The analytical Flory random coil is a simple-to-use reference model for unfolded and disordered proteins

Jhullian J. Alston[1,2]*, Garrett M. Ginell*[1,2], Andrea Soranno[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 to this work

✉ Corresponding author, e-mail: alex.holehouse@wustl.edu

ABSTRACT

Denatured, unfolded, and intrinsically disordered proteins (collectively referred to here as unfolded proteins) can be described using analytical polymer models. These models capture various polymeric properties and can be fit to simulation results or experimental data. However, the model parameters commonly require users' decisions, making them useful for data interpretation but less clearly applicable as stand-alone reference models. Here we use all-atom simulations of polypeptides in conjunction with polymer scaling theory to parameterize an analytical model of unfolded polypeptides that behave as ideal chains ($\nu = 0.50$). The model, which we call the analytical Flory Random Coil (AFRC), requires only the amino acid sequence as input and provides direct access to probability distributions of global and local conformational order parameters. The model defines a specific reference state to which experimental and computational results can be compared and normalized. As a proof-of-concept, we use the AFRC to identify sequence-specific intramolecular interactions in simulations of disordered proteins. We also use the AFRC to contextualize a curated set of 145 different radii of gyration obtained from previously published small-angle X-ray scattering experiments of disordered proteins. The AFRC is implemented as a stand-alone software package and is also available via a Google colab notebook. In summary, the AFRC provides a simple-to-use reference polymer model that can guide intuition and aid in interpreting experimental or simulation results.

INTRODUCTION

Proteins are finite-sized heteropolymers, and the application of polymer physics has provided a useful toolkit for understanding protein structure and function[1–9]. In particular, there has been significant interest in unfolded proteins under both native and non-native conditions[2,10–17]. Depending on the experimental techniques employed, a variety of polymeric properties can be measured, including the radius of gyration (R_g), the hydrodynamic radius (R_h), the end-to-end distance (R_e), and the apparent scaling exponent (ν_{app}). These and many other parameters can be calculated directly from all-atom simulations, and the synergy of simulation and experiment has provided a powerful approach for constructing large ensembles of unfolded proteins for greater insight into the unfolded state[15,18–28].

Polymers can be described in terms of scaling laws, expressions that describe how chain dimensions vary as a function of chain length[29–31]. Polymer scaling laws typically have the format $D = R_0N^{\nu}$. Here, D reports on chain dimensions, R_0 is a prefactor in units of spatial distance, and N is the number of monomers, which in the case of proteins is typically written in terms of the number of amino acids. ν (or, more accurately, ν_{app} when applied to finite-sized heteropolymers like proteins) is the (apparent) Flory scaling exponent. In principle, ν_{app} lies between 0.33 (as is obtained for a perfect spherical globule) and 0.59 (as obtained for a self-avoiding chain). However, for finite-sized polymers, values beyond 0.59 can be obtainable for self-repulsive chains [32–34]. The applicability of polymer scaling laws to describe real proteins assumes they are sufficiently long to display *bona fide* polymeric behavior and that they are sufficiently self-similar over a certain length scale, analogous to fractals. While this assumption often holds true, it is worth noting that sequence-encoded patterns in specific chemistries and/or secondary structure can lead to deviations from homopolymer-like behavior [18,35–37].

To what extent do polymer scaling laws apply to real proteins? For denatured polypeptides, Kohn *et al.* reported the ensemble-average radius of gyration using the scaling expressions $R_g = 1.927N^{0.59811}$. This result provides strong experimental evidence to support a model whereby denaturants unfolded proteins by uniformly weakening intramolecular protein-protein interactions[1]. A value for ν_{app} of 0.598 also agrees with the previously reported value of 0.57 by Wilkins *et al.* and earlier work by Damaschun[1,10,12]. In short, under strongly denaturing conditions, proteins appear to behave as polymers in a good solvent [1,32,38–41].

For proteins under native or native-like conditions, the apparent scaling exponents obtained for unfolded polypeptides are more variable. Marsh and Forman-Kay reported an average scaling expression of $R_h = 2.49N^{0.509}$, for a set of intrinsically disordered proteins, while Bernadó and Svergun found a similar average relationship in $R_g = 2.54N^{0.52}$ [42,43]. More recently, various means

to estimate ν_{app} for individual proteins have enabled values of ν_{app} between 0.42 and 0.60 to be measured for a wide range of unfolded proteins of different lengths and compositions[15,18,23,25,39,40,44–46]. An emerging consensus suggests that ν_{app} depends on the underlying amino acid sequence[2,17,47]. If sequence-encoded chemical biases enable intramolecular interactions, then ν_{app} may be lower than 0.5. Notably, despite clear conceptual limitations, the physics of homopolymers remains a convenient tool through which unfolded proteins can be assessed[15,18,36,37,48].

Given the variety in scaling exponents for unfolded proteins under native conditions, we felt that a sequence-specific reference model would be helpful for the field. Such a model could provide a touchstone for experimentally measurable polymeric parameters, including intermolecular distances, the radius of gyration, the end-to-end distance, and the hydrodynamic radius. Similarly, such a model would provide a simple reference state with which simulations could be directly compared and used to identify sequence-specific effects. Finally, a standard reference model could offer an easy way to compare unfolded proteins of different lengths to assess if they behave similarly despite different absolute dimensions.

Here, we perform sequence-specific numerical simulations for polypeptides as an ideal chain, so-called Flory Random Coil (FRC) simulations[2,31]. Under these conditions, chain-chain, chain-solvent, and solvent-solvent interactions are all equivalent, no long-range excluded volume contributions are included, and as such, the polypeptide behaves as a Gaussian chain with $\nu_{app} = 0.5$. Because our FRC implementation minimizes finite-chain artifacts, we can parameterize an analytical, sequence-specific model using standard approaches from scaling theory, a model we call the Analytical Flory Random Coil (AFRC). This model enables the calculation of distance distributions for the end-to-end distance and the radius of gyration, as well as a variety of additional parameters that become convenient for the analysis of all-atom simulations and experiments.

The AFRC is not a predictor of unfolded protein dimensions. Those dimensions depend on the complex interplay of chain:chain and chain:solvent interactions, which are themselves determined by sequence-encoded chemistry[49–53]. Instead, the AFRC provides a simple reference state that can aid in interpreting experimental and computational results without needing information other than the protein sequence. The AFRC is implemented in a stand-alone Python package and is also provided as a simple Google Colab notebook. We demonstrate the utility of this model by comparing experimental data and computational results.

The remainder of this paper is outlined as follows. First, we discuss the implementation details of the model, including a comparison against existing polymer models. Next, we analyzed

previously published all-atom simulations to demonstrate how the AFRC can identify signatures of sequence-specific intramolecular interactions in disordered ensembles. Finally, we use the AFRC model to re-interpret previously reported small-angle X-ray scattering data of intrinsically disordered proteins.

METHODS AND RESULTS

Implementation of a numerical model for sequence-specific ideal chain simulations

We used a Monte Carlo-based approach to construct sequence-specific atomistic ensembles of polypeptides as ideal chains. All-atom simulations with all non-bonded and solvation interactions scaled to zero were performed using a modified version of the CAMPARI Monte Carlo simulation engine using bond lengths and atomic radii defined by the ABSINTH-OPLS forcefield[2,54,55]. We modified CAMPARI to reproduce Flory's rotational isomeric state approximation[31,56]. In this method, an initial conformation of the polypeptide is randomly generated. Upon each Monte Carlo step, a residue is randomly selected, the backbone dihedrals are rearranged to one of a subset of allowed residue-specific psi/phi values (i.e., specific isomeric states), and the chain is rearranged accordingly (**Fig. 1A, B)**. Allowed phi/psi values are selected from a database of residue-specific allowed values as determined by all-atom simulations of peptide units, with the associated Ramachandran maps shown in **Fig. S1**. Importantly, the Monte Carlo moves in these simulations approach are rejection-free. That is, only allowed phi/psi angles are proposed, and no consideration of steric overlap in the resulting conformation is given. The ensemble generated by these simulations is referred to as the Flory Random Coil (FRC, **Fig. 1C**) and has been used as a convenient reference frame for comparing simulations of disordered and unfolded polypeptides for over a decade (as reviewed by Mao *et al.*[2])[15,57–61].

FRC simulations enable the construction of ensembles where each amino acid exists in a locally allowed configuration, yet no through-space interactions occur. This has two important implications for the construction of an ideal chain model. Firstly, each monomer has no preference for chain:chain vs. chain:solvent interactions (each monomer is "agnostic" to its surroundings). As a result, both internal and global dimensions show scaling behavior with an apparent scaling exponent (ν_{app}) of 0.5 (**Fig. 1D**), analogous to that of a polymer in a theta solvent. Secondly, terminal residues sample conformational space in the same way as residues internal to the chain (**Fig. S2**). This means that end-effects that emerge finite-chain effects are not experienced in terms of end effects (**Fig. 1E**). This is in contrast to finite-sized self-avoiding chains, in which internal scaling profiles reveal a noticeable and predictable "dangling end" finite-chain effect (**Fig 1E, Fig. S2**). In summary, FRC simulations enable us to generate

ensembles at all-atom resolution that are nearly fully approximations of ideal chains, reproducing the behavior of a hypothetical “ideal” polypeptide.

