UNIVERSITY OF CALIFORNIA, MERCED

Intrinsically disordered proteins as physicochemical sensors

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

in

Chemistry and Biochemistry

by

David Charles Moses

Committee in charge:
Professor Andy LiWang, Chair
Professor Ashok Deniz
Professor Victor Muñoz
Professor Shahar Sukenik, Advisor
Professor Tao Ye

2023

University of California, Merced
Graduate Division

The Dissertation of David Charles Moses is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

Professor Ashok Deniz Date

_____

Professor Victor Muñoz Date

_____

Professor Tao Ye Date

_____

Professor Shahar Sukenik, Advisor Date

_____

Professor Andy LiWang, Chair Date

2023

Dedication

To Shahar Sukenik and Alex Holehouse for the ideas on which I had the privilege of building my research. To Eduardo Flores for being an inspirational labmate and friend. To Karina Guadalupe for being a fun deskmate who never ever distracted me while I was working, for co-first authoring the *NSMB* paper, and for the great teamwork during the revision process. To Feng Yu for co-first authoring the *JPCL* paper, for the arguments during Happy Hours at the Tarmac, and for the most romantic vocal performance I've ever heard from someone dressed in a banana costume. And to Nora Shamoon for the generous spirit she brought to the lab.

Table of Contents

List of Figures

List of Abbreviations

IDR	intrinsically disordered protein region
GS	glycine-serine
FP	fluorescent protein
FRET	fluorescence resonance energy transfer
SAXS	small-angle X-ray scattering
SEC	size exclusion chromatography
CD	circular dichroism
E_f	FRET efficiency
E_f^{app}	apparent FRET efficiency (used for *in vitro* measurements)
E_f^{cell}	apparent FRET efficiency (used for live cell measurements)
E_f^{sim}	FRET efficiency calculated for a simulated construct
R_0	Förster radius
R_g	radius of gyration
R_e	end-to-end distance
χ	normalized end-to-end distance

Acknowledgements

I am indebted to my PI Shahar Sukenik for wise guidance and abundant opportunities to grow, to my committee members – Victor Muñoz, Tao Ye, Ashok Deniz and Andy LiWang – for asking excellent questions, to Erik Martin for important input regarding SAXS analysis, to Jesse Hopkins and Max Watkins at Argonne National Lab for stellar beamline work, to my labmates, to my collaborators, and to everyone at Castle and on the 2nd floor of BPS that made UC Merced a great place to develop as a scientist. As Poirot wrote to Hastings: They were good days.

This work was supported by a fellowship from NSF-CREST Center for Cellular and Biomolecular Machines (CCBM) at UC Merced, Grant No. NSF-HRD-1547848, by the NIH under award R35GM137926 to Shahar Sukenik, and by the Graduate Dean's Dissertation Fellowship from the Graduate Division at UC Merced. This research used the Advanced Photon Source at Argonne National Laboratory under Contract No. DE-AC02-06CH11357, Proposal No. 75514.

Permissions

Permission to use copyrighted material that appears in Chapter 1 has been granted by the authors. Sections 1.1 and 1.2 as well as all figures in Chapter 1 are excerpted from the following: David Moses, Garrett M. Ginell, Alex S. Holehouse and Shahar Sukenik (2023) Trends Biochem Sci. 2023 Aug 30:S0968-0004(23)00204-9.

Permission to use copyrighted material that appears in Chapter 2 has been 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.

Permission to use copyrighted material that appears in Chapter 3 has been granted by the authors. The material originally appeared in the following: David Moses, Karina Guadalupe, Feng Yu, Eduardo Flores, Anthony Perez, Ralph McAnelly, Nora M. Shamoon, Gagandeep Kaur, Estefania Cuevas-Zepeda, Andrea D. Merg, Erik W. Martin, Alex S. Holehouse and Shahar Sukenik (2023). Structural biases in disordered proteins are prevalent in the cell. Accepted for publication by Nature Structural and Molecular Biology.

Permission to use copyrighted material that appears in Chapter 4 has been granted by the authors. Chapter 4, sections 4.1 through 4.4 are excerpted from the following: 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 (2021) Nat. Commun. 12: 5438.

Curriculum Vitae

David Moses

EDUCATION

Ph.D., Biochemistry, University of California, Merced, 2023
M.S., Chemical Biology, Tamkang University (Tamsui, Taiwan), 2018

- Co-advisors: Professor Yau-Hung Chen, Professor Chi-Yuan Chou (National Yang-Ming University, Taipei, Taiwan)
- Thesis: Comparative structural and functional studies of wild-type and C271A mutant forms of severe acute respiratory syndrome coronavirus papain-like protease

M.A., East Asian Languages, University of California, Berkeley, 1994

- Advisor: Professor Stephen H. West
- Thesis: “Meaning is the master in writing”: literature and culture in the critical writings of Wang Ruo-Xu (1174–1273)

B.A., Zoology, University of California, Davis, 1989

RESEARCH EXPERIENCE

2018 – pres. Graduate Student Researcher – University of California, Merced

- Developed an experimental pipeline to allow high-throughput investigation of responses of disordered proteins to changes in their surrounding physicochemical environment.
- Developed python scripts to analyze and visualize data.
- Expressed proteins in E. coli and purified them using affinity and size-exclusion chromatography.
- Performed high-throughput FRET experiments using a ClarioStar plate reader.
- Worked as part of a team with live-cell microscopists and computational scientists and co-first-authored original papers including one accepted by Nature Structural and Molecular Biology.
- Wrote successful proposals for SAXS beamtimes at three national laboratories and collaborated with beamline scientists to design and plan experiments.
- Co-first author of a review of my field published in Trends in Biochemical Sciences.

- Successfully trained colleagues in protein expression/purification and biophysical assays.
- Participated extensively in the writing and editing of many scientific papers.
- Gave talks and presented posters at many conferences to promote a young lab.

2017 – 2018 Graduate Student Researcher (guest) – National Yang-ming University, Taiwan

- Performed FRET-based assays of enzyme activity.
- Expressed and purified proteins.

2016 – 2018 Graduate Student Researcher – Tamkang University, Taiwan

- Raised zebrafish.
- Imaged zebrafish cross-sections and evaluated for harm due to toxicity.

TEACHING EXPERIENCE

2018-2019 Teaching Assistant, Chemistry, UC Merced

Journal papers:

- David Moses*, Karina Guadalupe*, Feng Yu, Eduardo Flores, Anthony Perez, Ralph McAnelly, Nora M. Shamoon, Estefania Cuevas-Zepeda, Andrea Merg, Erik W. Martin, Alex S. Holehouse and Shahar Sukenik. Structural biases in disordered proteins are prevalent in the cell. 2023. Accepted for publication by Nature Structural and Molecular Biology. doi: 10.1038/s41594-023-01148-8. (Co-first author.)
- David Moses*, Garrett M. Ginell*, Alex S. Holehouse and Shahar Sukenik. Intrinsically disordered regions are poised to act as sensors of cellular chemistry. Trends in Biochemical Sciences. 2023;48(12):1019-1034. doi: 10.1016/j.tibs.2023.08.001. (Co-first author.)
- Cuevas-Velazquez CL, Vellosillo T, Guadalupe K, Schmidt HB, Yu F, Moses D, Brophy JAN, Cosio-Acosta D, Das A, Wang L, Jones AM, Covarrubias AA, Sukenik S, Dinneny JR. Intrinsically disordered protein biosensor tracks the physical-chemical effects of osmotic stress on cells. Nature Communications. 2021;12:5438. doi: 10.1038/s41467-021-25736-8.
- Moses D*, Yu F*, Ginell G, Shamoon N, Koenig P, Holehouse A, Sukenik S. Probing the Hidden Sensitivity of Intrinsically Disordered Proteins to their Chemical Environment. J Phys Chem Lett. 2020 Dec 3;11(23):10131-10136. doi: 10.1021/acs.jpclett.0c02822. (Co-first author.)

- Chu HF, Chen CC, Moses DC, Chen YH, Lin CH, Tsai YC, Chou CY. Porcine epidemic diarrhea virus papain-like protease 2 can be noncompetitively inhibited by 6-thioguanine. Antiviral Research. 2018;158:199-205. doi: 10.1016/j.antiviral.2018.08.011.
- Lin M-H, Moses DC, Hsieh C-H, Cheng S-C, Chen Y-H, Sun C-Y, Chou C-Y. Disulfiram can inhibit MERS and SARS coronavirus papain-like proteases via different modes. Antiviral Research. 2018;150:155-163. doi: 10.1016/j.antiviral.2017.12.015.
- Lai Y-H, Ding Y-J, Moses D, Chen Y-H. Teratogenic Effects of Topiramate in a Zebrafish Model. International Journal of Molecular Sciences. 2017;18(8):1721. doi:10.3390/ijms18081721.

* equal contribution

Honors and Awards:

2023 Dissertation Fellowship, UC Merced School of Natural Sciences

2023 Travel Fellowship, UC Merced Department of Chemistry and Biochemistry

2022 Travel Award, NSF workshop Find Your Inner Modeler, University of Illinois at Chicago

2022 Runner-up, graduate student poster competition, Gordon Research Seminar and Conference, Les Diablerets, Switzerland

2022 Travel Award, Gordon Research Seminar and Conference, Les Diablerets, Switzerland

2022 Travel Award, Biophysical Society Annual Meeting, San Francisco, CA

2021 UC Merced Department of Chemistry and Biochemistry Summer Research Fellowship

2020 First prize, poster session, NSF-CREST Center for Cellular and Biomolecular Machines (CCBM) Executive Advisory Board Meeting, University of California, Merced, virtual

2020 Travel Fellowship, UC Merced Department of Chemistry and Biochemistry

2020 UC Merced Department of Chemistry and Biochemistry Summer Research Fellowship

2020 NSF-CREST Center for Cellular and Biomolecular Machines (CCBM) Fellowship

2019 NSF-CREST Center for Cellular and Biomolecular Machines (CCBM) Fellowship

2017 Tamkang University Outstanding Foreign Student Fellowship

2016 Tamkang University Outstanding Foreign Student Fellowship

CONFERENCE PARTICIPATION

Discussion leader

2022 Discussion Leader, “IDPs and Biological Order Across Length Scales” session, Gordon Research Seminar, Les Diablerets, Switzerland

Talks

2023 “Structural Biases in Disordered Proteins are Prevalent in the Cell,” West Coast Structural Biology Workshop, Pacific Grove, CA

2022 “Structural Biases in Disordered Proteins are Prevalent in the Cell,” Gordon Research Seminar, Les Diablerets, Switzerland

2021 “Hidden Structure in Intrinsically Disordered Proteins is Revealed by Coupled FRET and SAXS Experiments,” NSF-CREST Center for Cellular and Biomolecular Machines (CCBM) Executive Advisory Board Meeting, University of California, Merced, virtual

2021 “Coupled FRET and SAXS Experiments Reveal Hidden Structure in Intrinsically Disordered Proteins,” 35th Gibbs Conference on Biothermodynamics, platform talk, virtual

2021 “Hidden Structure in Intrinsically Disordered Proteins is Revealed by Coupled FRET and SAXS Experiments,” Stanford University and Carnegie Institution for Science’s Intrinsically Disordered Protein Scientific Interest Group symposium, virtual

2021 “Probing the Hidden Sensitivity of Intrinsically Disordered Proteins to their Chemical Environment,” Biophysical Society Annual Meeting, platform talk, virtual

2021 “Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to their Chemical Environment,” NSF-CREST Center for Cellular and Biomolecular Machines (CCBM), Protein Metamorphosis and Responsive Nanodevices Research Thrust Meeting, Merced, CA, virtual

2020 “Environmental Sensing by Intrinsically Disordered Proteins is Encoded in Their Sequence,” Stanford University and Carnegie Institution for Science’s Intrinsically Disordered Protein Scientific Interest Group symposium, Stanford, CA

2019 “Solution Space Fingerprints of Intrinsically Disordered Regions,” NSF-CREST Center for Cellular and Biomolecular Machines Annual Retreat, Merced, CA

Posters

2023 “Structural Biases in Disordered Proteins are Prevalent in the Cell,” West Coast Structural Biology Workshop, Pacific Grove, CA

2023 “Intramolecular and Intermolecular Effects of Aromatic Sticker Residue Strength on Intrinsically Disordered Protein Regions,” Biophysical Society Annual Meeting, San Diego, CA

2022 “Structural Biases in Disordered Proteins are Prevalent in the Cell,” NSF workshop Find Your Inner Modeler, University of Illinois at Chicago

2022 “Structural Biases in Disordered Proteins are Prevalent in the Cell,” Gordon Research Conference, Les Diablerets, Switzerland, runner-up in graduate student poster competition

2022 “Hidden Structure in Disordered Proteins is Adaptive to Changes in the Cellular Environment,” Biophysical Society Annual Meeting, San Francisco, CA

2021 “Hidden Structure in Intrinsically Disordered Proteins is Revealed by Coupled FRET and SAXS Experiments,” Advanced Light Source User Meeting, Lawrence Berkeley National Laboratory, virtual

2021 “Hidden Structure in Intrinsically Disordered Proteins is Revealed by Coupled FRET and SAXS Experiments,” Protein Folding Consortium Workshop, virtual

2020 “Probing the Hidden Sensitivity of Intrinsically Disordered Proteins,” 34th Gibbs Conference on Biological Thermodynamics, virtual

2020 “Quantifying the Sensitivity of Intrinsically Disordered Proteins,” at NSF-CREST Center for Cellular and Biomolecular Machines (CCBM) Executive Advisory Board Meeting, University of California, Merced, virtual, awarded first prize in poster competition

2020 “Solution Space Fingerprints of Intrinsically Disordered Regions,” Biophysical Society Annual Meeting, San Diego, CA

2020 “Solution Space Fingerprints of Intrinsically Disordered Regions,” Stanford University and Carnegie Institution for Science’s Intrinsically Disordered Protein Scientific Interest Group symposium, Stanford, CA

2019 “Solution Space Fingerprints of Intrinsically Disordered Regions,” NSF-CREST Center for Cellular and Biomolecular Machines (CCBM) Workshop on Emerging Themes in Cellular and Biomolecular Machines, University of California, Merced

Abstract

Intrinsically disordered protein regions (IDRs) play a variety of essential roles in regulating cellular function. Instead of folding into a single stable structure, an IDR exists in an ensemble of interconverting conformations biased by local and long-range intramolecular interactions. The absence of a fixed 3D structure combined with high solvent accessibility makes IDRs sensitive to changes in their physical-chemical surroundings. This dissertation describes investigations related to the ability of IDRs to sense and respond to changes in their surrounding solution conditions, and proposes a new paradigm of IDRs as sensors of cellular physicochemistry. Chapter 1 introduces the concept of IDRs as physicochemical sensors. Chapter 2 describes a project in which I developed a pipeline to allow medium-throughput FRET assays of IDR global dimensions in a wide variety of solution conditions, as well as benchmarking and analytical tools to allow meaningful comparison of the behavior of diverse IDRs, and used this system to characterize the end-to-end distance and solution sensitivity of several naturally-occurring IDRs. Chapter 3 describes a project in which my colleague and I performed parallel assays on naturally-occurring and synthetic IDRs *in vitro* and in live cells to test whether structural biases and sensitivity seen *in vitro* would be preserved when the same IDRs were tested in live cells. Chapter 4 describes a bioengineering project in which my colleagues and I used the systems described in Chapters 2 and 3 to contribute to the design and testing of a novel IDR-based biosensor that can monitor osmotic stress in live cells. Chapter 5 concludes the dissertation and suggests directions for further research.

David C. Moses
Chemistry and Biochemistry, Ph.D.
University of California, Merced, 2023

Committee Chair: Andy LiWang

1. Introduction: Intrinsically disordered regions as physicochemical sensors

1.1 Intrinsically disordered regions and their conformational biases

Intrinsically disordered regions (IDRs) make up around a third of most eukaryotic proteomes and play critical roles in various cellular functions[1]. Unlike folded domains, IDRs lack a fixed folded structure and instead exist in a set of interconverting conformations known as an ensemble (**Fig. 1.1A**). While IDRs are characterized as disordered, they are not "unstructured." Instead, IDRs possess conformational biases that are dependent on their amino acid sequence[2–4]. These conformational biases may be driven by polar, hydrophobic, electrostatic, cation-pi, or pi-pi interactions between amino acid side chains that lead to attraction or repulsion between distal regions of an IDR[5–12]. Such interactions tune intramolecular distance distributions and ensemble-average global dimensions. As an example, long-range electrostatic interactions driven by clusters of oppositely charged residues can tune IDR global dimensions[13–15], as in the case of the cell cycle inhibitor protein p27Kip1[14]. Alternatively, short-range transient secondary structure can manifest as specific conformational states that appear as distinct subpopulations within the overall ensemble[16], e.g., transient helicity within specific subregions of IDRs, as seen in the RNA binding protein TDP-43 or the transcription factor p53[17,18]. For any given IDR, the emergent combination of sequence-encoded attractive and repulsive molecular interactions will dictate its conformational biases.

Besides amino acid sequence, another factor that influences IDR conformational biases, and therefore ensemble properties, is their physicochemical environment[19,20]. Folded domains benefit from a network of intramolecular non-covalent bonds that determine a consistent molecular topology. In IDRs, the lack of such a network has two implications. First, the designation of "buried" and "surface-exposed" residues commonly made in reference to folded proteins is not applicable as, in general, all residues in an IDR will be at least transiently solvent-exposed (**Fig. 1.1A**). Thus, the entire sequence is in direct interaction with the solution and can sense any change in surrounding chemistry. A second implication is that the sparse interactions that exist in an IDR are often too weak to resist the push and pull of the chain's interactions with its surrounding solution. For example, interactions with denaturants like urea can pull apart the non-covalent bonds that maintain a protein's structure. However, denaturation of a folded protein often requires a high urea concentration (6-8 M as a standard) because a network of intramolecular bonds resists this pull. IDRs, on the other hand, can be dramatically extended even by urea concentrations that are almost an order of magnitude smaller (< 1 M)[19–22].

Why do IDR conformational biases matter? The emerging sequence-ensemble-function paradigm for IDRs, analogous to the familiar sequence-structure-function paradigm for folded proteins, posits that IDR function is at least partly dependent on an ensemble's conformational biases[2,16]. Conformational biases can prime IDRs for molecular recognition that involves folding of the IDR upon its binding to a binding partner[18,23,24]. Alternatively, they can tune global dimensions or facilitate the formation of fuzzy complexes, where a bound structure lacks a defined 3D orientation[24–26]. Specific examples of the sequence-ensemble-function relationship include regions that form binding motifs when they exist as a transient helix[18], global dimensions tuning motif binding accessibility[27], and tuning of interactions by changing the overall volume occupied by the ensemble[28,29]. Additionally, IDRs can themselves play key functional roles without directly interacting with partners. For example, when two globular domains are tethered by an intervening IDR, the "effective concentration" of the two domains with respect to each other, and therefore the extent of their interactions, can be tuned by changing the end-to-end distance of the IDR tether[30,31]. Recent work has highlighted the importance of effective concentration to function by revealing that IDR dimensions – without conservation of a specific amino acid sequence – can be under evolutionary selection to ensure optimal linker lengths in a model termed "conformational buffering"[32]. In short, the relationship between sequence and ensemble can be critical for the biological function of IDRs.

The importance of conformational ensembles to IDR function, coupled to the inherent sensitivity of IDRs to their physicochemical environment, gives rise to the possibility of IDRs acting as molecular sensors of their surrounding physicochemical environment. The broad palette of chemistry available through the twenty natural amino acids (plus their post-translational modifications) makes possible the evolution of chemically orthogonal IDRs that are differentially sensitive to a variety of distinct physicochemical changes[19,20,33,34]. Sensing based on IDR ensemble changes would bring obvious advantages to the cell. In contrast to, for example, kinase signaling, IDR ensemble changes require no expenditure of ATP. Also, given that IDRs undergo conformational rearrangement on timescales of 50-200 ns, sensing based on IDR ensemble changes could occur extremely rapidly[35]. These features position IDRs to be exceptionally efficient protein-based sensors.

1.2. The molecular basis of physicochemical sensing

For an IDR to act as a physicochemical sensor, it must reproducibly respond to changes in its physicochemical environment (**Fig. 1.1B**). These responses may take the form of global changes in ensemble conformations or changes in local transient structure. Although these are often coupled, for simplicity we will consider them independently in our discussion below.

