UNIVERSITY OF CALIFORNIA, MERCED

Decoding the Structural Response of Disordered Proteins to their Surrounding Environments

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Quantitative and Systems Biology

by

Feng Yu

Committee in charge:

Professor Andy Liwang, Chair
Professor Michael Colvin
Professor Daniel Beller
Professor Shahar Sukenik, Advisor

2023

The Dissertation of Feng Yu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

_____

Professor Michael Colvin Date

_____

Professor Daniel Beller Date

_____

Professor Shahar Sukenik, Advisor Date

_____

Professor Andy Liwang, Chair Date

University of California, Merced
2023

DEDICATION

This is for my dad Bingyi Yu, my mom Donghui Hao and my girlfriend Fengxian Wang.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

IDR	intrinsically disordered region
FRET	fluorescence resonance energy transfer
SAXS	small-angle X-ray scattering
CD	circular dichroism
R_g	radius of gyration
R_e	end-to-end distance

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Shahar Sukenik, for guidance and support throughout my Ph.D. studies. I also extend my sincere thanks to my committee members Dr. Andy Liwang, Dr. Michael Colvin, and Dr. Daniel Beller for their advice and for being there during my doctoral journey.

I would like to thank my dad, mom, girlfriend, and entire family for their support throughout my life. Additionally, a special thanks goes to my friends, particularly Yuxin Wang, for their companionship and support in facing life's challenges.

This work used the Delta cluster at UIUC and the FASTER at TAMU through allocation BIO230024 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The work covered in this dissertation is supported by NIH under Award R35GM137926, NSF IntBIO Grant 2128067, and the NSF-CREST Center for Cellular and Biomolecular Machines at UC Merced (NSF-HRD-1547848 and NSF-HRD-2112675). The simulation covered in this dissertation is also supported by computing time on the MERCED cluster at UC Merced, (NSF Grant ACI-1429783,) on the XSEDE computational infrastructure framework, Grant TG-MCB190103(NSF Grant ACI-154856), and on the Expanse cluster through HPC@UC, (NSF Grant OAC-1928224).

Chapter 2, in full, is a reprint of the material as it appears in David Moses, Feng Yu, Garrett M. Ginell, Nora M. Shamoon, Patrick S. Koenig, Alex S. Holehouse, and Shahar Sukenik The Journal of Physical Chemistry Letters 2020 11 (23), 10131-10136.

Chapter 3, in full, is a reprint of the material as it appears in Feng Yu and Shahar Sukenik. The Journal of Physical Chemistry B 2023 127 (19), 4235-4244.

Chapter 4, in part, has appeared in Cesar L. Cuevas-Velazquez, Tamara Vellosillo, Karina Guadalupe, Hermann Broder Schmidt, Feng Yu, David Moses, Jennifer A. N. Brophy, Dante Cosio-Acosta, Alakananda Das, Lingxin Wang, Alexander M. Jones, Alejandra A. Covarrubias, Shahar Sukenik and José R. Dinneny Nature Communication 2021 12: 5438.

Curriculum Vitae

Feng Yu

Education

University of California, Merced	Ph.D.	2023
University of California, Merced	M.S.	2021
Nanjing University	B.S.	2018

Working Experience

Bioinformatics Postdoctoral Fellow, SAXS Beamline
Lawrence Berkeley National Laboratory — Dec. 2023

Scientific Developer Intern - Molecular Modeling
OpenEye, Cadence Molecular Sciences — May. 2023 - Aug. 2023

Joint Genome Institute Internship
Lawrence Berkeley National Laboratory — Jun. 2022 - Aug. 2022

Data Science Challenge - Graduate Student Mentor
Lawrence Livermore National Laboratory — May. 2022 - Jun. 2022

Graduate Computational Mentor
NSF Center for Cellular and Biomolecular Machines — Oct. 2021 - Dec.2023

Research Grants

Research Enhancement Awards ($10,000) — Principle Investigator
NSF Center of Research Excellence in Science and Technology - Center for Cellular and Biomolecular Machines — 2023

Explore ACCESS Allocation BIO230024 — Principle Investigator
NSF Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support program — 2023

Research Credit ($1,000)
Google Cloud — 2022

Awards

Student Research Achievement Award
Biophysical Society 66th Annual Meeting — 2022

GSA Travel Award
Graduate Student Association, UC Merced — 2022

GRAD-EXCEL Peer Mentorship Program Award
Graduate Division, UC Merced — 2022-2023

QSB Travel Award
Quantitative and Systems Biology Program, UC Merced 2022

QSB Summer Research Fellowship
Quantitative and Systems Biology Program, UC Merced 2019,2020,2023

CCBM Scholar
NSF-CREST Center for Cellular and Biomolecular Machines 2019 - 2023

Third Prize, Basic STEM Research forum
Nanjing University 2017

Outstanding Social/Humanitarian Program
Nanjing University 2015

National Basic Subject Talent Scholarship
Nanjing University 2015

Professional Experience

Journal Reviewer
PeerJ - Life and Environment 2023

Undergraduate Poster Award Competition Judge
Biophysical Society 67th Annual Meeting 2023

Co-organizer, Saturday Night Thermo
The 36th Gibbs Conference on Biothermodynamics 2022

Poster Judge, JGI 25th Annual Meeting
Joint Genome Institute, Berkeley Lab 2022

Editor, IDP State Letter
Biophysical Society Intrinsically Disordered Proteins Subgroup 2021 - 2022

“ITxia” (IT Service Desk) Repair Group Lead
Nanjing University 2017

Physics Education in Community Program
Nanjing University Humanitarian Program 2015

Publications

Yu, F., Sukenik, S. (2023). Structural Preferences Shape the Entropic Force of Disordered Protein Ensembles. The Journal of Physical Chemistry B, 127(19), 4235–4244.

Moses, D.*, Yu, F.*, Ginell, G.M., Shamoon, N.M., Koenig, P.S., Holehouse, A.S.†, and Sukenik, S†. Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to their

Chemical Environment. The Journal of Physical Chemistry Letters 2020 11 (23), 10131-10136

Cuevas-Velazquez, C. L., Vellosillo, T., Guadalupe, K., Schmidt, H. B., Yu, F., Moses, D., …, Dinneny, J. R. (2021). Intrinsically disordered protein biosensor tracks the physical-chemical effects of osmotic stress on cells. Nature Communications, 12(1), 5438.

Moses, D., Guadalupe, K., Yu, F., Flores, E., Perez, A., Mcanelly, R., Shamoon, N. M., Cuevas-Zepeda, E., Merg, A. D., Martin, E. W., Holehouse, A. S., Sukenik, S. (n.d.). Structural biases in disordered proteins are prevalent in the cell. https://doi.org/10.1101/2021.11.24.469609 (Accepted by Nature Structural and Molecular Biology)

Biswas, S., Gollub, E., Yu, F., Ginell, G., Holehouse, A., Sukenik, S., Boothby, T. C. (n.d.). Helicity of a tardigrade disordered protein promotes desiccation tolerance.

Sanchez-Martinez, S., Nguyen, K., Biswas, S., Nicholson, V., Romanyuk, . . . Yu, F. . . . Boothby, T. C. (2023). Labile assembly of a tardigrade protein induces biostasis. BioRxiv, 2023.06.30.547219.

Hesgrove, C. S., Nguyen, K. H., Biswas, S., Childs, C. A., Shraddha, K. C., Medina, B. X., . . .Yu, F., . . .Boothby, T. C. (2021). Molecular Swiss Army Knives: Tardigrade CAHS Proteins Mediate Desiccation Tolerance Through Multiple Mechanisms. BioRxiv, 2021.08.16.456555.

Conference Presentations

Yu, F., Sukenik, S, and Holehouse, A.S. Scanning solution space to uncover intrinsically disordered protein response to changing solutions. Poster presentation delivered at the Protein Folding Consortium Workshop, St. Louis, MO, June 2019

Yu, F., Sukenik, S, and Holehouse, A.S. Entropic Force in Disordered Proteins. Oral and Poster presentation delivered at the Northern California Computational Biology Symposium, Sacramento, CA, October 2019

Yu, F., Sukenik, S, and Holehouse, A.S. Assessing the sensitivity of disordered proteins to changes in their surrounding solution. Poster presentation delivered at 2020 Symposium on Intrinsically Disordered Proteins: From Fundamental Biology to Human Disease Workshop, Stanford, CA, February 2020

Yu, F., Sukenik, S, The Dimensions of Intrinsically Disordered Proteins Determine their Solution Sensitivity. Poster presentation delivered at 34th Gibbs Conference on Biological Thermodynamics, Online, September 2020

Yu, F., Sukenik, S, Sequence and Chemical Environment Determine the Intrinsically Disordered Protein Ensemble Preference. Poster presentation delivered at Biophysical Society 65th Annual Meeting, Online, February 2021

Yu, F., Sukenik, S, Hidden structures influence intrinsically disordered protein solution sensitivity. Poster presentation delivered at the Protein Folding Consortium Workshop, St. Louis, MO, June 2021

Yu, F., Sukenik, S, Long-range hidden structures determine the sensitivity of intrinsically disordered proteins. Oral and poster presentation delivered at 35th Gibbs Conference on Biological Thermodynamics, Online, September 2021

Yu, F., Sukenik, S, Linking Disordered Protein Sequence and Ensemble Using Interaction Maps. Poster presentation delivered at Biophysical Society 66th Annual Meeting, San Francisco CA, February 2022 (Award Winner)

Yu, F., Sukenik, S, The entropic force exerted by disordered proteins is determined by their structural biases. Poster presentation delivered at The 36th Gibbs Conference on Biothermodynamics, Carbondale IL, October 2022

Yu, F., Sukenik, S, Structural Preferences Modulate the Entropic Force Exerted by Disordered Proteins. Oral presentation delivered at Biophysical Society 67th Annual Meeting, San Diego CA, February 2023

ABSTRACT OF THE DISSERTATION

Decoding the Structural Response of Disordered Proteins to their Surrounding Environments

by

Feng Yu

Doctor of Philosophy in Quantitative and Systems Biology

University of California, Merced, 2023

This dissertation examines how the sequence and the chemical environment of intrinsically disordered protein regions (IDRs) affect their structure. Unlike structured proteins, IDRs do not adopt a singular structure. Instead, they exist in a dynamic conformational ensemble. The conformations that make up this ensemble are shaped by a range of molecular interactions, both with the environment and within the IDR itself. The absence of a stable three-dimensional structure, along with their high level of exposure to the solvent environment, makes IDR exceptionally adaptable to changes in their physical and chemical surroundings. My research employs computational methods to investigate how IDRs respond to different chemical environments and physical constraints. I will discuss how we quantify the contribution of IDR structural preferences to their entropic forces and describe how we leverage IDR sensitivity to design biosensors.

Chapter 1: Introduction

Background

The three-dimensional structures of well-folded proteins are critical to their functions. However, this paradigm does not hold for the entire proteome. Intrinsically disordered protein regions (IDRs) make up over 30% of the human proteome and exist in organisms across all kingdoms of life[1–3]. Unlike well-folded proteins, IDRs do not adopt fixed native structures. Instead, they exist in a dynamic state, with constantly changing conformations that provide them with a unique set of functional capabilities[4]. These diverse and interchangeable conformations are collectively referred to as the IDR's ensemble (**Fig. 1.1A).**

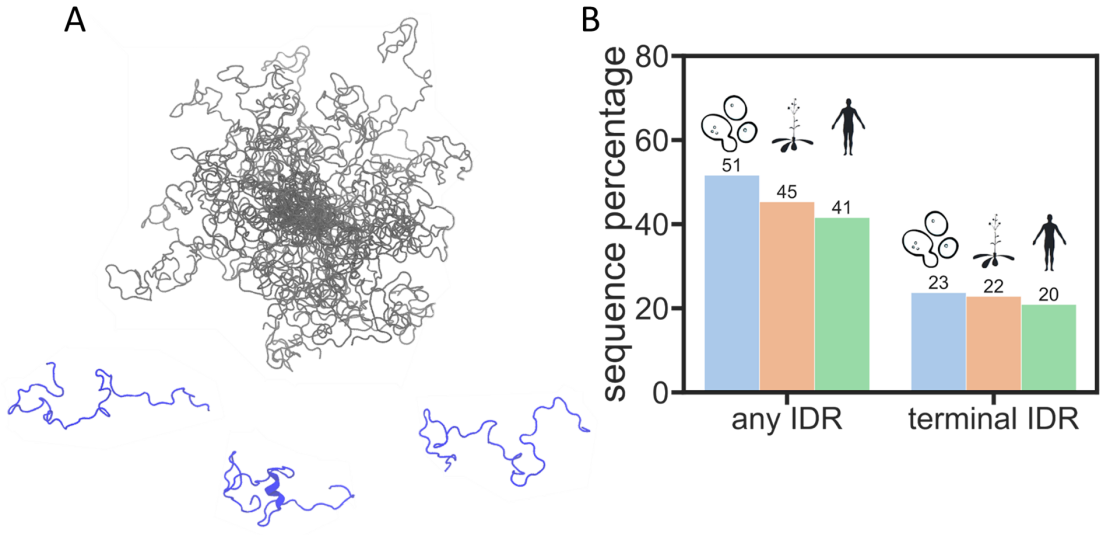


Figure 1.1. (**A**) IDR conformational ensembles. The top part represents the overlay of simulated p53-NTAD IDR conformations. Blue conformations are selected from the entire simulation to show the diversity of IDR conformations. (**B**) The percentage of proteins that have at least one IDR (longer than 30 amino acids) in the yeast, arabidopsis, or human proteomes. The analysis was conducted based on the AlphaFold2 database. Terminal IDRs represent IDRs located at the N-terminal/C-terminal of the entire protein sequence[5].

IDR amino acid sequences are typically rich in polar and charged residues but depleted in hydrophobic amino acids. This composition prevents the formation of a hydrophobic core that is inherent in well-folded proteins. Instead, this composition maintains the IDR in a dynamic and adaptable state[6].

The distinct amino acid composition of IDRs enabled computational biologists to predict the existence of thousands of IDR sequences in different genomes[3,7] (**Fig. 1.1B**). However, due to their lack of structure, the functional significance of IDRs was not recognized. In several works in the 1990s, several IDRs were shown to play a role in DNA recognition and transcription activation[8–11]. Since then, more IDRs with critical

functions have been identified by experimental methods, gaining the interest of the biophysics community[12–14].

IDR ensemble plasticity leads to important biological functions

The lack of stable structure and the ability to adopt multiple conformations of IDRs are referred to here as IDR structural plasticity[15]. This plasticity is a central feature of IDRs, intricately linked to their functions[16,17]. It allows them to alter their shapes and surfaces to interact with specific binding partners[18], and enables them to participate in and regulate complex protein networks[19–22]. For example, in signal transduction, Hypoxia-inducible factor 1-alpha (HIF-1α), can adopt different conformations to bind with interaction partners including DNAs to activate several survival pathways and induce the expression of hypoxia-related survival genes[23–25]. In another example, the dynamic ensemble of the p53 N-terminal IDR alternates between a free state and interaction with the DNA-binding domain, thereby enhancing its binding selectivity for target genes and inhibiting nonspecific binding[26]. As the functional mechanisms of IDRs are deeply intertwined with the structural plasticity of their ensemble, there is a growing interest in quantifying the dynamic behavior of IDR ensembles.

The conformational ensembles of IDRs

Describing the structural plasticity of IDRs requires quantifying the physical properties of their conformational ensembles. Experimentally, only the average properties of IDR conformational ensembles can be readily resolved. These include metrics such as the average helical propensity in the sequence, its radius of gyration (R_g), or its end-to-end distance (R_{ee}) (**Fig. 1.2A-D**). To examine specific conformations in the entire ensemble, molecular simulations are often combined with experimental data to obtain accurate structural properties[27–29].

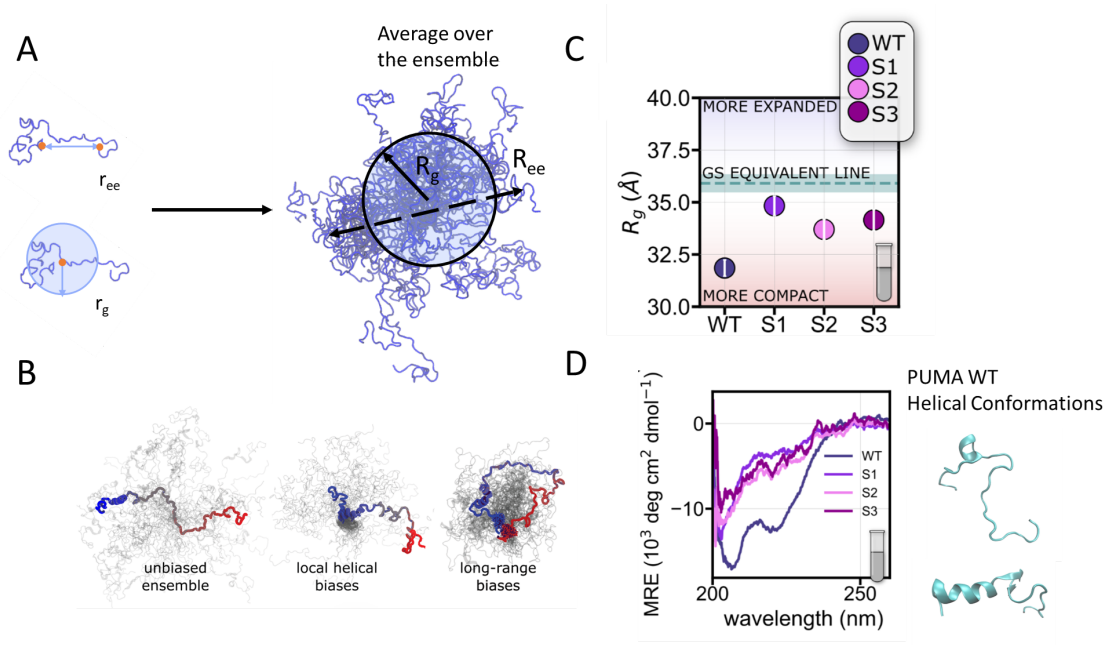


Figure 1.2. (**A**) R_g and R_{ee} can describe IDR global dimensions and structural preferences. Left: r_g and r_{ee} represent the radius of gyration and the end-to-end distance of a single IDR conformation. Right: schematic diagram of the average R_g and R_{ee} on a p53 N-terminal IDR ensemble consisting of 50 individual simulation conformations. (**B**) Comparison between IDR helical propensity with unique long-range structural biases. These structures are transient structures that can exist in the same IDR ensemble at different times. (**C**) PUMA scrambles demonstrated different global dimensions because of the sequence arrangement. (**D**) PUMA WT demonstrated helical propensity while other scrambles demonstrated no secondary structure despite similar amino acid composition[30].

IDR ensemble-function relationship

The central dogma in structural biology is that the 3D structure of proteins determines protein function[9,31]. Instead of a native structure, an IDR's function is impacted by its structural ensemble. With their dynamic structural preferences, IDRs mainly perform their function by binding with their binding partners. But how does a dynamic conformational ensemble lead to stable binding with other molecules?

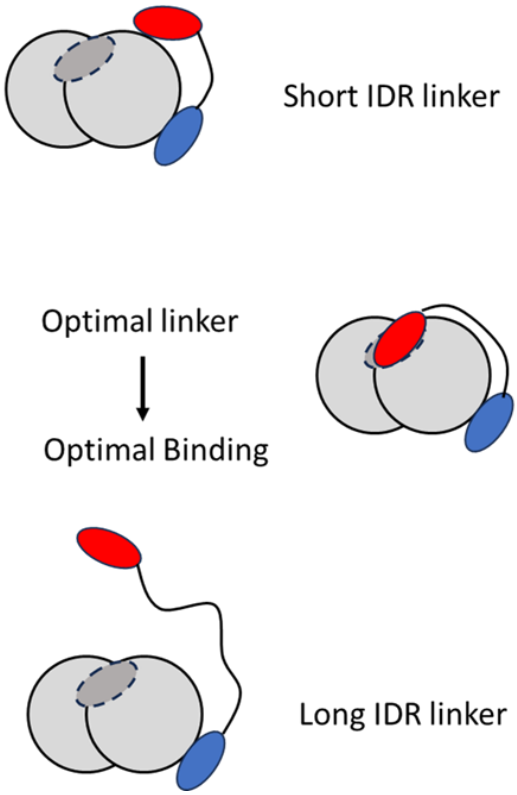


Figure 1.3. IDR linker global dimensions determine the binding affinity between adenovirus early region 1A (E1A) protein SLiMs and retinoblastoma (Rb) tumor suppressor[32].

One possible IDR binding mechanism is the "conformational selection" mechanism. An IDR ensemble adopts multiple pre-existing conformations in dynamic equilibrium and, upon binding to a ligand or another biomolecule, “selects” the conformation that has the highest affinity for the binding partner[33,34]. For example, the activator for thyroid hormone and retinoid receptors (ACTR) IDR ensemble forms an α-helix and may bind to the nuclear coactivator binding domain (NCBD) of the CREB binding protein. Mutations that increase helical propensity in the ACTR unbounded ensemble will stabilize the ACTR: NCBD binding complex[35]. Thus, increasing helical propensity in the sequence will enhance the affinity of ACTR to NCBD. In another example, the intrinsically disordered c-Myb activation domain of the cAMP-response element binding (CREB) protein will form helical structures when it binds to the KIX domain of the CREB-binding protein (CBP). It is discovered that the association rate of binding between different c-Myc mutants is correlated with the helix population of the IDR[36]. This suggests that changes in IDR sequences will alter IDR ensemble preferences and lead to IDR functional changes[37,38].

Another possible mechanism is “conformational buffering”. It suggests that dimensions of IDR linkers between binding sites are critical and optimized for the binding affinity between the binding partners. For example, the length of a disordered linker region can influence the binding affinity between the adenovirus early gene 1A (E1A) protein and the retinoblastoma (Rb) protein (**Fig. 1.3B**). E1A contains two important binding sites, E2F and LxCxE, that are separated by a linker IDR. It was found that the dimensions of

this linker mediate its binding to the Rb protein. A more compact IDR ensemble with a shorter linker may hinder the simultaneous binding of both sites, reducing the binding affinity. A more extended IDR ensemble with a longer linker may reduce the cooperativity between two binding sites and their binding affinity. Therefore, the E1A protein maintains the dimensions of the IDR ensemble at an optimal level, enhancing its binding effectiveness with the Rb protein. In diverse adenoviruses and across multiple hosts, the sequences of the linker regions between two binding sites show variations but molecular simulations indicate that R_{ee} of these linker regions remain conserved even across different species[32]. This result further suggests the importance of IDR ensemble preference in IDR functional mechanisms.

In summary, IDRs perform functions through specific ensemble preferences encoded within their sequences. As a result, a comprehensive understanding of IDR ensembles is crucial to link IDR sequences with function.

IDR ensembles in the context of their physical-chemical environments

The cellular environment experiences changes during various biological processes or due to external stress[39–41]. Such changes include physical-chemical parameters such as pH, temperature, ionic strength, and the composition of small and large solutes and biomolecules[42–46]. For example, in many cancer cells, there is a significant alteration in the cellular metabolism which will elevate the pH in the cellular environment (also known as the Warburg effect)[47,48]. In another example, intracellular pH will vary during each stage of the cell cycle, a change that is essential to the regulation of cell growth[49,50].

The examples above highlight the dynamic nature of cellular environments. Understanding the nature of IDR conformational changes induced by these environmental changes is crucial for elucidating the physiological roles and functional mechanisms of IDRs. It is also vital for developing biosensors or engineered biomaterials that can respond to cellular environmental changes. However, a biophysical understanding of the underlying molecular rules that define environmentally-driven structural changes in IDRs is lacking. This dissertation uncovers some of the rules that determine **how the surrounding environment changes the dimensions of IDR ensembles.**

IDR ensembles are sensitive to surrounding environments

As described above, IDRs lack stable intramolecular bonds, exhibit a flexible conformational ensemble, and have a high degree of surface area exposure, making them sensitive to surrounding environmental changes[51–53]. For example, denaturing agents like urea are commonly used to probe the stability of protein structures. They act by forming multiple weak attractions with the peptide backbone. These attractions cause the backbone of the protein to expand and unfold. However, for urea to act on well-folded proteins requires concentrations upwards of 8M. For IDRs, even low urea concentrations, below 1M, drive a measurable increase in ensemble dimensions. The disparity between the response of IDRs and well-folded proteins to urea underlines the

distinct nature of IDRs: their lack of stable tertiary and secondary structures means that the forces maintaining their conformational integrity are much weaker and more susceptible to disruption by weak interactions with their environment[54–57]. Indeed, variations in other solution parameters, including but not limited to pH, ionic strength, and temperature, can all induce shifts in the equilibrium of IDR conformations[58–60].

IDRs are notably affected by the spatial constraints of their environment, such as structured proteins and cell membranes. The existence of these biological constraint surfaces will lead to a reduction in the number of conformations IDRs can adopt and a corresponding decrease in conformational entropy. As a consequence, the IDR conformational ensemble will be changed and will also generate a force to regain the conformational entropy[5,61,62]. This will enable IDR membrane curvature sensing ability and alter the protein binding affinity.

These examples show that environmental changes in the physical-chemical environment of IDRs have a significant impact on their structural ensemble. Because of the role of the ensemble in the IDR function, understanding how IDRs respond to environmental factors is pivotal for decoding their role in the cell. The study of IDR ensemble structural sensitivity in different physical-chemical environments, even outside of the cell, may therefore reveal insights into the molecular basis of their involvement in regulatory processes. This understanding can also be instrumental in designing responsive biosensors.

Changing the surrounding environment can alter IDR functions

By changing IDR ensembles, surrounding environments can alter IDR functions. While few examples of this have been shown unequivocally, some studies show this clearly. For example, the binding of histone H1 and highly disordered prothymosin α (ProTα) is significantly impacted by the presence of salt, which affects the ionic strength of the environment. Their interaction, largely driven by electrostatic forces due to their large and opposite net charges, is sensitive to changes in ionic strength[63,64].

The surrounding chemical environment will alter IDR binding activity by screening electrostatic interaction with electrolytes. For example, association kinetics between PUMA IDR and Mcl-1 region is ion-dependent and salt-dependent. Overall, high ionic concentration will decelerate the association. Divalent cations demonstrate a larger impact on binding compared to monovalent cations such as potassium. The helical propensity of PUMA IDR is reduced by more than 15% within salt solutions. This finding suggests ionic strength alters IDR function by changing the IDR ensemble structure and interrupting the IDR folding-upon-binding mechanism[38].

Summary

Structural preferences of IDR ensembles are more sensitive to changes in their surrounding environment than the structure of well-folded proteins. With an established link between structure and function, it becomes necessary to have an accurate description of their ensemble in various physical-chemical contexts in order to understand their function. My Ph.D. research focuses on computational analysis to

describe IDR response to surrounding environmental change. In this dissertation, I aim to address how the physical-chemical environment influences IDR ensembles and functions using computational approaches.

In Chapter 2, I will explore how solution environments impact IDR ensembles. This chapter shows that an IDR's solution response is encoded in its sequence and is determined by intramolecular interactions. This builds the foundation for quantifying IDR responses to different chemical environments.

Chapter 3, I will delve into how IDRs generate entropic force when an IDR ensemble is limited by spatial constraints. This work demonstrates that an IDR's global dimensions are correlated with the strength of the entropic force it exerts.

In Chapter 4, I will combine the IDR solution response elucidated in the first two chapters to design an IDR biosensor of cellular osmotic pressure. The biosensor is designed based on a naturally occurring, desiccation-related IDR which can form transient secondary structure in changing chemical environments.

Finally, in Chapter 5, I will give a summary of my entire journey and discuss how these findings contribute to the IDR field. I will also give some insights into potential applications of my IDR solution sensitivity research.

Methodology

Importance of simulations in understanding IDR ensembles

As explained earlier, IDR ensembles can be key to understanding IDR function. However, several challenges exist in characterizing these ensembles structurally. First and foremost, the inability to crystallize IDRs has limited the use of one of the most critical historical tools in structural biophysics, X-ray crystallography[65]. Even newer structural methods such as Cryo-EM cannot resolve the flexible nature of the IDR ensemble[66]. Instead, several methods have emerged, including, nuclear magnetic resonance (NMR) spectroscopy, Förster resonance energy transfer (FRET), and small-angle X-ray scattering (SAXS), that have become the primary experimental methods for characterizing IDR structural preferences[67–70].

However, these experiments provide relatively low-resolution data for dynamic IDR ensembles[71,72]. SAXS can provide the average R_g of the IDR ensemble along with atom pair distance distributions[73]. FRET measures the average R_{ee} of the IDR ensemble[74]. NMR can provide more detailed atomistic information about the ensemble but still can only measure its average properties. To move beyond such average properties and understand the full range of structural preferences contained in an ensemble, there has been extensive use of molecular simulations[27–29,75,76].

Two common computational methods used to generate IDR ensembles are molecular dynamics (MD) and Monte Carlo (MC) simulations. MD simulations for IDR utilize force fields to apply Newton's laws of motion, translating interatomic forces into time-dependent changes in atom positions. By calculating these forces and integrating

them over time steps, MD explores the evolution over time of IDR dynamic structural ensembles. MC simulations, on the other hand, for IDRs involve using stochastic algorithms to sample the conformational space that IDRs can occupy.

MC simulations can capture the broad ensemble of conformations that IDPs adopt, offering insights into their structural properties. IDRs present a unique energy landscape characterized by many local minima, each representing a distinct preferred conformational state. This landscape is marked by small energy barriers that separate these minima, allowing the IDR to transition easily between different states. Exploring such a landscape poses significant challenges, as traditional molecular dynamics (MD) methods can often become trapped in these local minima, unable to efficiently sample the vast array of possible conformations. MC simulations, with their ability to randomly propose conformations of the IDR, can scan the entire conformational space more effectively than MD simulations, which might need multiple simulations with different starting conformations to fully scan the IDR conformational space. For MC simulation, we can introduce temperature variations to help the system escape the energy barriers, facilitating broader exploration of the conformational space.

MC simulations can be more computationally efficient for IDRs. Since they do not require the calculation of detailed atomic forces and velocities at each time step like MD, they can sample a wider range of conformations in a shorter amount of computational time. The MC algorithm I used for IDR simulation in this paper takes days on a single computational core and is suitable for high-throughput simulations. In comparison, MD simulation may require a hundred CPU cores and a GPU to simulate the conformational ensemble of a single IDR. Despite a faster simulation speed, MC simulations still generate accurate IDR ensembles, allowing for a good understanding of IDR structural preferences[77–79].

The central algorithm for MC simulations used in this dissertation was introduced in the 1950s and is widely known as the Metropolis–Hastings algorithm[80]. This algorithm is a specific variant within the broader category of MC simulation, and it plays a crucial role in generating thermodynamically accurate ensembles, especially in the context of studying molecular systems like proteins. This algorithm distinguishes itself by using the Boltzmann criterion for its accept/reject decisions. When a new conformation is proposed during the simulation, the algorithm evaluates it based on the Boltzmann distribution, which relates the probability of a state to its energy and the temperature of the system.

With an initial protein state A, a new random protein state B under the rigid constraints of the protein will be generated through the random change of the torsional angle between the residues of the protein[81]. This state B may not be accepted by the simulation algorithm and is thus called a trial move.

After each trial move, an energy calculation is performed using the all-atom force field. In my simulation, I will use the ABSINTH force field which will be introduced later. The move is accepted or rejected based on the total energy of the states.

1. If $E_B < E_A$, the state B will be accepted as a valid conformation of the protein.

2. If $E_B > E_A$, another random number $a \in [0, 1]$ will be generated. If $a < e^{-\frac{E_B - E_A}{kT}}$, the state B will be accepted otherwise it will be rejected.

Thus, the overall acceptance probability of state B is described as

$$p_{A->B} = min[1, exp(-\beta * \Delta E)] \quad (1.1)$$

Here, $\beta = \frac{1}{kT}$ and $\Delta E = E_B - E_A$. State B will serve as the subsequent starting point for the next step, and this process will persist until the number of simulation steps predetermined by the user is reached. All states accepted throughout this simulation will collectively constitute a thermodynamically accurate ensemble insofar as the forcefield used is accurate.

Solution Space Simulation

Currently, most all-atom simulation force fields are calibrated with experimentally measured IDR conformational ensembles in a dilute aqueous buffer. To simulate IDRs in different solution environments, we created the Solution Space (SolSpace) simulation method. This method is implemented with the CAMPARI MC simulation software[82]. To assess the energy of each conformation (E_A, E_B in **Eq. 1.1**) we use the ABSINTH force field. Energies are calculated using an effective Hamiltonian that quantifies the total energy of the system, which is the sum of 4 energy terms:

$$E_{total} = W_{solv} + U_{LJ} + W_{el} + U_{corr} \quad (1.2)$$

Here, U_{corr} is a correlation term applied to keep the peptide dihedral predominantly in the trans-configuration. W_{el} is the electric potential term that is the Coulomb potential between charged residues.

$$W_{el} = \sum_{i=1}^{N_{CG}} \sum_{k=1}^{n_i} \sum_{j=i+1}^{N_{CG}} \sum_{l=1}^{n_j} f_{ij} \frac{q_k^i q_l^j}{4\pi\varepsilon_0 r_{kl}} s_{kl} \quad (1.3)$$

Here, N_{CG} represent the number of charge groups of the molecule defined by the defined reference force field such as OPLSAA or CHARMM. n_i represent the number of point charges in the charge groups. The Coulomb interaction will not be calculated if the charge groups possess atoms that are (1-2)- or (1-3)-bonded to one another. In this case, f_{ij} will be zero otherwise unity. s_{kl} represent the solvent electrostatic screening effect on the Coulomb interaction between point charges.

In addition, U_{LJ} is the Lennard-Jones potential between protein residues.

$$U_{LJ} = 4 \sum_i \sum_{j<i} f_{ij} \varepsilon_{ij} [(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}] \quad (1.4)$$

r_{ij} describes the distance between atoms i and j. σ_{ij} is the distance where U_{LJ} is zero, referred to as a pairwise size parameter. ε_{ij} represents the dispersion energy. f_{ij} will be 1 if atom i and j are separated by more than one rotatable bond. Otherwise f_{ij} will be zero and we will ignore the Lennard-Jones interaction between those two atoms.

The Lennard-Jones potential is a simplified model to describe how the potential energy between two residues varies with distance: it predicts a strong repulsion as they get very close and a weaker attraction at moderate distances, attributable to transient induced dipole-dipole (van der Waals) forces[83,84].

W_{solv} is the effective solution-protein interaction energy function based on the implicit solvent model. In implicit solvent models, the solvent is not represented as individual molecules but rather through its averaged effects on the solute. In the ABSINTH model, the interactions between the solvent and the molecule are determined by considering the free volume surrounding the molecule's atoms. This approach calculates the solvent-accessible surface area based on the extent to which the space around each target atom is unoccupied by other atoms. When an area around a target atom is completely free of other atoms, it is considered to be fully solvated. This will simplify the calculation of solvent and reduce computational costs because it will neglect the interaction between solvent molecules but with some tradeoff in structural accuracy.[82] The fundamental discrete solvent-molecule interactions to continuum solvent interaction field approximation in the simplified model leads to the omission of the finite size of water molecules and tightly bound water molecules, which may be crucial for the function or stability of some conformations[85]. Below I give some explanation about the W_{solv} energy function term.

With SolSpace simulations, W_{solv} are modified based on transfer-free energy (TFE). TFE is the free energy cost of moving a molecule from one solution to another (**Fig. 1.4**)[86]. TFEs are calculated per amino acid surface area exposed to the solution. Thus, changes in solution might make one conformation less favorable than another, shifting the structural biases within IDR ensembles (**Fig. 1.4**). To account for this effect, the solvent accessible surface area (SASA) for each conformation of an IDR needs to be calculated, and the transfer free energy ΔG_{tr} calculated by[55] :

$$\Delta G_{tr} = \sum_{i=1}^{N_{RG}} \alpha_i \Delta g_i \qquad (1.5)$$

Here, N_{RG} is the total number of IDR amino acids. Solvation groups represent the type of amino acid side chain or the backbone units which all have the same transfer-free energy. α_i is the total surface area of the IDR solvation group i. Δg_i is the group transfer free energy per surface area in the solvation group i. The values of Δg_i for all amino acid side chains and backbone in some solvents have been previously measured experimentally[86,87].

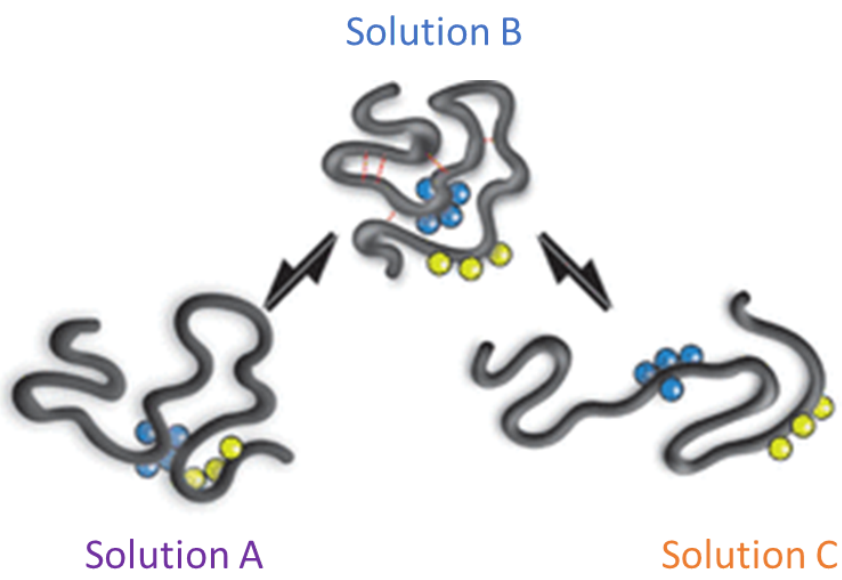


Figure 1.4. IDR may adopt different conformations in different solutions. TFE is the free energy change when the protein is moved from Solution A to Solution B.

To simulate transferring an IDR between different solution conditions, we need to eliminate the effect of conformational change and set up a reference point for the TFE calculation. Therefore, we defined the maximum possible TFE (MTFE) of the IDR corresponding to the TFE of the most extended IDR conformation. MTFE is calculated analytically for a given sequence using the following formula:

$$W_{solv}^{max} = \sum_{i=1}^{N_{SG}} (n_i \Delta g_i + n_i \theta_i \Delta g_{BB}) \qquad (1.6)$$

Here, N_{SG} represents the number of amino acid side chain groups. n_i is the element number in the i th side-chain group. Δg_i is the group transfer energy of the i th side-chain group to a given solution. Δg_{BB} is the group transfer energy of the backbone. θ_i is the correction factor that responds to the backbone SASA fraction of different amino acids. For example, glycine does not have a sidechain, so it will have $\theta = 1$ while $\Delta g_i = 0$. For phenylalanine θ is around 0.6. For a given sequence, MTFE represents W_{solv} for the most expanded conformation of the chain. This maximally expanded conformation does not change regardless of the solute identity, and so altering W_{solv}^{max} by a certain percentage difference corresponds to a change in IDP: solution interactions. This is measured by

$$\psi = \frac{W_{solv}^{max}(solution) - W_{solv}^{max}(water)}{W_{solv}^{max}(water)} \times 100\% \qquad (1.7)$$

If $\psi > 0$, the model solution is attractive to the IDR compared to the buffer condition. If $\psi < 0$, the model solution is repulsive to the IDR compared to the buffer condition. Using our model solution condition, we can observe IDR behavior in different solution conditions and compare solution sensitivity across IDR sequences.

In the upcoming chapters, I will utilize SolSpace simulations to analyze over 170 sequences across seven different solution conditions. This extensive dataset will enable us to explore the correlation and relationship between IDR sequences, their ensemble structures, and their environmental responses. Chapter 2 focuses on using this simulation data to characterize IDR solution response, specifically how IDR dimensions (R_g/R_{ee}) vary with changes in solution conditions. Chapter 3 will delve into the connection between solution sensitivity and IDR entropic force strength, again using SolSpace simulations for analysis. Finally, in Chapter 4, the emphasis will be on leveraging these insights from SolSpace simulations to advance the design of IDR biosensors.

References

(1) Ward, J. J.; Sodhi, J. S.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.* **2004**, *337* (3), 635–645.

(2) Fukuchi, S.; Hosoda, K.; Homma, K.; Gojobori, T.; Nishikawa, K. Binary Classification of Protein Molecules into Intrinsically Disordered and Ordered Segments. *BMC Struct. Biol.* **2011**, *11*, 29.

(3) Piovesan, D.; Necci, M.; Escobedo, N.; Monzon, A. M.; Hatos, A.; Mičetić, I.; Quaglia, F.; Paladin, L.; Ramasamy, P.; Dosztányi, Z.; Vranken, W. F.; Davey, N. E.; Parisi, G.; Fuxreiter, M.; Tosatto, S. C. E. MobiDB: Intrinsically Disordered Proteins in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D361–D367.

(4) Uversky, V. N.; Dunker, A. K. Understanding Protein Non-Folding. *Biochim. Biophys. Acta* **2010**, *1804* (6), 1231–1264.

(5) Yu, F.; Sukenik, S. Structural Preferences Shape the Entropic Force of Disordered Protein Ensembles. *bioRxiv* **2023**. https://doi.org/10.1101/2023.01.20.524980.

(6) van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; Kim, P. M.; Kriwacki, R. W.; Oldfield, C. J.; Pappu, R. V.; Tompa, P.; Uversky, V. N.; Wright, P. E.; Babu, M. M. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **2014**, *114* (13), 6589–6631.

(7) Dosztányi, Z.; Csizmok, V.; Tompa, P.; Simon, I. IUPred: Web Server for the Prediction of Intrinsically Unstructured Regions of Proteins Based on Estimated Energy Content. *Bioinformatics* **2005**, *21* (16), 3433–3434.

(8) Romero, P.; Obradovic, Z.; Kissinger, C. R.; Villafranca, J. E.; Garner, E.; Guilliot, S.; Dunker, A. K. Thousands of Proteins Likely to Have Long Disordered Regions. *Pac. Symp. Biocomput.* **1998**, 437–448.

(9) Wright, P. E.; Dyson, H. J. Intrinsically Unstructured Proteins: Re-Assessing the Protein Structure-Function Paradigm. *J. Mol. Biol.* **1999**, *293* (2), 321–331.

(10) Uversky, V. N.; Gillespie, J. R.; Fink, A. L. Why Are ?natively Unfolded? Proteins Unstructured under Physiologic Conditions? *Proteins* **2000**, *41* (3), 415–427.

(11) Kriwacki, R. W.; Hengst, L.; Tennant, L.; Reed, S. I.; Wright, P. E. Structural Studies of p21Waf1/Cip1/Sdi1 in the Free and Cdk2-Bound State: Conformational Disorder Mediates Binding Diversity. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93* (21), 11504–11509.

(12) Sickmeier, M.; Hamilton, J. A.; LeGall, T.; Vacic, V.; Cortese, M. S.; Tantos, A.; Szabo, B.; Tompa, P.; Chen, J.; Uversky, V. N.; Obradovic, Z.; Dunker, A. K. DisProt: The Database of Disordered Proteins. *Nucleic Acids Res.* **2007**, *35* (Database issue), D786–D793.

(13) Piovesan, D.; Tabaro, F.; Mičetić, I.; Necci, M.; Quaglia, F.; Oldfield, C. J.; Aspromonte, M. C.; Davey, N. E.; Davidović, R.; Dosztányi, Z.; Elofsson, A.; Gasparini, A.; Hatos, A.; Kajava, A. V.; Kalmar, L.; Leonardi, E.; Lazar, T.; Macedo-Ribeiro, S.; Macossay-Castillo, M.; Meszaros, A.; Minervini, G.; Murvai, N.; Pujols, J.; Roche, D. B.; Salladini, E.; Schad, E.; Schramm, A.; Szabo, B.; Tantos, A.; Tonello, F.; Tsirigos, K. D.; Veljković, N.; Ventura, S.; Vranken, W.; Warholm, P.; Uversky, V. N.; Dunker, A. K.; Longhi, S.; Tompa, P.; Tosatto, S. C. E. DisProt 7.0: A Major Update of the Database of Disordered Proteins. *Nucleic Acids Res.* **2017**, *45* (D1), D219–D227.