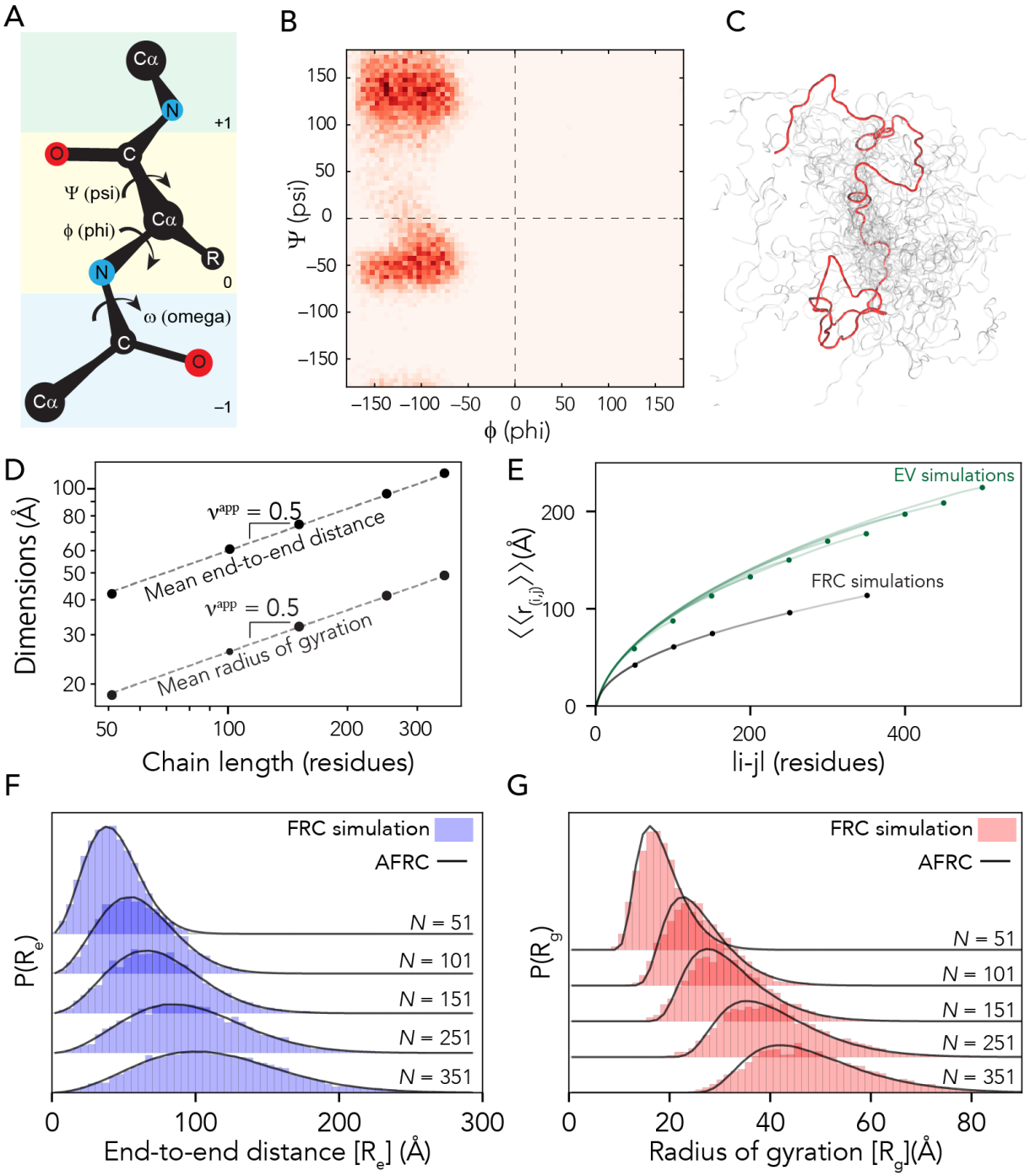

Figure 1: The AFRC is a pre-parameterized polymer model based on residue-specific polypeptide behavior. A. Schematic of the amino acid dihedral angles. **B.** Ramachandran map for alanine used to select acceptable backbone conformations for the FRC simulations. All twenty amino acids are shown in

Fig. S1. **C.** Graphical rendering of an FRC ensemble for a 100-residue homopolymer. **D.** Flory Random Coil (FRC) simulations performed using a modified version of the ABSINTH implicit model and CAMPARI simulation engine yield ensembles that scale as ideal chains (i.e., R_e and R_g scale with the number of monomers to the power of 0.5). **E.** Internal scaling profiles for FRC simulations and Excluded Volume (EV) simulations for poly-alanine chains of varying lengths (filled circles demark the end of profiles for different polymer lengths). Internal scaling profiles map the average distance between all pairs of residues $|i-j|$ apart in sequence space, where i and j define two residues. This double average reports on the fact we average over both all pairs of residues that are $|i-j|$ apart and do so over all possible configurations. EV simulations show a characteristic tapering ("dangling end" effect) for large values of $|i-j|$. All FRC simulation profiles superimpose on top of one another, reflecting the absence of finite chain effects. **F.** Histograms of end-to-end distances (blue) taken from FRC simulations vs. corresponding probability density profiles generated by the Analytical FRC (AFRC) model (black line) show excellent agreement. **G.** Histograms of radii of gyration (red) taken from FRC simulations vs. corresponding probability density profiles generated by the AFRC model (black line) also show excellent agreement.

Constructing an analytical description of the Flory Random Coil

Our FRC ensembles enable the calculation of a range of polymeric properties, including inter-residue distances, inter-residue contact probabilities, the hydrodynamic radius, or the radius of gyration. Comparing these properties with experiments or simulations is often convenient, offering a standard reference frame for normalization and biophysical context[2,15,17,36,37]. However, performing and analyzing all-atom simulations with CAMPARI necessitates a level of computational sophistication that may make these calculations inaccessible to many scientists. To address this, we next sought to develop a set of closed-form analytical expressions to reproduce these properties and implement them as an easy-to-use package available both locally and – importantly – via a simple web interface (Google colab notebook).

FRC simulations generate ensembles that – by definition – reproduce the statistics expected for an ideal chain. As mentioned, polymer scaling behavior generally takes the form;

$$D = R_0 N^{\nu} \quad (1)$$

For an ideal chain, ν_{app} should not depend on the amino acid sequence (as all chains should scale with $\nu_{app} = 0.5$). However, the prefactor R_0 can and will show sequence dependence. As such, computing polymeric properties from sequences necessitates a means to calculate sequence-specific prefactor values. Prefactor values were parameterized using homopolymer simulations of each amino acid (see supplementary information). The inter-residue distance prefactor A_0 was parameterized by fitting internal scaling profiles using equation (2);

$$\sqrt{\langle\langle r^2_{(i,j)}\rangle\rangle} = A_0|i-j|^{\nu} \quad (2)$$

In equation 2, |i-j| is the number of residues between residues at position i and j, the left-hand-side reports on the root-mean-square (RMS) distance between residues i and j in the chain, ν is the scaling exponent (in our case this is equal to 0.5), and A_0 is a prefactor for which we can directly solve for. The double angle brackets around the RMS distance reflect the fact we are averaging over all pairs of residues that are |i-j| apart and doing so for all chain configurations. Plotting |i-j| vs. the RMSD generates the internal scaling profile shown in **Fig. 1E**. By fitting homopolymers of the 20 amino acids, a set of residue-specific A_0 prefactors was determined, as listed in **Supplementary Table 1**.

For our homopolymers, we can calculate the root-mean-squared end-to-end distances using equation (3);

$$\sqrt{\langle r^2_e\rangle} = A_0 N^{\nu} \quad (3)$$

From this, we can then use the standard function for P(r) of a Gaussian chain to calculate the end-to-end distance distribution;

$$P(r) = 4\pi r^2\left(\frac{3}{2\pi\langle r^2_e\rangle}\right)^{3/2} e^{-\left(\frac{3r^2}{2\langle r^2_e\rangle}\right)} \quad (4)$$

After determining residue-specific A_0, a comparison of analytical and numerical simulation distributions show excellent agreement when homopolymer end-to-end distance distributions are compared between FRC simulations and the AFRC-derived values (**Fig. 1f**).

We next took a similar route to define the radius of gyration (R_g) distribution. While no closed-form solution for the distribution of the radius of gyration exists, Lhuillier previously defined a closed-form approximation for this distribution for a fractal chain[62];

$$P_{Rg}(x) \sim N^{-\nu d} f(x)\left(\frac{x}{N^{\nu}}\right) \quad (5)$$

Where;

$$f(x) \sim exp\left[-\left(\frac{N^{\nu}}{x}\right)^{\alpha d} - \left(\frac{x}{N^{\nu}}\right)^{\delta}\right] \quad (6)$$

And the variables α and δ are defined as:

$$\alpha = \frac{1}{(\nu d - 1)} \quad (7)$$

$$\delta = \frac{1}{(1-\nu)} \quad (8)$$

Here, x represents the distance in some arbitrary units (written as such to avoid confusion with r, which represents the distance in Angstroms [Å]), N and ν again represent the total number of residues and the scaling exponent (0.5.), while d is the dimensionality ($d=3$). This allows us to calculate α and δ exactly, given ν is fixed at 0.5. Consequently, we can recast equation 5 into units of Å using a sequence-specific normalization factor (X_0);

$$r = X_0 x \quad (8)$$

To calculate X_0, we fit numerically-generated $P(R_g)$ distributions from homopolymer simulations with a series of analytically generated distributions to identify the best-fitting amino acid-specific X_0 values. These prefactors are listed in **Supplementary Table 1**. As with the end-to-end distances, a comparison of numerically-generated $P(R_g)$ with analytically-generated $P(R_g)$ values are in extremely good agreement (**Fig. 1g**). Comparing ensemble average end-to-end distance and radii of gyration for homopolymers of all 20 amino acids in lengths from 50 to 350 amino acids revealed a Pearson correlation coefficient of 0.999 and a root mean square error (RMSE) of 0.8 Å and 0.3 Å for the end-to-end distance and radius of gyration, respectively (**Fig. S2**).

With analytical expressions for computing the end-to-end distance and radius of gyration probability distributions in hand, we can calculate additional polymeric properties. Given the fractal nature of the Flory Random Coil and the absence of end effects, we can calculate all possible inter-residue distances and, correspondingly, contact frequencies between pairs of residues (**Fig. 2a, b**). Similarly, using either the Kirkwood-Riseman equation or a recently derived empirical relationship, we can compute an approximation for the ensemble-average hydrodynamic radius[63–65]. In summary, the AFRC offers an analytic approach for calculating sequence-specific ensemble properties for unfolded homopolymers.

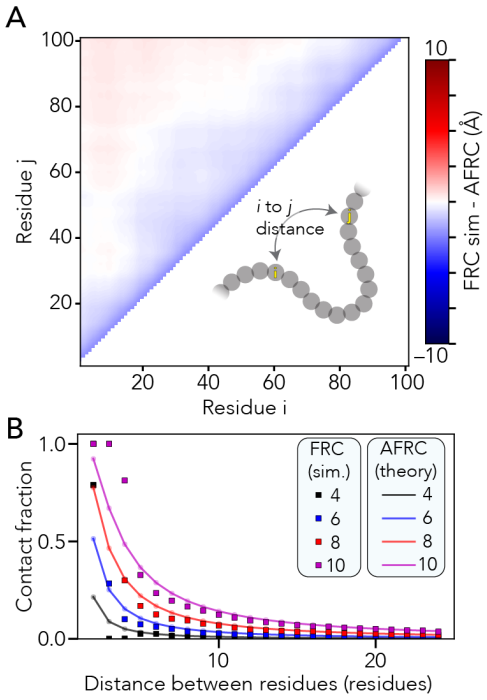

Figure 2. The AFRC enables the calculation of intra-residue distance distributions and expected distance-dependent contact fractions. A. We compared all-possible mean inter-residue distances obtained from FRC simulations with predictions from the AFRC. The maximum deviation across the entire chain is around 2.5 Å, with 92% of all distances having a deviation of less than 1 Å. **B.** Using the inter-residue distance, we can calculate the average fraction of an ensemble in which two residues are in contact (i.e., within some threshold distance). Here, we assess how that fractional contact varies with the contact threshold (different lines) and distance between the two residues. The AFRC does a somewhat poor job of estimating contact fractions for pairs of residues separated by 1,2 or 3 amino acids due to the discrete nature of the FRC simulations vis the continuous nature of the Gaussian chain distribution. However, the agreement is excellent above a sequence separation of three or more amino acids, suggesting that the AFRC offers a reasonable route to normalize expected contact frequencies.

