Published in final edited form as:
J Phys Chem Lett. 2020 May 07; 11(9): 3408–3415. doi:10.1021/acs.jpclett.0c00288.

Hydropathy Patterning Complements Charge Patterning to Describe Conformational Preferences of Disordered Proteins

Wenwei Zheng[†], **Gregory Dignon**[‡,§], **Matthew Brown**[†], **Young C. Kim**[¶], **Jeetain Mittal**[‡]

[†]College of Integrative Sciences and Arts, Arizona State University, Mesa, AZ 85212, USA

[‡]Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, PA 18015, USA

[¶]Center for Materials Physics and Technology, Naval Research Laboratory, Washington, DC 20375, United States

[§]Current address: Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794, USA

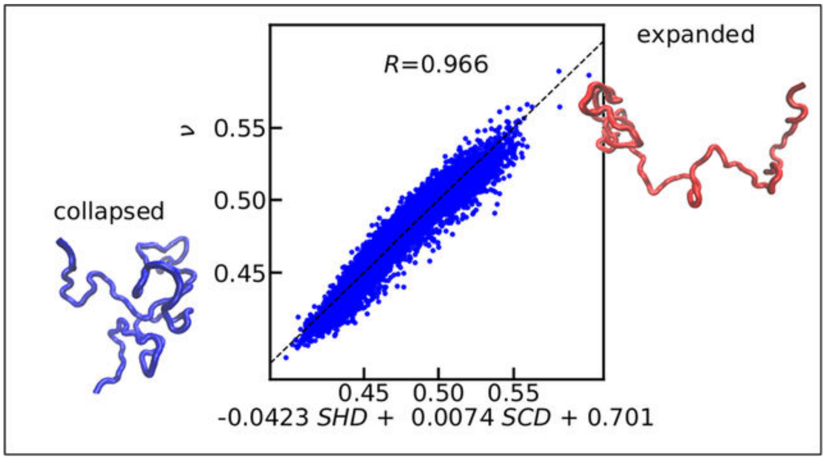
Abstract

Understanding the conformational ensemble of an intrinsically disordered protein (IDP) is of great interest due to its relevance to critical intracellular functions and diseases. It is now well established that the polymer scaling behavior can provide a great deal of information about the conformational properties as well as liquid-liquid phase separation of an IDP. It is, therefore, extremely desirable to be able to predict an IDP's scaling behavior from the protein sequence itself. The work in this direction so far has focused on highly charged proteins and how charge patterning can perturb their structural properties. As naturally occurring IDPs are composed of a significant fraction of uncharged amino acids, the rules based on charge content and patterning are only partially helpful in solving the problem. Here, we propose a new order parameter, sequence hydropathy decoration (SHD), which can provide a near quantitative understanding of scaling and structural properties of IDPs devoid of charged residues. We combine this with a charge patterning parameter, sequence charge decoration (SCD), to obtain a general equation, parameterized from extensive coarse-grained simulation data, for predicting protein dimensions from the sequence. We finally test this equation against available experimental data and find a semi-quantitative match in predicting the scaling behavior. We also provide guidance on how to extend this approach to experimental data, which should be feasible in the near future.

Graphical Abstract

wenweizheng@asu.edu; jeetain@lehigh.edu.

Supporting Information Available
Supporting methods, figures and tables.
This material is available free of charge via the Internet at http://pubs.acs.org/.



Intrinsically disordered proteins (IDPs) are of great interest in biology due to their involvement in important intracellular functions and pathological diseases.[1–4] These proteins lack a well-defined three-dimensional structure and are more appropriately described by a conformational ensemble in contrast to folded proteins with a single folded structure.[5,6] It is, therefore, nontrivial to study IDPs via the traditional structure-function relationship considering the heterogeneous nature of an IDP conformational ensemble. However, one still expects that the function of an IDP is determined by its sequence,[7,8] as observed in numerous cases.[9–13] It is important to identify sequence-dependent structural ensemble features capable of bridging the gap between sequence and function of an IDP, so that the structure-function paradigm can still be applied to IDPs.[14]

A variety of fundamental features ranging from average residue-level structural details to overall protein dimensions can be important for characterizing the conformational properties of an IDP. Nuclear magnetic resonance (NMR) experiments alone or coupled with all-atom simulations arguably provide the most detailed information on residual secondary structure properties and inter-residue interactions.[5,12,15–19] These data can help generate knowledge of how specific amino acids[20] and interactions between pairs of amino acid types may dictate the IDP properties.[19,21,22] Such empirical rules are significant for understanding the behavior of low complexity IDP sequences that are composed of only a few types of amino acids.[23] On the other hand, small-angle X-ray scattering (SAXS)[24] and Förster resonance energy transfer (FRET)[25] experiments provide estimates of global protein dimensions such as the radius of gyration (R_g). The interpretation of these experiments in terms of the polymer scaling behavior of proteins is helpful in applying existing analytical theories.[26–33]

Polymer scaling exponent (ν) is commonly used to characterize the relationship of the polymer size in solution with its chain length N as $R_g \propto N^{\nu}$. This variable also provides information on the solvent quality in terms of good, bad, or ideal solvent.[26,27] Despite the sequence heterogeneity of IDPs that contain twenty naturally occurring amino acids (and possibly many other non-canonical amino acids), there is increasing evidence that a single ν value may be used to characterize the conformational properties of disordered proteins. For instance, we recently showed that a ν-dependent distance distribution function based on a self-avoiding random walk model could help interpret experimental data from FRET[34] and

SAXS.[31] All-atom simulation data for more than 30 protein sequences further strengthened the understanding that on average IDPs display ideal chain behavior (ν=0.5) in aqueous solution, though the ν value is highly dependent on the protein sequence,[35] as expected. We also found that the θ-temperature (T_θ) of a single chain, the temperature at which ν=0.5, is strongly correlated with the critical temperature (T_c) of liquid-liquid phase separation (LLPS) of disordered proteins,[36,37] a result consistent with previous work from homopolymers.[28,38] This relationship provides a rapid method for approximating the behavior of IDPs in the context of LLPS, and aided in the development of a novel temperature-dependent interaction potential that explained upper- and lower critical solution temperature phase transitions based on temperature-dependent solvent-mediated interactions.[39]

Given the role of polymer scaling properties in dictating the conformational behavior of IDPs, there have been significant efforts to predict ν as a function of the protein sequence or order parameters representing important sequence characteristics. A protein's net charge and average hydropathy can help distinguish foldable sequences from disordered ones,[40] but it is essential also to consider other features such as the fraction of charged residues, and their patterning within the chain.[41–44] It is likely that the patterning of all amino acids, including uncharged ones, can contribute to the behavior of IDPs. Up to this point, however, this has not been studied in the context of twenty different amino acid types, even though it could be expected to be quite crucial, particularly for natural IDP sequences that contain a significant fraction of uncharged residues.[45–47]

In this work, we use our recently developed coarse-grained model of IDPs[48] to study the role of sequence patterning of uncharged residues in an extensive data set containing 5130 sequences. As expected, average hydropathy alone is not able to explain the sequence-dependent scaling behavior well, which leads us to develop a new sequence hydropathy decoration (SHD) parameter motivated by the extensively used sequence charge decoration (SCD) parameter.[43] The new SHD parameter reproduces the polymer scaling properties of these sequences, demonstrating the importance of patterning of twenty amino acids, beyond just charge patterning, in characterizing the size of the IDPs. We further find that the combination of SHD and SCD can capture the scaling properties of a more extensive data set (10260 sequences) containing all twenty amino acids remarkably well. Based on these results, we propose that a combination of SHD and SCD can be used to rapidly predict the scaling behavior of the disordered proteins and pave the way for high throughput screening of disordered sequences before wet lab investigation. We demonstrate this in the context of disordered protein sequences using the Disprot (disordered) database[23] and Top8000 (folded)[49] database as a control. We finally test predictions from our equation against existing experimental measurements of the size of several disordered proteins.

