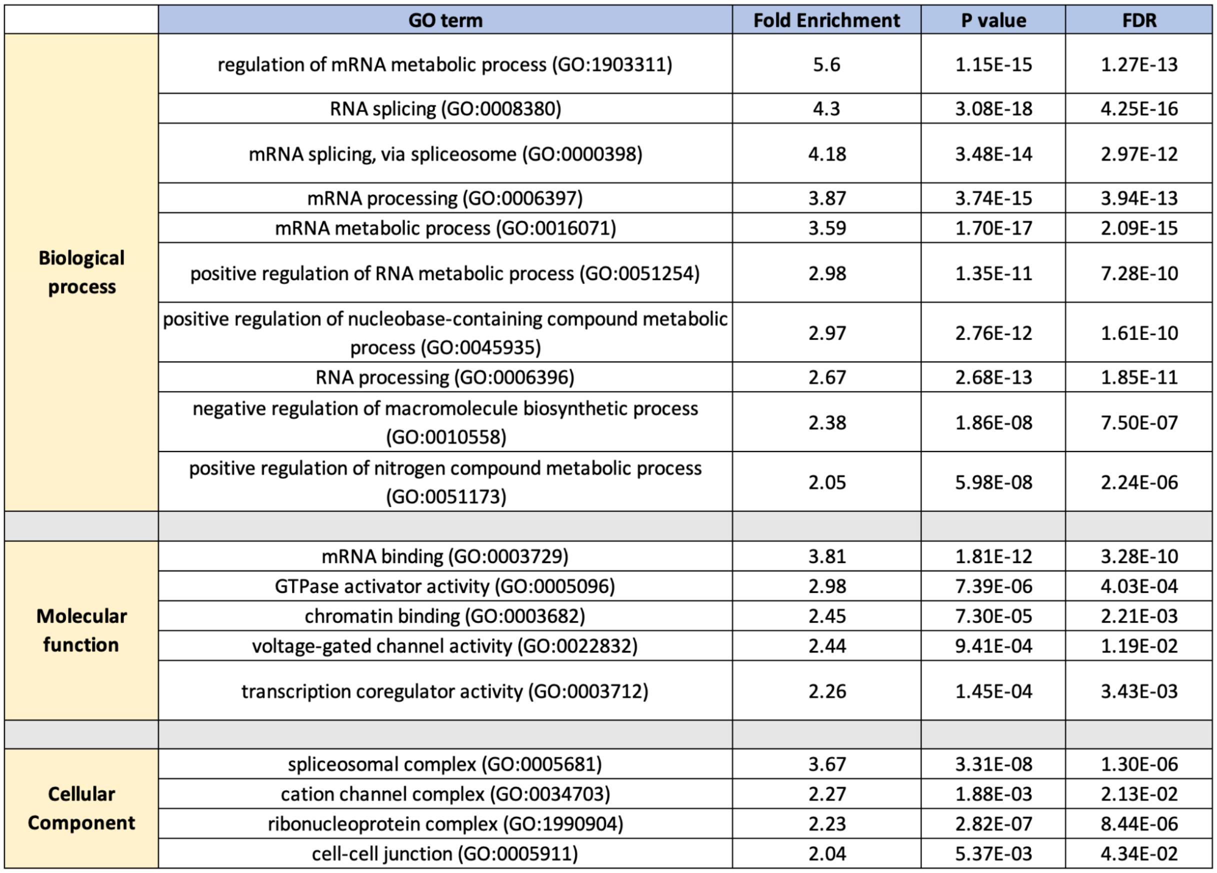


S9 Sequence property comparison for the biological and synthetic sequence libraries. A) Distribution of sequence charge decoration (SCD) values for the biological sequences (red) and synthetic sequences (blue). Note the SCD is plotted on a logarithmic scale as the synthetic sequence library covers a much broader distribution than the biological sequences. **B)** Distribution of sequence hydropathy decoration values for the biological sequences (red) and synthetic sequences (blue).