Generalization to heteropolymers

Our parameterization has thus far focused exclusively on homopolymer sequences. However, Flory's rotational isomeric state approach requires complete independence of each amino residue[31,56]. Consequently, we expected the prefactor associated with a given heteropolymer to reflect a weighted average of prefactors taken from homopolymers, where the sequence composition determines the weights.

To test this expectation, we compared numerical simulations with AFRC predictions for a set of different polypeptide sequences finding excellent agreement in both end-to-end distances and radii of gyration (**Fig. 3a, b** and **Fig. S3**). Similarly, given the absence of end-effects, our

analytical end-to-end distance expression works equally well for intramolecular distances in addition to the end-to-end distance. To assess this, we compared internal scaling profiles between FRC simulations and AFRC predictions (**Fig. 3c**). These profiles compare the ensemble average distance between each possible inter-residue distance and offer a convenient means to assess both short and long-range intramolecular distances. We performed FRC simulations for 320 different polypeptide sequences ranging in length from 10 to 500 amino acids with a systematic variation in amino acid composition. Across all internal scaling profile comparisons between FRC and AFRC simulations, the overall average RMSE was 0.5 Å, with almost all (92%) of individual comparisons revealing an RMSE under 1 Å (**Fig. 3D**). Similarly, the Pearson's correlation coefficient between internal scaling profiles for FRC vs. AFRC for all ten-residue chains was 0.9993, which was the worst correlation across all lengths (**Fig. S4**). In summary, the AFRC faithfully reproduces homo- and hetero-polymeric dimensions for polypeptides under the FRC assumptions.

Comparison with existing polymer models

For completeness, we compared the end-to-end distance distributions obtained from several other polymer models used throughout the literature for describing unfolded and disordered polypeptides. Previously-used polymer models offer a means to analytically fit experimental or computational results and benefit from taking one (or more) parameters that define the model's behavior. While the AFRC does not enable fitting to experimental or simulated data, it only requires an amino acid sequence as input. With this in mind, the AFRC serves a fundamentally different purpose than commonly used models.

We wondered if dimensions obtained from the AFRC would be comparable with dimensions obtained from other polymer models when using parameters used previously in the literature. We compared distributions obtained from the worm-like chain (WLC), the self-avoiding walk (SAW) model, and a recently-developed ν-dependent self-avoiding walk (SAW-ν)[23,66]. For the WLC model, we used a persistence length of 3.0 Å and an amino acid size of 3.8 Å (such that the contour length, l_c, is defined as $N \times 3.8$[66]). For the SAW model, we used a scaling prefactor of 5.5 Å (i.e., assuming $\langle R_e \rangle = 5.5N^{0.598}$)[23,32,66]. Finally, for SAW-ν, we computed distributions using a prefactor of 5.5 Å and using several different ν values[6,23]. These values were chosen because previous studies have used them to describe intrinsically disordered proteins.

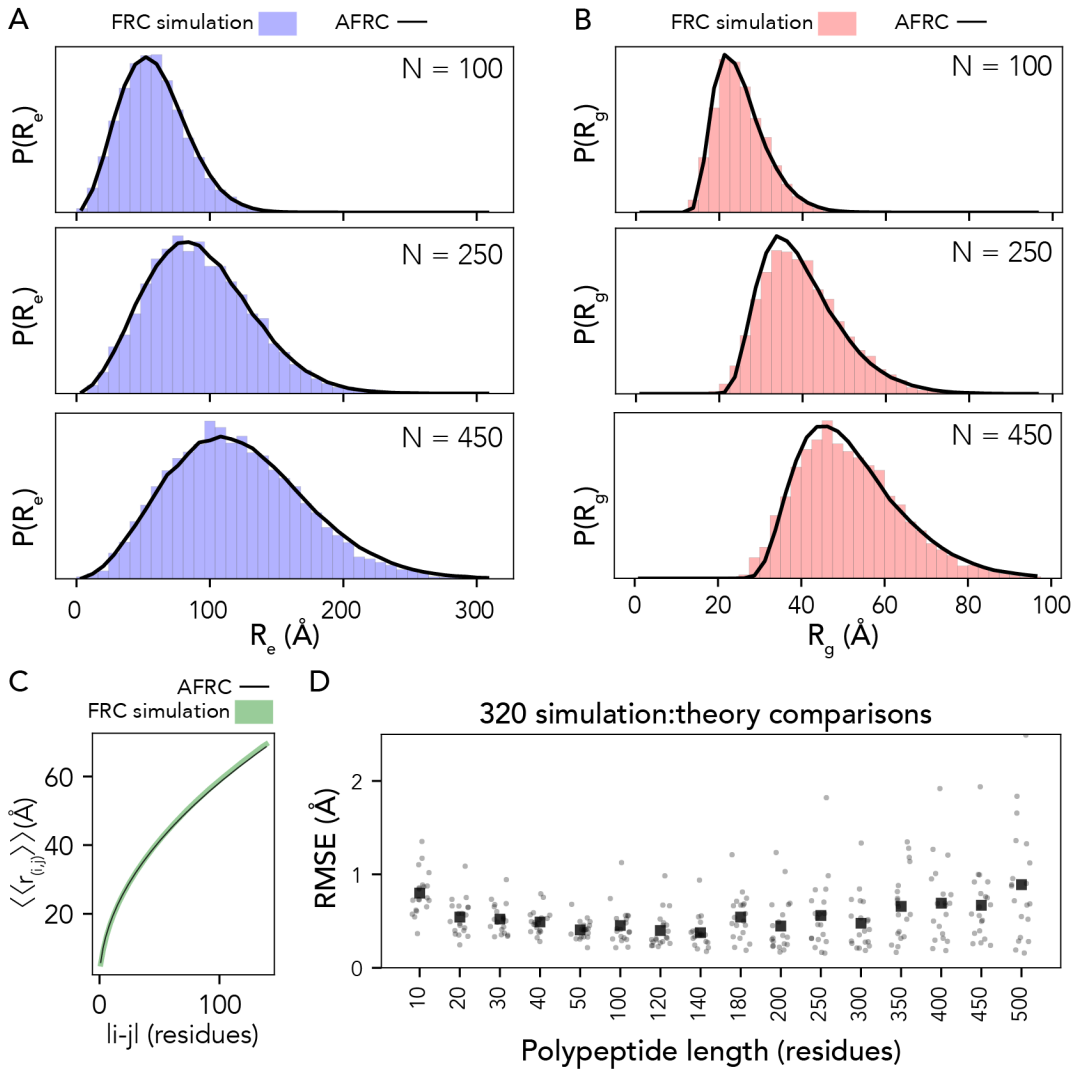

Fig. 3 The AFRC generalizes to arbitrary heteropolymeric sequences with the same precision and accuracy as it does for homopolymeric sequences. A. Representative examples of randomly polypeptide heteropolymers of lengths 100, 250, and 450, comparing the AFRC-derived end-to-end distance distribution (black curve) with the empirically-determined end-to-end distance histogram from FRC simulations (blue bars). **B**. The same three polymers, as shown in A, now compare the AFRC-derived radius of gyration distance distribution (black curve) with the empirically-determined radius of gyration histogram from FRC simulations (blue bars). **C** Comparison of AFRC vs. FRC simulation-derived internal scaling profiles for a 150-amino acid random heteropolymer. The deviation between FRC and AFRC for these profiles offers a measure of agreement across all length scales. **D** Comparison of root-mean-square error (RMSE) obtained from internal scaling profile comparisons (i.e., as shown in C) for 320 different heteropolymers straddling 10 to 500 amino acids in length. In all cases, the agreement with theory and simulations is excellent.

Fig. 4A shows comparisons of the AFRC distance distribution obtained for a 100-mer polyalanine (A_{100}) vs. the WLC and SAW (top) and vs. ν-dependent distributions (bottom). The AFRC is slightly more expanded than the WLC model using the parameters provided, although the persistence length can, of course, be varied to explore more compact (lower l_p) or more extended (higher l_p) distributions (**Fig. S6A**). The AFRC is substantially more compact than the SAW model. The comparison with the SAW model is important, as with a prefactor of 5.5 Å the SAW model describes a polypeptide as a self-avoiding random coil (ν=0.588), whereas the AFRC describes a polypeptide as an ideal chain (ν = 0.5), such that we should expect the SAW to be more expanded than the AFRC. Finally, in comparing the AFRC with the SAW-ν model, we find that the AFRC distribution falls almost completely top of the ν = 0.50 distribution. This indicates that both models arrive at nearly identical distance distributions despite being developed independently. This result is both confirmatory and convenient, as it means the AFRC and SAW-ν models can be used to analyze the same data without concern for model incompatibility.

We emphasize that this comparison with the existing polymer model is not presented to imply the AFRC is better than existing models but to highlight their compatibility. One can tune input parameters for all three models to arrive at qualitatively matching end-to-end distributions (**Fig. S6B**). The major difference between these three models and the AFRC is simply that the AFRC requires only amino acid sequence as input, making it a convenient reference point. For completeness, all four models are implemented in our Google colab notebook.

We also compared ensemble-average radii of gyration obtained from the various models with those obtained from the AFRC. While the WLC, SAW, and SAW-ν models do not provide approximate closed-form solutions for the radius of gyration distribution, they do enable an estimate of the ensemble-average radius of gyration to be calculated[23,66]. Using the same model parameters as was used in **Fig. 4A**, the AFRC falls between the SAW and the WLC. Moreover, the AFRC radii of gyration scale almost 1:1 with the SAW-ν derived radii as a function of chain length when ν = 0.50. As such, we conclude that the AFRC is consistent with existing polymer models yet benefits from being both parameter-free (for the user) and offering full distributions for the radius of gyration and intramolecular distance distributions per-residue contact fractions, convenient properties for normalization in simulations and experiment.