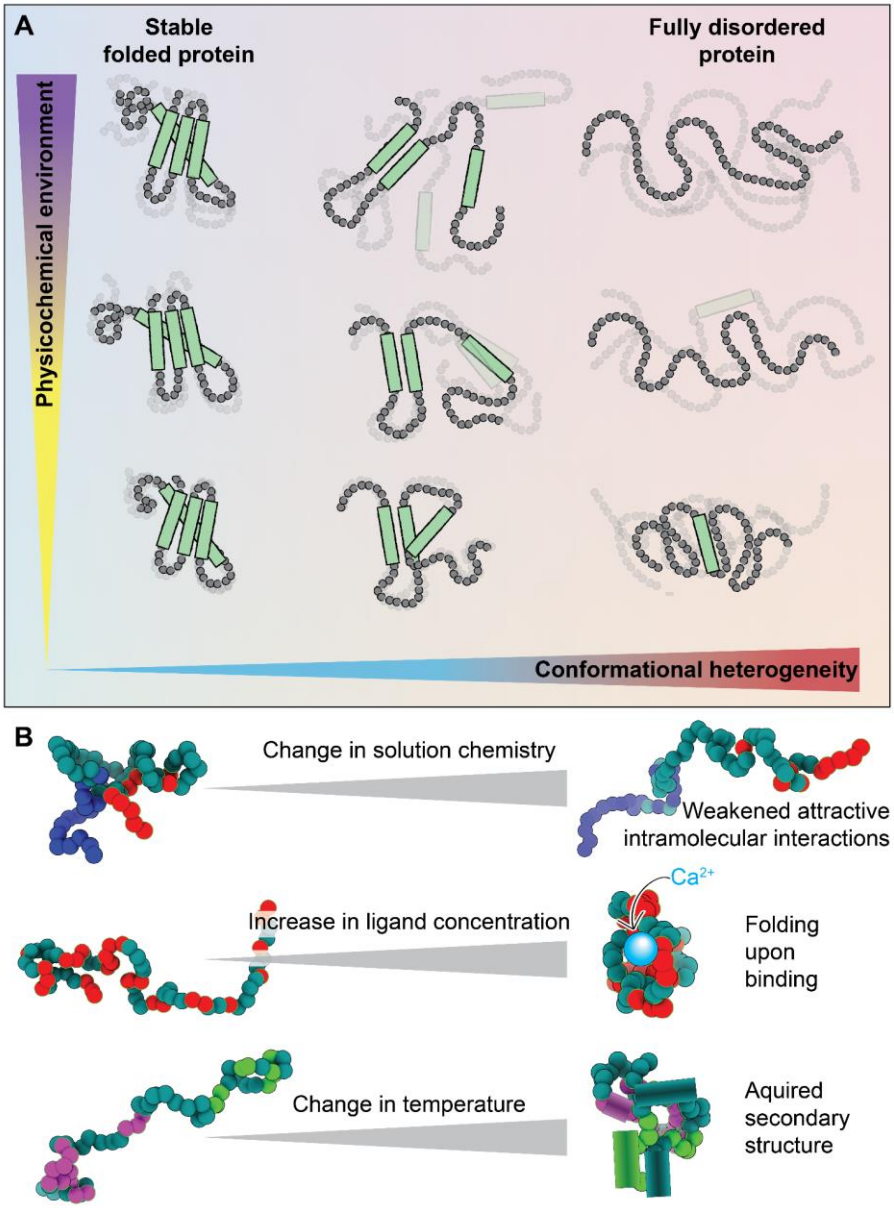

Figure 1.1. Disordered regions exist in an ensemble that is inherently sensitive to the physicochemical environment. (A) Protein conformational heterogeneity exists on a continuum,

whereby well-folded domains are at one extreme and fully disordered regions with no strong conformational biases are at the other. Regions that are highly conformationally heterogeneous contain fewer intramolecular bonds and are more solvent-exposed, and hence, in general, are more sensitive to even modest changes in the physicochemical environment. Here, the x-axis represents conformational heterogeneity while the y-axis represents some change in solution chemistry. **(B)** Scheme showing some examples of how changes in the physicochemical environment can alter IDR conformational biases. Changes in solution chemistry (salt, osmolytes, pH) may weaken (or strengthen) intramolecular interactions leading to a decrease (or increase) in transient intramolecular interactions. The presence (or loss) of ligands, including specific ions, small molecules, second messengers, and other biomacromolecules, can lead to the gain (or loss) of structure upon binding (or unbinding). Changes in physical parameters such as temperature or pressure can lead to the enhancement (or suppression) of intramolecular interactions, which can drive the acquisition (or loss) of secondary or even tertiary structure. These are just a handful of examples of how changes in physicochemistry can be sensed by IDRs.

1.2.1 Solution dependence of global conformational biases in IDRs

Changes in global IDR dimensions can be viewed through the lens of polymer physics[3,36]. If we represent an IDR as a homopolymer, its global dimensions depend on the balance between attractive and repulsive intramolecular interactions. This balance can be quantified as a single interaction energy that reflects the average overall attraction (or repulsion) of the polymer units (monomers) for one another, i.e., the mean-field self-interaction energy (ε) (**Fig. 1.2A**).

The mean-field self-interaction energy is inherently dependent on the solution environment. In a solution of polymer and solvent, increasing solvent:monomer repulsion is equivalent to increasing monomer:monomer attraction. Moving from a solution in which the mean-field self-interaction is repulsive to one in which it is attractive can manifest as a coil-to-globule transition[36] (**Fig. 1.2B**). The sharpness of this transition depends on the chain length and the magnitude of the change in self-interaction energy.

For homopolymers, only a single type of monomer unit is present, so there exists only a single type of pairwise interaction energy ($\varepsilon_{i,i}$) (**Fig. 1.2A**). For heteropolymers (like IDRs), chemically distinct monomers give rise to a matrix of pairwise interaction strengths ($E \equiv [\varepsilon_{i,i}, \varepsilon_{i,j,...,}\varepsilon_{k,n}]$) (**Fig. 1.2C,D**). A key concept in IDR sensitivity is that each of these individual pairwise interaction strengths may be modulated differently by changes in the physicochemical environment; that is, they may be chemically orthogonal (**Fig. 1.2D**). As a specific example, attractive pairwise interactions driven by electrostatics may be sensitive to salt, while attractive pairwise interactions driven by hydrogen bonding may not (**Fig. 1.2D,E**).

Two central conclusions emerge from this framework. First, IDR global dimensions must depend on amino acid sequence, as has been established by prior work[2,3,7,9,10,12,22,37]. Second, an IDR's sensitivity – that is, how much global dimensions change as a function of the changing physicochemical environment – depends on (a) the underlying IDR sequence, i.e., where on the coil-to-globule curve an ensemble begins (**Fig. 1.2E**); and (b) how the intramolecular interactions encoded in this sequence respond to their environment, i.e., how much the overall mean-field pairwise interaction energy changes in response to physicochemical changes[19] (**Fig. 1.3A**).

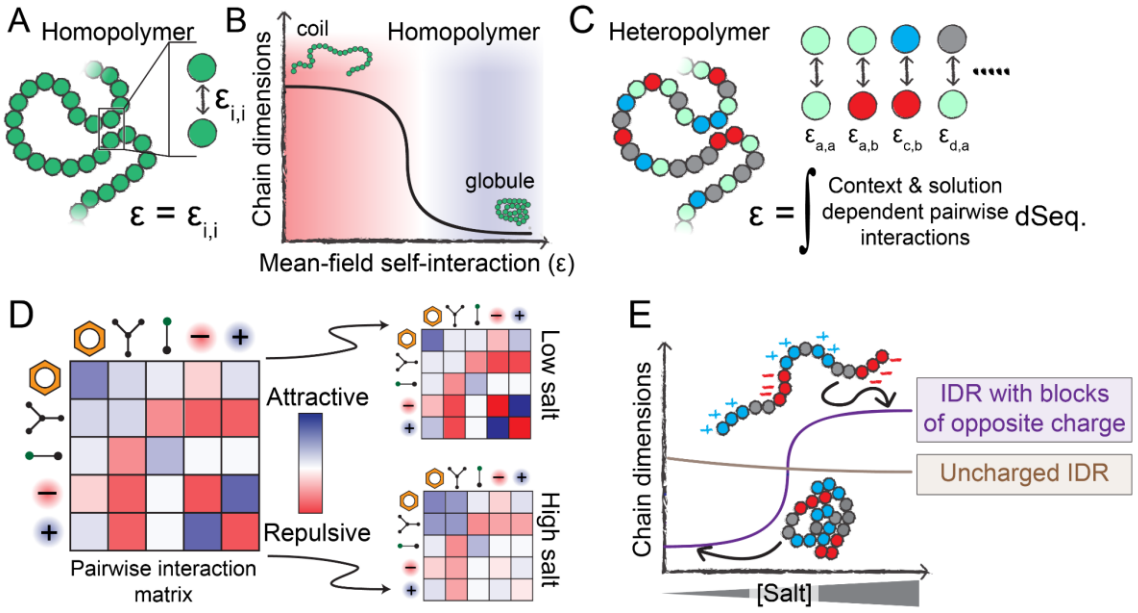

Figure 1.2. Physical principles that underlie sequence-specific IDR sensitivity to changes in physicochemistry. (A) Homopolymers are defined by a single interaction strength between each polymer unit, which also defines the mean-field self-interaction energy (ε). **(B)** If ε is repulsive, a homopolymer behaves as an extended coil with large chain dimensions, whereas if ε is attractive, a homopolymer behaves as a compact globule. The mean-field interaction energy can be varied by changing the chemical identity of the polymer unit, but can also be varied by altering the physicochemical environment the polymer finds itself in (temperature, pH, solutes, etc.). **(C)** Unlike homopolymers, heteropolymers consist of many chemically distinct units. A complete description of a heteropolymer requires knowledge of how each unique inter-residue interaction behaves, and the mean-field self-interaction energy (ε) is now defined in terms of the composition-weighted and context-dependent integral over all possible interactions. **(D)** The various types of interactions that may occur between residues in a heteropolymer can be, to first order, described by an interaction matrix. The strengths of these interactions depend on solution conditions. **(E)** The response of a heteropolymer to changes in the solution environment depends on the heteropolymer's chemistry. For example, a highly-charged IDR with blocks of oppositely-charged residues will be compact at

low salt due to strong intramolecular electrostatic interactions. However, under high-salt conditions, those attractive interactions are screened, leading to an expanded ensemble driven by the substantial solvation free energies associated with charged groups. In contrast, a charge-depleted heteropolymer may be relatively salt-insensitive and is relatively compact compared to the blocky IDR in the high-salt limit.

For IDR ensembles that begin near one of the baselines (either coil or globule), large changes in the mean-field energy can have a relatively small impact on chain dimensions, making them less sensitive (**Fig. 1.3A**, left)[19]. Analogously, for ensembles that begin in the middle of the coil-to-globule transition, relatively small changes in the mean-field interaction energy drive large changes in global dimensions, making them more sensitive (**Fig. 1.3A**, right). One could consider folded domains to be at the globular extreme of this transition, illustrating their lack of solution sensitivity. The upshot of this is that chain sensitivity peaks at the midpoint of the coil-to-globule transition (**Fig. 1.3B**). Indeed, prior work has shown that this conceptual framework is able to quantitatively normalize the solution dependence of IDRs across a wide range of different cosolutes (**Fig. 1.3C**)[19].

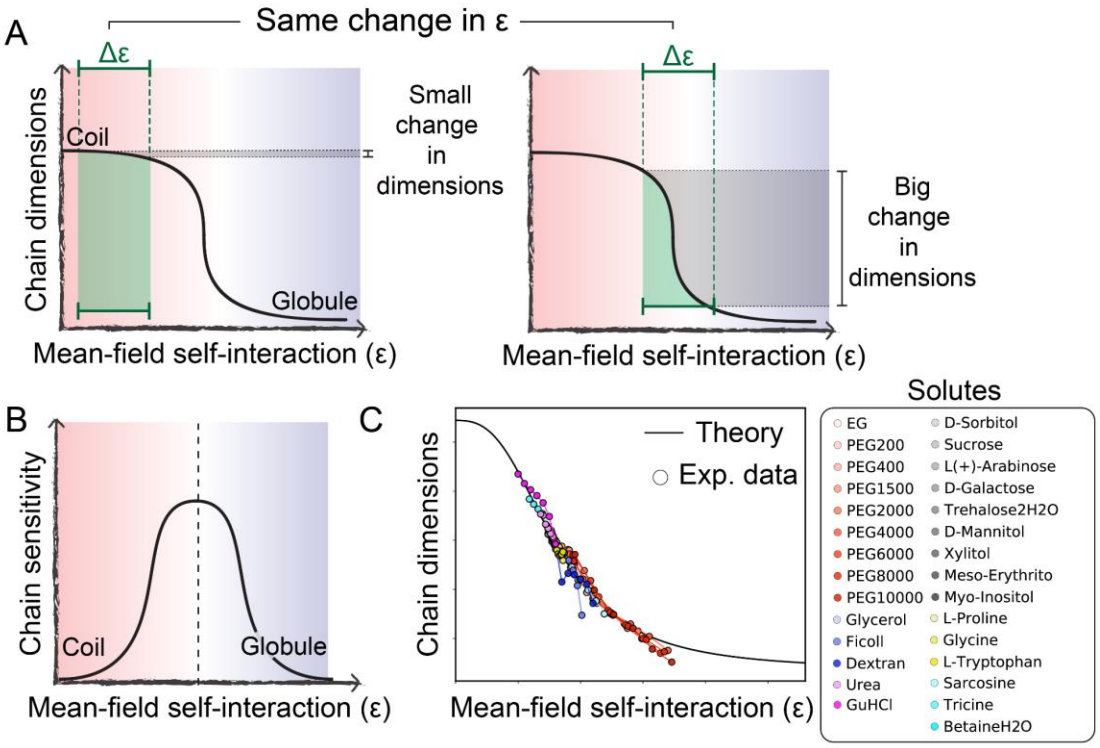

Figure 1.3. An IDR's sensitivity to the physicochemical environment depends on its intrinsic conformational biases. (A) The extent of change in chain dimensions (gray shaded region) in response to a change in mean-field self-interaction strength tuned by the physiochemical environment (green shaded region) depends on both the underlying sequence and the polymer's

behavior *prior* to the change. From equivalent changes in interaction strength, very different changes in polymer dimensions can emerge. **(B)** Chains at the midpoint of the coil-to-globule transition are most sensitive to changes in the solution environment. **(C)** Comparison of experimental data and analytical theory demonstrating the broad applicability of this framework (see **Fig. 2.4E**).

In short, baseline conformational behavior and sensitivity to environmental change, both of which depend on sequence, combine to determine an IDR's global dimensions (**Fig. 1.2E**). Together, these two features offer a quantitative framework through which IDR sensitivity can be interpreted and, looking forward, used as a design principle for the development of novel sensors.

1.2.2 Solution dependence of local conformational biases in IDRs

Local conformational biases, such as the gain or loss of transient secondary structure (especially transient helicity), can also be tuned by the environment[38,39] (**Fig. 1.4**). Importantly, the ability of ensemble conformations to change locally and not just globally means that ensembles can have different structural features even though global dimensions are the same[12,40,41]. This poses an additional challenge to experiments which often measure only a single global dimension.

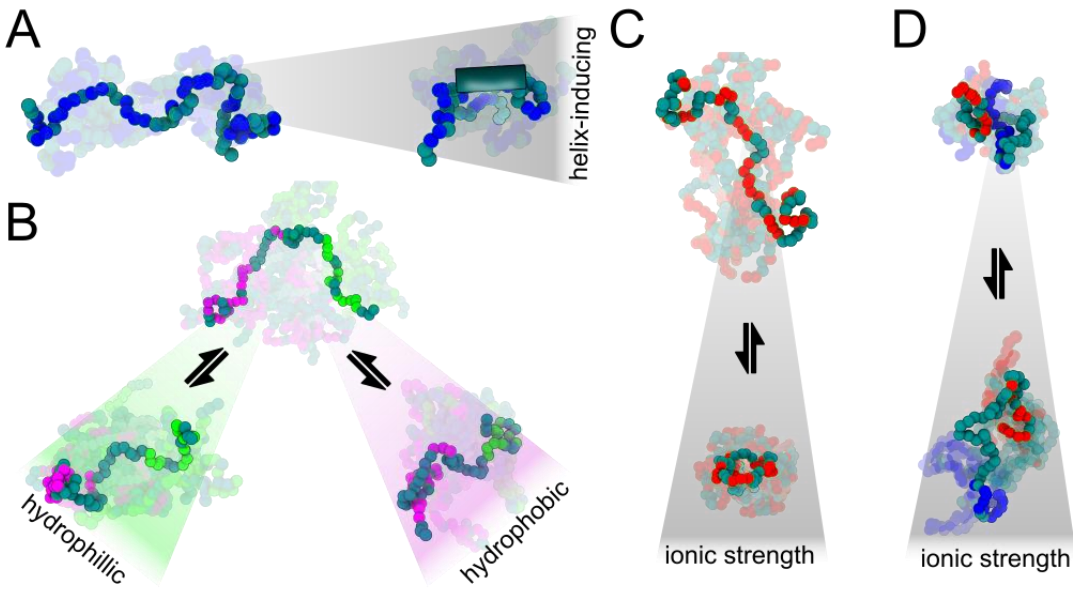

Figure 1.4. **Examples of physicochemically-driven changes in IDR ensembles**. **(A)** Promotion of secondary structural elements such as residual helicity (shown by the tube on the right) can form or dissolve binding motifs, modulating binding affinities. By prepaying an entropic cost for binding, the effective concentration of binding motifs can be rapidly enhanced or suppressed without the need to alter protein copy number. **(B)** Amphipathic sequences with patches of hydrophobic or

hydrophilic residues can compact or expand locally in different solutions, tuning accessibility of specific regions. **(C)** Sequences with high net charge (positive or negative) can compact when an increase in ionic strength screens out repulsive interactions. **(D)** Charged sequences with sequestered, opposite charges can expand at high ionic strength due to screening of attractive interactions.

1.3 Dynamic solution composition of biological systems

The cellular environment in which IDRs exist is primarily aqueous, but at the same time is crowded with macromolecules, osmolytes and ions[42]. Although cells employ many mechanisms to maintain homeostasis with respect to parameters such as pH, osmotic pressure, cell shape, and ion concentrations[43–46], the environment in the cell is nonetheless prone to changes in viscosity, crowding, ionic strength, and concentrations of various osmolytes and macromolecules[45,47]. These changes may come about as a result of routine cell-cycle events, stress, response to stress, or pathological conditions[48–51]. Changes in the concentrations of small and large cosolutes *in vitro* have been shown to alter the equilibrium stability of protein interactions[52] as well as the kinetics of protein folding and unfolding[53]; titration of small cosolutes into cells have been shown to produce changes in protein structure and aggregation propensity[54]; and rapid cell volume changes have been observed to alter the equilibrium between protein monomers and complexes, perhaps suggesting the modulation of weak protein-protein interaction networks by crowding[55]. Whether an IDR is part of a pathway critical to the life of a cell or critical to the ability of a virus to infect host cells, the cell or virus depends on the ability of the IDR to respond to changes in its surrounding environment by changing its conformation in a way that results in the required activity being performed.

1.4 IDR ensembles are modulated by changes in solution conditions

The ability of crowding to influence the structure and activity of macromolecules in the cellular cytoplasm, including their tendency to demix into separate phases, has long been acknowledged[56–58]. In recent years, a great deal of information has been obtained about the response of IDR structure and function to changes in solution conditions, such as that an IDR's response to changes in ionic strength is related to the IDR's net charge[7]; that, in accordance with polymer effects explained by Flory, IDRs existing above the θ-state in the absence of crowding should be more susceptible than IDRs existing below the θ-state in the absence of crowding to compaction by polymeric crowders[59]; that disordered proteins can mitigate the risks of solution exposure by forming interaction networks

that provide robustness in the face of solution changes[60]; and that changes in ionic strength or crowding can modulate the binding behavior of IDRs[39,60,61]. All of this is in addition to the evidence gathered in recent years linking protein intrinsic disorder to LLPS in response to cellular stress[62]. Even more recently, some groups have begun to attempt to systematically characterize the sensitivity of IDR ensembles to solution changes, whether by computationally modeling solutions that are more or less “attractive” or “repulsive” to given sequences[20], or through conducting low-resolution but high-throughput experiments placing IDRs in a variety of solution conditions and measuring ensemble-average structural changes[19]. However, despite the understanding gained thus far of the sensitivity of IDRs to their surrounding environment, and despite the pivotal roles IDRs play in the health and disease of cells and whole organisms, the modulation of IDR function by physical-chemical changes in the surrounding environment is still not well understood, and, in particular, there has been little effort devoted to systematically characterizing the sensitivity of IDR-mediated LLPS behavior to solution conditions.

1.5 Use of *in vitro* experiments to model cellular environment

My projects have involved *in vitro* experiments that place IDRs in surroundings that model phenomena common to cellular life. Crowding is one such phenomenon. Macromolecules crowd living cells at concentrations as high as 0.4 g/mL, or 40% by volume[57,63]. To simulate the experience of an IDR as its surrounding solution becomes more or less crowded with macromolecules, I used inert crowders such as Ficoll and PEG of various degrees of polymerization at concentrations ranging up to the estimated total concentration of macromolecules in living cells. The same logic applies to using widely varying concentrations of a single salt, such as NaCl, to model changes in ionic strength, or widely varying concentrations of a single osmolyte to model changes in total osmolyte concentration. Such *in vitro* experiments necessarily produce one-dimensional findings that neither fully reflect the true environment of an IDR within a living cell nor fully explain the mechanisms behind the findings. To try to fill in such gaps, I have worked with collaborators capable of investigating the behavior of the same IDRs in living cells, and with collaborators who specialize in computer simulations of macromolecular behavior.

1.6 *In vitro* methods for measuring IDR sensitivity

As discussed above, IDRs exist in a constantly changing ensemble of conformations, and their susceptibility to conformational change in response to their surrounding environment, specifically the dynamic environment in cells, has important implications regarding their function, and therefore the health and disease of the organisms in which they live. Hence it is necessary to understand how IDRs change their conformations in response to changes in their physical-chemical environment. As IDRs do not exist in a single native structure as in the case of folded proteins, they are not amenable to X-ray crystallography for structural study. High-resolution structural study of IDRs has been accomplished using single-molecule FRET and solution NMR, yielding knowledge of specific conformations adopted by IDRs, but these methods are not appropriate for high-throughput research. Therefore I settled on the use of ensemble FRET as a main method and SAXS as an orthogonal, model-free secondary method. These methods allowed me to obtain ensemble-average data describing global dimensions of IDRs, and allowed me to do so in a high-throughput fashion, enabling measurement of IDR behavior in a large number of solution conditions.