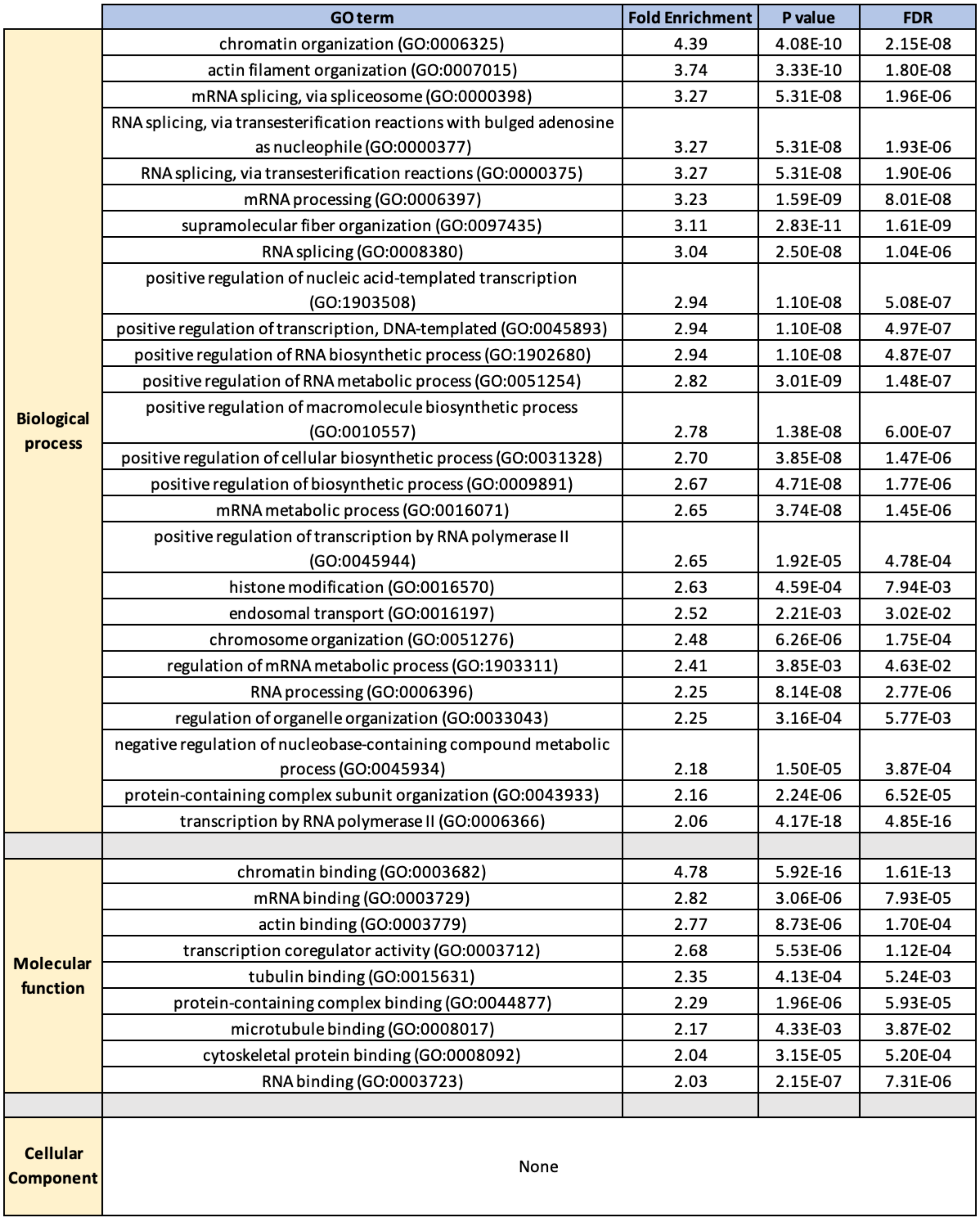
S10. Parameterizing an excluded volume model for the Mpipi-GG force field. A) Scaling behavior for poly-GS as represented in terms of Mpipi-GG (blue) Mpipi-GG as an excluded volume (EV) simulation, and the Analytical Flory Random Coil (AFRC) model. EV simulations are more extended than full Mipi-GG simulations. **B)** Wang-Frankel interaction potentials for a representative pair of beads (G:S). The full Wang-Frankel potentials for the interaction of glycine and serine in the Mpipi-GG forcefield - note the dip near r_{ij} of ~5.5, reflecting the attractive part of the potential. After tuning the $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$ parameters, we obtained a pairwise interaction potential with near zero attractive interactions (red) that match the same dimensions as the full Mpipi-GG forcefield.