Computational estimation of polymer scaling exponent ν of IDPs.

The advantage of using ν as opposed to R_g to characterize a protein's size is to eliminate the chain length dependence and to provide meaningful information on the solvent quality that can be useful in predicting protein LLPS.[36] We have recently shown that R_g of a single protein can be used to estimate the scaling exponent $\left(\nu_{R_g}\right)$: [34,50,51]

$$R_g = \sqrt{\frac{\gamma(\gamma+1)}{2\left(\gamma+2\nu_{R_g}\right)\left(\gamma+2\nu_{R_g}+1\right)}}\, bN^{\nu_{R_g}}, \tag{1}$$

where $\gamma \approx= 1.1615$,[52] $b = 0.55$ nm,[29,34] and N is the number of peptide bonds (i.e., one less than the number of residues). Alternatively, when analyzing molecular simulation data, ν_{fit} can be obtained using the following equation, which is based on the mean intrachain distance $R_{i,j}$ as a function of sequence separation $|i - j|$,[41,53]

$$< R_{i,\,j}^2 >^{1/2} = b\left|i - j\right|^{\nu_{\text{fit}}}, \tag{2}$$

where $b = 0.55$ nm as in the Eq. 1. In practice, Eq. 1 provides a more convenient way of estimating ν from simulation data set, and more importantly, from experimentally determined R_g. However, its validity over the whole range of compactness of IDPs has not yet been established. Thus, it is important for us to test whether these two definitions give consistent results in predicting ν.

For this purpose, we generated a large set of 10,260 random protein sequences, having chain lengths in the range from 30 to 200 residues, and with amino acid probabilities set equal to their relative abundance in natural IDPs (see database A in Fig. S1).[54] We then conducted simulations of all of these sequences using our recently developed coarse-grained (CG) model, which represents each amino acid as a single interaction site (See Supporting Methods and original literature[48]). We find that the two methods of calculating ν are highly correlated, as shown in Fig. 1. Slight deviations are observed at low and high ν values, which suggests that the two methods will yield somewhat different scaling exponents. We asked if these deviations are related to an easily identifiable source in terms of protein's sequence properties, such as the chain length. As shown in Fig. S2, chain length does seem to cause some discernible differences in the ν estimates based on the two methods. Further analysis suggests that for low ν values, the ν_{fit} estimate may not be appropriate as the intrachain distance fits are not optimal over the whole range of sequence separation (see Fig. S3). For higher ν values, one may have to use a different prefactor b while using the intrachain distance fits to obtain ν_{fit}. The parameters used in Eq. 1 (i.e. γ and b) are almost optimal for minimizing the averaging deviations between ν_{fit} and ν_{R_g} for the whole range of ν values as shown in Fig. S4. For simplicity and keeping in mind that the relative errors across the whole database are mostly less than 5%, we suggest using Eq. 1 to reliably estimate a protein's scaling behavior in this and future studies.

Sequence hydropathy decoration (SHD) parameter describes properties of uncharged IDPs.

Significant previous work has already highlighted the role of sequence charge patterning on the properties of IDPs and important order parameters, such as sequence charge decoration (SCD) and κ, are available to describe such effects.[42–44,55,56] Given the success of such strategies, we focus on developing a descriptor for the patterning of uncharged residues. Here we start with an amino-acid specific hydropathy value (λ),[57] normalized to a value between 0 and 1 (Table S1), as the relevant feature to describe or predict a protein's compactness, and equivalently the polymer scaling properties. To isolate the effects of non-electrostatic interactions on chain dimensions, we generated an additional protein database of 5130 sequences in which the charged residues are not incorporated, and otherwise preserving the relative abundance of the uncharged amino acids (database B in Fig. S1). Interestingly, in the absence of charged residues, the average hydropathy ($< \lambda >$) of the sequence does not predict its ν very well (Fig. 2A) with only a weak trend showing that higher $< \lambda >$ values are generally more compact. Thus, additional considerations are needed in order to better capture the sequence-dependent properties originating from the patterning of uncharged residues.

Motivated by the success of the *SCD* parameter in capturing the charge patterning effects on protein behavior,[36,43,58,59] we propose a similar strategy to describe effects of hydropathy patterning, terming the new parameter sequence hydropathy decoration (*SHD*). To derive *SHD*, we adopt the theoretical approach presented in detail by Sawle and Ghosh[43] for charged polymers. The excluded volume contribution to the end-to-end distance of a polymer can be expressed as Ω (see Eq. 13 of Ref.[43])

$$\Omega = N^{-1}\Sigma_i\Sigma_j\omega_{i,\,j}\left|i - j\right|^{-1/2}, \tag{3}$$

where $\omega_{i,j}$ is the excluded volume interaction parameter between the two residues i and j. In our IDP model, such interactions arise from the short-range pairwise interactions with the interaction strength of ($\lambda_i + \lambda_j$) for a pair of residues i and j. The simplest assumption, namely, $\omega_{i,j} = \lambda_i + \lambda_j$, then leads to the following equation,

$$SHD = N^{-1}\Sigma_i\Sigma_{j,\,j\,>\,i}\left(\lambda_i + \lambda_j\right)\left|j - i\right|^{\beta} \tag{4}$$

in which $\beta = -1/2$ accounts for the contribution of sequence separation. This choice of the exponent shows much greater predictive ability than just $< \lambda >$ with a much higher negative Pearson correlation coefficient (−0.974) between ν and *SHD* values (see Fig. S5). In general, a high value of *SHD* corresponds to sequences with higher $< \lambda >$, and more clustering (patchiness) of hydrophobic residues together, resulting in a more collapsed conformation.

Next, we ask if the correlation can be improved by changing the β value to account for factors not considered in Eq. 4. We find that *SHD* provides the best description of our data for the β value ranging from −1.1 to −1 as shown in Fig. 2B and Fig. S5. This β value suggests that $\omega_{i,j}$ can also be sequence separation dependent. Since the two residues will be less likely in contact when the average distance between them becomes large, we can assume $\omega_{i,j}$ is inversely proportional to the average distance between the two residues. We can, therefore obtain

$$\omega_{i,j} \propto \frac{\lambda_i + \lambda_j}{r_{i,j}} \propto \frac{\lambda_i + \lambda_j}{|i-j|^{\nu}}, \tag{5}$$

where the last part of the equation makes use of the expected distance dependence from the polymer scaling law. This leads to $\beta = -\nu - 1/2$, where for IDPs with $0.45 \le \nu \le 0.6$, one gets $-1.1 \le \beta \le -0.95$, which is in excellent agreement with β obtained from the simulation data (see Fig. 2C and Fig. S5). Therefore in principle β can be ν dependent and vary slightly depending on the selection of the sequences. Considering the minimal differences among β values between −1.1 and −1 (Fig. 2B), we set $\beta = -1$ for simplicity. Of course, a different value with a slightly better correlation can be used if one chooses so. This *SHD* is in good correlation with ν from simulations ($R = -0.991$) as shown in Fig. 2C. This is a huge improvement over the average hydropathy ($R = -0.249$).