(14) Tompa, P. Intrinsically Disordered Proteins: A 10-Year Recap. *Trends Biochem. Sci.*

2012, *37* (12), 509–516.
(15) Malaney, P.; Pathak, R. R.; Xue, B.; Uversky, V. N.; Davé, V. Intrinsic Disorder in PTEN and Its Interactome Confers Structural Plasticity and Functional Versatility. *Sci. Rep.* **2013**, *3*, 2035.
(16) Holmstrom, E. D.; Nettels, D.; Schuler, B. Conformational Plasticity of Hepatitis C Virus Core Protein Enables RNA-Induced Formation of Nucleocapsid-like Particles. *J. Mol. Biol.* **2018**, *430* (16), 2453–2467.
(17) Uversky, V. N. Intrinsically Disordered Proteins and Their Environment: Effects of Strong Denaturants, Temperature, pH, Counter Ions, Membranes, Binding Partners, Osmolytes, and Macromolecular Crowding. *Protein J.* **2009**, *28* (7-8), 305–325.
(18) Rogers, J. M.; Steward, A.; Clarke, J. Folding and Binding of an Intrinsically Disordered Protein: Fast, but Not "Diffusion-Limited." *J. Am. Chem. Soc.* **2013**, *135* (4), 1415–1422.
(19) Dunker, A. K.; Cortese, M. S.; Romero, P.; Iakoucheva, L. M.; Uversky, V. N. Flexible Nets. The Roles of Intrinsic Disorder in Protein Interaction Networks. *FEBS J.* **2005**, *272* (20), 5129–5148.
(20) Patil, A.; Nakamura, H. Disordered Domains and High Surface Charge Confer Hubs with the Ability to Interact with Multiple Proteins in Interaction Networks. *FEBS Lett.* **2006**, *580* (8), 2041–2045.
(21) Singh, G. P.; Ganapathi, M.; Dash, D. Role of Intrinsic Disorder in Transient Interactions of Hub Proteins. *Proteins* **2007**, *66* (4), 761–765.
(22) Bertolazzi, P.; Bock, M. E.; Guerra, C. On the Functional and Structural Characterization of Hubs in Protein-Protein Interaction Networks. *Biotechnol. Adv.* **2013**, *31* (2), 274–286.
(23) Masoud, G. N.; Li, W. HIF-1α Pathway: Role, Regulation and Intervention for Cancer Therapy. *Acta Pharm Sin B* **2015**, *5* (5), 378–389.
(24) Erler, J. T.; Cawthorne, C. J.; Williams, K. J.; Koritzinsky, M.; Wouters, B. G.; Wilson, C.; Miller, C.; Demonacos, C.; Stratford, I. J.; Dive, C. Hypoxia-Mediated down-Regulation of Bid and Bax in Tumors Occurs via Hypoxia-Inducible Factor 1-Dependent and -Independent Mechanisms and Contributes to Drug Resistance. *Mol. Cell. Biol.* **2004**, *24* (7), 2875–2889.
(25) Kilic, M.; Kasperczyk, H.; Fulda, S.; Debatin, K.-M. Role of Hypoxia Inducible Factor-1 Alpha in Modulation of Apoptosis Resistance. *Oncogene* **2007**, *26* (14), 2027–2038.
(26) Krois, A. S.; Dyson, H. J.; Wright, P. E. Long-Range Regulation of p53 DNA Binding by Its Intrinsically Disordered N-Terminal Transactivation Domain. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (48), E11302–E11310.
(27) Merchant, K. A.; Best, R. B.; Louis, J. M.; Gopich, I. V.; Eaton, W. A. Characterizing the Unfolded States of Proteins Using Single-Molecule FRET Spectroscopy and Molecular Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (5), 1528–1533.
(28) Nath, A.; Sammalkorpi, M.; DeWitt, D. C.; Trexler, A. J.; Elbaum-Garfinkle, S.; O'Hern, C. S.; Rhoades, E. The Conformational Ensembles of α-Synuclein and Tau: Combining Single-Molecule FRET and Simulations. *Biophys. J.* **2012**, *103* (9), 1940–1949.
(29) Henriques, J.; Arleth, L.; Lindorff-Larsen, K.; Skepö, M. On the Calculation of SAXS Profiles of Folded and Intrinsically Disordered Proteins from Computer Simulations. *J. Mol. Biol.* **2018**, *430* (16), 2521–2539.
(30) Moses, D.; Guadalupe, K.; Yu, F.; Flores, E.; Perez, A.; McAnelly, R.; Shamoon, N. M.; Cuevas-Zepeda, E.; Merg, A. D.; Martin, E. W.; Holehouse, A. S.; Sukenik, S.

Structural Biases in Disordered Proteins Are Prevalent in the Cell. *bioRxiv*, 2022, 2021.11.24.469609. https://doi.org/10.1101/2021.11.24.469609.
(31) Orengo, C. A.; Todd, A. E.; Thornton, J. M. From Protein Structure to Function. *Curr. Opin. Struct. Biol.* **1999**, *9* (3), 374–382.
(32) González-Foutel, N. S.; Glavina, J.; Borcherds, W. M.; Safranchik, M.; Barrera-Vilarmau, S.; Sagar, A.; Estaña, A.; Barozet, A.; Garrone, N. A.; Fernandez-Ballester, G.; Blanes-Mira, C.; Sánchez, I. E.; de Prat-Gay, G.; Cortés, J.; Bernadó, P.; Pappu, R. V.; Holehouse, A. S.; Daughdrill, G. W.; Chemes, L. B. Conformational Buffering Underlies Functional Selection in Intrinsically Disordered Protein Regions. *Nat. Struct. Mol. Biol.* **2022**, *29* (8), 781–790.
(33) Espinoza-Fonseca, L. M. Reconciling Binding Mechanisms of Intrinsically Disordered Proteins. *Biochem. Biophys. Res. Commun.* **2009**, *382* (3), 479–482.
(34) Dogan, J.; Gianni, S.; Jemth, P. The Binding Mechanisms of Intrinsically Disordered Proteins. *Phys. Chem. Chem. Phys.* **2014**, *16* (14), 6323–6331.
(35) Ješmantavičius, V.; Dogan, J.; Jemth, P. Helical Propensity in an Intrinsically Disordered Protein Accelerates Ligand Binding. *Angew. Chem. Int. Ed Engl.* **2014**.
(36) Arai, M.; Sugase, K.; Dyson, H. J.; Wright, P. E. Conformational Propensities of Intrinsically Disordered Proteins Influence the Mechanism of Binding and Folding. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (31), 9614–9619.
(37) Vacic, V.; Markwick, P. R. L.; Oldfield, C. J.; Zhao, X.; Haynes, C.; Uversky, V. N.; Iakoucheva, L. M. Disease-Associated Mutations Disrupt Functionally Important Regions of Intrinsic Protein Disorder. *PLoS Comput. Biol.* **2012**, *8* (10), e1002709.
(38) Wicky, B. I. M.; Shammas, S. L.; Clarke, J. Affinity of IDPs to Their Targets Is Modulated by Ion-Specific Changes in Kinetics and Residual Structure. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (37), 9882–9887.
(39) McGuffee, S. R.; Elcock, A. H. Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS Comput. Biol.* **2010**, *6* (3), e1000694.
(40) Yancey, P. H. Organic Osmolytes as Compatible, Metabolic and Counteracting Cytoprotectants in High Osmolarity and Other Stresses. *J. Exp. Biol.* **2005**, *208* (Pt 15), 2819–2830.
(41) Smoyer, C. J.; Jaspersen, S. L. Breaking down the Wall: The Nuclear Envelope during Mitosis. *Curr. Opin. Cell Biol.* **2014**, *26*, 1–9.
(42) Record, M. T.; Thomas Record, M.; Zhang, W.; Anderson, C. F. Analysis of Effects of Salts and Uncharged Solutes on Protein and Nucleic Acid Equilibria and Processes: A Practical Guide to Recognizing and Interpreting Polyelectrolyte Effects, Hofmeister Effects, and Osmotic Effects of Salts. *Advances in Protein Chemistry*. 1998, pp 281–353. https://doi.org/10.1016/s0065-3233(08)60655-5.
(43) Zhang, Y.; Kitazawa, S.; Peran, I.; Stenzoski, N.; McCallum, S. A.; Raleigh, D. P.; Royer, C. A. High Pressure ZZ-Exchange NMR Reveals Key Features of Protein Folding Transition States. *J. Am. Chem. Soc.* **2016**, *138* (46), 15260–15266.
(44) Zhou, H.-X.; Rivas, G.; Minton, A. P. Macromolecular Crowding and Confinement: Biochemical, Biophysical, and Potential Physiological Consequences. *Annu. Rev. Biophys.* **2008**, *37*, 375–397.
(45) Ebbinghaus, S.; Dhar, A.; McDonald, J. D.; Gruebele, M. Protein Folding Stability and Dynamics Imaged in a Living Cell. *Nat. Methods* **2010**, *7* (4), 319–323.
(46) Monteith, W. B.; Pielak, G. J. Residue Level Quantification of Protein Stability in Living Cells. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (31), 11335–11340.
(47) Koltai, T. The Ph Paradigm in Cancer. *Eur. J. Clin. Nutr.* **2020**, *74* (Suppl 1), 14–19.

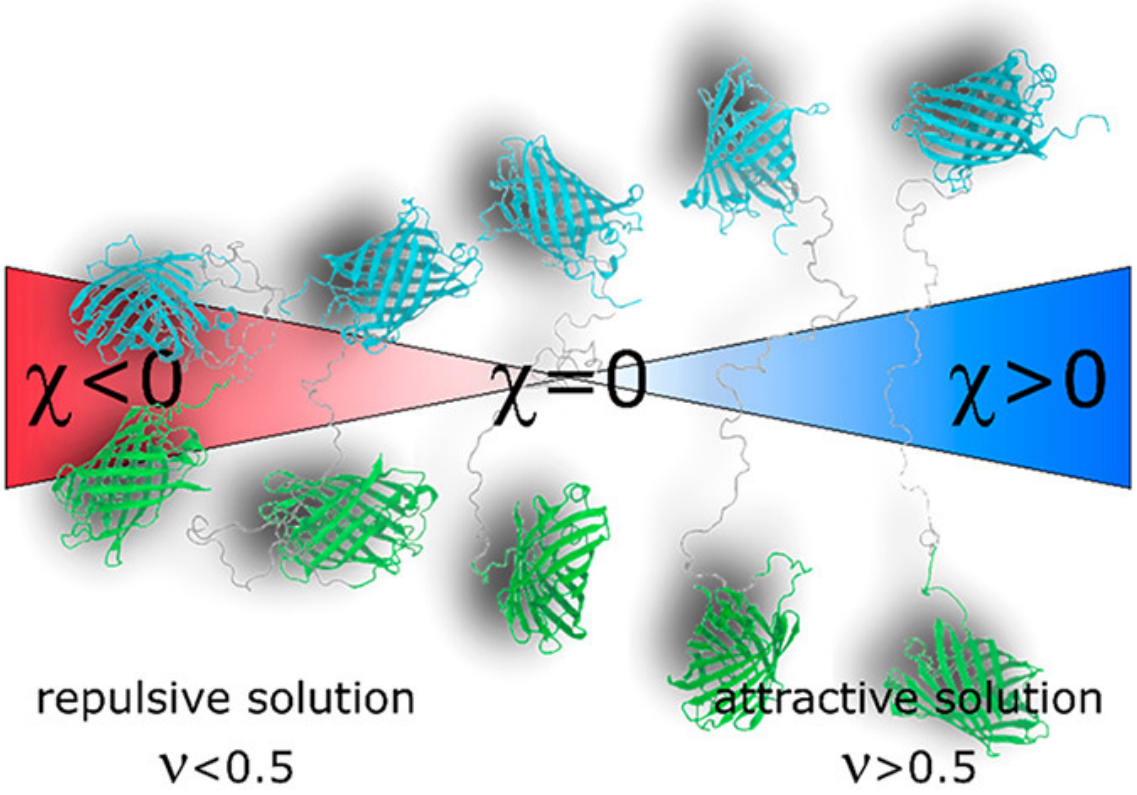
(48) Bensinger, S. J.; Christofk, H. R. New Aspects of the Warburg Effect in Cancer Cell Biology. *Semin. Cell Dev. Biol.* **2012**, *23* (4), 352–361.
(49) Spear, J. S.; White, K. A. Single-Cell Intracellular pH Dynamics Regulate the Cell Cycle by Timing the G1 Exit and G2 Transition. *J. Cell Sci.* **2023**, *136* (10). https://doi.org/10.1242/jcs.260458.
(50) Schwartz, L.; Peres, S.; Jolicoeur, M.; da Veiga Moreira, J. Cancer and Alzheimer's Disease: Intracellular pH Scales the Metabolic Disorders. *Biogerontology* **2020**, *21* (6), 683–694.
(51) Banks, A.; Qin, S.; Weiss, K. L.; Stanley, C. B.; Zhou, H.-X. Intrinsically Disordered Protein Exhibits Both Compaction and Expansion under Macromolecular Crowding. *Biophys. J.* **2018**, *114* (5), 1067–1079.
(52) Babu, M. M.; van der Lee, R.; de Groot, N. S.; Gsponer, J. Intrinsically Disordered Proteins: Regulation and Disease. *Curr. Opin. Struct. Biol.* **2011**, *21* (3), 432–440.
(53) Mansouri, A. L.; Grese, L. N.; Rowe, E. L.; Pino, J. C.; Chennubhotla, S. C.; Ramanathan, A.; O'Neill, H. M.; Berthelier, V.; Stanley, C. B. Folding Propensity of Intrinsically Disordered Proteins by Osmotic Stress. *Mol. Biosyst.* **2016**, *12* (12), 3695–3701.
(54) Moses, D.; Yu, F.; Ginell, G. M.; Shamoon, N. M.; Koenig, P. S.; Holehouse, A. S.; Sukenik, S. Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to Their Chemical Environment. *J. Phys. Chem. Lett.* **2020**, 10131–10136.
(55) Holehouse, A. S.; Sukenik, S. Controlling Structural Bias in Intrinsically Disordered Proteins Using Solution Space Scanning. *J. Chem. Theory Comput.* **2020**, *16* (3), 1794–1805.
(56) Borgia, A.; Zheng, W.; Buholzer, K.; Borgia, M. B.; Schüler, A.; Hofmann, H.; Soranno, A.; Nettels, D.; Gast, K.; Grishaev, A.; Best, R. B.; Schuler, B. Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *J. Am. Chem. Soc.* **2016**, *138* (36), 11714–11726.
(57) Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B. Polymer Scaling Laws of Unfolded and Intrinsically Disordered Proteins Quantified with Single-Molecule Spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (40), 16155–16160.
(58) Kjaergaard, M.; Brander, S.; Poulsen, F. M. Random Coil Chemical Shift for Intrinsically Disordered Proteins: Effects of Temperature and pH. *J. Biomol. NMR* **2011**, *49* (2), 139–149.
(59) Kjaergaard, M.; Nørholm, A.-B.; Hendus-Altenburger, R.; Pedersen, S. F.; Poulsen, F. M.; Kragelund, B. B. Temperature-Dependent Structural Changes in Intrinsically Disordered Proteins: Formation of Alpha-Helices or Loss of Polyproline II? *Protein Sci.* **2010**, *19* (8), 1555–1564.
(60) Müller-Späth, S.; Soranno, A.; Hirschfeld, V.; Hofmann, H.; Rüegger, S.; Reymond, L.; Nettels, D.; Schuler, B. Charge Interactions Can Dominate the Dimensions of Intrinsically Disordered Proteins. *Proceedings of the National Academy of Sciences* **2010**, *107* (33), 14609–14614.
(61) Keul, N. D.; Oruganty, K.; Schaper Bergman, E. T.; Beattie, N. R.; McDonald, W. E.; Kadirvelraj, R.; Gross, M. L.; Phillips, R. S.; Harvey, S. C.; Wood, Z. A. The Entropic Force Generated by Intrinsically Disordered Segments Tunes Protein Function. *Nature* **2018**, *563* (7732), 584–588.
(62) Zeno, W. F.; Thatte, A. S.; Wang, L.; Snead, W. T.; Lafer, E. M.; Stachowiak, J. C. Molecular Mechanisms of Membrane Curvature Sensing by a Disordered Protein. *J. Am. Chem. Soc.* **2019**, *141* (26), 10361–10371.

(63) Borgia, A.; Borgia, M. B.; Bugge, K.; Kissling, V. M.; Heidarsson, P. O.; Fernandes, C. B.; Sottini, A.; Soranno, A.; Buholzer, K. J.; Nettels, D.; Kragelund, B. B.; Best, R. B.; Schuler, B. Extreme Disorder in an Ultrahigh-Affinity Protein Complex. *Nature* **2018**, *555* (7694), 61–66.
(64) Vancraenenbroeck, R.; Harel, Y. S.; Zheng, W.; Hofmann, H. Polymer Effects Modulate Binding Affinities in Disordered Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (39), 19506–19512.
(65) Konrat, R. NMR Contributions to Structural Dynamics Studies of Intrinsically Disordered Proteins. *J. Magn. Reson.* **2014**, *241* (100), 74–85.
(66) Nwanochie, E.; Uversky, V. N. Structure Determination by Single-Particle Cryo-Electron Microscopy: Only the Sky (and Intrinsic Disorder) Is the Limit. *Int. J. Mol. Sci.* **2019**, *20* (17). https://doi.org/10.3390/ijms20174186.
(67) Kikhney, A. G.; Svergun, D. I. A Practical Guide to Small Angle X-Ray Scattering (SAXS) of Flexible and Intrinsically Disordered Proteins. *FEBS Lett.* **2015**, *589* (19 Pt A), 2570–2577.
(68) Salmon, L.; Nodet, G.; Ozenne, V.; Yin, G.; Jensen, M. R.; Zweckstetter, M.; Blackledge, M. NMR Characterization of Long-Range Order in Intrinsically Disordered Proteins. *J. Am. Chem. Soc.* **2010**, *132* (24), 8407–8418.
(69) 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.
(70) Gomes, G.-N. W.; Krzeminski, M.; Namini, A.; Martin, E. W.; Mittag, T.; Head-Gordon, T.; Forman-Kay, J. D.; Gradinaru, C. C. Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *J. Am. Chem. Soc.* **2020**, *142* (37), 15697–15710.
(71) Haas, E. Ensemble FRET Methods in Studies of Intrinsically Disordered Proteins. *Methods Mol. Biol.* **2012**, *895*, 467–498.
(72) Kachala, M.; Valentini, E.; Svergun, D. I. Application of SAXS for the Structural Characterization of IDPs. *Adv. Exp. Med. Biol.* **2015**, *870*, 261–289.
(73) Putnam, C. D.; Hammel, M.; Hura, G. L.; Tainer, J. A. X-Ray Solution Scattering (SAXS) Combined with Crystallography and Computation: Defining Accurate Macromolecular Structures, Conformations and Assemblies in Solution. *Q. Rev. Biophys.* **2007**, *40* (3), 191–285.
(74) Ohashi, T.; Galiacy, S. D.; Briscoe, G.; Erickson, H. P. An Experimental Study of GFP-Based FRET, with Application to Intrinsically Unstructured Proteins. *Protein Sci.* **2007**, *16* (7), 1429–1438.
(75) Kim, S. J.; Dumont, C.; Gruebele, M. Simulation-Based Fitting of Protein-Protein Interaction Potentials to SAXS Experiments. *Biophys. J.* **2008**, *94* (12), 4924–4931.
(76) Paissoni, C.; Jussupow, A.; Camilloni, C. Determination of Protein Structural Ensembles by Hybrid-Resolution SAXS Restrained Molecular Dynamics. *J. Chem. Theory Comput.* **2020**, *16* (4), 2825–2834.
(77) Cumberworth, A.; Bui, J. M.; Gsponer, J. Free Energies of Solvation in the Context of Protein Folding: Implications for Implicit and Explicit Solvent Models. *J. Comput. Chem.* **2016**, *37* (7), 629–640.
(78) Warner, J. B., 4th; Ruff, K. M.; Tan, P. S.; Lemke, E. A.; Pappu, R. V.; Lashuel, H. A. Monomeric Huntingtin Exon 1 Has Similar Overall Structural Features for Wild-Type and Pathological Polyglutamine Lengths. *J. Am. Chem. Soc.* **2017**, *139* (41), 14456–14469.
(79) Mittal, A.; Holehouse, A. S.; Cohan, M. C.; Pappu, R. V. Sequence-to-Conformation

Relationships of Disordered Regions Tethered to Folded Domains of Proteins. *J. Mol. Biol.* **2018**, *430* (16), 2403–2421.
(80) Metropolis, N.; Rosenbluth, A. W. Equation of State Calculations by Fast Computing Machines. *The journal of* **1953**.
(81) Vitalis, A.; Pappu, R. V. Chapter 3 Methods for Monte Carlo Simulations of Biomacromolecules. In *Annual Reports in Computational Chemistry*; Wheeler, R. A., Ed.; Elsevier, 2009; Vol. 5, pp 49–76.
(82) Vitalis, A.; Pappu, R. V. ABSINTH: A New Continuum Solvation Model for Simulations of Polypeptides in Aqueous Solutions. *J. Comput. Chem.* **2009**, *30* (5), 673–699.
(83) Huang, J.; MacKerell, A. D., Jr. Force Field Development and Simulations of Intrinsically Disordered Proteins. *Curr. Opin. Struct. Biol.* **2018**, *48*, 40–48.
(84) Durell, S. R.; Brooks, B. R.; Ben-Naim, A. Solvent-Induced Forces between Two Hydrophilic Groups. *J. Phys. Chem.* **1994**, *98* (8), 2198–2202.
(85) Onufriev, A. Chapter 7 - Implicit Solvent Models in Molecular Dynamics Simulations: A Brief Overview. In *Annual Reports in Computational Chemistry*; Wheeler, R. A., Spellmeyer, D. C., Eds.; Elsevier, 2008; Vol. 4, pp 125–137.
(86) Auton, M.; Bolen, D. W. Additive Transfer Free Energies of the Peptide Backbone Unit That Are Independent of the Model Compound and the Choice of Concentration Scale. *Biochemistry* **2004**, *43* (5), 1329–1342.
(87) Auton, M.; Bolen, D. W. Predicting the Energetics of Osmolyte-Induced Protein Folding/unfolding. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (42), 15065–15068.

Chapter 2: Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to their Chemical Environment

Adapted with permission granted by the American Chemical Society. The material originally appeared in the following: David Moses*, Feng Yu*, Garrett M. Ginell, Nora M. Shamoon, Patrick S. Koenig, Alex S. Holehouse, and Shahar Sukenik (2020) J. Phys. Chem. Lett. 11: 10131-10136.



Abstract

Intrinsically disordered protein regions (IDRs) make up roughly 30% of the human proteome and are central to a wide range of biological processes. Given a lack of persistent tertiary structure, all residues in IDRs are, to some extent, solvent-exposed. This extensive surface area, coupled with the absence of strong intramolecular contacts, makes IDRs inherently sensitive to their chemical environment. We report a combined experimental, computational, and analytical framework for high-throughput characterization of IDR sensitivity. Our framework reveals that IDRs can expand or compact in response to changes in their solution environment. Importantly, the direction and magnitude of conformational change depend on both protein sequence and cosolute identity. For example, some solutes such as short polyethylene glycol chains exert an expanding effect on some IDRs and a compacting effect on others. Despite this complex behavior, we can rationally interpret IDR responsiveness to solution composition changes using relatively simple polymer models. Our results imply that solution-responsive IDRs are ubiquitous and can provide an additional layer of regulation to biological systems.

Introduction

Intrinsically disordered proteins and protein regions (IDRs) play key roles in mediating cellular signaling, transcriptional regulation, and homeostatic functions[1]. IDRs differ from well-folded proteins in that they exist in an ensemble of rapidly changing configurations (**Fig. 2.1A**). This conformational ensemble is often tied to IDR function[2–4]. IDR ensembles have extensive surface area exposed to the surrounding solution, and few non-covalent intramolecular bonds that constrain their structure. As such, IDR ensembles are highly malleable and can be strongly affected by the chemistry of their surrounding environment[5]. Inside the cell, the chemical composition can change due to routine cell-cycle events or external stress[6–10]. The plasticity of IDR ensembles makes them ideal sensors and actuators of these changes[11–13], but perhaps also impairs their activity in deleterious environments such as metabolically rewired cancer cells[14]. Still, little effort has been directed at systematically characterizing IDR sensitivity to solution changes.

The effects of solution chemical changes on protein structure can be likened to a "tug-of-war" between intra-protein interactions and interactions between protein moieties and the surrounding solution. This tug-of-war is a balance that can be shifted by changes to sequence (mutations or post-translational modifications)[15,16], but also by changes in the physical-chemical composition of the intracellular environment[9,14]. While the sensitivity of IDRs to solution composition has been discussed[13,17,18], it has not been systematically characterized. Here we set out to systematically evaluate the sensitivity of IDRs to changes in their surrounding environment.

Methods

For the discussion about the experimental and computational methods of this section, please refer to **Appendix A**.

Results and Discussion

A high-throughput approach to reveal IDR dimensions using ensemble FRET

We use "solution space" scanning to characterize IDR sensitivity to solution composition changes. This is analogous to "sequence space" scanning, but uses different chemical environments instead of sequence mutations to probe protein behavior. To scan IDRs in solution space at high throughput, we developed a protocol that leverages ensemble FRET to report on changes in the average distance between their termini. We use a protein construct comprising an IDR of interest sandwiched between two fluorescent proteins (FPs) that together form a Förster resonance energy transfer (FRET) pair (**Fig. 2.1A**). The FPs selected were mTurquoise2[19] (donor) and mNeonGreen[20] (acceptor)[21,22]. We chose four IDRs whose ensembles play functional roles: the 61-residue N-terminal transactivation domain of p53 (p53)[4]; the 34-residue BH3 domain of apoptosis regulating protein PUMA (PUMA)[3]; the 83-residue C-terminal domain of the yeast transcription factor Ash1 (Ash1)[23]; the 40-residue N-terminal domain of the adenoviral hub protein

E1A (E1A)[24]. For each of these constructs the FRET efficiency, E_f, was determined as described in **Appendix A1.5**.

To derive changes in IDR dimensions from E_f we began by measuring a series of Gly-Ser (GS) repeats in our FRET backbone to generate a length-dependent point of reference. E_f for these constructs scales linearly with GS repeats as expected (**Fig. 2.1B**, and see **Appendix A1.6**), allowing us to interpolate E_f for a GS linker of a given length to create a ratio χ:

$$\chi = \frac{R_e^i}{R_e^{GS}} - 1 \tag{2.1}$$

where R_e is the end-to-end distance between donor and acceptor FPs obtained from E_f as described in **Appendix A1.7**, and the superscript i or GS refers to a specific IDR sequence or a GS linker of the equivalent length, respectively. Thus, a negative χ value indicates the chain is more compact, while a positive value indicates it is more expanded, than a GS linker of the equivalent length. Conveniently, χ allows us to plot IDRs of different lengths on the same axes. Our calculations of χ in neat buffer (i.e. in buffer without additional co-solutes) for different FRET constructs reveal a range of behaviors, with Ash1 and p53 having more expanded, and PUMA and E1A more compact, ensembles (**Fig. 2.1C**).

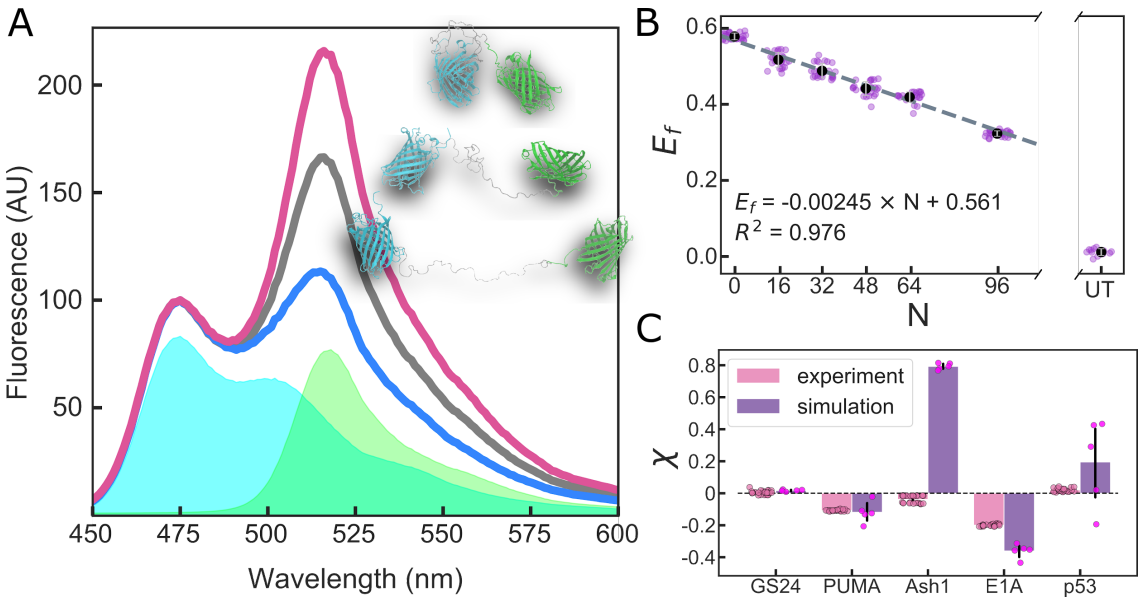


Figure 2.1. (**A**) Fluorescence spectra normalized to donor peak intensity of a FRET construct in compacting (red), buffer (black), and expanding (blue) solutions. Cyan and green areas are base spectra of donor and acceptor FPs, respectively. Inset shows single configurations for various degrees of expansion. (**B**) FRET efficiency of Gly-Ser repeat linkers vs. number of residues (*N*) in a buffer solution. UT is a solution of untethered, equimolar donor and acceptor. Dashed line shows linear fit of the data. (**C**) Calculated χ for FRET constructs in buffer determined by experiment (average of four repeats with 6 replicates each) and simulation (average of five repeats). Error bars are SD of all replicates/repeats.

IDR ensemble dimensions are sensitive to solution composition and protein sequence, but not to length

We next investigate how IDR dimensions change in different chemical environments. The solutions we use are *not* representative of the cellular environment. Instead, solutions containing osmolytes, polymeric crowders, polyols, free amino acids, denaturants and salts probe IDR structure by “pushing” or “pulling” against the attractions or repulsions of intra-protein interactions. We calculated χ for each combination of IDR/solution as described in **Appendix A1.7**. The resulting changes in χ reveal a distinctive solution-space "fingerprint" for each IDR (**Fig. 2.2A**) and highlight that different sequences have different sensitivities to the same solute[25,26]. This is in sharp contrast to the sensitivity of GS linkers, which all display a similar fingerprint regardless of length (**Fig. A1**).

Focusing on several solute archetypes reveals interesting trends (**Fig. 2.2B**). Short polyethylene glycol (PEG) chains, such as PEG200, display disparate effects on different sequences, causing only Ash1 to compact, and the rest to expand, in line with other observations[27]. Larger polymers such as PEG2000 and Ficoll appear to compact the dimensions of all IDRs as shown for other disordered proteins[28], with a sequence-dependent magnitude that is stronger for Ash1 and PUMA as previously reported[29]. Smaller solutes like sarcosine and tricine also reveal a linear expanding or compacting effect, but show that different proteins expand or compact by different magnitudes under the same solution. Salts like NaCl display a characteristic non-monotonic effect, as described previously[8,30,31]. In ionic solutes, the initial expansion likely stems from screening of attractive electrostatic interactions that may in fact arise not only from the IDR chain but also from the FP tags, as indicated by the effect on uncharged GS linkers (**Fig. A1**), while the compaction trend stems from specific ion effects, and differs between protein types[32]. Overall, the picture that emerges is that different solution environments affect IDRs in a way that strongly depends on sequence composition and arrangement, but much less on length. The full dataset is available in **Tables S1-S2** mentioned in **Appendix A**.

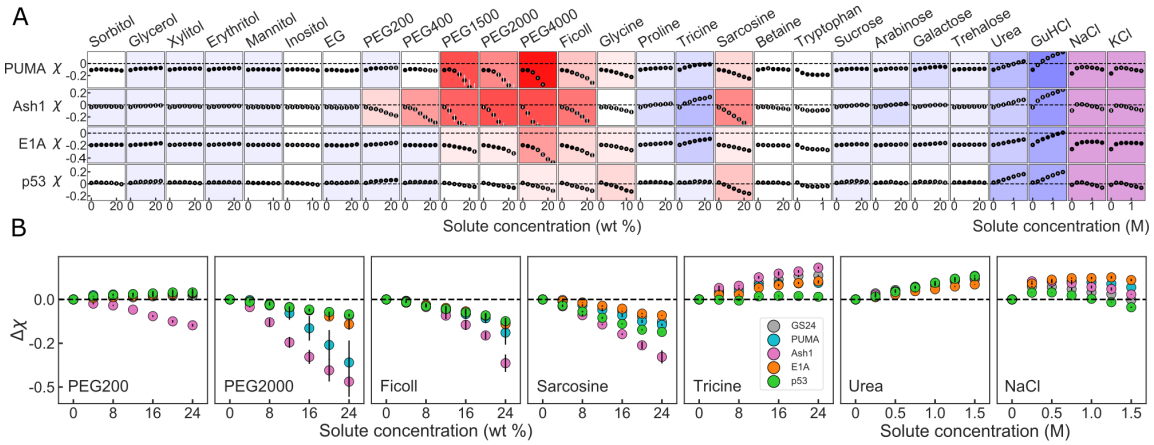


Figure 2.2. (**A**) Solution space scans of IDRs. Each data point shows the average χ vs. concentration of a specific solute for each protein taken from two repeats. Vertical grey bars show spread of data, and are often too small to see. Proteins vary down columns, and solutes across rows. Background color represents the sensitivity of change to solute addition: stronger colors imply higher sensitivity, red hues indicate compaction, and blue hues indicate expansion. Purple background indicates non-monotonic behavior. (**B**) Differential response of IDRs to individual solutes. Each panel point shows $\Delta\chi = \chi([\text{solute}]) - \chi([\text{solute}] = 0)$ vs. concentration from two repeats of a specific solute for several different constructs. Vertical lines are the spread of the data.

IDR dimensions in neat buffer predict sensitivity to solution changes

To see how sensitivities play out in a larger range of IDRs, we turn to all-atom simulations. We use the ABSINTH forcefield that has previously been shown to reproduce experimentally measured IDR ensembles (see **Appendix A Section 2.1**)[23,33–35]. To maintain connection with experiments, we start by simulating GS linkers of various lengths (**Fig. A2**), and use ensemble-averaged R_e to calculate χ for simulated IDRs according to **Eq. 2.1**. The simulation-derived χ for the four different proteins used in our experiments qualitatively agrees with our FRET experiments, aside from Ash1 which is significantly more expanded in simulations than our FRET experiments show (**Fig. 2.1C**). It is important to note that the absence of FPs in these simulations dictates that the value of χ is necessarily different between experiment and simulations, and a quantitative match is not expected.

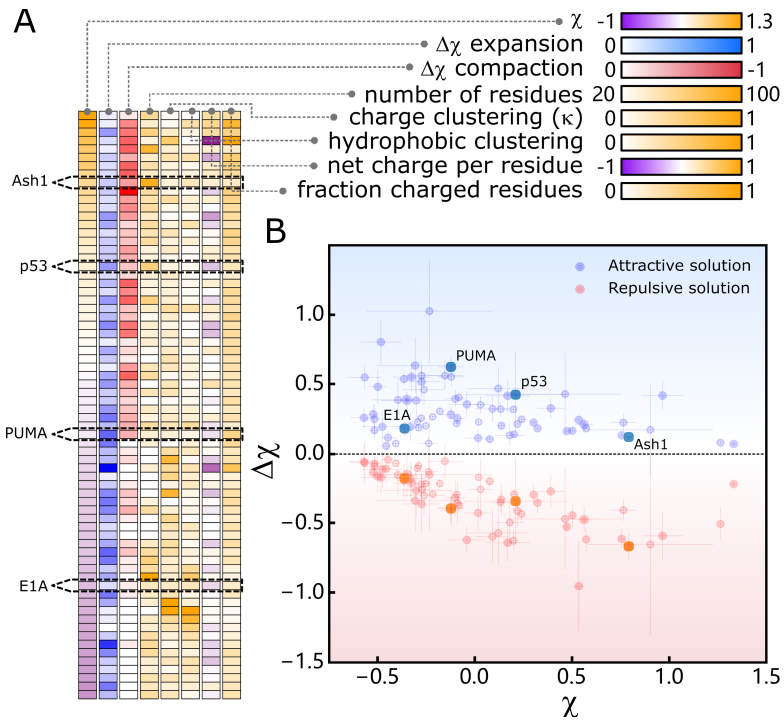


Figure 2.3. All-atom simulations of IDR sensitivity to solutions. (**A**) Heatmap of protein sensitivity and molecular features. Protein identity varies from top to bottom across cells, and molecular features vary left to right. Colormaps are shown for each molecular feature. (**B**) The magnitude $\Delta\chi$ in attractive (blue) or repulsive (red) solutions as a function of χ in aqueous solution for each protein in (A). Darker points represent proteins shows in **Fig. 2.1C.** Error bars calculated from SD of 5 repeats. 70 points are plotted in each condition. All data available in **Table S3**.

We next wanted to see how other naturally occurring disordered sequences would respond to different solution conditions. We have previously designed and calibrated an approach to perform computational solution space scanning with ABSINTH[12]. We selected 70 experimentally identified IDRs[36], and used our computational solution space scanning approach to change interactions between the solvent and the backbone of these proteins, akin to the effect of osmolytes and denaturants[37–39] (data in **Table S3).** We quantified the sensitivity of the protein to compacting or expanding solutions based on the extent of change in χ (**Fig. A3**). The dataset is sorted from compact to expanded (negative to positive χ) in **Fig. 2.3A** and shows little correlation with many sequence-based parameters, but relatively strong correlation with the change $\Delta\chi = \chi([\text{solute}]) - \chi([\text{solute}] = 0)$ in solutions that cause the sequence to compact (repulsive solutions) or expand (attractive solutions). We refer to this value as the "solution sensitivity" of the protein. We plot the solution sensitivity ,$\Delta\chi$, vs. protein dimensions in buffer, χ, in **Fig. 2.3B**. As expected from **Fig. 2.3A**, sequences with a negative χ have a larger tendency to expand, but a limited ability to compact, and vice versa for positive χ. Remarkably, both compaction and expansion show the same dependence on χ, even at different solution interaction strengths (see **Fig. A4**).

Predicting the extent of solution sensitivity in intrinsically disordered chains

To see if the non-monotonic trend shown in **Fig. 2.3B** can be generalized, we measured solution-induced expansion and compaction in a lattice-based heteropolymer model detailed in **Appendix A Section 2.2**.[23,35]. We simulated a total of 10^4 sequences with lengths ranging from 20 to 100 residues in 11 solution conditions, and quantified χ and $\Delta\chi$ for each sequence/solution pair (**Fig. A5**). The trends from all-atom simulations, re-drawn as a density map in **Fig. 2.4A,** match the coarse-grained simulations shown in **Fig. 2.4B**. A non-monotonic change in $\Delta\chi$ is observed, with the inflection point centered approximately around χ = 0.0 and a 'dead zone' in the center of the plot. For naturally expanded chains ($\chi \rightarrow 0.4$) solution sensitivity is minimized, while for naturally compact chains ($\chi \rightarrow -0.4$) a broad distribution of sensitivity is observed with respect to expansion, while sensitivity through compaction trends to zero.

Based on these results, we developed an analytical homopolymer model to relate changes in chain-solvent interaction to chain dimensions (see **Appendix A Section 2.3**). Using this model we generated chains with a specific χ value in buffer and perturbed the chain-solvent interactions, and directly calculated $\Delta\chi$ (**Fig. 2.4C, Fig. A6**). Despite being a simplified homopolymer model, our analytical expression revealed the same phenomenological pattern as obtained in our all-atom and coarse-grained simulations.

The χ-dependence of the chain-solvent interaction strength is shown in **Fig. 2.4D** (black line), which reveals that $\Delta\chi$ depends on both the strength of the change in chain-solvent interaction and the χ value in an aqueous solution. Our model offers direct physical intuition as to the origin of the complex relationship between χ and Δχ. Perhaps most importantly, it implies that while expanded or compact proteins display a wide range of sensitivities, IDRs where $\chi \sim 0$ display a basal sensitivity to solution interactions. In this region, where most IDRs fall[40], even small changes to solution composition are predicted to have a measurable effect on IDR dimensions and/or residual structure.[12]

Under the assumption that cosolute-protein interactions scale linearly[25,41], we globally fit our experimental data onto our analytical model leveraging the fact that all experimental measurements start in the same neat buffer (**Fig. 2.4E**, **Fig. A7**). All solution perturbations can be rationally interpreted as driving sequence-dependent shifts along the coil-to-globule transition, in which the magnitude of the shift maps directly to modulation of chain-solvent interactions. The scaling factors required for this mapping qualitatively mirror known co-solute interaction coefficients, and reveal quantitative sequence-dependent differences in the solution response (**Fig. A8**). Chain dimensions can also be represented using an apparent scaling exponent (ν^{app}) (see **Appendix A Section 2.4**)[42,43]. The solvent-induced changes observed are substantial, and for many solutes drive changes equal to or greater than changes observed in IDRs due to mutagenesis[35].

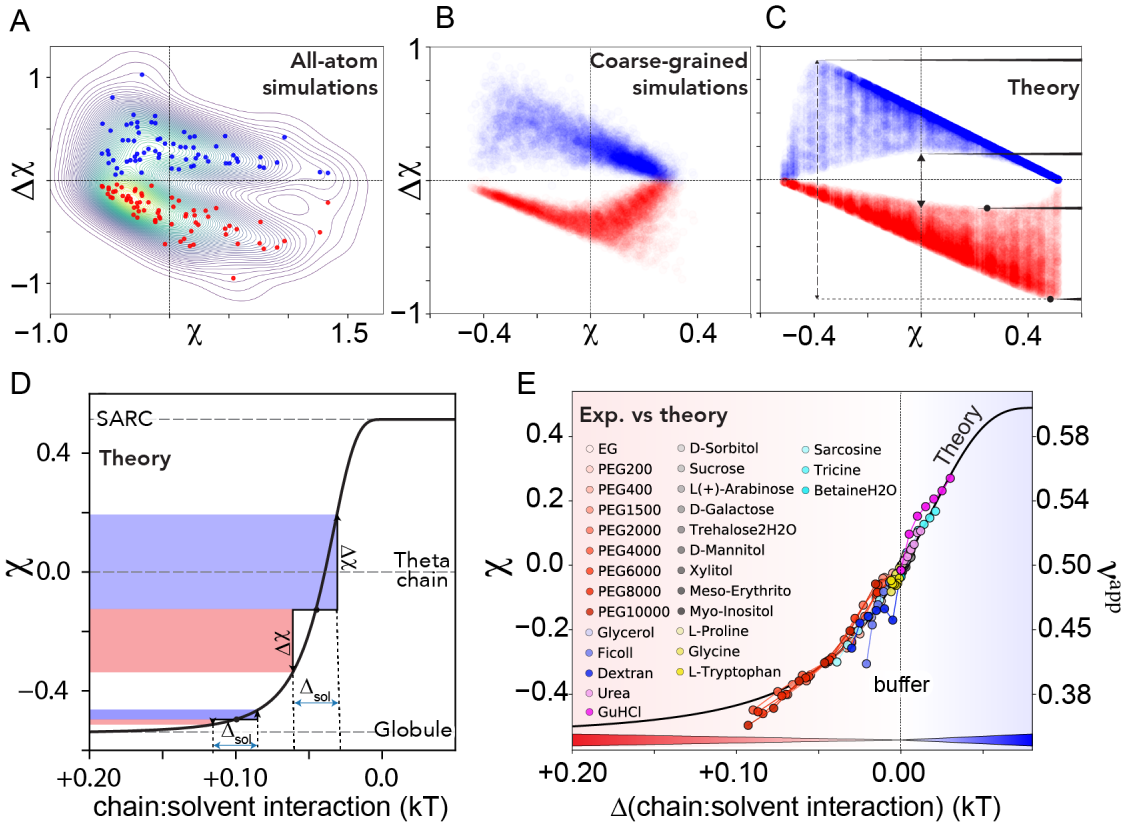


Fig. 2.4 (A-C). Density maps of all-atom simulations shown in **Fig, 2.3B** (**A**), PIMMS coarse-grained simulations **(B)**, and an analytical model (**C**) for solution sensitivity $\Delta\chi$ vs dimensions in the aqueous buffer χ. **(D)** Coil-to-globule transition obtained from an analytical model (SARC = self-avoiding random coil). $\Delta\chi$ is measured as the height of the blue (contraction) or red (expansion) shaded regions. When the same chain-solvent perturbation (Δ_{sol}) is applied to a 100-residue chain with different starting χ values, very different $\Delta\chi$ are expected. **(E)** Projection of experimental data for Ash1 onto the analytical model from (**D**), with solute concentrations scaled to the change in mean-field chain-solvent interaction as compared with neat buffer. The X axis here represents the same units as in panel D but reports on the change in chain:solvent interaction relative to aqueous solvent, which is set to 0. Chain dimensions are also shown by their apparent scaling exponent ν^{app}. The mapping of other proteins is shown in **Fig. A7**.

In this work, we set out to measure the ability of IDRs to respond to chemical composition changes in their surrounding solution. Although the solutions used here do not represent real cellular environments, they reveal that IDR ensembles carry an inherent, sequence-encoded sensitivity to changes in their chemical environment. This sensitivity can stem from different molecular features, and as far as we determined correlates only with the dimensions of the sequence (χ) in the aqueous buffer. IDR function through conformational selection has been reported for numerous proteins. In this mechanism, the function is linked to the conformational ensemble of the IDR, directly linking environment-induced ensemble changes to IDR activity. The most exciting idea our data suggests is that changes in the chemical composition that commonly occur in the cell can tune the function (or malfunction) of intrinsically disordered proteins.