1.7 Overview of the work

Chapter 2 describes a project in which I first developed our FRET-based pipeline to allow medium-throughput assays of IDR global dimensions in a wide variety of solution conditions, as well as benchmarking and analytical tools to allow meaningful comparison of the behavior of diverse IDRs, and then used this system to characterize the end-to-end distance and solution sensitivity of several naturally-occurring IDRs. The project described in Chapter 2 synergistically combines this *in vitro* system with *in silico* experiments by my colleague Feng Yu. Building on this, Chapter 3 describes a project in which my colleague Karina Guadalupe and I performed parallel assays on naturally occurring and synthetic IDRs *in vitro* and in live cells to test whether structural biases and sensitivity seen *in vitro* would be preserved when the same IDRs were tested in live cells. To ensure accuracy of the *in vitro* FRET results, I also carried out SEC-SAXS experiments on each IDR and a colleague performed simulations of the IDR ensembles we used for benchmarking. To show a practical application of these methods, Chapter 4 describes a bioengineering project in which my colleagues and I used the systems described in Chapters 2 and 3 to contribute to the design and testing of a novel IDR-based biosensor that can monitor osmotic stress in live cells. Finally, Chapter

5 concludes the dissertations by describing some implications of IDRs as physicochemical sensors and suggesting directions for further research.

2. Revealing the hidden sensitivity of intrinsically disordered proteins

2.1 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. Despite the sensitivity of IDR ensembles to their surrounding solution and their demonstrated link to IDR function, our understanding of how these ensembles are influenced by their chemical environment is limited. We report a combined experimental, computational, and analytical framework for high-throughput characterization of IDR sensitivity we call solution space scanning. Our framework reveals a sequence-dependent sensitivity of IDRs to solution chemistry, with complex behavior that can be interpreted through relatively simple polymer models. These results imply that solution-responsive IDRs are ubiquitous and can provide an additional layer of biological regulation.

2.2 Introduction

Intrinsically disordered protein regions (IDRs) play key roles in mediating cellular signalling, transcriptional regulation, and homeostatic functions[64]. 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[18,39,65]. 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[66]. Inside the cell, the chemical composition can change due to routine cell-cycle events or external stress[47–49,54,60]. The plasticity of IDR ensembles makes them ideal sensors and actuators of these changes[20,67,68], but perhaps also impairs their activity in deleterious environments such as metabolically rewired cancer cells[69]. 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)[70,71], but also by changes in the physical-chemical composition of the

intracellular environment[47,69]. While the sensitivity of IDRs to solution composition has been discussed[1,68,72], it has not been systematically characterized. Here we set out to systematically evaluate the sensitivity of IDRs to changes in their surrounding environment.

2.3 Results

2.3.1 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[73] (donor) and mNeonGreen[74] (acceptor)[30,75]. We chose four IDRs whose ensembles play functional roles: the 61-residue N-terminal transactivation domain of p53 (p53)[18]; the 34-residue BH3 domain of apoptosis regulating protein PUMA (PUMA)[39]; the 83-residue C-terminal domain of the yeast transcription factor Ash1 (Ash1)[12]; the 40-residue N-terminal domain of the adenoviral hub protein E1A (E1A)[76] (see **Table A1.1** for amino acid sequences). For each of these constructs the FRET efficiency, E_f, was determined as described in Appendix 1, Section A1.1.5.

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 1, Section A1.1.6), allowing us to interpolate E_f for a GS linker of a given length to create a ratio χ:

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

where R_e is the end-to-end distance between donor and acceptor FPs obtained from E_f as described in Appendix 1, Section A1.1.7, 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**).

2.3.2 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 SI Section S1.7. 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[77,78]. This is in sharp contrast to the sensitivity of GS linkers, which all display a similar fingerprint regardless of length (**Fig. A1.1**).

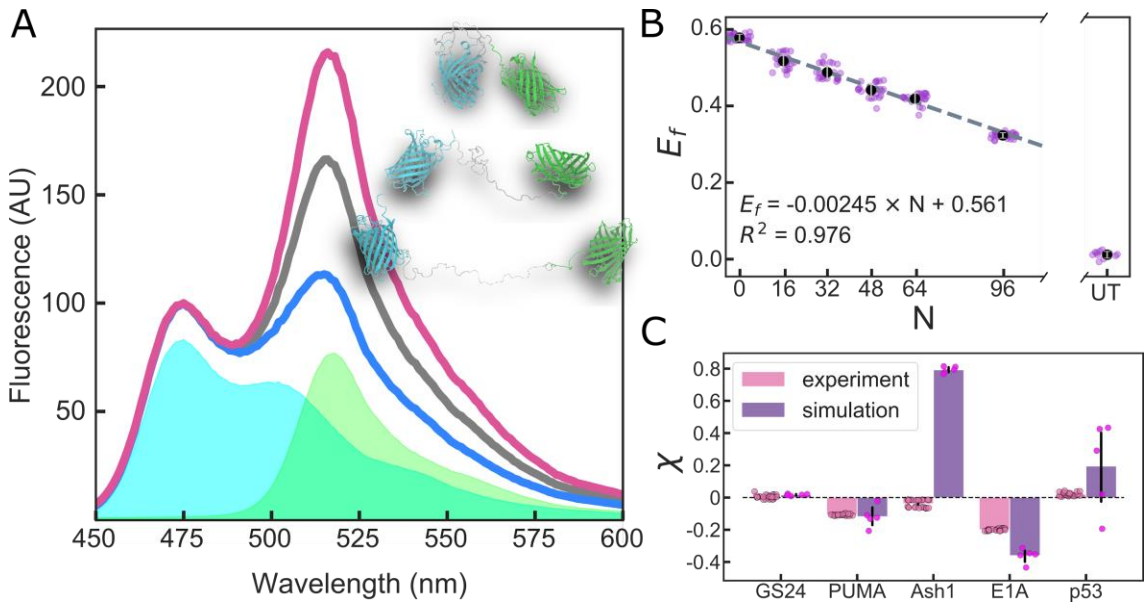

Figure 2.1. Quantifying IDR ensemble dimensions. (**A**) Fluorescence spectra normalized to donor peak intensity of a FRET construct in compacting (red), buffer (black), and expanding (blue) solutions. Cyan and green areas are base spectra of donor and acceptor FPs, respectively. Inset shows single configurations 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.

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[79]. Larger polymers such as PEG2000 and Ficoll appear to compact the dimensions of all IDRs as shown for other disordered proteins[61], with a sequence-dependent magnitude that is stronger for Ash1 and PUMA as previously reported[59]. 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[60,80,81]. 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. A1.1**), while the compaction trend stems from specific ion effects, and differs between protein types[82]. 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.

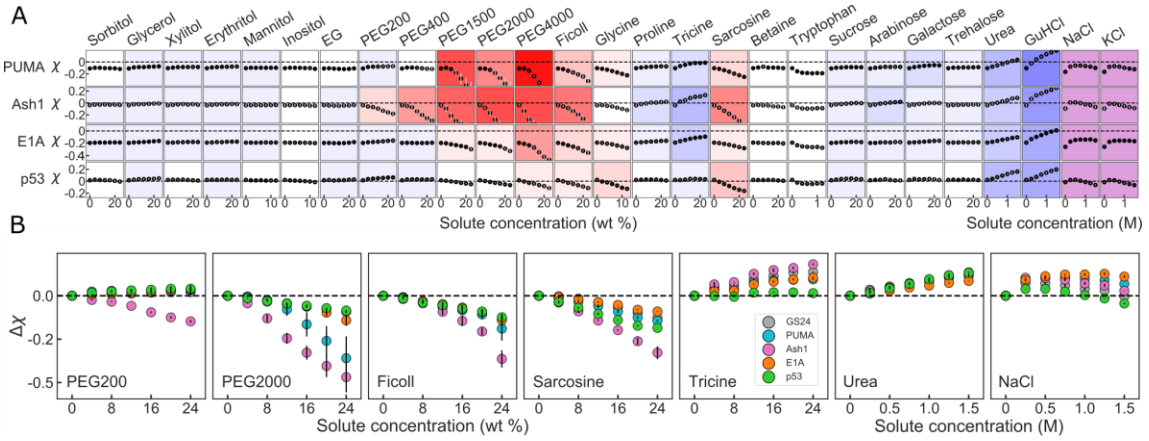

Figure 2.2. Solution space scans of IDRs. (**A**) Solution space scans of four naturally occurring IDRs. Each data point shows the average χ vs. concentration of a specific solute for each protein taken from two repeats. Vertical grey bars show spread of data, and are often too small to see. Proteins 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 $\Delta\chi = \chi([solute]) - \chi([solute] = 0)$ vs. concentration from two repeats of a specific solute for several different constructs. Vertical lines are the spread of the data.

2.3.3 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 **Section A1.2.1**).[10,12,14,83] To maintain connection with experiments, we start by simulating GS linkers of various lengths (**Fig. A1.2**), and use ensemble-averaged R_e to calculate χ for simulated IDRs according to **Eq. 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.

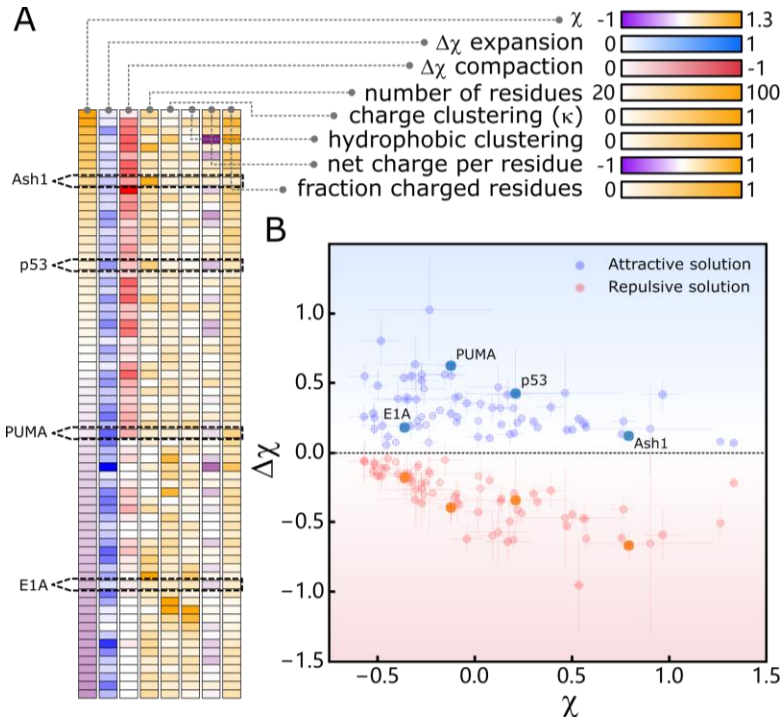

Figure 2.3. All-atom simulations of IDR sensitivity to solutions. (**A**) Heatmap of protein sensitivity and molecular features. Protein identity varies from top to bottom across cells, and molecular features vary left to right. Colormaps are shown for each molecular feature. (**B**) The magnitude $\Delta\chi$ 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.

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.[20] We selected 70 experimentally identified IDRs[84], 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[85–87]**.** We quantified the sensitivity of the protein to compacting or expanding solutions based on the extent of change in χ (**Fig. A1.3**). 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 $\Delta\chi = \chi([solute]) - \chi([solute] = 0)$ 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. A1.4**).

2.3.4 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 **Section A1.2.2**.[10,12]. 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. A1.5**). 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 **Section**

A1.2.3). 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. A1.6**). 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 $\Delta\chi$. Perhaps most importantly, it implies that while expanded or compact proteins display a wide range of sensitivities, IDRs where χ ~ 0 display a basal sensitivity to solution interactions. In this region, where most IDRs fall[9], even small changes to solution composition are predicted to have a measurable effect on IDR dimensions and/or residual structure.[20]

Under the assumption that cosolute-protein interactions scale linearly[77,88], 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. A1.7**). 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. A1.8**). Chain dimensions can also be represented using an apparent scaling exponent (ν^{app}) (see **Section A1.2.4**)[89,90]. 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[10].

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 aqueous buffer. IDR function through conformational selection has been reported for numerous proteins. In this mechanism, 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 which commonly occur in the cell can tune the function (or malfunction) of intrinsically disordered proteins.

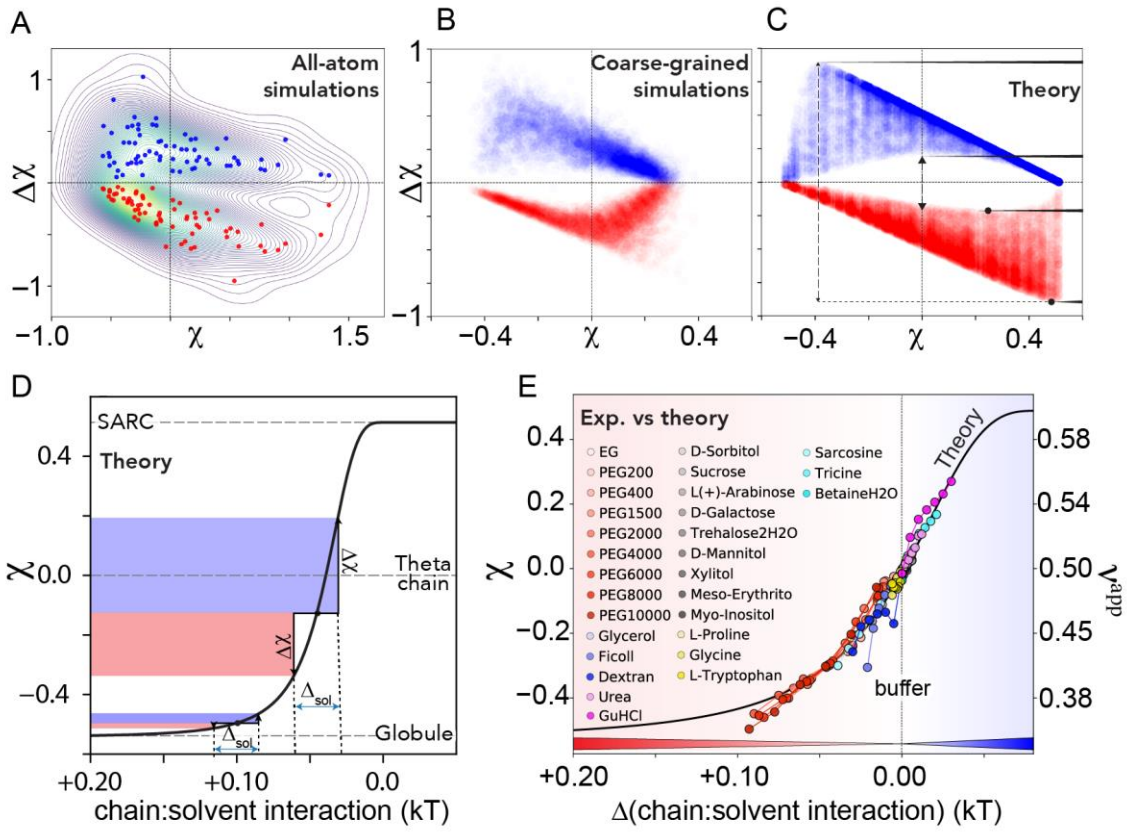

Figure 2.4. Using all-atom simulations to predict solution sensitivity of IDRs. (A-C). Density maps of all-atom simulations shown in **Fig. 2.3B**. (A), PIMMS coarse-grained simulations (B), and an analytical model (C) for solution sensitivity $\Delta\chi$ vs. dimensions in aqueous buffer χ. **(D)** 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 (Δsol) 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}. Mapping of other proteins is shown in **Fig. A1.7**.

2.4 Methods and supplementary figures

Experimental methods and supplementary figures for in this study are in Appendix 1. Full *in vitro* and simulation data sets can be found online at https://github.com/sukeniklab/IDP_structural_bias.

2.5. Author contributions

S.S. and A.S.H. conceptualized and led the project. D.M. designed and performed all *in vitro* experiments and analysis with the help of N.M.S. and P.K. A.S.H., F.Y. and G.G. designed, ran and analyzed simulations. S.S., D.M., F.Y. and A.S.H. wrote and revised the manuscript.

3. Structural biases in disordered proteins are prevalent in the cell

3.1 Abstract

Intrinsically disordered protein regions (IDRs) are essential to cellular function in all proteomes. Unlike folded proteins, IDRs exist in an ensemble of rapidly interchanging conformations. IDR sequences encode interactions that create structural biases within the ensemble. Such structural biases determine the three-dimensional shape of IDR ensembles and can affect their activity. However, the plasticity and sensitivity of IDR ensembles means that structural biases, often measured *in vitro*, may differ in the dynamic and heterogeneous intracellular environment. Here we reveal that structural biases found *in vitro* in well-studied IDRs are recapitulated inside human-derived cells. We further show that ensemble structural biases can change in a sequence-dependent manner due to changes in the intracellular milieu, subcellular localization, and interactions with tethered well-folded domains. We propose that the structural sensitivity of IDR ensembles can be leveraged for biological function, be the underlying cause of IDR-driven pathology, or be used to design disorder-based biosensors and actuators.

3.2 Introduction

Intrinsically disordered protein regions (IDRs) play key roles in many cellular pathways and are vital to cellular function in all kingdoms of life[64,91]. Compared to folded proteins, IDRs lack a stable tertiary structure, have fewer intramolecular interactions, and expose a greater area of their sequence to the surrounding solution[1]. As a result, an IDR exists in an ensemble of conformations that can change rapidly in response to the physical-chemical characteristics of its surroundings[19,20].

Despite being highly dynamic, IDR ensembles often contain structural biases, or preferences for certain subsets of conformations within the ensemble[2]. Such structural biases may arise from short- or long-range interactions within the protein sequence (**Fig. 3.1A**)[16]. An extensive body of work has established the importance of IDR structural biases to their function[18,39,64,65,92–94]. For example, local biases that form transient *α*-helical segments modulate binding affinity in PUMA[39] and p53[18,95] and the liquid-liquid phase separation properties of TDP-43[94]. Changes to long-range structural biases were found to influence IDR function in p53[96], BMAL1[97] and Myc[98]. Thus, uncovering the structural biases of IDR ensembles is a prerequisite for understanding IDR function[14,32,62,99]. With few exceptions[100–102], studies linking IDR ensemble structure to its function are performed *in vitro*. The differences between an aqueous buffer and the cellular

environment are dramatic[47,103], casting doubt as to whether or not structural biases linked to function *in vitro* persist in the cell.

The structural malleability of IDR ensembles, coupled with the dynamic nature of the cellular environment, prompts two major unanswered questions: (1) To what degree are IDR structural biases observed *in vitro* preserved inside the cell? (2) How do IDR structural biases respond to physical-chemical changes in the dynamic intracellular environment?

Answering these questions requires studying IDRs inside the cellular environment. However, resolving IDR structural biases inside the heterogeneous and dynamic cellular environment is especially challenging. Changes to the cellular physical-chemical composition occur regularly during the cell cycle[19,20,60,104] (e.g., the breakdown of the nuclear envelope during mitosis[48]). Alternatively, these changes may result from pathology, such as the elevation in intracellular pH and rewiring of metabolic pathways common to nearly all cancer cells[69,105]. Such compositional changes are known to affect even well-folded proteins[47,103,106,107], but their effect on IDR structural biases has not been studied.

Here we demonstrate that IDR structural biases observed *in vitro* persist in live cells, but display sequence-dependent sensitivities to physical-chemical changes in the cellular environment. Our observations rely on ensemble fluorescence resonance energy transfer (FRET). To obtain a structural metric for IDR ensembles, we place sequences of interest between two fluorescent proteins (FPs) that form a FRET pair, mTurquoise2 and mNeonGreen (**Fig. 3.1B**)[19,30,75]. Ensemble FRET provides, among other advantages, unmatched throughput and ease-of-use when working in live cells, but suffers from drawbacks when it comes to accurate quantification of distances. To mitigate these drawbacks, we have established a characterization pipeline that combines ensemble FRET (FRET, **Fig. 3.1C**), analytical size exclusion chromatography (SEC, **Fig. 3.1C**), small angle X-ray scattering (SAXS, **Fig. 3.1C**), changes in solution composition[19,20] (**Fig. 3.1D**), and molecular simulations to identify structural biases of IDRs *in vitro*. We then leverage this characterization to examine the same constructs inside live cells using FRET microscopy (**Fig. 3.1E**). Finally, we perturb the cellular ensembles by subjecting cells to osmotic challenges that rapidly change cell volume, and measure the response of IDR ensembles through changes in FRET signal (**Fig. 3.1F**).

We first validate our pipeline using dipeptide Gly-Ser (GS) repeats, establishing these sequences as homopolymer benchmarks that contain no significant structural biases[30,108,109]. We next compare the BH3 domain of PUMA, a naturally occurring IDR containing well-defined helical structural biases, against three variants where the wild-type sequence is scrambled. By scrambling a sequence, we confirm that structural biases are encoded by amino acid sequence, rather than amino acid composition. Finally, we investigate five well-studied naturally occurring IDRs whose conformational biases have been linked to their physiological function. We find that in all cases, the structural biases that define the ensemble *in vitro* also exist inside the cell. Furthermore, we highlight cases where IDRs respond in a sequence-dependent manner to a changing environment: through osmotic challenges, changes in subcellular localization, or interaction with a folded domain, some IDRs show sequence-encoded ensemble changes that are not observed in GS repeats.

Our work offers clear evidence that sequence-encoded structural biases exist in IDR ensembles in living cells, and that these biases can be tuned by changes to protein sequence or to the cellular environment. The existence of structural biases in IDR ensembles inside the cell suggests that they are subject to evolutionary pressure, and that IDRs can be rationally designed to create disorder-based sensors and actuators.