Considering a reasonable correlation ($R = -0.950$) obtained using *SHD* with $\beta = 0$ (Fig. S5), which is equivalent to rescaling the average hydropathy value by the chain length, it is clear that the average hydropathy does not work well due to not properly taking into account the chain length dependence. We have further tested two mean-field descriptions of *SHD* in which average hydropathy instead of residue specific hydropathy is used (Fig. S6). Both give similar correlation coefficients in comparison to the *SHD* with $\beta = -0.5$, suggesting for a well-mixed sequence mean-field approximation is reasonable. However, we want to note that such a mean-field description of *SHD* is not likely to be very useful for protein chains that are significantly more patterened. This can be easily seen in the data based on binary sequences with identical composition but different arrangement of amino acids (Fig. S7 and Table S2) We also find that this empirical approach to obtaining the sequence separation exponent (β in Eq. 4) recovers the known exponent value for *SCD* (0.5 as derived by Sawle and Ghosh;[43] Fig. S8). The observed dependence of hydropathy patterning on the sequence separation is weaker ($\beta = -1$) as compared to the charge patterning ($\beta = 0.5$), which could be expected considering their differences in interaction range. Thus we find that by developing the *SHD* parameter, we are able to make accurate predictions of IDP scaling behavior simply from the sequence, assuming the absence of charged amino acids. There are many other hydropathy scales available in the literature[60,61] that can potentially be used to compute *SHD* as well as to parameterize the CG model. The relative assessment of these different scales will be a topic of future study.

Predicting scaling behavior from sequence descriptors.

We then investigate how *SHD* compares with the other sequence descriptors (Table S3) to characterize ν, particularly in the case of sequences containing charged residues as well. We first look at the correlation between all sequence descriptors (independent of each other) and ν (Fig. S9A) and find that the most representative descriptors are *SHD*, $<\lambda>$, and *SCD*. The importance of $<\lambda>$ is consistent with previous work showing that it can be used to categorize disordered proteins.[40] However *SHD* and *SCD* stand out, which is probably due to the detailed nature of these two descriptors, accounting not only for the average value, but also patterning.

We expect that at least two sequence descriptors – one relevant to the amino acid charges and the other describing amino acid hydropathy – will be needed to describe the properties of IDPs. To test whether these metrics can work cooperatively to predict ν, we scan every pair of sequence descriptors using multilinear regression. In Fig. 3A, we show the Pearson correlation coefficients between the predicted ν from each pair of sequence descriptors and the simulated ν using the sequence database A (which contains charged amino acids; Fig. S1). While $< \lambda >$ scored higher than SCD in the single parameter regression (Fig. S7A), what it provides is redundant when used with SHD, so the combination of $< \lambda >$ and SHD does not significantly improve predictions over just SHD. The pairing of SHD and SCD results in the highest Pearson correlation coefficient (0.966, Fig. 3B) between the predicted and simulated ν. The multilinear equation for using SHD and SCD to predict ν is,

$$\nu = -0.0423_3 SHD + 0.0074_2 SCD + 0.701_2 \tag{6}$$

in which the subscripts show the errors of the last digit. The errors were estimated by randomly splitting the sequences into five groups for obtaining the standard deviation of the fitting parameters and repeating the random selection 100 times for averaging the errors. A linear regression using only SHD gives similar fitting parameters in comparison to Eq. 6 for the SHD prefactor and the constant term (Fig. S9B). By combining this multilinear equation with the equation between ν and R_g (Eq. 1), we can therefore predict the R_g directly from the sequence as shown in Fig. 3C.

Eq. 6 should also work for IDP sequences without any charged amino acids since SCD goes to zero. However, when the fraction of charged amino acids ($< |q| >$) increases, we expect that the contribution of SCD to the compactness of the chain should also become more important. This can be verified by performing the multilinear regression for subsets of our IDP sequence database with different values of $< |q| >$. As shown in Fig. S10, we find that the three fitting parameters do not change that much for $< |q| >$ values from 0.2 to 0.3. The SCD prefactor starts to increase when $< |q| >$ is greater than 0.3, as expected. We further assess relative effectiveness of different sequence descriptors and ν for different ranges of $< |q| >$ values (Fig. S11). We see that the charge patterning descriptors, SCD and κ, become increasingly important in determining the chain properties with increasing charge content ($< |q| >$). This is also consistent with previous literature that for sequences with all charged amino acids, charge patterning parameters are most important in characterizing the dimension of the chain.[42,43,62] However, because a large fraction of disordered proteins have $< |q| >$ smaller than 0.3, one needs for the role of hydropathy patterning, which we propose can be accomplished using SHD.

Experimental verification of the simplified equation based on *SHD* and *SCD*.

Since we find that SHD and SCD together can be used to predict ν from the simulation data set based on a simplified CG model of IDPs, we would like to test the model's transferability by using known disordered and folded protein sequences. We select the disordered protein sequences from the Disprot database,[23] excluding sequences having the disordered region shorter than 30 residues, as the polymer scaling law description may not work well for shorter chain lengths. For each sequence, only the longest disordered region is selected resulting in a total of 557 disordered protein sequences. We select folded protein sequences from a protein database Top8000,[49] in which the structure of every sequence has been solved with a high-resolution experimental method. We exclude the sequences in the database for which multiple chains are present, resulting in a total of 2360 folded protein sequences. We show in Fig. 4 that Eq. 6, using a combination of SHD and SCD obtains an average ν value close to 0.5 for the disordered proteins, consistent with previous knowledge in the field that disordered proteins in aqueous conditions on average behave similar to a Gaussian chain. It also suggests that there are half of the sequences with a scaling exponent of less than 0.5 and some are with values close to globular structures, similar to what a previous literature seen taking into account charge patterning of these sequences.[63] As a control, the ν values for the folded proteins predicted using the model are generally smaller.

It is likely that the use of a simple CG model to parameterize Eq. 6 would introduce errors based on the limitations of the model. Thus, we would like to further ask if one can directly use experimental data to parameterize Eq. 6 and how many sequences with experimentally determined R_g are necessary to obtain such an optimal predictive equation. We estimate the number of experimental sequences needed by splitting our computational database into two sets–a training set for fitting ν and a test set for checking the accuracy of the resulting model. For consistency, the number of sequences in the test set is fixed at a quarter of the total number of sequences (2565 of 10260 sequences) in database A (Fig. S1) while reducing the number of sequences in the training set. The process of randomly selecting the sequences to form the training and test sets is repeated 100 times to obtain the averaging errors of the model. We observe a typical L-shape plot for the relative errors as a function of the number of sequences (Fig. S12), which suggests that about 100 sequences will be sufficient to obtain an accurate predictive model. We expect that the actual number of sequences may differ as the estimate above is based on randomly generated protein sequences that may not capture the diversity of naturally occurring protein sequences, which are not completely random due to pressure from natural selection. Still, one expects the number of protein sequences necessary to obtain an experimentally validated predictive model to be within reach, especially if these sequences are carefully designed. Interestingly, the relative error is still quite reasonable, even for considerably smaller training sets (Fig. S12). Data-driven methods may prove to be quite helpful in this regard.

