The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

Fine structures of intrinsically disordered proteins

Cite as: J. Chem. Phys. 160, 014902 (2024); doi: 10.1063/5.0176306
Submitted: 12 September 2023 • Accepted: 5 December 2023 •
Published Online: 2 January 2024

Swarnadeep Seth, Brandon Stine, and Aniket Bhattacharya[a]

AFFILIATIONS

Department of Physics, University of Central Florida, Orlando, Florida 32816-2385, USA

[a]Author to whom correspondence should be addressed: Aniket.Bhattacharya@ucf.edu

ABSTRACT

We report simulation studies of 33 single intrinsically disordered proteins (IDPs) using coarse-grained bead-spring models where interactions among different amino acids are introduced through a hydropathy matrix and additional screened Coulomb interaction for the charged amino acid beads. Our simulation studies of two different hydropathy scales (HPS1, HPS2) [Dignon *et al.*, PLoS Comput. Biol. **14**, e1005941 (2018); Tesei *et al.* Proc. Natl. Acad. Sci. U. S. A. **118**, e2111696118 (2021)] and the comparison with the existing experimental data indicate an optimal interaction parameter $\epsilon = 0.1$ and 0.2 kcal/mol for the HPS1 and HPS2 hydropathy scales. We use these best-fit parameters to investigate both the universal aspects as well as the fine structures of the individual IDPs by introducing additional characteristics. (i) First, we investigate the polymer-specific scaling relations of the IDPs in comparison to the universal scaling relations [Bair *et al.*, J. Chem. Phys. **158**, 204902 (2023)] for the homopolymers. By studying the scaled end-to-end distances $\langle R_N^2\rangle/(2L\ell_p)$ and the scaled transverse fluctuations $\tilde{l}_\perp^2 = \sqrt{\langle l_\perp^2\rangle}/L$, we demonstrate that IDPs are broadly characterized with a Flory exponent of $\nu \simeq 0.56$ with the conclusion that conformations of the IDPs interpolate between Gaussian and self-avoiding random walk chains. Then, we introduce (ii) Wilson charge index ($\mathcal{W}$) that captures the essential features of charge interactions and distribution in the sequence space and (iii) a skewness index ($\mathcal{S}$) that captures the finer shape variation of the gyration radii distributions as a function of the net charge per residue and charge asymmetry parameter. Finally, our study of the (iv) variation of $\langle R_g\rangle$ as a function of salt concentration provides another important metric to bring out finer characteristics of the IDPs, which may carry relevant information for the origin of life.

I. INTRODUCTION

Intrinsically disordered proteins (IDPs) are characterized by a low proportion of hydrophobic residues and a high content of polar and charged amino acids, which make them distinct from those that fold.[1–7] IDPs lack well-defined three-dimensional structures and do not participate in forming α-helices or β-strands and other secondary or tertiary structures. Since their discovery almost three decades ago, the number of IDPs has been growing at a steady rate.[8] It is now known that almost 30% of the proteins are either IDPs or folded proteins that have intrinsically disordered regions (IDR),[9] which play crucial roles in numerous biological processes, such as regulating signaling pathways, helping in molecular recognition, initiating protein–protein interactions, and serve as molecular switches.[10,11] The conformal flexibility of IDPs helps mediate interactions with binding partners to form components of macromolecular complexes.[12,13] The flexibility and often faster dynamics allow IDPs to bind to multiple different proteins.[11,14] The IDP complexes[12] have also been realized to play a central to the pathology of several degenerative diseases: α-synuclein (Parkinson's disease), tau (Alzheimer's disease), and IAPP (Type II Diabetes).[15,16] IDPs form membrane-less intra-cellular compartments, have been identified to be the key drivers of liquid–liquid phase separation in the cell,[17–19] and, therefore, have generated tremendous interest in the underlying phase separation in IDPs, formation of IDP complexes, and the role of charge separation in such systems using well-established concepts of polymer physics.[12,20–22]

Evidently, the study of IDPs in the past two decades has been an active area in various branches of science. Despite tremendous growth and interest in studying IDPs, the discovery of new IDPs and their fast dynamics made it difficult to study experimentally using small-angle x-ray scattering (SAXS),[23] single molecule Förster resonance energy transfer (smFRET),[22,24–26] and solution nuclear magnetic resonance (sNMR),[27] which have produced conflicting results. The conformational information, such as end-to-end distances and gyration radii, is also available for a limited number of IDPs. Thus, an integrative structural biology approach that combines experimental techniques[28] using nuclear magnetic resonance

(NMR) spectroscopy SAXS, smFRET, and combined with computational methods seems to be a practical and feasible approach to unravel the conformational properties and interactions of IDPs, shedding light on their structural ensembles.

Historically, computer simulation studies of coarse-grained (CG) models of polymers have played an important role as a stand-alone discipline between theory and experiments successfully predicting conformational and dynamic properties of neutral and charged polymers.[29] Similar studies have been generalized for the IDPs taking into account different sizes, charges, and hydropathy indices of the 20 different CG amino acid beads.[30–35] Data-driven approaches aided by machine learning algorithms have helped to build hydropathy indices for numerical studies of IDPs.[32,36] In this article, we enlarge the scope of validity of a subset of these models (HPS and M3) by comparing simulation results obtained for a large number of IDPs using these two models with experimental data[6,28,37–40] and with the results contained using other CG models.[33,34]

II. COARSE-GRAINED MODELS OF IDPs

One of the hallmarks of IDPs is their characterization using the Uversky plot,[1] where it has been shown that when the mean net absolute charge $\langle Q \rangle$ of a polypeptide chain at neutral pH is plotted against the mean side chain hydropathy $\langle H \rangle$, measured using the Kyte–Doolittle[41] hydrophobicity scale, a boundary line

$$\langle Q \rangle = 2.785 \langle H \rangle - 1.151 \tag{1}$$

separates the compact (natively folded or globular) and expanded (coil-like or pre-molten globular) conformations.[1,42] Habchi *et al.*[42] improved Eq. (1) but the basic observation remains the same. Evidently, a larger charge and a smaller hydropathy ensure the extended structure of the IDPs. The simplicity has been appealing to build CG models of IDPs based on hydropathy, where the standard bead-spring model of a homopolymer has been generalized to incorporate the relative well depth between any two amino acids through a hydropathy matrix.[30] Mittal and coworkers[31,43] have used this HPS model to compare the gyration radii for several IDPs and found a reasonably good agreement. A slightly different version has been used by Tesei *et al.*[32] Unlike the HPS models[30–32] where hydropathy is introduced directly, other implicit solvent CG models have been used to study various properties of IDPs. Pappu and coworkers developed a software called ABSINTH (Assembly of Biomolecules Studied by an Implicit, Novel, and Tunable Hamiltonian) to study phase transitions in IDPs.[33] Thirumalai and coworkers used another CG model called SOP-IDP [self-organized polymer (SOP) coarse-grained model for IDPs] with a finer level of granularity where, except for glycine and alanine, the rest of the amino acid residues are represented using a backbone bead and a side-chain (SC) bead.[34] All these models are computationally more efficient compared to the models with explicit solvent molecules and, hence, can be used to study macromolecular condensates of IDPs leading to liquid–liquid phase separation[17–19,44–48] in membraneless organelles.

Since their discovery, progressively more IDPs are being cataloged.[8] Compared to the models for the folded proteins, the CG models of the IDPs are relatively new. Due to their flexibility and faster dynamics, the experimental studies of the IDPs are relatively limited and often very difficult to interpret. Thus, studies of several CG models with convergence to the experimental results are important aspects of developing a better understanding of the fundamental physics of the IDPs that share properties of the polyelectrolytes and polyampholytes but exhibit very different sequence-specific behaviors. In this article, we critically examine the parameters of two hydropathy models of the IDPs by enlarging the scope of previous work.[31,32] The gyration radii obtained from the simulation then have been reanalyzed with reference to the universal properties of the homopolymer model. Another important aspect of this work is that we have introduced new physically motivated metrics to analyze these fine structures of the IDPs. Finally, we provide specific examples in Sec. V, where IDPs clustered around different regions of the charge–hydropathy space may exhibit markedly different characteristics.

III. SIMULATION MODEL

The CG amino acid residues interact among themselves by a modified van der Waals interaction potential, first introduced by Ashbaugh and Hatch,[30] given by

$$U_{VdW}(r_{ij}) = \begin{cases} U_{LJ}(r_{ij}) + (1-\lambda_{ij})\epsilon_{ij}, & r_j \le 2^{\frac{1}{6}}\sigma_{ij}, \\ \lambda_{ij} U_{LJ}(r_{ij}), & \text{otherwise}, \end{cases} \tag{2}$$

where U_{LJ} is the Lennard-Jones (LJ) potential,

$$U_{LJ}(r_{ij}) = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], \tag{3}$$

where $r_{ij} = |\vec{r}_i - \vec{r}_j|$ and $\sigma_{ij} = \frac{1}{2}(\sigma_i + \sigma_j)$ are the distance and the effective diameter between the amino acid beads of diameter σ_i and σ_j with indices i and j positioned at $\vec{r}_i$ and $\vec{r}_j$, respectively. The strength of the van der Waal interaction $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j} \cong \epsilon$ is chosen to be the same, and $\lambda_{ij} = \frac{1}{2}(\lambda_i + \lambda_j)$ is the average hydropathy factor between any two amino acids with indices i and j. A harmonic bond potential

$$U_b(r_{ij}) = \frac{k_b}{2}\left(\frac{r_{ij} - r_{ij}^0}{\sigma_{ij}}\right)^2 \tag{4}$$

acts between two consecutive amino acid residues i and $j = i \pm 1$. The spring constant $k_b = 8033$ kJ/(mol nm^2) = 1920 kcal/(mol nm^2), and the equilibrium bond length is $r_{ij}^0 = r_0 = 0.38$ nm.