3.3 Results

3.3.1 Glycine-serine repeats are an unbiased, model-free standard to quantify IDR ensembles

The structure of a folded protein is commonly described in terms of its “native” conformation discerned through X-ray crystallography. For an IDR, no single structure can be obtained. Instead, IDR structure is often described with reference to well-established homopolymer models[22,59]. However, no models exist for our dumbbell-shaped FRET construct, especially not models that are relevant in the cellular environment. We therefore wanted to create an empirical standard against which we could compare IDRs of arbitrary lengths.

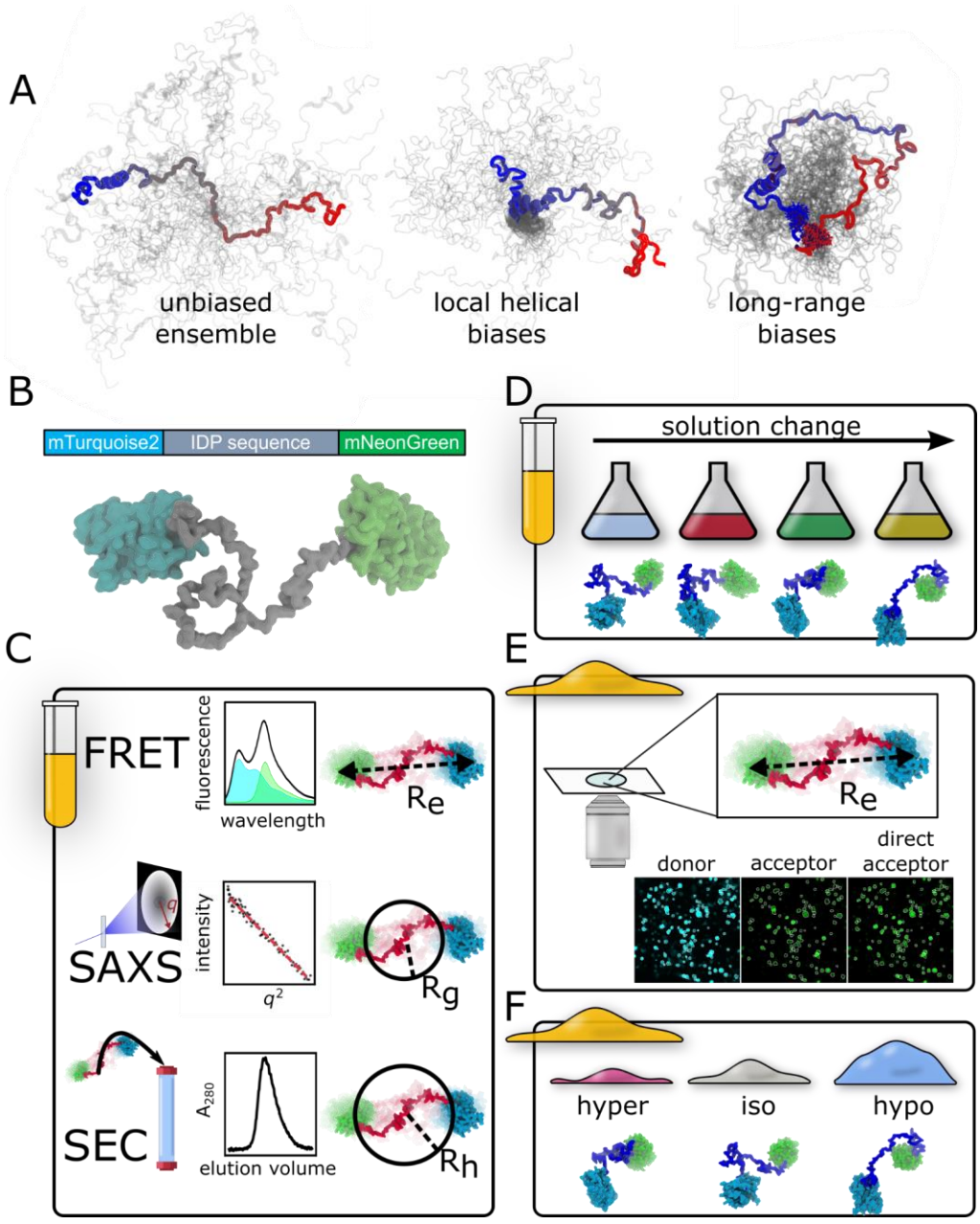

Figure 3.1. Diagrams of methods used in this study. (A) IDR ensembles with and without structural biases. In all schemes, a single conformation is shown in color and other conformations are shown in gray. Structural biases increase the density in specific regions of the ensemble and alter its average dimensions. (**B**) FRET construct consisting of an IDR between two fluorescent proteins that serve as a FRET donor and a FRET acceptor. **(C)** *In vitro* experiments. Top: fluorescence resonance energy transfer (FRET). Middle: small angle X-ray scattering (SAXS). Bottom: analytic size-exclusion chromatography. **(D)** *In vitro* solution space scanning measures the FRET signal of a sequence in the presence of denaturing (urea, guanidinium), stabilizing (sucrose, ethylene glycol, glycine, sarcosine), and crowding (PEG400, PEG2k, Ficoll) solutes, as well as salts (NaCl, KCl) that screen electrostatic interactions. **(E)** Live-cell FRET microscopy is performed on

cells expressing the same constructs used *in vitro*. **(F)** Changes in ensemble dimensions are measured using live-cell FRET following rapid hyperosmotic and hypoosmotic challenges.

As a benchmark against which to compare properties of naturally occurring heteropolymeric IDRs, we inserted homopolymeric dipeptide repeats into our FRET construct (**Table A2.1**). Specifically, we chose glycine-serine (GS) repeats for benchmarking because (1) they lack hydrophobicity, charge, and aromaticity which makes them easy to express and highly soluble[2], (2) they have been shown to lack local and long-range structural biases, instead behaving as expected for a random coil across the range of lengths studied in our work[109,110], and (3) they have been shown to behave as real-chain mimics of ideal Gaussian chains in aqueous solutions[110–112].

Ensemble FRET experiments provide an apparent FRET efficiency (E_f^{app}), which is inversely proportional to the distance between the two FPs. When the FPs are close together, E_f^{app} is high, and when they are far apart, E_f^{app} is low. Increasing and decreasing E_f^{app} can indicate compaction and expansion of an ensemble. As previously reported, E_f^{app} decreased linearly with the number of GS repeats in a dilute buffer solution[19] (**Fig. 3.2A,B, A2.1**). However, the three-dimensional structure of the ensemble cannot be resolved by a single-distance measurement[40,41,113].

To obtain additional, orthogonal measurements that can inform about the structure of the ensemble, we performed size-exclusion chromatography coupled with small-angle X-ray scattering (SEC-SAXS) on the same constructs as we measured using FRET[114–116]. The chromatograms obtained from SEC showed a consistent, linear, size-dependent increase in elution volume (**Figs. 3.2C,D, A2.2**), indicating that the proteins increase in dimension with GS repeat length. Analysis of SAXS intensity curves showed a similar linear dependence on GS length (**Figs. 3.2E,F, A2.3, A2.4**), displaying linearly increasing radii of gyration (R_g, **Fig. 3.2F**) in agreement with our other results.

Finally, we conducted all-atom simulations of all GS-repeat sequences to enable a molecular benchmark between SAXS and FRET results. Our simulations assumed that the FPs only take up space (i.e., are non-interacting) and that GS repeats behave like homopolymers. From these simulations, ensembles were selected to quantitatively match the SAXS scattering data (**Fig. A2.5**). These ensembles reproduced the GS length-dependent E_f^{app} values as well, indicating

that the simulation conditions at least managed to reproduce our experimental results (**Fig. 3.2B,F**). The application of one experimental dataset as a constraint to assess simulations against an orthogonal experimental dataset has been used previously to assess unfolded protein ensembles to great effect[116,117].

Taken together, our methods consistently show the same length-dependent trend for the GS repeats, and that the length of the sequence, rather than, e.g., intramolecular interactions between or with FPs, is the dominant factor affecting these dimensions. The excellent quantitative agreement with our simulations further indicates that GS repeats behave like ideal homopolymers, which lack structural biases[40,113].

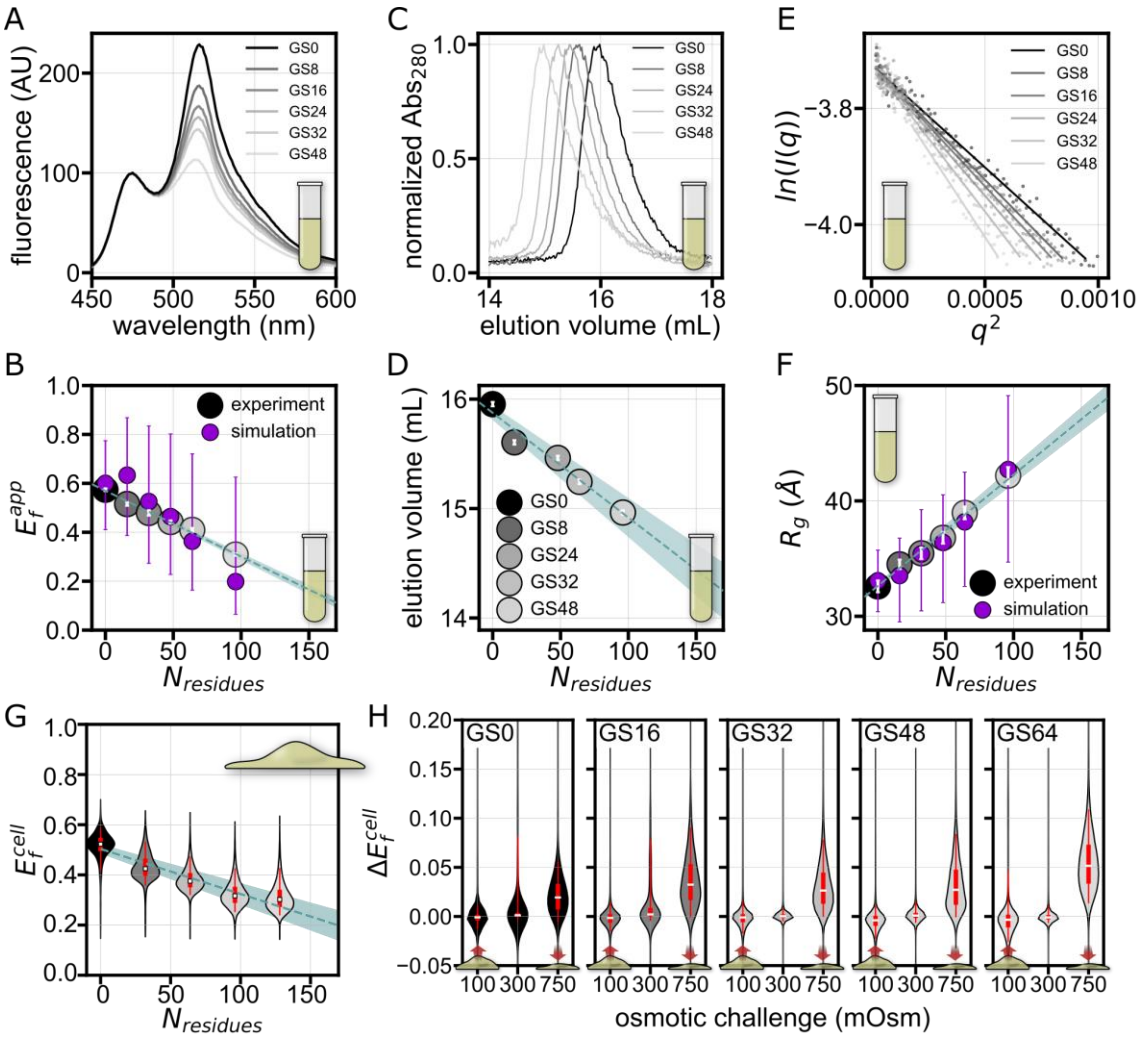

Figure 3.2. Characterization of GS repeat standards. (A) Fluorescence spectra from *in vitro* measurements of FRET GSX constructs, where X indicates the number of Gly-Ser repeats. **(B)** Average apparent *in vitro* FRET efficiencies (E_f^{app}) of GS repeats. Error bars represent the SD

(N=12). Here and in **Fig. 3.2D**, **3.2F** and **3.2G**, dashed lines represent expected values for GS repeats of corresponding lengths based on a linear fit. Here and elsewhere, the blue shaded region represents the standard error of the linear fit. All-atom simulations of GS repeats are shown in purple, with error bars representing the median 50% of simulation results. **(C)** SEC chromatograms for GS repeats. **(D)** SEC elution volumes, expressed as the position of the peak in mL, vs. number of residues in the GS-repeat sequence. Error is assumed to be one frame in each direction. **(E)** Guinier regions and fitted lines from SAXS experiments for GS repeats. **(F)** Radii of gyration (R_g) derived from Guinier analysis of SAXS data for GS repeats. White error bars represent error from fitting lines to Guinier plots. The same all-atom simulations of GS repeats shown in **Fig. 3.2B** are used to calculate the simulated R_g, shown in purple, with purple error bars representing the median 50% of simulation results. **(G)** FRET efficiencies of GS repeats measured in live cells (E_f^{cell}). In all live-cell results, violin plots span the entire dataset and their thickness represents the probability. The median is shown as a white square, and the median 50% and 95% are shown as thick and thin lines at the center of the violin, respectively. The blue line is a linear fit of the medians, and fit error is shown by the shaded region. **(H)** Response to osmotic challenge expressed as change in E_f^{cell} from before to after the challenge (ΔE_f^{cell}).

To further verify that GS repeats do not contain structural biases, we conducted FRET-based solution space scanning of GS-repeat constructs[19,20]. Solution space scanning probes structural biases in the ensemble by modulating interactions between the sequence and the solution. We reason that if structural biases exist, different GS-repeat lengths will show a different structural response to the same solution. The solutes we added as chemical probes in this case were salts, amino acids, polymeric crowders and their monomeric units, and denaturants. We stress that these solutes were not intended to directly mimic the cellular environment, but rather to probe the response of the ensemble to changes in solution chemistry. We measured the change in FRET efficiency $\Delta E_f^{app} = E_{f,solute}^{app} - E_{f,buffer}^{app}$ for all GS repeat lengths (**Fig. A2.6**). As expected, GS repeats of all lengths responded identically to each of the solution conditions we created (**Fig. A2.6**). Overall, the internal consistency of the results from our orthogonal characterization methods establishes GS repeats as a model-free homopolymer standard which lacks structural biases.

3.3.2 Live-cell measurements recapitulate *in vitro* results for GS repeat ensembles

We next sought to establish GS repeats as a bias-free standard in live cells. To facilitate direct and straightforward comparison with our *in vitro* experiments, we used the same genetically encoded FRET constructs as we had used *in vitro*. GS-repeat FRET constructs were expressed in HEK293T cells, which all showed

similar morphology and expression levels regardless of the construct being expressed (**Fig. A2.7**).

Our live-cell measurements of GS repeats showed trends in FRET efficiency calculated from live-cell imaging (E_f^{cell}) that are in quantitative agreement with *in vitro* measurements (**Fig. 3.2B,G**). Notably, in live cells our FRET constructs showed a much broader distribution of E_f^{cell} compared to the distribution of E_f^{app} shown *in vitro*. This variability may be caused by a range of factors, including cell-to-cell differences in composition, cell state, and construct expression levels. Despite this, the remarkable agreement with *in vitro* data indicates that the lack of structural biases for GS repeats detected *in vitro* persists inside live cells.

To test whether GS ensemble dimensions are sensitive to the cellular environment, we subjected cells to osmotic challenge. To resolve their immediate effects on a protein, these perturbations are performed rapidly and measured as quickly as possible to prevent any kind of transcriptional response[54,55]. We use rapid osmotic challenges induced by the addition of NaCl (hyperosmotic, 750 mOsm) or water (hypoosmotic, 100 mOsm) to media (isosmotic, 300 mOsm). Osmotic challenges were previously shown to produce robust and reproducible changes in cellular volume through the efflux or influx of water[54,55,118]. We report on the difference in FRET signal of each cell following this perturbation, $\Delta E_f^{cell} = E_{f,after}^{cell} - E_{f,before}^{cell}$. The measurements before and after the challenge are collected within a span of 45 seconds or less (**Fig. 3.1F**).

Hyperosmotic perturbations resulting in cell shrinkage caused a positive ΔE_f^{cell} that scaled with the length of the construct (**Fig. A2.8**). This is in line with previous studies of IDRs in crowded conditions and in the cell[59,104,118], and can be explained by the increased ability of longer sequences to compact. Hypoosmotic perturbations, on the other hand, produced no significant change in E_f^{cell} (**Fig. A2.8**). This lack of response was surprising, especially considering the fact that GS polymers are capable of expansion *in vitro* (**Fig. A2.6**). Regardless, our osmotic challenge experiments define a standard for the response of bias-free IDR ensembles to osmotically induced changes in cellular volume.

3.3.3 Amino acid sequence determines IDR structural biases and their response to changes in solution composition

Having established a reliable homopolymer standard *in vitro* and in live cells, we set out to investigate how a naturally occurring IDR compares with GS repeats. We chose the sequence of the PUMA BH3 domain (wild-type (WT) PUMA) (**Fig. 3.3A,B, Table A2.1**) because its residual helicity is a well-studied example of functionally linked structural biases in IDRs[23,39]. We first established the previously reported short-range helical structural biases of the unlabeled sequence[119] as indicated by the characteristic double minima in the circular dichroism (CD) spectrum (**Fig. 3.3B,C**). Next, we measured the E_f^{app}, R_g, and SEC elution volume of WT PUMA using our *in vitro* pipeline (WT in **Fig. 3.3D-F**). Although in SEC WT PUMA eluted near the same volume as would be expected of GS repeats of the same length (**Fig. 3.3E**), SAXS and FRET showed WT PUMA to be significantly more compact than corresponding GS repeats (**Fig. 3.3D,F**), confirming that we are able to detect local structural biases present in WT PUMA but absent in GS repeats.

Is residual helicity like that observed in WT PUMA a prerequisite for detectable structural biases? To answer this question, we generated sequence scrambles of WT PUMA (**Fig. 3.3A, Table A2.1**) and measured their ensembles *in vitro*. Sequence scrambles retain the amino acid composition but change their order, disrupting structural biases present in the wild type[40,41]. The three scrambles of WT PUMA were designed to have varying degrees of charge clustering, as measured by the parameter κ (kappa) in CIDER[120] (sequences S1-3, **Fig. 3.3A,B**). To test for the existence of helical structural biases in the scrambled sequences, we measured the secondary structure of the label-free IDRs using CD. As expected, the CD spectra of the scrambles showed no double minima (**Fig. 3.3C, A2.9**), indicating that the helical structural biases of WT PUMA were no longer present.

We next characterized ensemble dimensions of the scrambles using FRET (**Fig. 3.3D**), SEC (**Fig. 3.3E**), SAXS (**Fig. 3.3F**), and all-atom Monte Carlo simulations (**Fig. A2.10**). FRET and SAXS show that not only are the scrambles more compact than a GS repeat of the same length, but they also all differ from each other despite having similar CD spectra and identical amino acid composition (**Fig. 3.3A-C**). The overall agreement between trends from FRET and SAXS measurements shows that the WT PUMA sequence ensemble is the most compact, followed by S2, S3, and finally S1. This trend is also recapitulated in label-free all-

atom simulations, indicating that tethering to the two fluorescent protein labels does not change the trends in ensemble dimensions for this measurement (**Fig. A2.10**). SEC data shows a different trend, with all sequences appearing more expanded than a GS linker and S3 showing an almost equal compaction to the WT (**Fig. 3.3E**). This may be due to chemical interactions between the constructs and the SEC column matrix[121]. However, since all four sequences contain the same amino acid composition, even these different interactions indicate sequence-dependent structuring within the ensemble.

The differences shown by all methods, between WT PUMA and the three scrambles and also among the scrambles, demonstrate not only that the WT PUMA ensemble is uniquely more compact than the scrambles, but also that structural biases exist even in the absence of the helical structural biases in the WT sequence. These results also show that, in this case, charge patterning alone does not dictate ensemble dimensions, since S3 has similar patterning to WT but is significantly more expanded according to FRET and SAXS results (although not according to SEC measurements).

We hypothesized that different structural biases in PUMA and its scrambles would also manifest in their response to different solutions. To test this, we performed solution space scans for all four PUMA variants (**Fig. 3.3H, A2.11**). We compare ΔE_f^{app} of each sequence against the interpolated ΔE_f^{app} of GS repeats of the same length in the same solution condition (**Fig. 3.3H, A2.12**). Deviations from ΔE_f^{app} of length-equivalent GS repeats indicate higher/lower sensitivities of the sequences (indicated by red/blue backgrounds, respectively) (**Fig. 3.3H**). We were surprised to find that despite having the most compact ensemble, WT PUMA showed the highest sensitivity of all scrambles (as indicated by the stronger blue and red backgrounds in **Fig. 3.3H**). Specifically, the WT sequence displayed stronger compaction in response to polymeric crowders (specifically PEG2000) and stronger expansion in response to denaturants (urea and GuHCl) compared to both the corresponding GS-repeat sequence and the three sequence scrambles. The three scrambles showed milder responses, with a notable difference for S2, which was significantly less sensitive to all solutes. These differences indicate that IDRs possess sequence-encoded sensitivity to the chemical composition of their environment. Furthermore, the presence of structural biases does not preclude ensemble sensitivity to the surrounding solution, and may even amplify it.