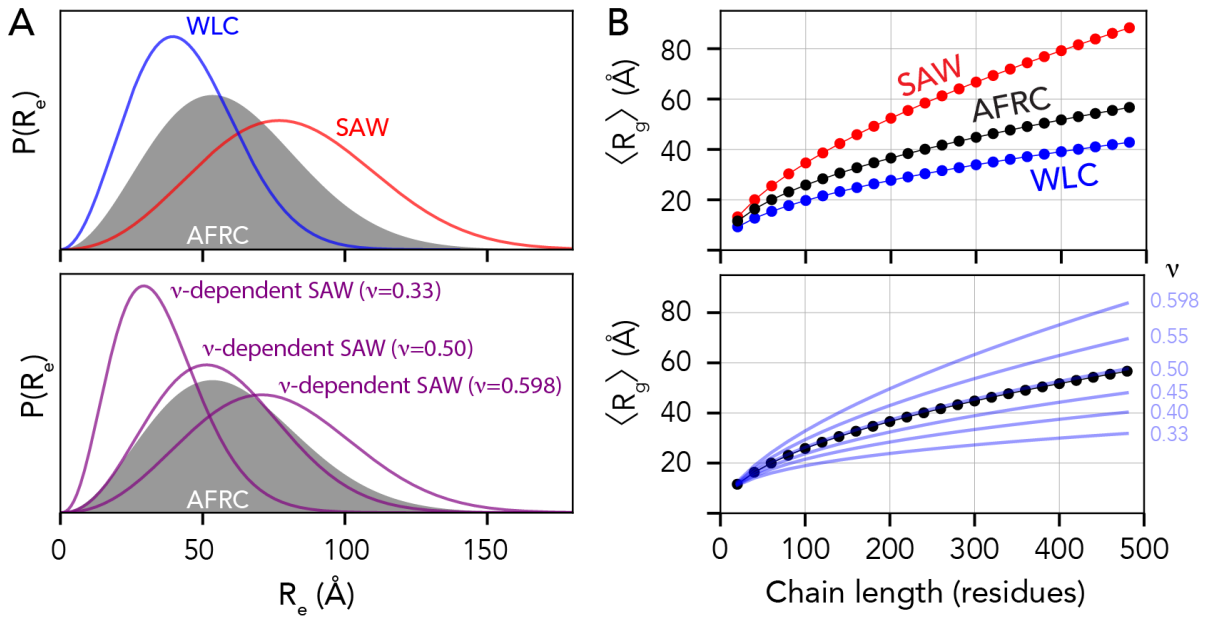

Fig. 4 The AFRC is complementary to existing polymer models. (A) Comparison of end-to-end distance distributions for several other analytical models, including the Wormlike Chain (WLC), the self-avoiding walk (SAW), and the ν-dependent SAW model (SAW-ν). The AFRC behaves like a ν-dependent SAW with a scaling exponent of 0.5. **(B)** Comparisons of ensemble-average radii of gyration as a function of chain length for the same sets of polymer models. The AFRC behaves as expected and again is consistent with a ν-dependent SAW with a scaling exponent of 0.5.

Comparison with all-atom simulations

Our work thus far has focussed on developing and testing the robustness of the AFRC. Having done this, we next sought to ask how similar (or dissimilar) distributions obtained from the AFRC are compared to all-atom simulations. We used simulations generated via all-atom molecular dynamics with the Amber99-disp forcefield and all-atom Monte Carlo simulations with the ABSINTH-OPLS forcefield[25,55,67–71]. Specifically, we examined nine different fully disordered proteins: The unfolded Drosophila Drk N-terminal SH2 domain (DrkN, 59 residues)[67,72,73], the ACTR domain of p160 (ACTR, 71 residues)[39,40,67,74], a C-terminal disordered subregion of the yeast transcription factor Ash1 (Ash1, 83 residues)[68], the N-terminal disordered regions of p53 (p53, 91 residues)[71,75], the C-terminal IDR of p27 (p26, 107 residues)[70], the intrinsically disordered intracellular domain of the notch receptor (Notch, 132 residues)[69], the C-terminal disordered domain of the measles virus nucleoprotein (Ntail, 132 residues)[67,76], the C-terminal low-complexity domain of hnRNPA1 (A1-LCD, 137 residues)[25], and full-length alpha-synuclein (asyn, 140 residues)[67,77,78].

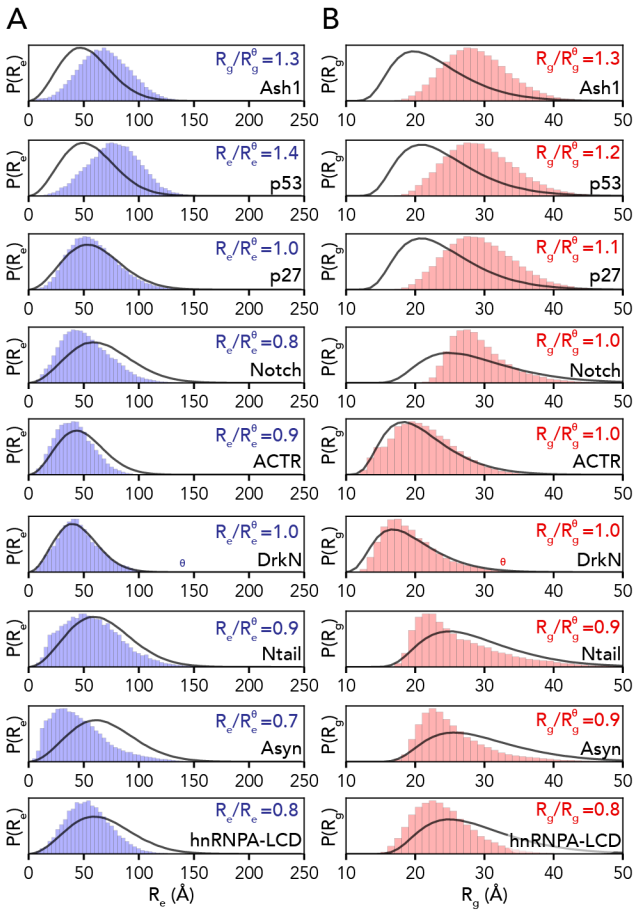

Fig. 5 AFRC-derived distance distributions enable simulations to be qualitatively compared against a null model. A. Comparison of the AFRC-derived end-to-end distance distributions (black line) with the simulation-derived end-to-end distribution (blue bars) for all-atom simulations of nine different disordered proteins. **B.** Comparison of the AFRC-derived radius of gyration distributions (black line) with the simulation-derived radius of gyration distribution (red bars) for all-atom simulations of nine different disordered proteins.

We compared distributions for the end-to-end distance and radius of gyration for our all-atom simulations with analogous distributions generated by the AFRC (**Fig. 5**). These comparisons revealed that while the general shape of the distributions recovered from simulations was not dissimilar from the AFRC-derived end-to-end distance and radius of gyration distributions, the width and mean were often different. This is hardly surprising, given that the global dimensions of an unfolded protein depend on the underlying amino acid sequence. The ratio of the mean end-to-end distance divided by the AFRC-derived mean end-to-end distance (or the corresponding ratio for the radius of gyration) was found to range between 0.7 and 1.4. In some cases, the end-to-end distance ratio or radius of gyration ratio varied within the same protein. For example, for the 132-residue intracellular-domain IDR from Notch (Notch), the end-to-end

distance ratio was 0.8 (i.e., smaller than predicted by the AFRC), while the radius of gyration ratio was 1.0. Similarly, in alpha-synuclein (Asyn), the corresponding ratios were 0.7 and 0.9, again reporting a smaller end-to-end distance than radius of gyration. As suggested previously, discrepancies in end-to-end distance vs. radius of gyration vs. expectations from homopolymer models are diagnostic of sequence-encoded conformational biases[18,35,36,79].