Data Availability

All experimental and computational methods, as well as equations for analytical models are available in **Appendix A**. All code and data used to prepare the figures are available for download from https://github.com/sukeniklab/HiddenSensitivity.

References

(1) Wright, P. E.; Dyson, H. J. Intrinsically Disordered Proteins in Cellular Signalling and Regulation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16* (1), 18–29.

(2) Arai, M.; Sugase, K.; Dyson, H. J.; Wright, P. E. Conformational Propensities of Intrinsically Disordered Proteins Influence the Mechanism of Binding and Folding. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (31), 9614–9619.

(3) Wicky, B. I. M.; Shammas, S. L.; Clarke, J. Affinity of IDPs to Their Targets Is Modulated by Ion-Specific Changes in Kinetics and Residual Structure. *Proceedings of the National Academy of Sciences* **2017**, No. 24, 201705105.

(4) Borcherds, W.; Theillet, F.-X.; Katzer, A.; Finzel, A.; Mishall, K. M.; Powell, A. T.; Wu, H.; Manieri, W.; Dieterich, C.; Selenko, P.; Loewer, A.; Daughdrill, G. W. Disorder and Residual Helicity Alter p53-Mdm2 Binding Affinity and Signaling in Cells. *Nat. Chem. Biol.* **2014**, *10* (12), 1000–1002.

(5) Babu, M. M.; van der Lee, R.; de Groot, N. S.; Gsponer, J. Intrinsically Disordered Proteins: Regulation and Disease. *Curr. Opin. Struct. Biol.* **2011**, *21* (3), 432–440.

(6) Smoyer, C. J.; Jaspersen, S. L. Breaking down the Wall: The Nuclear Envelope during Mitosis. *Curr. Opin. Cell Biol.* **2014**, *26*, 1–9.

(7) Stewart, M. P.; Helenius, J.; Toyoda, Y.; Ramanathan, S. P.; Muller, D. J.; Hyman, A. A. Hydrostatic Pressure and the Actomyosin Cortex Drive Mitotic Cell Rounding. *Nature* **2011**, *469* (7329), 226–230.

(8) Vancraenenbroeck, R.; Harel, Y. S.; Zheng, W.; Hofmann, H. Polymer Effects Modulate Binding Affinities in Disordered Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2019**. https://doi.org/10.1073/pnas.1904997116.

(9) Davis, C. M.; Gruebele, M.; Sukenik, S. How Does Solvation in the Cell Affect Protein Folding and Binding? *Curr. Opin. Struct. Biol.* **2018**, *48*, 23–29.

(10) Sukenik, S.; Salam, M.; Wang, Y.; Gruebele, M. In-Cell Titration of Small Solutes Controls Protein Stability and Aggregation. *J. Am. Chem. Soc.* **2018**, *140* (33), 10497–10503.

(11) Tompa, P. The Principle of Conformational Signaling. *Chem. Soc. Rev.* **2016**, *45* (15), 4252–4284.

(12) Holehouse, A. S.; Sukenik, S. Controlling Structural Bias in Intrinsically Disordered Proteins Using Solution Space Scanning. *J. Chem. Theory Comput.* **2020**. https://doi.org/10.1021/acs.jctc.9b00604.

(13) Banks, A.; Qin, S.; Weiss, K. L.; Stanley, C. B.; Zhou, H. X. Intrinsically Disordered Protein Exhibits Both Compaction and Expansion under Macromolecular Crowding. *Biophys. J.* **2018**, *114* (5), 1067–1079.

(14) Hsu, P. P.; Sabatini, D. M. Cancer Cell Metabolism: Warburg and beyond. *Cell* **2008**, *134* (5), 703–707.

(15) Wu, D.; Zhou, H.-X. Designed Mutations Alter the Binding Pathways of an Intrinsically Disordered Protein. *Sci. Rep.* **2019**, *9* (1), 6172.

(16) Bah, A.; Forman-Kay, J. D. Modulation of Intrinsically Disordered Protein Function by Post-Translational Modifications. *J. Biol. Chem.* **2016**, *291* (13), 6696–6705.

(17) van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; Kim, P. M.; Kriwacki, R. W.; Oldfield, C. J.; Pappu, R. V.; Tompa, P.; Uversky, V. N.; Wright, P. E.; Babu, M. M. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **2014**, *114* (13), 6589–6631.

(18) Mansouri, A. L.; Grese, L. N.; Rowe, E. L.; Pino, J. C.; Chennubhotla, S. C.; Ramanathan, A.; O'Neill, H. M.; Berthelier, V.; Stanley, C. B. Folding Propensity of Intrinsically Disordered Proteins by Osmotic Stress. *Mol. Biosyst.* **2016**, *12* (12), 3695–3701.
(19) Goedhart, J.; von Stetten, D.; Noirclerc-Savoye, M.; Lelimousin, M.; Joosen, L.; Hink, M. A.; van Weeren, L.; Gadella, T. W. J., Jr; Royant, A. Structure-Guided Evolution of Cyan Fluorescent Proteins towards a Quantum Yield of 93%. *Nat. Commun.* **2012**, *3*, 751.
(20) Shaner, N. C.; Lambert, G. G.; Chammas, A.; Ni, Y.; Cranfill, P. J.; Baird, M. A.; Sell, B. R.; Allen, J. R.; Day, R. N.; Israelsson, M.; Davidson, M. W.; Wang, J. A Bright Monomeric Green Fluorescent Protein Derived from Branchiostoma Lanceolatum. *Nat. Methods* **2013**, *10* (5), 407–409.
(21) Mastop, M.; Bindels, D. S.; Shaner, N. C.; Postma, M.; Gadella, T. W. J., Jr; Goedhart, J. Characterization of a Spectrally Diverse Set of Fluorescent Proteins as FRET Acceptors for mTurquoise2. *Sci. Rep.* **2017**, *7* (1), 11999.
(22) Sørensen, C. S.; Kjaergaard, M. Effective Concentrations Enforced by Intrinsically Disordered Linkers Are Governed by Polymer Physics. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (46), 23124–23131.
(23) Martin, E. W.; Holehouse, A. S.; Grace, C. R.; Hughes, A.; Pappu, R. V.; Mittag, T. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J. Am. Chem. Soc.* **2016**, *138* (47), 15323–15335.
(24) Ferreon, A. C. M.; Ferreon, J. C.; Wright, P. E.; Deniz, A. A. Modulation of Allostery by Protein Intrinsic Disorder. *Nature* **2013**, *498* (7454), 390–394.
(25) Sukenik, S.; Sapir, L.; Gilman-Politi, R.; Harries, D. Diversity in the Mechanisms of Cosolute Action on Biomolecular Processes. *Faraday Discuss.* **2013**, *160*, 225–237; discussion 311–327.
(26) Senske, M.; Törk, L.; Born, B.; Havenith, M.; Herrmann, C.; Ebbinghaus, S. Protein Stabilization by Macromolecular Crowding through Enthalpy rather than Entropy. *J. Am. Chem. Soc.* **2014**, *136* (25), 9036–9041.
(27) Kozer, N.; Kuttner, Y. Y.; Haran, G.; Schreiber, G. Protein-Protein Association in Polymer Solutions : From Dilute to Semidilute to Concentrated. *Biophys. J.* **2007**, *92* (6), 2139–2149.
(28) Zosel, F.; Soranno, A.; Buholzer, K. J.; Nettels, D.; Schuler, B. Depletion Interactions Modulate the Binding between Disordered Proteins in Crowded Environments. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (24), 13480–13489.
(29) Soranno, A.; Koenig, I.; Borgia, M. B.; Hofmann, H.; Zosel, F.; Nettels, D.; Schuler, B. Single-Molecule Spectroscopy Reveals Polymer Effects of Disordered Proteins in Crowded Environments. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (13), 4874–4879.
(30) Pegram, L. M.; Wendorff, T.; Erdmann, R.; Shkel, I.; Bellissimo, D.; Felitsky, D. J.; Record, M. T., Jr. Why Hofmeister Effects of Many Salts Favor Protein Folding but Not DNA Helix Formation. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (17), 7716–7721.
(31) Sarkar, M.; Li, C.; Pielak, G. J. Soft Interactions and Crowding. *Biophys. Rev.* **2013**, *5* (2), 187–194.
(32) Pegram, L. M.; Record, M. T., Jr. Thermodynamic Origin of Hofmeister Ion Effects. *J. Phys. Chem. B* **2008**, *112* (31), 9428–9436.
(33) Vitalis, A.; Pappu, R. V. ABSINTH: A New Continuum Solvation Model for

Simulations of Polypeptides in Aqueous Solutions. *J. Comput. Chem.* **2009**, *30* (5), 673–699.

(34) Das, R. K.; Huang, Y.; Phillips, A. H.; Kriwacki, R. W.; Pappu, R. V. Cryptic Sequence Features within the Disordered Protein p27Kip1 Regulate Cell Cycle Signaling. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (20), 5616–5621.

(35) Martin, E. W.; Holehouse, A. S.; Peran, I.; Farag, M.; Incicco, J. J.; Bremer, A.; Grace, C. R.; Soranno, A.; Pappu, R. V.; Mittag, T. Valence and Patterning of Aromatic Residues Determine the Phase Behavior of Prion-like Domains. *Science* **2020**, *367* (6478), 694–699.

(36) Piovesan, D.; Tabaro, F.; Mičetić, I.; Necci, M.; Quaglia, F.; Oldfield, C. J.; Aspromonte, M. C.; Davey, N. E.; Davidović, R.; Dosztányi, Z.; Elofsson, A.; Gasparini, A.; Hatos, A.; Kajava, A. V.; Kalmar, L.; Leonardi, E.; Lazar, T.; Macedo-Ribeiro, S.; Macossay-Castillo, M.; Meszaros, A.; Minervini, G.; Murvai, N.; Pujols, J.; Roche, D. B.; Salladini, E.; Schad, E.; Schramm, A.; Szabo, B.; Tantos, A.; Tonello, F.; Tsirigos, K. D.; Veljković, N.; Ventura, S.; Vranken, W.; Warholm, P.; Uversky, V. N.; Dunker, A. K.; Longhi, S.; Tompa, P.; Tosatto, S. C. E. DisProt 7.0: A Major Update of the Database of Disordered Proteins. *Nucleic Acids Res.* **2017**, *45* (D1), D219–D227.

(37) Liu, Y.; Bolen, D. W. The Peptide Backbone Plays a Dominant Role in Protein Stabilization by Naturally Occurring Osmolytes. *Biochemistry* **1995**, *34* (39), 12884–12891.

(38) Auton, M.; Holthauzen, L. M. F.; Bolen, D. W. Anatomy of Energetic Changes Accompanying Urea-Induced Protein Denaturation. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (39), 15317–15322.

(39) O'Brien, E. P.; Ziv, G.; Haran, G.; Brooks, B. R.; Thirumalai, D. Effects of Denaturants and Osmolytes on Proteins Are Accurately Predicted by the Molecular Transfer Model. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (36), 13403–13408.

(40) Marsh, J. A.; Forman-Kay, J. D. Sequence Determinants of Compaction in Intrinsically Disordered Proteins. *Biophys. J.* **2010**, *98* (10), 2383–2390.

(41) Harries, D.; Rosgen, J. A Practical Guide on How Osmolytes Modulate Macromolecular Properties. *Methods Cell Biol.* **2008**, *84* (07), 679–735.

(42) Holehouse, A. S.; Pappu, R. V. Collapse Transitions of Proteins and the Interplay Among Backbone, Sidechain, and Solvent Interactions. *Annu. Rev. Biophys.* **2018**, *47*, 19–39.

(43) Riback, J. A.; Bowman, M. A.; Zmyslowski, A. M.; Knoverek, C. R.; Jumper, J. M.; Hinshaw, J. R.; Kaye, E. B.; Freed, K. F.; Clark, P. L.; Sosnick, T. R. Innovative Scattering Analysis Shows That Hydrophobic Disordered Proteins Are Expanded in Water. *Science* **2017**, *358* (6360), 238–241.

Chapter 3: Structural Preferences Shape the Entropic Force of Disordered Protein Ensembles

Adapted with permission granted by the American Chemical Society. The material originally appeared in the following: Feng Yu and Shahar Sukenik. The Journal of Physical Chemistry B 2023 127 (19), 4235-4244.

Abstract

Intrinsically disordered protein regions (IDRs) make up over 30% of the human proteome and exist in a dynamic conformational ensemble instead of a native, well-folded structure. Tethering IDRs to a surface (for example, the surface of a well-folded region of the same protein) can reduce the number of accessible conformations in these ensembles. This reduces the ensemble's conformational entropy, generating an effective entropic force that pulls away from the point of tethering. Recent experimental work has shown that this entropic force causes measurable, physiologically relevant changes to protein function. But how the magnitude of this force depends on the IDR sequence remains unexplored. Here we use all-atom simulations to analyze how structural preferences in IDR ensembles contribute to the entropic force they exert upon tethering. We show that sequence-encoded structural preferences play an important role in determining the magnitude of this force: compact, spherical ensembles generate an entropic force that can be several times higher than more extended ensembles. We further show that changes in the surrounding solution's chemistry can modulate IDR entropic force strength. We propose that the entropic force is a sequence-dependent, environmentally tunable property of terminal IDR sequences.

Introduction

Intrinsically disordered proteins and protein regions (IDRs) do not have a native structure. Instead, IDRs exist in a constantly interchanging conformational ensemble that contains transient and relatively weak intramolecular interactions. These interactions define the structural preferences and the resulting average shape of the ensemble. Decades of work have linked the structural preferences of IDRs to their biological functions[1–4].

Unlike well-folded proteins, IDR ensembles have a high conformational entropy. This conformational entropy can be reduced by covalently linking, or tethering, the IDR through one of its termini to a surface (**Fig. 3.1A**). In this case, entropy is reduced due to the constraint placed upon the ensemble by the surface it is tethered to. As a result, upon tethering an IDR will try to maximize its conformational entropy by producing an effective force that pulls up and away from the point of tethering, gaining entropy by increasing its number of accessible conformations generating an entropic force (**Fig. 3.1B**)[5,6].

This tethering scenario may seem rare when considering naturally occurring proteins, but it is rather common: in eukaryotes, IDRs are often tethered to a more rigid surface that constrains the chain's conformational entropy, and this tethering results in measurable effects. For example, IDRs tethered to a cell membrane can sense the curvature of the membrane and help to facilitate the endocytosis process through entropic force[7–10]. The same entropic force can also help translocate IDRs through the bacterial cell wall to the extracellular environment, an essential process for bacterial infection[11,12]. An even more prevalent scenario occurs when disordered N- or C-terminal IDRs are attached to a well-folded protein region (**Fig. 3.1A**). The entropic force exerted by such disordered terminal regions can influence protein function, including ligand binding affinity and thermodynamic stability[13,14]. These examples suggest that entropic force may be an important and prevalent mechanism unique to IDRs that mediates biological function.

To address the dynamics of IDR ensembles, previous studies have successfully applied analytical polymer models to describe IDR structural preferences (e.g. self-avoiding random chains and worm-like chains). These models can predict the average properties of IDR ensembles and have been systematically validated using experimental methods including small-angle X-ray scattering and single-molecule FRET experiments[3,15–21]. In addition, polymer models have been used previously to understand how chains exert an entropic force. Beyond the steric and chemical features of the monomers, these studies have implicated the length and the geometry of the constraining surface as major factors affecting polymer entropic force strength[5,6,22].

Thus, previous entropic force studies of IDRs also focused on the role of sequence length and the geometry of the constraining surface[9,11,13]. IDR length is indeed a critical factor in determining entropic force magnitude[8,13], since the longer the chain, the higher the number of conformations available. But is chain length always the most dominant factor affecting entropic force magnitude? Previous research has shown that, unlike homopolymers, IDR ensembles have distinct sequence-encoded structural

preferences[17,18,20,23–28]. These structural biases affect the average shape occupied by IDR ensembles[29,30], but their role in determining IDR entropic force strength has not been tested.

To link IDR structural biases with entropic force strength, we use all-atom Monte Carlo simulations to sample the conformational ensembles of over 90 experimentally validated IDR sequences. To gauge the magnitude of the entropic force sequences can exert, we measure the reduction in the number of allowed conformations upon tethering their ensembles to a flat surface. Our simulations show that the entropic force depends not only on the length of the IDR but also on its sequence-encoded ensemble shape, with more compact ensembles exerting a stronger entropic force. To further test this finding, we alter the dimensions of each ensemble by changing their interaction with the surrounding solution (while keeping the sequence intact). We show that solution-induced compaction also increases the entropic force, but only for a subset of the sequences. Our findings reveal how sequence-encoded intramolecular and protein:solution interactions combine to modulate the magnitude of the entropic force exerted by tethered IDR. They also suggest that the entropic force can be tuned by evolution to exert an optimized effect on full-length proteins.

Methods

Intrinsically disordered protein prediction with AlphaFold database

Systematic evaluations of AlphaFold2 (AF2) previously showed that it is a good predictor of intrinsically disordered regions[31–33]. We downloaded the predicted structures of three different proteomes (Saccharomyces cerevisiae: UP000002311, Arabidopsis thaliana: UP000006548, Homo sapiens: UP000005640) from the AF2 database version 3.[34] The disorder predictions are obtained from AF2's pLDDT score. Based on a previous report[31], we used 30 consecutive residues with pLDDT < 50% as an indicator for IDRs. Detected IDRs are labeled as terminal if they start at the N-terminal or end at the C-terminal of the protein in the AF2 database.

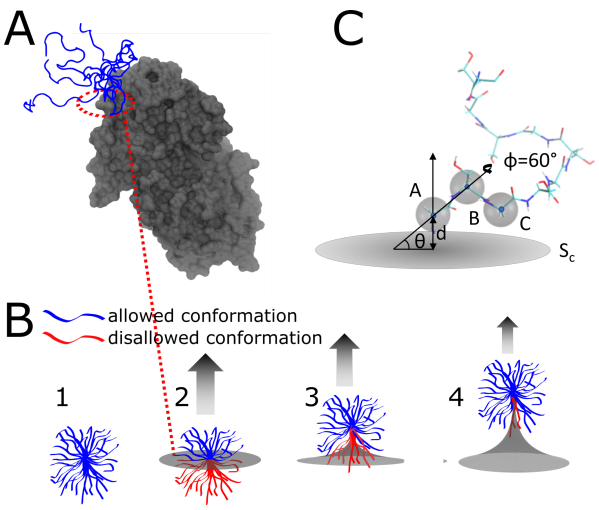


Figure 3.1. Dynamic IDR conformational ensemble generates an entropic force (**A**) IDR tethered to a well-folded domain. Here, the C-terminal IDR tail of the UDP-glucose 6-dehydrogenase (UGDH) protein is shown in blue (with 5 overlapping conformations to illustrate the variability in the ensemble) tethered to the main folded domain of the enzyme (in grey). Structure obtained from Alphafold2[32]. (**B**) Schematic showing how a constraining surface alters the conformational entropy of an IDR ensemble. (1) A few representative conformations from an IDR ensemble (blue) occupy an extended volume. (2) as the ensemble is tethered at the terminal to a surface (grey), some conformations clash with the surface (colored in red), causing them to be disallowed and lowering the conformational entropy. (3 and 4) The number of accessible tethered states (Ω_T) can be regained by "pulling up" against and pinching the surface (arrow). The ratio between the allowed and total number of conformations for a given ensemble is proportional to the entropic force strength (see **Eq. 3.3**). (**C**) Enhanced conformational sampling. All conformations of an IDR are aligned along the vector AB connecting the first two C_α atoms. The distance *d* between the constraining surface S_c and point A is varied to represent tether flexibility. The angle between vector AB and the constraint surface, θ, is varied to represent one degree of rotation for the ensemble, and a second angle, ϕ, represents the rotation angle along the AB vector.

All-atom Monte-Carlo simulation

All IDRs were simulated with the ABSINTH implicit solvent force field using the CAMPARI simulation suite v2_09052017.[35] We chose the ABSINTH forcefield and the CAMPARI simulation suite because of their extensive benchmarking and their computational efficiency, allowing us to simulate sequences that are 30-100 residues long in 2-4 days on a single processor. The example parameter file and simulation settings are provided in the GitHub repository. Simulations were conducted at 310 K with 10^7 steps of equilibration. After calibration, production conformations were written every 12,500 steps. For each IDR, we performed five independent simulations with ~5,600 individual conformations in each repeat. This leads to a total of ~28,000 conformations for each IDR (details in **Table S1 in Appendix B**). For PUMA scrambles, we performed three individual repeats, leading to ~16,800 conformations.

Calculation of ensemble properties

Normalized end-to-end distance. The end-to-end distance R_{ee} of polymers can be calculated based on the number of residues in the chain (N) using $R_{ee} = R_0 N^{\nu}$.[17] The scaling law ν can have a range of fractional values. Specifically, $\nu = 0.59$ for expanded chains, $\nu = 0.5$ for ideal (or θ-state) polymers, and $\nu = 0.33$ for collapsed/compacted polymers[17,36]. For homopolymers, the prefactor R_0 is constant and depends on the segment length of the monomer[17,37]. Seven glycine-serine dipeptide repeat (GS-repeats) sequences with 8, 16, 24, 32, 40, 48, and 64 GS segments were simulated and analyzed as described above with five individual repeats. We use GS-repeats because it maintains a consistent, experimentally validated point of reference for the entire dataset. The GS-repeat R_{ee} data were fitted using Scipy curve_fit function to the power law equation above (**Fig. B1**). The results of the fit gave a prefactor $R_0 = 0.55 \pm 0.06$ and an exponent $\nu = 0.48 \pm 0.03$ for GS repeats, demonstrating ideal polymer behavior. This is in agreement with the previous experimental data[38,39]. We use this fitted curve to normalize IDR R_{ee} for comparison across IDRs of different lengths. We interpolate and extrapolate corresponding GS-repeats R_{ee} based on the length of the IDR of interest. We calculate the normalized R_{ee} using the following equation.

$$\overline{R}_{ee} = \frac{R_{ee}}{R_{ee}^{GS}} - 1 \qquad (3.1).$$

Here, $\overline{R}_{ee}$ is the normalized end-to-end distance, R_{ee} is the end-to-end distance of the target IDR, and R_{ee}^{GS} is the calculated end-to-end distance of a GS-repeat sequence of the same length as the target IDR (obtained from the fit shown in **Fig. B1**).

Asphericity. IDR ensemble properties were analyzed using the MDtraj python library[40]. R_{ee} was calculated between the C_α of the first and last residue of the IDR. Helicity was calculated using the DSSP algorithm integrated into MDtraj[41]. Asphericity was calculated using the gyration tensor of the simulated IDR ensemble as described previously[42–44].

$$\delta = 1 - 3\left(\frac{\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1}{(\lambda_1+\lambda_2+\lambda_3)^2}\right) \qquad (3.2).$$

Here, δ is the asphericity, and $\lambda_{1,2,3}$ are the three principal moments of the gyration tensor. The standard deviation of all these properties was calculated based on the averages from the five independent repeats. Analysis scripts are available at the accompanying GitHub repository at https://github.com/sukeniklab/Entropic_Force.

Entropy analysis

To calculate the effect of tethering on IDR conformational entropy we count the number of allowed conformations in the ensemble upon tethering (**Fig. 3.1B**). To do this, we first tether each conformation of each simulated IDR ensemble to a single point on a flat surface and then calculate the number of allowed conformations Ω_T from the total number of conformations in the simulated ensemble Ω_U. Tethering is done relative to the first, second, and third C_α coordinates of each conformation, labeled here as A, B, C (**Fig. 3.1C**). For each conformation, we move A to the origin of the coordinate system. We plot the constraint surface, S_c, perpendicular to the surface containing atoms A, B, and C.

Enhanced sampling. In order to better understand the spatial relationship between the ensemble and the constraining surface, we perform several geometric transformations on each sampled conformation for calculating Ω_T: (1) To account for the possibility of stretching at the point of tethering, we vary the distance d between point A and S_c. (2) To account for the possibility of rotation around the point of tethering, we vary the half-angle θ formed between the norm vector to S_c with vector AB. (3) We rotate the vector AB with an angle ϕ. All the coordinates specified here are illustrated in **Fig. 3.1C.** In total, we make 36 transformations (3 values for d, 3 values for θ, and 6 values for ϕ) for each conformation of each simulated ensemble.

Entropy calculation. We consider the interaction between IDR and the constraint surface as a hard sphere interaction. Accessible conformations are defined as those with no C_α that is positioned below the constraining surface. We use the dot product between the norm vector of S_c and the coordinate of C_α to calculate and determine the relative position of the C_α to the surface S_c. We then count the number of all accessible conformations in the tethered, original ensemble and all d, θ, and ϕ permutations. Finally, we sum the number of accessible states from these perturbations and calculate the entropic force strength. The entropic force is then given by:[13]

$$\Delta S = k_B ln(\Omega_T/\Omega_U) \qquad (3.3).$$

Here k_B is the Boltzmann constant, Ω_T is the total number of possible IDR conformations when the ensemble is tethered to a surface and Ω_U is the total number of conformations sampled for the same IDR ensemble when untethered. The entropic force strength is proportional to the ΔS. The transformation and analysis scripts are provided as Jupyter notebooks at https://github.com/sukeniklab/Entropic_Force.

Ensemble XZ-projections. For each IDR conformation, we move A to the origin of the coordinate system and rotate the conformation to make AB fall on the Z-axis (Z>0). XZ-coordinate of each C_α will provide an ensemble projection of IDR ensemble on the XZ

plane. The C_α density was normalized by the number of amino acids in the sequence, the frame number of trajectories, and the bin size.

Solution Space Scanning simulations

Solution space scanning simulations are conducted as described previously[23,35,45]. Briefly, we modify the effective Hamiltonian of the ABSINTH force field to alter protein backbone:solvent interactions. The ABSINTH Hamiltonian is a sum of four energy terms:

$$E_{total} = W_{solv} + U_{LJ} + W_{el} + U_{corr} \quad (3.4).$$

U_{LJ} , W_{el} , and U_{corr} represent Lennard-Jones (LJ) potential, electrostatic interaction, and torsional correction terms for dihedral angles. W_{solv} is the solvation free-energy and equal to the transfer free energy between vacuum and diluted aqueous solution. Changing the free energy term W_{solv} such that results in a change in the protein:solvent relative interaction strength, defined by

$$\frac{W_{solv}^{max}(solution) - W_{solv}^{max}(water)}{W_{solv}^{max}(water)} \times 100\% \quad (3.5).$$

W_{solv}^{max} is the solvation free-energy calculated based on fully extended protein conformation in different solution conditions. Negative values of protein:solvent interaction represent solutions that are attractive to the protein backbone, such as urea solutions, while positive values represent solutions that are repulsive to the protein backbone, such as those containing protective osmolytes. A value of 0 represents a buffered, aqueous solution with no cosolutes. We simulated seven different solution conditions for each IDR with a protein:solvent relative interaction strength ranging from +3% (equivalent roughly to 1 M TMAO) to -3% (equivalent roughly to 1.5 M Urea)[45]. It is important to note, however, that even the most attractive solutions used here are not sufficient to unfold well-folded protein domains. We use the same temperature and sampling method for each solution condition as we do for aqueous solutions. The simulation averages of ensemble properties and entropic force in all solution conditions, as well as sequence details, are reported for all IDRs in **Table S1 in Appendix B**.

Limitations and drawbacks of entropic force calculations

In our calculations, we completely neglect any interactions between the IDR and the surface other than steric, hard-core repulsions. We also assume that the constraining surface is completely flat. In the context of an actual, full-length protein, constraining surfaces will have distinct chemical moieties, including hydrophobic, polar, and charged residues. Specific surface chemistries will introduce an enthalpic component to the free energy change upon IDR tethering which can alter, and sometimes completely reverse, the force induced by tethering. These effects are very important as shown in several cases, especially when charges are introduced[8,46,47].

Another limitation is that the constraint surface we use is fixed, flat, and does not change over time. The surface of folded domains displays irregular shapes, fluctuations and motions that may change the number of allowed conformations or change the overall entropy of the entire system which we didn't consider here. Indeed, some of the solution chemistry changes we use in this work may also act to alter these fluctuations.

To mitigate these limitations, we stress that the entire dataset was obtained using the same methods and analysis, and compared against the same GS-repeat benchmarks. This self-consistency is what allows us to probe the role of the ensemble itself on the entropic force, all other factors being held constant.

Results and Discussion

The human proteome is rich in disordered terminal sequences

We define terminal IDRs as those that exist at the N or C termini of proteins, and reason that with one free end, such IDRs can exert an entropic force against the more rigid, folded protein domain to which they are connected (**Fig. 3.1A**). To see if terminal IDRs are common in proteomes, we tested their prevalence in the yeast, arabidopsis, and human proteomes. using the AlphaFold Protein Structure Database v3 [48] (**Fig. 3.2A**). The confidence score of AlphaFold2 (pLDDT) has been previously shown to be a good indicator of potential disordered regions[31] and so was used to identify disordered regions in the three proteomes. A protein segment was marked as disordered when it had more than 30 consecutive residues with a 'very low' pLDDT score (< 50%). For the proteomes we tested, over 40% of proteins have at least one disordered segment, in line with previous studies[49] (**Fig. 3.2A**, left). In the human proteome specifically, over half of the proteins that contain IDRs have at least one at either the N- or C terminal (**Fig. 3.2A**, right). This result indicates that terminal-tethered IDRs exist widely in eukaryotes and that the entropic force scenario described above can occur in many proteins.

Based on past work, we reasoned that length is a factor that contributes strongly to the entropic force mechanism in these IDRs[8,13]. We therefore wanted to test if there is a significant difference in the length distribution of terminal vs. non-terminal IDRs[50,51]. Our analysis reveals that the length distribution is roughly the same between the terminal and non-terminal IDRs (**Fig. 3.2B**).

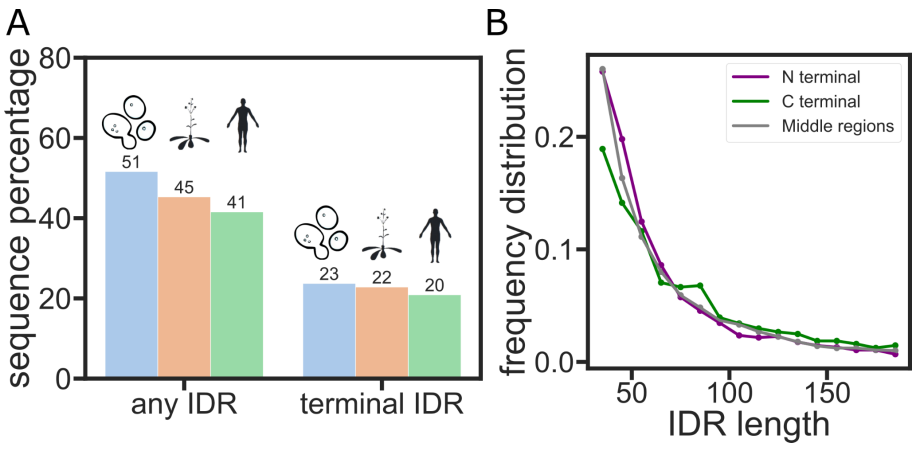


Figure 3.2. Entropic force may be a widely existing IDR function mechanism in the proteome. (**A**) The percentage of proteins that have a terminal IDR in the yeast, arabidopsis, or human proteomes. (**B**) Distribution of the number of amino acids in the IDRs of the human proteome.

An IDR simulation database reveals structural diversity

With IDR sequence length being roughly the same in both terminal and non-terminal sequences, we turned our attention to the structural preferences of their ensemble. Ensemble average end-to-end distance (R_{ee}) has been widely used to quantify the global dimensions and the internal structure of dynamic IDR conformational ensembles[17,18]. Since ensemble dimensions cannot be accurately predicted from the sequence, we used the ABSINTH forcefield to gain an atomic-level simulation of over 90 IDR ensembles. Most of these sequences are experimentally validated IDR sequences from the DisProt database[52] (**Table S1 in Appendix B**). These sequences have a diverse distribution of properties including the length, fraction of charged residues (FCR), and net charge per residue (NCPR) (**Fig. 3.3A-C**).

Simulations reveal a large distribution of R_{ee} (sometimes more than a factor of 2 for sequences with the same number of amino acids), indicating distinct structural preferences in these sequences. To compare different IDRs of various lengths across the proteome, we use Gly-Ser repeat peptides (GS-repeats) as a homopolymer point-of-reference. It has been shown experimentally that GS-repeats have a similar ensemble to an ideal homopolymer (a polymer where R_{ee} scales as $N^{0.5}$)[17,36,53]. We simulated several different lengths of GS-repeat sequences using the ABSINTH forcefield. Our simulation data shows R_{ee} of GS-repeats follows a scaling law with an exponent of 0.48 ± 0.03 (**Fig. B1**), which matches previously reported experimental results[38]. Our analysis shows that a large majority of the sequences measured deviate from the GS-repeat line (**Fig. 3.3D**).

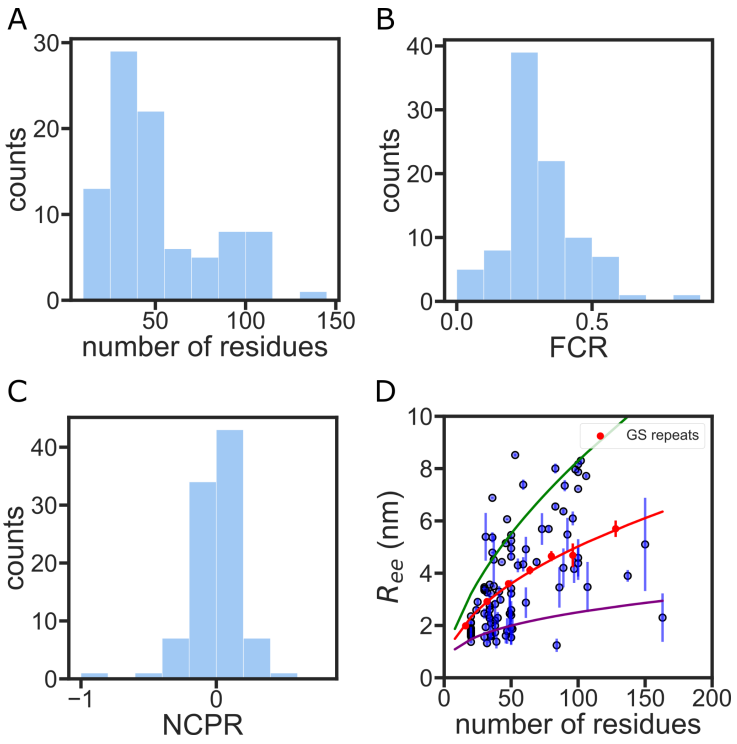


Figure 3.3. IDR simulation database shows diverse sequence properties and structural preferences. (**A**) The sequence length distribution of the IDR simulation database. (**B**) The fraction of charged residues (FCR) distribution of the IDR simulation database. (**C**) The net charge per residue (NCPR) distribution of the IDR simulation database. (**D**) End-to-end distance vs the number of residues for each simulated IDR. Error bars are calculated from five independent simulations of the same sequence. GS-repeat simulations are shown in red. The red curve is a power law fit of the GS-repeat data (see also **Fig. B1**). The green curve is the R_{ee} prediction of the GS-repeats with an exponent of 0.59 and the purple curve is the prediction of the GS-repeats with an exponent of 0.33, which represents the limits of an extended and compact homopolymer[36].

Quantifying the entropic force of disordered ensembles using enhanced sampling

We next wanted to probe if these structural preferences alter the magnitude of the entropic force these sequences exert. To assess how ensemble structural preferences change the entropic force, we quantified the change in IDR conformational entropy upon tethering the simulated ensemble to a flat surface and the change in allowed conformations/accessible states (as described in **Methods** and in **Eq. 3.3**). The change in conformational entropy upon tethering, $\Delta S/k_B$, is directly correlated to the magnitude of the entropic force (**Fig. 3.1B**).

To obtain the number of allowed conformations in the tethered state, Ω_T, we tethered our simulated IDR conformational ensemble to a flat surface through the N-terminal C_α. Beyond the conformations included in the ensemble, the geometry of the tethering point can also affect the magnitude of $\Delta S/k_B$. To account for this, we introduced an enhanced

sampling method to vary tethering configurations and measure the entropic force at various ensemble orientations relative to the tethered surface (**Fig. 3.1C, Methods**). With these variations, we generate additional conformations and plot an accessible state heatmap to visualize the number of allowed conformations in each orientation (**Fig. 3.4A**). To obtain a measure of the entropic force that will be comparable between all sequences, we sum the number of allowed conformations in all different orientations to provide a single entropic force strength for each sequence (**Fig. 3.4B**).

Validation of the entropic force calculation using experimental data

Several studies have highlighted the importance of IDR length on the entropic force it exerts[9,11]. A recent study by Keul et al. demonstrated that the length of a terminal IDR tail was the only factor determining its functional effect on the folded enzyme to which it was tethered[13]. The study focused on the C-terminal IDR of a key glycolytic enzyme, UDP glucose 6-dehydrogenase (UGDH). The study showed that the C-terminal IDR acts, through the entropic force it exerts, as an allosteric switch that alters the affinity of the protein to its allosteric feedback inhibitor UDP-xylose. The authors discovered that the entropic force (and the measured binding affinity) depend solely on the length of the terminal IDR, and not on its amino acid composition or sequence (**Table S1**). As a test of our method, we wanted to see if this length-dependent behavior for the UGDH IDR sequence is reproduced in our simulations.

The homopolymeric GS-repeat entropic force was fitted to an exponential decay function, indicating it is solely determined by the sequence length[5]. In agreement with Kuel et al.'s observations, UGDH-derived sequences of different lengths also fell on the same line as the GS-repeats (**Fig. 3.4B**). This indicates that the terminal UGDH IDR has entropic force strength similar to that of a homopolymer. However, UGDH might be a special case resulting from the specific amino acid composition. Indeed, two other IDR sequences display significantly different $\Delta S/k_B$ despite having the same number of residues (**Fig. 3.4B**). For example, we selected a disordered region of the type II methyltransferase (M.PvuII, Disprot ID: DP00060r010) from the DisProt database, and compared it to the C-terminal intracellular region of the mu-type opioid receptor (MOR-1, Disprot ID: DP00974r002). Both sequences are 38 residues long. Despite this, the C-terminal region of the MOR-1 has half as many accessible states as M.PvuII when tethered to a constraining surface, generating a stronger entropic force (**Fig. 3.4B**).

Is the magnitude of the entropic force dependent on amino acid composition alone, or on the sequence of the IDR? To answer this question, we generated a library of scrambled sequences of a naturally occurring sequence, the BH3 IDR domain of the p53-upregulated modulator of apoptosis (PUMA)[54] (**Fig. 3.4C**). Despite having the same sequence length and same amino acid composition, scrambles of the PUMA sequence demonstrated a significant difference in entropic force strength. The maximum entropic force of PUMA scrambles is more than two times the minimum force. We observed that scrambled sequences can exert both a stronger and a weaker entropic force upon tethering compared to the wild-type sequence. This result suggests that the order of amino acids in an IDR sequence, and not just amino acid composition, plays a vital role in determining entropic force strength (**Fig. 3.4C**).

Overall, our simulations recapitulated experimental observables that implicate IDR length as a key factor affecting IDR entropic force but also highlighted the role of amino acid composition and sequence in the magnitude of this force.

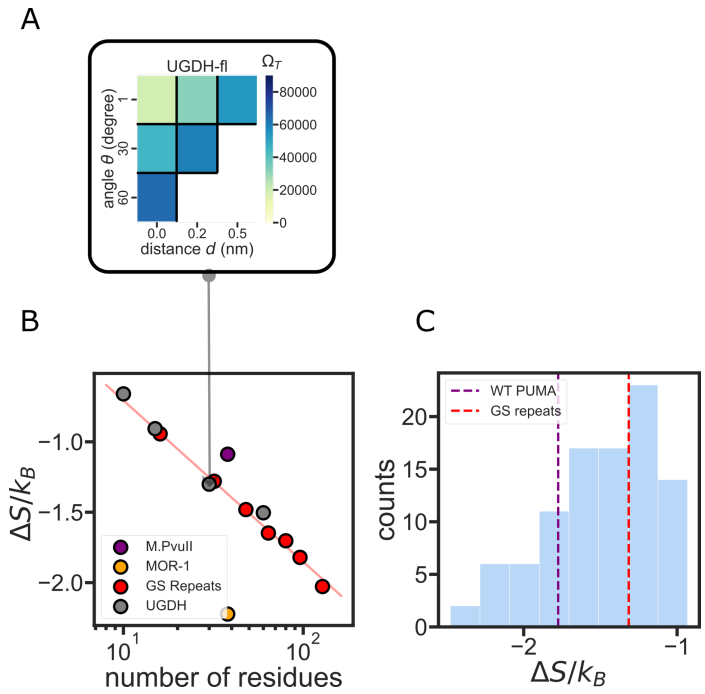


Figure 3.4. The role of IDR sequence length in determining entropic force strength. (**A**) The variables d and θ are varied discreetly to assess the number of allowed states Ω_T for the ensemble when tethered to the constraint surface. The color in each position on the grid represents the number of allowed states Ω_T from 6 different ϕ values. The total number of accessible states is used to calculate the entropic force strength for each construct. (**B**) Sequence length determines the entropic force strength of homopolymer-like IDRs. Red curve: an exponential fit of the GS-repeats entropic force strength. Grey dots: UGDH segments as measured in Ref. 13 show a similar entropic force as the equivalent GS-repeat homopolymer. (**C**) A histogram of the entropic force of 96 PUMA scramble sequences. The red dashed line shows the entropic force strength of the same-length GS-repeat sequence.

Systematic analysis of IDR entropic force

We next wanted to understand the role that sequence plays in determining IDR entropic force. We looked for sequence feature correlations with entropic force but found no strong correlations with most individual sequence features[18,19] (**Fig. B2**). One exception was the hydropathy decoration parameter proposed by Mittal and co-workers[21], which showed a strong negative correlation with entropic force (**Fig. B2H**), though this can be largely attributed to the length dependence of this metric (**Fig. B2I**). We therefore focused our attention on IDR ensemble dimensions, which are encoded in the sequence but are difficult to predict from structure[3,18]. We applied our enhanced sampling analysis to 94 IDR sequences obtained from the DisProt database. We observed IDRs generating

both higher and lower entropic force compared to GS-repeats, despite having the same length (**Fig. 3.5A**).

To ascertain how ensemble dimensions may play a role in determining $\Delta S/k_B$, we must first find a way to compare the ensembles of IDRs of various lengths. To do this, we normalize the average R_{ee} of all IDRs against the R_{ee} of a GS-repeat sequence of the same length to get normalized end-to-end distance $\overline{R}_{ee}$ (**Eq. 3.1, Methods**). $\overline{R}_{ee}$ has a negative value when the ensemble is more compact than a GS-repeat, and a positive value when an ensemble is more expanded. We plot $\Delta S/k_B$ for each sequence as a function of this normalized distance in **Fig. 3.5B**. It is immediately noticeable that the vertical red line drawn at $\overline{R}_{ee} = 0$ separates sequences with a higher entropic force (purple markers) from those with a weaker entropic force (green markers). This means ensembles that are on average more compact than an equivalent GS-repeat (as indicated by a negative $\overline{R}_{ee}$) tend to generate a stronger entropic force, while more expanded ensembles tend to generate a weaker entropic force than equivalent GS-repeats.

This seemed counterintuitive since our initial thought was that an expanded ensemble should take up more space and would therefore lose more conformational entropy upon tethering to the constraint surface. However, a more expanded ensemble will tend to have a higher persistence length and a more ellipsoid shape[55]. These properties mean that the backbone will point away from the tethered surface (because of this longer persistence length), reducing the number of conformations that will sterically clash with the surface. To validate this hypothesis, we calculated the average asphericity of the IDR ensemble[42]. Similar to $\overline{R}_{ee}$, ensembles with low asphericity have a lower entropic force, and ensembles with a high asphericity have a stronger entropic force than that of GS-repeats (**Fig. B3, B4**). This suggests that a more spherical ensemble tends to have a higher possibility of clashing with the constraining surface and thus generates a stronger entropic force, while a more elongated ensemble tends to have less interaction with the constraining surface. To verify this, we visualized the position of C_α atoms on an XZ plane that is normal to the constraining surface for several sequences (**Fig. 3.5C**). This visualization highlights how spherical ensembles with a low asphericity tend to have more atoms at or under the constraint surface (located at Z=0) while ellipsoidal ensembles with a high asphericity tend to expand with a higher atom density above the constraint surface.