A screened-Coulomb (SC) interaction acts between any two charged amino acids,

$$U_{SC}(r_{\alpha\beta}) = \frac{q_\alpha q_\beta e^2}{4\pi\epsilon_0\epsilon_r}\left(\frac{e^{-\kappa r_{\alpha\beta}}}{r_{\alpha\beta}}\right), \tag{5}$$

where the indices α and β refer to the subset of the indices i and j for the charged amino acids, ϵ_r is the dielectric constant of water, and κ is the Debye screening length.[49] The inverse Debye length κ^{-1} is dependent on the ionic concentration (I) and expressed as

$$\kappa^{-1} = \sqrt{8\pi l_B I N_A \times 10^{-24}}, \tag{6}$$

where N_A is Avogadro's number and l_B is the Bjerrum length,

$$l_B = \frac{e^2}{4\pi\epsilon_0\epsilon_r k_B T}. \quad (7)$$

At higher temperatures, the dielectric constant typically decreases, which affects the strength of the electrostatic interactions. If the dielectric constant does not account for temperature effects, the electrostatic interactions may be overestimated, leading to unrealistic protein conformations or interactions. Hence, we implement the temperature-dependent dielectric constant of water as expressed by the empirical relation,[50]

$$\epsilon_r(T) = \frac{5321}{T} + 233.76 - 0.9297T + 1.147 \times 10^{-3}T^2 - 8.292 \times 10^{-7}T^3. \quad (8)$$

IV. HYDROPATHY SCALES

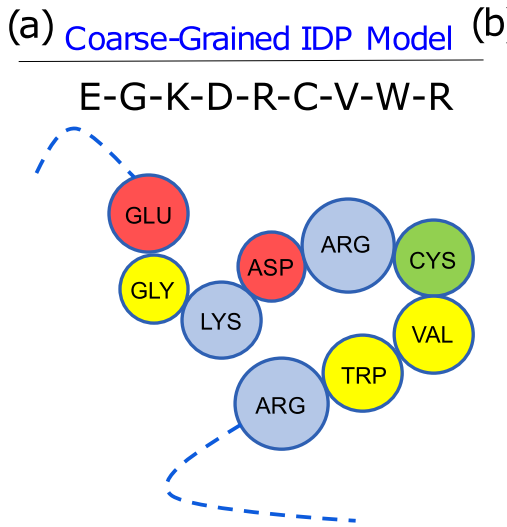
Historically, many hydropathy scales have been introduced to model the properties of amino acids and provide a quantitative measure of the hydrophilicity or hydrophobicity of amino acids based on their propensity to reside in a water-soluble or water-insoluble environment.[41,51–54] Each of these scales assigns a numerical value to each amino acid, reflecting its hydrophobic or hydrophilic nature. The scores obtained from hydropathy scales are useful in predicting protein structure and function. Recently, specific hydropathy scales are employed to study the formation of condensates and liquid–liquid phase separation (LLPS) behavior in IDPs. Dignon *et al.* proposed the HPS[31] hydropathy scale where proline and phenylalanine are considered to be the most hydrophobic with $\lambda^{HPS} = 1$ and arginine is the least hydrophobic with $\lambda^{HPS} = 0$. All the amino acids' hydropathy are scaled to fit in the range. Later, Tesei *et al.* used the Bayesian parameter-learning procedure to further optimize the hydropathy values and showed M3[32] hydropathy scale performs better to produce radius of gyration values closer to the experiments.

A. HPS1 scale

Dignon *et al.* proposed HPS[31] hydropathy values that are listed in the sixth column in Fig. 1(b). The original HPS model uses the constant dielectric constant of water $\epsilon_r = 80$ for simulating the IDPs. However, in some cases, such as, for CoINT (row 13 in Table I), experiments are performed at a lower temperature. To account for these temperature changes, we used the analytic expression of the temperature-dependent dielectric constant [see Eq. (8)], which makes our model slightly different from the original HPS model. However, the amino acid hydropathy scale remains identical. We denote it as the HPS1 scale in our simulation.

B. HPS2 scale

Tesei *et al.* used the Bayesian parameter-learning procedure to further optimize the hydropathy values and showed M3[32] hydropathy scale performs better to produce radius of gyration values closer to the experiments. In the original M3 model, the charges of the end amino acids of an IDP are modified and the charge of the Histidine residue is tuned as a function of the pH. However, in our simulation, we do not alter the end charges and the charge of the Histidine residue to make a consistent comparison of the results with the other existing models. We denote this hydropathy scale of the amino acids as HPS2 and are shown in the seventh column in Fig. 1(b).



(b) Amino Acids and Hydropathy Scales

One Lett.	Amino Acid	Mass (amu)	Charge	σ (nm)	λ^{HPS1}	λ^{HPS2}
A	ALA	71.08	0	0.504	0.73	0.003
R	ARG	156.2	1	0.656	0	0.723
N	ASN	114.1	0	0.568	0.432	0.16
D	ASP	115.1	-1	0.558	0.378	0.002
C	CYS	103.1	0	0.548	0.595	0.4
Q	GLN	128.1	0	0.602	0.514	0.468
E	GLU	129.1	-1	0.592	0.459	0.022
G	GLY	57.05	0	0.45	0.649	0.784
H	HIS	137.1	0.5	0.608	0.514	0.487
I	ILE	113.2	0	0.618	0.973	0.687
L	LEU	113.2	0	0.618	0.973	0.335
K	LYS	128.2	1	0.636	0.514	0.095
M	MET	131.2	0	0.618	0.838	0.993
F	PHE	147.2	0	0.636	1	0.871
P	PRO	97.12	0	0.556	1	0.471
S	SER	87.08	0	0.518	0.595	0.487
T	THR	101.1	0	0.562	0.676	0.274
W	TRP	186.2	0	0.678	0.946	0.753
Y	TYR	163.2	0	0.646	0.865	0.984
V	VAL	99.07	0	0.586	0.892	0.428
	Average	118.885	0.025	0.59	0.677	0.471

FIG. 1. (a) Coarse-grained model of a portion of IDP containing nine amino acid residues with one letter code EGKDRCVWR. Each amino acid residue is represented by the corresponding three letter code beads of different diameters σ (not to the scale) listed on the table. (b) The table lists the one-letter and three-letter codes of the amino acids in the first and second columns. The mass, charge, and diameter of the amino acid residues are tabulated in the third, fourth, and fifth columns, respectively. Finally, the HPS1 and HPS2 hydropathy scales are shown in the sixth and seventh columns, respectively.

TABLE I. Comparisons of gyration values from experimental (column 11) and simulation studies obtained using HPS1 (column 12) and HPS2 (column 14) models. The symbols are explained in the main text. Experimental values listed in column 11 correspond to the IDP references in the second column obtained from Baul *et al.*[34] (red-colored circle), Dignon *et al.*[31] (blue-colored box), Tesei *et al.*[32] (green-colored diamond), and Dannenhoffer-Lafage and Best[36] (purple-colored triangle). The 12th and 14th columns represent the simulation radius of gyration results obtained by the HPS1 and the HPS2 hydropathy scale for ϵ = 0.1 and 0.2 kcal/mol, respectively, producing least MSE [see Eq. (9), and Fig. 2] with the corresponding percentage errors in the 13th and 15th columns.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Row	ID	IDP Name	N	I	T(K)	Q	N_q	q_{abs}	q_{net}	f^*	R_g^{expt}	R_g^{HPS1}	ΔR_g^{HPS1}	R_g^{HPS2}	ΔR_g^{HPS2}	$\mathcal{W}$
1	α	K32 (○)	198	150	293	22.5	56	0.265	0.114	0.049	4.15	4.595	11%	4.611	11%	23.49
2	β	K16 (○)	176	150	293	20.5	47	0.253	0.116	0.054	3.90	4.278	10%	4.225	8%	21.78
3	γ	K27 (○)	167	150	293	19	48	0.269	0.114	0.048	3.70	4.096	11%	4.171	13%	20.13
4	δ	K17 (○)	145	150	293	17	39	0.255	0.117	0.054	3.60	3.86	7%	3.705	3%	18.48
5	ε	K44 (○)	283	150	293	14.5	84	0.284	0.051	0.009	5.20	5.222	0%	5.090	2%	26.59
6	ζ	sNase (□)	136	17	298	13	49	0.346	0.096	0.026	2.12	3.64	72%	3.682	74%	9.50
7	η	K18 (○)	130	150	293	12.5	37	0.265	0.096	0.035	3.80	3.46	9%	3.446	9%	13.63
8	θ	SIC1 (Δ)	90	150	293	11	11	0.122	0.122	0.122	3.21	2.789	13%	2.820	12%	9.00
9	ι	p15PAF (◊)	111	150	298	9.5	34	0.302	0.086	0.024	2.81	3.217	14%	3.157	12%	11.14
10	κ	K10 (○)	168	150	293	9.5	47	0.259	0.057	0.012	4.00	3.971	1%	3.936	2%	12.60
11	λ	K19 (○)	99	150	293	9	29	0.273	0.091	0.03	3.50	3.017	14%	3.065	12%	10.85
12	μ	HST5 (○, ◊)	24	150	293	8.5	16	0.521	0.354	0.241	1.38	1.436	4%	1.394	1%	10.14
13	ν	CoINT (◊)	98	400	277	7.5	23	0.179	0.077	0.033	2.83	2.888	2%	2.715	4%	7.25
14	ξ	K23 (○)	254	150	293	6	73	0.276	0.024	0.002	4.90	4.81	2%	4.823	2%	20.59
15	π	K25 (○)	185	150	293	5.5	55	0.289	0.030	0.003	4.10	4.035	2%	4.012	2%	11.27
16	ρ	SH4UD (Δ)	85	217	293	5	18	0.200	0.059	0.017	2.82	2.676	5%	2.662	6%	3.76
17	q	hCyp (Δ)	115	42	298	4	30	0.243	0.035	0.005	2.51	3.119	24%	2.950	18%	2.69
18	p	An16 (Δ)	185	0	293	3	6	0.016	0.016	0.016	4.40	3.931	11%	3.721	15%	0.34
19	o	R15 (Δ)	114	128	298	2.5	40	0.338	0.022	0.001	2.33	3.079	32%	3.145	35%	-3.97
20	n	CspTm (Δ)	67	42	298	2	23	0.328	0.030	0.003	1.47	2.138	45%	2.223	51%	-3.44
21	m	Nucleoporin153 (○)	81	162	293	0	0	0.000	0.000	0.000	2.40	2.506	4%	2.424	1%	0.00
22	l	IN (Δ)	60	50	298	-2	20	0.168	0.014	0.004	2.16	2.203	2%	2.203	2%	-1.02
23	k	FhuA (◊)	143	150	298	-2	25	0.300	0.033	0.001	3.34	3.535	6%	3.484	4%	0.52
24	j	L (□)	64	128	298	-3	17	0.266	0.047	0.008	1.65	2.253	37%	2.341	42%	-3.61
25	i	R17 (Δ)	100	128	298	-6	38	0.360	0.060	0.010	2.37	2.884	22%	3.075	30%	-5.43
26	h	ACTR (□)	71	199	298	-7.5	19	0.261	0.106	0.043	2.51	2.476	1%	2.558	2%	-8.98
27	g	ERMTADn (○)	122	239	293	-8	38	0.295	0.066	0.015	3.96	3.301	17%	3.210	19%	-15.62
28	f	α-synuclein (□)	140	156	298	-8.5	40	0.282	0.061	0.013	3.30	3.584	9%	3.787	15%	-1.54
29	e	hNHE1cdt (Δ)	131	85	298	-12	41	0.305	0.092	0.027	3.63	3.41	6%	3.485	4%	-17.71
30	d	p53 (○)	93	208	293	-15	19	0.204	0.161	0.127	2.87	2.909	1%	3.051	6%	-18.64
31	c	OPN (◊)	219	150	298	-28.5	81	0.354	0.130	0.048	5.13	4.405	14%	4.436	14%	-14.56
32	b	ProTaC (□)	129	199	298	-40	69	0.512	0.310	0.188	3.63	3.932	8%	4.272	18%	-50.33
33	a	ProTaN (Δ)	112	42	298	-43	63	0.563	0.384	0.262	3.78	4.625	22%	4.671	24%	-54.39
												Average Error:	14.0%		14.0%	