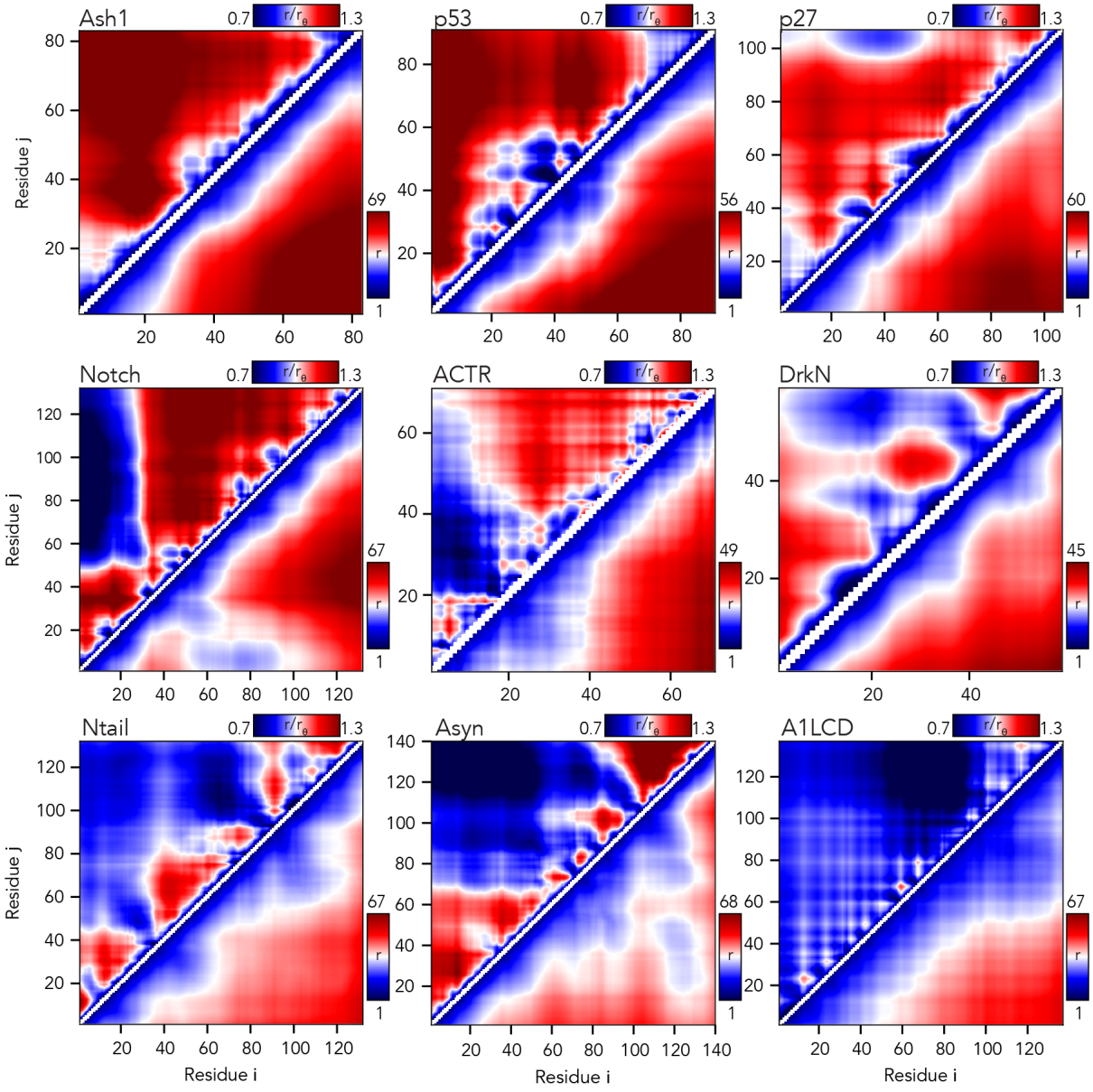

Fig. 6 The AFRC enables a consistent normalization of intra-chain distances to identify specific sub-regions that are closer or further apart than expected. Inter-residue scaling maps (top left) and distance maps (bottom right) reveal the nuance of intramolecular interactions. Scaling maps (top left) report the average distance between each pair of residues (i,j) divided by the distance expected for an AFRC-derived distance map, providing a unitless parameter that varies between 0.7 and 1.3 in these

simulations. Distance maps (bottom right) report the absolute distance between each pair of residues in angstroms. While distance maps provide a measure of absolute distance in real space, scaling maps provide a cleaner, normalized route to identify deviations from expected polymer behavior, offering a convenient means to identify sequence-specific effects. For example, in Notch and alpha-synuclein, scaling maps clearly identify end-to-end distances as close than expected. Scaling maps also offer a much sharper resolution for residue-specific effects - for example, in p53, residues embedded in the hydrophobic transactivation domains are clearly identified as engaging in transient intramolecular interactions, leading to sharp deviations from expected AFRC distances.

We also used the AFRC to calculate scaling maps. Scaling maps are non-redundant matrices of inter-residue distances obtained from simulations and normalized by the expected inter-residue distances obtained by the AFRC (**Fig. 6**)[68]. We compared these scaling maps (top left triangle of each panel) against absolute distances (bottom right triangle). This comparison highlights the advantage that using a reference polymer model offers. Long-range sequence-specific conformational biases are much more readily visualized as deviations from an expected polymer model. Moreover, the same dynamic range of values can be used for chains of different lengths, normalizing the units from Å to a unitless ratio.

Returning to the notch simulations, both types of intramolecular distance analysis clearly illustrate a strong long-range interaction between the N-terminal residues 1-30 and the remainder of the sequence. The long-range interaction between chain ends influences the end-to-end distance much more substantially than it does the radius of gyration (**Fig. 6**). Similarly, in alpha-synuclein, we observed long-range interactions between the negatively charged C-terminus and the positively-charged residues 20-50, leading to a reduction in the end-to-end distance. In short, the AFRC provides a convenient approach to enable direct interrogation of sequence-to-ensemble relationships in all-atom simulations.

Finally, we calculated per-residue contact scores for each residue in our nine proteins (**Fig. 7**). These contact scores sum the length-normalized fraction of the simulation in which each residue is in contact with any other residue in the sequence[25]. While this collapses information on residue-specificity into a single number, it integrates information from the typically-sparse contact maps for IDR ensembles to identify residues that may have an outside contribution towards short (<6 Å) range molecular interactions. We and others have previously used this approach to identify “stickers” - regions or residues in IDRs that have an outsized contribution to intra- and inter-molecular interactions[25,61,80,81].

In some proteins, specific residues or subregions were identified as contact hotspots. This includes the aliphatic residues in ACTR, and hydrophobic residues in the p53 transactivation

domains, in line with recent work identifying aliphatic residues as driving intramolecular interactions[61,82]. Most visually noticeable, aromatic residues in the A1-LCD appear as spikes that uniformly punctuate the sequence, highlighting their previously-identified role as evenly-spaced stickers[25]. Intriguingly, in alpha-synuclein, several regions in the aggregation-prone non-amyloid core (NAC) region (residues 61-95) appear as contact score spikes, potentially highlighting the ability of intramolecular interactions to guide regions or residues that may mediate inter-molecular interaction.

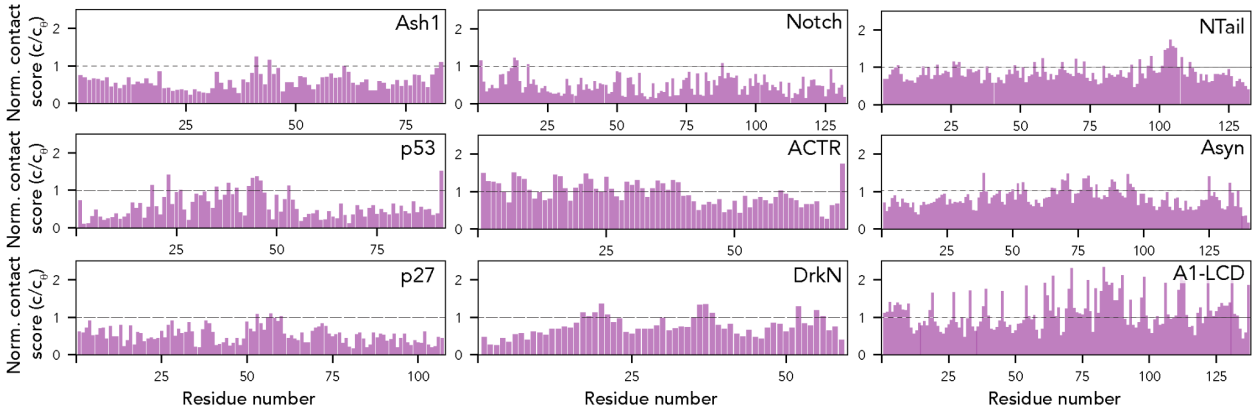

Fig. 7 The AFRC enables an expected contract fraction to be calculated, such that normalized contact frequencies can be easily calculated for simulations. Across the nine different simulated disordered proteins, we computed the contact fraction (i.e., the fraction of simulations each residue is in contact with any other residue) and divided this value by the expected contact fraction from the AFRC model. This analysis revealed subregions within IDRs that contribute extensively to intramolecular interactions, mirroring finer-grain conclusions obtained in **Fig. 6**.

Comparison with SAXS-derived radii of gyration

Having compared AFRC-derived parameters with all-atom simulations, we next sought to determine if AFRC-derived polymeric properties compared reasonably with experimentally-measured values. As a reminder, the AFRC is not a predictor of IDR behavior; instead, it offers a null model against which IDR dimensions can be compared. To perform a comparison with experimentally derived data, we curated a dataset of 145 examples of radii of gyration measured by small-angle X-ray scattering (SAXS) of disordered proteins. We choose to use SAXS data because SAXS-derived radii of gyration offer a label-free, model-free means to determine the overall dimensions of a disordered protein. That said, SAXS-derived measurements are not without their caveats (see discussion), and where possible, we re-analyzed primary scattering data to ensure all radii of gyration reported here are faithful and accurate.

To assess our SAXS-derived radii of gyration, we calculated expected dimensions for denatured proteins, folded globular domains, or AFRC chains by fitting scaling laws with the form $R_g = R_0 N^{\nu}$ against different polymer models. We used a denatured-state polymer model (ν = 0.59, R_0 = 1.98, as defined by Kohn *et al.*) and a folded globular domain model (ν = 0.33, R_0 = 2.86, as obtained from PDBSELECT25 originally plotted by Holehouse & Pappu)[11,48,83]. We also calculated the AFRC-derived radii of gyration for all 145 chains and fitted a polymer scaling model to the resulting data where the only free parameter was R_0 (ν = 0.50, R_0 = 2.50). This analysis showed that the majority of the 145 proteins have a radius of gyration above that of the AFRC-derived radius of gyration (see discussion), with some even exceeding the expected radius of gyration of a denatured protein (**Fig. 8A**). Based on these data, we determined an empirical upper and lower bound for the biologically accessible radii of gyration given a chain length (see discussion). This threshold suggests that, for a sequence of a given length, there is a wide range of possible IDR dimensions accessible (**Fig. 8B**, **Fig. S5**).

Finally, we wondered how well the AFRC-derived radii of gyration would correlate with experimentally-measured values. Based on the upper and lower bounds shown in **Fig. 8B**, we excluded four radii of gyration that appear to be spuriously large, leaving 141 data points. For these 141 points, we calculated the Pearson correlation coefficient (r) and the RMSE between the experimentally-measured radii of gyration and the AFRC-derived radii of gyration. This analysis yielded a correlation coefficient of 0.91 and an RMSE of 6.4 Å (**Fig. 8C**). To our surprise, these metrics outperform several established coarse-grained models for assessing intrinsically disordered proteins, as reported recently[84]. We again emphasize that the AFRC is not a predictor of IDR dimensions. However, we tentatively suggest that this result demonstrates that a reasonably good correlation between amino acid sequence and global dimensions can be obtained solely by recognizing that disordered proteins are flexible polymers. With this in mind, we conclude that the AFRC provides a convenient and easily-accessible control for experimentalists measuring the global dimensions of disordered proteins.