3.3.4 Sequence-dependent structural biases seen *in vitro* persist in live cells

We next wanted to see if the structural biases measured *in vitro* for WT PUMA and its scrambles were retained in the cellular environment. We expected helical structural biases to persist in the cell due to the intrinsic stability of helical secondary structure[122], but reasoned that biases within the scrambled sequences were weaker and therefore might not be retained. To test this, we performed our live-cell FRET imaging experiments on WT PUMA and the three variants (**Fig. 3.3G**). Our live-cell FRET experiments showed striking agreement with the FRET measurements done in dilute aqueous buffers (**Fig. 3.3D**). Specifically, both the relative magnitude and the trend in E_f^{app} measured *in vitro* was replicated in live cells, with WT > S2 > S3 > S1. Overall, E_f^{cell} reveals that the structural biases found in these sequences *in vitro* persist inside the cell, even in the absence of short-range helical structural biases (which occur only in WT).

Our next goal was to measure whether these ensembles differ in their response to changes in the cellular environment. We again used osmotically-triggered cell volume perturbations as a means to reproducibly change the concentration of all cellular solutes. The change in value of each cell's average FRET signal following osmotic challenge, ΔE_f^{cell}, is reported and compared with the interpolated ΔE_f^{cell} for a GS-repeat of the same length (**Fig. 3.3I**). We were surprised to find that the WT sequence, which displayed more sensitivity than a corresponding GS-repeat sequence to certain solutes *in vitro*, showed a response similar to that of GS repeats under both cell volume increase and decrease. Remarkably, this similarity to GS-repeat sensitivity in live cells was seen in all sequences except S2, which displayed a markedly lower tendency to compact under hyperosmotic conditions (as indicated by the lack of overlap between the median 50% of the data and the GS-repeat equivalent). The lower sensitivity of S2 was also observed *in vitro* (**Fig. 3.3H**). This result indicates that IDR ensemble sensitivity to changes in the cellular environment is encoded in sequence, but is difficult to predict since it may or may not correlate with the sensitivity measured in dilute buffers.

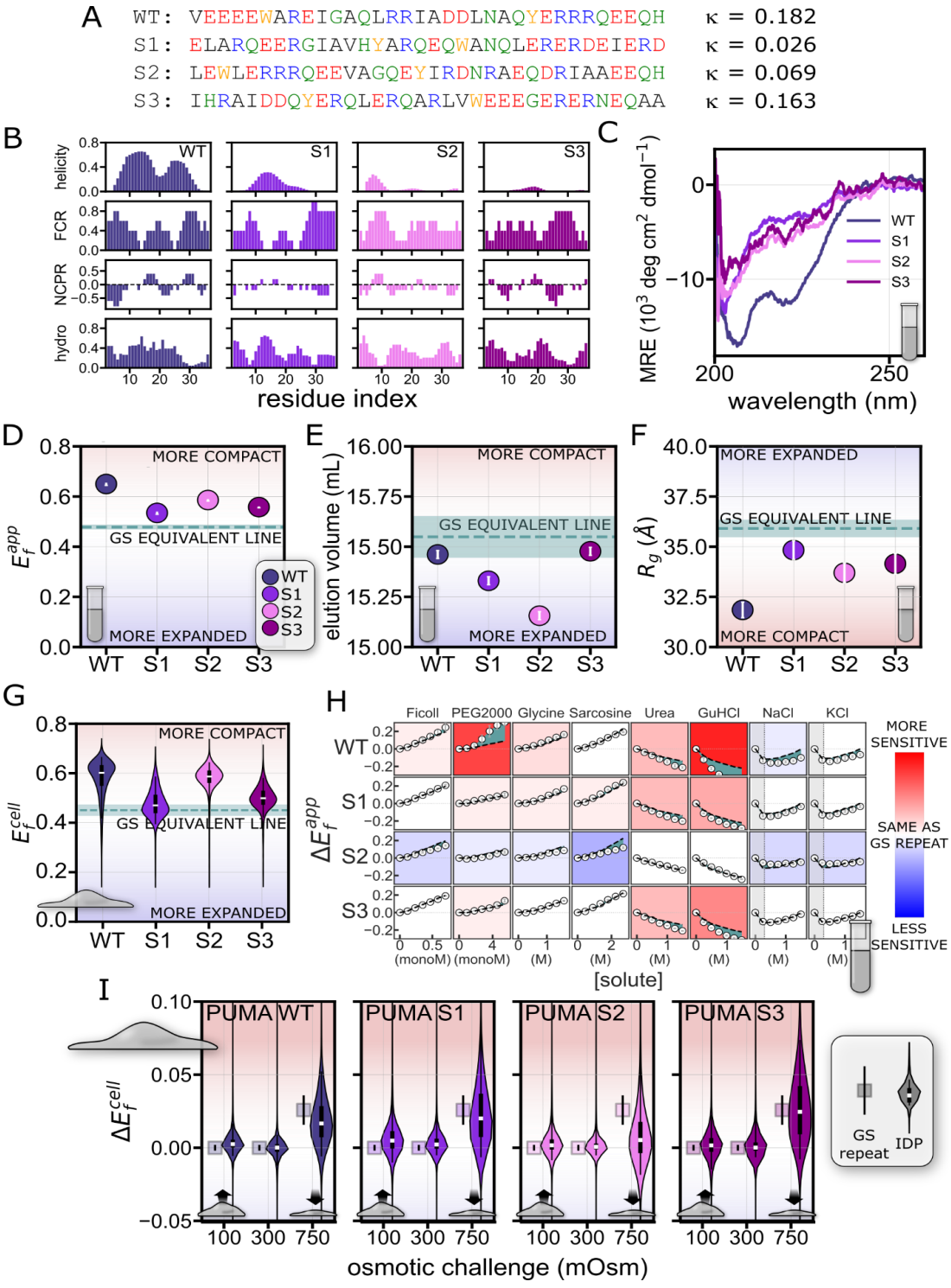

Figure 3.3. Comparison of global dimensions and solution sensitivity of GS repeats and PUMA variants. (A) Sequence of wild-type PUMA BH3 domain (WT PUMA) and three sequences (S1, S2, S3) derived by shuffling WT PUMA's sequence. Red: negative charge; blue: positive charge; black: hydrophobic; green: polar; orange: aromatic. κ measures charge clustering in the

sequence (positive indicates a better separation of charges across the sequence) **(B)** Molecular features of WT PUMA and shuffles. Predicted helicity was calculated using all-atom simulations. Other parameters were evaluated by localCIDER[120]. FCR: fraction of charged residues. NCPR: net charge per residue. hydro: Kyte-Doolittle hydrophobicity. Value at each position on the x axis represents the average of the indicated residue and its four nearest neighbors. **(C)** CD spectroscopy signatures of PUMA variants without flanking FPs. See also **Fig. A2.9**. **(D)** Average E_f^{app} of PUMA constructs. Error bars represent the SD (N=12). In all panels the horizontal dashed line represents the expected value for a GS-repeat construct of the same length (34 residues). **(E)** SEC elution volume for PUMA constructs. Error is assumed to be one frame in each direction. **(F)** SAXS-derived R_g of PUMA constructs. Error bars represent error from fitting lines to Guinier plots. **(G)** E_f^{cell} of PUMA constructs. Features are as in **Fig. 3.2G**. **(H)** Solution space scans of PUMA constructs, with results expressed as ΔE_f^{app}. White circles: ΔE_f^{app} of IDR. Black dashed lines: interpolated ΔE_f^{app} of a GS-repeat sequence of the same length as the IDR (**Fig. A2.12**). Green shaded regions are differences between ΔE_f^{app} of IDR and GS repeats. Red/white/blue background shows more/same/less sensitivity (more expansion or compaction) than a GS-repeat sequence of the same length, respectively. Shaded regions on left side for solutes NaCl and KCl: approximate range of concentrations within which electrostatic screening is the dominant effect; the leftmost two points of each series, since they are within that range, are not used in the assignment of background color. **(I)** Osmotic challenge of HEK293T cells expressing PUMA constructs. Violin plots represent the data for PUMA constructs and squares represent ΔE_f^{cell} of a GS-repeat equivalent. Features are as in **Fig. 3.2G,H**.

3.3.5 Structural biases in naturally occurring IDRs persist inside the cell

Having seen that structural biases *in vitro* persist inside the cell for PUMA and its scrambles, we wanted to see whether this is a general property of other IDR sequences. We inserted a range of well-studied IDRs of different lengths into our construct and characterized them *in vitro* and in live cells. We tested the N-terminal disordered region of p53 (residues 1-61, p53)[18] which contains the N-terminal activation domain (NTAD)[18], the low-complexity domain of FUS (residues 1-163, FUS)[123], the N-terminal region of the adenovirus hub protein E1A (residues 1-40, E1A)[76], and the C-terminal region of the yeast transcription factor Ash1 (residues 418-500, Ash1)[12] (**Fig. A2.13, Table A2.1**). Importantly, the ensemble structure of each of these IDRs has previously been characterized *in vitro* and has been shown or proposed to determine IDR function (see Discussion).

Using our *in vitro* characterization pipeline, we found clear divergence in nearly all constructs from GS repeats. Our FRET experiments show that three sequences (PUMA, E1A, FUS) are more compact than a GS-repeat sequence of the same dimensions (indicated by E_f^{app} averages whose standard deviation is above the GS line, **Fig. 3.4A**). The two that fell close to the GS line, p53 and Ash1,

have been reported to be relatively expanded in other studies[12,18]. A similar trend was observed for SAXS-derived R_g values (**Fig. 3.4C**). SEC data (**Fig. 3.4B**) shows mostly similar trends, though PUMA, E1A, and p53 appear to be more expanded than GS repeats. As before, the deviations from the GS-equivalent line, together with the changes in trends between characterization methods, highlight the differences in structural biases between different IDR sequences.

Our next goal was to determine the extent to which the structural biases observed *in vitro* for these constructs persist in the cell. As before, we expressed the same constructs in HEK293T cells, and used live-cell imaging to quantify the in-cell FRET efficiency, E_f^{cell} (**Fig. 3.4D**). Good agreement was observed between E_f^{app} measured *in vitro* and the in-cell E_f^{cell} values (**Fig. 3.4A,D**). As before, this agreement indicates that structural biases that determine IDR ensemble shape *in vitro* largely exist inside the cell.

We next wanted to see how the localization of IDRs in the cell might affect their ensembles. We reasoned that different organelles have different physical-chemical compositions, and this may affect the ensemble preferences encoded in IDR sequences[124]. To test this idea, we measured E_f^{cell} in the cytoplasm and nucleus of U-2 OS cells for all our sequences. GS repeats showed the same E_f^{cell} in both cytoplasm and nucleus within error, indicating their ensemble is unaffected by changes in localization (**Fig. A2.14**). All E_f^{cell} measurements were normalized to a GS repeat of the same length (**Fig. 3.4E**). Most sequences showed no significant difference between the cytoplasm and the nucleus. This is in line with our results thus far: if moving these sequences from aqueous buffers to the cellular environment induced little change in ensemble structure, we expect the same to happen moving from the cytoplasm to the nucleus. An exception was observed for the FUS low-complexity domain which was significantly more expanded in the nucleus ($p \leq 0.0001$, **Fig. 3.4E**). This might be due to its ability to interact with nuclear-abundant RNA[125,126].

3.3.6 Naturally occurring IDRs differ in their sensitivity to solution changes

Next, we performed solution space scanning on PUMA, FUS, p53, Ash1 and E1A (**Fig. 3.4F, A2.15**). In particular, we wondered whether Ash1 and p53, despite their similarity in dimensions to GS repeats of the same length, would display different behavior in different solutions. As expected, different sequences showed markedly different sensitivities to the solutes used. PUMA and Ash1

showed an outlying degree of sensitivity, with larger changes compared to GS repeats of the same length in both compacting and expanding solutes, while E1A appeared to be less sensitive to the same solutes (**Fig. 3.4F**). The response to salts also showed deviations, with less response to high salt concentrations for E1A (indicated by the constant value in ΔE_f^{app} at concentrations above 0.25 M salt). Interestingly, p53, whose dimensions were closest to those of its GS equivalent in dilute buffer (**Fig. 3.4A**), also displayed sensitivity most similar to its GS equivalent (**Fig. 3.4F**). In line with our previous results[19], we found that PEG 2000 produces greater increases in E_f^{app} than the smaller PEG 400 at equal monomer-molar concentrations, and that the monomer units of the crowders (sucrose, ethylene glycol) produce relatively small changes in the dimensions of the IDRs. This wide range of responses to changes in solution conditions further supports the existence of sequence-dependent structural biases found in our FRET, SAXS, and SEC results. Moreover, the different IDR ensembles show differing and specific sensitivities to changes in their chemical environment.

Finally, we wanted to measure the response of these IDRs to changes in intracellular composition. We subjected cells to hypoosmotic or hyperosmotic challenges and followed the changes in average FRET signal for each cell, ΔE_f^{cell} (**Fig. 3.4G**). We compare these to the changes expected for GS repeats of the same length, shown as the squares adjacent to each violin plot. We observe that some sequences behaved as expected from GS repeats—namely PUMA, Ash1, FUS, and p53 all fall within the range expected of a GS-repeat equivalent. FUS displayed a similar behavior to GS repeats upon hyperosmotic challenge, but showed an outlying ability compared to the other naturally occurring IDRs to expand in hypoosmotic conditions. However, most striking was E1A's response to cellular perturbations. Expansion of IDRs under increased crowding has been previously reported *in vitro*[68], and may be caused inside the cell by protein-protein interactions such as chaperone binding[127] or post-translational modifications[71].

Taken together, these results show not only that structural biases in IDR ensembles exist both *in vitro* and inside the cell, but also that IDR ensembles are able to sense and respond to changes in the composition of their environment. This ability is encoded in sequence, and occurs both in the test tube and in the cell. However, despite the agreement between IDR structural biases in a dilute solution *in vitro* and in isosmotic conditions in the cell, comparing *in vitro* and in-cell solution sensitivity is not straightforward.

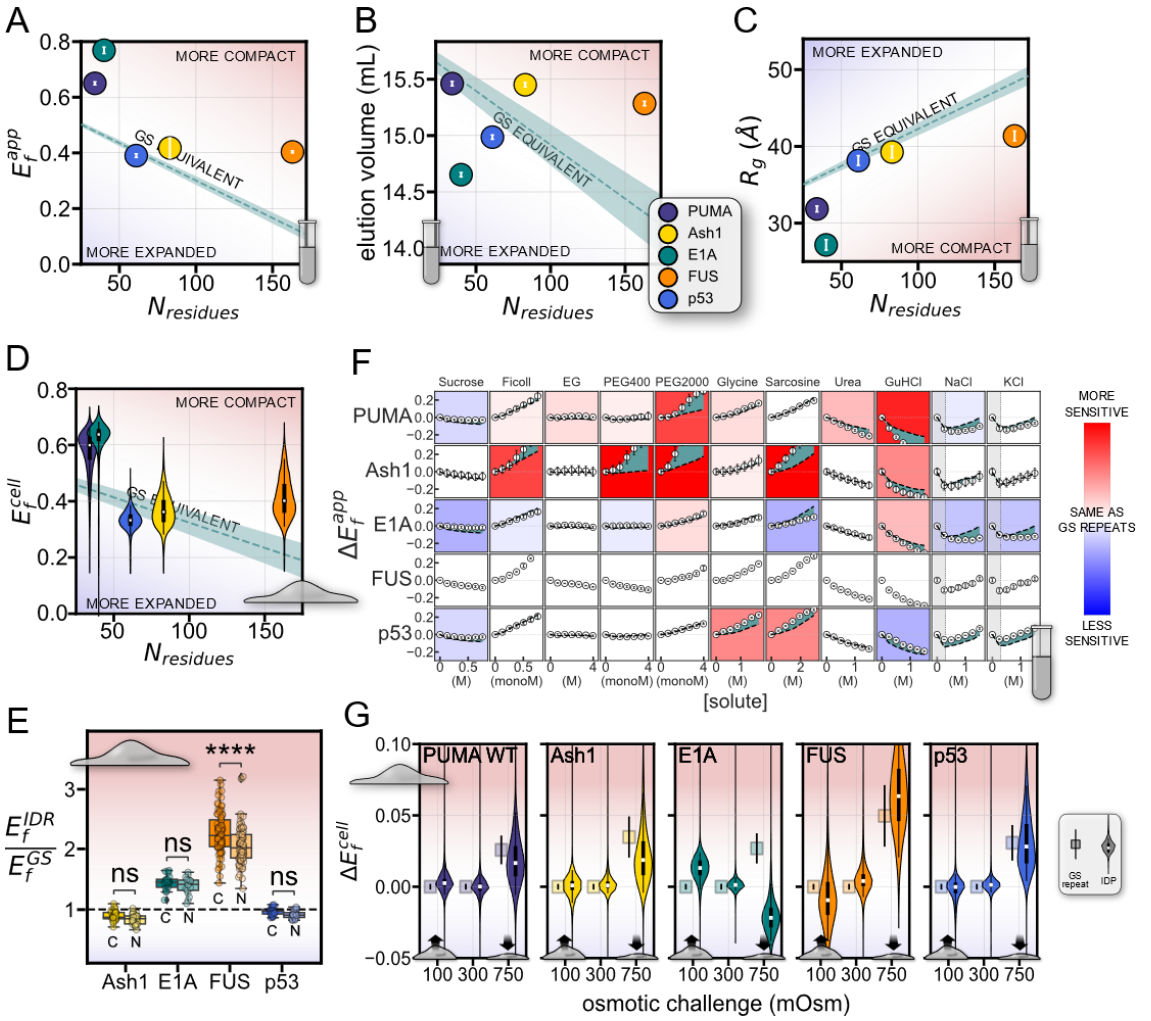

Figure 3.4. Comparison of global dimensions and solution sensitivity of GS repeats and naturally occurring IDRs. (A) Average E_f^{app} of IDR constructs. Error bars represent the SD (N=12). Here and in Fig. 3.4B-D, the dashed line represents expected values for GS repeat sequences of corresponding lengths. (B) SEC elution volume of IDR constructs. Error is assumed to be one frame in each direction. (C) R_g of IDR constructs. Errors are from fitting lines to Guinier plots. (D) E_f^{cell} of IDR constructs. Features are as in Fig. 3.2G. (E) E_f^{cell} of four IDR constructs measured in the cytoplasm (C) and nucleus (N) of U-2 OS cells and normalized to the E_f^{cell} of an equivalent GS linker. Each box represents the 25th and 75th percentiles with the median shown as the black line and the whiskers showing the minimum and maximum for each construct. Each circle corresponds to a single cell. Asterisks denote the significance between distributions determined by a Mann-Whitney test (**** indicates $p < 0.0001$). (F) Solution space scans of IDR constructs. Features are as in Fig. 3.3H, except for FUS for which the GS repeat trends could not be accurately extrapolated. (G) Osmotic challenge of IDR constructs. Features are as in Fig. 3.2G,H.

3.3.7 Interactions between IDRs and their tethered folded domains

One alternative possibility that could explain the aberrant behavior of E1A is that the IDR interacts with one or both of the FPs in our FRET construct, and that cellular perturbations disrupt this interaction. To test whether IDR ensemble structural biases are influenced by interactions with the tethered FPs, we repeated our FRET experiments using constructs with the locations of the FPs flipped from their original locations (**Fig. 3.5A**). We reasoned that since the surface of each FP (**Fig. A2.16A**), their termini (**Fig. A2.16B**), and the termini of the IDR differ, changes in FRET signal in the flipped vs. the original construct would indicate the involvement of interactions between the IDR and the FPs in determining E_f^{app}.

As with previous experiments, we first started with a GS-repeat sequence. In this case, the IDR termini are identical, and any difference would be a result of changes in the FPs themselves rather than a difference in IDR:FP interactions. Our *in vitro* measurements showed a higher E_f^{app} for the flipped GS16 construct, indicating a more compact conformation (**Fig. 3.5B**). Further NaCl titration experiments and analysis of raw fluorescence spectra showed that (1) electrostatic interactions do not account for the difference in E_f^{app} (**Fig. A2.17A-C**), and that (2) the difference in E_f^{app} between the original and flipped construct is likely a result of changes in the structure of the mNeonGreen tail tethered to the IDR (**Fig. A2.17D**). As described above, our analysis indicates that E_f^{app} of GS-repeat homopolymers is not driven by IDR:FP interactions. When measured in live cells, flipped GS16 again displayed similar results to those seen *in vitro*, with a higher E_f^{cell} for the flipped GS16 construct (**Fig. 3.5C**).

We next compared the basal *in vitro* E_f^{app} and live-cell E_f^{cell} distributions of the original and flipped versions of three previously measured constructs (**Fig. 3.4G**): E1A (whose original version had shown a significantly different response than GS repeats to osmotic challenge), Ash1 (whose original version had only shown a difference in hyperosmotic conditions compared to GS repeats), and p53 (whose original version had shown a similar response to GS repeats to osmotic challenge). Unlike GS16, these naturally occurring IDRs contain different sequences at their N and C termini, as well as charged residues that could contribute to electrostatic interactions between the FPs and IDR. Both *in vitro* and in cells, E1A displayed a dramatic reduction in FRET efficiency, while flipped Ash1 and p53 showed little change compared to the original constructs (**Fig. 3.5D**). This points to interactions between one or both of the FPs and the sequence of E1A.

As with GS16, further tests of emission peak wavelengths implicated mNeonGreen as the FP with significant changes to spectra upon tethering. It also showed a different trend in peak wavelength shift for E1A compared to Ash1 and p53 (**Fig. A2.17D**).