V. RESULTS

We studied 33 different IDPs with varying numbers of amino acids (N = 24–283) with both net positive and net negative charges (see the sixth column in Table I). All these IDPs have been studied earlier by different CG models.[31,32,34] However, some of the IDPs, such as CspTm, Protein-L, hCyp, R15, and R17, are either actually intrinsically disordered regions (IDRs) of the folded proteins or unfolded state of normally folded proteins.[22,26] Within the range of our studied IDPs, An16 is a polyelectrolyte (PE) containing only six positively charged histidine residues, SIC1 contains six positively charged lysine and five positively charged arginine residues, Nucleoporin153 contains only uncharged residues, and the rest of the 30 IDPs are polyampholytes (PAs). The table is sorted according to their net charge Q (sixth column) from highly positive (in red) to highly negative (in dark green) values. The first row of the table lists K32 that contains 14% lysine (+1) and 4% aspartic acid (−1), which makes it highly positive. On the other hand, the bottom row is ProTa-N that contains 20% glutamic acid (−1), 17% aspartic acid (−1), and 8% Lysine (+1), which makes it highly negative. We assign unique letter codes for each of the IDPs, Greek letters α–ρ, in the ascending order starting from highly positively charged IDPs and in descending order with the alphabets starting from the negatively charged IDPs. The second to sixteen columns in Table I list the names of the IDPs, the length (N), the ionic concentration (I) in mM, the temperature [T(K)] at Kelvin scale, total charge Q,

number of charged beads (N_q), the absolute and net charge per unit length q_{abs} and q_{net} [Eq. (21)], the charge asymmetry parameter f^* [Eq. (18)], R_g^{expt}, R_g^{HPS1}, ΔR_g^{HPS1}, R_g^{HPS2}, ΔR_g^{HPS2} in units of nm, and the Wilson parameter $\mathcal{W}$ (Fig. 4), respectively.

The aforementioned two hydropathy scales HPS1[31] and HPS2[32] are used to study 33 IDPs using the coarse-grained simulation method described in Sec. II. The gyration radii R_g^{HPS1} and R_g^{HPS2} are obtained from simulation using hydropathy scales HPS1 and HPS2 as a function of ϵ (in kcal/mol) are then used to calculate the mean square error (MSE),

$$MSE = \frac{1}{N_{tot}} \sum_{i=1}^{N_{tot}} \left[R_g^{expt}(i) - R_g^k(i) \right]^2, \tag{9}$$

to compare the deviation from the experimental gyration radii R_g^{expt}. Here, N_{tot} is the number of different IDPs studied and k refers to either HPS1 or HPS2. In Fig. 2, we show the scatter plots of the experimental vs gyration radii for five different values of ϵ (ϵ = 0.08–0.13 kcal/mol for the HPS1 model and ϵ = 0.1–0.3 kcal/mol for the HPS2 model) and conclude that ϵ = 0.1 and 0.2 kcal/mol has the lowest errors (MSE = 0.23 and 0.26) for the HPS1 and HPS2 models, respectively. The gyration values from these models are nearly identical as shown in the inset (left) in Fig. 3(a). In this paper, we use HPS1 hydropathy scale with ϵ = 0.1 kcal/mol to further analyze the properties of the IDPs. However, we have carried out the same simulation studies using the HPS2 scale using ϵ = 0.2 kcal/mol (not shown in the manuscript) and the results are nearly identical to those from HPS1 model. Both the scales show ≈14% error (Table I, columns 13 and 15) for the set used here.

A. Universal scaling properties of the IDPs

Despite the fact the IDPs are mostly described as polyampholytes (PAs) or polyelectrolytes (PEs),[55] a fraction of experimental and theoretical studies using the HPS model by Dignon *et al.* describe IDPs as Gaussian chains;[43] however, in a recent publication, Thirumalai and co-workers using a two-bead CG model calculated the RMS $R_g \equiv \sqrt{\langle R_g^2 \rangle}$ and the end-to-end distance $R_N \equiv \sqrt{\langle R_N^2 \rangle}$, and concluded that globally the IDPs are described not as the Gaussian chains, but described as fully flexible excluded volume (EV) self-avoiding-walk (SAW) chains that obey the Flory scaling $R_g = aL^{0.59}$ in three dimensions (3D), where L is the contour length of the IDP.[35] We investigate this point further to find out to what extent the properties of the IDPs are universal. From theoretical arguments following Schaefer *et al.*[56] and Nakanishi,[57] it is established that a proper description of a semi-flexible EV chain characterized by a contour length L and a persistence length ℓ_p in d spatial dimensions is given by

$$\sqrt{\langle R_N^2 \rangle} \simeq b_l^{\frac{d-2}{d+2}} N^{\frac{3}{d+2}} \ell_p^{\frac{1}{d+2}} = b_l^{\frac{d+1}{d+2}} \left(\frac{L}{b_l} \right)^{\nu} \ell_p^{\frac{1}{d+2}}, \tag{10}$$

where N is the number of monomers of the chain so that length of the IDP is $L = (N-1)b_l \simeq Nb_l$ (for $N \gg 1$), b_l is the bond length between two neighboring monomers, and the mean-field Flory exponent $\nu = 3/(d+2)$ in 2D = 0.75 and in 3D = 0.60 (≈0.588 actual), respectively. This EV chain accurately describes the limit $L/\ell_p \gg 1$ and supersedes the Worm-like-chain model,[58]

$$\frac{\langle R_N^2 \rangle}{L^2} = \frac{2\ell_p}{L} \left(1 - \frac{\ell_p}{L} \left[1 - \exp\left(-L/\ell_p \right) \right] \right), \tag{11}$$

which does not take into account the EV effect, and hence saturates to $\langle R_N^2 \rangle = 2L\ell_p$ even when $L/\ell_p \gg 1$. In a previous publication, we have shown that scaled end-to-end distance $\langle R_N^2 \rangle/(2L\ell_p)$ and the scaled transverse fluctuation $\sqrt{\langle l_\perp^2 \rangle}/L$ as a function of L/ℓ_p collapse

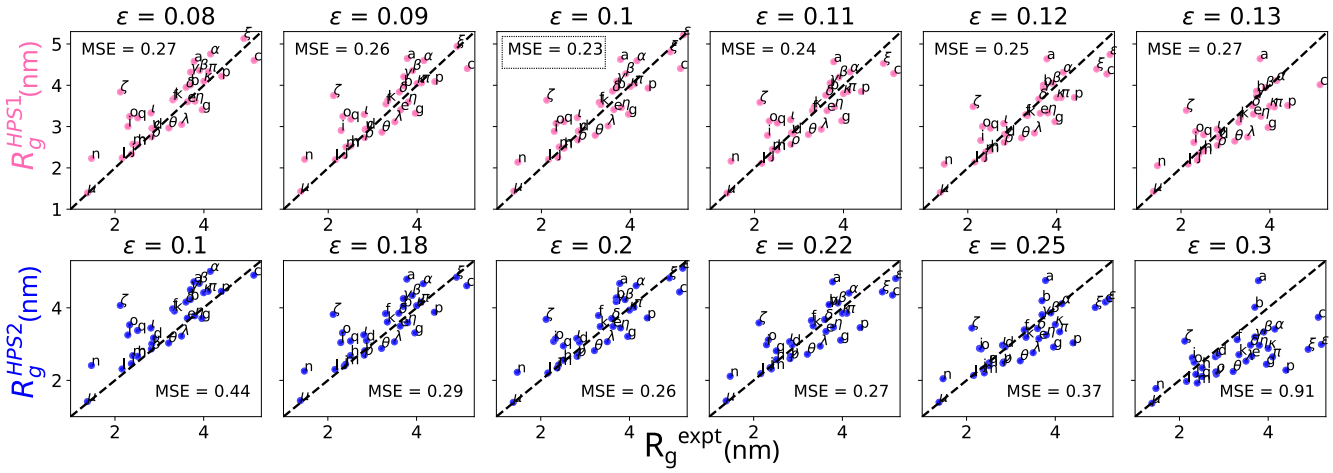


FIG. 2. The comparison of R_g^{sim} with R_g^{expt} for two different hydropathy models—the first row shows the simulation results with HPS1 hydropathy scale[31] and the HPS2 hydropathy scale[32] results are shown in the second row for different values of LJ interaction strengths ϵ in kcal/mol unit. The black dashed line serves as a guide to show the highest positive correlation between the experiment and simulation results with a unit slope, and the deviations are calculated as mean square errors, shown as MSE, in each subplot.