Changes in solution chemistry alter IDR entropic force strength

An alternative way to change ensemble dimensions, and one that does not involve a change in IDR sequence is to expose IDRs to different solution environments[23,45]. Previously, we found that IDRs tend to be more sensitive than folded proteins to changes in the chemical composition of their surrounding solution. We designed the Solution Space Scanning method to simulate IDR ensemble structural preferences under changing solution conditions[45]. Briefly, the method alters IDR ensembles by tuning the protein backbone:solvent interactions of the ABSINTH forcefield to be more or less

repulsive than the value for water (see **Methods**). Usually, IDRs have a more compact conformational ensemble in repulsive solutions (e.g. in the presence of an osmolyte or a more crowded environment). In attractive solutions (e.g. urea or other denaturants), IDRs have an expanded conformational ensemble. However, this general trend can be mitigated and sometimes even reversed based on the IDR sequence[23,24,45].

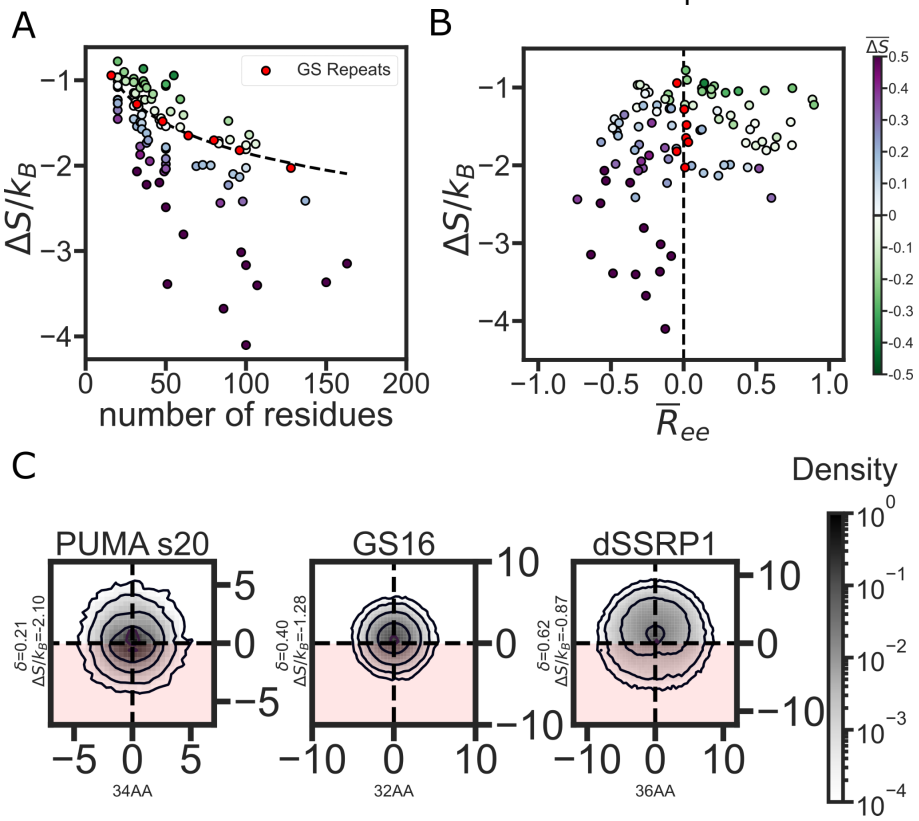


Figure 3.5. IDR structural preferences divide between weak and strong entropic force. (**A**) Entropic force *vs.* the number of residues in 94 different IDRs. The black curve is an exponential fit of GS-repeat data. Each point represents the entropic force of a single sequence calculated from 5 independent repeats. The color-coding shows the entropic difference between the IDR and the same-length GS-repeat ($\Delta\bar{S} = \Delta S/\Delta S^{GS} - 1$), with purple (green) markers showing a stronger (weaker) entropic force compared to the equivalent GS-repeat. (**B**) Entropic force *vs* the GS-repeat normalized end-to-end distance $\overline{R}_{ee}$ (see **Eq. 3.3**). Each marker represents a single IDR color-coded as in (A). (**C**) XZ-projections of C_α density for 3 different IDRs with increasing asphericity. The constraint plane is normal to Z=0 such that the density at Z>0 will avoid the surface and the density at Z<0 clashes with the surface (the disallowed region is indicated by the red color).

To see how solution-induced changes in the ensemble affect entropic force, we used Solution Space Scanning to simulate the ensemble average R_{ee} of the proteins shown in **Fig. 3.2B** and **Fig. 3.5** in five different solution conditions. We observed significant compaction of the ensemble in the repulsive solution, and the ensemble change is correlated with protein:solvent interaction strength (**Fig. 3.6A**). To quantify how ΔS changes with solution condition change, we use the change in entropic force between solute and buffer with the following equation:

$$\Delta\Delta S/k_B = \Delta S/k_B^{solution} - \Delta S/k_B^{aqueous} \quad (7).$$

Here, $\Delta\Delta S/k_B$ represents the change in the entropic force in different protein:solvent interactions. We calculate the entropic force change between the buffer/aqueous condition and other solution conditions. Our analysis shows that, on average, IDRs will generate a stronger entropic force when their ensemble is compacted due to the presence of a repulsive solution (**Fig. 3.6B**). This result strengthens our conclusion that compact IDR ensembles tend to exert a larger entropic force than extended ensembles.

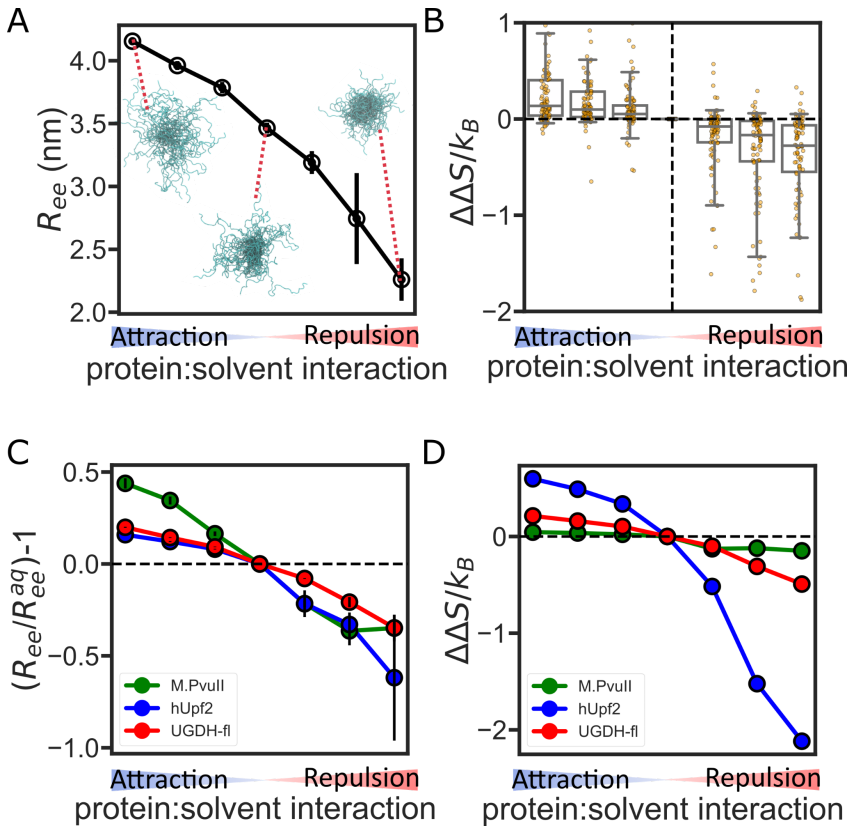


Figure 3.6 Solution conditions alter IDR entropic force. (**A**) End-to-end distance for UGDH-fl as a function of backbone:solvent interactions. The blow-up ensembles show representative conformations in attractive, neutral (aqueous), and repulsive solutions. (**B**) Box plot showing the change in entropic force due to change in protein backbone:solvent interactions. Boxes show the median as a central line, the median 50% as the box limits, and the median 90% of the data as the whiskers. Individual sequences are shown as points overlaid on each box. (**C**) Solution sensitivity of three IDR ensembles. Solution sensitivity is quantified using relative R_{ee} compared to the R_{ee} of the same IDR in the neutral (aqueous) solution. (**D**) The change in entropic force due to solution condition changes for the three IDR ensembles.

However, not every IDR is sensitive to solution condition changes. We observed that some IDRs do not have a significant entropic force change, despite significant changes in their ensemble (**Fig. 3.6C, D**). For example, M.PvuII displays a significant change in

R_{ee}, but almost no change in entropic force (**Fig. 3.6C**). On the other hand, the ensemble of the IDR of the regulator of nonsense transcripts 2 (hUpf2, Disprot ID: DP00949r013) is very sensitive to solution changes, and $\Delta S/k_B$ changes accordingly. This suggests that solution-driven changes in entropic force response are highly sequence-dependent. Different sequences encode diverse structural ensembles that in turn influence IDR environmental response. Interestingly, the UGDH IDR has a low sensitivity of entropic force despite its high sensitivity ensemble. Considering UGDH performs an allosteric function through its entropic force, this indicates some sequences may evolve to generate a stable entropic force for performing their function. This has been previously proposed as a general property of proline-rich domains in the Wnt signaling pathway[56].

Conclusions

Here we report on a computational method to quantify the conformational entropic force of tethered IDRs using all-atom Monte Carlo simulations. Compared to coarse-grained or analytical models, this method offers an accurate, quantitative metric of how IDR entropic force is determined by sequence-encoded conformational ensemble preferences. Our method is compared against and qualitatively matches previously published experimental measurements of entropic force (**Fig. 3.4B**). Our results also support the current literature and highlight that IDR sequence length is indeed a key factor in the entropic force it exerts (**Fig. 3.4B**). Despite its drawbacks and limitations (see the section in **Methods**), our method offers an accessible description of the entropic force which is computationally easy to calculate and a self-consistent dataset from which to draw conclusions linking between IDR sequence and entropic force.

Our simulations show that there is more to the story of entropic force than just the length of the sequence. We reveal that IDR structural preferences can determine the magnitude of entropic force strength. We show that the structural preferences of IDR ensembles are encoded not just in amino acid composition but also in their arrangement in the sequence, which can be an important factor in determining entropic force strength. Perhaps counterintuitively, we find that more expanded IDR ensembles can extract a weaker entropic force than more compact IDR ensembles when tethered to a flat surface (**Fig. 3.5A, 3.5B, B4**).

We also show that the entropic force exerted by an IDR can change when the surrounding chemical environment changes. By modulating protein backbone:solvent interactions, we altered IDR ensembles and showed that the entropic force magnitude of most IDRs increased as their ensembles became more compact, validating the trend shown for different sequences (**Fig. 3.6B**). This result also suggests the possibility of manipulating IDR entropic force by altering the physical-chemical composition of the cellular environment [23,24,57].

Since the dimensional properties of IDR sequences are sequence-encoded[58], we propose that some sequences have evolved to exert an outsized entropic force on the protein they are tethered to, while other sequences have evolved to exert a weak force. Our study further suggests that this entropic force can be modulated by post-translational modifications, binding of small molecules or other proteins[59], and changes in the cellular environment that are known to alter IDR ensembles[1,2,24,60]. Taken

together, the entropic force is a sequence-encoded, tunable function that may be more common than previously realized in IDR-containing proteins.

Data Availability

Appendix B contains **Figures B1-B4**. **Appendix B** provided the link to **Table S1** which contains all IDR sequences, the ensemble analysis result, and the entropic force analysis result of the simulation database. All code and data used to generate the figures in this chapter are provided at: https://github.com/sukeniklab/Entropic_Force

References

(1) Wright, P. E.; Dyson, H. J. Intrinsically Unstructured Proteins: Re-Assessing the Protein Structure-Function Paradigm. *J. Mol. Biol.* **1999**, *293* (2), 321–331.

(2) Dyson, H. J.; Wright, P. E. Intrinsically Unstructured Proteins and Their Functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6* (3), 197–208.

(3) van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.et al. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **2014**, *114* (13), 6589–6631.

(4) González-Foutel, N. S.; Glavina, J.; Borcherds, W. M.; Safranchik, M.; Barrera-Vilarmau, S.; Sagar, A.; Estaña, A.; Barozet, A.; Garrone, N. A.; Fernandez-Ballester, G. et al. Conformational Buffering Underlies Functional Selection in Intrinsically Disordered Protein Regions. *Nat. Struct. Mol. Biol.* **2022**, *29* (8), 781–790.

(5) Polson, J. M.; MacLennan, R. G. Entropic Force of Cone-Tethered Polymers Interacting with a Planar Surface. *Phys. Rev. E* **2022**, *106* (2), 024501.

(6) Maghrebi, M. F.; Kantor, Y.; Kardar, M. Entropic Force of Polymers on a Cone Tip. *EPL* **2011**, *96* (6), 66002.

(7) McMahon, H. T.; Gallop, J. L. Membrane Curvature and Mechanisms of Dynamic Cell Membrane Remodelling. *Nature* **2005**, *438* (7068), 590–596.

(8) Zeno, W. F.; Thatte, A. S.; Wang, L.; Snead, W. T.; Lafer, E. M.; Stachowiak, J. C. Molecular Mechanisms of Membrane Curvature Sensing by a Disordered Protein. *J. Am. Chem. Soc.* **2019**, *141* (26), 10361–10371.

(9) Zeno, W. F.; Baul, U.; Snead, W. T.; DeGroot, A. C. M.; Wang, L.; Lafer, E. M.; Thirumalai, D.; Stachowiak, J. C. Synergy between Intrinsically Disordered Domains and Structured Proteins Amplifies Membrane Curvature Sensing. *Nat. Commun.* **2018**, *9* (1), 4152.

(10) Fakhree, M. A. A.; Blum, C.; Claessens, M. M. A. E. Shaping Membranes with Disordered Proteins. *Arch. Biochem. Biophys.* **2019**, *677*, 108163.

(11) Halladin, D. K.; Ortega, F. E.; Ng, K. M.; Footer, M. J.; Mitić, N. S.; Malkov, S. N.; Gopinathan, A.; Huang, K. C.; Theriot, J. A. Entropy-Driven Translocation of Disordered Proteins through the Gram-Positive Bacterial Cell Wall. *bioRxiv*, 2020, 2020.11.24.396366. https://doi.org/10.1101/2020.11.24.396366.

(12) Pizarro-Cerdá, J.; Cossart, P. Bacterial Adhesion and Entry into Host Cells. *Cell* **2006**, *124* (4), 715–727.

(13) Keul, N. D.; Oruganty, K.; Schaper Bergman, E. T.; Beattie, N. R.; McDonald, W. E.; Kadirvelraj, R.; Gross, M. L.; Phillips, R. S.; Harvey, S. C.; Wood, Z. A. The Entropic Force Generated by Intrinsically Disordered Segments Tunes Protein Function. *Nature* **2018**, *563* (7732), 584–588.

(14) Dave, K.; Gasic, A. G.; Cheung, M. S.; Gruebele, M. Competition of Individual Domain Folding with Inter-Domain Interaction in WW Domain Engineered Repeat Proteins. *Phys. Chem. Chem. Phys.* **2019**, *21* (44), 24393–24405.

(15) Nettels, D.; Gopich, I. V.; Hoffmann, A.; Schuler, B. Ultrafast Dynamics of Protein Collapse from Single-Molecule Photon Statistics. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (8), 2655–2660.

(16) O'Brien, E. P.; Morrison, G.; Brooks, B. R.; Thirumalai, D. How Accurate Are Polymer Models in the Analysis of Förster Resonance Energy Transfer Experiments on Proteins? *J. Chem. Phys.* **2009**, *130* (12), 124903.

(17) Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B. Polymer Scaling Laws of Unfolded and Intrinsically Disordered Proteins Quantified with Single-Molecule Spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (40), 16155–16160.
(18) Das, R. K.; Pappu, R. V. Conformations of Intrinsically Disordered Proteins Are Influenced by Linear Sequence Distributions of Oppositely Charged Residues. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (33), 13392–13397.
(19) Firman, T.; Ghosh, K. Sequence Charge Decoration Dictates Coil-Globule Transition in Intrinsically Disordered Proteins. *J. Chem. Phys.* **2018**, *148* (12), 123305.
(20) 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.
(21) Zheng, W.; Dignon, G.; Brown, M.; Kim, Y. C.; Mittal, J. Hydropathy Patterning Complements Charge Patterning to Describe Conformational Preferences of Disordered Proteins. *J. Phys. Chem. Lett.* **2020**, *11* (9), 3408–3415.
(22) Nowicki, W.; Nowicka, G.; Narkiewicz-Michałek, J. Influence of Confinement on Conformational Entropy of a Polymer Chain and Structure of Polymer–Nanoparticles Complexes. *Polymer* **2009**, *50* (9), 2161–2171.
(23) Moses, D.; Yu, F.; Ginell, G. M.; Shamoon, N. M.; Koenig, P. S.; Holehouse, A. S.; Sukenik, S. Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to Their Chemical Environment. *J. Phys. Chem. Lett.* **2020**, 10131–10136.
(24) Moses, D.; Guadalupe, K.; Yu, F.; Flores, E.; Perez, A.; McAnelly, R.; Shamoon, N. M.; Cuevas-Zepeda, E.; Merg, A. D.; Martin, E. W. et al. Structural Biases in Disordered Proteins Are Prevalent in the Cell. *bioRxiv*, 2022, 2021.11.24.469609. https://doi.org/10.1101/2021.11.24.469609.
(25) Brady, J. P.; Farber, P. J.; Sekhar, A.; Lin, Y.-H.; Huang, R.; Bah, A.; Nott, T. J.; Chan, H. S.; Baldwin, A. J.; Forman-Kay, J. D. et al. Structural and Hydrodynamic Properties of an Intrinsically Disordered Region of a Germ Cell-Specific Protein on Phase Separation. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (39), E8194–E8203.
(26) Sottini, A.; Borgia, A.; Borgia, M. B.; Bugge, K.; Nettels, D.; Chowdhury, A.; Heidarsson, P. O.; Zosel, F.; Best, R. B.; Kragelund, B. B. et al. Polyelectrolyte Interactions Enable Rapid Association and Dissociation in High-Affinity Disordered Protein Complexes. *Nat. Commun.* **2020**, *11* (1), 5736.
(27) Theillet, F.-X.; Binolfi, A.; Bekei, B.; Martorana, A.; Rose, H. M.; Stuiver, M.; Verzini, S.; Lorenz, D.; van Rossum, M.; Goldfarb, D. et al. Structural Disorder of Monomeric α-Synuclein Persists in Mammalian Cells. *Nature* **2016**, *530* (7588), 45–50.
(28) Martin, E. W.; Holehouse, A. S.; Grace, C. R.; Hughes, A.; Pappu, R. V.; Mittag, T. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J. Am. Chem. Soc.* **2016**, *138* (47), 15323–15335.
(29) Baul, U.; Chakraborty, D.; Mugnai, M. L.; Straub, J. E.; Thirumalai, D. Sequence Effects on Size, Shape, and Structural Heterogeneity in Intrinsically Disordered Proteins. *J. Phys. Chem. B* **2019**, *123* (16), 3462–3474.
(30) Song, J.; Li, J.; Chan, H. S. Small-Angle X-Ray Scattering Signatures of Conformational Heterogeneity and Homogeneity of Disordered Protein Ensembles. *J. Phys. Chem. B* **2021**, *125* (24), 6451–6478.
(31) Reid Alderson, T.; Pritišanac, I.; Moses, A. M.; Forman-Kay, J. D. Systematic Identification of Conditionally Folded Intrinsically Disordered Regions by AlphaFold2. *bioRxiv*, 2022, 2022.02.18.481080.

https://doi.org/10.1101/2022.02.18.481080.
(32) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A. et al. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596* (7873), 590–596.
(33) Piovesan, D.; Monzon, A. M.; Tosatto, S. C. E. Intrinsic Protein Disorder and Conditional Folding in AlphaFoldDB. *Protein Sci.* **2022**, e4466.
(34) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A. et al. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res.* **2022**, *50* (D1), D439–D444.
(35) Vitalis, A.; Pappu, R. V. ABSINTH: A New Continuum Solvation Model for Simulations of Polypeptides in Aqueous Solutions. *J. Comput. Chem.* **2009**, *30* (5), 673–699.
(36) Flory, P. J. The Configuration of Real Polymer Chains. *J. Chem. Phys.* **1949**, *17* (3), 303–310.
(37) Hammouda, B. SANS from Homogeneous Polymer Mixtures: A Unified Overview. In *Polymer Characteristics*; Springer Berlin Heidelberg: Berlin, Heidelberg, 1993; pp 87–133.
(38) Sørensen, C. S.; Kjaergaard, M. Effective Concentrations Enforced by Intrinsically Disordered Linkers Are Governed by Polymer Physics. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (46), 23124–23131.
(39) Basak, S.; Sakia, N.; Dougherty, L.; Guo, Z.; Wu, F.; Mindlin, F.; Lary, J. W.; Cole, J. L.; Ding, F.; Bowen, M. E. Probing Interdomain Linkers and Protein Supertertiary Structure In Vitro and in Live Cells with Fluorescent Protein Resonance Energy Transfer. *J. Mol. Biol.* **2021**, *433* (5), 166793.
(40) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109* (8), 1528–1532.
(41) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*. 1983, pp 2577–2637. https://doi.org/10.1002/bip.360221211.
(42) Vitalis, A. *Probing the Early Stages of Polyglutamine Aggregation with Computational Methods*; Washington University in St. Louis, 2009.
(43) Steinhauser, M. O. A Molecular Dynamics Study on Universal Properties of Polymer Chains in Different Solvent Qualities. Part I. A Review of Linear Chain Properties. *J. Chem. Phys.* **2005**, *122* (9), 094901.
(44) Tran, H. T.; Pappu, R. V. Toward an Accurate Theoretical Framework for Describing Ensembles for Proteins under Strongly Denaturing Conditions. *Biophys. J.* **2006**, *91* (5), 1868–1886.
(45) Holehouse, A. S.; Sukenik, S. Controlling Structural Bias in Intrinsically Disordered Proteins Using Solution Space Scanning. *J. Chem. Theory Comput.* **2020**, *16* (3), 1794–1805.
(46) Knotts, T. A., 4th; Rathore, N.; de Pablo, J. J. An Entropic Perspective of Protein Stability on Surfaces. *Biophys. J.* **2008**, *94* (11), 4473–4483.
(47) Taneja, I.; Holehouse, A. S. Folded Domain Charge Properties Influence the Conformational Behavior of Disordered Tails. *Curr Res Struct Biol* **2021**, *3*, 216–228.
(48) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A. et al. Highly Accurate

Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589.
(49) Deiana, A.; Forcelloni, S.; Porrello, A.; Giansanti, A. Intrinsically Disordered Proteins and Structured Proteins with Intrinsically Disordered Regions Have Different Functional Roles in the Cell. *PLoS One* **2019**, *14* (8), e0217889.
(50) Vuzman, D.; Levy, Y. Intrinsically Disordered Regions as Affinity Tuners in Protein–DNA Interactions. *Mol. BioSyst.* 2012, pp 47–57. https://doi.org/10.1039/c1mb05273j.
(51) Moesa, H. A.; Wakabayashi, S.; Nakai, K.; Patil, A. Chemical Composition Is Maintained in Poorly Conserved Intrinsically Disordered Regions and Suggests a Means for Their Classification. *Mol. Biosyst.* **2012**, *8* (12), 3262–3273.
(52) Piovesan, D.; Tabaro, F.; Mičetić, I.; Necci, M.; Quaglia, F.; Oldfield, C. J.; Aspromonte, M. C.; Davey, N. E.; Davidović, R.; Dosztányi, Z.; Elofsson, A. et al. DisProt 7.0: A Major Update of the Database of Disordered Proteins. *Nucleic Acids Res.* **2017**, *45* (D1), D219–D227.
(53) van Rosmalen, M.; Krom, M.; Merkx, M. Tuning the Flexibility of Glycine-Serine Linkers To Allow Rational Design of Multidomain Proteins. *Biochemistry* **2017**, *56* (50), 6565–6574.
(54) Wicky, B. I. M.; Shammas, S. L.; Clarke, J. Affinity of IDPs to Their Targets Is Modulated by Ion-Specific Changes in Kinetics and Residual Structure. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (37), 9882–9887.
(55) Chin, A. F.; Toptygin, D.; Elam, W. A.; Schrank, T. P.; Hilser, V. J. Phosphorylation Increases Persistence Length and End-to-End Distance of a Segment of Tau Protein. *Biophys. J.* **2016**, *110* (2), 362–371.
(56) Liu, C.; Yao, M.; Hogue, C. W. V. Near-Membrane Ensemble Elongation in the Proline-Rich LRP6 Intracellular Domain May Explain the Mysterious Initiation of the Wnt Signaling Pathway. *BMC Bioinformatics*. 2011. https://doi.org/10.1186/1471-2105-12-s13-s13.
(57) Cuevas-Velazquez, C. L.; Vellosillo, T.; Guadalupe, K.; Schmidt, H. B.; Yu, F.; Moses, D.; Brophy, J. A. N.; Cosio-Acosta, D.; Das, A.; Wang, L. et al. Intrinsically Disordered Protein Biosensor Tracks the Physical-Chemical Effects of Osmotic Stress on Cells. *Nat. Commun.* **2021**, *12* (1), 5438.
(58) Das, R. K.; Ruff, K. M.; Pappu, R. V. Relating Sequence Encoded Information to Form and Function of Intrinsically Disordered Proteins. *Curr. Opin. Struct. Biol.* **2015**, *32*, 102–112.
(59) De Los Rios, P.; Ben-Zvi, A.; Slutsky, O.; Azem, A.; Goloubinoff, P. Hsp70 Chaperones Accelerate Protein Translocation and the Unfolding of Stable Protein Aggregates by Entropic Pulling. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (16), 6166–6171.
(60) Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W. et al. Intrinsically Disordered Protein. *J. Mol. Graph. Model.* **2001**, *19* (1), 26–59.

Chapter 4 : Leveraging IDR Solution Sensitivity for Biosensor Design

Adapted with permission granted by Springer Nature. The material originally appeared in the following:
Cuevas-Velazquez, C.L., Vellosillo, T., Guadalupe, K., Schmit, B., Yu, F., et al. Intrinsically disordered protein biosensor tracks the physical-chemical effects of osmotic stress on cells. Nat Commun 12, 5438 (2021).
Structural biases in disordered proteins are prevalent in the cells Moses, D., Guadalupe, K., Yu, F. et al. Nat. Struct. Mol. Biol. In press (2024).

Abstract

In the preceding chapters of this dissertation, we have explored the complex landscape of IDR conformational ensembles and showed their sensitivity to the surrounding physical environment. Building on this foundation, this chapter explores the significance of IDR solution sensitivity within the cellular environment. We focus particularly on the context of desiccation protection-related IDRs. This examination sheds light on how these dynamic IDR ensembles respond to the effects of cellular dehydration and osmotic pressure changes. We take advantage of SolSpace simulations and FRET experiments to engineer a novel IDR-based biosensor, Sensor Expressing Disordered Protein 1 (SED1). SED1 employs the Arabidopsis AtLEA4-5 IDR attached to FRET fluorescent-protein pairs to monitor cellular responses to osmotic pressure. The ability of this construct to sense its surrounding environment is then validated across various organisms. This innovative biosensor utilizes IDRs to monitor and respond to environmental changes, showcasing the practical application of our fundamental research on IDR ensemble and solution sensitivities. This work underscores the potential of leveraging IDR solution sensitivity to create functional protein and can be a key to unlocking new bioengineering capabilities.

Introduction

IDR transient secondary structure enables IDR desiccation protection

Understanding how desiccation affects molecular biology is critical for several reasons. It provides insights into how the loss of water at the cellular level affects biomolecules and cellular structures. It is also vital for comprehending how organisms adapt to and survive in environments with limited water availability. This is especially pertinent in the context of climate change, where increasing areas of the globe are experiencing drought conditions.

During desiccation, the cellular environment undergoes a dramatic transformation that poses significant challenges to organismal survival. As water is progressively lost from the cell, it leads to a decrease in cell volume and can cause the environment of the cytoplasm to become crowded and viscous[1]. The loss of water significantly disrupts the balance of ions within the cell, a critical aspect of maintaining cellular homeostasis. This ionic imbalance can have detrimental effects on the cell's metabolic processes[2–4]. Furthermore, the dehydrated state poses a risk to the structural integrity and function of cellular structures and enzymes[5]. Proteins, for instance, may denature or aggregate in such stressful conditions, losing their functional shape and thus impairing the cell's biochemical processes[1,6–8].

Despite their sensitivity to the drastic changes in desiccating environments, IDRs have been reported to act as essential protectants against desiccation in organisms from all kingdoms of life[9–11]. In this context, it is particularly intriguing to ask how the flexible and dynamic ensemble of protective IDRs responds and adapts to the dramatic environmental changes brought about by desiccation. Can we link our IDR solution response to explain their protective function?

With thousands of IDRs reported to be linked to desiccation, we focus on IDR sequences from the Arabidopsis Late Embryogenesis Abundant (AtLEA) proteins. The LEA protein family has been involved in desiccation protection across different organisms[10,12]. LEA proteins were first identified in plant seeds during the late stages of embryogenesis, particularly under conditions of water deficit. Most LEAs are predicted to be fully disordered and classified into different groups based on their unique motifs[13,14]. In some cases, their presence was shown to enhance the plant tolerance to desiccation and water deficit[15–17]. Despite the overall structural disorder in aqueous solutions, previous research on AtLEA4-5 suggests that the N-terminal region can form a transient amphipathic α-helix structure under water stress[10]. We suggest that the formation of α-helices in AtLEA4-5 upon dehydration leads to compaction of their structural ensemble. This conformational change, triggered by environmental stressors, is likely critical to their functional mechanism in conferring desiccation tolerance[18,19].

Leveraging this ensemble structural change, we have developed an innovative IDR sensor that uses the AtLEA4-5 protein as a basis, to detect and measure changes in osmotic pressure in the cellular environment. This sensor is the first example of how the environmental sensitivity of IDRs can be a tool for understanding and monitoring cellular responses to changes in cellular physical chemistry.

Methods

For a detailed description of the experimental and computational methods for this section please refer to **Appendix C** and **D**.

Results and Discussion

IDR solution sensitivity enables IDR environmental sensing

We have already demonstrated that fluorescence resonance energy transfer (FRET) can detect changes in IDR ensembles in different chemical environments in **Chapter 2**. Our *in vitro* FRET construct can be coupled to IDR whose ensembles are sensitive to solution chemistry changes to create FRET-based sensors for monitoring the cellular environment, as has been done for well-folded proteins[20]. The high osmotic pressure on cells may lead to high macromolecular crowding in the cellular environment. In **Chapter 2**, IDRs demonstrated a compact ensemble under crowding triggered by the osmolytes. Thus, we propose that the FRET sensor expressed in cells will form a compact ensemble due to the solution sensitivity of the IDR under a high osmotic pressure environment and thus generate a stronger FRET signal compared to under a normal environment. Therefore, observing the change in FRET signal can monitor the osmotic pressure change around the cell.

To validate that the AtLEA4-5 has a high solution sensitivity and is a good candidate for the biosensor, we first ran SolSpace simulations of the AtLEA4-5 sequence and its scrambles (**Fig. 4.1A**). Our results have proven highly encouraging, as they indicate that AtLEA4-5 exhibits a high degree of solution sensitivity when compared to other IDRs in our extensive simulation database (gray curves in **Fig. 4.1C**, also described in **Chapter 2**). In addition, we generated 5 sequence scrambles with the same length and the same amino-acid composition with AtLEA4-5. None of those demonstrated a solution sensitivity higher than the wild-type sequence (**Fig. 4.1C**). This suggests the AtLEA4-5 solution sensitivity is high compared to other IDRs and determined by its unique sequence rather than the amino acid composition.

Based on these findings, we proposed that the sensitivity of AtLEA4-5 to the chemical environment could be harnessed to develop biosensors for detecting osmotic pressure changes surrounding cells. To turn the sequence into a biosensor, we combined the AtLEA4-5 with mCerulean3 and mCitrine fluorescence proteins, attached to the N- and C-terminal, respectively. These two fluorescence proteins create a FRET pair that facilitates sensing. We named this construct Sensor Expressing Disordered Protein 1 (SED1) (**Fig. 4.1B**).

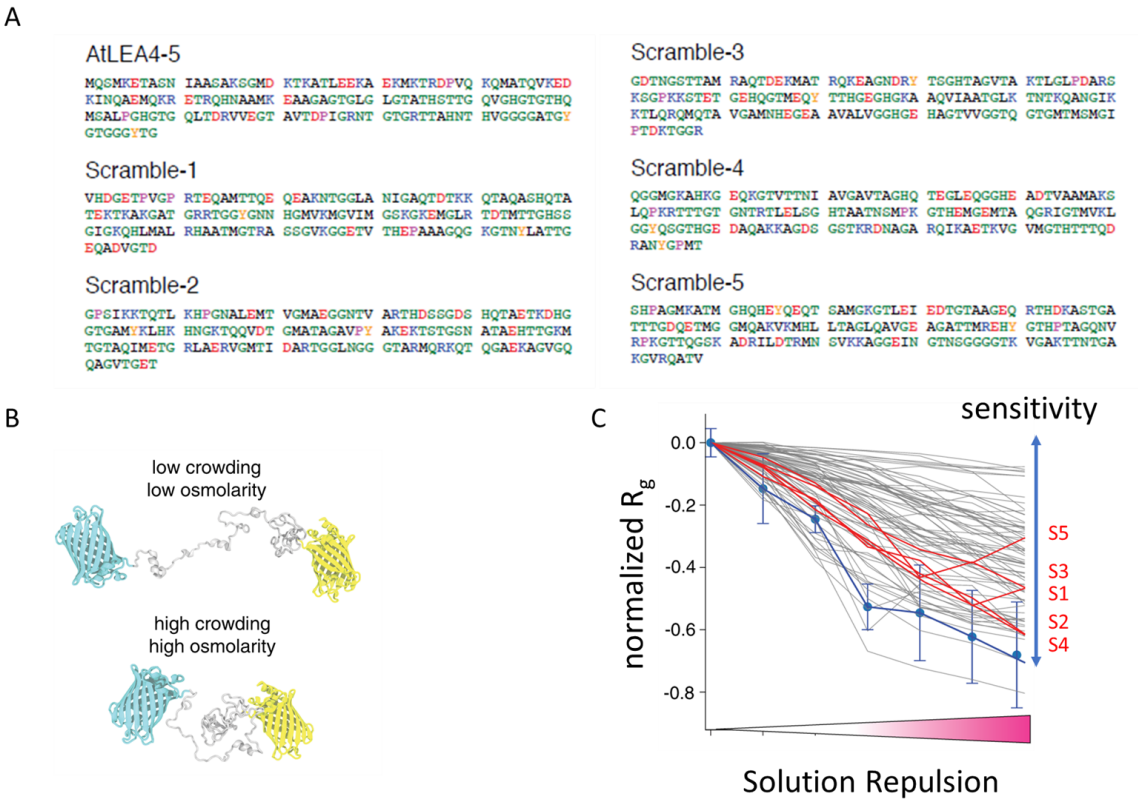


Figure 4.1. AtLEA4-5 has a high solution sensitivity. (**A**) AtLEA4-5 and its random scramble sequences as the candidate for the biosensor (**B**) Representative conformations of the SED osmotic pressure FRET sensor in an expanded (top) and compact state (bottom). (**C**) Computational solution space scan of the normalized radius of gyration (R_g) of AtLEA4-5 (blue), five different scrambled sequences shown in (**A**) (red), and 70 different naturally occurring IDRs (gray) under different solution repulsion levels.

To validate the effectiveness of our biosensor design, we've expressed the biosensor in a range of organisms, including yeast, E. coli, and mammalian cells. To measure the degree of sensing, we introduced osmotic stress to the cellular environment by changing the concentration of NaCl in the solution. When exposed to a high NaCl concentration that induces elevated osmotic pressure inside the cell, SED1 FRET efficiency increases in all organisms tested (**Fig. 4.2A, B, C**). This result proves that when exposed to hyperosmotic conditions, the FRET signal intensifies significantly, whereas hypoosmotic conditions result in a slight reduction in the FRET signal (**Fig. C3**). The construct and experimental details are described in **Appendix C**.

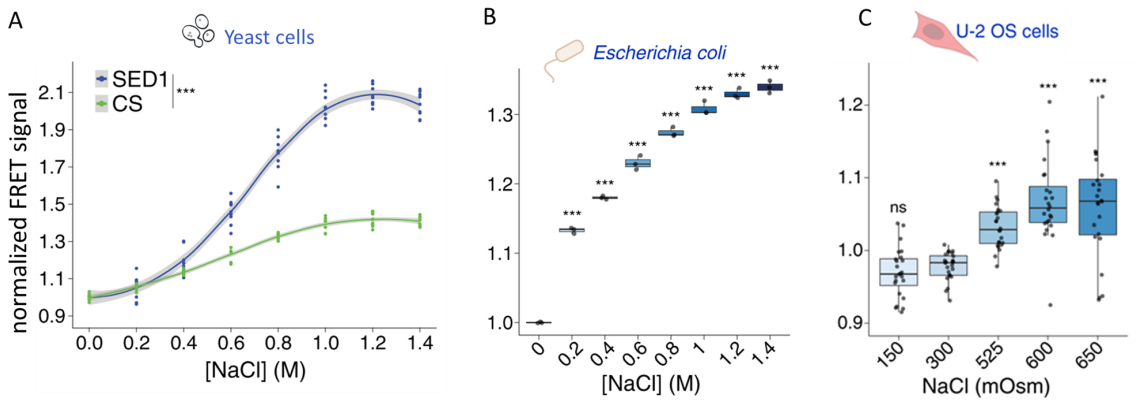


Figure 4.2. SED1 osmotic pressure sensing in different cell types. (**A**) SED1 can sense the osmotic pressure in yeast cells. The X-axis demonstrated the osmotic pressure modified by NaCl concentration. The y-axis is the normalized FRET signal described in Appendix C. A previously designed Crowding biosensor (CS) is based on an artificially synthesized helix-turn-helix structured motif[21]. (**B**) SED1 sensing of osmotic pressure in E. coli cells. (**C**) SED1 sensing of osmotic pressure in human-derived U-2 OS cells.

The creation of IDR biosensors is a critical milestone in our ability to leverage the unique properties of IDRs for technological uses. It underscores the significance of IDR ensemble sensitivity as a key aspect of IDR functionality. By harnessing IDR sensitivity, we are not only advancing our knowledge of these dynamic protein regions but also opening up new avenues for exploring and manipulating cellular environments with profound implications for both research and practical applications.

Discussion

We hypothesized that the SED1 biosensor utilized the transient helical propensity of AtLEA4-5 to sense the surrounding environment. However, many IDR sequences and their ensemble have no tendency to form a helical structure. How do their IDR ensembles sense the surrounding environmental change? Is there any other hidden structure that can determine IDR solution response?

To address this, we created three sequence scrambles for the BH3 IDR of the p53 upregulated apoptosis modulator (PUMA). The PUMA wild-type IDR is reported to exhibit helical propensity in an unbound state and to form a stable helical structure upon binding with the Mcl-1 region[22]. Scrambling the PUMA sequence maintains the same amino acid composition but disrupts the helical propensity. By comparing the solution responses of these sequences, we aim to verify whether α-helix formation is crucial for the IDR solution response.

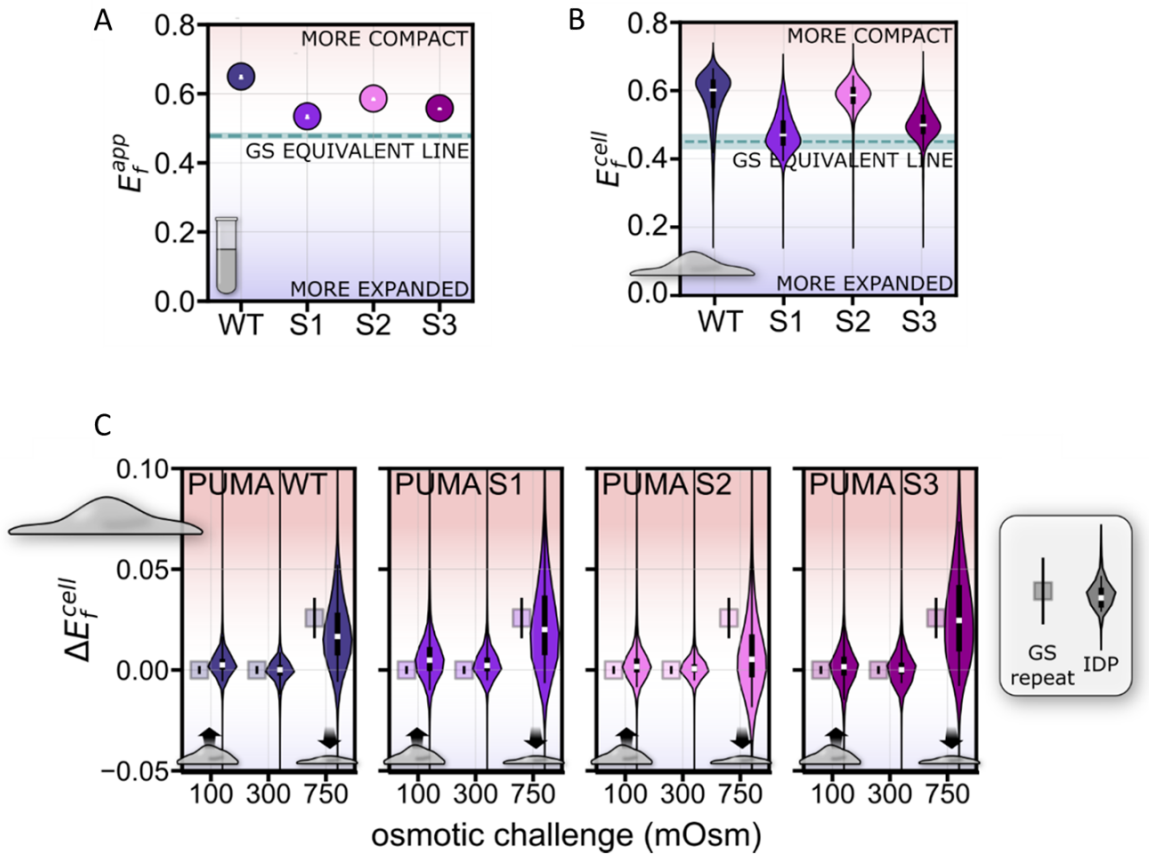


Figure 4.3. Assessing the sensitivity of PUMA and its scrambles to cellular osmotic challenges. (**A**) E_f^{app} of PUMA constructs. E_f^{app} represents FRET efficiency in vitro. High E_f^{app} represents a low average R_{ee} between FRET pairs and a compact IDR ensemble. Error bars represent errors from two repeats. Here the dashed line represents the expected value for a GS-repeat (described in **Chapter 2**) construct of the same length (34 residues) as PUMA WT and scrambles. (**B**) E_f^{cell} of PUMA constructs. E_f^{cell} represents FRET efficiency in vivo. (**C**) The osmotic challenge of HEK293T cells expressing PUMA constructs. Violin plots represent the data for PUMA constructs and squares represent the change of FRET efficiency of a GS-repeat equivalent.

To test the design of PUMA scrambles, I conducted CD experiments on PUMA variant sequences. Under buffer conditions, the PUMA wild-type (WT) sequences demonstrated a concentration-dependent helical propensity as indicated by the minimum at 208 nm and 222 nm on the CD spectra. Higher concentrations of PUMA led to greater helical propensity in the buffer. In contrast, the PUMA scrambles showed little tendency to form a secondary structure, as designed and predicted (**Figure D10, Figure 1.2D** in **Chapter 1).**

We further measured the ensemble dimensions of the PUMA scrambles (S1, S2, S3) to uncover their hidden structural preferences. The ensemble global dimension of PUMA scrambles is compared to the same length ideal polymer behavior GS-repeats reported in Chapter 2 and Chapter 3. The GS-repeats are homopolymeric sequences that have been reported to have an unbiased IDR ensemble[23]. Thus a more compact ensemble

compared to GS-repeats suggests attractive intramolecular interactions. A more expanded ensemble compared to GS-repeats suggests repulsive intramolecular interactions. The relative R_{ee} of IDR ensembles was using the in-vitro FRET method mentioned in **Chapter 2** (**Fig. 4.3A**). Here, a higher E_f represents a compact ensemble with a small R_{ee}. A smaller E_f represents a more expanded ensemble with a large R_{ee} (The conversion and calculation of E_f can be found in **Appendix D**.). The result shows that PUMA WT possesses the most compact ensemble. This may be due to the high helical propensity. Despite lacking helical propensity, the ensemble of scrambles showed significant differences. The S1 ensemble exhibits a more compact formation, whereas the S2 ensemble presents a more expanded ensemble. To verify this finding, the average R_g of the FRET constructs was measured using Size-Exclusion Chromatography coupled with Small-Angle X-ray Scattering (SEC-SAXS) (**Figure 1.2B** in **Chapter 1**). The trend of ensemble dimension is the same compared to the FRET method as the PUMA WT shows the most compact ensemble while the S2 possesses the most expanded ensemble.