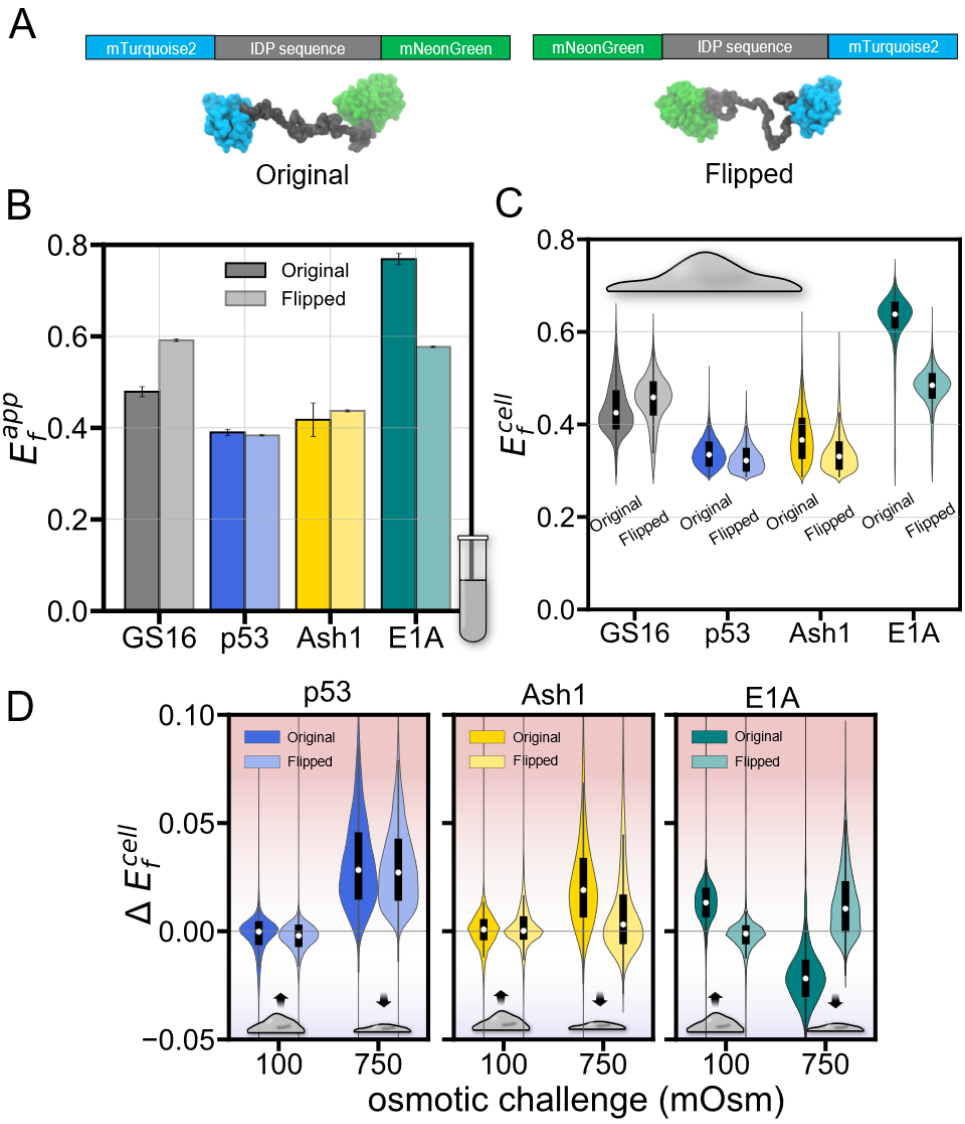

Figure 3.5. Determination of FRET pair influence on IDR ensemble dimensions. (**A**) Original FRET construct (left) consisting of an IDR between two fluorescent proteins that serve as a FRET donor and a FRET acceptor and the flipped construct (right) with the FRET pairs on the opposite ends. (**B**) E_f^{app} of selected constructs measured *in vitro* for the original (darker color) and flipped (lighter color) constructs. (**C**) E_f^{cell} of selected constructs measured in HEK293T cells for original (darker color) and flipped (lighter color) constructs. (**D**) Response to osmotic challenge of each construct expressed as change in E_f^{cell} before and after the challenge (ΔE_f^{cell}).

We hypothesized that if there were significant changes to the ensemble in the flipped construct, it would also alter the response to changes in cell volume. Testing this, we indeed found that p53, but not Ash1, displayed similar responses to changes in cell volume (**Fig. 3.5E**). This is despite p53 and Ash1 having similar dimensions between the original and flipped constructs. E1A, on the other hand, showed a completely opposite response between the flipped and original constructs (**Fig. 3.5E**). These results indicate that IDR:folded-domain interactions can alter the ensemble's response to changes in the cellular environment. But regardless of these differences between the constructs, the ensemble dimensions as measured by FRET efficiency remain similar *in vitro* and in the cell.

3.3.8 Limitations and drawbacks

One drawback of this work is the use of fluorescent proteins (FPs) in our constructs. There are many advantages to genetically encoded FRET constructs. They can be produced easily in *E. coli* with no need for further labeling, or transiently or stably expressed in any genetically tractable cell line. Additionally, the FPs flanking the sequence increase solubility and signal from scattering methods, and hinder aggregation and phase separation. However, as indicated for E1A, the presence of bulky folded domains tethered to the IDR of interest may affect our results through intramolecular interactions of the FPs with each other or with the IDR sequence.

Nonetheless, concerns regarding artifacts from our use of FPs are mitigated by (1) the use of the same FPs for all constructs and the comparison against GS-repeat constructs, which facilitate meaningful comparison between all sequences; and (2) the agreement between our experiments and all-atom simulations of the GS-repeats (**Figs. 3.1B, 3.1F,** and **A2.10**).

As a final point, we acknowledge that interactions between the studied IDRs and the FPs that make up our FRET construct exist and likely affect the dimensions of our measured ensembles. To address this, we point to the fact that nearly all studied IDRs (including those in this work) are excised from full-length proteins in which they would be tethered to folded domains. Our results point to the importance of the intramolecular context of an IDR: interactions with a tethered folded domain can alter IDR ensembles, as well as their response to changes in the cell. The importance of IDR:folded domain interactions has already been pointed out in several recent studies[128–131]. Despite all this, our results show that

even where FP:IDR interactions exist, the structural biases shaping disordered protein ensembles *in vitro* are recapitulated in the cell.

3.4 Discussion

The study of disordered proteins requires shifting from the classical sequence-structure-function paradigm to one where the structural biases of the ensemble beget function[2]. While an extensive body of work has established the existence of structural biases in IDR ensembles *in vitro*, few studies have attempted to do so in the cell across many constructs in a self-consistent manner. Our results systematically show that structural biases are prevalent in IDR sequences, are encoded in amino acid sequence rather than composition, and exist even in the absence of local secondary structural biases (e.g., local helical preference, **Fig. 3.1A**).

The cell is often treated as a chemically monolithic environment, yet spatial and temporal regulation of volume, water content, pH, ions, and metabolites accompany key processes and pathology in cell biology[132–134]. Our in-cell study establishes that IDR structural biases observed *in vitro* also occur in live cells for almost all cases reported here. Furthermore, both in cells and *in vitro*, IDR structural biases can reshape in response to changes in the surrounding environment. This provides a mechanistic explanation for numerous cases where IDRs sense and actuate a response to such changes[28,33,135], since a change in structural bias in response to physical-chemical changes can alter IDR function. Importantly, sensing and actuating through this mechanism occurs at the speed of protein conformational changes (milliseconds or less[119]) and requires no additional energy (e.g., ATP).

The importance of IDR ensembles for molecular function has been shown or proposed for all of the naturally occurring IDRs characterized in this study. The structural preferences of the PUMA BH3 ensemble have been shown to affect its binding kinetics to MCL1 - a key event in the function of PUMA as a modulator of p53, and it has further been shown that this structural change can be induced by changing the composition of the solution[39]. Changing the structural preferences of the p53 N-terminal ensemble affects its binding affinity to MDM2, a potent inhibitor of p53's protective function, altering downstream p53 function[18]. FUS low-complexity region can undergo phase separation *in vitro* and *in vivo*. Recent work has shown that for monomeric low-complexity sequences, chain dimensions dominated by intramolecular interactions can quantitatively inform on

intermolecular interactions in the context of phase transitions[136]. The Ash1 ensemble has been shown to remain largely unperturbed by phosphorylation, indicating the need for robust activity of this key transcription factor in yeast[12]. Finally, a region proximal to the E1A sequence used here has been shown to be highly conserved in terms of the average end-to-end distance of its ensemble, and this length critical to its function, implicating strong selection for ensemble properties across the whole protein[32].

Given that IDR ensemble sensitivity can be encoded by amino acid sequence, we suggest that this sensitivity could also be subject to evolutionary selection. We propose that certain sequences have evolved to act as sensors and actuators of changes in the cellular environment. In the same vein, other disordered sequences have the ability to resist structural changes (as shown for the case of PUMA S2). Indeed, changes in ensemble structure provide a rapid, specific, and energetically efficient way for IDRs to sense and respond to changes in the cellular environment. This sensing capability of IDRs has been demonstrated not only for changes in solution conditions and osmotic pressure as studied here, but also for changes in other conditions such as membrane curvature[28], water availability[137], and temperature[138]. As our understanding of IDR sensing expands, we expect to uncover novel functions for this important class of proteins. In addition, learning to predict and control this sensitivity will allow for the design of IDR-based sensors targeting specific physicochemical intracellular conditions, as has already been demonstrated for the case of osmotic pressure sensing[33].

An additional implication of the evolved ability to sense and respond to changes in the environment is that a misregulated intracellular environment may disparately affect IDR function. Metabolic rewiring, a hallmark of cancer, viral infection, and other pathologies, can dramatically alter the physicochemical composition of the cell[139,140]. Even if this change would alter the activity of only a small subset of IDRs, their role as central signaling hubs could cause widespread cellular malfunction. In this way, IDR sequences can be drivers of pathology in a deleterious cellular environment, even in the absence of mutations. We propose that this phenomenon is a previously overlooked cause of IDR-driven proteopathies.

3.5 Methods and supplementary figures

Experimental methods and supplementary figures for in this study are in Appendix 2. Full *in vitro*, live cell and simulation data sets can be found online at https://github.com/sukeniklab/IDP_structural_bias.

3.6 Author contributions

S.S. conceptualized and led the project. D.M. designed and performed all *in vitro* experiments and analysis with the help of R.M. K.G. designed and performed all live cell experiments and analysis with the help of N.M.S. and G.K. E.F., assisted by E.C.Z., performed *in vitro* controls. A.R.P. synthesized, purified and characterized unlabeled PUMA peptides with assistance from A.D.M. E.W.M. assisted D.M. with SAXS analysis. A.S.H. and F.Y. designed, ran and analyzed simulations. S.S., D.M., K.G. and A.S.H. wrote and revised the manuscript.

4. Leveraging biophysical understanding to aid in the design of a biosensor based on an intrinsically disordered region

4.1 Abstract

Cell homeostasis is perturbed when dramatic shifts in the external environment cause the physical-chemical properties inside the cell to change. Experimental approaches for dynamically monitoring these intracellular effects are currently lacking. Here, we leverage the environmental sensitivity and structural plasticity of intrinsically disordered protein regions (IDRs) to develop a FRET biosensor capable of monitoring rapid intracellular changes caused by osmotic stress. The biosensor, named SED1, utilizes the Arabidopsis intrinsically disordered AtLEA4-5 protein expressed in plants under water deficit. Computational modeling and *in vitro* studies reveal that SED1 is highly sensitive to macromolecular crowding. SED1 exhibits large and near-linear osmolarity-dependent changes in FRET inside living bacteria, yeast, plant, and human cells, demonstrating the broad utility of this tool for studying water-associated stress. This study demonstrates the remarkable ability of IDRs to sense the cellular environment across the tree of life and provides a blueprint for their use as environmentally-responsive molecular tools.

4.2 Introduction

Intracellular osmotic fluctuations are one of the most common physicochemical perturbations cells experience throughout their life[141]. In the absence of external stressors, the metabolic activity of the cell can induce large changes in the concentration of different metabolites that alter intracellular osmolarity[142]. Additional osmotic variations can be caused by the activity of ion channels that change the total concentration of free inorganic ions (K^+, Na^+, Mg^{2+}, etc.)[143]. Severe intracellular osmotic perturbations are readily caused by environmentally-induced stress conditions, where the osmolarity outside of the cells changes dramatically. For instance, a decrease in water content in the exterior of a cell increases extracellular osmolarity in a way that causes the passive efflux of water out of the cell. This results in an immediate collapse of cell volume and concomitant increase in the concentration of solutes, macromolecular crowding, and the viscosity of the cell interior, impacting a plethora of molecular and cellular functions[55,144–146].

Despite the importance of osmotic regulation on cell function, our mechanistic understanding of how cells sense such conditions, particularly in multicellular organisms, is limited[147,148]. One of the main barriers to better understanding the intracellular effects of osmotic stress is the lack of methods to

reliably monitor physical-chemical changes that occur in single cells, in real time, and in a non-destructive manner[147,149,150].

Genetically encoded fluorescent biosensors are optical tools that enable the dynamic visualization and quantification of biochemical events that occur in living cells at various scales, from single cells to whole organisms[151]. Fluorescent biosensors are chimeric proteins composed of at least one fluorescent protein fused to a sensing domain. The selection of the sensing domain is based on its ability to specifically change its conformation in the presence or absence of an analyte[152]. The conformational change of the sensing domain then causes a change in the fluorescence readout that can be quantified. As of today, there are dozens of different fluorescent biosensors used to track small molecules, phosphorylation events, neurotransmitters, posttranslational modifications, and hormones; however, just a small fraction of biosensors are designed to report changes in the physical-chemical properties of the environment[153–157]. The main challenge for developing environmentally-responsive biosensors is in sourcing sensory domains capable of specifically and reversibly altering their structure in response to changes in a specific physical-chemical property.

Intrinsically disordered protein regions (IDRs) are protein domains that lack a stable tertiary structure and instead behave as ensembles of dynamic and rapidly changing conformations[64]. Because IDRs have a more extended surface area than globular proteins, they are highly sensitive to the physical-chemical properties of the solvent. Conditions such as pH, temperature, redox state, and high osmolarity induce conformational changes in some IDRs[124]. Recent work shows that environmental sensitivity is a shared property of many IDRs[19,20]. Furthermore, it has been proposed that the environmental sensitivity of IDRs could be used to regulate their activity, potentially allowing them to function as sensors of the environment[135,150,158]. Based on the aforementioned properties, we propose that IDRs are promising candidates for use in the development of environmentally-responsive biosensors.

Here, we report the design, development and implementation of a Förster Resonance Energy Transfer (FRET) biosensor that tracks the effects of osmotic stress on living cells from a wide variety of organisms. For the sensory domain we tested a group of hyperosmotic-induced IDRs from plants. Specifically, we used group 4 LATE EMBRYOGENESIS ABUNDANT (LEA) proteins from the model plant Arabidopsis thaliana[159], which we previously reported are intrinsically disordered proteins that exhibit conformational rearrangements *in vitro* upon

changes in the osmolarity of the solution[137]. Initial screening of prototype biosensors in budding yeast revealed that these proteins, particularly AtLEA4-5, undergo a rapid and reversible conformational change in response to hyperosmotic treatments with different osmolytes in cells. All-atom simulations and *in vitro* experiments revealed that the physical dimensions of the AtLEA4-5 conformational ensemble can change dramatically when the composition of the surrounding solution is altered, supporting our findings in living cells. The resulting FRET biosensor, named SENSOR EXPRESSING DISORDERED PROTEIN 1 (SED1), can dynamically monitor the response of budding yeast to osmotic stress at the cellular level. SED1 can also be used to track the effects of osmotic stress on live bacteria (*Escherichia coli*) cells, plant (*Nicotiana benthamiana*) cells, and U-2 OS human cells. The use of fluorescent biosensors such as SED1 will aid in understanding how cells sense, respond, and acclimate to dynamic environmental fluctuations caused by water-associated stress.

4.3 Results

4.3.1 Design of a biosensor for studying the effects of osmotic stress on living cells

To track the effects of osmotic stress on living cells, we sought to combine the power of osmo-sensitive IDRs and ratiometric FRET readouts to build a genetically encoded fluorescent biosensor. For the sensory domain, we tested two members of the group 4 LEA proteins from the model plant Arabidopsis thaliana[159]. Group 4 LEA proteins are intrinsically disordered proteins that exhibit a reversible disorder-to-folded transition in response to increased osmolarity *in vitro*[137]. We hypothesized that such osmolarity-dependent conformational changes would also occur inside living cells, making them excellent candidates for environmentally-responsive biosensor development.

To test the ability of group 4 LEA protein structure to change in response to osmotic stress *in vivo*, we fused either AtLEA4-2 or AtLEA4-5 ORFs between the coding sequences of a FRET-compatible pair of fluorophores (mCerulean3 as the donor and mCitrine as the acceptor) (**Fig. 4.1A**). These constructs were expressed in live budding yeast (Saccharomyces cerevisiae) cells and treated with NaCl to induce hyperosmotic shock. Both constructs exhibited a NaCl-concentration-dependent increase in the acceptor-to-donor emission ratio[33]. We observed that the treatment displayed typical FRET behavior with an increase in fluorescence intensity of the acceptor (donor excitation-acceptor emission; DxAm) coupled to a

decrease in fluorescence intensity of the donor (donor excitation-donor emission; DxDm), leading to a higher acceptor to donor emission ratio (DxAm/DxDm)[33]. The FRET ratio change was significantly smaller when we tested a globular protein (arabinose-binding protein, ABP) as a reference[160]. Hyperosmotic treatment with increasing concentrations of other ionic and non-ionic osmolytes showed that the change in FRET of both constructs is osmolarity-dependent and not osmolyte-specific[33]. Since AtLEA4-5 exhibited the largest FRET change in response to osmotic shock, we continued our characterization with this construct. The fluorescence intensity of single mCerulean3 or mCitrine fused to AtLEA4-5 was not significantly affected by hyperosmotic shock induced with different solutes, demonstrating the stability of the fluorophores in such conditions[33]. Finally, testing ten different FRET pairs revealed that the mCerulean3-mCitrine pair had the highest dynamic range among monomeric fluorescent proteins[33].

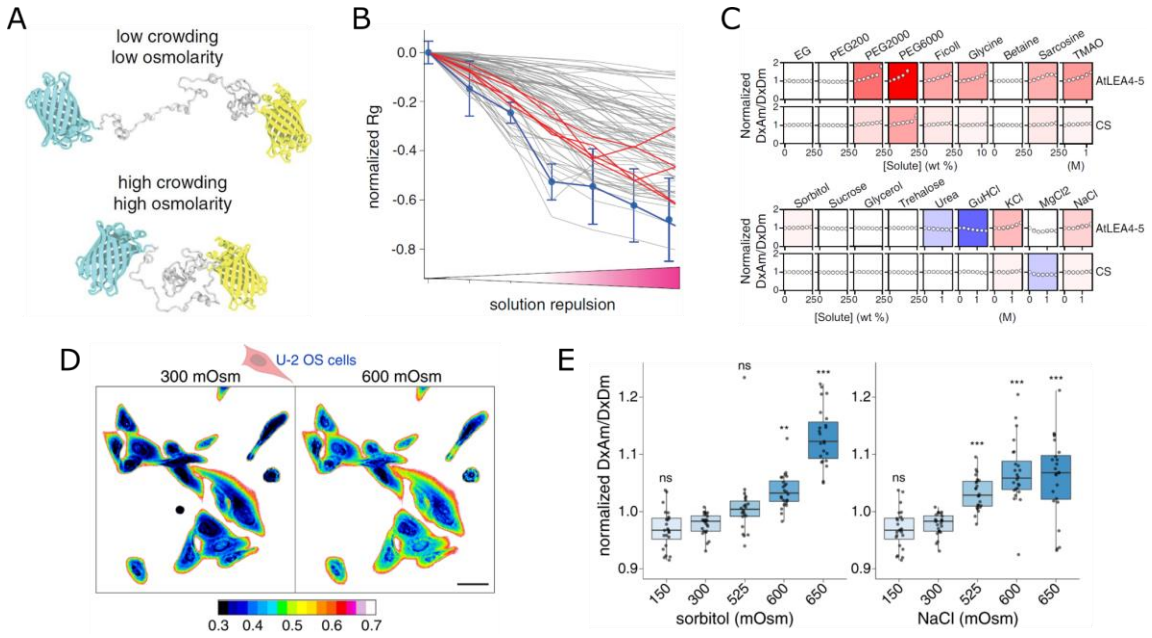

Figure 4.1. Contribution of our system to the design of a biosensor. (**A**) Schematic representation of the biosensor design under low and high macromolecular crowding/osmolarity—prevalent intracellular conditions upon hypoosmotic or hyperosmotic stress, respectively. The conformations are selected from the ensemble of all-atom simulations of AtLEA4-5 in the corresponding conditions. Cyan: mCerulean3. Yellow: Citrine. Gray: AtLEA4-5. (**B**) Computational solution space scan of the normalized radius of gyration (Rg) of AtLEA4-5 (blue), five different scrambled sequences (red), and 70 different naturally occurring IDRs (gray) under different solution repulsion levels (low to high solution repulsion of the protein backbone). Mean ± SD from n = 5 independent simulations. (**C**) Experimental solution space scan of AtLEA4-5 and CS. Open circles show the normalized FRET ratio (DxAm/DxDm) for the indicated concentration of each solute, with two points (that often overlap) for each concentration taken from separate repeats, highlighting the

reproducibility of the data. Background color intensity represents sensitivity to the addition of solute. Stronger colors indicate stronger sensitivity. Red: compaction; blue: expansion; white: no change. Solution concentrations are given in weight percent (0–25 or 0–12 wt%) or molar (0–1.5 M). (**D**) Ratiometric image of live SED1-expressing U-2 OS cells at 300 mOsm (isosmotic) or 600 mOsm (hyperosmotic) treated with sorbitol. Scale bar = 50 μm. Calibration bar represents the normalized FRET ratio (DxAm/DxDm). (**E**) Normalized FRET ratio of SED1-expressing U-2 OS cells exposed to different osmotic treatments with sorbitol and NaCl. One-way ANOVA. *p < 0.05, **p < 0.01, ***p < 0.001.