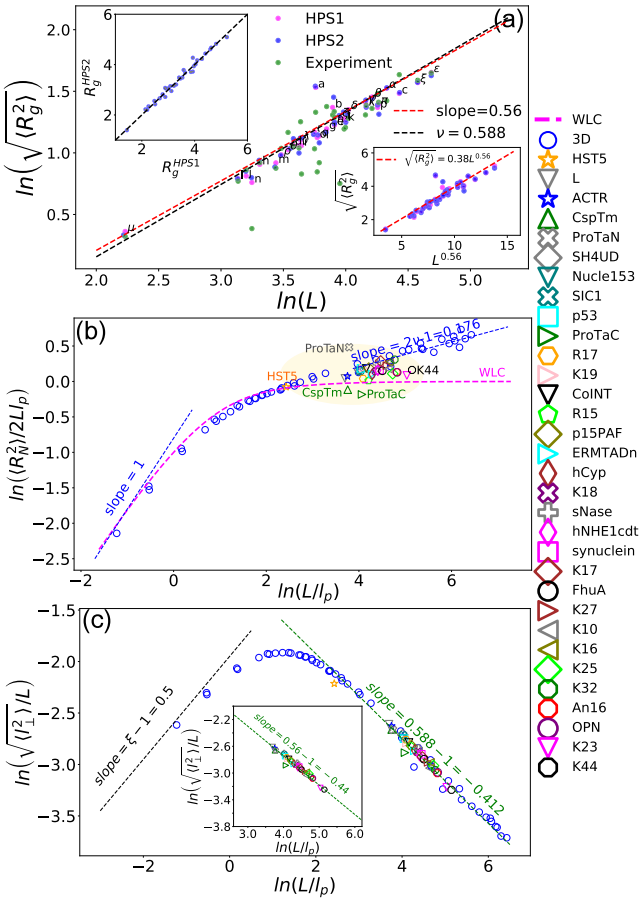


FIG. 3. (a) Log–log plot of IDPs radius of gyration as a function of chain length L obtained from HPS1 (magenta circles) and HPS2 (blue circles) models and from the experiments green circles). The red dashed line shows the best fit of the simulation data. (top) The inset (top-left) shows the comparison of the gyration radii from HPS1 and HPS2 models and the black dashed line indicates the line of unit slope. The other inset (bottom-right) shows the simulation gyration radii $\sqrt{\langle R_g^2\rangle}$ as a function of $L^{0.56}$ and the red dashed line corresponds to $\sqrt{\langle R_g^2\rangle} = 0.38L^{0.56}$. (b) Log–log plot of the scaled end-to-end distances, $\langle R_N^2\rangle/2L\ell_p$, as a function of L/ℓ_p for homopolymer chains in a blue-colored circle for a variety of combinations of L and ℓ_p for different IDPs in colored symbols for the HPS1 model. The dashed purple line in each figure shows the behavior of the WLC model [Eq. (11)]. (c) Log–log plot of the scaled transverse fluctuation $\sqrt{\langle l_\perp^2\rangle}/L$, as a function of L/l_p with the symbols having the same meaning as (b).

onto the same master curve[59,60] for all ratios of L/ℓ_p spanning rod to Gaussian and the EV limit. We would like to discuss our simulation findings for the IDPs in the context of these universal scaling plots (Fig. 3).

Figure 3(a) summarizes the simulation results from HPS1 (magenta solid circles) and HPS2 (blue solid circles) models. For comparison, we have also included the experimental points (green circles). The red dashed line corresponds to $\langle R_g\rangle = 0.38L^{0.56}$ [in nm shown explicitly as a power-law at the inset (bottom-right)]. Please note that the prefactor 0.38 nm is exactly the same as the average bond length obtained from the simulation. The dashed line (black) $\langle R_g\rangle \approx L^{0.588}$ corresponds to the excluded volume chain in 3D.

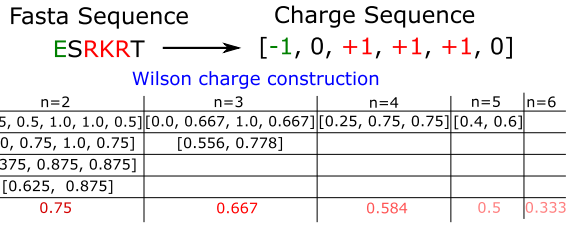
To get a clearer perspective, we have calculated the length of the IDPs and compared the scaled end-to-end distances $\langle R_n^2\rangle/2L\ell_p$ in units of dimensionless length L/ℓ_p[59,60] (blue circles) in Fig. 3(b). These data points will serve as a guide and help readers visualize the deviation of the scaling properties of the IDPs in reference to the universal master plot for the self-avoiding chain. The data points for the IDPs show that while H5T5, CspTm, and ProTaC lie in the Gaussian regime, the rest of the IDPs are in between Gaussian (WLC) and self-avoiding-walks in 3D but closer to being represented as self-avoiding-walks in 3D, consistent with the conclusion from Fig. 3(a). It is worth mentioning that a slightly lower exponent ($0.56 < 0.588$) may be attributed partly due to the finite size effect for the chain lengths $24 \le N \le 283$ used in this study. Our conclusion is further strengthened in the inset plot in Fig. 3(c), where we find that the scaled transverse fluctuations collapse with a slope of 0.56–1 = −0.44. These results possibly bridge and explain both the previous findings by Dignon *et al.*[43] where they identified the Gaussian behavior of the IDPs and by Baul *et al.*[34] where they described IDPs as self-avoiding chains. We will see in Sec. V C that IDPs can be further classified in terms of their skewness indices ($\mathcal{S}$-index captures the overall shape). $\mathcal{S}$-index will also affect an IDP's global positioning with respect to the universal master plot. For example, CspTm and ProTa-C having larger $\mathcal{S}$-index lie on the WLC line. This finer distinction makes IDPs more interesting. One can use the universal curve to understand why they are more Gaussian or behave more like a self-avoiding chain. In Secs. V B and V C, we will introduce such characteristics to further classify IDPs.

B. Wilson index $\mathcal{W}$ of the IDPs

Now that we have demonstrated an approximate universal description of the IDPs in terms of the HPS1 model, we want to demonstrate the sequence-specific features that make each IDP distinct and may exhibit very different behaviors from their homopolymer counterparts. Evidently, the charges present along the sequence play a crucial role in shaping the structure and dynamics of the IDPs. Out of 20 amino acids, only five of them carry a charge. Specifically, in the HPS1 model, both "R" and "K" have a +1 charge, "H" possesses a +0.5 charge, and "D" and "E" have charge of −1. However, the charges are randomly distributed along the chain backbone, and therefore, IDPs, in general, can be classified either as PA or PE.[7,21] Das and Pappu showed that weak polyampholytes form globules, whereas for a strong polyampholyte, the net charge per residue and their linear distribution along the chain backbone control their conformational preferences.[21] FRET experiments have explicitly demonstrated the expansion of charged IDPs at low ionic concentration.[65] Extensive research has been conducted to study this phenomenon in the existing literature. Nonetheless, our study delves deeper to investigate the positional implications of the amino acid sequence in terms of Wilson index ($\mathcal{W}$) as described below.

Unlike a homopolymer, an IDP can have varied degrees of local stiffness and flexibility resulting in the amino acids in different seg-

ments interacting with neighboring sequences even if they are far apart in the sequence space. To capture these potential dynamical interactions, we employ the concept of Wilson Renormalization extensively used to study the spin systems.[61] This renormalization approach allows us to analyze the sequence of charges and their unfolding interactions, considering interactions up to the next nearest neighbor. Figure 4 illustrates a hypothetical example of a short IDP sequence "ESRKRT" of length 6, showcasing the presence of a negative charge at the beginning followed by three positive charges in the middle, and the remaining amino acids being neutral. To initiate the averaging procedure, we select a window of length n. The simplest case $\tilde{\mathcal{W}}(2)$ considers sliding averages of window length $n = 2$ and denotes the next neighbor interactions. The window length for averaging can range from 2 to N, where N represents the number of amino acids in the IDP sequence. For the general case of $\tilde{\mathcal{W}}(n)$, where n consecutive charges are averaged, the process begins by sliding an averaging window from one end of the sequence toward the other. After the first step of averaging, denoted as S1, we obtain a new sequence of length $N - n + 1$ and use the new sequence to carry on the averaging procedure as



	n=2	n=3	n=4	n=5	n=6
S1	[-0.5, 0.5, 1.0, 1.0, 0.5]	[0.0, 0.667, 1.0, 0.667]	[0.25, 0.75, 0.75]	[0.4, 0.6]	
S2	[0.0, 0.75, 1.0, 0.75]	[0.556, 0.778]			
S3	[0.375, 0.875, 0.875]				
S5	[0.625, 0.875]				
$\tilde{\mathcal{W}}$(n)	0.75	0.667	0.584	0.5	0.333

FIG. 4. The schematic diagram shows the derivation of Wilson charge $\tilde{\mathcal{W}}(n)$ of a hypothetical fasta sequence ESRKRT of length 6. The table shows the different $\tilde{\mathcal{W}}(n) = 0.75 - 0.333$ for different window lengths $n = 2$–6 averages.

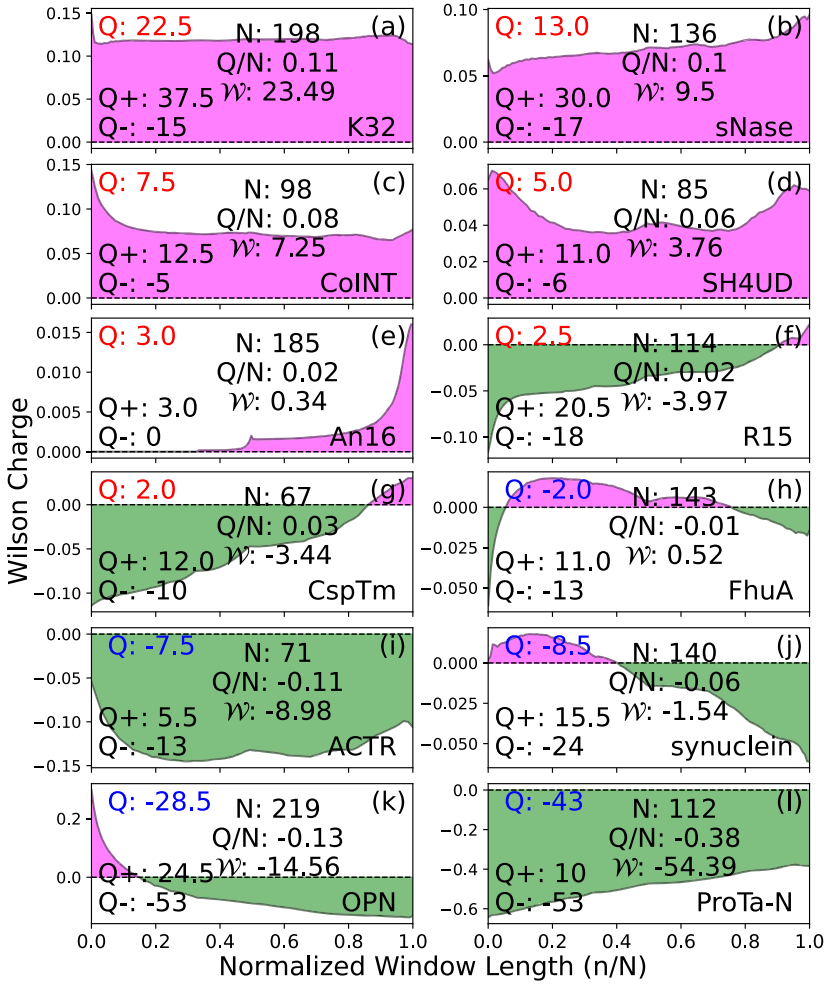


FIG. 5. The Wilson construction of charge distribution plotted against the corresponding window length for (a) K32, (b) sNase (c) CoINT (d) SH4UD, (e) An16, (f) R15, (g) CspTm, (h) FhuA, (i) ACTR, (j) α-synuclein, (k) OPN, and (l) ProTa-N in the order of positive to negative net charge content of the IDPs. Q denotes the net charge of the IDP along with positive and negative charges as $Q+$ and $Q-$, respectively. $\mathcal{W}$ denotes the area under the Wilson constructed curve. The color shade magenta/green shows the positive/negative intensity of the Wilson charge as a function of window length.

$$S0 : [a_1, a_2, \ldots, a_N], \tag{12a}$$

$$S1 : \left[\frac{1}{n}\sum_{i=1}^{n} a_i,\ \frac{1}{n}\sum_{i=2}^{n+1} a_i, \ldots,\ \frac{1}{n}\sum_{i=N-n+1}^{N} a_i\right], \tag{12b}$$

$$SN : \tilde{\mathcal{W}}(n) = \frac{1}{n}\sum_{i=1}^{n} a_i. \tag{12c}$$

We continue this procedure iteratively until we reach the final average value, represented as $\tilde{\mathcal{W}}(n)$. If the length of the charge sequence becomes less than the window length n during the averaging process, we terminate the procedure and calculate a global average to obtain $\tilde{\mathcal{W}}(n)$.