We then test currently available data on IDPs from the existing literature to validate our predictions. There has been an increasing number of experimental measurements on the compactness of disordered proteins, using either FRET or SAXS. However, there is clear difficulty of directly using the available data for parameterizing an empirical equation due to difficulties in interpreting experimental measurements. Recent work has shown that interpreting R_g or ν from FRET or SAXS experiments is not trivial due to the heterogeneous conformations disordered proteins can adopt.[64,65] FRET experiments tended to underestimate the R_g due to the assumptions about underlying distance distribution, whereas SAXS experiments overestimated the R_g due to a non-linearity in the Guinier plot.[31,64] We have identified a list of disordered sequences from a series of recent publications (see details in Table S4)[29,64,66–70] for assessing the computational model against the available experimental data. We reanalyzed the FRET data using our recently published method:

SAW-ν,[34] in which a ν-dependent distance distribution function is used to adapt the variation of chain dimension. The included SAXS data have been analyzed using recent approaches that employ a wide range of SAXS intensity instead of only Guinier region.[67–70] As shown in Fig. 5, the predicted ν for these proteins using Eq. 6, which was solely parameterized based on a CG IDP model with simple electrostatic interactions and no backbone potentials, are in reasonable agreement with the experimental ν values. It is even more remarkable if we consider the simplicity of our linear model combined with the lack of parameterization to account for different ionic strengths in these experiments. We believe this can be a first step towards the future refinement of the model based on experimental data by accounting for solution conditions appropriately.

Conclusion

Intrinsically disordered proteins perform a myriad of biological functions and are also involved in several debilitating disease conditions, but the sequence-structure-(mis)function relationships of these proteins are not well understood. The first step in developing such relationships is to understand better how the conformational preferences of disordered proteins originate from their sequence. Previous work has highlighted the role of charge content and patterning in developing sequence-structure relationships of highly charged proteins to capture the effects of electrostatic interactions. There has been relatively little progress in accounting for the role of other types of interactions such as van der Waals interactions and hydrogen bonding, through which uncharged amino acids interact. We propose that the amino acid hydropathy value can serve as a useful proxy to capture the average interactions of different amino acids, and how it affects the protein dimensions as part of a chain. To describe the presence and arrangement of amino acids with varying values of hydropathy, we propose a sequence hydropathy decoration parameter that can quantitatively capture the sequence-structure relationship for an extensive set of disordered proteins (lacking charged residues) simulated using a coarse-grained model. We combine this new parameter with the existing sequence charge decoration parameter to quantitatively predict protein dimensions simply based on the protein sequence. We anticipate that the predictive equation can serve as a quick screening tool to design new protein sequences with tunable properties as well as allow for future rapid optimization of coarse-grained models to better reproduce experimental results. Most importantly, we can already describe the scaling behavior of many proteins for which experimental data are available from single-molecule FRET and SAXS. This work should significantly contribute towards a quantitative understanding of a disordered proteins’ sequence-structure relationship, which we expect to apply to a better understanding of protein function as well.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This work was partially supported by the National Institutes of Health grants R01GM118530 and R01GM120537. Use of the high-performance computing capabilities of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation, project no. TG-MCB120014, is also gratefully acknowledged. W.Z. acknowledges the Research Computing at Arizona State University for providing HPC and storage that have contributed to the research results reported within this paper. Y.C.K is supported by the Office of Naval Research via the U.S. Naval Research Laboratory base program.

References

1. Tompa P Intrinsically unstructured proteins. Trends Biochem. Sci 2002, 27, 527–533. [PubMed: 12368089]
2. Uversky VN; Oldfield CJ; Midic U; Xie H; Xue B; Vucetic S; Iakoucheva LM; Obradovic Z; Dunker AK Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. BMC Genomics 2009, 10, S7.
3. Van Der Lee R; Buljan M; Lang B; Weatheritt RJ; Daughdrill GW; Dunker AK; Fuxreiter M; Gough J; Gsponer J; Jones DT et al. Classification of intrinsically disordered regions and proteins. Chem. Rev 2014, 114, 6589–6631. [PubMed: 24773235]
4. Wright PE; Dyson HJ Intrinsically disordered proteins in cellular signalling and regulation. Nat. Rev. Mol. Cell Biol 2015, 16, 18–29. [PubMed: 25531225]
5. Mittag T; Forman-Kay JD Atomic-level characterization of disordered protein ensembles. Curr. Opin. Struct. Biol 2007, 17, 3–14. [PubMed: 17250999]
6. Ozenne V; Bauer F; Salmon L; Huang J-r.; Jensen, M. R.; Segard, S.; Bernadó, P.; Charavay, C.; Blackledge, M. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics 2012, 28, 1463–1470. [PubMed: 22613562]
7. Forman-Kay JD; Mittag T From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. Structure 2013, 21, 1492–1499. [PubMed: 24010708]
8. Das RK; Ruff KM; Pappu RV Relating sequence encoded information to form and function of intrinsically disordered proteins. Curr. Opin. Struct. Biol 2015, 32, 102–112. [PubMed: 25863585]
9. Patel A; Lee HO; Jawerth L; Maharana S; Jahnel M; Hein MY; Stoynov S; Mahamid J; Saha S; Franzmann TM et al. A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. Cell 2015, 162, 1066–1077. [PubMed: 26317470]
10. Conicella AE; Zerze GH; Mittal J; Fawzi NL ALS mutations disrupt phase separation mediated by α-helical structure in the TDP-43 low-complexity C-terminal domain. Structure 2016, 24, 1537–1549. [PubMed: 27545621]
11. Wang J; Choi J-M; Holehouse AS; Lee HO; Zhang X; Jahnel M; Maharana S; Lemaitre R; Pozniakovsky A; Drechsel D et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. Cell 2018, 174, 688–699. [PubMed: 29961577]
12. Peng Y; Cao S; Kiselar J; Xiao X; Du Z; Hsieh A; Ko S; Chen Y; Agrawal P; Zheng W et al. A metastable contact and structural disorder in the estrogen receptor transactivation domain. Structure 2019, 27, 229–240. [PubMed: 30581045]
13. Schuster B; Dignon GL; Tang WS; Kelley F; Ranganath AK; Jahnke CN; Simpkins AG; Regy RM; Hammer DA; Good MC et al. Identifying Sequence Perturbations to an Intrinsically Disordered Protein that Determine Its Phase Separation Behavior. bioRxiv preprints 2020,
14. Wright PE; Dyson HJ Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol 1999, 293, 321–331. [PubMed: 10550212]
15. Demarest SJ; Martinez-Yamout M; Chung J; Chen H; Xu W; Dyson HJ; Evans RM; Wright PE Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. Nature 2002, 415, 549–553. [PubMed: 11823864]
16. Rogers JM; Wong CT; Clarke J Coupled folding and binding of the disordered protein PUMA does not require particular residual structure. J. Am. Chem. Soc 2014, 136, 5197–5200. [PubMed: 24654952]
17. Martin EW; Holehouse AS; Grace CR; Hughes AJ; Pappu RV Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation. J. Am. Chem. Soc 2016, 138, 15323–15335. [PubMed: 27807972]
18. Adamski W; Salvi N; Maurin D; Magnat J; Milles S; Jensen MR; Abyzov A; Moreau CJ; Blackledge M A Unified Description of Intrinsically Disordered Protein Dynamics nder