The ensemble difference of PUMA scramble suggests that helicity is not required to compact the IDR ensemble and thus proves the presence of other intramolecular interactions between residues that shape the ensemble, even in live cells. These interactions may include long-range interactions between charged or aromatic residues performing like intramolecular “stickers”, as indicated by polymer models and other experimental research[24]. The *in vivo* FRET analysis correlated well with the outcomes from the in-vitro method and SAXS data, indicating that the compactness of the ensemble is a fundamental characteristic of the IDR sequence. Additionally, this reinforces the accuracy of measuring the dimensions of the IDR ensemble using FRET pairs under both setups.

Furthermore, these differences in hidden structure significantly affect the IDR solution sensitivity. Using a construct similar to the SED1 biosensor, we measured the response of these scrambles to osmotic pressure changes in cells. Under the buffer condition and in HEK293T cell, the average ensemble dimension of each construct measured with E_f shows the same trend as *in vitro* (**Fig. 4.3A, B**). When we introduced the osmotic pressure change surrounding the cell, we compared the solution response of PUMA to the solution response of GS-repeat peptides with the same length. The PUMA solution sensitivity is measured by the FRET signal change (ΔE_f^{cell}) under different osmotic pressures. Compared to the PUMA WT, which tends to form a secondary structure, S1 and S3 demonstrated greater ensemble change and sensitivity to hyperosmotic pressure. Meanwhile, S2 showed a stable ensemble with little change in global dimensions under both hyperosmotic and hypoosmotic conditions (**Fig. 4.3C**). This suggests that PUMA scrambles possess a similar ability to sense change in osmotic pressure, even without the formation of helices. Thus the ability of IDR to sense the surrounding environment is encoded by its ensemble flexibility and plasticity rather than just the tendency to form secondary structures.

Our findings revealed that IDRs can still show significant changes in response to osmotic pressure changes despite a lack of helical propensity. This suggests that other forms of intramolecular interactions are influential in IDR behavior, pointing at a more complex

mechanism underlying IDR environmental sensitivity. Future computational and experimental research can be focused on uncovering the mechanism and design of IDR intramolecular interactions to improve the efficiency of biosensors. This application also suggests the IDR biosensor FRET construct can be used as a tool to track IDR ensemble change in vivo which broadens the application of the biosensor.

Conclusions

This chapter applies the research of the previous 2 chapters and demonstrates how IDR ensemble sensitivity and structural preference impact IDR function in rapidly changing environments. Building on our understanding of IDR sensitivity, we have engineered a novel IDR-based FRET biosensor, SED1, based on AtLEA4-5 protein, which is a desiccation protection IDR that can form helix under osmotic stress. My simulation data verified that AtLEA4-5 has a high solution sensitivity under different solution conditions. This makes AtLEA4-5 a good IDR to monitor cellular responses to osmotic pressure across various organisms. Furthermore, SED1 shows ideal sensing ability in E. coli, yeast, and mammalian cells suggesting the success of biosensor design. With the biosensor construct and PUMA scrambles, we discovered that intramolecular interactions can determine IDR solution sensitivity instead of helicity, which points out the new direction for improving and designing the biosensor. Overall, this research not only creates a unique IDR biosensor that can be applied to multiple species but also provides a solid *in vivo* experimental method to measure IDR solution sensitivity to discover more potential of IDR solution response in the bioengineering field.

My computational research can be leveraged across multiple disciplines to create disorder-based sensors. With the advancement of IDR research, we anticipate that such computational models and simulations will find utility in a diverse range of scientific inquiries, from the design of novel biosensors to the understanding of IDR ensemble in cells This cross-disciplinary potential underscores the transformative power of computational research in IDR field.

References

(1) Romero-Perez, P. S.; Dorone, Y.; Flores, E.; Sukenik, S.; Boeynaems, S. When Phased without Water: Biophysics of Cellular Desiccation, from Biomolecules to Condensates. *Chem. Rev.* **2023**, *123* (14), 9010–9035.

(2) Chamberlin, M. E.; Strange, K. Anisosmotic Cell Volume Regulation: A Comparative View. *Am. J. Physiol.* **1989**, *257* (2 Pt 1), C159–C173.

(3) Burg, M. B.; Kwon, E. D.; Kültz, D. Regulation of Gene Expression by Hypertonicity. *Annu. Rev. Physiol.* **1997**, *59*, 437–455.

(4) Liu, B.; Poolman, B.; Boersma, A. J. Ionic Strength Sensing in Living Cells. *ACS Chem. Biol.* **2017**, *12* (10), 2510–2514.

(5) Hoekstra, F. A.; Golovina, E. A.; Buitink, J. Mechanisms of Plant Desiccation Tolerance. *Trends Plant Sci.* **2001**, *6* (9), 431–438.

(6) Oliver, M. J.; Farrant, J. M.; Hilhorst, H. W. M.; Mundree, S.; Williams, B.; Bewley, J. D. Desiccation Tolerance: Avoiding Cellular Damage During Drying and Rehydration. *Annu. Rev. Plant Biol.* **2020**, *71*, 435–460.

(7) Bray, E. A. Plant Responses to Water Deficit. *Trends Plant Sci.* **1997**, *2* (2), 48–54.

(8) Potts, M. Desiccation Tolerance: A Simple Process? *Trends Microbiol.* **2001**, *9* (11), 553–559.

(9) Chakrabortee, S.; Meersman, F.; Schierle, G. S. K.; Bertoncini, C. W.; McGee, B.; Kaminski, C. F.; Tunnacliffe, A. Catalytic and Chaperone-like Functions in an Intrinsically Disordered Protein Associated with Desiccation Tolerance. *Proceedings of the National Academy of Sciences* **2010**, *107* (37), 16084–16089.

(10) Cuevas-Velazquez, C. L.; Saab-Rincón, G.; Reyes, J. L.; Covarrubias, A. A. The Unstructured N-Terminal Region of Arabidopsis Group 4 Late Embryogenesis Abundant (LEA) Proteins Is Required for Folding and for Chaperone-like Activity under Water Deficit. *J. Biol. Chem.* **2016**, *291* (20), 10893–10903.

(11) Boothby, T. C.; Tapia, H.; Brozena, A. H.; Piszkiewicz, S.; Smith, A. E.; Giovannini, I.; Rebecchi, L.; Pielak, G. J.; Koshland, D.; Goldstein, B. Tardigrades Use Intrinsically Disordered Proteins to Survive Desiccation. *Mol. Cell* **2017**, *65* (6), 975–984.e5.

(12) Hand, S. C.; Menze, M. A.; Toner, M.; Boswell, L.; Moore, D. LEA Proteins during Water Stress: Not Just for Plants Anymore. *Annu. Rev. Physiol.* **2011**, *73*, 115–134.

(13) Bremer, A.; Wolff, M.; Thalhammer, A.; Hincha, D. K. Folding of Intrinsically Disordered Plant LEA Proteins Is Driven by Glycerol-Induced Crowding and the Presence of Membranes. *FEBS J.* **2017**, *284* (6), 919–936.

(14) Covarrubias, A. A.; Romero-Pérez, P. S.; Cuevas-Velazquez, C. L.; Rendón-Luna, D. F. The Functional Diversity of Structural Disorder in Plant Proteins. *Arch. Biochem. Biophys.* **2020**, *680*, 108229.

(15) Battaglia, M.; Olvera-Carrillo, Y.; Garciarrubio, A.; Campos, F.; Covarrubias, A. A. The Enigmatic LEA Proteins and Other Hydrophilins. *Plant Physiol.* **2008**, *148* (1), 6–24.

(16) Hundertmark, M.; Hincha, D. K. LEA (late Embryogenesis Abundant) Proteins and Their Encoding Genes in Arabidopsis Thaliana. *BMC Genomics* **2008**, *9*, 118.

(17) Olvera-Carrillo, Y.; Campos, F.; Reyes, J. L.; Garciarrubio, A.; Covarrubias, A. A. Functional Analysis of the Group 4 Late Embryogenesis Abundant Proteins Reveals Their Relevance in the Adaptive Response during Water Deficit in Arabidopsis. *Plant Physiol.* **2010**, *154* (1), 373–390.

(18) Shimizu, T.; Kanamori, Y.; Furuki, T.; Kikawada, T.; Okuda, T.; Takahashi, T.; Mihara,

H.; Sakurai, M. Desiccation-Induced Structuralization and Glass Formation of Group 3 Late Embryogenesis Abundant Protein Model Peptides. *Biochemistry* **2010**, *49* (6), 1093–1104.
(19) Hughes, S. L.; Schart, V.; Malcolmson, J.; Hogarth, K. A.; Martynowicz, D. M.; Tralman-Baker, E.; Patel, S. N.; Graether, S. P. The Importance of Size and Disorder in the Cryoprotective Effects of Dehydrins. *Plant Physiol.* **2013**, *163* (3), 1376–1386.
(20) Liu, B.; Åberg, C.; van Eerden, F. J.; Marrink, S. J.; Poolman, B.; Boersma, A. J. Design and Properties of Genetically Encoded Probes for Sensing Macromolecular Crowding. *Biophys. J.* **2017**, *112* (9), 1929–1939.
(21) Boersma, A. J.; Zuhorn, I. S.; Poolman, B. A Sensor for Quantification of Macromolecular Crowding in Living Cells. *Nat. Methods* **2015**, *12* (3), 227–229, 1 p following 229.
(22) Rogers, J. M.; Steward, A.; Clarke, J. Folding and Binding of an Intrinsically Disordered Protein: Fast, but Not “Diffusion-Limited.” *J. Am. Chem. Soc.* **2013**, *135* (4), 1415–1422.
(23) Sørensen, C. S.; Kjaergaard, M. Effective Concentrations Enforced by Intrinsically Disordered Linkers Are Governed by Polymer Physics. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (46), 23124–23131.
(24) Choi, J.-M.; Holehouse, A. S.; Pappu, R. V. Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions. *Annu. Rev. Biophys.* **2020**, *49*, 107–133.

Chapter 5: Conclusion

This dissertation has provided a comprehensive exploration of how the surrounding physical-chemical environment impacts IDR ensembles. Using simulations and other computational approaches, we discovered that IDR solution response is correlated with IDR ensemble dimensions and encoded in the arrangement of amino acids rather than their composition. Then I applied these findings to create biosensors based on IDRs, proving the value of my computational approach.

Chapter 1 introduced IDR ensembles and how their structural preferences play important roles in determining their function. IDR ensembles are sensitive to changes in the surrounding cellular environment, including factors such as ionic strength and temperature. This sensitivity is due to their high degree of surface area exposure and a lack of intramolecular interactions. To quantify IDR ensemble response to the surrounding environments, I used Monte-Carlo simulations and SolSpace scans to investigate IDR ensemble compaction and repulsion in varying physical-chemical environments.

Chapter 2 revealed how solution responses are encoded within IDR sequences rather than amino acid composition. I identified a correlation between IDR solution sensitivity and IDR ensemble dimensions which was verified by *in vitro* experiments. This sets the stage for appreciating how IDRs interact with their environment, emphasizing the predictive power of computational analysis in understanding IDR behavior. Based on this result, we hypothesize that IDR solution responses may be coded in their structural preferences and intramolecular interactions.

Moving forward, Chapter 3 expanded upon the understanding of chemical solution sensitivity to physical constraint by examining the generation of entropic force by IDR when confined by surrounding membrane/macromolecules. The discovery that an IDR's global dimensions are intimately linked to the strength of the entropic force it generates further underscores the importance of IDR ensemble dimensions. This has profound implications for the functional mechanisms of IDRs, suggesting that their shape and extent are not merely passive features but are active determinants of their role within crowding environments such as cytoplasm or cell membranes.

Finally, Chapter 4 leveraged the work shown in the previous chapters and showed a proof-of-principle for the design of an IDR biosensor to detect osmotic pressure change around cells. Starting with desiccation protection IDRs, we investigated how to manipulate IDR solution sensitivity to design IDR biosensors and understand IDR functional mechanisms. The creation of biosensors is a significant step forward for the IDR research community, offering a platform to detect cellular environmental change under osmotic pressure.

Looking ahead, the IDR simulation dataset and simulation protocols in this dissertation were developed to provide large datasets for exploring IDR ensemble dynamics. The expansion of this dataset can lead to the development of predictive models that further link IDR solution sensitivity with its sequence. By utilizing these models and datasets, researchers can strategically design IDR sequences to modify IDR ensembles within various cellular environments. In this regard, I hope my computational framework can advance drug development strategies that target IDR function in cellular environments.

Appendix A

Supplementary information for Publication:
Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to their Chemical Environment

The material originally appeared in the following: David Moses, Feng Yu, Garrett M. Ginell, Nora M. Shamoon, Patrick S. Koenig, Alex S. Holehouse, and Shahar Sukenik The Journal of Physical Chemistry Letters 2020 11 (23), 10131-10136.

Table S1-S4 can be found at this link:
https://github.com/sukeniklab/HiddenSensitivity

A1 Experimental Methods

A1.1 FRET construct design and cloning

FRET backbone (called fIDR_pET-28a(+)-TEV, **Fig. A9**) was prepared by ligating mTurquoise2 and mNeonGreen into pET28a-TEV backbone using 5' NdeI and 3' XhoI restriction sites. Genes encoding for IDR regions were obtained from GenScript, and ligated between the two fluorescent proteins using 5' NdeI and 3' HindIII restriction sites. Cloned plasmids were amplified in XL1 Blue (Invitrogen) cell lines using manufacturer supplied protocol. Sequences of all IDR sequence inserts are available in **Table S4**.

A1.2 FRET construct expression and purification

Plasmids encoding for FRET constructs were expressed in BL21 (DE3) cells in LB medium with 50 µg/mL kanamycin. Cultures were incubated at 37 °C while shaking at 225 rpm until OD600 of 0.6 was reached (approx. 3 h), then induced with 1 mM IPTG and incubated for 20 h at 16 °C while shaking at 225 rpm. Cells were harvested by centrifugation for 15 min at 3,000 rcf, the supernatant was discarded, and the cells were lysed in lysis buffer (50 mM NaH_2PO_4, pH 8, 0.5 M NaCl) using an Avestin Emulsiflex C3 homogenizer. Lysate was centrifuged for 1 h at 20,000 rcf and the supernatant collected and flowed through a column packed with Ni-NTA beads (Qiagen). FRET construct was eluted with 50 mM NaH_2PO_4, pH 8, 0.5 M NaCl, 250 mM imidazole, and further purified using size-exclusion chromatography on a Superdex 200 PG column (GE Healthcare) in an AKTA go protein purification system (GE Healthcare). The purified FRET constructs were aliquoted into 200 µL aliquots, flash-frozen in liquid nitrogen, and stored at -80 °C in 20 mM sodium phosphate buffer, pH 7.4, with the addition of 100 mM NaCl. Protein concentration was measured after thawing and before use using UV-vis absorbance at 434 and 506 nm (the peak absorbance wavelengths for mTurquoise2 and mNeonGreen,

respectively; the molar absorbance coefficients for mTurquoise2 and mNeonGreen are 30,000 $cm^{-1}M^{-1}$ and 116,000 $cm^{-1}M^{-1}$, respectively.[1] Calculations of concentration based on λ = 434 nm produced slightly higher values than calculations based on λ = 506 nm, so the concentrations based on the measurement at λ = 506 nm were used), and purity was assessed by SDS-PAGE after thawing and before use.

A1.3 Solution preparation and specifics

Solutes were purchased from Alfa Aesar (Dextran, Xylitol, L-Tryptophan, Sarcosine, PEG200, PEG400, PEG1500, PEG2000, PEG4000, PEG6000, PEG8000, PEG10000), VWR (D-Sorbitol), GE Healthcare (Ficoll), TCI (Meso-Erythritol, D-(+)-Trehalose Dihydrate), Thermo Scientific (Guanidine Hydrochloride), Acros Organics (D-Mannitol, Betaine Monohydrate, L(+)-Arabinose), Sigma-Aldrich (Myo-Inositol, Taurine), and Fisher BioReagents (Ethylene Glycol, D-Galactose, Glycerol, Glycine, L-Proline, Tricine, Potassium Chloride, Sodium Chloride, Urea), and used without further purification. Stock solutions were made by mixing the solute with 20 mM sodium phosphate buffer, pH 7.4, with the addition of 100 mM NaCl except for NaCl and KCl solutions, which were free of additional salt. The same buffer was used for all dilutions.

A1.4 FRET experiments

FRET experiments were conducted in black plastic 96-well plates (Nunc) using a CLARIOstar plate reader (BMG LABTECH). Buffer, stock solution, and purified protein solution were mixed in each well to reach a volume of 150 µL containing the desired concentrations of the solute and the FRET construct, with a final concentration of 1 µM protein (or of each FP in the case of the “untethered” control). Fluorescence measurements were taken from above, at a focal height of 5.7 mm, with gain fixed at 1020 for all samples. For each FRET construct, two repeats with 12 replicates each were performed for each protein in neat buffer, and at least two repeats were done in every other solution condition. Fluorescence spectra were obtained for each FRET construct in each solution condition by exciting the sample in a 16-nm band centered at λ = 420 nm, with a dichroic at λ = 436.5 nm, and measuring fluorescence emission from λ = 450 to 600 nm, averaging over a 10 nm window moved at intervals of 0.5 nm. Base donor and acceptor spectra for each solution condition were obtained using the same excitation and emission parameters on solutions containing 1 µM mTurquoise2 or mNeonGreen alone, and measuring fluorescence emission from 450 to 600 nm[1,2].

A1.5 Calculation of E_f

The process of calculating the FRET efficiency E_f for a FRET construct in one solute at a range of concentrations is summarized in **Fig. A10**. Specifically, E_f of each FRET construct in each solution condition was calculated by linear regression of the fluorescence spectrum of the FRET construct with the spectra of the separate donor and

acceptor emission spectra (in order to correct for solute-dependent effects on fluorophore emission) in the same solution conditions. E_f was calculated using :

$$E_f = 1 - \frac{F_d}{\frac{Q_d f_d}{Q_a f_a} F_s + F_d}$$

where F_d is the decoupled donor contribution, F_s is the decoupled acceptor contribution, f_d is the area-normalized donor spectrum, f_a is the area-normalized acceptor spectrum, $Q_d = 0.93$ is the quantum yield of the donor, and $Q_a = 0.8$ is the quantum yield of the acceptor[2,3].

More specifically, the data for each series of solution conditions consisting of increasing concentrations of a single solute was processed in the following manner:

1. Raw spectra for the free donor and free acceptor in the various solution conditions were loaded, and the averages of all repeats in each solution condition were computed. These averages are referred to as the "raw" donor and acceptor spectra below because they will be further corrected.

2. The donor and acceptor peak intensities were assumed to change in a linear fashion with increasing solute concentration, so peak height of donor or acceptor-only spectra *vs.* concentrations were linearly fit.

3. To correct for artifacts (such as variations in FRET construct concentration between different wells) that may contribute to unexpected differences in fluorescence intensity, a correction factor was applied to each raw donor and acceptor spectrum to bring the peak intensity to the linear fit described in step 2, resulting in "corrected" donor and acceptor spectra. Importantly, while this corrected well-to-well variations in raw data, it did little to affect the overall values or trends in χ (e.g., without this correction **Fig. 2.1** and **2.2** would vary by less than 5%).

4. The raw FRET construct fluorescence spectra for the series were loaded.

5. To compensate for unintended direct excitation of the acceptor by donor excitation frequency, the corrected acceptor spectrum for each solution condition was subtracted from the FRET construct spectrum for each solution condition, resulting in "corrected" FRET construct spectra.

6. The corrected donor, acceptor and FRET construct spectrum for each solution condition was fitted with a linear regression function to determine the decoupled contributions of the donor and acceptor to the FRET construct spectrum.

7. E_f of each FRET construct in each solution condition was calculated using the equation shown above.

A1.6 Assessment of the expected scaling behavior for interprotein distances

For flexible polymers, the end-to-end distance (R_e) and radius of gyration (R_g) follow well-defined scaling relationships defined by $R = AN^{\nu}$. Here, R is a physical distance (i.e., R_e or R_g), A is a constant in units of distance, N is the unitless degree of polymerization (i.e., number of residues) v is a unitless scaling exponent[4,5]. In the limit of finite-sized polymers, v is more correctly written as ν^{app}. For constructs with two fluorescent proteins connected by a flexible linker, in the limit of infinitely long linkers, the inter-fluorescent protein distance will approximately equal the end-to-end distance of the intervening linker. However, in the limit of finite-length linkers where the linker dimensions are on par with the dimensions of the fluorescent proteins, we anticipated that deviations from conventional scaling theory might arise due at least in part to the excluded volume effects of the fluorescent proteins.

To assess the role of excluded volume effects in deviation, we examined the expected intra-fluorescent protein distance dependence on linker length for a well-defined self-avoiding random coil system. Such a model is convenient in that the dependence of the end-to-end distance for a flexible self-avoiding polymer is well-defined analytically as $R_e = BN^{0.59}$.

We built a series of fluorescent-protein linker constructs with linkers of various lengths and performed simulations at all-atom resolution using the CAMPARI simulation engine and the ABSINTH implicit solvent model (see also **SI Section A2.1**). To achieve behavior in the true self-avoiding random coil limit, the Hamiltonian (which here refers to the instantaneous potential energy function) used to generate the ensemble does not experience a contribution from the attractive portion of the Lennard-Jones potential for short-range non-bonded interactions, nor solvation effects, nor electrostatic interactions, as described previously[6]. The backbone dihedral angles associated with residues in the two fluorescent proteins were held fixed, while the backbone dihedral angles associated with residues in the linker were allowed to vary. All side chain angles were fully flexible. In effect, this provides a “toy” system in a well-defined polymer limit which allows us to assess the impact of the fluorescent proteins without any confounding concerns for forcefield accuracy, sampling challenges, etc.

We first established that a flexible linker between two FPs indeed scales as expected for a self-avoiding random coil. The scaling exponent obtained by fitting a number of GS repeats vs. intra-chain distances revealed a scaling exponent of 0.61 - extremely close to the value of 0.59 expected from analytical theory (**Fig. A11**).

We then repeated the same analysis for the same system assessing the inter-domain distance between the chromophores in the fluorescent proteins - i.e., the inter-fluorescent protein distance (**Fig. A12**). Unlike the intra-chain distances (**Fig. A11**), the inter-domain distances showed a linear dependence on linker length. This behavior is readily explained by the excluded volume impact of the fluorescent proteins. For shorter chains, the inter-fluorescent protein distance is much larger than the distance between the ends of an analogous flexible polymer because the excluded volume from the fluorescent proteins effectively acts as repulsors at the chain ends. However, as

chain length increases this effect becomes less significant, the offset becomes negligible and the system returns to a power-law dependence. This behavior is not specific to the self-avoiding random coil, and as such we expected an approximately linear dependence of inferred distance on the number of GS repeats. Indeed, this linear dependence mirrors what we observed experimentally, providing confidence that our experimentally-derived distances are following expected trends given the physical nature of the setup.

A1.7 Calculation of χ

For each FRET construct in each solution condition, χ was calculated in three steps:

1. The mean FRET efficiency values for 24 replicates (in 4 repeats) each for linkers of 8, 16, 24, 32 and 48 GS repeats (16, 32, 48, 64 and 96 amino acids in length) in a buffer solution (20 mM NaH_2PO_4, 100 mM NaCl) were linearly fit to arrive at a relation between FRET efficiency in buffer to the number of amino acids (N) in the GS linker.

2. The resulting slope and y-intercept (shown in **Fig. 2.1B**) were used to interpolate an implied FRET efficiency (E_f^{GS}) for a GS linker of the same length N as the IDR of interest.

3. χ was then calculated as:

$$\chi = \frac{R_{ee}^{i}}{R_{ee}^{GS}} - 1 = \frac{R_0^i(1/E_f^i - 1)^{\frac{1}{6}}}{R_0^{GS}(1/E_f - 1)^{\frac{1}{6}}} - 1 = \frac{n^i(1/E_f - 1)^{\frac{1}{6}}}{n^{GS}(1/E_f^{GS} - 1)^{\frac{1}{6}}} - 1$$

where R_0 is the Förster distance, defined as the distance between the FPs at which E_f = 0.5, the superscript i indicates the IDR we are measuring, the superscript GS indicates a GS linker of length equivalent to that of IDR i, and n is the refractive index of the solution in which the IDR is measured. We have tried modulating the refractive index between 1.33 (for neat buffer) and 1.37 (the refractive index of 24 w/w% PEG10000)[7] and noticed no significant changes in the trends of our data, and an absolute change of < 5% in absolute values of χ. We therefore decided not to use this correction for the work presented in **Fig. 2.1** and **2.2**.

A1.8 Impact of macromolecular crowding

To assess the impact of macromolecular crowding, we computed the overlap concentration using the established scaling relationship for PEG derived by Devanand and Selser[8]. Specifically, this states that $R_g = 0.0215M^{0.583}$ where M is the PEG molecular weight and R_g is measured in nanometers. Using PEG-dependent R_g values we computed the overlap concentration in molars, first by computing the chain volume:

$$V_l = 1000\left(\frac{4\pi R_g^3}{3}\right)$$

where R_g is the radius of gyration in meters and V_l is the chain volume in liters. The overlap concentration is then defined as:

$$c^* = \frac{1}{V_l \times N_A}$$

where N_a is Avogadro's number and c* is the overlap concentration in moles per liter. We then made the approximation that molarity and molality are sufficiently close under the concentration regimes explored, allowing us to determine the overlap concentration in weight/weight (%).

In parallel, we computed the end-to-end distance of the shorter synthetic construct examined (GS8) using all-atom simulations in which the linker was allowed to move freely (see **Fig. A11** and **A12**). The radius of gyration of this system is approximately 4.6 nm. In comparison, the radius of gyration of the largest PEG used (PEG10000) is computed to be 4.6 nm. As such, in essential every scenario the crowder is equal to or smaller than the size of the protein reporter of interest.

For each protein, we assessed how χ varies as a function of PEG with the PEG-dependent overlap concentrations annotated (**Fig. A17-A18**). For GS-linker constructs, we observe a systematic drop in χ as a function of PEG concentration (**Fig. A17**). While this decrease becomes increasingly pronounced as a function of PEG molecular weight, there is minimal dependence on the number of GS repeats. Moreover, the overlap concentration does not represent an obvious threshold but instead demarks the beginning of a regime where a gradual drop in χ is observed as a function of concentration. For example, χ values for systems in which the PEG concentration is at 12% are relatively similar, regardless of whether the PEG concentration is above the overlap concentration (PEG1500) or far below the overlap concentration (PEG10000). The same cannot be said at higher PEG concentrations, however, where crowding-induced compaction affects longer GS linkers more substantially than shorter linkers, as expected.

For non-GS IDRs, more complex behavior is observed, notably sharper or weaker dependencies, depending on the sequence (**Fig. A18**). For example, p53 χ values show a shallow and essentially linear dependence on PEG concentration, where the χ-dependence becomes steeper as PEG becomes larger. In contrast, E1A χ values show a non-linear decrease, implying that E1A is substantially more sensitive to crowding-induced compaction than p53.

A2 Computational Methods

A2.1 All-atom simulations

All-atom Monte Carlo-based simulations were performed using the CAMPARI simulation suite, with the ABSINTH implicit solvent model[9]. In CAMPARI, the effective Hamiltonian is a combination of 4 energy terms:

$$E_{total} = W_{solv} + U_{LJ} + W_{el} + U_{corr}$$

Here U_{LJ} is the Lennard-Jones potential between protein residues, W_{el} is the electric potential term based on coulombic potential, U_{corr} is a term applied to the dihedral angles, and W_{solv} is a solution-protein interaction term based on the ABSINTH implicit solvent model[10], which is equivalent to a transfer free energy from a vacuum to a dilute aqueous solution.

Our solution space scanning method is carried out as described previously[11]. Briefly, the implicit solvation term, W_{solv}, is first calculated for each sequence based on its fully extended protein conformation. This represents the maximum transfer free energy (W_{solv}^{max}) since it is the most exposed configuration accessible to the protein. Solution space is then probed by modulating W_{solv}^{max} by changing the attraction/repulsion of different protein moieties in relation to the implicit solvent. We express the total strength of solution interaction by the parameter ψ where

$$\psi = \frac{W_{solv}^{max}(solution) - W_{solv}^{max}(water)}{W_{solv}^{max}(water)} \times 100\%$$

In this paper we change ψ by making interactions with the backbone less or more attractive (negative or positive ψ values, respectively). Our previous calibration based on helix-to-coil transition has shown that a 1 M urea solution is equivalent to $\psi \approx +1.2\%$.[11]

Our all-atom simulation dataset consists of 70 proteins (not including GS repeats). We selected sequences that were shown experimentally to be disordered, as collected on the DisProt server[12]. All sequences were simulated at 310 K with 10^7 steps of equilibration, followed by 7×10^7 steps of production. Conformations were written every 12,500 steps, resulting in a total of ~ 5,000 conformations for every simulation. Each sequence was simulated in five independent repeats, resulting in an ensemble containing 20,000 conformations per sequence. The MDtraj python library[13] was used to calculate the radius of gyration and end-to-end distance of the ensemble. Data from analyzed all-atom trajectories for each sequence is available in **Table S3**.

A2.2 Coarse-grained simulations

Our coarse-grained depiction of heteropolymer IDPs uses the PIMMS simulation framework[14]. PIMMS is a lattice-based Monte Carlo simulation engine in which inter-bead interactions are determined by nearest-neighbor interactions. All bead interactions are anisotropic along on-lattice and diagonal directions. The system evolves through a collection of moves that include individual crankshaft moves, chain translation/rotation, and chain pivot moves. For our purposes, residues are represented as beads, and a simple heteropolymer amino acid alphabet was used to generate chains of various lengths with a heteropolymeric distribution of residues that are similar to polar, hydrophobic, and charged amino acid residues. We emphasize that the parameters generated here, shown in **Fig. A5A**, are phenomenological and not meant to reflect specific amino acids. The set of PIMMS sequences used are available upon request.

The parameters chosen demonstrate sequence-specific coil-to-globule transitions, as shown in **Fig. A5B**. The simulation temperature was set to be units of k_BT. Accordingly, the total energy of the system in a given state is calculated based on a summation over pairwise interactions involving nearest-neighbor, non-bonded contacts, or a solution interaction in the case that no neighbors are present. Moves are accepted or rejected via a standard Metropolis criterion whereby the acceptance ratio is $\min\{1,\exp(-\Delta E/k_BT)\}$ where k_B = 1, ΔE is the energy difference between the current and proposed configurations, and T has the same units as the contact energies thus making the ratio $\Delta E/k_BT$ a dimensionless quantity. This conversion makes the point that the parameterized interactions that reproduce the observed experimental data are in fact relatively weak, being less than k_BT, depending of course on the simulation temperature.

For each chain length, 2000 sequences were randomly generated, and each sequence simulated in 10 solution interaction strengths (plus “buffer” condition) for a total of 11 trajectories per sequence. Each simulation consisted of a 20-step equilibration followed by a 1020-step production run at T = 70. Upon each step, tens to hundreds of thousands of individual Monte Carlo accept or reject moves are performed (on average 1000 local chain perturbations per bead in the chain per step). Simulation analysis was performed every 10 steps, and the reported distances are averages of the entire trajectory. Simulations were performed in a sufficiently large box to avoid finite-size effects. Average end-to-end distances vs. solution interaction strengths for the entire dataset are shown in **Fig. A5B**.

A2.3 Analytical model for the solution-driven coil-to-globule transition of a polymer

We developed a simple and generic analytical model to characterize the coil-to-globule transition of a homopolymer as assessed by a mean-field net inter-monomer interaction parameter. This model was then parameterized using homopolymeric PIMMS simulations performed for a range of chain lengths and interaction strengths to provide an analytical expression that relates the inter-monomer interaction strength to the degree of compaction/expansion as measured by the parameter χ . While we “parameterize” using PIMMS simulations, the simulations essentially tailor the model parameters to

reproduce the interaction strengths and dimensions as are native to PIMMS. In principle, any polymer model could be used to obtain key numerical parameters that dictate spatial and interaction features.

This model is built on the assumption that the coil-to-globule transition can be empirically mapped as a cooperative transition in which the cooperativity and midpoint show an exponential dependence on chain length, and the end-points reflect defined expected χ values for a flexible polymer in the globule (compact) or coil (expanded) limits. Specifically, we define χ as

$$\chi = a + b\left(\frac{1}{(m/e)^{\theta}}\right),$$

Where

$$\theta = c\log(L) + d$$

and

$$m = \gamma L^{\gamma}$$

$$a = \left(\frac{L^{0.33}}{L^{0.50}}\right) - 1$$

$$b = \left[\left(\frac{L^{0.59}}{L^{0.50}}\right) - 1\right] - a$$

The parameters in this model are defined as follows:

- *L* is chain length
- *e* is the apparent net inter-monomer interaction energy (measured in k_BT)
- *θ* is a measure of the cooperativity of the coil-to-globule transition, and itself depends logarithmically on *L* and two free parameters (*c* and *d*)
- *m* is a measure of the midpoint of the coil-to-globule transition and depends exponentially on chain length and one free parameter (γ)

The free parameters (*c*, *d*, and γ) are obtained by fitting to homopolymeric PIMMS simulations where χ is calculated directly from the simulations (**Fig. A13-A14**). The specific values for these three parameters will depend on the physical nature of the polymer model but do not ultimately influence the limiting behavior or trends of the model behavior, assuming they retain physically realistic values. These parameters depend on chain stiffness and monomer valence.

This model was chosen to provide a simple analytical description, under the simplifying assumption that chain solvent-dependence can, to a first-order approximation, be described using a simple homopolymer that expands/compacts as reported by χ. Chain-solvent interactions are captured in terms of an apparent intra-bead interaction parameter (*e*), which reports on the net favorable energy associated with monomer-monomer interaction in a given solution.

In the limit of a self-avoiding chain, the coil-to-globule transition is entirely determined by the chain-solvent interaction strength. In the limit of a chain where chain-solvent interactions are set to zero, the coil-to-globule transition is entirely determined by the monomer-monomer interaction strength. Real chains sit somewhere between these two limits, where both chain-chain and chain-solvent interactions contribute to the chain dimensions. Our model is formally parameterized in the non-interacting chain-solvent limit, but this can be recast as the non-interacting chain-chain limit in which the apparent chain-solvent interactions are defined as half the chain-chain interactions. In this way, we can write the coil-to-globule transition as a function of either chain-chain interactions or chain-solvent interactions, as is shown in **Fig. A15.** For simplicity, we have leveraged the chain-solvent representation, as most easily dovetails with our experimental work.

In its current format, the maximum chain expansion reflects the self-avoiding chain limit in which chain-solvent interactions are set to zero. Note that for polypeptides with charged residues, further expansion is possible via electrostatic repulsion[15]. These longer-range repulsive interactions are not captured by our analytical model nor by the model parameters used for our PIMMS simulations. However, they are evident in our all-atom simulations, offering an explanation as to why the χ axes for the all-atom simulations extend to substantially larger values than in either the theory or coarse-grained simulations.

A2.4 Converting from χ to ν^{app}

As in **Eq. 1** we define χ as

$$\chi_i = \frac{R_e^i}{R_e^{GS}} - 1$$

.

R_e can also be written as

$$R_e^i = BN^{\nu_i^{app}}$$

where *B* is a prefactor in units of distance, and the apparent scaling exponent (ν^{app}) is a measure of the apparent solvent quality for the chain[4,16]. In both our simulations and prior experiments, a GS linker in neat buffer behaves as a polymer in a theta solvent, a reference state in which chain-chain and chain-solvent interactions are counterbalanced, and where $\nu^{app} = 0.50$.[17]

Operating under this assumption, we can rewrite χ as:

$$\chi_i = \frac{BN^{\nu_i^{app}}}{BN^{0.5}} - 1$$

And more simply as

$$\chi = \frac{N^{\nu_i^{app}}}{N^{0.5}} - 1$$

As such, it is trivial to convert between χ and the apparent scaling exponent (ν^{app}) for the chain of a given length *N* in the limit of a homopolymer instantiation of our model under the simplifying assumption of a fixed, sequence-independent and ν-independent prefactor (*B*). For heteropolymers this assumption may not be valid, but as applied to our simple homopolymer model this is a reasonable set of approximations.

The major advantage of using χ over ν (or ν^{app}, as we have described here) reflects the fact that while ν is derived from polymer scaling theory, χ is simply a ratio whereby the denominator is some directly measurable reference state. ν has precise mathematical meaning in the context of analytical polymer physics. Unfortunately, this meaning frequently fails to hold true in the context of finite-sized heteropolymers, necessitating finite-sized corrections[18–22]. Moreover, approaches for calculating ν in finite-size polymers (leading to the apparent scaling exponent, ν^{app}) can be method-dependent due to necessary assumptions regarding the nature of the scaling prefactor, end-effects, heteropolymeric interactions, and the intrinsic uncertainty associated with finite-sized polymers[5,18,19,23–26]. Taken together, the application of scaling theory to finite-sized polymers can be misleading unless *bona fide* scaling behavior can be shown in terms of the dependence of global chain dimensions as a function of chain length over a sufficiently large number of long polymers[4,26,27]. In contrast, χ is simply a mathematical ratio of measured values. It imposes no assumptions other than the fact that the denominator reflects a reference value measured for a length-matched glycine-serine (GS) linker in aqueous (neat) buffer. Even the explicit polymeric behavior of the GS linker is relatively unimportant, although prior work has established that GS linkers behave in a manner at least qualitatively if not quantitatively as a flexible random coil[17,28].

In the context of our FRET-based assay, the application of homopolymer theory raises an additional challenge since our system is by definition outside of a regime in which homopolymer physics can be easily applied owing to the relative size of the fluorescent proteins compared to the disordered regions (**Fig. A11-A12**). Using χ allows us to bypass the clear limitations that making homopolymer-based assumptions would necessitate.

A3 Supplementary Figures

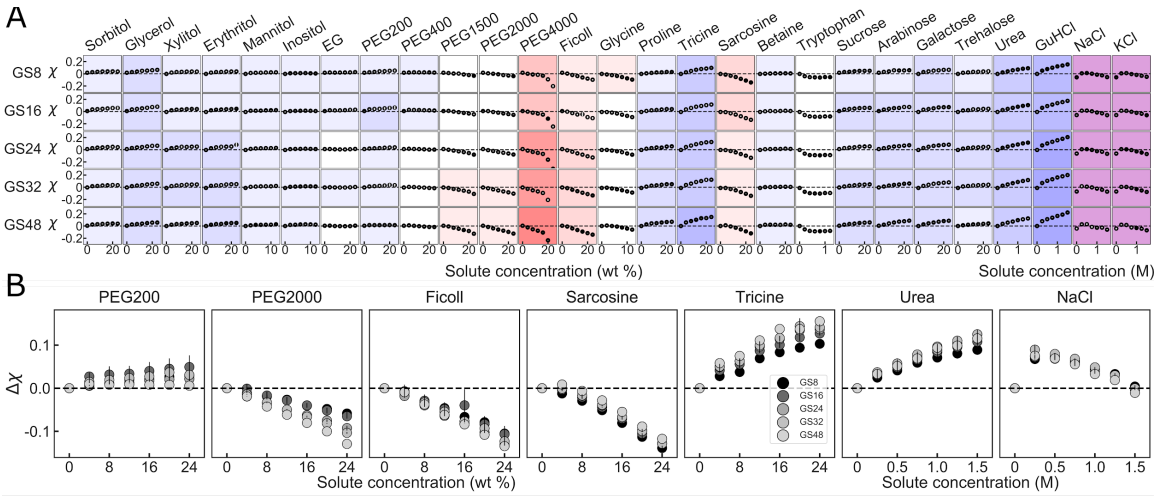


Figure A1. (**A**) Solution space scans of Gly-Ser linkers. Each data point shows the average χ vs. concentration of a specific solute for a specific IDR taken from two repeats. Vertical lines show the spread of repeats, and are often too small to see. IDRs vary down columns, and solutes vary along rows. Background color represents the sensitivity of change to solute addition: stronger colors imply higher sensitivity, red hues indicate compaction, and blue hues indicate expansion. Purple background indicates non-monotonic behavior. (**B**) Identical response of GS linkers to individual solutes contrasts with the differential response of other sequences shown in **Fig. 2.2B**. Each panel point is the average of the solution-induced change in χ vs. concentration of a specific solute and construct from two repeats. Vertical lines are the spread of the data.

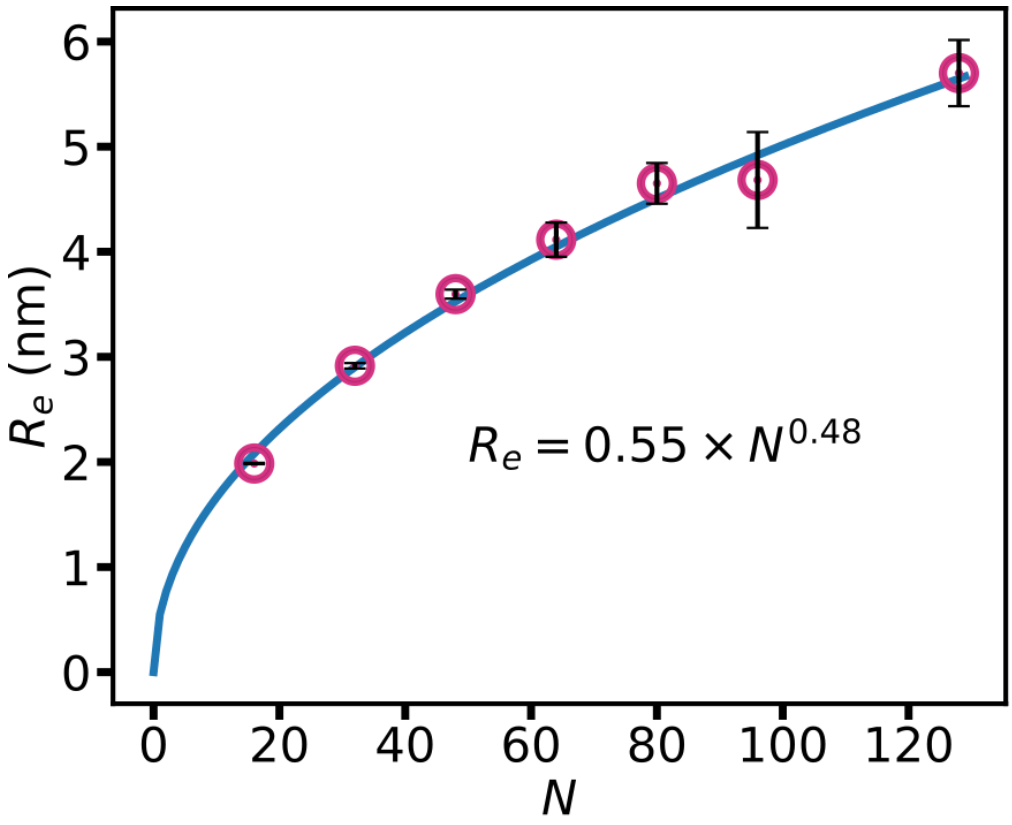


Figure A2. The end-to-end distance of Gly-Ser repeat sequences as a function of their total number of residues *N*, obtained from all-atom simulations in aqueous solution. Each data point is an average of five individual repeats, with lines being the standard deviation of the data. The blue curve is a power-law fit of the data, shown in the inset. The fitted exponent, 0.48 ± 0.03 is within error of the exponent expected of an ideal polymer (0.5).