These averaging procedures with varied window sizes $n \in [2, N]$ can effectively capture the combination of charge interactions at different length scales. One can show that $\tilde{\mathcal{W}}(2)$ consider binomial interactions among the charges and expressed as

$$\tilde{\mathcal{W}}(n=2) = \frac{1}{2^{N-1}}\sum_{m=0}^{N-1}\binom{N-1}{m} a_m. \tag{13}$$

The higher order window averaging considers interactions of varying magnitudes, which can have an impact on determining dynamic conformations of IDPs. In Fig. 5, we explore $\tilde{\mathcal{W}}(n)$ for 12 IDPs with different total charges from highly positive (a) K32 (Q = 22.5), (g) CspTm (Q = 2.0) to highly negatively charged IDP (l) ProTa-N (Q = −43.0) as a function of normalized window length n/N. In the case of highly positively and negatively charged IDPs, the Wilson curves consistently remain above or below the zero line respectively. However, for IDPs with lower net charges, we sometimes observe the Wilson curve crossing from negative to positive in the case of (f) R15, (g) CspTm, and from positive to negative in the case of (j) α-synuclein, and (k) OPN. The area under the Wilson curve is denoted by $\mathcal{W} = \sum_{n=2}^{N} \tilde{\mathcal{W}}(n)$ and listed in the 16th column in Table I. It is conceivable that when plotted in normalized unit length scale, IDPs with similar Wilson charge $\tilde{\mathcal{W}}(n)$ will behave the same way and, thus, can be used as their fingerprints.

C. Charge patches and the local persistence length

The presence of charge patches introduces varying degrees of local stiffness along the chain backbone. During the BD simulation, we use a discrete chain, and the persistence length is calculated from[58]

$$\ell_p/\sigma = -\frac{1}{\ln(\cos\theta_i)}, \tag{14}$$

where θ_i is the angle between two bond vectors connecting the ith bead to the $(i \pm 1)$th beads.[60] We have checked that, for a homopolymer chain, this matches well with the continuum description of persistence length,[62]

$$\ell_p/\sigma = \kappa/k_B T \quad (3D). \tag{15}$$

IDPs with very similar net charges can have markedly different distributions of charges. An IDP containing correlated charge patches will have increased chain stiffness along that region that will affect its conformations and dynamics. To demonstrate this, we calculate the local persistence length (l_p) in nm along the chain using Eq. (14) for a few IDPs shown in Fig. 6. For example, CspTm has sparsely distributed charged residues with less net charge compared to ProTa-N

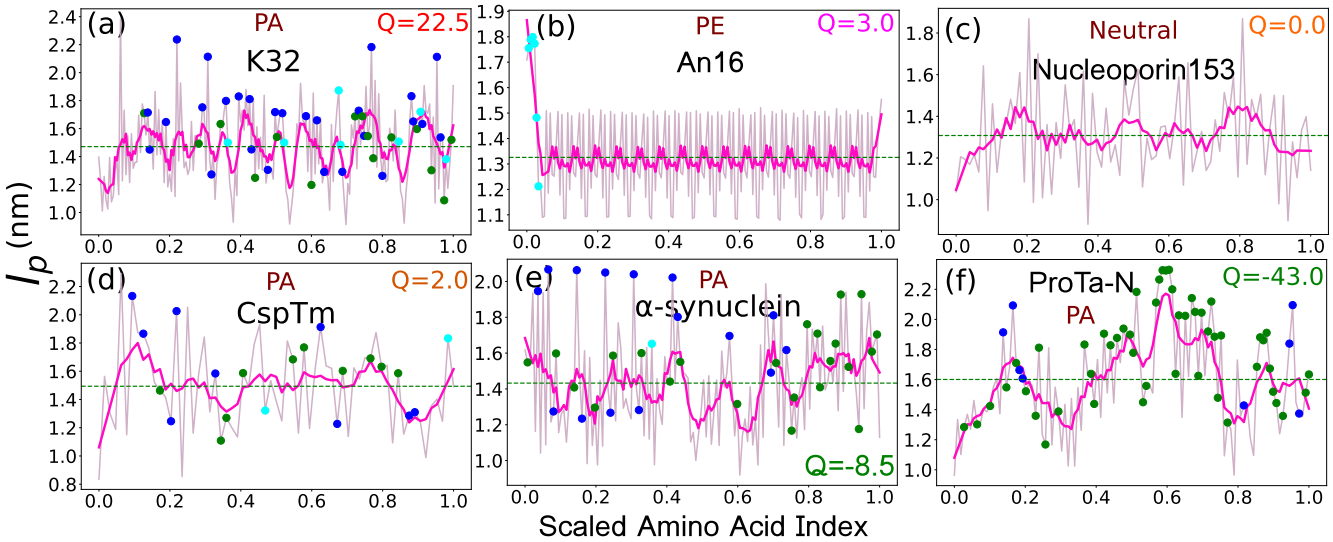


FIG. 6. The persistence length l_p is shown as a function of the normalized amino acid index for (a) K32, (b) An16, (c) Nucleoporin153, (d) CspTm, (e) α-synuclein, and (f) ProTa-N. In each case, the light pink line represents the persistence length from Eq. (14) while the magenta line corresponds to the sliding average $\langle l_p \rangle = \left(\sum_{i-2}^{i+2} l_p(i)\right)/5$. The positions of charged residues along the chain backbone are indicated by the blue-filled circle, cyan-filled circle, and darkgreen-filled circle for +1, +0.5, and −1.0 charged residues, respectively. The green dashed lines denote the average persistence length in each case.

containing mostly negatively charged residues in patches, and we observe an increase in l_p on those regions. The electrostatic repulsion among the same charge residues makes the chain locally stiffer and possibly has a deeper effect on their participation in biophysical processes.

D. Skewness index ($\mathcal{S}$-index) of the radius of gyration

The variation of the chain persistence length due to different charge species along the chain backbone is manifested in the shapes of the corresponding gyration radii that we measure in terms of the $\mathcal{S}$-index. The $\mathcal{S}$-index is obtained by fitting the distribution $P(\bar{R}_g)$ of the scaled gyration radii $\bar{R}_g = \sqrt{R_g^2}/\sqrt{\langle R_g^2\rangle}$ with an exponentially modified Gaussian distribution (exponnorm)[63] function given by

$$f(x,K) = \frac{1}{2K}\exp\left(\frac{1}{2K^2} - \frac{x}{K}\right)\mathrm{erfc}\left(-\frac{x-1/K}{\sqrt{2}}\right), \tag{16}$$

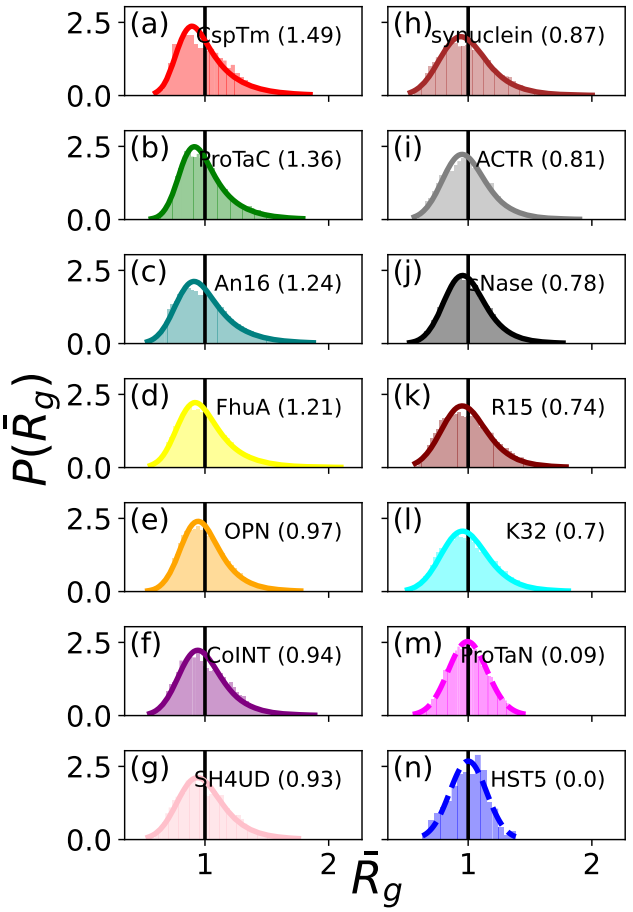


FIG. 7. $P(\bar{R}_g)$ for several IDPs as a function of the $\mathcal{S}$-index. In each figure, the solid colored line denotes the exponentially modified Gaussian fit for the $P(\bar{R}_g)$ histograms (please see the text) with the solid lines for $\mathcal{S} \geq 0.1$ while the dotted line for the near Gaussian distribution ($\mathcal{S} \leq 0.1$). The corresponding $\mathcal{S}$-index is written under parentheses in the legends.

where x is a real number, $K > 0$, and erfc is the complementary error function. The $\mathcal{S}$-index is then obtained from the following equation:

$$\mathcal{S} = \frac{2K^3}{\left(1+K^2\right)^{3/2}}, \tag{17}$$

where K corresponds to the shape parameter of the exponnorm distribution of Eq. (16). As shown below, the usefulness of the $\mathcal{S}$-index is manifested in capturing the overall shape of the distributions of the gyration radii, which then can be further analyzed as a function of the amino acid compositions and their net charge content. We observe in Fig. 7 that the shapes of the distribution of the gyration radii vary from being near Gaussian ($\mathcal{S} \to 0$) to an exponentially modified Gaussian distribution (that exhibits a tail for larger value of $\mathcal{S}$). Figure 7 confirms that most of the IDPs have long exponential tails ($\mathcal{S} \geq 1.0$), such as CspTm, ProTa-C, An16, FhuA, OPN, CoINT, SH4UD, α-synuclein, ACTR, sNase, R15, and K32. The highly charged IDP ProTa-N and HST5 are observed to have almost perfect Gaussian distribution with $\mathcal{S} \to 0$. Moreover, for the highly skewed distributions, the peaks shift toward the left that signifies the median is smaller than the mean and there is a propensity of these IDPs to expand occasionally. This skewness parameter can be utilized to further classify IDPs broadly into three categories [three colored regions in Fig. 8(a)] that characterize the propensity of expansions.