3. Supplementary Tables



	GO term	Fold Enrichment	P value	FDR
Biological process	regulation of mRNA metabolic process (GO:1903311)	5.6	1.15E-15	1.27E-13
	RNA splicing (GO:0008380)	4.3	3.08E-18	4.25E-16
	mRNA splicing, via spliceosome (GO:0000398)	4.18	3.48E-14	2.97E-12
	mRNA processing (GO:0006397)	3.87	3.74E-15	3.94E-13
	mRNA metabolic process (GO:0016071)	3.59	1.70E-17	2.09E-15
	positive regulation of RNA metabolic process (GO:0051254)	2.98	1.35E-11	7.28E-10
	positive regulation of nucleobase-containing compound metabolic process (GO:0045935)	2.97	2.76E-12	1.61E-10
	RNA processing (GO:0006396)	2.67	2.68E-13	1.85E-11
	negative regulation of macromolecule biosynthetic process (GO:0010558)	2.38	1.86E-08	7.50E-07
	positive regulation of nitrogen compound metabolic process (GO:0051173)	2.05	5.98E-08	2.24E-06
Molecular function	mRNA binding (GO:0003729)	3.81	1.81E-12	3.28E-10
	GTPase activator activity (GO:0005096)	2.98	7.39E-06	4.03E-04
	chromatin binding (GO:0003682)	2.45	7.30E-05	2.21E-03
	voltage-gated channel activity (GO:0022832)	2.44	9.41E-04	1.19E-02
	transcription coregulator activity (GO:0003712)	2.26	1.45E-04	3.43E-03
Cellular Component	spliceosomal complex (GO:0005681)	3.67	3.31E-08	1.30E-06
	cation channel complex (GO:0034703)	2.27	1.88E-03	2.13E-02
	ribonucleoprotein complex (GO:1990904)	2.23	2.82E-07	8.44E-06
	cell-cell junction (GO:0005911)	2.04	5.37E-03	4.34E-02

Table S1. Gene ontology analysis for proteins with compact IDRs. Only those ontology annotations that contained 100 or more entries in the basis set and showed 2-fold or higher enrichment are shown. With the exception of a few terms with questionable FDR rates, every annotated term pertains to RNA in some way, across the three classes of gene ontology.



	GO term	Fold Enrichment	P value	FDR
Biological process	chromatin organization (GO:0006325)	4.39	4.08E-10	2.15E-08
	actin filament organization (GO:0007015)	3.74	3.33E-10	1.80E-08
	mRNA splicing, via spliceosome (GO:0000398)	3.27	5.31E-08	1.96E-06
	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile (GO:0000377)	3.27	5.31E-08	1.93E-06
	RNA splicing, via transesterification reactions (GO:0000375)	3.27	5.31E-08	1.90E-06
	mRNA processing (GO:0006397)	3.23	1.59E-09	8.01E-08
	supramolecular fiber organization (GO:0097435)	3.11	2.83E-11	1.61E-09
	RNA splicing (GO:0008380)	3.04	2.50E-08	1.04E-06
	positive regulation of nucleic acid-templated transcription (GO:1903508)	2.94	1.10E-08	5.08E-07
	positive regulation of transcription, DNA-templated (GO:0045893)	2.94	1.10E-08	4.97E-07
	positive regulation of RNA biosynthetic process (GO:1902680)	2.94	1.10E-08	4.87E-07
	positive regulation of RNA metabolic process (GO:0051254)	2.82	3.01E-09	1.48E-07
	positive regulation of macromolecule biosynthetic process (GO:0010557)	2.78	1.38E-08	6.00E-07
	positive regulation of cellular biosynthetic process (GO:0031328)	2.70	3.85E-08	1.47E-06
	positive regulation of biosynthetic process (GO:0009891)	2.67	4.71E-08	1.77E-06
	mRNA metabolic process (GO:0016071)	2.65	3.74E-08	1.45E-06
	positive regulation of transcription by RNA polymerase II (GO:0045944)	2.65	1.92E-05	4.78E-04
	histone modification (GO:0016570)	2.63	4.59E-04	7.94E-03
	endosomal transport (GO:0016197)	2.52	2.21E-03	3.02E-02
	chromosome organization (GO:0051276)	2.48	6.26E-06	1.75E-04
	regulation of mRNA metabolic process (GO:1903311)	2.41	3.85E-03	4.63E-02
	RNA processing (GO:0006396)	2.25	8.14E-08	2.77E-06
	regulation of organelle organization (GO:0033043)	2.25	3.16E-04	5.77E-03
	negative regulation of nucleobase-containing compound metabolic process (GO:0045934)	2.18	1.50E-05	3.87E-04
	protein-containing complex subunit organization (GO:0043933)	2.16	2.24E-06	6.52E-05
	transcription by RNA polymerase II (GO:0006366)	2.06	4.17E-18	4.85E-16
Molecular function	chromatin binding (GO:0003682)	4.78	5.92E-16	1.61E-13
	mRNA binding (GO:0003729)	2.82	3.06E-06	7.93E-05
	actin binding (GO:0003779)	2.77	8.73E-06	1.70E-04
	transcription coregulator activity (GO:0003712)	2.68	5.53E-06	1.12E-04
	tubulin binding (GO:0015631)	2.35	4.13E-04	5.24E-03
	protein-containing complex binding (GO:0044877)	2.29	1.96E-06	5.93E-05
	microtubule binding (GO:0008017)	2.17	4.33E-03	3.87E-02
	cytoskeletal protein binding (GO:0008092)	2.04	3.15E-05	5.20E-04
	RNA binding (GO:0003723)	2.03	2.15E-07	7.31E-06
Cellular Component	None			