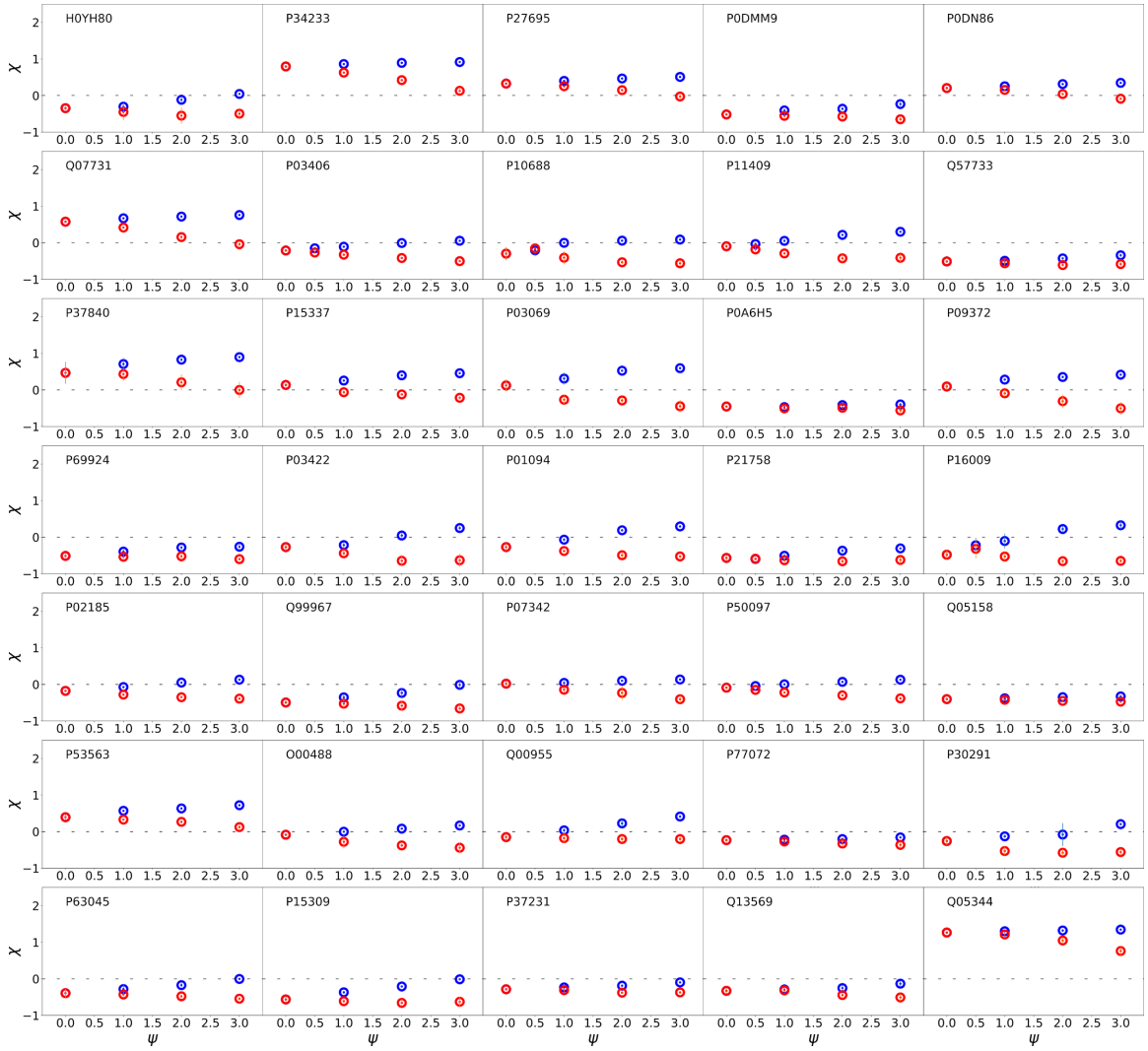


Figure A3. χ vs strength of solution interactions ψ (see **Section A2.1**) for each of the 70 IDRs shown in Fig. 3. Each subplot represents a single IDR. Blue points are attractive solutions ($\psi > 0$) and red points are repulsive solutions ($\psi < 0$). IDs are UniProt ID when available. All protein names, sequences, and data for each IDR are available in **Table S3**.

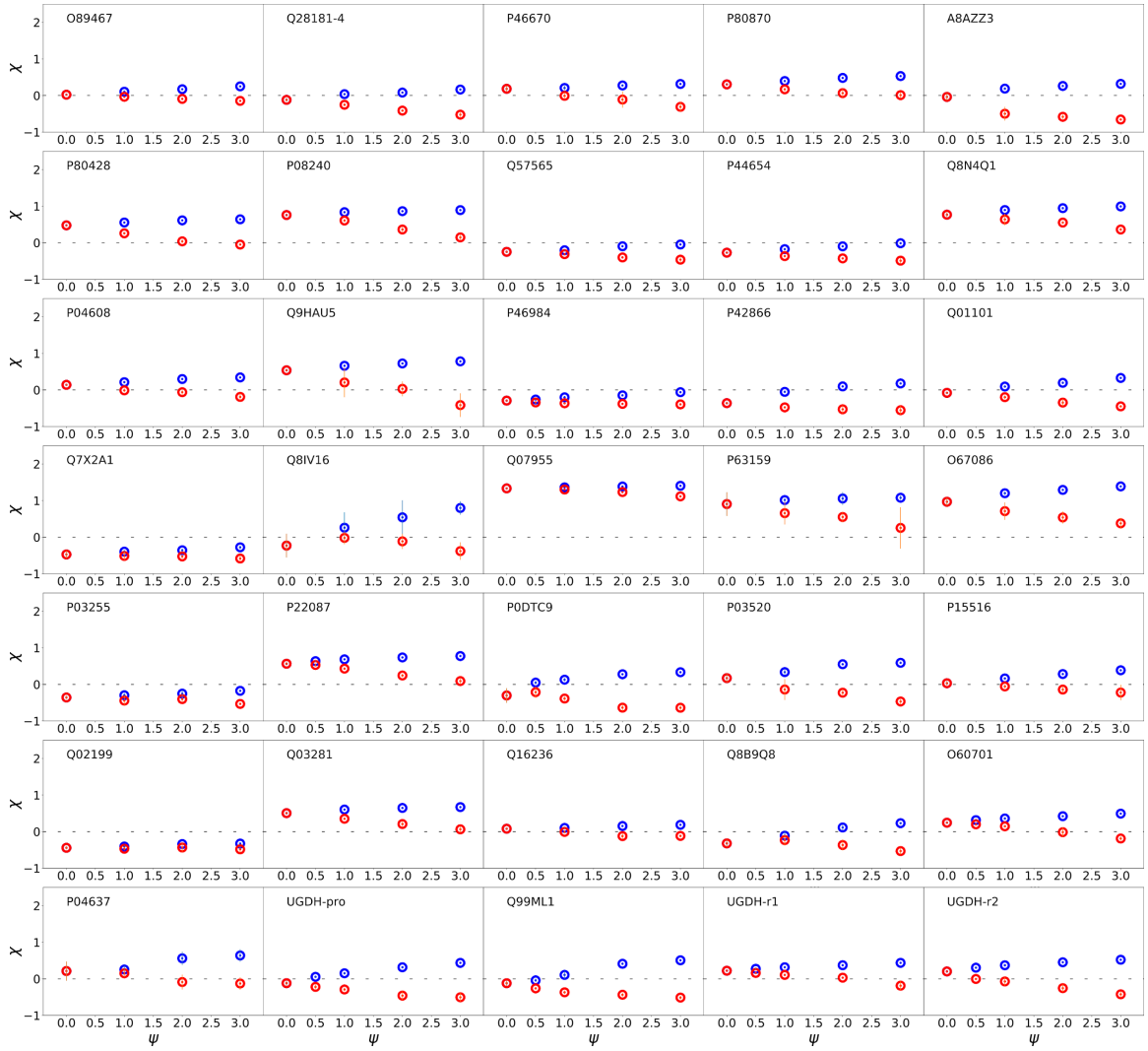


Figure A3 (cont.). χ vs. strength of solution interactions ψ (see **Section A2.1**) for each of the 70 IDRs shown in Fig. 3. Each subplot represents a single IDR. Blue points are attractive solutions ($\psi > 0$) and red points are repulsive solutions ($\psi < 0$). IDs are UniProt IDwhen available. All protein names, sequences, and data for each IDR are available in **Table S3**.

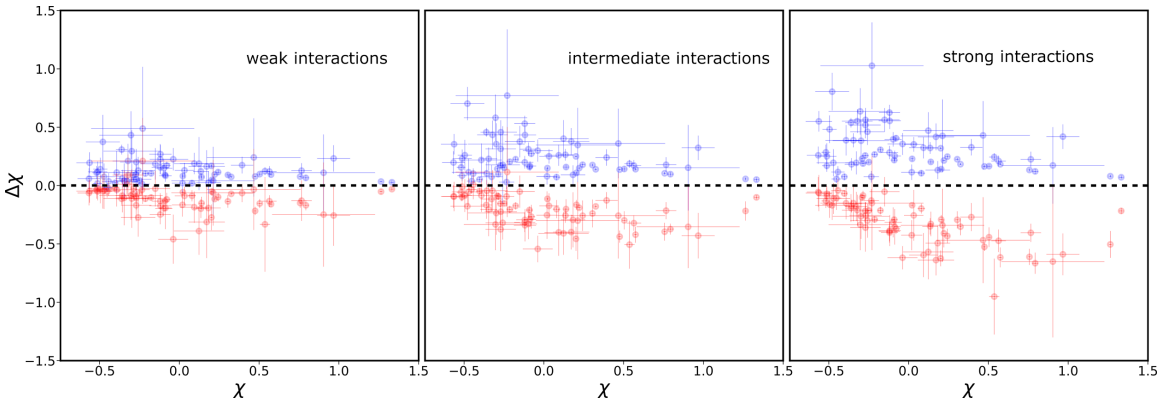


Figure A4. Solution sensitivity of IDRs shown in **Fig. A3**. Each point represents the solution-induced change in χ ($\Delta\chi$), for $\psi = \pm 1$ (weak interactions), ± 2 (intermediate interactions) or ± 3 (strong interactions). Blue points represent the response to repulsive solutions and red points represent the response to attractive solutions. Error bars are calculated from five independent simulations. See also **Fig. A6**.

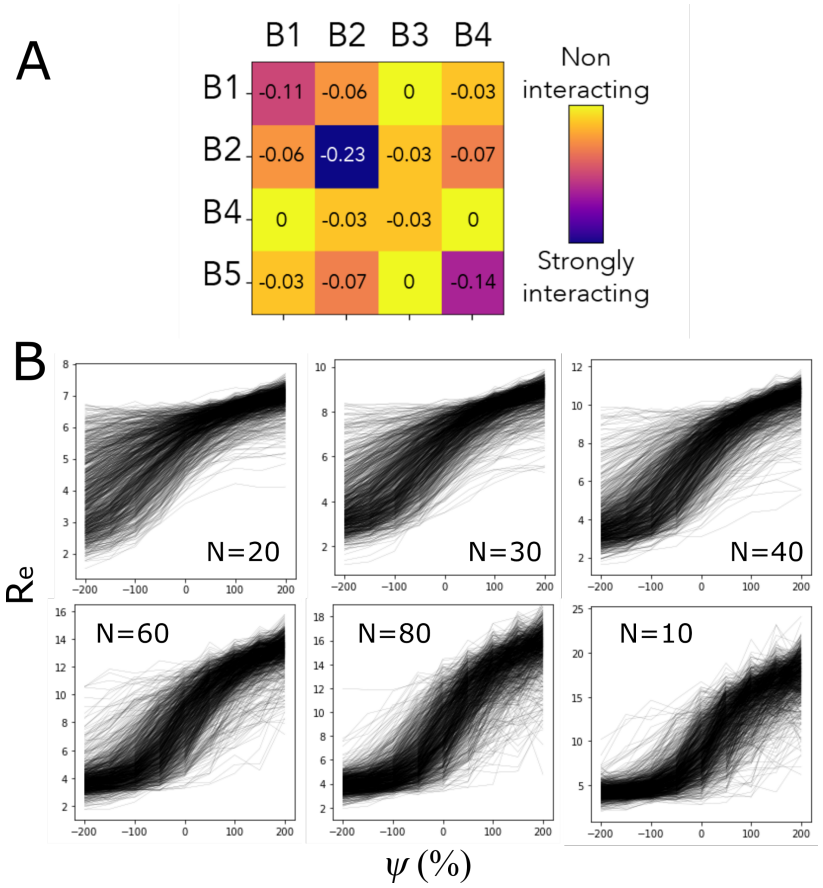


Figure A5. **(A)** Summary of the PIMMS parameters that were used for heteropolymer simulations. Interaction energies are defined in units of k_BT and were selected to approximate the chemical diversity observed in polypeptides. B1-B4 are "bead" 1 to "bead" 4. **(B)** End-to-end distances (in grid units) for PIMMS coarse-grained simulations of various sequences and chain lengths N. These curves were used to produce **Fig. 2.4B**.

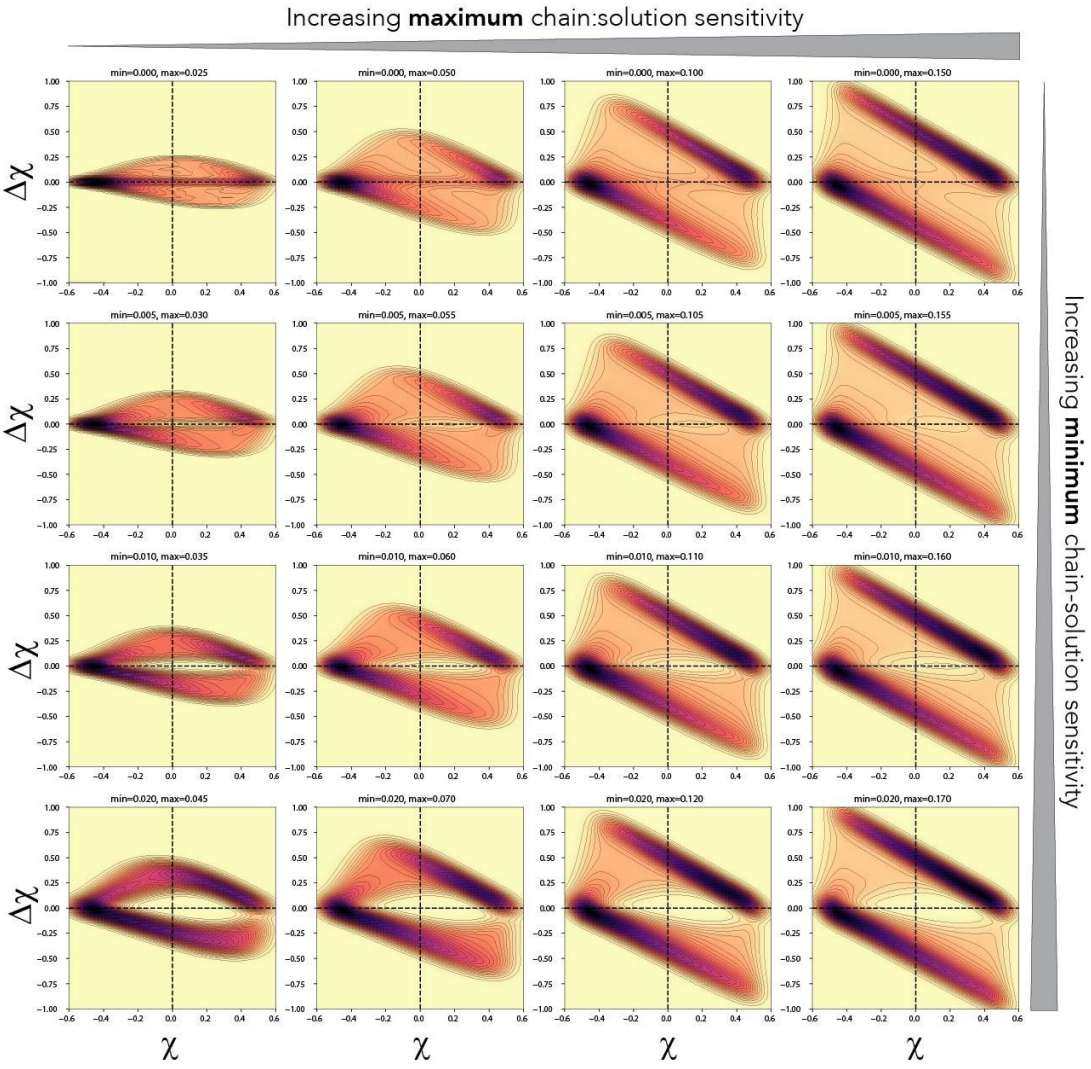


Figure A6. Dependence of Δχ vs. χ as a function of the most and least sensitive chains in an ensemble of sequences. Each figure defines the maximum and minimum perturbation to the chain-solvent interaction. As the maximum perturbation grows (left to right), Δχ becomes larger in a uniform manner along the χ axis. As the minimum perturbation grows (top to bottom), the opening of a central "pore" region emerges. These two phenomena can be understood intuitively. At the limit of the minimum perturbation being zero, this effectively means there exist chains that are fully insensitive to changes in the solution, such that Δχ is zero. As that minimum increases, *every* chain is somewhat sensitive, with a minimum sensitivity defined by this minimum value. Chains along the coil-to-globule transition are more sensitive than at the coil or globule limits (**Fig. 2.4D**) such that the pore is centered around χ = 0. The maximum perturbation defines the magnitude of Δχ, but is bounded by the chain dimensions, such that Δχ has upper and lower bounds. As the maximum is increased, more perturbations push up against that maximum, such that increasing Δχ density is observed at the bounds (*i.e.*, see top right).

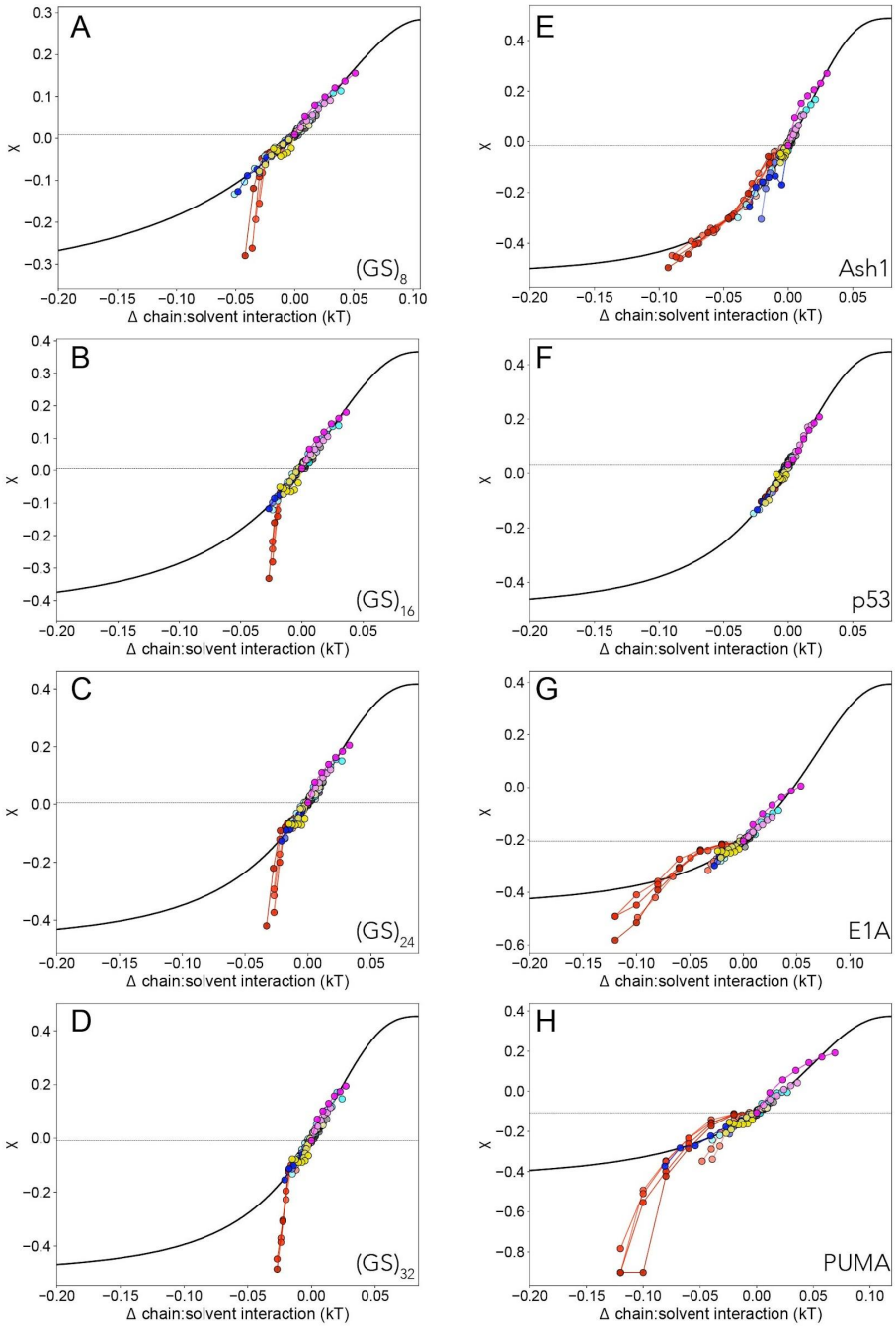


Figure A7. Full fit curves for all eight IDRs. Horizontal dashed lines reflect the χ value as measured in buffer. Black curve is a length-derived prediction from our analytic model. Note that for many of the curves the high-molecular weight PEG solutions lead to substantial deviations from the master curve, as expected as chain behavior enters the semidilute regime[29], the concentration regime in which PEG chains begin to overlap with one another. PUMA shows the worst agreement with the analytical model; a behavior interpreted as being due to its considerable residual helical structure.

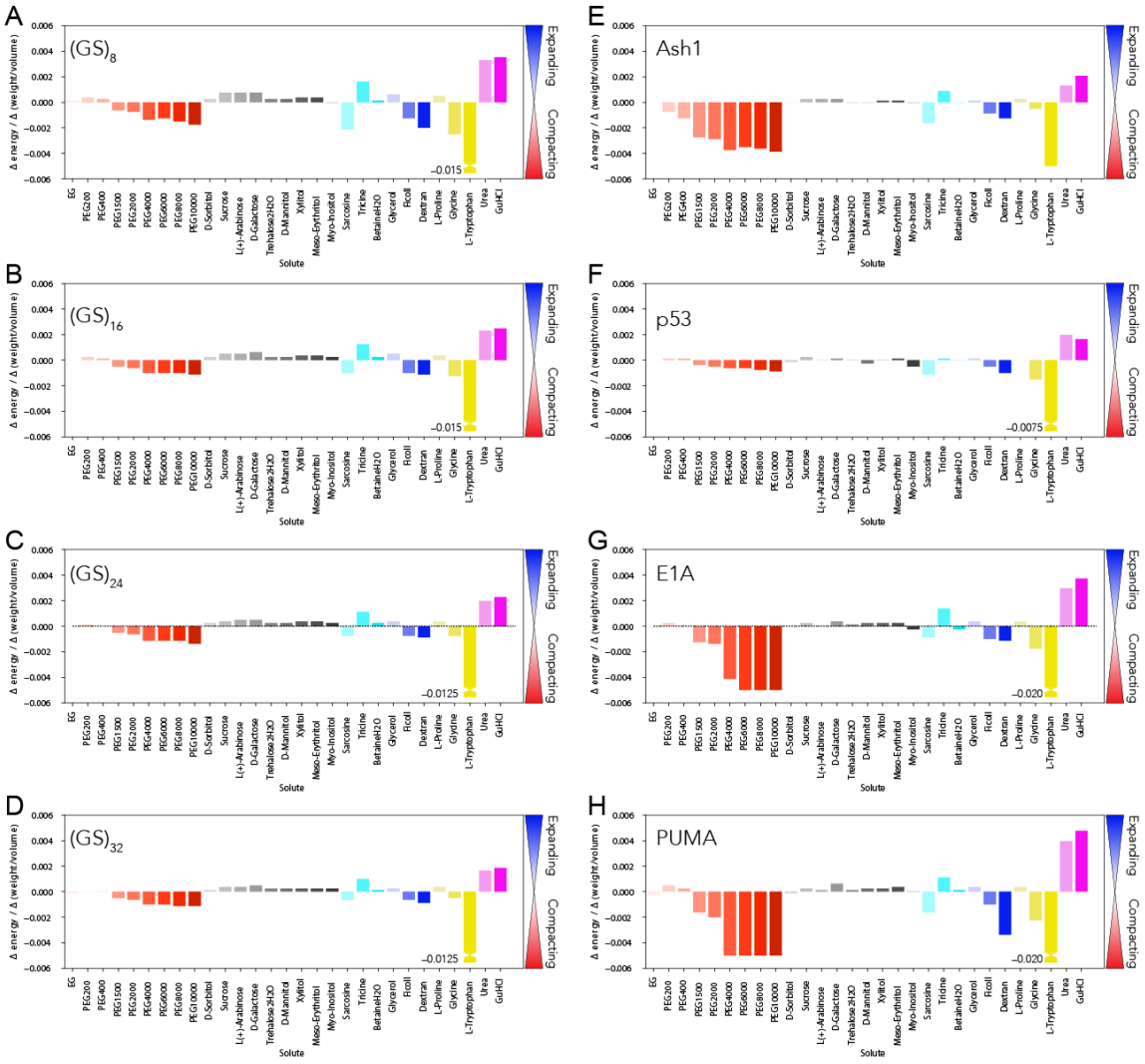


Figure A8. Derived solute-specific scalar factors that relate change in chain-solute interaction strength to a change in χ. More positive values lead to chain expansion while more negative values lead to chain compaction.

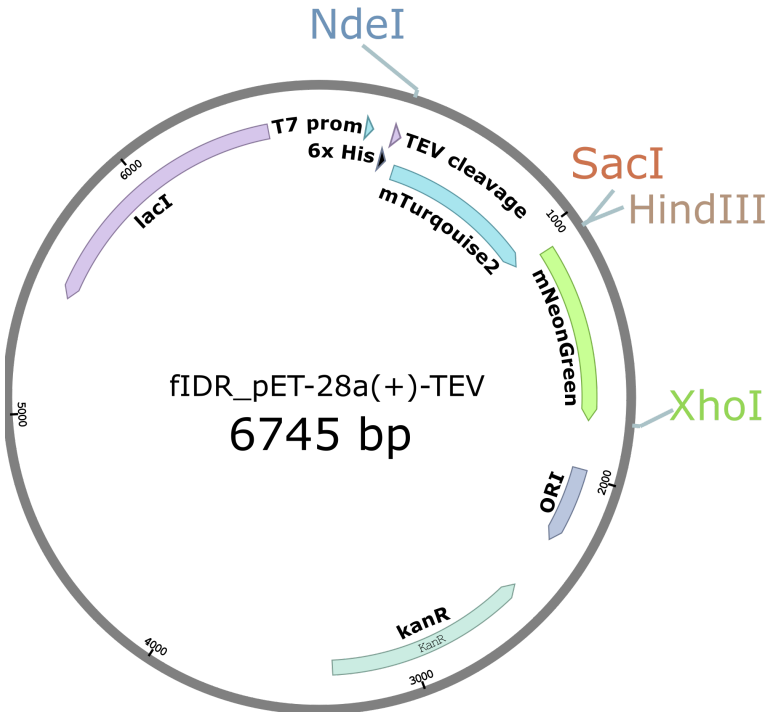


Figure A9. Plasmid map for FRET construct bacterial expression vector. Disordered sequences from **Table S4** are inserted between 5' SacI and 3' HindIII restriction sites.

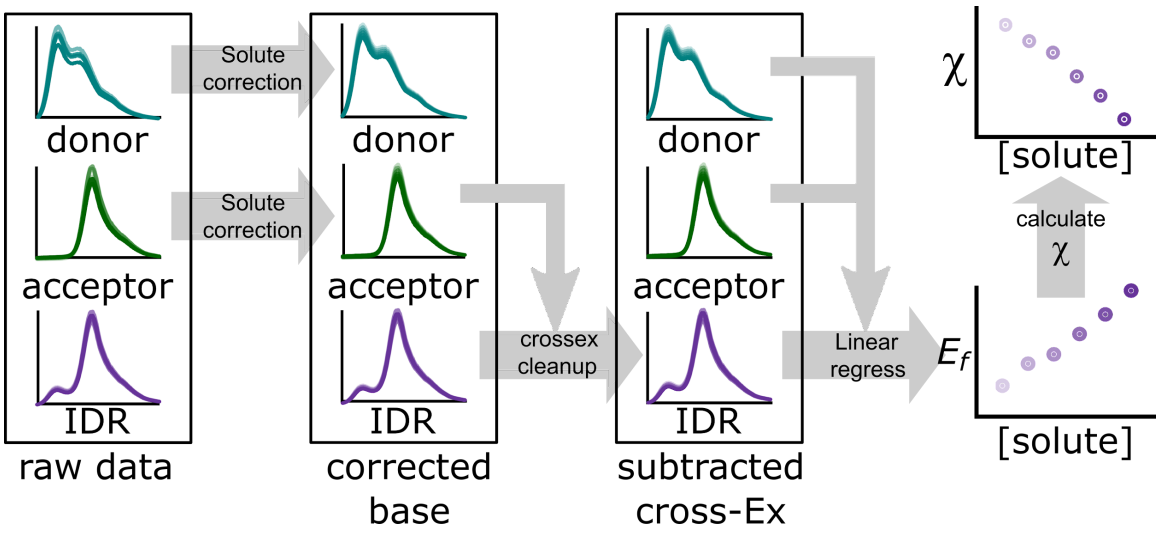


Figure A10. Visual summary of the data processing procedure detailed in **Appendix Section 1.5**. All panels show intensity vs. wavelength data for solutions containing donor-only, acceptor-only, and IDR construct (unless specified otherwise). Spectra are arranged from light to dark going from buffer to high concentrations of solute. Beginning from raw data, base spectra are corrected for pipetting error and protein absorbance to the plate to get corrected base spectra. The acceptor channel is then subtracted from the raw IDR data to remove cross-excitation artifacts. After this, both corrected base spectra are used to fit the corrected IDR spectrum by linear regression. Results of the linear regression are used to calculate the FRET efficiency, E_f, as described in **Appendix Section 1.5**, and E_f is used to calculate χ as described in **Appendix Section 1.7**.

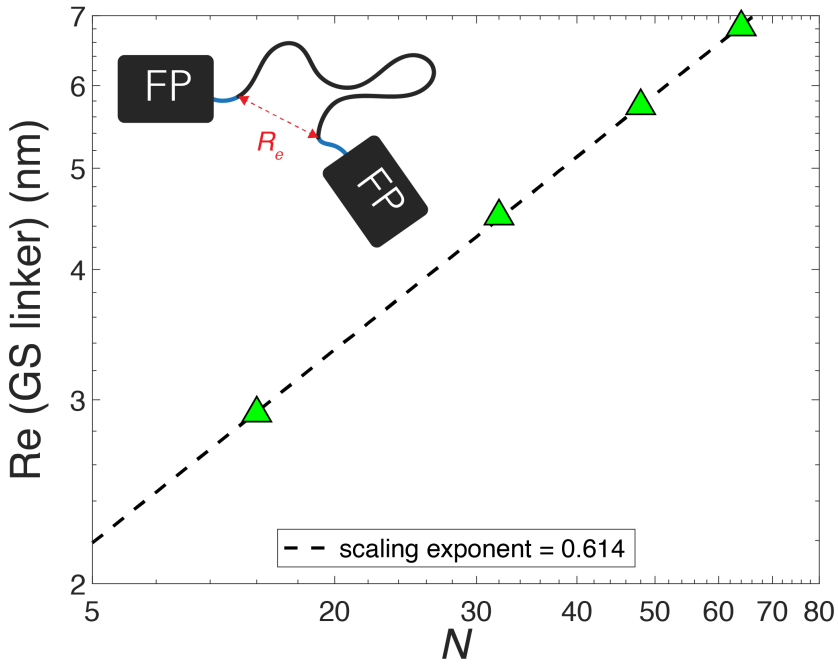


Figure A11. The intra-chain distance of glycine-serine (GS) linkers connecting two fluorescent proteins in a system that rigorously behaves as a self-avoiding random coil. GS linker end-to-end distance is measured between the first and last residue in the GS repeat region. Note that short (3-7 residue) cloning scars are also present in our model to replicate the actual experimental construct, and these do not contribute residues to the GS linkers in this analysis. Cloning scars are shown as teal parts of the linker.

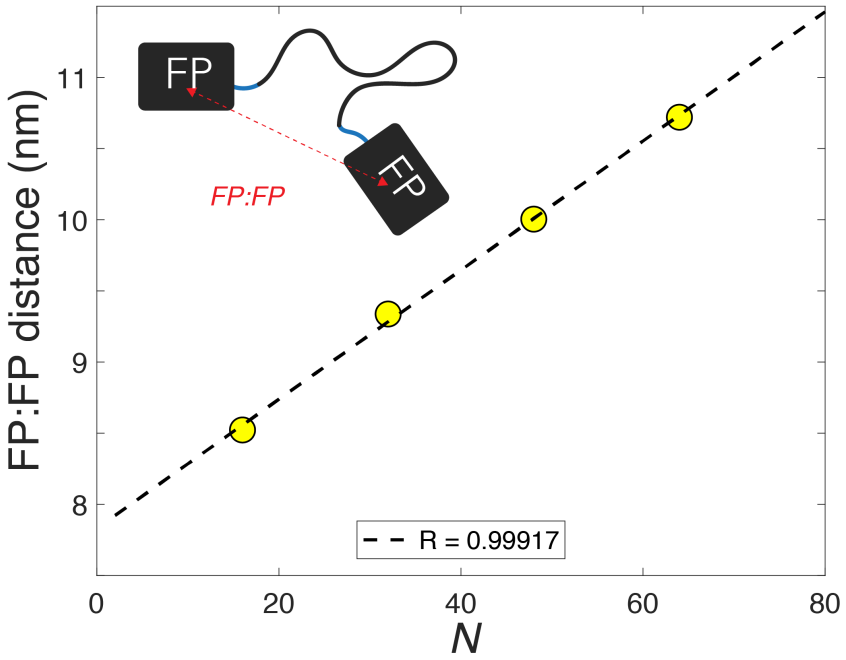


Figure A12. Intra-fluorescent protein distance for the same system as in **Fig. A11**. The distance here is measured between the two chromophore centers in each of the two fluorescent proteins. Note that when intra-fluorescent protein distances are measured, we obtain a linear relationship (as opposed to a power law relationship as in **Figs. A11** and **A2**).

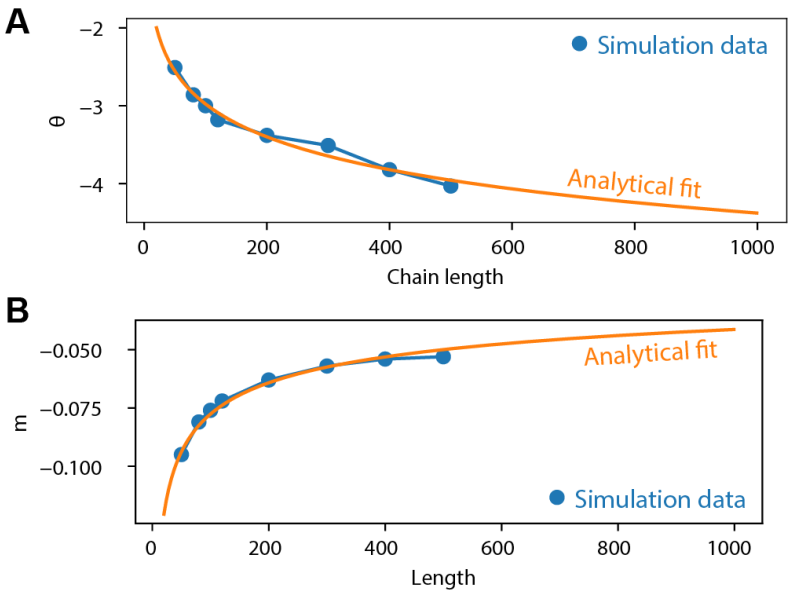


Figure A13. Fit of length-dependent model parameters to match PIMMS homopolymer simulations. The orange curves represent the analytical expressions defined in the **Methods** using the best fit parameters to fit to the experimentally measured values. **(A)** The fitting of the parameters *c* and *d* to reproduce the experimentally-derived length dependence of the cooperativity of the coil-to-globule transition, as quantified by *θ*. **(B)** The fitting of the parameter γ to reproduce the experimentally-derived length dependence of the midpoint on the coil-to-globule transition, as quantified by *m*.

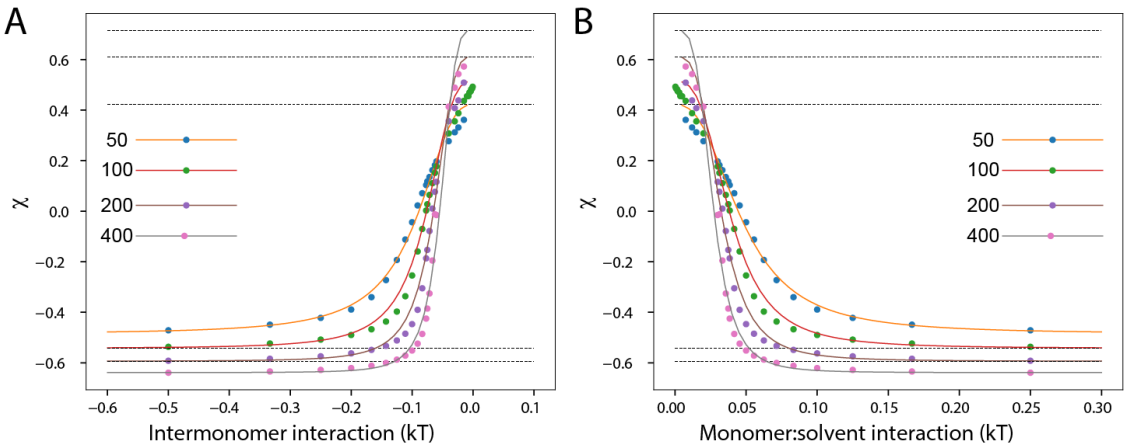


Figure A14. Best fit of floating parameters for analytical model (line) to PIMMS simulations (filled circles). (**A)** Data plotted in terms of inter-monomer interaction strength (assuming neutral chain-solvent interactions). **(B)** Same data plotted in terms of chain-solvent interaction strength (assuming neutral inter-monomer interactions). Colors denote chain lengths as specified in the legends.

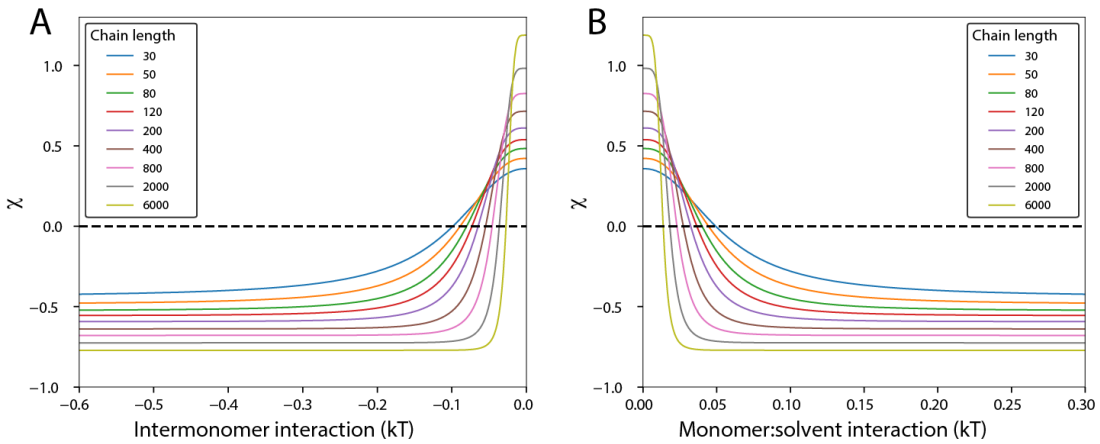


Figure A15. Relationship between inter-monomer interaction strength and χ. As chain length increases, cooperativity of the coil-to-globule transition increases. Note that the maximum and minimum χ values show a modest but well-defined length dependence. (**A)** Data plotted in terms of inter-monomer interaction strength (assuming neutral chain-solvent interactions). **(B)** Same data plotted in terms of chain-solvent interaction strength (assuming neutral inter-monomer interactions).

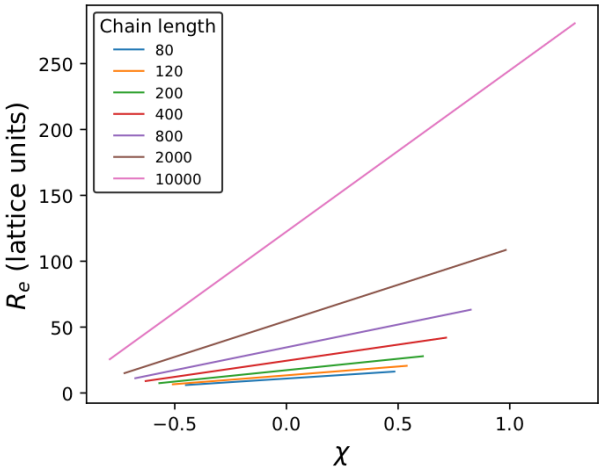


Figure A16. Dependence of the end-to-end distance (R_e) on χ. As the chain becomes longer, both the maximum and the steepness of the R_e-dependence on χ become larger.

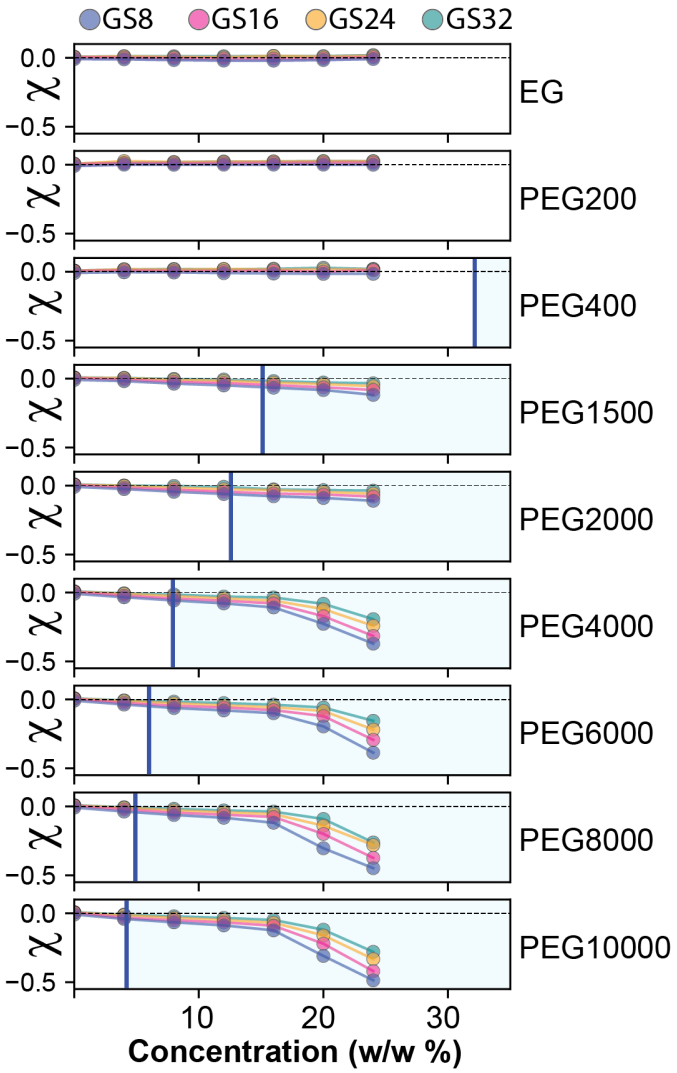


Figure A17: PEG-dependence of GS-linkers plotted on same axes. Blue line represents overlap concentration (c^*), with concentrations higher than c^* identified by the light-blue shaded regime.

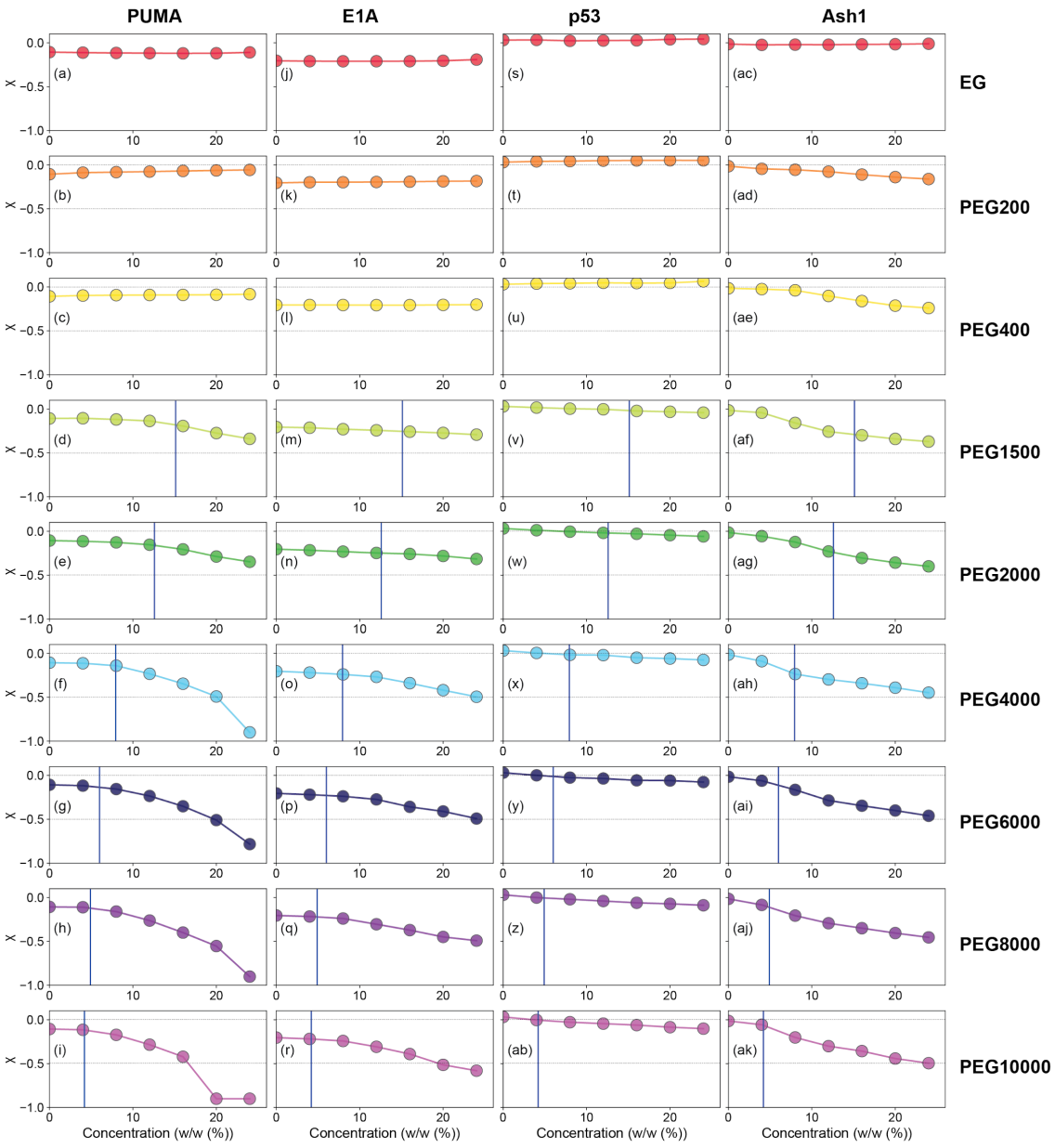


Figure A18: PEG-dependence of χ from naturally occurring IDRs as a function of PEG concentration and PEG molecular weight. Where present, horizontal lines are PEG-specific overlap concentration (c^*). Each of the four columns represents a distinct IDR and each row is a distinct PEG solution. **(a-i)** PEG-dependence of χ for PUMA. **(j-r)** PEG-dependence of χ for E1A. **(s-ab)** PEG-dependence of χ for p53. **(ac-ak)** PEG-dependence of χ for Ash1.