We now study how the $\mathcal{S}$-index depends on the charge content of the IDPs in more detail. In a previous study, Pappu *et al.*[20,21]

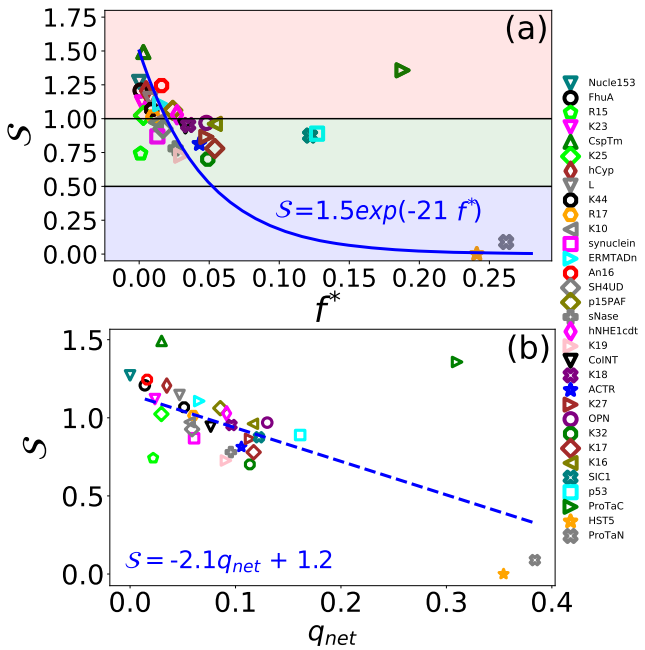


FIG. 8. (a) $\mathcal{S}$-index as a function of f^*. The solid blue line is an exponential fit. Light pink-, green-, and blue-shaded regions denote the high-, moderate-, and low-skewed radius of gyration of the IDPs, respectively. (b) f^* as a function of net charge. The straight line is a linear fit through the points.

demonstrated that the radius of gyration depends on the charge asymmetry parameter,

$$f^* = \frac{(f^+ - f^-)^2}{f^+ + f^-}, \tag{18}$$

where f^+/f^- is the net positive/negative charge per residue of an IDP. In Fig. 8(a), we plotted $\mathcal{S}$-index as a function of the f^* and based on the data it appears that there is an exponential decay of $\mathcal{S}$-index, although a power law fit could not be excluded. However, when plotted as a function of the net charge q_{net} (column 9 in Table I), a simpler relation is obtained [Fig. 8(b)]. The simulation data indicate that the $\mathcal{S}$-index decreases linearly with q_{net}. The linear fit that can be simplified to

$$\mathcal{S} \approx -2q_{net} + 1 \tag{19}$$

can serve as a useful relation. Since $\mathcal{S}$ cannot be negative, one observes a general relation

$$\lim_{q_{net} \to 0.5} \mathcal{S} \to 0 \text{ (Gaussian)}. \tag{20}$$

From Fig. 7 (last two rows of the right column), one observes that the shapes of the IDPs Prota-N and HST5 with $q_{net} = 0.384$ and 0.354 are almost Gaussian with $\mathcal{S} = 0.09$ and 0, respectively. This general conclusion will be tested when data for more IDPs will be available and analyzed in the future.

E. Chain conformations and ionic concentration

Solvent conditions, such as pH, temperature, ionic strength, and the presence of specific molecules, can significantly influence the conformations and, hence, the behaviors of the IDPs, particularly in a cellular environment. The robustness of the IDPs under external conditions can also be associated with the origin of life. Previous experimental studies[64–66] and simulation studies using CG models[67,68] have revealed conformational changes and salt-induced phase transition and looked at the liquid-liquid phase transitions in IDPs. IDPs are described either as PEs or PAs with varying amounts of net charge.[20] Thus, it is conceivable that screening will affect the conformational aspects in a significant way. Intuitively one can understand the behavior by using the idea of screening. The IDPs that are PE, an increase in salt concentration, will screen the net charge reducing the electrostatic repulsion, and hence, all the PEs with a net positive or negative charge will have reduced gyration radii as a function of increased screening. The case of PAs is a bit more subtle depending on an IDP's not only the net charge per residue q_{net} but also the fraction of the residues that are charged q_{abs} defined as follows:

$$q_{net} = \frac{\|Q_+| - |Q_-\|}{N} \quad \text{and} \quad q_{abs} = \frac{|Q_+| + |Q_-|}{N}, \tag{21}$$

where Q_+, Q_-, and N represent the total positive and negative charges and the number of amino acids in the IDP, respectively. In the literature, q_{net} and q_{abs} are previously denoted as NCPR and FCR, respectively.[21] For the PA, the loss/gain in electrostatic energy and entropy ultimately controls the show. We have made an extensive study of the dependence of gyration radii of the 33 IDPs (listed in Table I) on salt concentration under physiological conditions ranging from 0 to 300 mM shown in Fig. 9. The IDPs can be placed on any one of the four quadrants (I, II, III, and IV) of (q_{abs}, q_{net}) to study their dependency on salt concentration [Fig. 9(a)]. Based on the values of (q_{abs}, q_{net}), two decision boundary lines $q_{abs} = 0.275$ and $q_{net} = 0.13$ place the IDPs into four subclasses. Figure 9(a) displays scatter plots of the 33 IDPs classified into four quadrants, repre-

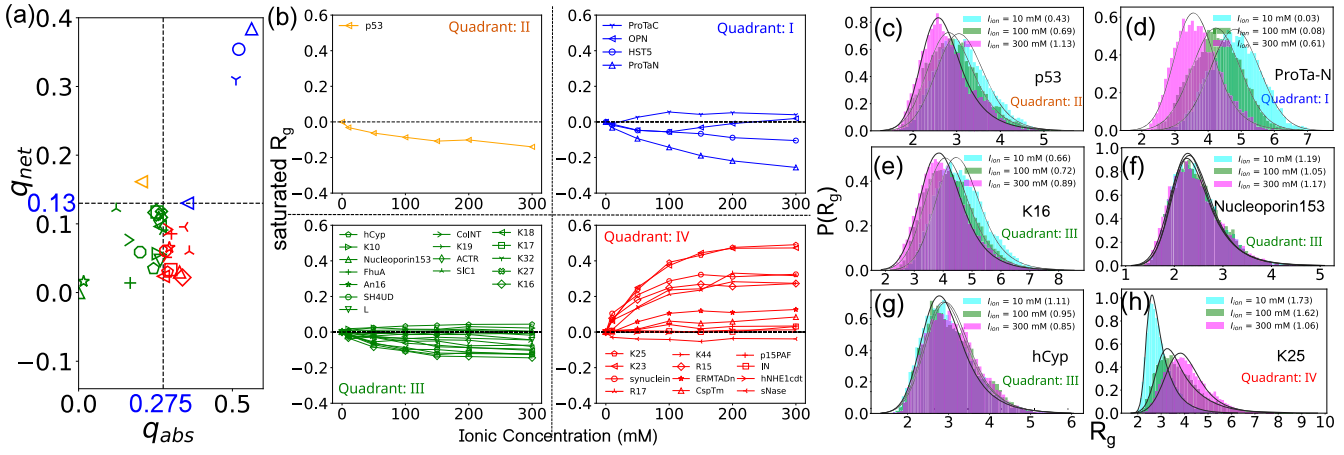


FIG. 9. (a) Characterization of IDPs using their net charge per residue q_{net} as a function of the absolute charge per residue q_{abs}. The two dotted black horizontal and vertical lines $q_{net} = 0.13$ and $q_{abs} = 0.275$, respectively, further identify the IDPs in terms of their placement into one of the four quadrants (I–IV). (b) $\sqrt{R_g^2/R_g^2(0)} - 1$ as a function of ionic concentration I. Separate colors and in each quadrant with different symbols for each IDP clearly identify different dependencies of the IDPs on the salt concentration. $P(\bar{R}_g)$s for p53 (c), ProTa-N (d), K16 (e), Nucleoporin153 (f), hCyp (g), and K25 (h) are shown for three different ionic concentrations 10 mM (cyan), 100 mM (green), and 300 mM (magenta), respectively. The black lines denote the exponentially modified Gaussian fit of the histograms and the corresponding $\mathcal{S}$-indices are denoted on the legends.

sented by blue, green, orange, and red symbols corresponding to the I, II, III, and IV quadrants, respectively. For each quadrant, a plot of the saturation values at each concentration is used to plot $\sqrt{R_g^2/R_g^2(0)} - 1$ as a function of ionic concentration, where $R_g(0)$ corresponds to the radius of gyration under ion-free conditions. Quadrants I and IV are easy to understand. A strong PA, such as Prota-N, lies in the quadrant-I as expected. Other PAs (HST5, OPN, ProTa-C) with large q_{net} and q_{abs} belong here. In this case, as the salt concentration increases, charge screening occurs, leading to a decrease in their gyration radii. This condition holds when only one type of charged residue is abundant in number. On the other hand, in quadrant IV, q_{abs} is high, but q_{net} is low, corresponding to a situation where there is a higher number of charged residues, yet they are almost equal in numbers. As both types of charges are present, at low salt concentration, the attraction between opposite charges reduces their radius of gyration due to electrostatic interactions. However, with increasing salt concentration, the charge screening effect comes into play, and the strength of electrostatic attraction among the oppositely charged residues decreases. Consequently, we observe a swelling of the IDPs, leading to a higher radius of gyration. Twelve IDPs belong to this category. In quadrants II and III, the q_{abs} value is low, indicating a low content of charged residues. In quadrant II, we find that p53 is the only IDP out of the 33 that falls into this category but has a high value of q_{net}. This pathological case is characterized by having only 17 negative charge residues (GLU and ASP) and two positively charged residues (ARG and LYS). Due to the charge screening effect mostly on the negatively charged residues, it can be inferred that the radius of gyration will decrease and that is indeed true as observed from the plot. In quadrant III, the q_{net} is low corresponding to IDPs that have less net charge per residue. We find 15 IDPs belong to this category. In this case, gyration radii of hCyp, FhuA, and K10 increase while gyration radii of K18, K17, K32, k27, and K16 decrease as a function of salt concentration. On the other hand, the remaining five IDPs, namely, Protein-L, SH4UD, An16, Nucleoporin153, and CoINT, are robust to the variation of ionic concentration as they are mostly low charge containing IDPs. The segregation of IDPs into four subsections unravels insights about their responses to salt concentration and provides a framework to classify other unknown proteins based on how they will behave under a wide range of salt solutions.