Physiological Conditions Using NMR Spectroscopy. J. Am. Chem. Soc 2019, 141, 17817–17829. [PubMed: 31591893]

19. Murthy AC; Dignon GL; Kan Y; Zerze GH; Parekh SH; Mittal J; Fawzi NL Molecular interactions underlying liquid- liquid phase separation of the FUS low-complexity domain. Nat. Struct. Mol. Biol 2019, 26, 637. [PubMed: 31270472]
20. Marsh JA; Forman-Kay JD Sequence determinants of compaction in intrinsically disordered proteins. Biophys. J 2010, 98, 2383–2390. [PubMed: 20483348]
21. Vernon RM; Chong PA; Tsang B; Kim TH; Bah A; Farber P; Lin H; Forman-Kay JD Pi-Pi contacts are an overlooked protein feature relevant to phase separation. Elife 2018, 7, e31486. [PubMed: 29424691]
22. Qamar S; Wang G; Randle SJ; Ruggeri FS; Varela JA; Lin JQ; Phillips EC; Miyashita A; Williams D; Ströhl F et al. FUS phase separation is modulated by a molecular chaperone and methylation of arginine cation-π interactions. Cell 2018, 173, 720–734. [PubMed: 29677515]
23. Sickmeier M; Hamilton JA; LeGall T; Vacic V; Cortese MS; Tantos A; Szabo B; Tompa P; Chen J; Uversky VN et al. DisProt: the database of disordered proteins. Nucleic Acids Res 2006, 35, D786–D793. [PubMed: 17145717]
24. Bernadó P; Svergun DI Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. Mol. BioSyst 2012, 8, 151–167. [PubMed: 21947276]
25. 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. [PubMed: 27145874]
26. Flory PJ The configuration of real polymer chains. J. Chem. Phys 1949, 17, 303–310.
27. de Gennes P-G Scaling Concepts in Polymer Physics; Cornell University Press, 1978.
28. Rubinstein M; Colby RH Polymer physics; Oxford university press New York, 2003; Vol. 23.
29. Hofmann H; Soranno A; Borgia A; Gast K; Nettels D; Schuler B Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. Proc. Natl. Acad. Sci. U.S.A 2012, 109, 16155–16160. [PubMed: 22984159]
30. Brangwynne CP; Tompa P; Pappu RV Polymer physics of intracellular phase transitions. Nat. Phys 2015, 11, 899.
31. Zheng W; Best RB An extended Guinier analysis for intrinsically disordered proteins. J. Mol. Biol 2018, 430, 2540–2553. [PubMed: 29571687]
32. Vancraenenbroeck R; Harel YS; Zheng W; Hofmann H Polymer effects modulate binding affinities in disordered proteins. Proc. Natl. Acad. Sci. U.S.A 2019, 116, 19506–19512. [PubMed: 31488718]
33. Baul U; Chakraborty D; Mugnai ML; Straub JE; Thirumalai D Sequence effects on size, shape, and structural heterogeneity in Intrinsically Disordered Proteins. J. Phys. Chem. B 2019, 123, 3462–3474. [PubMed: 30913885]
34. Zheng W; Zerze GH; Borgia A; Mittal J; Schuler B; Best RB Inferring properties of disordered chains from FRET transfer efficiencies. J. Chem. Phys 2018, 148, 123329. [PubMed: 29604882]
35. Zerze GH; Zheng W; Best RB; Mittal J Evolution of All-atom Protein Force Fields to Improve Local and Global Properties. J. Phys. Chem. Lett 2019, 10, 2227. [PubMed: 30990694]
36. Dignon GL; Zheng W; Best RB; Kim YC; Mittal J Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. Proc. Natl. Acad. Sci. U.S.A 2018, 115, 9929–9934. [PubMed: 30217894]
37. Panagiotopoulos AZ; Wong V; Floriano MA Phase equilibria of lattice polymers from histogram reweighting Monte Carlo simulations. Macromolecules 1998, 31, 912–918.
38. Wang R; Wang Z-G Theory of polymer chains in poor solvent: Single-chain structure, solution thermodynamics, and Θ point. Macromolecules 2014, 47, 4094–4102.
39. Dignon GL; Zheng W; Kim YC; Mittal J Temperature-Controlled Liquid–Liquid Phase Separation of Disordered Proteins. ACS Cent. Sci 2019, 5, 821. [PubMed: 31139718]
40. Uversky VN; Gillespie JR; Fink AL Why are “natively unfolded” proteins unstructured under physiologic conditions? Proteins 2000, 41, 415–427. [PubMed: 11025552]

41. Mao AH; Crick SL; Vitalis A; Chicoine C; Pappu RV Net charge per residue modulates conformational ensembles of intriniscally disordered proteins. Proc. Natl. Acad. Sci. U.S.A 2010, 107, 8183–8188. [PubMed: 20404210]

42. Das RK; Pappu RV Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc. Natl. Acad. Sci. U.S.A 2013, 110, 13392–13397. [PubMed: 23901099]

43. Sawle L; Ghosh K A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. J. Chem. Phys 2015, 143, 085101. [PubMed: 26328871]

44. Samanta HS; Chakraborty D; Thirumalai D Charge fluctuation effects on the shape of flexible polyampholytes with applications to intrinsically disordered proteins. J. Chem. Phys 2018, 149, 163323. [PubMed: 30384718]

45. Khokhlov AR; Khalatur PG Conformation-dependent sequence design (engineering) of AB copolymers. Phys. Rev. Lett 1999, 82, 3456.

46. Ashbaugh HS Tuning the globular assembly of hydrophobic/hydrophilic heteropolymer sequences. J. Phys. Chem. B 2009, 113, 14043–14046. [PubMed: 19799382]

47. Statt A; Casademunt H; Brangwynne CP; Panagiotopoulos AZ Model for disordered proteins with strongly sequence-dependent liquid phase behavior. J. Chem. Phys 2020, 152, 075101. [PubMed: 32087632]

48. Dignon GL; Zheng W; Kim YC; Best RB; Mittal J Sequence determinants of protein phase behavior from a coarse-grained model. PLoS Comput. Biol 2018, 14, e1005941. [PubMed: 29364893]

49. Chen VB; Arendall WB; Headd JJ; Keedy DA; Immormino RM; Kapral GJ; Murray LW; Richardson JS; Richardson DC MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D 2010, 66, 12–21. [PubMed: 20057044]

50. Fisher ME Shape of a Self-Avoiding Walk or Polymer Chain. J. Chem. Phys 1966, 44, 616–622.

51. Des Cloizeaux J Lagrangian theory for a self-avoiding random chain. Phys. Rev. A 1974, 10, 1665.

52. Le Guillou J; Zinn-Justin J Critical exponents for the n-vector model in three dimensions from field theory. Phys. Rev. Lett 1977, 39, 95.