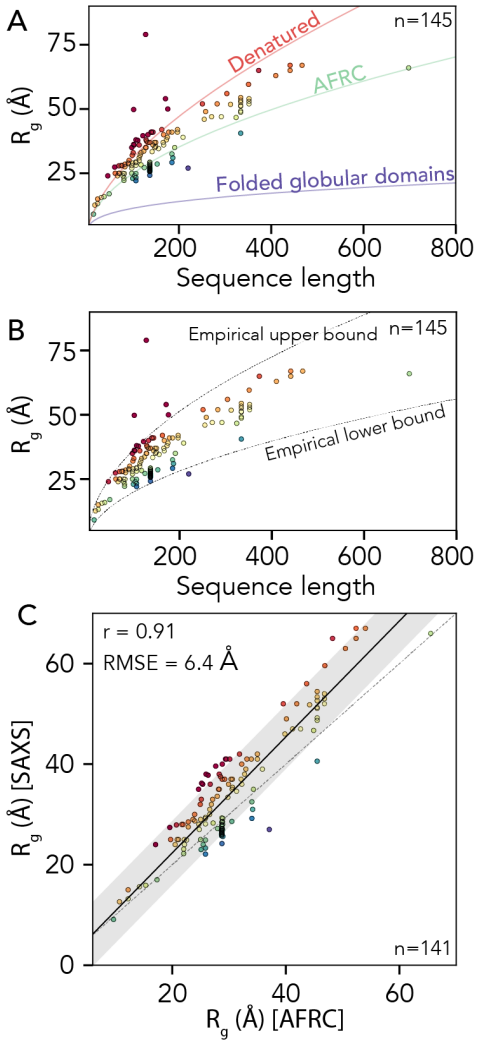

Fig 8. Comparison of AFRC-derived radii of gyration with experimentally-measured values. A. We compared 145 experimentally-measured radii of gyration against three empirical polymer scaling models that capture the three classes of polymer scaling (ν = 0.33 [globular domains], ν = 0.5 [AFRC], and ν = 0.59 [denatured state]). Individual points are colored by their normalized radius of gyration (SAXS-derived radius of gyration divided by AFRC-derived radius of gyration). **B.** The same data as in panel A with the empirically defined upper and lower bound. As with panel A, individual points are colored by their normalized radius of gyration. **C.** Comparison of SAXS-derived radii of gyration and AFRC-derived radii of gyration, as with panels A and B, individual points are colored by their normalized radius of gyration.

Reference implementation and distribution

Computational and theoretical tools are only as useful as they are usable. To facilitate the adoption of the AFRC as a convenient reference ensemble, we provide the AFRC as a stand-alone Python package distributed through PyPI (`pip install afrc`). We also implemented the additional polymer modes described in **Fig. 4** with a consistent programmatic interface, making it relatively straightforward to apply these models to analyze and interpret

computational and experimental data. Finally, to further facilitate access, we provide an easy-to-use Google colab notebook for calculating expected parameters for easy comparison with experiments and simulations. All information surrounding access to the AFRC model is provided at https://github.com/idptools/afrc.

DISCUSSION & CONCLUSION

In this work, we have developed and presented the Analytical Flory Random Coil (AFRC) as a simple-to-use reference model for comparing against simulations and experiments of unfolded and disordered proteins. We demonstrated that the AFRC behaves as a truly ideal chain and faithfully reproduces homo- and hetero-polymeric inter-residue and radius of gyration distributions obtained from explicit numerical simulations. We also compared the AFRC against several previously-established analytical polymer models, showing that ensemble-average or distribution data obtained from the AFRC are interoperable with existing models. Finally, we illustrated how the AFRC could be used as a null model for comparing data obtained from simulations and from experiments.

The AFRC differs from established polymer models in two key ways. While existing models define functional forms for polymeric properties, they do not prescribe specific length scales or parameters for those models. This is not a weakness - it simply reflects how analytical models work. However, the need to provide ‘appropriate’ parameters to ensure these models recapitulate behaviors expected for polypeptides places the burden on selecting and/or justifying those parameters on the user. The AFRC combines several existing analytical models (the Gaussian chain and the Lhuillier approximation for the radius of gyration distribution) with specific parameters obtained from numerical simulations to provide a “parameter-free” polymer model defined by its reference implementation (as opposed to the mathematical form of the underlying distributions). We place parameter free in quotation marks because the freedom from parameters is at the user level - the model itself is explicitly parameterized to reproduce polypeptides dimensions. However, from the user's perspective, no information is needed other than the amino acid sequence.

Although the AFRC was explicitly parameterized to recapitulate numerical FRC simulations, sequence-specific effects do not generally have a major impact on the resulting dimensions. For example, **Fig. S6** illustrates the radius of gyration or end-to-end distance obtained for varying lengths of poly-alanine and poly-glycine. This behavior is not a weakness of the model - it *is* the model. This relatively modest sequence dependence reflects the fact that for an ideal chain, both the second and third virial coefficients are set to zero (*i.e.*, the integral of Mayer f-function should equal 0)[85]. As such, the AFRC does not enable explicitly excluded volume contributions to the chain’s dimensions from sidechain volume, although this is captured implicitly based on

the allowed isomeric states (compare glycine to alanine in **Fig. S1**). In summary, the AFRC does not offer any new physics, but it does encapsulate previously derived physical models along with numerically-derived sequence-specific parameters to make it easy to construct null models explicitly for comparison with polypeptides.

In comparing AFRC-derived polymeric properties with those obtained from all-atom simulations, we recapitulate sequence-to-ensemble features identified previously [25,28,67,69]. When comparing the normalized radii of gyration (R_g^{Sim}/ R_g^{AFRC}), we noticed the lower and upper bounds obtained here appear to be approximately 0.8 and 1.4, respectively. To assess if this trend held true for experimentally derived radii of gyration, we calculated the normalized radii of gyration for the 141 values reported in **Fig. 8C**, recapitulating a similar range (0.8 to 1.46). Based on these values, we defined an empirical boundary condition for the anticipated range in which we would expect to see a disordered chain's radius of gyration as between $0.8R_g^{AFRC}$ and $1.45R_g^{AFRC}$ (**Fig. 8B**). We emphasize this is not a hard threshold. However, it offers a convenient rule-of-thumb, such that measured radii of gyration can be compared against this value to assess if a potentially spurious radius of gyration has been obtained (either from simulations or experiments). Such a spurious value does not necessarily imply a problem, but may warrant further investigation to explain its physical origins.

Our comparison with experimental data focussed on radii of gyration obtained from SAXS experiments. We chose this route given the wealth of data available and the label-free and model-free nature in which SAXS data are collected and analyzed. Given the AFRC offers the expected dimensions for a polypeptide behaving qualitatively as if it is in a theta solvent, it may be tempting to conclude from these data that the vast majority of disordered proteins are found in a good solvent environment (**Fig. 9A**). The solvent environment reflects the mean-field interaction between a protein and its environment. In the good solvent regime, protein:solvent interactions are favored, while in the poor solvent regime protein:protein interactions are favored [2,6,44,48]. However, it is worth bearing in mind that SAXS experiments generally require relatively high concentrations of protein to obtain reasonable signal-to-noise[43]. Recent advances in size exclusion chromatography (SEC) coupled SAXS have enabled the collection of scattering data for otherwise aggregation-prone proteins with great success[86]. However, there is still a major acquisition bias in the technical need of these experiments to work with high concentrations of soluble proteins when integrated over all existing measured data. By definition, such highly soluble proteins experience a good solvent environment. Given this acquisition bias, we remain agnostic as to whether these results can be used to extrapolate to the solution behavior of all IDRs.

Prior work has implicated the presence of charged and proline residues as mediating IDR chain expansion [33,34,42,46,49,57,68,87–90]. We took advantage of the fact that the AFRC enables a length normalization of experimental radii of gyration and assessed the normalized radius of gyration vs. the fraction of charged and proline residue (**Fig. 9B**). Our data support this conclusion as a first approximation, but also clearly demonstrate that while this trend is true on average, there is variance in this relationship. Notably, for IDRs with a fraction of charged and proline residues between 0.2 and 0.4, the full range of possible IDR dimensions are accessible. The transition from (on average) more compact to (on average) more expanded chains occurs around a fraction of proline and charged residues of around 0.25 – 0.30, in qualitative agreement with prior work exploring the fraction of charge residues required to drive chain expansion [33,34,42]. However, we emphasize that there is massive variability observed on a per-sequence basis. In summary, while the presence of charged and proline residues clearly influences IDR dimensions, complex patterns of intramolecular interactions can further tune this behavior [2,17,28].

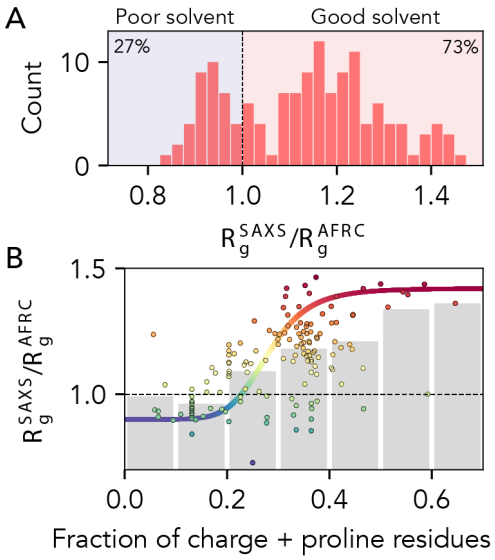

Fig 9. AFRC-normalized radii of gyration from experimentally-measured proteins. A. Histogram showing the normalized radii of gyration for 141 different experimentally-measured sequences. **B.** Comparison of normalized radii of gyration for 141 different experimentally-measured sequences against the fraction of charge and proline residues in those sequences. Individual points are colored by their normalized radius of gyration. Grey bars reflect the average radius of gyrations obtained by binning sequences with the corresponding fraction of charge and proline residues. The colored sigmoidal curve is included to guide the eye across the transition region, suggesting that – on average – the midpoint of this transition is at a fraction of charged and proline residues of ~0.25. The Pearson correlation coefficient (r) for the fraction of charged and proline residues vs. normalized radius of gyration is 0.58).

In summary, the AFRC offers a convenient, analytical approach to obtain a well-defined reference state for comparing and contrasting simulations and experiments of unfolded and disordered proteins. It can be easily integrated into complex analysis pipelines, or used for one-off analysis via a Google Colab notebook without requiring any computational expertise at all.

ACKNOWLEDGEMENTS

We thank members of the Pappu lab and Holehouse lab for many useful discussions over the years. We are indebted to Dr. Nick Lyle for the original implementation of the CAMPARI-based FRC engine. We thank Dr. Erik Martin for bringing the work of Lhuillier to our attention. Funding for this work was provided by the National Institute on Allergic and Infectious Diseases with R01AI163142 to A.S.H. and A.S., by the National Science Foundation with 2128068 to A.S.H., by the Longer Life Foundation, an RGA/Washington University in St. Louis Collaboration to A.S.H., and by the National Cancer Institute with F99CA264413 to J.J.A. We also thank members of the Water and Life Interface Institute (WALII), supported by NSF DBI grant #2213983, for helpful discussions.

REFERENCES

(1) Dill, K. A.; Shortle, D. Denatured States of Proteins. *Annu. Rev. Biochem.* **1991**, *60*, 795–825.

(2) Mao, A. H.; Lyle, N.; Pappu, R. V. Describing Sequence–ensemble Relationships for Intrinsically Disordered Proteins. *Biochem. J* **2013**, *449* (2), 307–318.