We further studied the accompanying variation in the shape of the distribution of the gyration radii by monitoring the $\mathcal{S}$-index as progressively more screening is introduced for the reason discussed in Sec. V D. Some examples are shown in Figs. 9(c)–9(h). The skewness of the distributions for K25 for 10, 100, and 300 mM ionic concentration changes from 1.73, 1.62, and 1.06, respectively, and they span a larger conformational space. On the other hand, ProTa-N, while the gyration radii decrease at higher salt concentrations, becomes more compact without altering the distribution shape. We also observe that gyration radii for a few IDPs (Protein-L, SH4UD, An16, Nucleoporin153, and CoINT) in Fig. 9(a) remain unaffected within the low salt limit of our study. With the change in the salt concentration, our simulation studies show that the degree of alteration in the shape of IDPs is different. The shape deformation is drastic in K25 compared to ProTa-N.

F. Summary and outlook

In conclusion, we used two different CG models (HPS1 and HPS2) to study both universal and fine structures of 33 IDPs and compared our results with available experimental data as well as simulation results for the same IDPs using other CG models. Our systematic studies of IDPs with fairly disparate levels of absolute and net charge (q_{abs}, q_{net}) and net hydropathy add many interesting characteristics to the prior studies using similar models. For the HPS1 model, the use of a larger set of IDPs and new experimental data[36] converges on the interaction parameter $\epsilon \approx 0.1$ kcal/mol, as opposed to 0.2 kcal/mol if we use the original set of Dignon *et al.*[31] (please refer to the supplementary material). For the HPS2 model, our optimized parameter 0.2 kcal/mol using 33 IDPs concurs with that of Tesei *et al.*[32,69] and seems to be more robust compared to the HPS1 model. Analysis of the results using the expanded dataset establishes a more robust and reliable framework for studying IDPs in bulk. The experimental results converge well with the simulation results for the optimum ϵ for both HPS1 and HPS2 models.

A natural question that has been addressed in the community that if sequence specificity makes every IDP distinct from each other, or they share some universal characteristics of homopolymers described by Flory's theory. We have been able to address both issues. A comparison of the scaled end-to-end distance and transverse fluctuations in reference to our recently established universal results,[59] and we observe that IDPs studied here interpolate from being Gaussian chains to self-avoiding chains to a different degree depending upon their amino acid composition.

We then study in detail how the absolute and net charge per residue (q_{abs}, q_{net}) manifest themselves in the finer characteristics of the IDPs. We come up with several new metrics that reveal the uniqueness of each IDP but leave room for making further classification of IDPs in different categories. The first one is the Wilson index $\mathcal{W}$ that on a normalized unit length scale demonstrates the uniqueness of each IDP and hence can be used as their fingerprints (e.g., We believe that using the $\mathcal{W}$ index of the newly discovered IDPs in future will allow us to predict their behaviors with those of similar $\mathcal{W}$ index.

Likewise, we demonstrate how the charge patches control the local stiffness and hence the overall conformations of the IDPs that we further characterize by introducing a skewness index $\mathcal{S}$. We first demonstrate that the distribution of the scaled gyration radii for highly charged IDPs with $\mathcal{S} \to 0$ are near-perfect Gaussians and that moderately charged IDPs with a larger value of the $\mathcal{S} \to 1.0 - 1.5$ are characterized by exponential tails. This observation immediately leads us to the observation of a simple linear dependence of $\mathcal{S}$ on q_{net} as approximated in Eq. (19). We related $\mathcal{S}$ parameter to the net charge asymmetry parameter f^* introduced by Das and Pappu,[21] where we observe a stronger exponential dependence. We believe that Fig. 8 along with Eq. (19) can serve as a good reference to understand the properties of the individual IDPs.

An important classification of the IDPs emerges from the study of salt dependence on the conformations of the IDPs. We find that IDPs exhibit very different characteristics and can be broadly placed in four different regions in the (q_{abs}, q_{net}) space.

We conclude with some comments that may promote further studies to perfect the CG models. We and many others used the

isotropic radius of gyration as the sole physical quantity for comparison as this is mostly available from the experiments. The CG model can be refined by introducing other quantities. For example, Wohl *et al.*[68] studied the salting-out effect on the liquid–liquid phase separation (LLPS) of IDPs by introducing a salt-dependent term into the hydropathy used in the HPS1 model. Maity *et al.*[67] introduced the molecular transfer model to study the salt-induced transitions. In the low concentrations of salts (≤1M), IDP conformations are affected by the degree of screening of electrostatic interactions of the charged residues and are independent of the specific salt identity, which is likely the regime that we have studied. However, at high concentrations, salts affect IDP conformations through salt-specific Hofmeister effects.[70,71] Thus, our studies will be useful in refining the existing hydropathy models for a wider range of parameter space.

We note that some of the IDPs remain unaffected with the variation of salt concentrations and, thus, can be compared with the behavior of other simpler amino acids identified and studied in the context of the origin of life.[72] Recently, Tesei *et al.* developed an efficient model to generate conformational ensembles of IDRs and reported conformational properties 28 058 IDRs from sequence only.[73] We believe that some of the new ideas introduced here can be tested for a much larger set and will open up several exciting avenues for future research to obtain a deeper understanding of the unique properties and behavior of IDPs.

SUPPLEMENTARY MATERIAL

A comparison of the gyration radii data from our simulation based on HPS1 hydropathy scale with the experimental data for a smaller subset of IDPs[31] is included in the supplementary material.

ACKNOWLEDGMENTS

All computations were carried out using STOKES High-Performance Computing Cluster at UCF. We thank Professor Kresten Lindorff-Larsen and Dr. Giulio Tesei from the University of Copenhagen for discussion, comparison of simulation results, and bringing Ref. 36 to our attention. We sincerely thank both the referees for their critical comments on the manuscript. Simulation results were obtained using HOOMD-blue.[74]

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Swarnadeep Seth: Conceptualization (equal); Data curation (lead); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Supervision (supporting); Validation (equal); Visualization (lead); Writing – original draft (equal); Writing – review & editing (equal). **Brandon Stine**: Data curation (equal); Investigation (supporting); Software (supporting); Writing – review & editing (equal). **Aniket Bhattacharya**: Conceptualization (lead); Data curation (supporting); Formal analysis (equal); Investigation (equal); Methodology (equal); Project administration (lead); Resources (lead); Software (supporting); Supervision (lead); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (lead).

DATA AVAILABILITY

Simulation data will be available upon request.

REFERENCES

[1] V. N. Uversky, J. R. Gillespie, and A. L. Fink, "Why are 'natively unfolded' proteins unstructured under physiologic conditions?," Proteins: Struct., Funct., Genet. **41**, 415–427 (2000).
[2] C. J. Oldfield and A. K. Dunker, "Intrinsically disordered proteins and intrinsically disordered protein regions," Annu. Rev. Biochem. **83**, 553–584 (2014).
[3] R. B. Best, "Computational and theoretical advances in studies of intrinsically disordered proteins," Curr. Opin. Struct. Biol. **42**, 147–154 (2017).
[4] K. Ghosh, J. Huihui, M. Phillips, and A. Haider, "Rules of physical mathematics govern intrinsically disordered proteins," Annu. Rev. Biophys. **51**, 355–376 (2022).
[5] T. Ehm *et al.*, "Intrinsically disordered proteins at the nano-scale," Nano Futures **5**, 022501 (2021).
[6] R. Evans, S. Ramisetty, P. Kulkarni, and K. Weninger, "Illuminating intrinsically disordered proteins with integrative structural biology," Biomolecules **13**, 124 (2023).
[7] R. van der Lee *et al.*, "Classification of intrinsically disordered regions and proteins," Chem. Rev. **114**, 6589–6631 (2014).
[8] M. Sickmeier *et al.*, "DisProt: The database of disordered proteins," Nucleic Acids Res. **35**, D786–D793 (2007).
[9] A. Deiana, S. Forcelloni, A. Porrello, and A. Giansanti, "Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell," PLoS One **14**, e0217889 (2019).
[10] P. E. Wright and H. J. Dyson, "Intrinsically disordered proteins in cellular signalling and regulation," Nat. Rev. Mol. Cell Biol. **16**, 18–29 (2014).
[11] J. J. Ferrie, J. P. Karr, R. Tjian, and X. Darzacq, "'Structure'-function relationships in eukaryotic transcription factors: The role of intrinsically disordered regions in gene regulation," Mol. Cell **82**, 3970–3984 (2022).
[12] A. Borgia *et al.*, "Extreme disorder in an ultrahigh-affinity protein complex," Nature **555**, 61–66 (2018).
[13] A. Sottini *et al.*, "Polyelectrolyte interactions enable rapid association and dissociation in high-affinity disordered protein complexes," Nat. Commun. **11**, 5736 (2020).
[14] H. Y. J. Fung, M. Birol, and E. Rhoades, "IDPs in macromolecular complexes: The roles of multivalent interactions in diverse assemblies," Curr. Opin. Struct. Biol. **49**, 36–43 (2018).
[15] V. N. Uversky, C. J. Oldfield, and A. K. Dunker, "Intrinsically disordered proteins in human diseases: Introducing the D_2 concept," Annu. Rev. Biophys. **37**, 215–246 (2008).
[16] O. Coskuner-Weber, O. Mirzanli, and V. N. Uversky, "Intrinsically disordered proteins and proteins with intrinsically disordered regions in neurodegenerative diseases," Biophys. Rev. **14**, 679–707 (2022).
[17] C. P. Brangwynne, P. Tompa, and R. V. Pappu, "Polymer physics of intracellular phase transitions," Nat. Phys. **11**, 899–904 (2015).
[18] A. C. Murthy *et al.*, "Molecular interactions underlying liquid–liquid phase separation of the FUS low-complexity domain," Nat. Struct. Mol. Biol. **26**, 637–648 (2019).