53. Zheng W; Borgia A; Buholzer K; Grishaev A; Schuler B; Best RB Probing the action of chemical denaturant on an intrinsically disordered protein by simulation and experiment. J. Am. Chem. Soc 2016, 138, 11702–11713. [PubMed: 27583687]

54. Coeytaux K; Poupon A Prediction of unfolded segments in a protein sequence based on amino acid composition. Bioinformatics 2005, 21, 1891–1900. [PubMed: 15657106]

55. Thirumalai D; Samanta HS; Maity H; Reddy G Universal nature of collapsibility in the context of protein folding and evolution. Trends in Biochem. Sci 2019, 44, 675. [PubMed: 31153683]

56. Lytle TK; Chang L-W; Markiewicz N; Perry SL; Sing CE Designing Electrostatic Interactions via Polyelectrolyte Monomer Sequence. ACS Cent. Sci 2019, 5, 709–718. [PubMed: 31041391]

57. Kapcha LH; Rossky PJ A simple atomic-level hydrophobicity scale reveals protein interfacial structure. J. Mol. Biol 2014, 426, 484–498. [PubMed: 24120937]

58. Lin Y-H; Chan HS Phase Separation and Single-Chain Compactness of Charged Disordered Proteins Are Strongly Correlated. Biophys. J 2017, 112, 2043–2046. [PubMed: 28483149]

59. Das S; Amin AN; Lin Y-H; Chan HS Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters. Phys. Chem. Chem. Phys 2018, 20, 28558–28574. [PubMed: 30397688]

60. Kyte J; Doolittle RF A simple method for displaying the hydropathic character of a protein. J. Mol. Biol 1982, 157, 105–132. [PubMed: 7108955]

61. Huang F; Oldfield CJ; Xue B; Hsu W-L; Meng J; Liu X; Shen L; Romero P; Uversky VN; Dunker AK Improving protein order-disorder classification using charge-hydropathy plots. BMC Bioinformatics 2014, 15, S4.

62. Lin Y-H; Forman-Kay JD; Chan HS Sequence-specific polyampholyte phase separation in membraneless organelles. Phys. Rev. Lett 2016, 117, 178101. [PubMed: 27824447]

63. Firman T; Ghosh K Sequence charge decoration dictates coil-globule transition in intrinsically disordered proteins. J. Chem. Phys 2018, 148, 123305. [PubMed: 29604827]

64. Borgia A; Zheng W; Buholzer K; Borgia MB; Schuler A; Hofmann H; Soranno A; Nettels D; Gast K; Grishaev A et al. Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. J. Am. Chem. Soc 2016, 138, 11714–11726. [PubMed: 27583570]

65. Song J; Gomes G-N; Shi T; Gradinaru CC; Chan HS Conformational Heterogeneity and FRET Data Interpretation for Dimensions of Unfolded Proteins. Biophys. J 2017, 113, 1012–1024. [PubMed: 28877485]

66. Fuertes G; Banterle N; Ruff KM; Chowdhury A; Mercadante D; Koehler C; Kachala M; Girona GE; Milles S; Mishra A et al. Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. Proc. Natl. Acad. Sci. U.S.A 2017, 114, E6342–E6351. [PubMed: 28716919]

67. Riback JA; Bowman MA; Zmyslowski AM; Knoverek CR; Jumper JM; Hinshaw JR; Kaye EB; Freed KF; Clark PL; Sosnick TR Innovative scattering analysis shows that hydrophobic proteins are expanded in water. Science 2017, 358, 238–241. [PubMed: 29026044]

68. Martin EW; Holehouse AS; Peran I; Farag M; Incicco JJ; Bremer A; Grace CR; Soranno A; Pappu RV; Mittag T Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science 2020, 367, 694–699. [PubMed: 32029630]

69. Rieloff E; Skepö M Phosphorylation of a disordered peptide–structural effects and force field inconsistencies. J. Chem. Theory Comput 2020, 16, 1924–1935. [PubMed: 32050065]

70. Cragnell C; Durand D; Cabane B; Skepö, M. Coarse-grained modeling of the intrinsicallydisordered protein Histatin 5 in solution:Monte Carlo simulations in combinationwith SAXS. Proteins 2016, 84, 777–791. [PubMed: 26914439]