References

(1) Cranfill, P. J. *et al.* Quantitative assessment of fluorescent proteins. *Nat. Methods* **13**, 557–562 (2016).

(2) Lambert, T. J. FPbase: a community-editable fluorescent protein database. *Nat. Methods* **16**, 277–278 (2019).

(3) Mastop, M. *et al.* Characterization of a spectrally diverse set of fluorescent proteins as FRET acceptors for mTurquoise2. *Sci. Rep.* **7**, 11999 (2017).

(4) Rubinstein, M. & Colby, R. H. *Polymer Physics*. (Oxford University Press, 2003).

(5) Peran, I. *et al.* Unfolded states under folding conditions accommodate sequence-specific conformational preferences with random coil-like dimensions. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12301–12310 (2019).

(6) Holehouse, A. S., Garai, K., Lyle, N., Vitalis, A. & Pappu, R. V. Quantitative assessments of the distinct contributions of polypeptide backbone amides versus side chain groups to chain expansion via chemical denaturation. *J. Am. Chem. Soc.* **137**, 2984–2995 (2015).

(7) Mohsen-Nia, M., Modarress, H. & Rasa, H. Measurement and modeling of density, kinematic viscosity, and refractive index for poly (ethylene glycol) aqueous solution at different temperatures. *J. Chem. Eng. Data* **50**, 1662–1666 (2005).

(8) Devanand, K. & Selser, J. C. Asymptotic behavior and long-range interactions in aqueous solutions of poly(ethylene oxide). *Macromolecules* **24**, 5943–5947 (1991).

(9) Vitalis, A. & Pappu, R. V. ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* **30**, 673–699 (2009).

(10) Mittal, A., Das, R., Vitalis, A. & Pappu, R. ABSINTH Implicit Solvation Model and Force Field Paradigm for Use in Simulations of Intrinsically Disordered Proteins. *Computational Approaches to Protein Dynamics: From Quantum to Coarse-Grained Methods* 181 (2014).

(11) Holehouse, A. S. & Sukenik, S. Controlling Structural Bias in Intrinsically Disordered Proteins Using Solution Space Scanning. *J. Chem. Theory Comput.* (2020). doi:10.1021/acs.jctc.9b00604.

(12) Piovesan, D. *et al.* DisProt 7.0: A major update of the database of disordered proteins. *Nucleic Acids Res.* **45**, D219–D227 (2017).

(13) McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528–1532 (2015).

(14) Martin, E. W. *et al.* Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367**, 694–699 (2020).

(15) Hofmann, H. *et al.* Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proceedings of the National Academy of Sciences* **109**, 16155–16160 (2012).

(16) Holehouse, A. S. & Pappu, R. V. Collapse Transitions of Proteins and the Interplay Among Backbone, Sidechain, and Solvent Interactions. *Annu. Rev. Biophys.* **47**, 19–39 (2018).

(17) Sørensen, C. S. & Kjaergaard, M. Effective concentrations enforced by intrinsically disordered linkers are governed by polymer physics. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 23124–23131 (2019).

(18) Gomes, G.-N. W. *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).
(19) Song, J., Gomes, G.-N., Shi, T., Gradinaru, C. C. & Chan, H. S. Conformational heterogeneity and FRET data interpretation for dimensions of unfolded proteins. *Biophys. J.* **113** (2017).
(20) Zheng, W. *et al.* Inferring properties of disordered chains from FRET transfer efficiencies. *J. Chem. Phys.* **148**, 123329 (2018).
(21) Fuertes, G. *et al.* Comment on 'Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water'. *Science* **361** (2018).
(22) Riback, J. A. *et al.* Response to Comment on 'Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water'. *Science* **361** (2018).
(23) Stenzoski, N. E. *et al.* The Cold-Unfolded State Is Expanded but Contains Long- and Medium-Range Contacts and Is Poorly Described by Homopolymer Models. *Biochemistry* **59**, 3290–3299 (2020).
(24) Song, J., Gomes, G.-N., Gradinaru, C. C. & Chan, H. S. An Adequate Account of Excluded Volume Is Necessary To Infer Compactness and Asphericity of Disordered Proteins by Forster Resonance Energy Transfer. *J. Phys. Chem. B* **119**, 15191–15202 (2015).
(25) Meng, W., Lyle, N., Luan, B., Raleigh, D. P. & Pappu, R. V. Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2123–2128 (2013).
(26) Thirumalai, D., Samanta, H. S., Maity, H. & Reddy, G. Universal Nature of Collapsibility in the Context of Protein Folding and Evolution. *Trends Biochem. Sci.* **44**, 675–687 (2019).
(27) de Gennes, P. G. *Scaling concepts in polymer physics*. (Cornell University Press, 1979).
(28) Dyla, M. & Kjaergaard, M. Intrinsically disordered linkers control tethered kinases via effective concentration. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 21413–21419 (2020).
(29) Kozer, N., Kuttner, Y. Y., Haran, G. & Schreiber, G. Protein-Protein Association in Polymer Solutions: From Dilute to Semidilute to Concentrated. *Biophys. J.* **92**, 2139–2149 (2007)**.**

Appendix B

Supporting Information for Publication:

Structural Preferences Shape the Entropic Force of Disordered Protein Ensembles

The material originally appeared in the following: Feng Yu and Shahar Sukenik. The Journal of Physical Chemistry B 2023 127 (19), 4235-4244

Table S1 can be found at this link:
https://github.com/sukeniklab/Entropic_Force

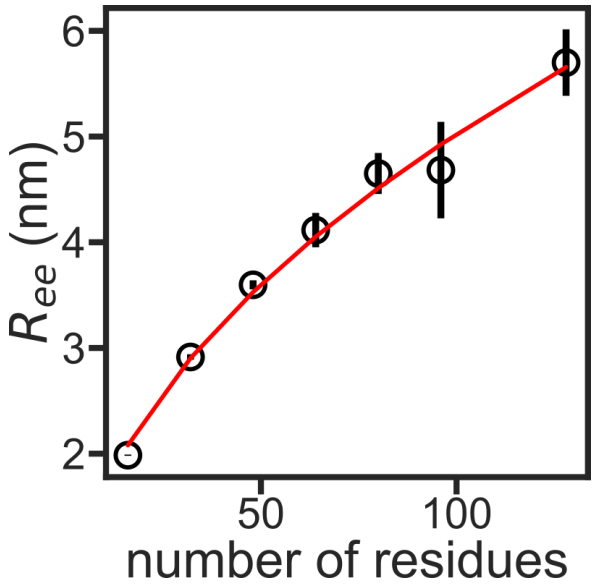


Figure B1. GS repeats match homopolymer scaling law under buffer conditions. The average end-to-end distance from five repeats vs the total number of residues for a series of Gly-Ser repeats, The error bars are the standard deviation of the five repeats. The red curve is the result of fitting to $R_{ee} = R_0 N^{\nu}$, with $R_0 = 0.55 \pm 0.06$ nm and $\nu = 0.48 \pm 0.03$. Errors are obtained from the fit.

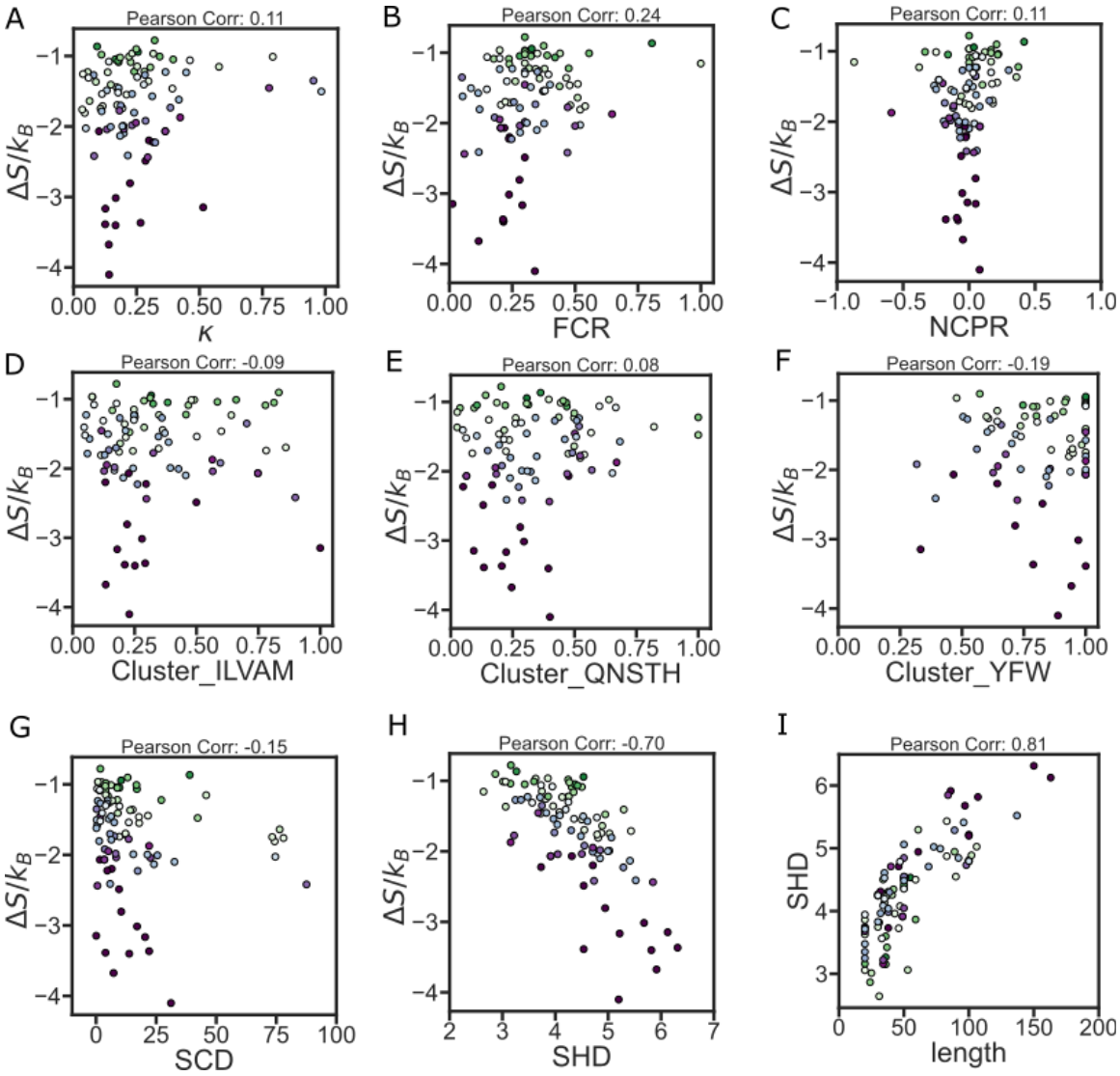


Figure B2. Sequence features correlated with the entropic force strength. The entropic force strength is plotted vs several sequence features. Unless stated otherwise, all sequence features are calculated using the localCIDER python package. (**A**) κ: a metric for mixing of charged amino acids[18] (**B**) FCR: fraction of charged residues, (**C**) NCPR: net charge per residue, (**D**) Cluster_ILVAM: hydrophobic amino acid mixing calculated using the same algorithm as κ, (**E**) Cluster_QNSTH: polar amino acid mixing calculated using the same algorithm as κ, (**F**) Cluster_YFW, aromatic amino acid mixing calculated using the same algorithm as κ. (**G**) SCD: sequence charge decoration (**H**) SHD: sequence hydropathy decoration (**I**) SHD vs sequence length shows a strong correlation with the sequence length. This may explain at least some of the correlation with entropic force, which is also shown to correlate with sequence length (**Fig. 3.5A**).

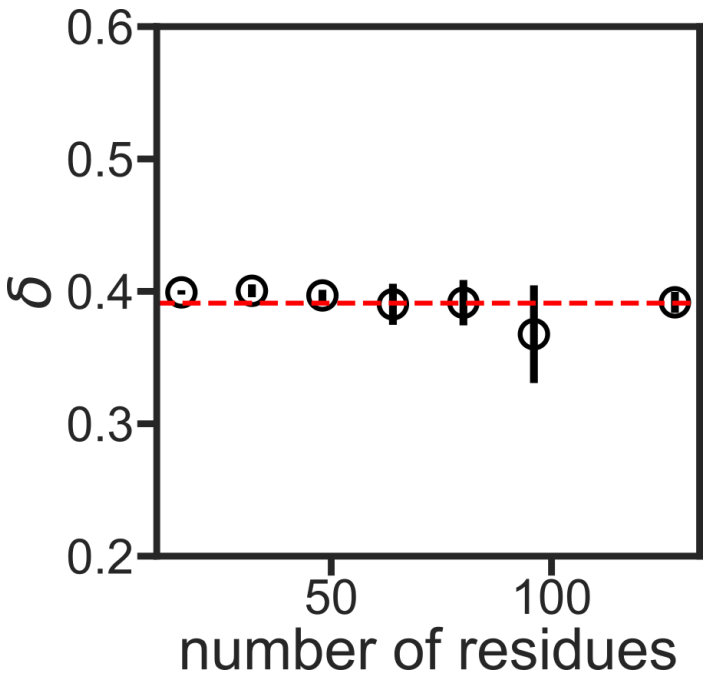


Figure B3. GS repeat asphericity is independent of length. The average asphericity of GS repeats vs the number of residues in the sequence. The mean of all seven data points is shown by the red line, with $\delta = 0.39 \pm 0.01$.

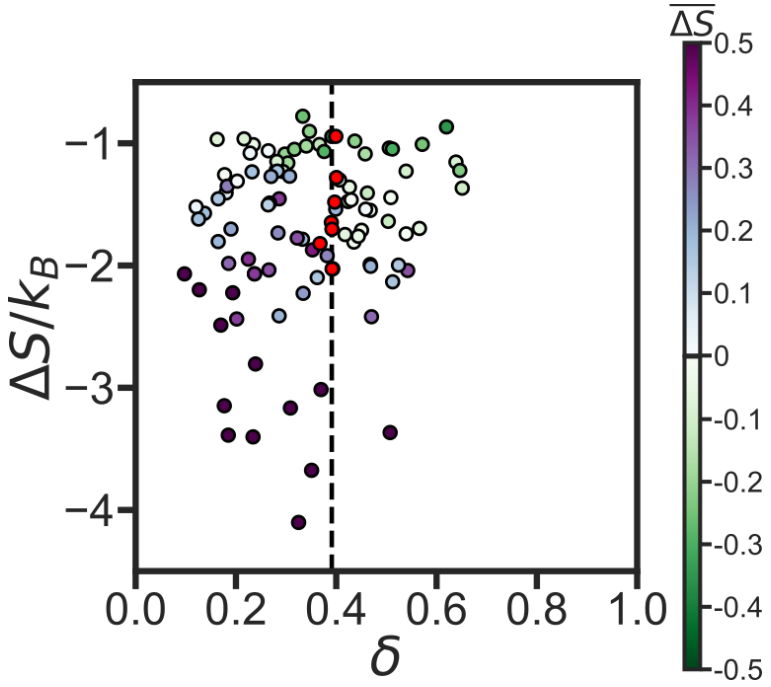


Figure B4. Entropic force as a function of average asphericity. The black line represents the length-independent asphericity of GS repeats shown in **Fig. B3**. Each marker represents a single IDR, color-coded as in **Fig. 3.5A** , with stronger purple (green) markers showing a stronger (weaker) entropic force compared to the GS repeat of the same size.

Appendix C

Supporting Information for Chapter 4 Results:

The material originally appeared in the following: Cuevas-Velazquez, C.L., Vellosillo, T., Guadalupe, K., Schmit, B., Yu, F., et al. Intrinsically disordered protein biosensor tracks the physical-chemical effects of osmotic stress on cells. Nat Commun 12, 5438 (2021).

Methods

The all-atom simulation and structural analysis are done by the author of this dissertation. The protein construction and FRET measurement are done by the collaborators.

All-atom simulation

Simulations of the AtLEA4-5 protein, its scrambles, and other IDRs were done using Solution Space Scanning, an all-atom Monte Carlo simulation method based on the ABSINTH force field[1], that has been previously described[2]. Briefly, the Hamiltonian function to be evaluated in each step can be written as the following representation.

$$E_{total} = W_{solv} + U_{LJ} + W_{el} + U_{corr}$$

Here, W_{solv} is the energy describing the interaction between the protein surface and the surrounding solution. By changing the W_{solv} term, we can alter this interaction and sample a protein's conformations in different solution conditions.

For each combination of solution condition and protein (AtLEA4-5 and each of its sequence scrambles), we ran five independent simulations consisting of 5×10^7 Monte Carlo steps (following 1×10^7 steps of equilibration) starting from random conformations to ensure proper sampling. Protein conformations were written out every 12,500 steps. The dataset of 70 other IDRs shown in Fig. (need figure number) was obtained using the same methods, is publicly available on https://github.com/sukeniklab/HiddenSensitivity, and has been previously described[3]. We analyzed the average radii of gyration of the simulated conformation ensembles using the MDTraj Python library[4]. Standard deviations were calculated based on the average of five individual repeats. Each radius of gyration was then normalized based on the most expanding solution to highlight solution sensitivity.

Transgene constructs

pDRFLIP38 backbone was used for biosensors yeast expression47. This plasmid contains the constitutive promoter pPMA1, and was provided by Dr. Alexander M. Jones. The vector was digested with XbaI (NEB) and EcoRI (NEB) to clone the open reading frames (ORFs) of mCreluan3, AtLEA4-5, and Citrine downstream of the pPMA1 promoter. The biosensor construct was cloned using the Gibson Assembly cloning method (NEB) by mixing the XbaI-EcoRI-digested pDRFLIP38 with the PCR-amplified

ORFs containing overlapping ends. The ORFs of the other fluorescent proteins (t7.eCFP.t9, Aphrodite.t9, t7.TFP.t9, mTFP.t9, Cerulean, edCerulean, edCitrine, edAphrodite.t9) used in this study were cloned in the same way. The sensory domains tested (AtLEA4-2, ABP, CS, N-AtLEA4-5, C-AtLEA4-5, Scramble-1, Scramble-2, Scramble-3, Scramble-4, Scramble-5) were cloned between mCerulean3 and Citrine ORFs. To do this, pDRFLIP38-AtLEA4-5 was digested with SacI and BglII to remove the AtLEA4-5 ORF. The digested plasmid was mixed with the different PCR-amplified sensory domains-ORFs containing overlapping ends using the Gibson Assembly method (NEB). AtLEA4-2, AtLEA4-5, N-AtLEA4-5, and C-AtLEA4-5 ORFs were amplified from pTrc99A-AtLEA4-2 and pTrc99A-AtLEA4-5 plasmids provided by Dr. Alejandra A. Covarrubias25. ABP ORF was amplified from pGW1araF.Ec plasmid provided by Dr. Wolf B. Frommer26. CS ORF was amplified from Cr1-pRSET-A provided by Dr. Arnold Boersma14. Scrambled versions were randomly designed using the Scrambler tool of PeptideNexus (https://peptidenexus.com/). Scrambles were chosen based on disorder propensity, α-helix prediction (AGADIR web server http://agadir.crg.es/) and charge mixing (Kappa value)27. All AtLEA4-5 Scrambled ORFs were synthesized as gene fragments (GenScript).

For bacterial expression, the pDEST-HisMBP backbone was used (Addgene #11085). This plasmid contains the Tac IPTG-inducible promoter for protein expression with an N-terminal 6x His tag and an MBP tag. The full SED1 ORF was cloned into pENTR-D-TOPO (Thermo Fisher Scientific). Recombination of pENTR-D-TOPO-SED1 and pDEST-HisMBP was done using Gateway technology to produce pDEST-HisMBP-SED1. The same strategy was followed for the full CS ORF to produce pDEST-HisMBP-CS.

Transgene expression

The constructs indicated in the main text were transformed into Saccharomyces cerevisiae protease-deficient yeast strain (BJ5465 lacking Pep4 and Prb1) using the lithium acetate transformation method49. Transformed colonies were selected in plates containing 6.8 g/L YNB media (Sigma-Aldrich) supplemented with 5 g/L glucose and 1.92 g/L synthetic drop-out medium without uracil (Sigma-Aldrich). Positive clones were confirmed by colony PCR. SED1 was also transformed into wild-type and hog1Δ::G418 and pbs2Δ::G418 mutant backgrounds of the Saccharomyces cerevisiae BY4742 strain (provided by Dr. Hugo Tapia). Transformation and selection were done as described above.

pDEST-HisMBP-SED1 was transformed into *Escherichia coli* BL21 (DE3) strain using the standard heat shock transformation protocol. Transformed colonies were selected in plates containing LB media supplemented with ampicillin (100 µg/mL). Positive clones were confirmed by colony PCR. The same strategy was followed for pDEST-HisMBP-CS.

Fluorescence analysis of live *Escherichia coli* cells

3 mL of SED1-expressing *Escherichia coli* culture was grown at 37 °C in liquid LB supplemented with ampicillin to OD600 ~1–2. No IPTG induction was needed since the fluorescence obtained from the leaking expression of the Tac promoter was sufficient for

measurements. Cells were centrifuged and washed twice with 50 mM MES, pH 6, and resuspended in 3 mL of the same buffer. 50 µL of the cell suspension was loaded into individual wells of a 96-well black F-bottom clear microplate (Greiner). 150 µL of treatment solution (see Chapter 4 main text) was added to the cell suspension, mixing was performed by pipetting up and down, and the fluorescence was measured immediately after. Fluorescence readings were acquired using a Safire fluorimeter (Tecan) for donor fluorophore (mCerulean3 excitation 433 nm, mCerulean3 emission 480 nm, abbreviated DxDm), acceptor fluorophore (Citrine excitation 510 nm, Citrine emission 525 nm, abbreviated AxAm), and energy transfer from donor to acceptor (mCerulean3 excitation 433 nm, Citrine emission 525 nm, abbreviated DxAm). Fluorescence emission scans from 460 nm to 550 nm (step size 5 nm) with an excitation wavelength of 433 nm were acquired. For all fluorescence measurements, the bandwidth was set to 7.5 nm, the number of flashes was 10, the integration time was 40 µs, and the gain was 100. Three independent measurements were acquired for each treatment and construct.

Fluorescence analysis of live Saccharomyces cerevisiae cells

5 mL of yeast cells expressing the indicated constructs (see main text) were grown at 30 °C in liquid YNB media (6.8 g/L) (Sigma-Aldrich) supplemented with 5 g/L glucose and 1.92 g/L synthetic drop-out medium without uracil (Sigma-Aldrich) until OD600 ~ 1–2. Cells were centrifuged and washed twice with 50 mM MES, pH 6 and resuspended in 5 mL of the same buffer. 50 µL of the cell suspension was loaded into individual wells of a 96-well black F-bottom clear microplate (Greiner). 150 µL of treatment solution (see main text) was added to the cell suspension, mixing was performed by pipetting up and down, and the fluorescence was measured immediately after. Fluorescence readings were acquired using a Safire fluorimeter (Tecan) for donor fluorophore (mCerulean3 excitation 433 nm, mCerulean3 emission 480 nm, abbreviated DxDm), acceptor fluorophore (Citrine excitation 510 nm, Citrine emission 525 nm, abbreviated AxAm), and energy transfer from donor to acceptor (mCerulean3 excitation 433 nm, Citrine emission 525 nm, abbreviated DxAm). Fluorescence emission scans from 460 nm to 550 nm (step size 5 nm) with an excitation wavelength of 433 nm were acquired. For all fluorescence measurements, bandwidth was set to 5 nm (7.5 nm for the emission scan), number of flashes was 10, integration time was 40 µs, and gain was 100. For time course measurements, the 96-well plate was kept inside the plate reader for the duration of the experiment. Measurements were acquired every 60 s for a period of 120 to 150 min. Shake (linear) duration was set to 3 s before every measurement. Nine independent measurements were acquired for each treatment and construct. Experiments were repeated three times.

U-2 OS cell culture

All U-2 OS (ATCC HTB-96) and HEK-293T (ATCC CRL-3216) cell lines used in this study were cultured at 37 °C in 5% CO2 in high-glucose DMEM (GE Healthcare) supplemented with 10% FBS (Atlanta Biologicals), 1 mM sodium pyruvate (Gibco), 2 mM L-glutamine (Gemini Biosciences), 1x MEM non-essential amino acids (Gibco), 40 U/ml penicillin and 40 µg/ml streptomycin (Gemini Biosciences). Stable U-2 OS SED1-expressing cell lines were generated by lentiviral transduction. To produce

lentiviral particles, the SED1 construct was first subcloned into EcoRV-HF (NEB)-digested pLenti-CMV Puro DEST (Addgene #17452) using the NEBuilder HiFi DNA Assembly master mix (NEB), and then transfected into HEK-293T cells together with pMD2.G (Addgene #12259) and psPAX2 (Addgene #12260). Virus was harvested 48 h after transfection, filtered through non-binding 45 µm syringe filters (Pall Corporation) and used to transduce U-2 OS cells. After 24 h, the virus-containing medium was removed and replaced with selection medium containing 2 µg/ml Puromycin (Sigma–Aldrich). After 7 days of selection, single-cell clones were derived by sorting for the top ~60% fluorescent cells using a Sony SH800 flow cytometer. Two individual clones were randomly selected for further use.

U-2 OS sample preparation

U-2 OS cells expressing SED1 were cultured in Corning treated flasks with Dulbecco's modified Eagle's medium (DME:F-12 1X from Hyclone Cat No SH30023.01) supplemented with 10% FBS (Gibco REF 16000-044) and 1% penicillin/streptomycin (Gibco REF 15140-122). Cells were incubated at 37 °C and 5% CO2. Sorbitol (VWR CAS 50-70-4) and NaCl (Fisher Bioreagents CAS 7647-14-5) stock solutions of 3 M and 5 M respectively were prepared by dissolving the corresponding amounts of sorbitol or NaCl in autoclaved DI water and filtering using a 0.2 µm filter. The solutions used for perturbations were obtained by diluting the stock solutions with autoclaved DI water.

Prior to imaging, 13,000 cells were plated in a µ-Plate 96-well black treated imaging plate (Ibidi) and allowed to adhere overnight (~16 h) before perturbations. Cells were stained with DAPI (Thermo). To prepare the stain, a 14.3 mM DAPI stock dissolved in DI water was diluted to a final concentration of 300 µM with complete media. The media from the cells was aspirated and DAPI-containing media was added to the cells, which were then incubated for 15 min at 37 °C and 5% CO2. After the incubation period, the cells were rinsed twice with PBS and 200 µL of PBS was added.

U-2 OS fluorescence microscopy

Imaging was done on a Zeiss epifluorescent microscope using a 40 × 0.9 NA dry objective. Excitation was done with a Colibri LED excitation module and data were collected on dual Hamamatsu Flash v3 sCMOS cameras. The cells were imaged at room temperature before and less than 1 min following perturbation with 300 ms exposure times. Imaging was done by exciting DAPI (385 nm) under donor excitation (Dx, 430 nm) or acceptor excitation (Ax, 511 nm). Emitted light was passed on to the camera using a triple bandpass dichroic (467/24, 555/25, 687/145). When measuring FRET, emitted light was split into two channels using a downstream beamsplitter with a 520 nm cutoff. For each perturbation, the cells were focused using the DAPI channel, and imaged with two channels using Dx, in one channel using Ax. The final osmolarities that were used for the perturbations were: 150 mOsm, 300 mOsm (isosmotic), 525 mOsm, 600 mOsm, and 650 mOsm with sorbitol or NaCl as the osmotic agents. From each well in the 96-well plate, 4-5 cells were analyzed. Each perturbation was replicated at least 3 times in a single plate, and the data reported are combined from at least two plates prepared on different days.

U-2 OS image analysis

The images were analyzed using ImageJ. For each cell, 5 ROIs were selected: (1) background ROI, located where no cells were present, to measure any background changes that may have occurred due to media changes; (2–5) four ROIs in the cytoplasm of each cell. For each ROI, the background signal was subtracted, and average intensity values were reported in four channels: (a) donor emission under donor excitation (DxDm), (b) acceptor emission under donor excitation (DxAm), (c) acceptor emission under acceptor excitation (AxAm), and (d) DAPI emission under DAPI excitation. To correct for donor bleedthrough, cells were plated and stained as previously mentioned. Cells were imaged, the acceptor was photobleached under prolonged direct acceptor excitation, and the cells were imaged again. ROIs of all the cells present in the plane of view were measured. A correlation plot of donor emission against acceptor emission was generated to determine percent bleedthrough, as shown in **Fig C3B**.

Quantification and statistical analysis

Data were analyzed with one-way ANOVA or two-way ANOVA for all experiments with more than two samples, as indicated in the figure legends, with Tukey's multiple comparison test. For experiments with two samples, data were analyzed using unpaired Student's t-test. Symbols *, **, and *** indicate p-values < 0.05, 0.01, and 0.001, respectively, unless specified differently in the figure legends.

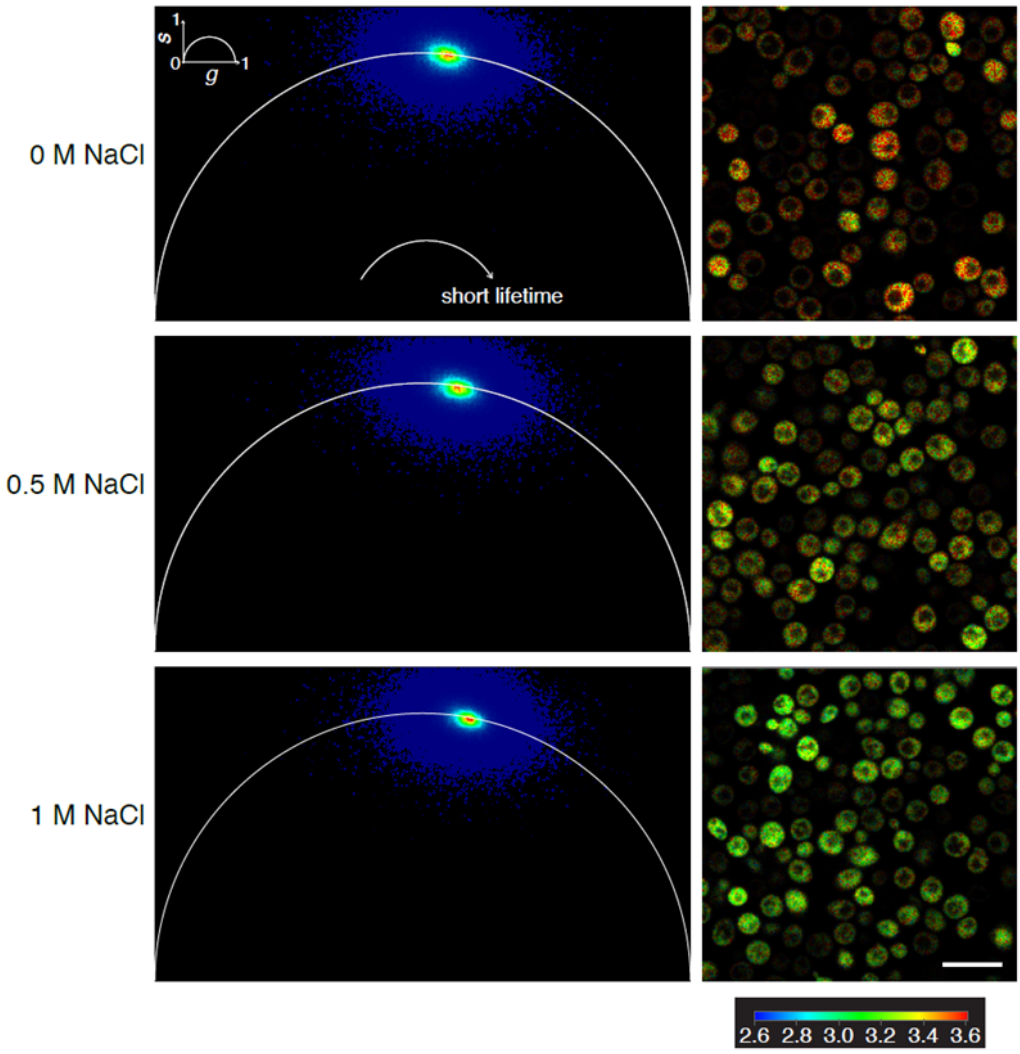


Figure C1. Phasor plots (left) and donor fluorescence lifetime images (right) of live yeast cells expressing AtLEA4-5 fused to mCerulean3 (donor-only control) under 0 M, 0.5 M, and 1 M NaCl. Signals shifted to the left side of the phasor plot represent longer fluorescence lifetimes, whereas those shifted to the right side represent shorter fluorescence lifetimes. Scale bar = 10 μm. The calibration bar represents the donor fluorescence lifetime in nanoseconds (ns). The experiment was repeated 3 times independently with similar results.

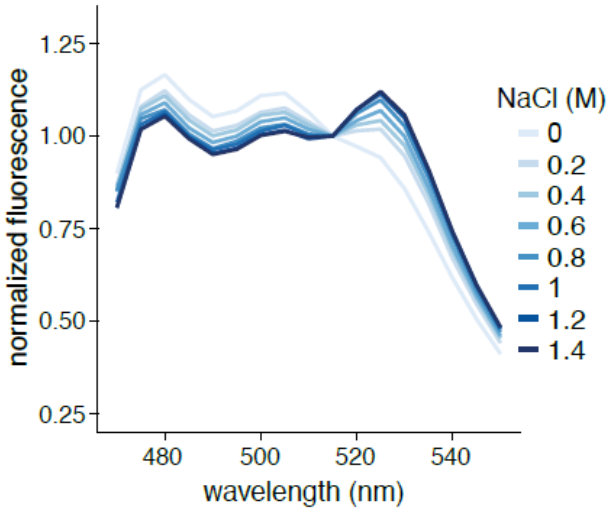


Figure C2. Fluorescence emission spectra of NaCl-treated live Escherichia coli cells expressing SED1. Fluorescence values were normalized to the value at 515 nm.

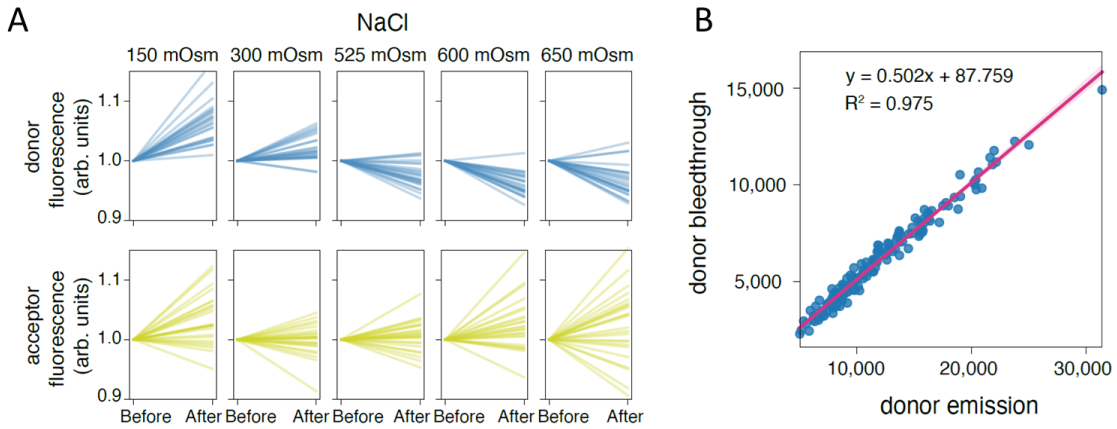


Figure C3. (A) Donor and acceptor fluorescence trajectories before and after the treatment of SED1-expressing U-2 OS single cells with NaCl at the indicated osmolarities. (B) A correlation plot of donor (DxDm) against acceptor (DxAm) emission was used to determine the bleedthrough correction. Multiple wells were imaged and measurements of all cells present in the plain of view were taken from the bleached images. Source Data are provided in SI Tables.

References

(1) Vitalis, A.; Pappu, R. V. ABSINTH: A New Continuum Solvation Model for Simulations of Polypeptides in Aqueous Solutions. *J. Comput. Chem.* **2009**, *30* (5), 673–699.
(2) Holehouse, A. S.; Sukenik, S. Controlling Structural Bias in Intrinsically Disordered Proteins Using Solution Space Scanning. *J. Chem. Theory Comput.* **2020**, *16* (3), 1794–1805.
(3) Moses, D.; Yu, F.; Ginell, G. M.; Shamoon, N. M.; Koenig, P. S.; Holehouse, A. S.; Sukenik, S. Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to Their Chemical Environment. *J. Phys. Chem. Lett.* **2020**, 10131–10136.
(4) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109* (8), 1528–1532.

Appendix D

Supporting Information for Chapter 4 Discussion:

The material originally appeared in the following: Structural biases in disordered proteins are prevalent in the cells David Moses, Karina Guadalupe, Feng Yu. et al. bioRxiv (2021)

Table S1-S3 can be found at this link:
https://www.biorxiv.org/content/10.1101/2021.11.24.469609v2.supplementary-material
Table S4 can be found at this link:
https://github.com/sukeniklab/IDP_structural_bias

Methods

The CD experiments and data analysis are done by the author of this dissertation. The other works mentioned here are done by the collaborators.

FRET construct design and cloning

The FRET backbone for bacterial expression (fIDP_pET-28a(+)-TEV) or for mammalian expression (fIDP_pCDNA3.1(+)) was prepared by ligating mTurquoise2 and mNeonGreen into pET28a-TEV or pCDNA backbone using 5' NdeI and 3' XhoI restriction sites. Genes encoding for IDP regions were obtained from GenScript (Piscataway, NJ) and ligated between the two fluorescent proteins using 5' SacI and 3' HindIII restriction sites. Cloned plasmids were amplified in XL1 Blue (Invitrogen) cell lines using manufacturer-supplied protocol. Sequences of all IDP inserts are available in **Table S1**.

FRET construct expression and purification

BL21 (DE3) cells were transformed with fIDP_pET-28a(+)-TEV plasmids according to manufacturer protocol and grown in LB medium with 50 µg/mL kanamycin. Cultures were incubated at 37 °C while shaking at 225 rpm until OD600 of 0.6 was reached (approx. 3 h), then induced with 1 mM IPTG and incubated for 20 h at 16 °C while shaking at 225 rpm. Cells were harvested by centrifugation for 15 min at 3,000 rcf, the supernatant was discarded, and the cells were lysed in lysis buffer (50 mM NaH_2PO_4, pH 8, 0.5 M NaCl) using a QSonica Q700 Sonicator (QSonica, Newtown, CT). Lysate was centrifuged for 1 h at 20,000 rcf and the supernatant collected and flowed through a column packed with Ni-NTA beads (Qiagen). The FRET construct was eluted with 50 mM NaH_2PO_4, pH 8, 0.5 M NaCl, 250 mM imidazole, and further purified using size-exclusion chromatography on a Superdex 200 PG column (GE Healthcare) in an AKTA go protein purification system (GE Healthcare). The purified FRET constructs were divided into 200 µL aliquots, flash-frozen in liquid nitrogen, and stored at -80 °C in 20 mM sodium phosphate buffer, pH 7.4, with the addition of 100 mM NaCl. Protein concentration was measured after thawing and before use using UV-vis absorbance at 434 and 506 nm (the peak absorbance wavelengths for mTurquoise2 and mNeonGreen, respectively; the molar absorbance coefficients for mTurquoise2 and mNeonGreen are

30,000 $cm^{-1}M^{-1}$ and 116,000 $cm^{-1}M^{-1}$, respectively. Calculations of concentration based on λ = 434 nm produced slightly higher values than calculations based on λ = 506 nm, so the concentrations based on the measurement at λ = 506 nm were used), and purity was assessed by SDS-PAGE after thawing and before use. To verify the brightness of the FPs, we measured the UV-Vis absorbance of both donor and acceptor molecules before each FRET assay. We used only samples that displayed an absorbance ratio Abs_{506}/Abs_{434} = ratio of 2.8 ± 0.2, a reasonable ratio given the difference in the molar extinction coefficients of mTurquoise2 and mNeonGreen. Samples where the ratio deviated from this value were discarded.

Preparation of solutions for solution-space scanning

Solutes were purchased from Alfa Aesar (Sarcosine, PEG2000), GE Healthcare (Ficoll), Thermo Scientific (Guanidine Hydrochloride), and Fisher BioReagents (Glycine, Potassium Chloride, Sodium Chloride, Urea), and used without further purification. Stock solutions were made by mixing the solute with 20 mM sodium phosphate buffer, pH 7.4, with the addition of 100 mM NaCl except for experiments where the concentration of NaCl and KCl were varied, which began free of additional salt. The same buffer was used for all dilutions.

***In vitro* FRET experiments**

In vitro FRET experiments were conducted in black plastic 96-well plates (Nunc) with clear bottom using a CLARIOstar plate reader (BMG LABTECH). Buffer, stock solution, and purified protein solution were mixed in each well to reach a volume of 150 µL containing the desired concentrations of the solute and the FRET construct, with a final concentration of 1 µM protein. Fluorescence measurements were taken from above, at a focal height of 5.7 mm, with gain fixed at 1020 for all samples. For each FRET construct, two repeats from different expressions with 6 or 12 replicates each were performed in neat buffer, and two repeats from different expressions were done in every other solution condition. Fluorescence spectra were obtained for each FRET construct in each solution condition by exciting the sample in a 16 nm band centered at λ = 420 nm, with a dichroic at λ = 436.5 nm, and measuring fluorescence emission from λ = 450 to 600 nm, averaging over a 10 nm window moved at intervals of 0.5 nm. Base donor and acceptor spectra for each solution condition were obtained using the same excitation and emission parameters on solutions containing 1 µM mTurquoise2 or mNeonGreen alone, and measuring fluorescence emission from 450 to 600 nm.[1,2]

Calculation of FRET efficiencies and end-to-end distances

The apparent FRET efficiency (E_f^{app}) of each FRET construct in each solution condition was calculated by linear regression of the fluorescence spectrum of the FRET construct with the spectra of the separate donor and acceptor emission spectra in the same solution conditions (in order to correct for solute-dependent effects on fluorophore emission). (E_f^{app}) was calculated using the equation[3]:

$$E_f^{app} = 1 - \frac{F_d}{\frac{Q_d f_d}{Q_d f_a} F_s + F_d}$$

where F_d is the decoupled donor contribution, F_s is the decoupled acceptor contribution, f_d is the area-normalized donor spectrum, f_a is the area-normalized acceptor spectrum, Q_d = 0.93 is the quantum yield of mTurquoise2, and Q_a = 0.8 is the quantum yield of mNeonGreen[2,4].

The data for each series of solution conditions consisting of increasing concentrations of a single solute was processed in the following manner:

1. Raw spectra for the free donor and free acceptor in the various solution conditions were loaded, and the averages of all repeats in each solution condition were computed. These averages are referred to as the “raw” donor and acceptor spectra below because they will be further corrected.
2. The donor and acceptor peak intensities were assumed to change in a linear fashion with increasing solute concentration, so peak height of donor- or acceptor-only spectra vs. concentrations were linearly fit.
3. To correct for artifacts (such as variations in FRET construct concentration between different wells) that may contribute to unexpected differences in fluorescence intensity, a correction factor was applied to each raw donor and acceptor spectrum to bring the peak intensity to the linear fit described in step 2, resulting in “corrected” donor and acceptor spectra. Importantly, we have seen in our previous work that this correction corrects well-to-well variations in raw data but has a negligible effect on overall values and trends[5].
4. The raw FRET construct fluorescence spectra for the series were loaded.
5. To compensate for unintended direct excitation of the acceptor by excitation at the donor excitation frequency, the corrected acceptor spectrum for each solution condition was subtracted from the FRET construct spectrum for each solution condition, resulting in “corrected” FRET construct spectra.
6. The corrected donor, acceptor and FRET construct spectrum for each solution condition was fitted with a linear regression function to determine the decoupled contributions of the donor and acceptor to the FRET construct spectrum.
7. E_f^{app} of each FRET construct in each solution condition was calculated using the equation shown above.