[19]S. Das, Y.-H. Lin, R. M. Vernon, J. D. Forman-Kay, and H. S. Chan, "Comparative roles of charge, π, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins," Proc. Natl. Acad. Sci. U. S. A. **117**, 28795–28805 (2020).
[20]A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, and R. V. Pappu, "Net charge per residue modulates conformational ensembles of intrinsically disordered proteins," Proc. Natl. Acad. Sci. U. S. A. **107**, 8183–8188 (2010).
[21]R. K. Das and R. V. Pappu, "Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues," Proc. Natl. Acad. Sci. U. S. A. **110**, 13392–13397 (2013).
[22]H. Hofmann *et al.*, "Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy," Proc. Natl. Acad. Sci. U. S. A. **109**, 16155–16160 (2012).
[23]P. Bernadó and D. I. Svergun, "Structural analysis of intrinsically disordered proteins by small-angle x-ray scattering," Mol. BioSyst. **8**, 151–167 (2012).
[24]B. Schuler, A. Soranno, H. Hofmann, and D. Nettels, "Single-molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins," Annu. Rev. Biophys. **45**, 207–231 (2016).
[25]B. Schuler, "Perspective: Chain dynamics of unfolded and intrinsically disordered proteins from nanosecond fluorescence correlation spectroscopy combined with single-molecule FRET," J. Chem. Phys. **149**, 010901 (2018).
[26]K. A. Merchant, R. B. Best, J. M. Louis, I. V. Gopich, and W. A. Eaton, "Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations," Proc. Natl. Acad. Sci. U. S. A. **104**, 1528–1533 (2007).
[27]S. Kosol, S. Contreras-Martos, C. Cedeño, and P. Tompa, "Structural characterization of intrinsically disordered proteins by NMR spectroscopy," Molecules **18**, 10802–10828 (2013).
[28]G.-N. W. Gomes *et al.*, "Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET," J. Am. Chem. Soc. **142**, 15697–15710 (2020).
[29]A. D. Sokal, in *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*, edited by K. Binder (Oxford University Press, New York, 1995), Chap. 2.
[30]H. S. Ashbaugh and H. W. Hatch, "Natively unfolded protein stability as a coil-to-globule transition in charge/hydropathy space," J. Am. Chem. Soc. **130**, 9536–9542 (2008).
[31]G. L. Dignon, W. Zheng, Y. C. Kim, R. B. Best, and J. Mittal, "Sequence determinants of protein phase behavior from a coarse-grained model," PLoS Comput. Biol. **14**, e1005941 (2018).
[32]G. Tesei, T. K. Schulze, R. Crehuet, and K. Lindorff-Larsen, "Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties," Proc. Natl. Acad. Sci. U. S. A. **118**, e2111696118 (2021).
[33]J. M. Lalmansingh, A. T. Keeley, K. M. Ruff, R. V. Pappu, and A. S. Holehouse, "SOURSOP: A Python package for the analysis of simulations of intrinsically disordered proteins," J. Chem. Theory Comput. **19**, 5609–5620 (2023).
[34]U. Baul, D. Chakraborty, M. L. Mugnai, J. E. Straub, and D. Thirumalai, "Sequence effects on size, shape, and structural heterogeneity in intrinsically disordered proteins," J. Phys. Chem. B **123**, 3462–3474 (2019).
[35]M. L. Mugnai *et al.*, "Sizes, conformational fluctuations, and SAXS profiles for intrinsically disordered proteins," bioRxiv:2023.04.24.538147 (2023).
[36]T. Dannenhoffer-Lafage and R. B. Best, "A data-driven hydrophobicity scale for predicting liquid–liquid phase separation of proteins," J. Phys. Chem. B **125**, 4046–4056 (2021).
[37]C. L. Johnson *et al.*, "The two-state prehensile tail of the antibacterial toxin colicin N," Biophys. J. **113**, 1673–1684 (2017).
[38]J. A. Riback *et al.*, "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water," Science **358**, 238–241 (2017).
[39]G. Platzer *et al.*, "The metastasis-associated extracellular matrix protein osteopontin forms transient structure in ligand interaction sites," Biochemistry **50**, 6113–6124 (2011).
[40]S. Jephthah, L. Staby, B. B. Kragelund, and M. Skepö, "Temperature dependence of intrinsically disordered proteins in simulations: What are we missing?," J. Chem. Theory Comput. **15**, 2672–2683 (2019).
[41]J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," J. Mol. Biol. **157**, 105–132 (1982).
[42]J. Habchi, P. Tompa, S. Longhi, and V. N. Uversky, "Introducing protein intrinsic disorder," Chem. Rev. **114**, 6561–6588 (2014).
[43]D. S. Devarajan *et al.*, "Effect of charge distribution on the dynamics of polyampholytic disordered proteins," Macromolecules **55**, 8987–8997 (2022).
[44]S. Alberti, A. Gladfelter, and T. Mittag, "Considerations and challenges in studying liquid–liquid phase separation and biomolecular condensates," Cell **176**, 419–434 (2019).
[45]S. Alberti and D. Dormann, "Liquid–liquid phase separation in disease," Annu. Rev. Genet. **53**, 171–194 (2019).
[46]J. McCarty, K. T. Delaney, S. P. O. Danielsen, G. H. Fredrickson, and J.-E. Shea, "Complete phase diagram for liquid–liquid phase separation of intrinsically disordered proteins," J. Phys. Chem. Lett. **10**, 1644–1652 (2019).
[47]S. Das and M. Muthukumar, "Microstructural organization in α-synuclein solutions," Macromolecules **55**, 4228–4236 (2022).
[48]H.-Y. Chou and A. Aksimentiev, "Single-protein collapse determines phase equilibria of a biological condensate," J. Phys. Chem. Lett. **11**, 4923–4929 (2020).
[49]J. N. Israelachvili, *Intermolecular and Surface Forces*, 3rd ed. (Elsevier, 2011).
[50]G. C. Akerlof and H. I. Oshry, "The dielectric constant of water at high temperatures and in equilibrium with its vapor," J. Am. Chem. Soc. **72**, 2844–2847 (1950).
[51]D. M. Engelman, T. A. Steitz, and A. Goldman, "Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins," Annu. Rev. Biophys. Biophys. Chem. **15**, 321–353 (1986).
[52]T. P. Hopp and K. R. Woods, "A computer program for predicting protein antigenic determinants," Mol. Immunol. **20**, 483–489 (1983).
[53]D. Eisenberg, E. Schwarz, M. Komaromy, and R. Wall, "Analysis of membrane and surface protein sequences with the hydrophobic moment plot," J. Mol. Biol. **179**, 125–142 (1984).
[54]J. L. Cornette *et al.*, "Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins," J. Mol. Biol. **195**, 659–685 (1987).
[55]V. Yamakov, A. Milchev, H. Jörg Limbach, B. Dünweg, and R. Everaers, "Conformations of random polyampholytes," Phys. Rev. Lett. **85**, 4305–4308 (2000).
[56]D. W. Schaefer, J. F. Joanny, and P. Pincus, "Dynamics of semiflexible polymers in solution," Macromolecules **13**, 1280–1289 (1980).
[57]H. Nakanishi, "Flory approach for polymers in the stiff limit," J. Phys. **48**, 979–984 (1987).
[58]M. Rubinstein and R. H. Colby, *Polymer Physics* (Oxford University Press, 2003).
[59]J. Bair, S. Seth, and A. Bhattacharya, "Universality in conformations and transverse fluctuations of a semi-flexible polymer in a crowded environment," J. Chem. Phys. **158**, 204902 (2023).
[60]A. Huang, A. Bhattacharya, and K. Binder, "Conformations, transverse fluctuations, and crossover dynamics of a semi-flexible chain in two dimensions," J. Chem. Phys. **140**, 214902 (2014).
[61]S. K. Ma, *Modern Theory of Critical Phenomena* (Taylor & Francis, 1976).
[62]L. D. Landau and E. M. Lifshitz, *Statistical Physics, Part 1*, 3rd ed. (Pergamon Press, 1980).
[63]See http://scipy.github.io/devdocs/reference/generated/scipy.stats.exponnorm.html for scipy.stats.exponnorm—SciPy v1.12.0.dev Manual. (n.d.).
[64]H. Matsunaga and H. Ueda, "Stress-induced non-vesicular release of prothymosin-α initiated by an interaction with S100A13, and its blockade by caspase-3 cleavage," Cell Death Differ. **17**, 1760–1772 (2010).
[65]S. Müller-Späth *et al.*, "Charge interactions can dominate the dimensions of intrinsically disordered proteins," Proc. Natl. Acad. Sci. U. S. A. **107**, 14609–14614 (2010).

[66]R. Vancraenenbroeck, Y. S. Harel, W. Zheng, and H. Hofmann, "Polymer effects modulate binding affinities in disordered proteins," Proc. Natl. Acad. Sci. U. S. A. **116**, 19506–19512 (2019).
[67]H. Maity, L. Baidya, and G. Reddy, "Salt-induced transitions in the conformational ensembles of intrinsically disordered proteins," J. Phys. Chem. B **126**, 5959–5971 (2022).
[68]S. Wohl, M. Jakubowski, and W. Zheng, "Salt-dependent conformational changes of intrinsically disordered proteins," J. Phys. Chem. Lett. **12**, 6684–6691 (2021).
[69]G. Tesei and K. Lindorff-Larsen, "Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range," bioRxiv:2022.07.09.499434 (2022).
[70]M. T. Record, E. Guinn, L. Pegram, and M. Capp, "Introductory Lecture: Interpreting and predicting Hofmeister salt ion and solute effects on biopolymer and model processes using the solute partitioning model," Faraday Discuss. **160**, 9–44 (2013).
[71]L. M. Pegram *et al.*, "Why Hofmeister effects of many salts favor protein folding but not DNA helix formation," Proc. Natl. Acad. Sci. U. S. A. **107**, 7716–7721 (2010).
[72]A. Pohorille, M. A. Wilson, and G. Shannon, "Flexible proteins at the origin of life," Life **7**, 23 (2017).
[73]G. Tesei *et al.*, "Conformational ensembles of the human intrinsically disordered proteome: Bridging chain compaction with function and sequence conservation," bioRxiv:2023.05.08.539815 (2023).
[74]J. A. Anderson, J. Glaser, and S. C. Glotzer, "HOOMD-blue: A Python package for high-performance molecular dynamics and hard particle Monte Carlo simulations," Comput. Mater. Sci. **173**, 109363 (2020).