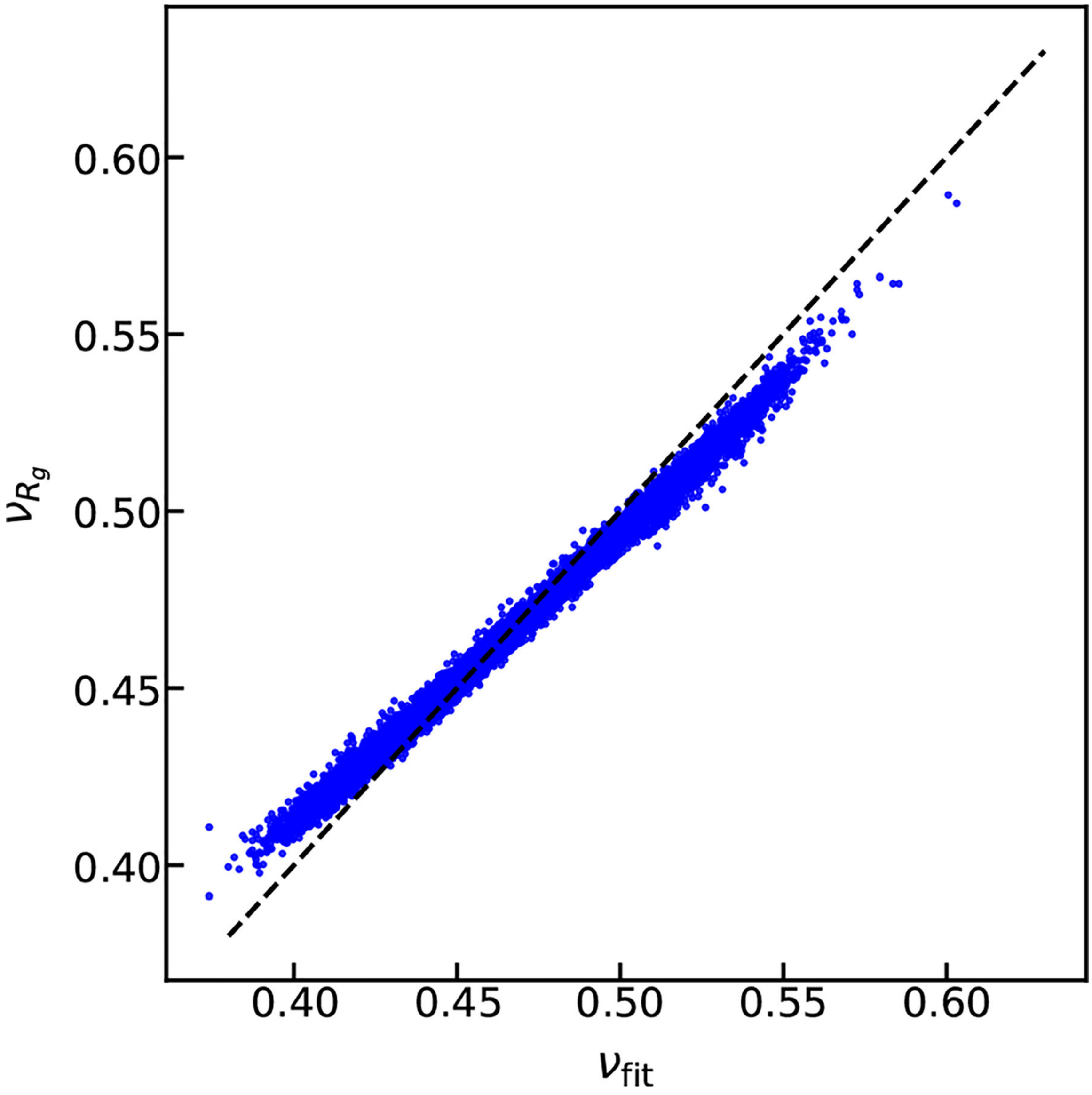


Figure 1:

Comparison between the polymer scaling exponents obtained by fitting intramolecular distances (ν_{fit}) or by using Eq. 1$\left(\nu_{R_g}\right)$.

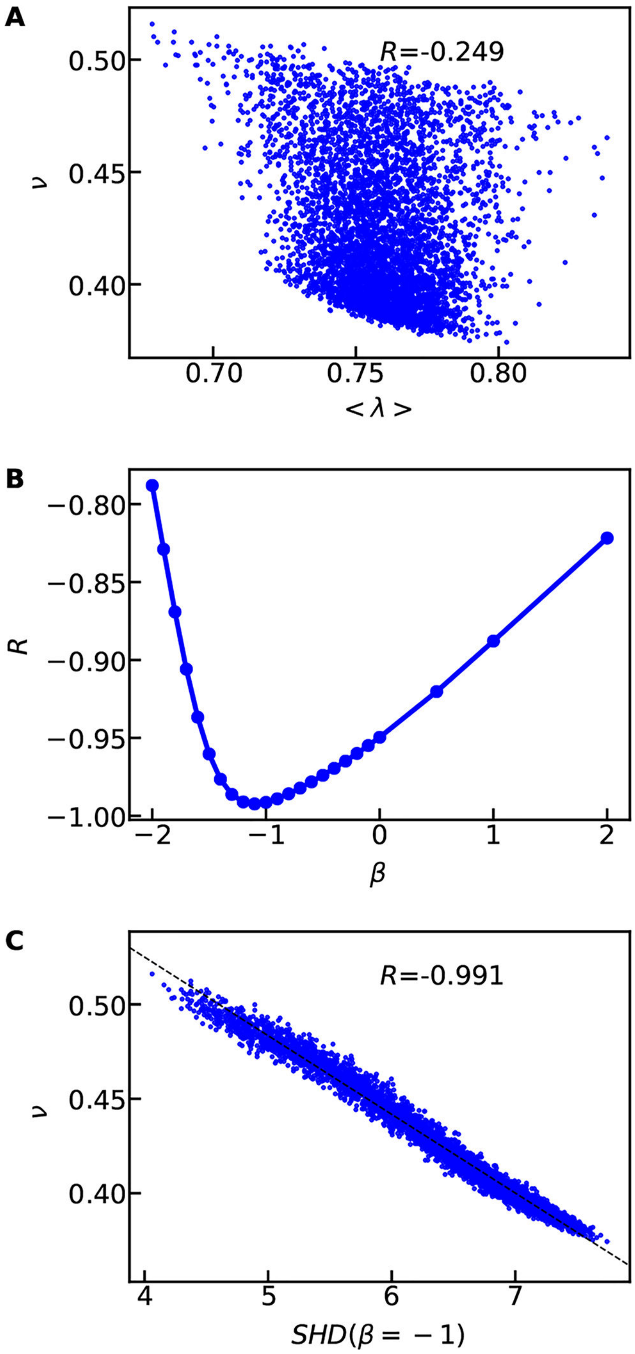


Figure 2:

A) Using the mean hydropathy ($<\lambda>$) to capture the scaling exponents of random uncharged sequences. B) Pearson correlation coefficient between *SHD* and ν when varying β in Eq. 4. C) Using the hydropathy patterning parameter *SHD* with $\beta = -1$ to capture the scaling exponents. The dashed line show the linear fitting between *SHD* and ν and the legends show the Pearson correlation coefficients.

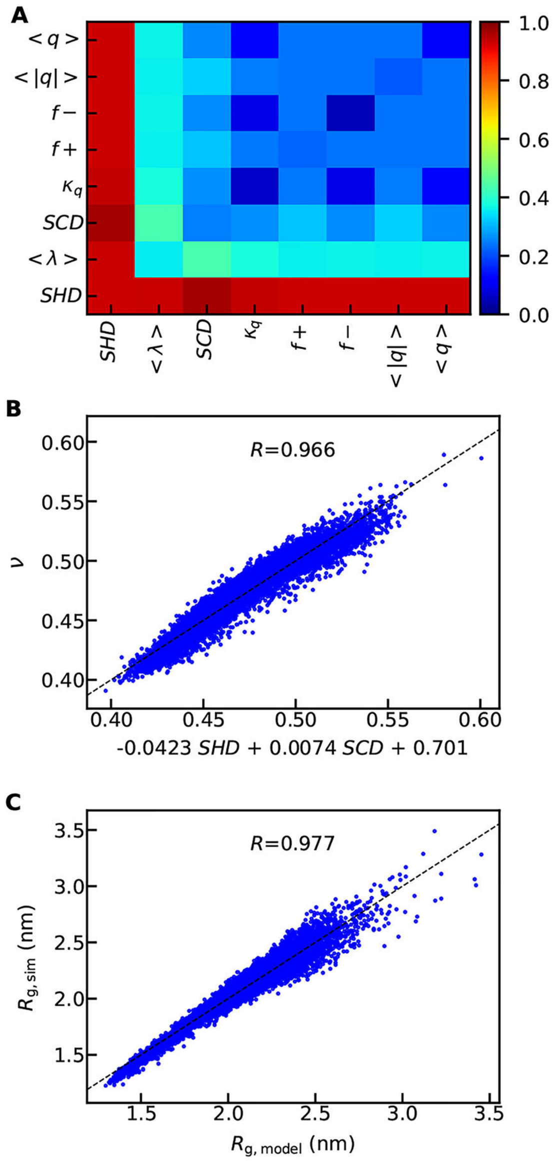


Figure 3:

Capturing the scaling exponents (ν) using linear models of two sequence descriptors. A) Pearson correlation coefficients between the linearly modelled and simulated ν. B) The comparison between the simulated ν and the predicted ν based on the best pair of sequence descriptors with the linear equation shown in labels of x-axis. C) The comparison between the simulated and the predicted R_g using the best pair of sequence descriptors.