Table S3. Gene ontology analysis for proteins with expanded IDRs. Only those ontology annotations that contained 100 or more entries in the basis set and showed a 2-fold or higher enrichment are shown. A variety of IDR-associated annotations are shown, which include RNA-associated functions, but also include chromatin binding, cytoskeletal regulation, and cellular

organization, in good agreement with analogous analysis from Tesei & Trolle, despite the two analyses being done in different ways [20].

4. Supplementary References

1. Wang, X., Ramírez-Hinestrosa, S., Dobnikar, J. & Frenkel, D. The Lennard-Jones potential: when (not) to use it. *Phys. Chem. Chem. Phys.* **22,** 10624–10633 (2020).
2. 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).
3. 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).
4. Auton, M., Holthauzen, L. M. F. & Bolen, D. W. Anatomy of energetic changes accompanying urea-induced protein denaturation. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 15317–15322 (2007).
5. 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).
6. Yarawsky, A. E., English, L. R., Whitten, S. T. & Herr, A. B. The Proline/Glycine-Rich Region of the Biofilm Adhesion Protein Aap Forms an Extended Stalk that Resists Compaction. *J. Mol. Biol.* **429,** 261–279 (2017).
7. Crick, S. L., Jayaraman, M., Frieden, C., Wetzel, R. & Pappu, R. V. Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed

structures in aqueous solutions. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 16764–16769 (2006).

8. Sørensen, C. S. & Kjaergaard, M. Effective concentrations enforced by intrinsically disordered linkers are governed by polymer physics. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 23124–23131 (2019).
9. Moses, D., Guadalupe, K., Yu, F., Flores, E., Perez, A., McAnelly, R., Shamoon, N. M., Cuevas-Zepeda, E., Merg, A. D., Martin, E. W., Holehouse, A. S. & Sukenik, S. Structural biases in disordered proteins are prevalent in the cell. *bioRxiv* 2021.11.24.469609 (2022). doi:10.1101/2021.11.24.469609
10. Moses, D., Yu, F., Ginell, G. M., Shamoon, N. M., Koenig, P. S., Holehouse, A. S. & Sukenik, S. Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to their Chemical Environment. *J. Phys. Chem. Lett.* **11,** 10131–10136 (2020).
11. Holehouse, A. S., Garai, K., Lyle, N., Vitalis, A. & Pappu, R. V. Quantitative assessments of the distinct contributions of polypeptide backbone amides versus side chain groups to chain expansion via chemical denaturation. *J. Am. Chem. Soc.* **137,** 2984–2995 (2015).
12. Holehouse, A. S. & Pappu, R. V. Collapse Transitions of Proteins and the Interplay Among Backbone, Sidechain, and Solvent Interactions. *Annu. Rev. Biophys.* **47,** 19–39 (2018).
13. Vitalis, A. & Pappu, R. V. in *Annual Reports in Computational Chemistry* (ed. Wheeler, R. A.) **5,** 49–76 (Elsevier, 2009).
14. 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. *bioRxiv* (2023). doi:10.1101/2023.02.16.528879
15. Thomas, P. D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L.-P. & Mi, H. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci.* **31,** 8–22 (2022).
16. 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).

17. Boze, H., Marlin, T., Durand, D., Pérez, J., Vernhet, A., Canon, F., Sarni-Manchado, P., Cheynier, V. & Cabane, B. Proline-rich salivary proteins have extended conformations. *Biophys. J.* **99,** 656–665 (2010).

18. Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V. & Mittag, T. Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat. Chem.* **14,** 196–207 (2022).

19. 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).

20. Tesei, G., Trolle, A. I., Jonsson, N., Betz, J., Pesce, F., Johansson, K. E. & Lindorff-Larsen, K. Conformational ensembles of the human intrinsically disordered proteome: Bridging chain compaction with function and sequence conservation. *bioRxiv* (2023).