(3) Chan, H. S.; Dill, K. A. Polymer Principles in Protein Structure and Stability. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 447–490.

(4) Pappu, R. V.; Wang, X.; Vitalis, A.; Crick, S. L. A Polymer Physics Perspective on Driving Forces and Mechanisms for Protein Aggregation - Highlight Issue: Protein Folding. *Arch. Biochem. Biophys.* **2008**, *469* (1), 132–141.

(5) Schuler, B.; Soranno, A.; Hofmann, H.; Nettels, D. Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins. *Annu. Rev. Biophys.* **2016**, *45*, 207–231.

(6) Soranno, A. Physical Basis of the Disorder-Order Transition. *Arch. Biochem. Biophys.* **2020**, *685*, 108305.

(7) Lin, Y.-H.; Forman-Kay, J. D.; Chan, H. S. Theories for Sequence-Dependent Phase Behaviors of Biomolecular Condensates. *Biochemistry* **2018**, *57* (17), 2499–2508.

(8) Thirumalai, D.; O'Brien, E. P.; Morrison, G.; Hyeon, C. Theoretical Perspectives on Protein Folding. *Annu. Rev. Biophys.* **2010**, *39*, 159–183.

(9) Cubuk, J.; Soranno, A. Macromolecular Crowding and Intrinsically Disordered Proteins: A Polymer Physics Perspective. *ChemSystemsChem* **2022**. https://doi.org/10.1002/syst.202100051.

(10) Wilkins, D. K.; Grimshaw, S. B.; Receveur, V.; Dobson, C. M.; Jones, J. A.; Smith, L. J. Hydrodynamic Radii of Native and Denatured Proteins Measured by Pulse Field Gradient NMR Techniques. *Biochemistry* **1999**, *38* (50), 16424–16431.

(11) 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.* **2004**, *101* (34), 12491–12496.

(12) Damaschun, G.; Damaschun, H.; Gast, K.; Zirwer, D.; Bychkova, V. E. Solvent Dependence of Dimensions of Unfolded Protein Chains. *Int. J. Biol. Macromol.* **1991**, *13* (4), 217–221.

(13) Calmettes, P.; Durand, D.; Desmadril, M.; Minard, P.; Receveur, V.; Smith, J. C. How Random Is a Highly Denatured Protein? *Biophys. Chem.* **1994**, *53* (1-2), 105–113.

(14) Mok, Y. K.; Kay, C. M.; Kay, L. E.; Forman-Kay, J. NOE Data Demonstrating a Compact Unfolded State for an SH3 Domain under Non-Denaturing Conditions. *J. Mol. Biol.* **1999**, *289* (3), 619–638.

(15) Peran, I.; Holehouse, A. S.; Carrico, I. S.; Pappu, R. V.; Bilsel, O.; Raleigh, D. P. Unfolded States under Folding Conditions Accommodate Sequence-Specific Conformational Preferences with Random Coil-like Dimensions. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (25), 12301–12310.

(16) Meng, W.; Luan, B.; Lyle, N.; Pappu, R. V.; Raleigh, D. P. The Denatured State Ensemble Contains Significant Local and Long-Range Structure under Native Conditions: Analysis of the N-Terminal Domain of Ribosomal Protein L9. *Biochemistry* **2013**, *52* (15), 2662–2671.

(17) Das, R. K.; Ruff, K. M.; Pappu, R. V. Relating Sequence Encoded Information to Form and Function of Intrinsically Disordered Proteins. *Curr. Opin. Struct. Biol.* **2015**, *32* (0), 102–112.

(18) Fuertes, G.; Banterle, N.; Ruff, K. M.; Chowdhury, A.; Mercadante, D.; Koehler, C.; Kachala, M.; Estrada Girona, G.; Milles, S.; Mishra, A.; Onck, P. R.; Gräter, F.; Esteban-Martín, S.; Pappu, R. V.; Svergun, D. I.; Lemke, E. A. Decoupling of Size and Shape Fluctuations in Heteropolymeric Sequences Reconciles Discrepancies in SAXS vs. FRET Measurements.

Proc. Natl. Acad. Sci. U. S. A. **2017**, *114* (31), E6342–E6351.
(19) Milles, S.; Mercadante, D.; Aramburu, I. V.; Jensen, M. R.; Banterle, N.; Koehler, C.; Tyagi, S.; Clarke, J.; Shammas, S. L.; Blackledge, M.; Gräter, F.; Lemke, E. A. Plasticity of an Ultrafast Interaction between Nucleoporins and Nuclear Transport Receptors. *Cell* **2015**, *163* (3), 734–745.
(20) 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.* **2020**, *142* (37), 15697–15710.
(21) Zosel, F.; Mercadante, D.; Nettels, D.; Schuler, B. A Proline Switch Explains Kinetic Heterogeneity in a Coupled Folding and Binding Reaction. *Nat. Commun.* **2018**, *9* (1), 3332.
(22) Borgia, A.; Borgia, M. B.; Bugge, K.; Kissling, V. M.; Heidarsson, P. O.; Fernandes, C. B.; Sottini, A.; Soranno, A.; Buholzer, K. J.; Nettels, D.; Kragelund, B. B.; Best, R. B.; Schuler, B. Extreme Disorder in an Ultrahigh-Affinity Protein Complex. *Nature* **2018**, *555* (7694), 61–66.
(23) Zheng, W.; Zerze, G. H.; Borgia, A.; Mittal, J.; Schuler, B.; Best, R. B. Inferring Properties of Disordered Chains from FRET Transfer Efficiencies. *J. Chem. Phys.* **2018**, *148* (12), 123329.
(24) Chung, H. S.; Piana-Agostinetti, S.; Shaw, D. E.; Eaton, W. A. Structural Origin of Slow Diffusion in Protein Folding. *Science* **2015**, *349* (6255), 1504–1510.
(25) Martin, E. W.; Holehouse, A. S.; Peran, I.; Farag, M.; Incicco, J. J.; Bremer, A.; Grace, C. R.; Soranno, A.; Pappu, R. V.; Mittag, T. Valence and Patterning of Aromatic Residues Determine the Phase Behavior of Prion-like Domains. *Science* **2020**, *367* (6478), 694–699.
(26) Alston, J. J.; Soranno, A.; Holehouse, A. S. Integrating Single-Molecule Spectroscopy and Simulations for the Study of Intrinsically Disordered Proteins. *Methods* **2021**, *193*, 116–135.
(27) Bottaro, S.; Lindorff-Larsen, K. Biophysical Experiments and Biomolecular Simulations: A Perfect Match? *Science* **2018**, *361* (6400), 355–360.
(28) 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**. https://doi.org/10.1101/2023.02.16.528879.
(29) Rubinstein, M.; Colby, R. H. *Polymer Physics*; Oxford University Press: New York, 2003.
(30) de Gennes, P. G. *Scaling Concepts in Polymer Physics*; Cornell University Press: Ithaca, N.Y., 1979.
(31) Flory, P. J. *Statistical Mechanics of Chain Molecules*; Oxford University Press: New York, 1969.
(32) Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B. Polymer Scaling Laws of Unfolded and Intrinsically Disordered Proteins Quantified with Single-Molecule Spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (40), 16155–16160.
(33) Mao, A. H.; Crick, S. L.; Vitalis, A.; Chicoine, C. L.; Pappu, R. V. Net Charge per Residue Modulates Conformational Ensembles of Intrinsically Disordered Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (18), 8183–8188.
(34) 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.* **2010**, *107* (33), 14609–14614.
(35) Ruff, K. M.; Holehouse, A. S. SAXS versus FRET: A Matter of Heterogeneity? *Biophys. J.* **2017**, *113* (5), 971–973.
(36) Song, J.; Gomes, G.-N.; Shi, T.; Gradinaru, C. C.; Chan, H. S. Conformational Heterogeneity and FRET Data Interpretation for Dimensions of Unfolded Proteins. *Biophys. J.* **2017**, *113* (5), 1012–1024.
(37) Stenzoski, N. E.; Zou, J.; Piserchio, A.; Ghose, R.; Holehouse, A. S.; Raleigh, D. P. The

Cold-Unfolded State Is Expanded but Contains Long- and Medium-Range Contacts and Is Poorly Described by Homopolymer Models. *Biochemistry* **2020**, *59* (36), 3290–3299.
(38) Canchi, D. R.; García, A. E. Cosolvent Effects on Protein Stability. *Annual Reviews of Physical Chemistry* **2013**, *64*, 273–293.
(39) Borgia, A.; Zheng, W.; Buholzer, K.; Borgia, M. B.; Schüler, A.; Hofmann, H.; Soranno, A.; Nettels, D.; Gast, K.; Grishaev, A.; Best, R. B.; Schuler, B. Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *J. Am. Chem. Soc.* **2016**, *138* (36), 11714–11726.
(40) Zheng, W.; Borgia, A.; Buholzer, K.; Grishaev, A.; Schuler, B.; Best, R. B. Probing the Action of Chemical Denaturant on an Intrinsically Disordered Protein by Simulation and Experiment. *J. Am. Chem. Soc.* **2016**, *138* (36), 11702–11713.
(41) Tran, H. T.; Pappu, R. V. Toward an Accurate Theoretical Framework for Describing Ensembles for Proteins under Strongly Denaturing Conditions. *Biophys. J.* **2006**, *91* (5), 1868–1886.
(42) Marsh, J. A.; Forman-Kay, J. D. Sequence Determinants of Compaction in Intrinsically Disordered Proteins. *Biophys. J.* **2010**, *98* (10), 2383–2390.
(43) Bernadó, P.; Svergun, D. I. Structural Analysis of Intrinsically Disordered Proteins by Small-Angle X-Ray Scattering. *Mol. Biosyst.* **2011**, *8* (1), 151–167.
(44) Riback, J. A.; Bowman, M. A.; Zmyslowski, A. M.; Knoverek, C. R.; Jumper, J. M.; Hinshaw, J. R.; Kaye, E. B.; Freed, K. F.; Clark, P. L.; Sosnick, T. R. Innovative Scattering Analysis Shows That Hydrophobic Disordered Proteins Are Expanded in Water. *Science* **2017**, *358* (6360), 238–241.
(45) Aznauryan, M.; Delgado, L.; Soranno, A.; Nettels, D.; Huang, J.-R.; Labhardt, A. M.; Grzesiek, S.; Schuler, B. Comprehensive Structural and Dynamical View of an Unfolded Protein from the Combination of Single-Molecule FRET, NMR, and SAXS. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (37), E5389–E5398.
(46) 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.* **2022**, *14* (2), 196–207.
(47) Martin, E. W.; Holehouse, A. S. Intrinsically Disordered Protein Regions and Phase Separation: Sequence Determinants of Assembly or Lack Thereof. *Emerg Top Life Sci* **2020**, *4* (3), 307–329.
(48) Holehouse, A. S.; Pappu, R. V. Collapse Transitions of Proteins and the Interplay Among Backbone, Sidechain, and Solvent Interactions. *Annu. Rev. Biophys.* **2018**, *47*, 19–39.
(49) Huihui, J.; Ghosh, K. An Analytical Theory to Describe Sequence-Specific Inter-Residue Distance Profiles for Polyampholytes and Intrinsically Disordered Proteins. *J. Chem. Phys.* **2020**, *152* (16), 161102.
(50) Sawle, L.; Ghosh, K. A Theoretical Method to Compute Sequence Dependent Configurational Properties in Charged Polymers and Proteins. *J. Chem. Phys.* **2015**, *143* (8), 085101.
(51) Firman, T.; Ghosh, K. Sequence Charge Decoration Dictates Coil-Globule Transition in Intrinsically Disordered Proteins. *J. Chem. Phys.* **2018**, *148* (12), 123305.
(52) Das, S.; Lin, Y.-H.; Vernon, R. M.; Forman-Kay, J. D.; Chan, H. S. Comparative Roles of Charge, π, and Hydrophobic Interactions in Sequence-Dependent Phase Separation of Intrinsically Disordered Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (46), 28795–28805.
(53) Lin, Y.-H.; Chan, H. S. Phase Separation and Single-Chain Compactness of Charged Disordered Proteins Are Strongly Correlated. *Biophys. J.* **2017**, *112* (10), 2043–2046.
(54) Vitalis, A.; Pappu, R. V. Chapter 3 Methods for Monte Carlo Simulations of Biomacromolecules. In *Annual Reports in Computational Chemistry*; Wheeler, R. A., Ed.; Elsevier, 2009; Vol. 5, pp 49–76.