Size exclusion chromatography and small-angle X-ray scattering experiments

Small-angle X-ray scattering (SAXS) experiments were performed at BioCAT (beamline 18ID at the Advanced Photon Source, Chicago). The experiments were performed with in-line size exclusion chromatography (SEC-SAXS) (**Fig. D1**) to separate monomeric protein from aggregates and improve the accuracy of buffer subtraction. Experiments were conducted at 20 °C in 20 mM sodium phosphate, pH 7.4, with 100 mM NaCl. Samples of approximately 300 µL were loaded, at concentrations in mg/mL approximately equal to 240 divided by the molecular weights of the constructs in kD (for example, a typical construct of molecular weight 60 kD would have a target concentration for SEC-SAXS of 240/60 = 4 mg/mL), onto a Superdex 200 Increase 10/300 column (GE Life Sciences) and run at 0.6 mL/min using an ÄKTA Pure FPLC

system (Cytiva). The column eluent passed through the UV monitor and proceeded through the SAXS flow cell which consists of a 1.5 mm ID quartz capillary with 10 µm walls. The column to X-ray beam dead volume was approximately 0.1 mL. Scattering intensity was recorded using a Pilatus3 1M (Dectris) detector placed 3.5 m from the sample providing access to a q-range from 0.003-0.35 $Å^{-1}$. 0.5 second exposures were acquired every 2 seconds during the elution. Data was reduced at the beamline using BioXTAS RAW version 2.1.1[6,7]. The contribution of the buffer to the X-ray scattering curve was determined by averaging frames from the SEC eluent which contained baseline levels of integrated X-ray scattering, UV absorbance and conductance. Frames were selected as close to the protein elution as possible and, ideally, frames pre- and post-elution were averaged. Multiple peaks for GS48, WT PUMA, E1A, and FUS were deconvolved using evolving factor analysis (EFA) (**Fig. D2**)[8,9] and the peak with calculated molecular weight corresponding to the monomer was chosen for further analysis. Final scattering profiles were generated by subtracting the average buffer trace from all elution frames and averaging curves from elution volumes close to the maximum integrated scattering intensity; these frames were statistically similar in both small and large angles. Buffer subtraction and subsequent Guinier fits (**Fig. D3**), as well as Kratky transformations (**Fig. D4**), deconvolution of peaks using EFA, molecular weight calculations based on volume of correlation[10] and Porod volume[11] (**Table S1**), and pair distance distribution (P(r)) analysis using the indirect Fourier transform (using the algorithm in the GNOM program by Svergun and Semyenuk) were done in BioXTAS RAW. Radii of gyration (R_g) were calculated from the slope of the fitted line of the Guinier plot at maximum $q \times R_g = 1$using the equation[12]:

$$ln[I(q)] = ln[I(0)] - (\frac{R_g^2}{3})q^2$$

Mammalian cell culture

HEK293T cells were cultured in Corning treated flasks with Dulbecco's modified Eagle medium (Advanced DMEM:F12 1X, Gibco Cat. No. 12634-010) supplemented with 10% FBS (Gibco Cat. No. 16000-044) and 1% penicillin/streptomycin (Gibco Cat. No. 15140-122). For live-cell microscopy experiments, 5,000 cells were plated in a µ-Plate 96-well black treated imaging plate (Ibidi Cat. No. 89626) and allowed to adhere overnight (~16 hours) before transfection. Cells were incubated at 37 °C and 5% CO2. Before transfection, the media was switched out with new warmed media. XtremeGene HP (Sigma Cat. No. 6366236001) was used to transfect FRET construct plasmids into HEK293T cells per manufacturer's protocol. Cells were incubated at 37 °C and 5% CO2 for 48 hours. NaCl stock solution was prepared by dissolving NaCl (Fisher Bioreagents CAS 7647-14-5) in 1X PBS (Gibco Cat. No. 70011-044) and filtering using a 0.2 µm filter. The solutions used for perturbations were obtained by diluting the imaging media (1X PBS) with autoclaved DI water to achieve hypoosmotic (100 mOsm) conditions or by adding NaCl stock solution for hyperosmotic (750 mOsm) conditions. Isoosmotic (300 mOsm) conditions were obtained by adding 1X PBS. To prepare for imaging, cells were rinsed once with 1X PBS and left in 200 µL PBS (300 mOsm) for imaging.

Live-cell microscopy

Imaging was done on a Zeiss epifluorescent microscope using a 10X 0.3 NA dry objective. Excitation was done with a Colibri LED excitation module and data was collected on a duocam setup with two linked Hamamatsu Flash v3 sCMOS cameras. The cells were imaged at room temperature before and after perturbation with 150 ms exposure times. Imaging was done by exciting mTurquoise2 at 430 nm (donor and acceptor channels) or mNeonGreen at 511 nm (direct acceptor channel). Emitted light was passed on to the camera using a triple bandpass dichroic (467/24, 555/25, 687/145). When measuring FRET, emitted light was split into two channels using a downstream beamsplitter with a 520 nm cutoff. For each perturbation, the cells were focused using the acceptor channel and imaged before manually adding water (hypoosmotic condition), PBS (isosmotic condition), or NaCl solution (hyperosmotic condition) and pipetting up and down 10 times to ensure mixing. The final osmolarities that were used for the perturbations were: 100 mOsm, 300 mOsm (isosmotic), and 750 mOsm with NaCl as the osmotic agent. Imaging was typically completed in ~ 45 seconds.

Image analysis

Images were analyzed using ImageJ[13]. Images collected before and after osmotic challenge, containing three channels each, were stacked and aligned using the StackReg plugin with rigid transformation (**Fig. D5**)[14]. The aligned image was segmented based on the donor channel before perturbation. Segmentation was done using several methods to ensure that the results were robust. The methods included the ImageJ built-in implementations of the Triangle and MinError algorithm, as well as a fixed threshold that selected only pixels with intensities between 1,500 - 40,000. All methods gave nearly identical results, so the fixed threshold method was finally selected for the data shown in all live cell figures. The resulting mask was processed using the Open and Watershed binary algorithms of ImageJ. Cells were selected using the Analyze Particles option of ImageJ, picking only those with an area between 65 - 845 μm^2, and with a circularity of 0.1 - 1.0. The resulting regions of interest were averaged in each channel at each timepoint. The resulting cells were filtered to remove cells with an intensity over 10,000 (to correlate with *in vitro* experiment concentrations, see **Fig. D6**) and cells where the absolute change in direct acceptor emission was over 2,000 (which tended to be cells that moved or lifted off the coverslip during measurement). To correct for donor bleedthrough and cross-excitation, cells were transfected with the mTurquoise2 or mNeonGreen construct only, the cells were imaged and analyzed using the same protocol as previously mentioned, and correlation plots were generated to determine percent bleedthrough and cross-excitation (**Fig. D7**). The final filtering step removed cells with a corrected donor/acceptor ratio that was negative or higher than 6. Cell FRET efficiency before and after perturbation ($E^{cell}_{f,before}$ and $E^{cell}_{f,after}$ respectively) was calculated by $E^{cell}_{f} = \frac{F_A}{F_D + F_A}$. The resulting dataset is available as **Table S2**. The number of cells measured for each construct and condition from this dataset are summarized in **Table S3**. Analysis code is available as an ImageJ macro at https://github.com/sukeniklab/IDP_structural_bias.

Label-free peptide synthesis and purification

WT PUMA and shuffled sequences were prepared via standard microwave-assisted solid-phase peptide synthesis protocols using a Liberty Blue automated microwave peptide synthesizer (CEM, NC, USA) and ProTide Rink Amide resin (CEM). Fmoc-deprotection was achieved by treatment with 4-methylpiperidine (20% v/v) in dimethylformamide (Sigma-Aldrich), and Fmoc-amino acids were activated using N,N'-Diisopropylcarbodiimide (Sigma-Aldrich) and Oxyma Pure (CEM). Peptides were N-terminally acetylated and C-terminally amidated. After synthesis, the peptidyl resins were filtered and rinsed with acetone and air-dried. The crude peptides were cleaved from the resin for 4 hours at room temperature with a 92.5% trifluoroacetic acid (TFA), 2.5% H2O, 2.5% 3,6-dioxa1,8-octane-dithiol, 2.5% triisopropylsilane cleavage solution, precipitated with cold diethyl ether, and centrifuged at 4000 rpm for 10 min at 4 °C. After centrifugation, the supernatants were discarded, and the pellets were dried under vacuum overnight. Crude peptides were purified by high-performance liquid chromatography (HPLC) using an Agilent 1260 Infinity II HPLC instrument equipped with a preparative scale Phenomenex Kinetex XB-C18 column (250 × 30 mm, 5 µm, 100 Å) (**Fig. D8**). Peptides were eluted with a linear gradient of acetonitrile-water with 0.1% TFA. The target fractions were collected, rotovapped, and lyophilized. Purified peptides were analyzed by mass spectrometry using a Q-Exactive Hybrid Quadrupole-Orbitrap mass spectrometer (Thermo Scientific) (**Fig. D9, Table S4**).

CD spectroscopy

Lyophilized protein constructs were weighed and dissolved in a 20 mM sodium phosphate, 100 mM NaCl buffer at pH 7.4 to make a 200 µM stock. The stock was diluted into a concentration series to measure the CD spectra. CD spectra were measured using a JASCO J-1500 CD spectrometer with a 1 cm quartz cell for 1 µM and 2 µM protein concentration and 0.1 cm quartz cell for other concentrations (Starna Cells, Inc., Atascadero, CA) using a 0.1 nm step size, a bandwidth of 1 nm, and a scan speed of 200 nm/min between 260 to 190 nm. Each spectrum was measured 7 times and averaged to increase the signal-to-noise ratio. The buffer control spectrum was subtracted from each protein spectrum. CD spectra were normalized using UV 280 nm absorbance to eliminate the small concentration difference between different protein constructs.

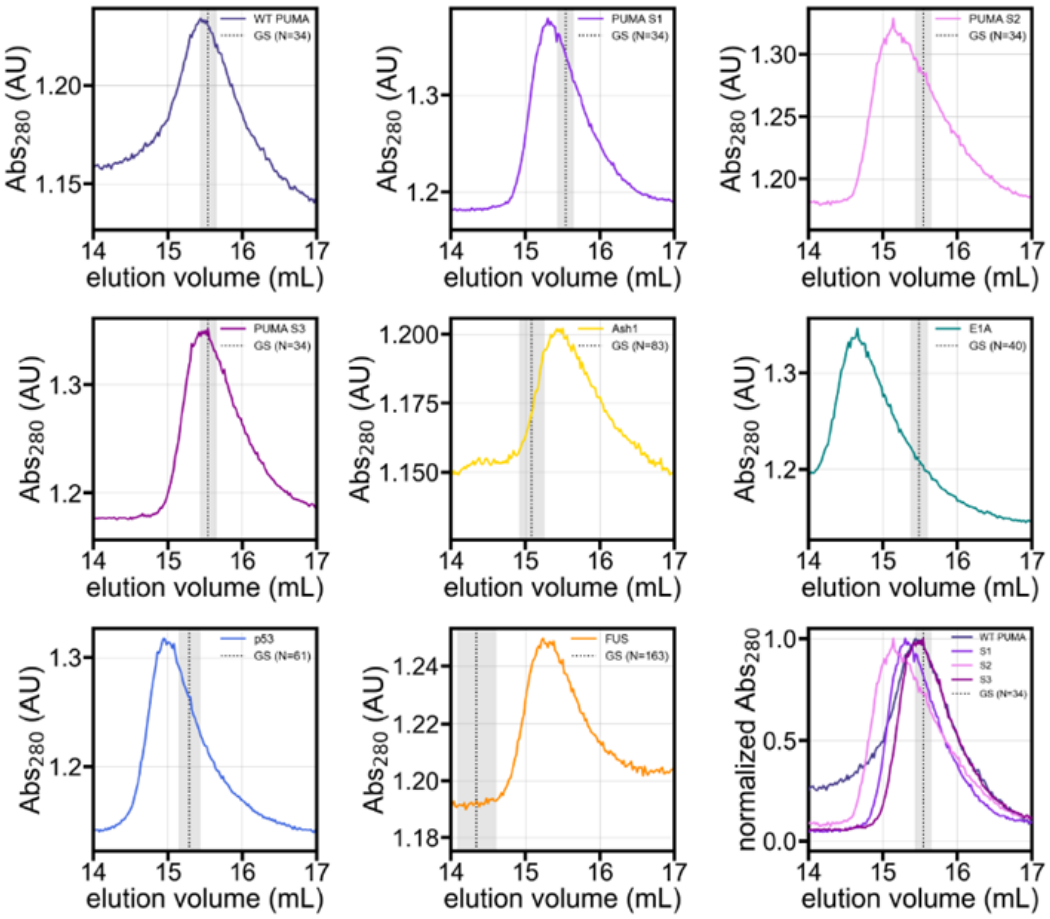


Figure D1. Chromatograms from SEC-SAXS experiments in which the samples were donor-IDP-acceptor FRET constructs in a dilute phosphate buffer solution. The vertical dotted line labeled “GS” in each panel represents the expected elution peak position of a FRET construct containing a GS-repeat sequence equal in length to the IDP, where N refers to the number of amino acids. The shaded region in each panel represents the standard error of the expected GS peak position.

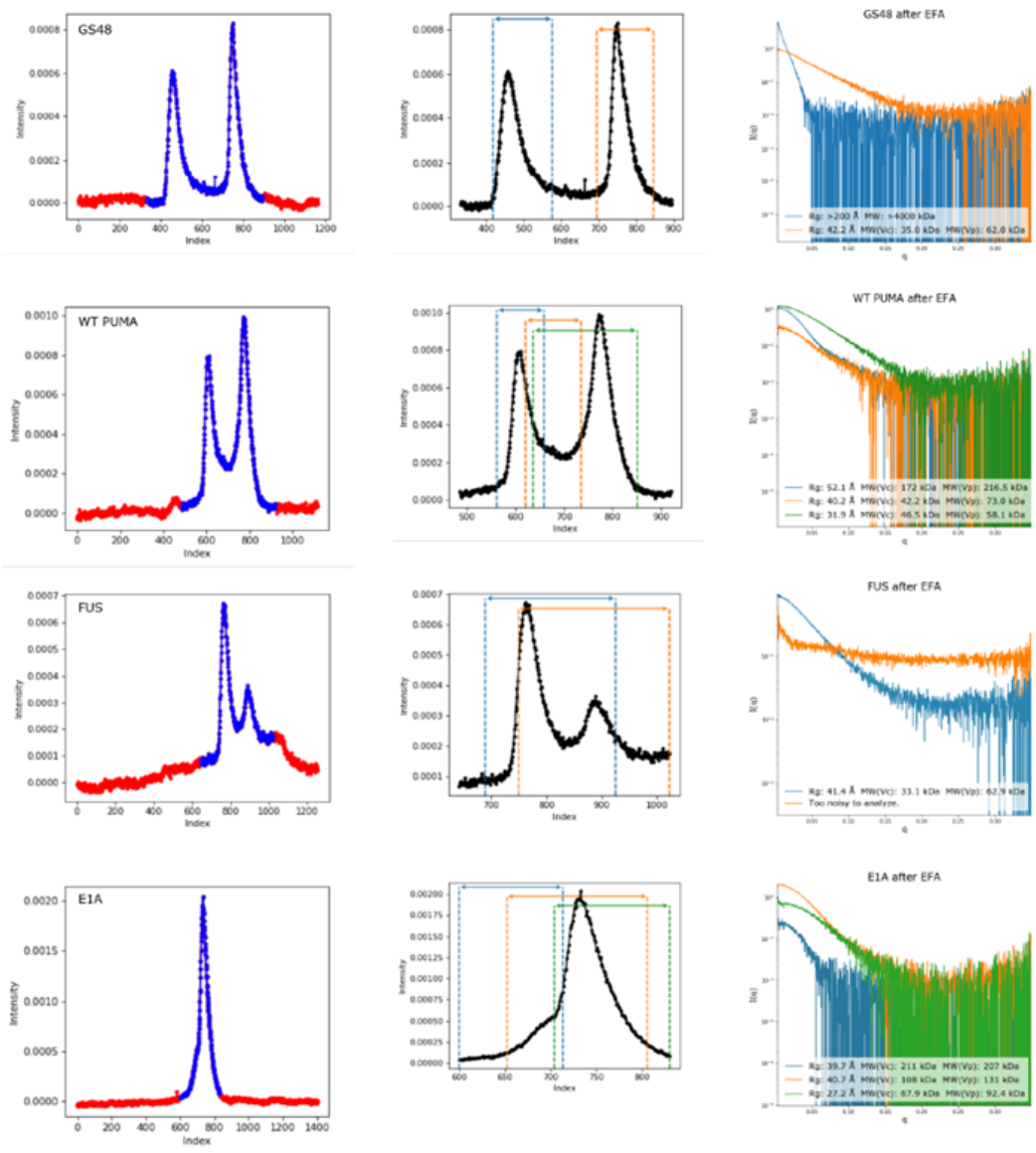


Figure D2. Screens from BioXTAS RAW software showing process of deconvolution of SEC peaks using evolving factor analysis. Left: raw chromatograms. Center: ranges of deconvoluted peaks. Right: $I(q)$ vs q series, calculated radius of gyration, and calculated molecular weight for each deconvoluted peak. Same colors in center and right panels represent the same deconvoluted peaks.

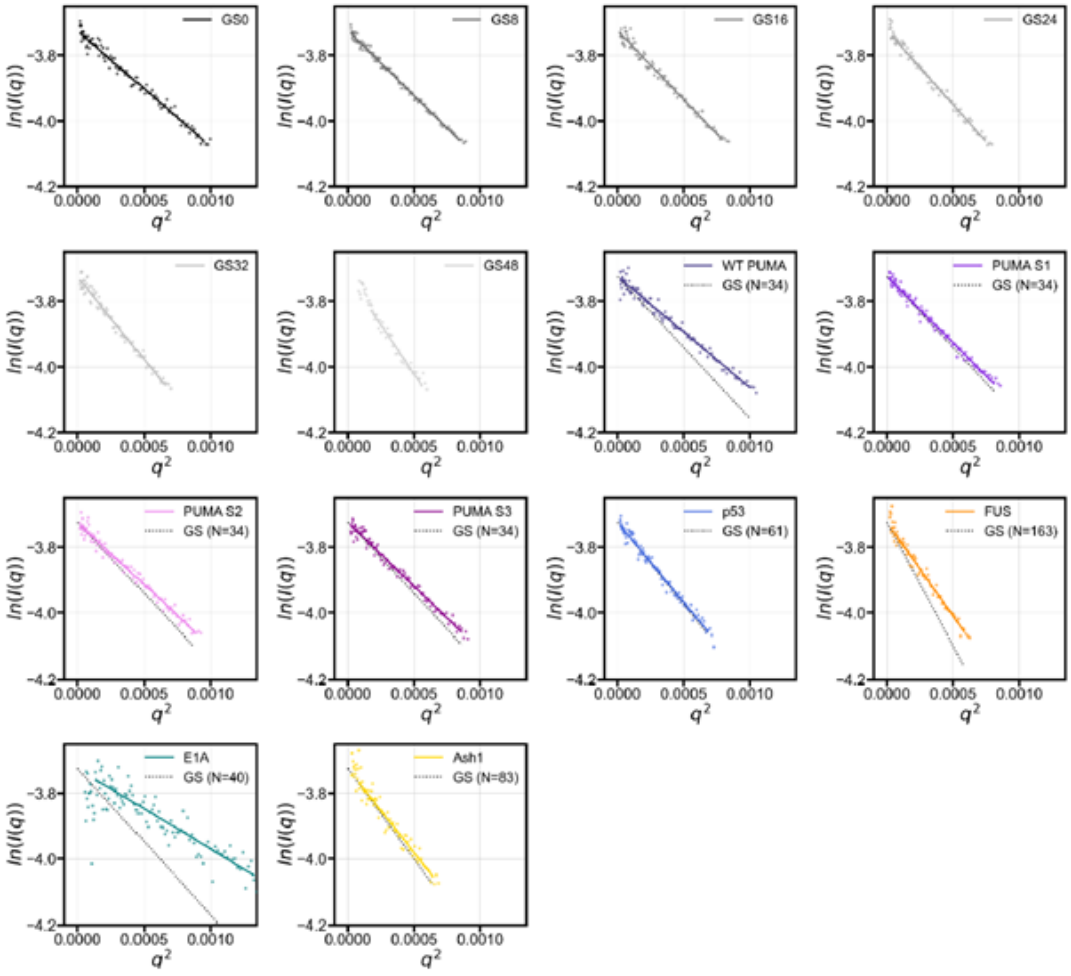


Figure D3. Guinier plots for donor-IDP-acceptor FRET constructs from SEC-SAXS experiments. For IDRs that are not GS-repeat sequences, the fitted line is compared with the expected fitted line for a construct containing a GS-repeat sequence of the same length (black dotted lines), where N refers to the number of amino acids. Lines are fitted to a maximum $q * R$ value of 1.

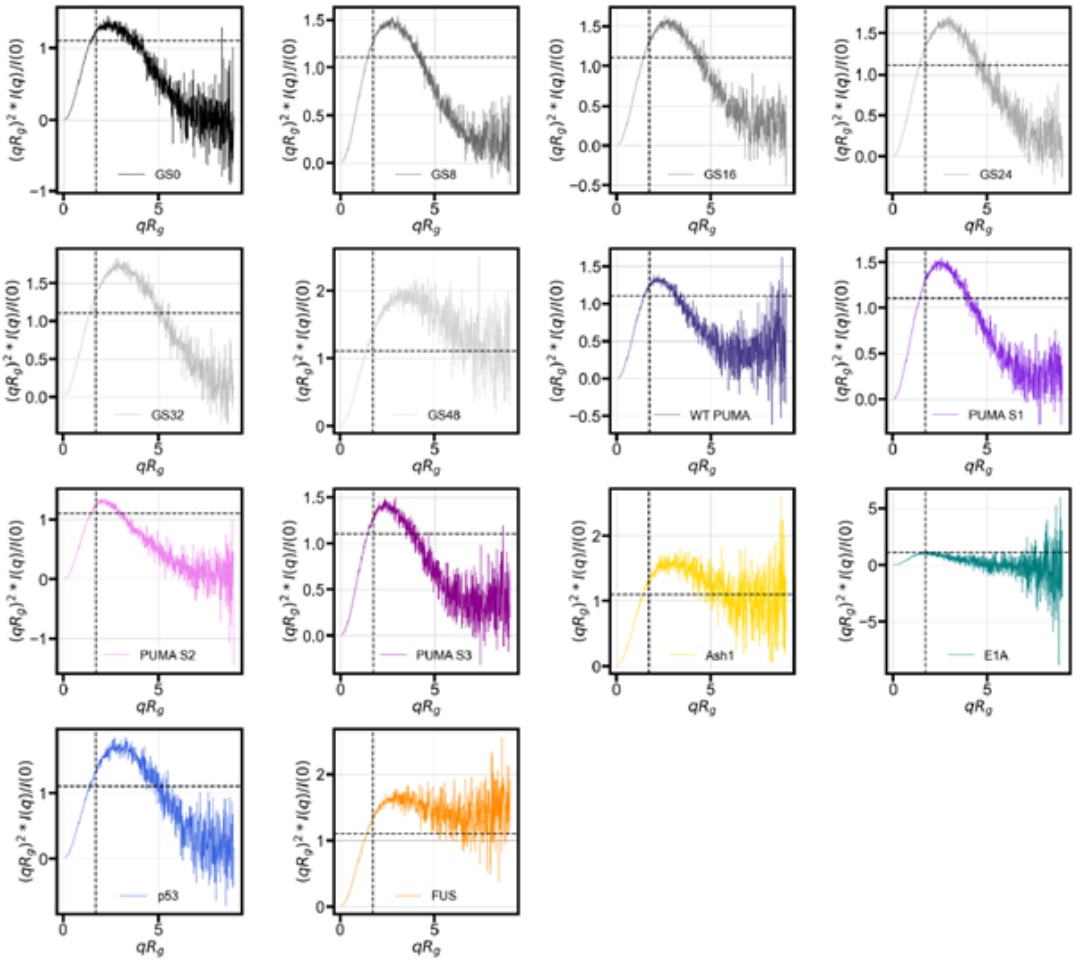


Figure D4. Dimensionless Kratky plots derived by transforming the scattering profiles from which the values reported in the main text were calculated. For a globular protein, the peak position should be at $qR_g = \sqrt{3} \sim 1.73$ (shown by vertical dashed line) and the peak height should be $(qR_g)^2 * I(q)/I(0) = 3/e \sim 1.1$ (shown by horizontal dashed line).

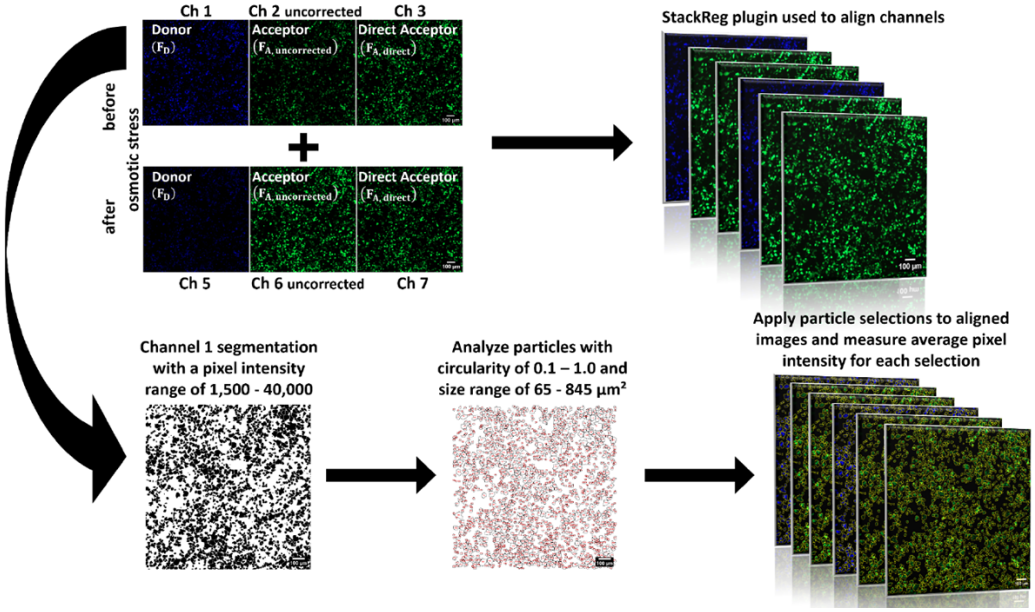


Figure D5. Analysis pipeline for live cell data. The donor channel before perturbation (Ch 1)was segmented using a fixed threshold to include any pixels with an intensity value between 1,500 - 40,000. The ImageJ "analyze particles" algorithm was used to select thresholded regions with a circularity between 0.1 -1.0 and a size of 65 - 845 µm². All channels were aligned using the StackReg plugin before segmented regions were applied and measured. Final measurements were corrected for bleedthrough and cross-excitation using slopes obtained from Fig. D3. The complete dataset can be found in Table S2. In this table, channels 1, 2, 2 uncorrected and 3 correspond to donor (F_D), corrected acceptor (F_A), uncorrected acceptor ($F_{A,\,uncorrected}$) and direct acceptor ($F_{A,\,direct}$) before osmotic stress, respectively. Channels 5, 6, 6 uncorrected and 7 correspond to donor (F_D), corrected acceptor (F_A), uncorrected acceptor ($F_{A,\,uncorrected}$) and direct acceptor ($F_{A,\,direct}$) after osmotic stress, respectively.

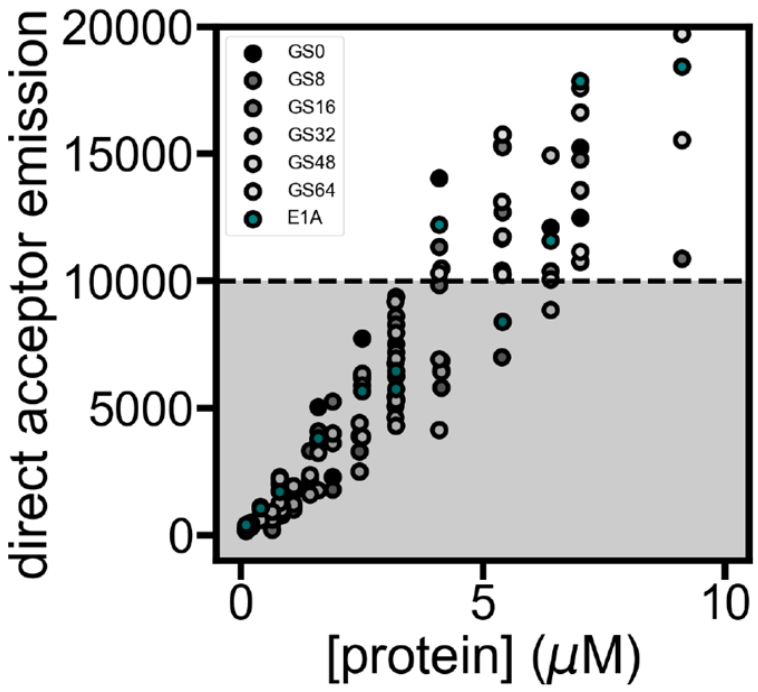


Figure D6. In vitro measurement of direct acceptor emission for known recombinant, purified proteins measured on the same setup as the live cells. The dashed line shows the emission cutoff used to select cells with a concentration range around 5 μM or lower to correlate with in vitro experiments.

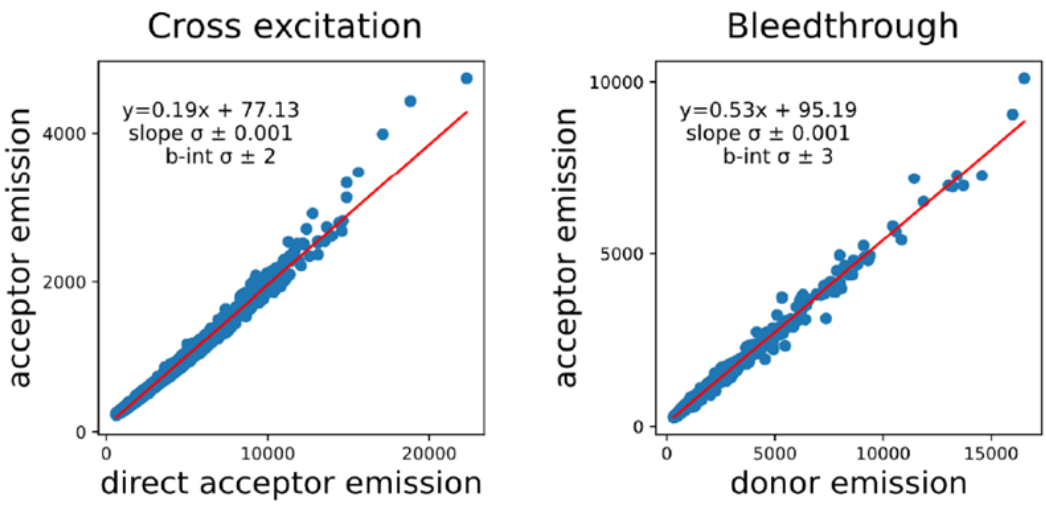


Figure D7. Measurements of cross-excitation (left) and bleedthrough (right) from donor to acceptor channel. To calculate cross-excitation, cells expressing mNeonGreen only were imaged. To calculate bleedthrough, cells expressing mTurquoise2 only were imaged. In both cases, the same imaging settings as those used for FRET constructs were used. (left) The x-axis shows acceptor emission under acceptor excitation. (right) The x-axis shows donor emission under donor excitation. In both figures, the y-axis shows acceptor emission under donor excitation. The slopes of these two values were used to correct the signal from the FRET construct according to the following equation:

$$F_A = F_{A,\,uncorrected} - (0.19 \times F_{A,\,corrected} + 0.53 \times F_D)$$

where is F_A used to calculate E_f^{cell}. The numbers 0.19 ± 0.001 and 0.53 F ± 0.001 are the slopes from the figures above. Additionally, we performed photobleaching experiments where mNeonGreen of various FRET constructs were bleached. These bleached constructs were used to measure and calculate bleedthrough and similar results were obtained (slope of 0.51 ± 0.007).

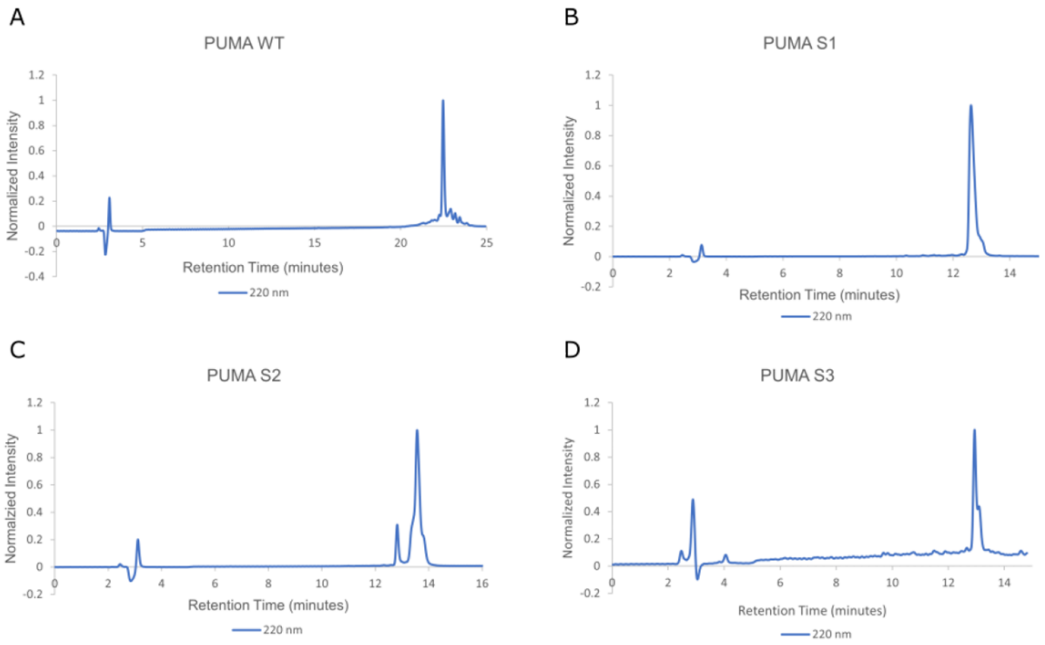


Figure D8. HPLC traces from the purification of label-free peptides. (A) PUMA WT. (B) PUMA S1. (C) PUMA S2. (D) PUMA S3.

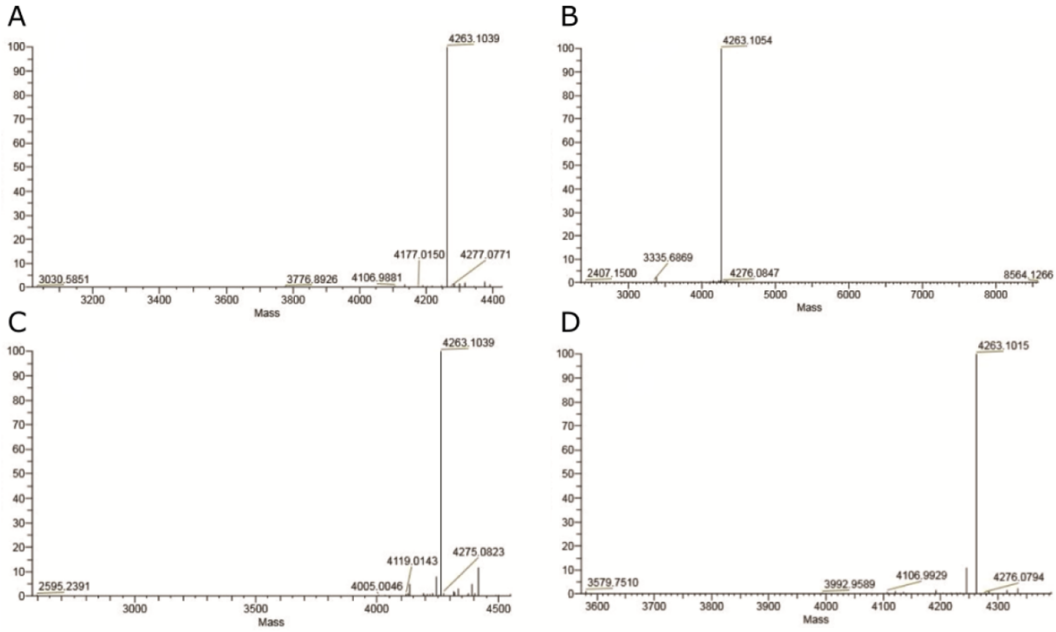


Figure D9. High-resolution ESI mass spectra of purified label-free peptides. (A) PUMA WT. (B) PUMA S1. (C) PUMA S2. (D) PUMA S3. Calculated and experimental masses are shown in **Table S4.**

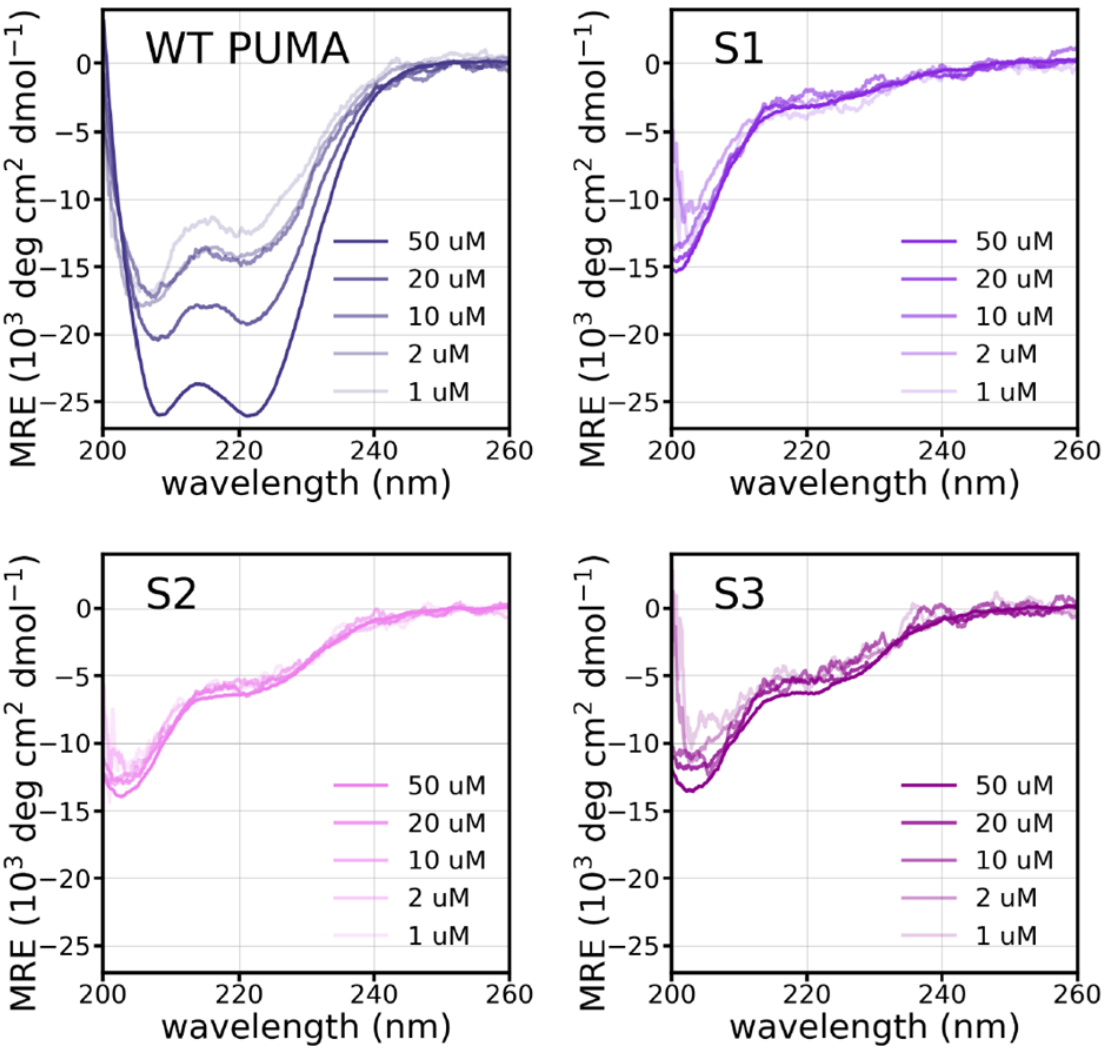


Figure D10. Concentration dependence of circular dichroism measurements of PUMA WT and sequence scrambles.

References

(1) Cranfill, P. J.; Sell, B. R.; Baird, M. A.; Allen, J. R.; Lavagnino, Z.; de Gruiter, H. M.; Kremers, G.-J.; Davidson, M. W.; Ustione, A.; Piston, D. W. Quantitative Assessment of Fluorescent Proteins. *Nat. Methods* **2016**, *13* (7), 557–562.

(2) Lambert, T. J. FPbase: A Community-Editable Fluorescent Protein Database. *Nat. Methods* **2019**, *16* (4), 277–278.

(3) Wlodarczyk, J.; Woehler, A.; Kobe, F.; Ponimaskin, E.; Zeug, A.; Neher, E. Analysis of FRET Signals in the Presence of Free Donors and Acceptors. *Biophys. J.* **2008**, *94* (3), 986–1000.

(4) Mastop, M.; Bindels, D. S.; Shaner, N. C.; Postma, M.; Gadella, T. W. J., Jr; Goedhart, J. Characterization of a Spectrally Diverse Set of Fluorescent Proteins as FRET Acceptors for mTurquoise2. *Sci. Rep.* **2017**, *7* (1), 11999.

(5) Moses, D.; Yu, F.; Ginell, G. M.; Shamoon, N. M.; Koenig, P. S.; Holehouse, A. S.; Sukenik, S. Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to Their Chemical Environment. *J. Phys. Chem. Lett.* **2020**, 10131–10136.

(6) Nielsen, S. S.; Toft, K. N.; Snakenborg, D.; Jeppesen, M. G.; Jacobsen, J. K.; Vestergaard, B.; Kutter, J. P.; Arleth, L. BioXTAS RAW, a Software Program for High-Throughput Automated Small-Angle X-Ray Scattering Data Reduction and Preliminary Analysis. *J. Appl. Crystallogr.* **2009**, *42* (5), 959–964.

(7) Hopkins, J. B.; Gillilan, R. E.; Skou, S. BioXTAS RAW: Improvements to a Free Open-Source Program for Small-Angle X-Ray Scattering Data Reduction and Analysis. *J. Appl. Crystallogr.* **2017**, *50* (Pt 5), 1545–1553.

(8) Maeder, M.; Zilian, A. Evolving Factor Analysis, a New Multivariate Technique in Chromatography. *Chemometrics Intellig. Lab. Syst.* **1988**, *3* (3), 205–213.

(9) Meisburger, S. P.; Taylor, A. B.; Khan, C. A.; Zhang, S.; Fitzpatrick, P. F.; Ando, N. Domain Movements upon Activation of Phenylalanine Hydroxylase Characterized by Crystallography and Chromatography-Coupled Small-Angle X-Ray Scattering. *J. Am. Chem. Soc.* **2016**, *138* (20), 6506–6516.

(10) Rambo, R. P.; Tainer, J. A. Accurate Assessment of Mass, Models and Resolution by Small-Angle Scattering. *Nature* **2013**, *496* (7446), 477–481.

(11) Piiadov, V.; Ares de Araújo, E.; Oliveira Neto, M.; Craievich, A. F.; Polikarpov, I. SAXSMoW 2.0: Online Calculator of the Molecular Weight of Proteins in Dilute Solution from Experimental SAXS Data Measured on a Relative Scale. *Protein Sci.* **2019**, *28* (2), 454–463.

(12) Martin, E. W.; Hopkins, J. B.; Mittag, T. Chapter Seven - Small-Angle X-Ray Scattering Experiments of Monodisperse Intrinsically Disordered Protein Samples close to the Solubility Limit. In *Methods in Enzymology*; Keating, C. D., Ed.; Academic Press, 2021; Vol. 646, pp 185–222.

(13) Abramoff; Magalhães, P. J.; Ram, S. J. Image Processing with ImageJ. *Biophotonics Int.* **2004**, *11* (7), 36–42.

(14) Thévenaz, P.; Ruttimann, U. E.; Unser, M. A Pyramid Approach to Subpixel Registration Based on Intensity. *IEEE Trans. Image Process.* **1998**, *7* (1), 27–41.