4.3.2 AtLEA4-5 is highly sensitive to the chemical composition of the solution

As water leaves the cell during hyperosmotic shock, a number of physical-chemical properties of the intracellular environment change; in particular, the concentration of organic and inorganic solutes rises, as does the extent of macromolecular crowding. Any of these properties could underlie the biophysical mechanism driving the conformational changes in AtLEA4-5. Macromolecular crowding is a general condition of the cell interior that gets exacerbated under hyperosmotic conditions due to water loss[161–163]. In order to further investigate the mechanism of AtLEA4-5 responsiveness observed in cells, we designed an approach to test AtLEA4-5 sensitivity to different solutions *in silico* and *in vitro*.

First, we performed all-atom Monte Carlo simulations to sample the conformational landscape of AtLEA4-5 under a wide range of solution conditions. This class of simulation, known as solution space (SolSpace) scanning, has been used to investigate the solution-protein interactions of dozens of IDRs[19,20]. We used this method to exert a compacting force on a range of IDRs and compared the tendency of the different sequences to compact. We observed that AtLEA4-5 shows an enhanced sensitivity to such compaction compared to the scrambled versions of the sequence, in agreement with our *in vivo* observations[33]. Furthermore, a comparison with 70 different naturally occurring IDRs shows that AtLEA4-5 is an outlier in terms of its sensitivity (**Figure 4.1B**).

We further investigated the solution sensitivity of AtLEA4-5 *in vitro*. We used the FRET efficiency of purified full-length AtLEA4-5 fused to mCerulean3 and mCitrine as a proxy for the end-to-end distance of the construct under different solution conditions. We induced macromolecular crowding with solutions of different molecular weight polyethylene glycol (PEG) isoforms at various concentrations, and compared these results to a previously reported macromolecular crowding biosensor (CS) as a reference[157]. The CS sensory domain is a synthetic, helical peptide with a hinge-like topology thought to compact

in response to higher macromolecular crowding. Our experiments showed that PEG induced the compaction of AtLEA4-5 in a concentration and size-dependent manner (**Fig. 4.1C**). The PEG-induced compaction was more prominent in AtLEA4-5 than in CS, confirming the relative sensitivity of AtLEA4-5 to macromolecular crowding. Together, these data show that despite its intrinsic disorder, the conformational ensemble of AtLEA4-5 is highly responsive to changes in the chemical composition of the solution, particularly macromolecular crowding, *in silico* and *in vitro*, and that these properties are based on both topology and amino acid sequence.

4.3.3 SED1 tracks changes in osmolarity in a wide set of organisms

Given the ability of SED1 to report the effects of osmotic stress on budding yeast [not excerpted here; please see the original paper], we sought to apply it to other biological systems. We first expressed SED1 in the bacteria *Escherichia coli*. Similar to what we found in yeast, we observed a hyperosmotic stress-dependent increase in the FRET readout[33]. Next, we tested SED1 in two evolutionarily distant multicellular organisms: plants and humans.

Plants heavily rely upon water to provide structural support and to facilitate gas exchange with the environment[150]. To test the utility of SED1 in this context, we transiently expressed a nuclear-localized SED1 transgene in tobacco leaves. Small discs of leaf tissue were placed onto 96-well plates, in wells containing hyperosmotic (sorbitol or NaCl) or hypoosmotic (water) solutions. We found that when SED1-expressing leaf discs were incubated with sorbitol or NaCl, the FRET readout increased over time, with an increase in fluorescence intensity of the acceptor and a concomitant decrease in fluorescence of the donor[33].

We further tested SED1 in human cells. To do so, we stably introduced SED1 into human epithelial (U-2 OS) cells and measured the SED1 FRET signal in response to sorbitol and NaCl treatments at different osmolarities using live-cell confocal microscopy. We observed that both treatments induced an increased FRET ratio immediately after the addition of the solution (**Fig. 4.1D,E**). The increased fluorescence of the acceptor and decreased fluorescence of the donor after the treatments, along with the acceptor photobleaching control, confirmed the expected FRET behavior. This data demonstrates that SED1 is responsive in live human cells.

In conclusion, we showed that SED1 is a versatile, genetically-encoded optical tool that can be used to dynamically track the response to osmotic stress of living cells from a plethora of organisms in an inherently quantitative manner. This opens new avenues to investigate the poorly-understood impact of environmental perturbations on the regulation of cellular function.

4.4 Discussion

The use of SED1 to monitor osmotic variations in living cells has the potential to reveal new fundamental aspects of cell biology. SED1 might be used to 1) dynamically track the macromolecular crowding of individual cells during perception, response, and acclimation to osmotic stress; 2) screen for mutants disrupted in the sensing and response mechanisms to osmotic shock; 3) test whether other kinds of stressors induce intracellular osmotic variation; 4) generate osmolarity and/or macromolecular crowding maps of different cell types of multicellular organisms. This has the potential to revolutionize our understanding of the biological processes that enable desiccation survival, extreme salt tolerance, and rehydration. The ability of SED1 to work in evolutionarily distant organisms means that these processes can be studied across the tree of life to broaden our understanding of the ways in which water regulates life on earth.

4.5 Relation to other chapters in this dissertation

Chapter 2 of this dissertation describes how I, my colleague Feng Yu, and other colleagues developed a combined *in vitro* and *in silico* “solution-space scanning” system for investigating the physicochemical sensing capabilities of IDRs. Chapter 3 describes how I, my colleague Karina Guadalupe, and other colleagues built on this system to carry out parallel FRET experiments *in vitro* and in live cells. This chapter describes a practical application of this system, focusing on the contribution that Feng, Karina and I were able to make to the design and testing of a novel biosensor based on the sensing capabilities of a plant IDR. Please see the original paper for further discussion of the design, testing and implications of this biosensor[33].

4.6 Methods

The methods for the experiments excerpted here are described in Appendix C. For other methods used in this study, please see the original paper[33].

4.7 Author contributions

To show a practical application of the systems developed in Chapters 2 and 3, this excerpt concentrates on data obtained *in vitro* by David Moses, *in silico* by Feng Yu, and in U-2 OS human cells by Karina Guadalupe. **Fig. 4.1B** shows data obtained by F.Y., **Fig 4.1C** shows data obtained by D.M. and **Fig. 4.1D** and **E** show data obtained by K.G. The excerpted paragraphs were written and edited by Cesar L. Cuevas-Velazquez, Shahar Sukenik and José R. Dinneny with assistance from the other co-authors. The method descriptions in Appendix C were written by D.M., K.G. and F.Y. and edited by C.L.C.V. and S.S.

5. Conclusion

5.1 Summary of this work

During the past five years, my collaborators and I have worked to establish a new paradigm of IDRs as physicochemical sensors in the cell. To qualify as a sensor of a given stimulus, an IDR must respond differently than other IDRs to that stimulus, and also must respond differently to that stimulus than it does to other stimuli[164]. With that in mind, in the study described in Chapter 2, building on the concept of "solution-space scanning" proposed earlier the same year by Holehouse and Sukenik[20], my colleague Feng Yu and I, along with other colleagues used ensemble FRET experiments *in vitro* along with computer simulations to systematically and rigorously quantify the sensitivity of IDRs to a variety of changes in solution conditions[19]. We found enough variation among IDRs in responses to varying concentrations of salts, osmolytes, crowders, etc. (**Fig. 2.2**), to validate our hypothesis and cause us to push forward in the development of this paradigm.

The question of whether conformational biases of IDRs seen *in vitro* accurately represents the conformational biases of the same IDRs in the cellular environment has been an important gap in the disordered protein subfield. Believing that we could contribute to filling this gap in knowledge, Karina Guadalupe and I, along with other colleagues, embarked on the study described in Chapter 3[34]. We built a solid foundation by characterizing each IDR via FRET, SAXS, and analytical SEC *in vitro*, and complemented this experimental foundation with strategic simulations of our benchmark GS-repeat IDRs that provided a bridge between the FRET and SAXS results (**Fig. 3.2B,F**). Seeing excellent agreement between FRET and SAXS results, and very good agreement between those results and the results of analytical SEC and simulations, we felt confident moving our constructs into cells and running FRET experiments as similar as possible to what we had done *in vitro*. As shown in Chapter 3, we saw excellent agreement between our *in vitro* and live-cell results in almost all cases (**Figs. 3.2B,G, 3.3D,G, 3.4A,D**). This agreement held true even for IDRs such as the PUMA scrambles where disagreement would not have been at all surprising since we had abolished the local helical structure of the wild type, leaving only interactions not identifiable as secondary structure to cause any differences in conformational biases.

Chapter 4 described an immediate practical application of the parallel *in vitro*, *in silico* and live-cell IDR characterization and solution-space scanning systems my colleagues and I had developed, as our collaborator Cesar Cuevas-

Velazquez combined his previous research on LEA proteins with our systems to design a biosensor sensitive to osmotic stress[33] (**Fig. 4.1**). This biosensor fills an important gap in the field, enabling real-time monitoring of osmotic stress in live cells. Taken together, these studies show that IDRs sense their environment *in vitro*, that IDR conformational biases seen *in vitro* are likely to remain prevalent when the same IDR is moved into live cells, and that this knowledge could be harnessed to design a biosensor that can monitor a key process in real time in live cells.

5.2 Implications and future direction

The implications of IDRs serving as physicochemical sensors are far-reaching. In the context of healthy cells, sensing by IDRs offers cells a fast and energy-efficient way to respond to changes in intracellular conditions. For example, the metabolic profile of cells routinely changes over the cell cycle[165]. If IDRs can sense these changes, how do they respond, and how do these responses affect the health of the cell? In the context of disease, the sensitivity of IDRs means that deleterious changes to the cellular environment could produce changes in IDR ensembles, and therefore function, that could potentially drive pathology even in the absence of mutations. For example, the prevailing idea in cancer research is that genetic mutations change the function of proteins, causing healthy cells to become cancerous[166]. But given the sensing capabilities of wild-type p53 and PUMA that we have seen in Chapters 2 and 3, could their cancer-preventative functions be compromised, without mutation, by changes in their surrounding conditions? Also, the metabolic profiles of cells in various types of tumors are extensively altered compared to the metabolic profiles of healthy cells[167]. How might IDR ensembles be perturbed by these changes, and how might these responses contribute to the progress of cancer?

Just as exciting as these questions is the possibility of leveraging the versatile sensing capabilities of IDRs to design biological and medical tools. Could biosensors based on IDRs be designed that could sense differences between healthy and diseased cells? Could our understanding of IDR structural biases allow custom design of IDRs that would form biomolecular condensates capable of catalyzing desirable reactions or sequestering undesirable entities in the cell? Would it be possible to design IDR-based drug-delivery vehicles that could discern the correct destination at which to release the drug? Innovative labs have started down the road toward tackling such projects by developing algorithms for designing IDRs with specific structural or molecular properties[168,169] and showing

how changes to IDR sequence can modulate the material properties of biological condensates[170].

To gain a better understanding of the roles of IDRs in health and disease and to move forward with further inventions, one immediate need is to build our understanding of the relationship between IDR ensemble and function. Many proposed ensemble-function relationships that need further investigation are listed in my review article excerpted in Chapter 1[164]. I anticipate that better understanding of the sensory capabilities of IDRs, as well as the relationship between IDR ensemble and function, will lead to major innovations in disease therapies and biomolecular tools.

Appendices

Appendix 1

A1.1 Experimental Methods

A1.1.1 FRET construct design and cloning

FRET backbone (called fIDR_pET-28a(+)-TEV, **Fig. A1.9**) 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 used in this study are shown in **Table A1.1**.

A1.1.2 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.[171] 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.

A1.1.3 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.

A1.1.4 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[171,172].

A1.1.5 Calculation of FRET efficiency

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. S10. 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[75,172].

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 **Figs. 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.

A1.1.6 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) ν is a unitless scaling exponent[117,173]. In the limit of finite-sized polymers, ν 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 a 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 **Section A1.2**). 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[174]. 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. A1.11**).

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. A1.12**). Unlike the intra-chain distances (**Fig. A1.11**), 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.

A1.1.7 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:

$$\chi = \frac{R_e^i}{R_e^{GS}} - 1 = \frac{R_0^i\left(1/E_f^i - 1\right)^{\frac{1}{6}}}{R_0^{GS}\left(1/E_f - 1\right)^{\frac{1}{6}}} - 1 = \frac{n^i\left(1/E_f - 1\right)^{\frac{1}{6}}}{n^{GS}\left(1/E_f^{GS} - 1\right)^{\frac{1}{6}}} - 1$$

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)[175] 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 **Figs. 2.1** and **2.2**.

A1.1.8 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[176]. Specifically, this states that R_g = 0.0215 $M^{0.583}$ 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 **Figs. A1.11** and **A1.12**). 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 (Figs. A1.17, A1.18). For GS-linker constructs, we observe a systematic drop in χ as a function of PEG concentration (Fig. A1.17). 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. A1.18**). 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.

A1.2 Computational Methods

A1.2.1 All-atom simulations

All-atom Monte Carlo-based simulations were performed using the CAMPARI simulation suite, with the ABSINTH implicit solvent model[83]. 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[177], 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[20]. 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 $\psi \approx +1.2\%$.[20]

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[84]. 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[178] was used to calculate the radius of gyration and end-to-end distance of the ensemble.

A1.2.2 Coarse-grained simulations

Our coarse-grained depiction of heteropolymer IDRs uses the PIMMS simulation framework[10]. 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. A1.5A**, 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. A1.5B**. 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 $\min\{1, \exp(-\Delta \mathrm{E}/k_BT)\}$ where k_B = 1, $\Delta \mathrm{E}$ is the energy difference between the current and proposed configurations, and T has the same units as the contact energies thus making the ratio $\Delta \mathrm{E}/k_BT$ 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. A1.5B**.

A1.2.3 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
- ε is the apparent net inter-monomer interaction energy (measured in $k_B T$)
- 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 (**Figs. A1.13** and **A1.14**). 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 (ε), 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. A1.15**. 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[22]. 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.

A1.2.4 Converting from χ to ν^{app}

As in Section 2.3.1, 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[89,173]. 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 ν^{app} = 0.50.[30]

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[179–183]. 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[117,180,184–188]. Taken together, the application of scaling theory to finite-sized polymers can be misleading unless bona fide scaling behaviour 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[173,188,189]. 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[30,31].

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 (**Figs. A1.11** and **A1.12**). Using χ allows us to bypass the clear limitations that making homopolymer-based assumptions would necessitate.

Supplementary Figures

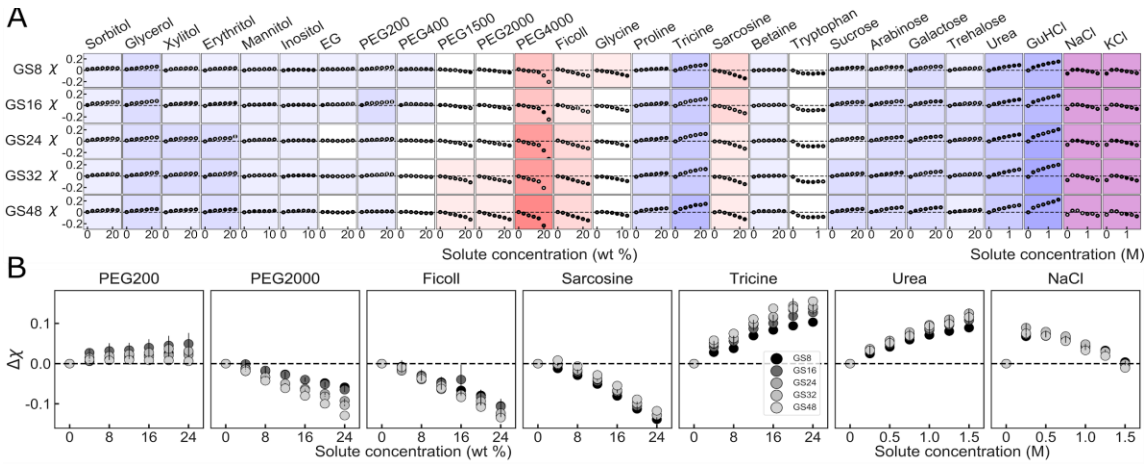

Figure A1.1 (**A**) Solution space scans of Gly-Ser linkers. Each data point shows the average χ vs. concentration of a specific solute for a specific IDR taken from two repeats. Vertical lines show the spread of repeats, and are often too small to see. IDRs vary down columns, and solutes vary along 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.

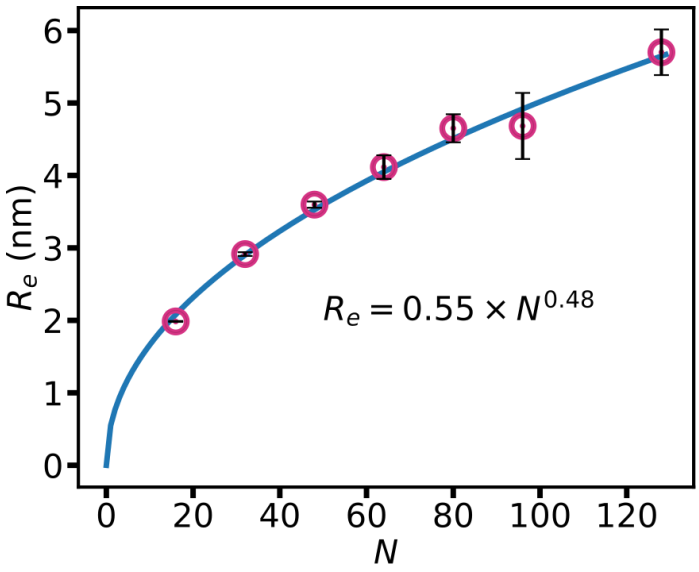

Figure A1.2. The end-to-end distance of Gly-Ser repeat sequences as a function of their total number of residues N, obtained from all-atom simulations in aqueous solution. Each data point is an average of five individual repeats, with lines being the standard deviation of the data. The blue curve is 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).

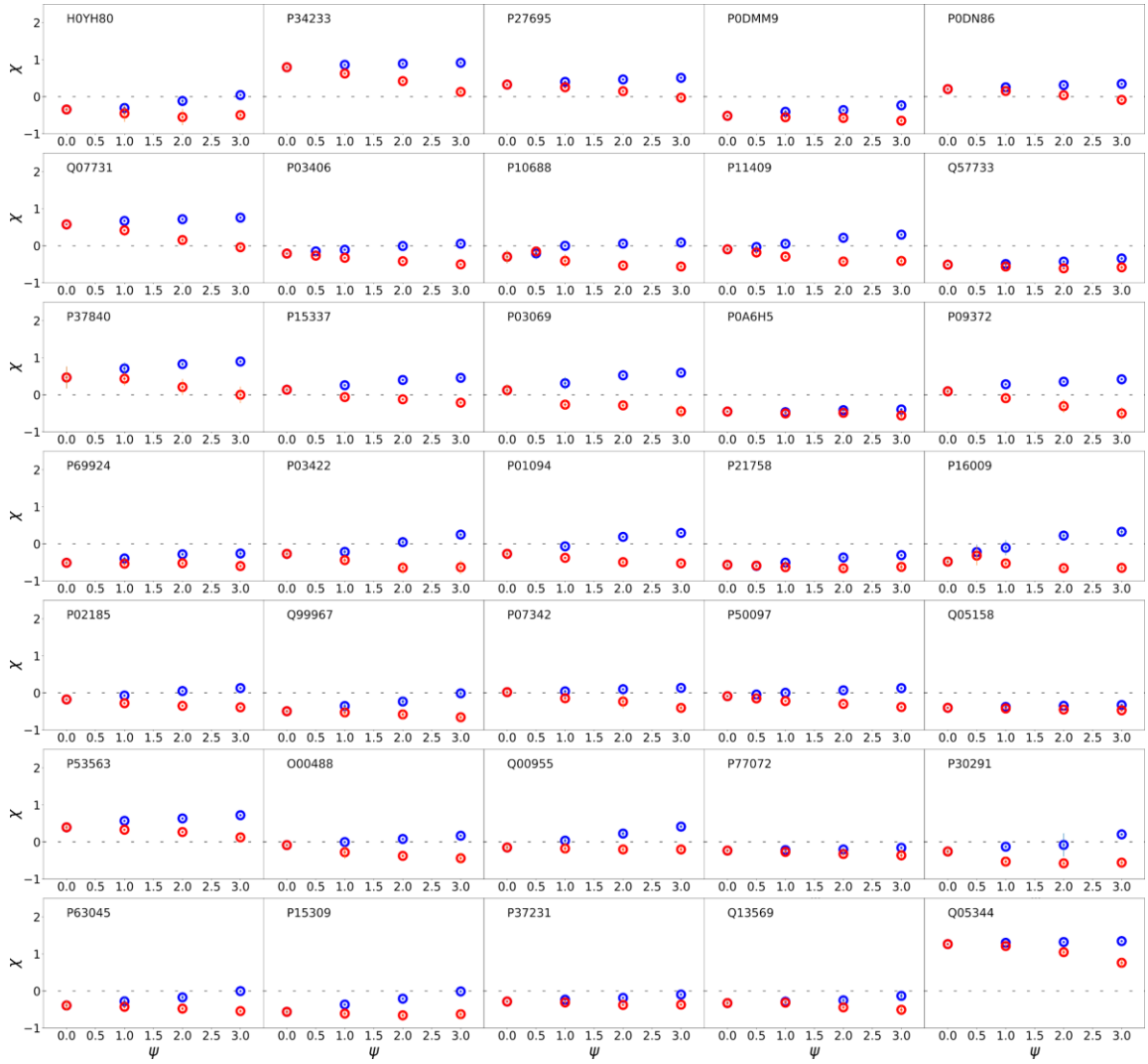

Figure A1.3. χ vs strength of solution interactions ψ (see **Section A1.2.1**) for each of the 70 IDRs shown in **Fig. 2.3**. Each subplot represents a single IDR. Blue points are attractive solutions (ψ > 0) and red points are repulsive solutions (ψ < 0). IDs are UniProt ID when available.