(55) Vitalis, A.; Pappu, R. V. ABSINTH: A New Continuum Solvation Model for Simulations of Polypeptides in Aqueous Solutions. *J. Comput. Chem.* **2009**, *30* (5), 673–699.
(56) Volkenstein, M. V. *Molecular Biophysics*; Academic Press, New York, 1977.
(57) Das, R. K.; Pappu, R. V. Conformations of Intrinsically Disordered Proteins Are Influenced by Linear Sequence Distributions of Oppositely Charged Residues. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (33), 13392–13397.
(58) Holehouse, A. S.; Garai, K.; Lyle, N.; Vitalis, A.; Pappu, R. V. Quantitative Assessments of the Distinct Contributions of Polypeptide Backbone Amides versus Side Chain Groups to Chain Expansion via Chemical Denaturation. *J. Am. Chem. Soc.* **2015**, *137* (8), 2984–2995.
(59) Kentsis, A.; Mezei, M.; Gindin, T.; Osman, R. Unfolded State of Polyalanine Is a Segmented Polyproline II Helix. *Proteins* **2004**, *55* (3), 493–501.
(60) Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V. Fluorescence Correlation Spectroscopy Shows That Monomeric Polyglutamine Molecules Form Collapsed Structures in Aqueous Solutions. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (45), 16764–16769.
(61) 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* **2022**, *82* (17), 3193–3208.e8.
(62) Lhuillier, D. A Simple-Model for Polymeric Fractals in a Good Solvent and an Improved Version of the Flory Approximation. *Journal De Physique* **1988**, *49* (5), 705–710.
(63) Nygaard, M.; Kragelund, B. B.; Papaleo, E.; Lindorff-Larsen, K. An Efficient Method for Estimating the Hydrodynamic Radius of Disordered Protein Conformations. *Biophys. J.* **2017**, *113* (3), 550–557.
(64) Kirkwood, J. G.; Riseman, J. The Intrinsic Viscosities and Diffusion Constants of Flexible Macromolecules in Solution. *J. Chem. Phys.* **1948**, *16* (6), 565–573.
(65) Pesce, F.; Newcombe, E. A.; Seiffert, P.; Tranchant, E. E.; Olsen, J. G.; Grace, C. R.; Kragelund, B. B.; Lindorff-Larsen, K. Assessment of Models for Calculating the Hydrodynamic Radius of Intrinsically Disordered Proteins. *Biophys. J.* **2022**. https://doi.org/10.1016/j.bpj.2022.12.013.
(66) O'Brien, E. P.; Morrison, G.; Brooks, B. R.; Thirumalai, D. How Accurate Are Polymer Models in the Analysis of Forster Resonance Energy Transfer Experiments on Proteins? *J. Chem. Phys.* **2009**, *130* (12), 124903.
(67) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (21), E4758–E4766.
(68) Martin, E. W.; Holehouse, A. S.; Grace, C. R.; Hughes, A.; Pappu, R. V.; Mittag, T. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J. Am. Chem. Soc.* **2016**, *138* (47), 15323–15335.
(69) Sherry, K. P.; Das, R. K.; Pappu, R. V.; Barrick, D. Control of Transcriptional Activity by Design of Charge Patterning in the Intrinsically Disordered RAM Region of the Notch Receptor. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (44), E9243–E9252.
(70) Das, R. K.; Huang, Y.; Phillips, A. H.; Kriwacki, R. W.; Pappu, R. V. Cryptic Sequence Features within the Disordered Protein p27Kip1 Regulate Cell Cycle Signaling. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (20), 5616–5621.
(71) Holehouse, A. S.; Sukenik, S. Controlling Structural Bias in Intrinsically Disordered Proteins Using Solution Space Scanning. *J. Chem. Theory Comput.* **2020**, *16* (3), 1794–1805.
(72) Marsh, J. A.; Neale, C.; Jack, F. E.; Choy, W.-Y.; Lee, A. Y.; Crowhurst, K. A.; Forman-Kay, J. D. Improved Structural Characterizations of the drkN SH3 Domain Unfolded State Suggest a Compact Ensemble with Native-like and Non-Native Structure. *J. Mol. Biol.* **2007**, *367* (5), 1494–1510.
(73) Bezsonova, I.; Singer, A.; Choy, W.-Y.; Tollinger, M.; Forman-Kay, J. D. Structural

Comparison of the Unstable drkN SH3 Domain and a Stable Mutant. *Biochemistry* **2005**, *44* (47), 15550–15560.
(74) Demarest, S. J.; Martinez-Yamout, M.; Chung, J.; Chen, H.; Xu, W.; Dyson, H. J.; Evans, R. M.; Wright, P. E. Mutual Synergistic Folding in Recruitment of CBP/p300 by p160 Nuclear Receptor Coactivators. *Nature* **2002**, *415* (6871), 549–553.
(75) Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen, M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. Structure of Tumor Suppressor p53 and Its Intrinsically Disordered N-Terminal Transactivation Domain. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (15), 5762–5767.
(76) Longhi, S.; Receveur-Bréchot, V.; Karlin, D.; Johansson, K.; Darbon, H.; Bhella, D.; Yeo, R.; Finet, S.; Canard, B. The C-Terminal Domain of the Measles Virus Nucleoprotein Is Intrinsically Disordered and Folds upon Binding to the C-Terminal Moiety of the Phosphoprotein. *J. Biol. Chem.* **2003**, *278* (20), 18638–18648.
(77) Syme, C. D.; Blanch, E. W.; Holt, C.; Jakes, R.; Goedert, M.; Hecht, L.; Barron, L. D. A Raman Optical Activity Study of Rheomorphism in Caseins, Synucleins and Tau. New Insight into the Structure and Behaviour of Natively Unfolded Proteins. *Eur. J. Biochem.* **2002**, *269* (1), 148–156.
(78) Theillet, F.-X.; Binolfi, A.; Bekei, B.; Martorana, A.; Rose, H. M.; Stuiver, M.; Verzini, S.; Lorenz, D.; van Rossum, M.; Goldfarb, D.; Selenko, P. Structural Disorder of Monomeric α-Synuclein Persists in Mammalian Cells. *Nature* **2016**, *530* (7588), 45–50.
(79) 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*, 2022, 2021.11.24.469609. https://doi.org/10.1101/2021.11.24.469609.
(80) Mohanty, P.; Shenoy, J.; Rizuan, A.; Mercado Ortiz, J. F.; Fawzi, N. L.; Mittal, J. Aliphatic Residues Contribute Significantly to the Phase Separation of TDP-43 C-Terminal Domain. *bioRxiv*, 2022, 2022.11.10.516004. https://doi.org/10.1101/2022.11.10.516004.
(81) Murthy, A. C.; Tang, W. S.; Jovic, N.; Janke, A. M.; Seo, D. H.; Perdikari, T. M.; Mittal, J.; Fawzi, N. L. Molecular Interactions Contributing to FUS SYGQ LC-RGG Phase Separation and Co-Partitioning with RNA Polymerase II Heptads. *Nat. Struct. Mol. Biol.* **2021**, *28* (11), 923–935.
(82) Rekhi, S.; Devarajan, D. S.; Howard, M. P.; Kim, Y. C.; Nikoubashman, A.; Mittal, J. Role of Strong Localized vs. Weak Distributed Interactions in Disordered Protein Phase Separation. *bioRxiv*, 2023, 2023.01.27.525976. https://doi.org/10.1101/2023.01.27.525976.
(83) Griep, S.; Hobohm, U. PDBselect 1992–2009 and PDBfilter-Select. *Nucleic Acids Res.* **2009**, *38* (Database issue), D318–D319.
(84) 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* **2021**, *1* (11), 732–743.
(85) Dill, K.; Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*; Garland Science, 2010.
(86) 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.* **2021**, *646*, 185–222.
(87) 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.* **2022**, *29* (8), 781–790.
(88) Zerze, G. H.; Best, R. B.; Mittal, J. Sequence- and Temperature-Dependent Properties of

Unfolded and Disordered Proteins from Atomistic Simulations. *J. Phys. Chem. B* **2015**, *119* (46), 14622–14630.

(89) Sørensen, C. S.; Kjaergaard, M. Effective Concentrations Enforced by Intrinsically Disordered Linkers Are Governed by Polymer Physics. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (46), 23124–23131.

(90) Riback, J. A.; Katanski, C. D.; Kear-Scott, J. L.; Pilipenko, E. V.; Rojek, A. E.; Sosnick, T. R.; Drummond, D. A. Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. *Cell* **2017**, *168* (6), 1028–1040.e19.