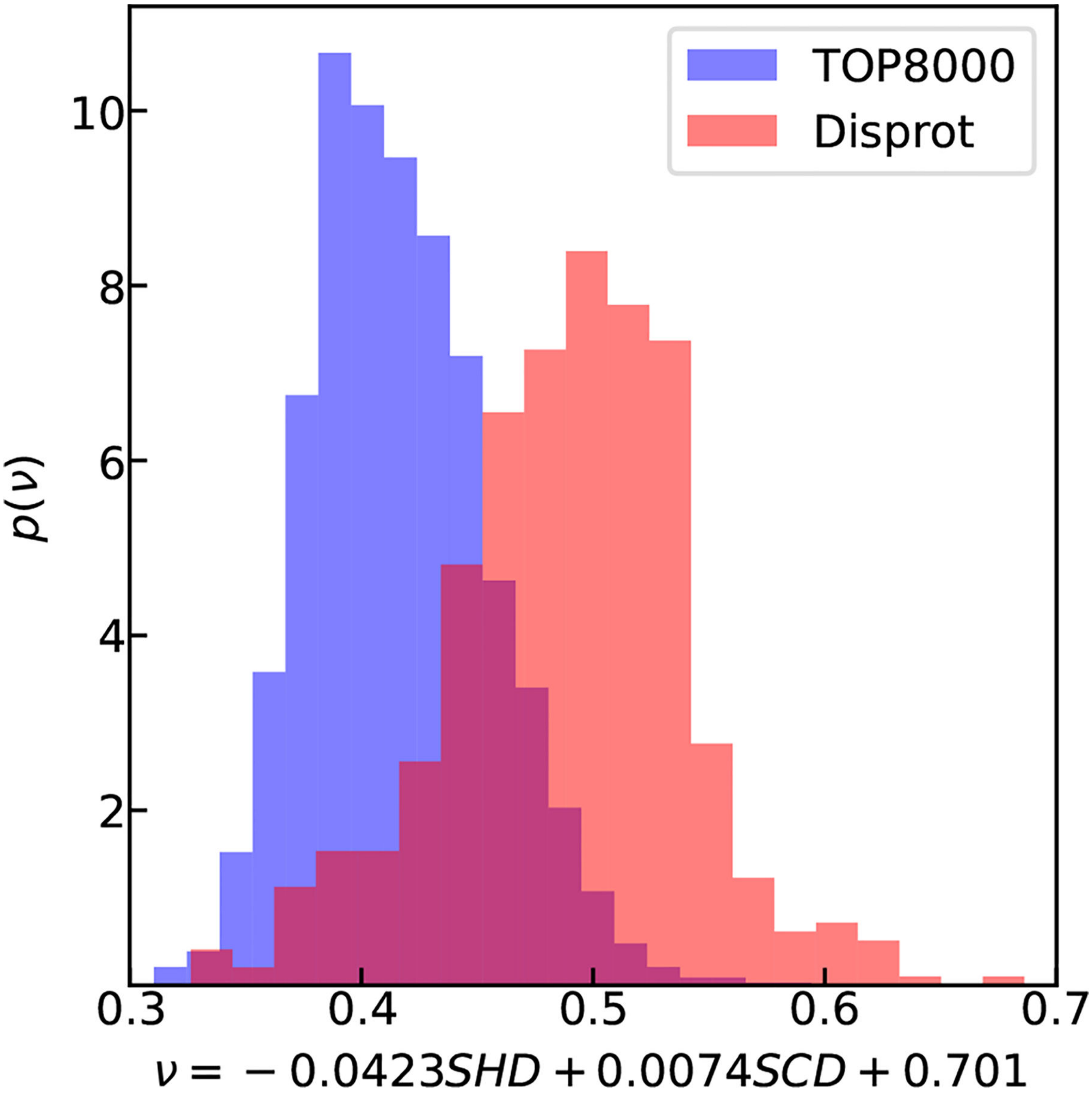


Figure 4:

Using *SHD* and *SCD* to predict ν of disordered sequences from Disprot database (red).[23] The folded sequences using TOP8000 database[49] are shown in blue as a control.

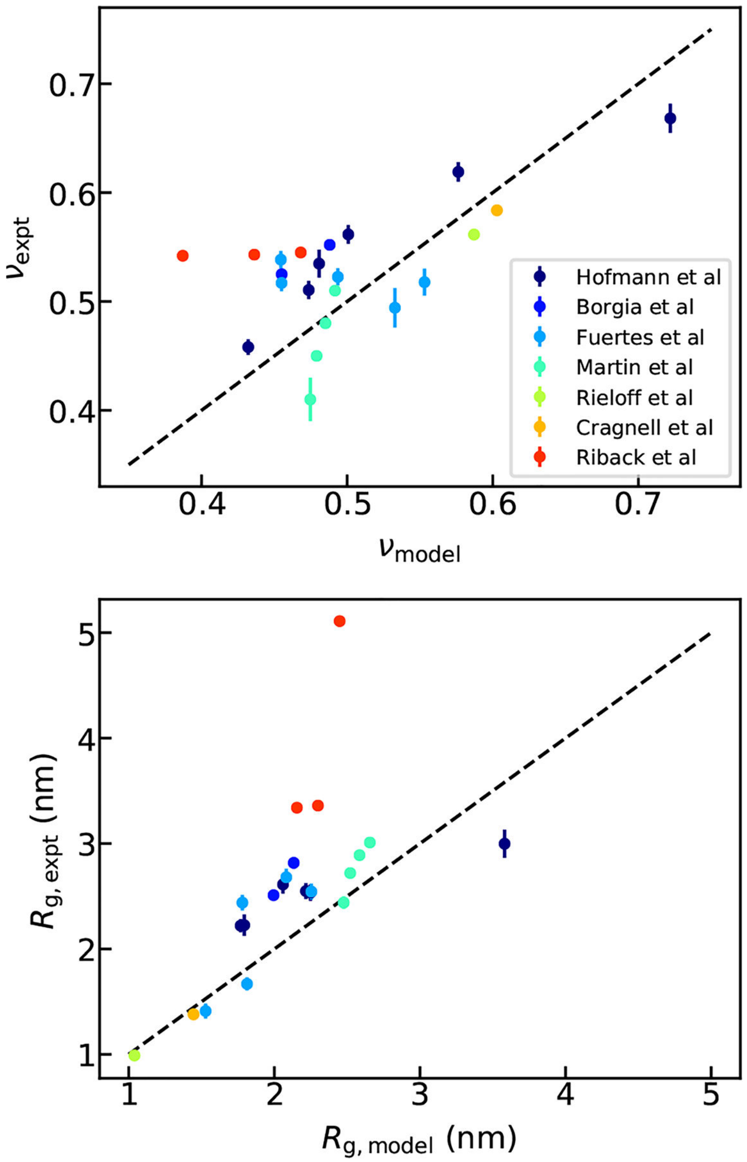


Figure 5:

Comparison between the R_g (A) and ν (B) from linear model using *SHD* and *SCD* and from FRET and SAXS experiments.