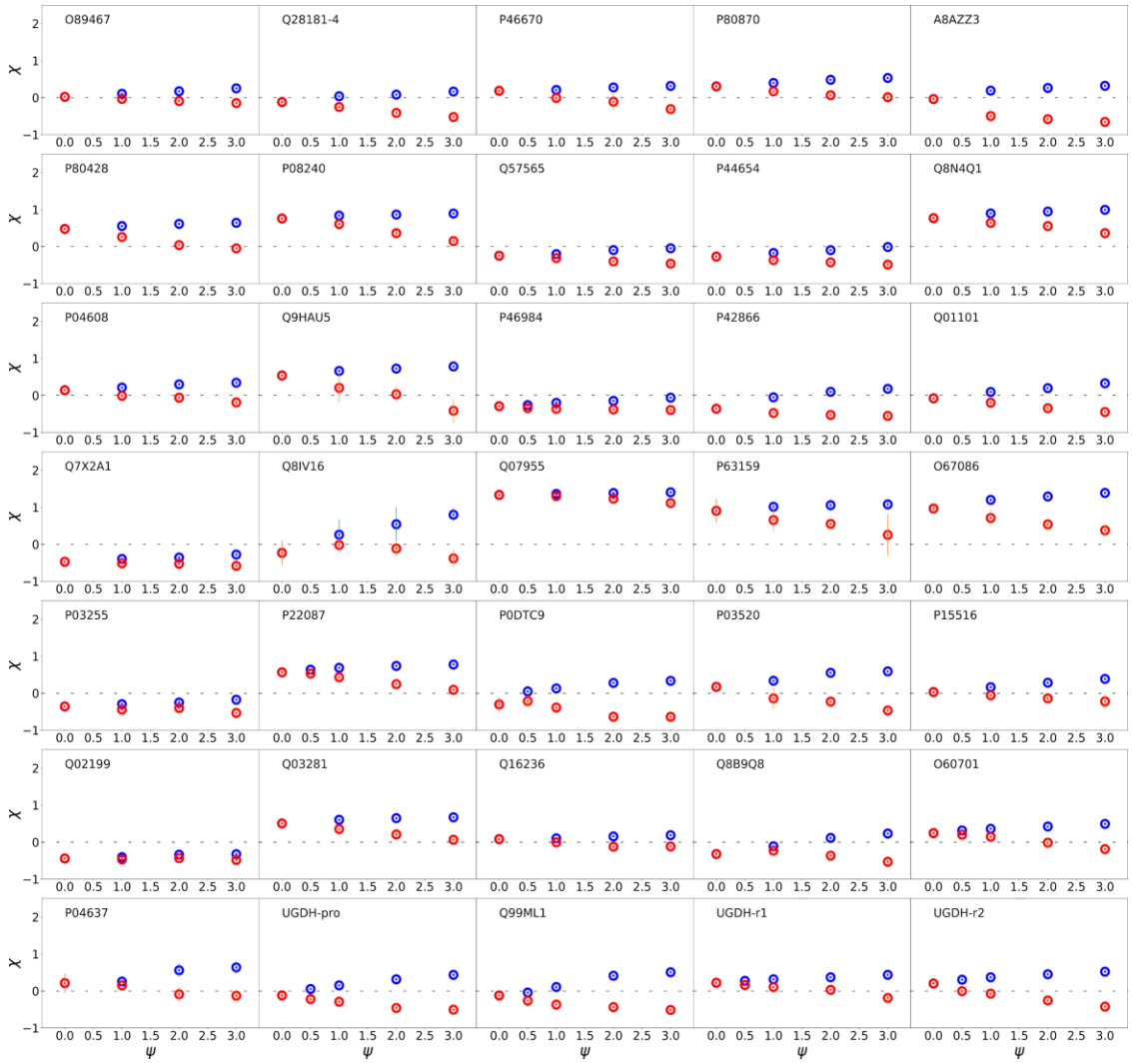

Figure A1.3 (cont.). χ vs. strength of solution interactions ψ (see **Section A1.2.1**) for each of the 70 IDRs shown in Fig. 3. Each subplot represents a single IDR. Blue points are attractive solutions ($\psi > 0$) and red points are repulsive solutions ($\psi < 0$). IDs are UniProt ID when available.

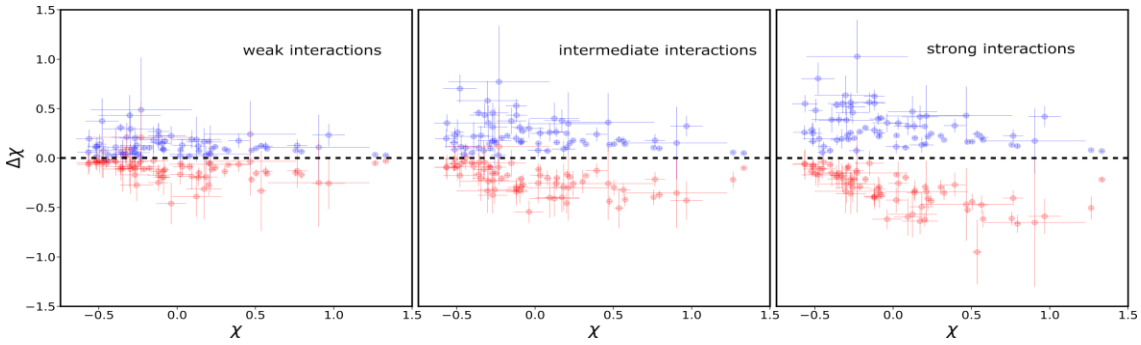

Figure A1.4. Solution sensitivity of IDRs shown in **Fig. A1.3**. Each point represents the solution-induced change in χ ($\Delta\chi$), for $\psi = \pm 1$ (weak interactions), ±2 (intermediate interactions) or ±3 (strong interactions). Blue points represent the response to repulsive solutions and red points represent the response to attractive solutions. Error bars are calculated from five independent simulations. See also **Fig. A1.6**.

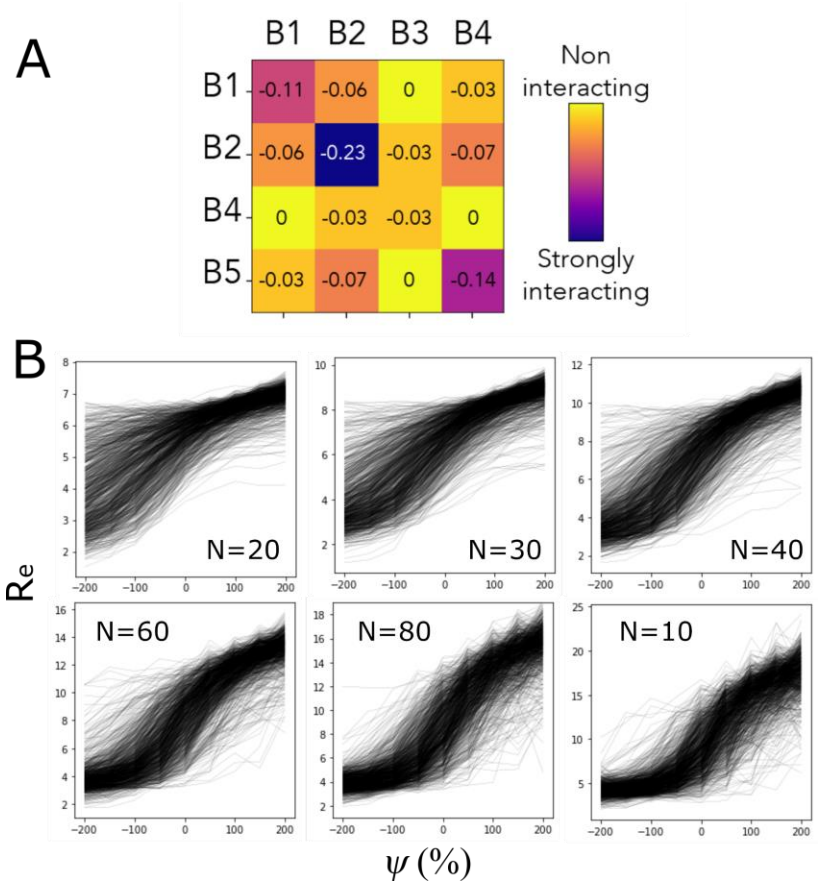

Figure A1.5. **(A)** Summary of the PIMMS parameters that were used for heteropolymer simulations. Interaction energies are defined in units of k_BT and were selected to approximate the chemical diversity observed in polypeptides. B1-B4 are "bead" 1 to "bead" 4. **(B)** End-to-end distances (in grid units) for PIMMS coarse-grained simulations of various sequences and chain lengths N. These curves were used to produce **Fig. 2.4B**.

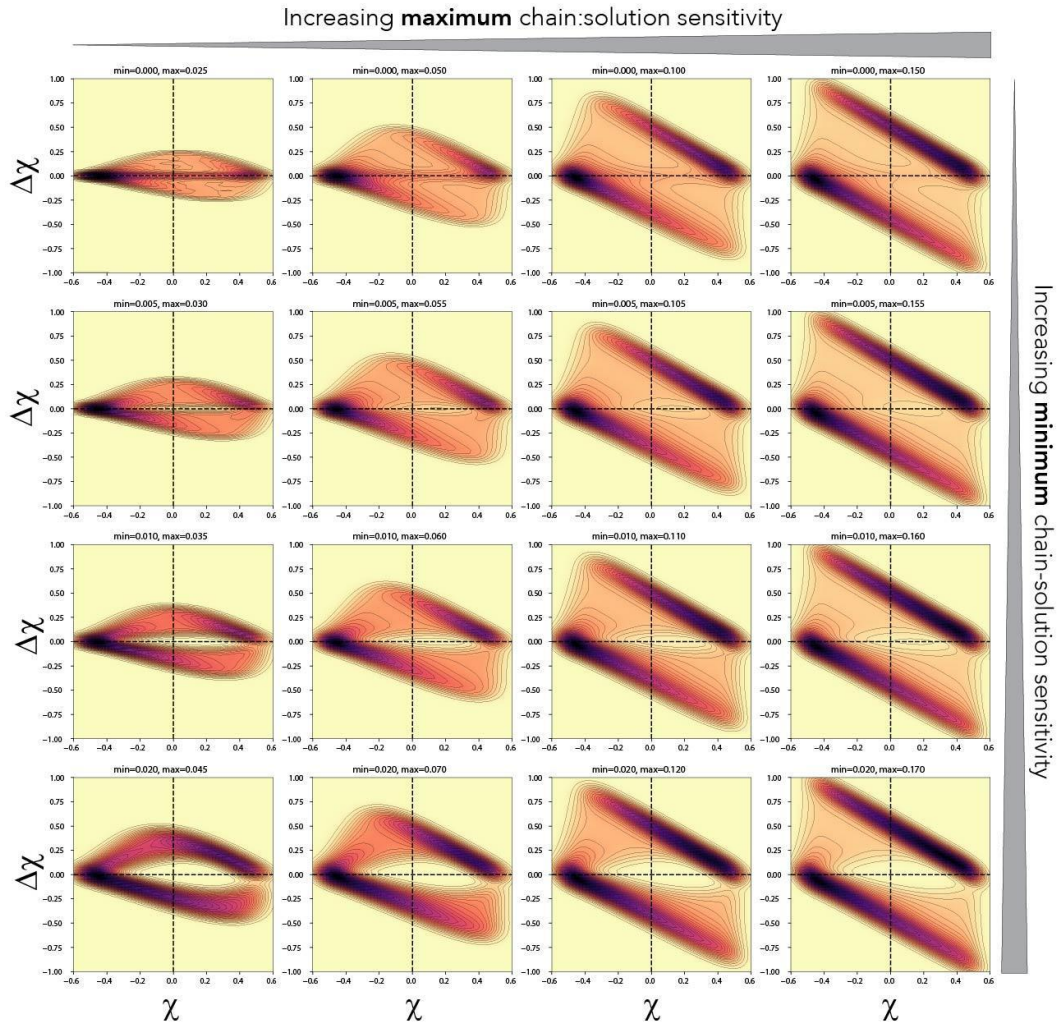

Figure A1.6. Dependence of $\Delta\chi$ vs. χ as a function of the most and least sensitive chains in an ensemble of sequences. Each figure defines the maximum and minimum perturbation to the chain-solvent interaction. As the maximum perturbation grows (left to right), $\Delta\chi$ becomes larger in 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 $\Delta\chi$ 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 $\chi = 0$. The maximum perturbation defines the magnitude of $\Delta\chi$, but is bounded by the chain dimensions, such that $\Delta\chi$ has upper and lower bounds. As the maximum is increased, more perturbations push up against that maximum, such that increasing $\Delta\chi$ density is observed at the bounds (i.e., see top right).

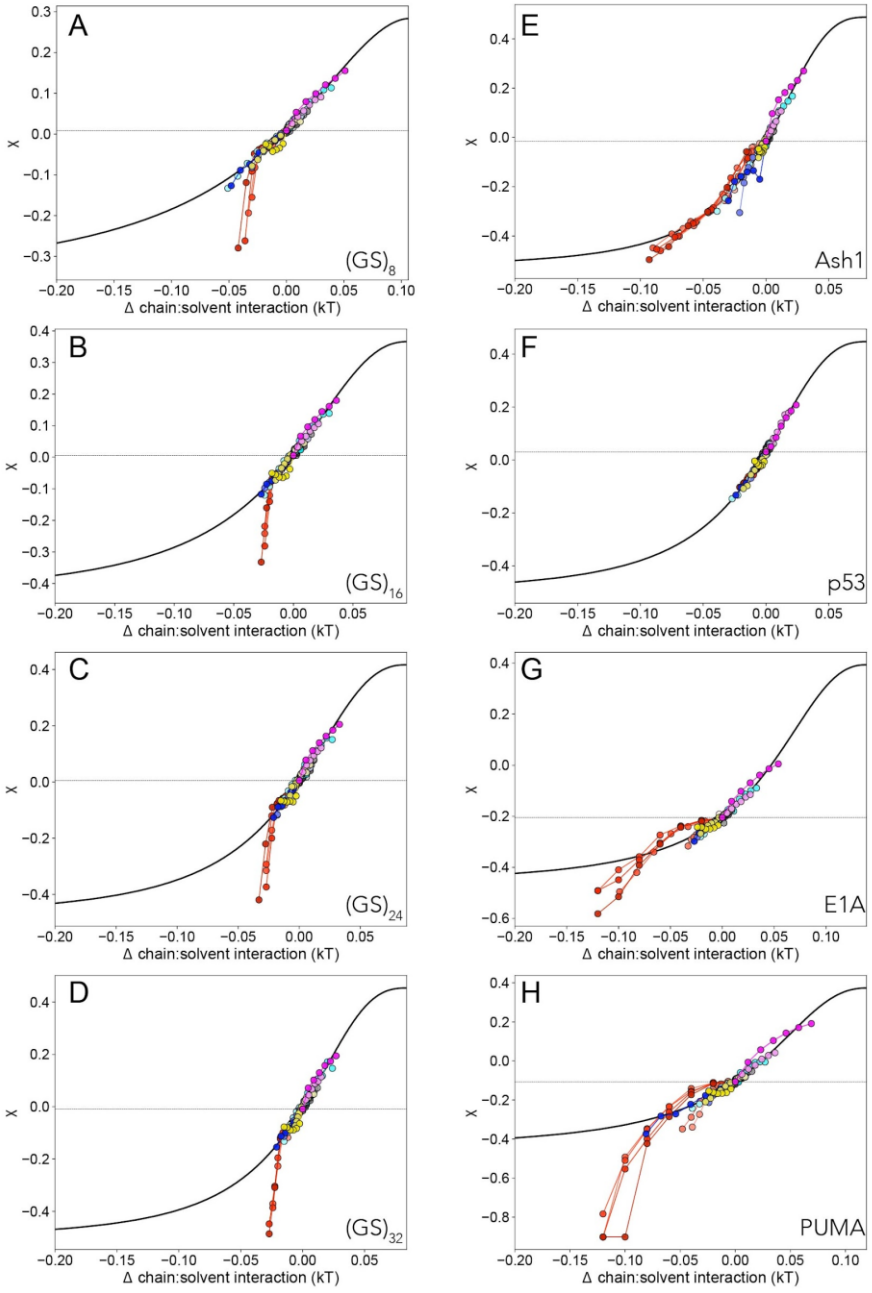

Figure A1.7. Full fit curves for all eight IDRs. Horizontal dashed lines reflect the χ value as measured in buffer. Black curve is a length-derived prediction from our analytic model. Note that for many of the curves the high-molecular weight PEG solutions lead to substantial deviations from the master curve, as expected as chain behavior enters the semidilute regime[79], 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.

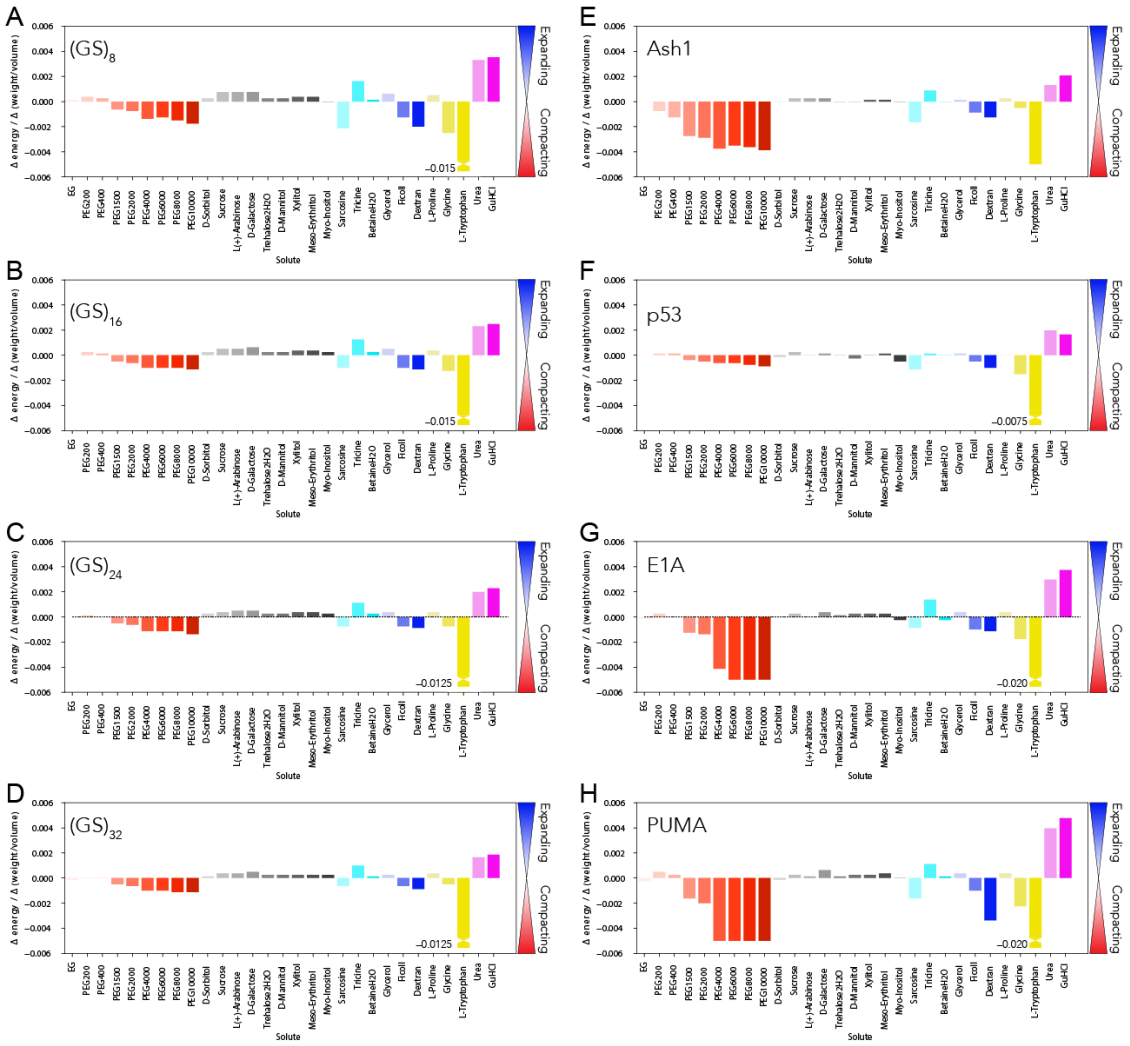

Figure A1.8. Derived solute-specific scalar factors that relate change in chain-solute interaction strength to a change in χ. More positive values lead to chain expansion while more negative values lead to chain compaction.

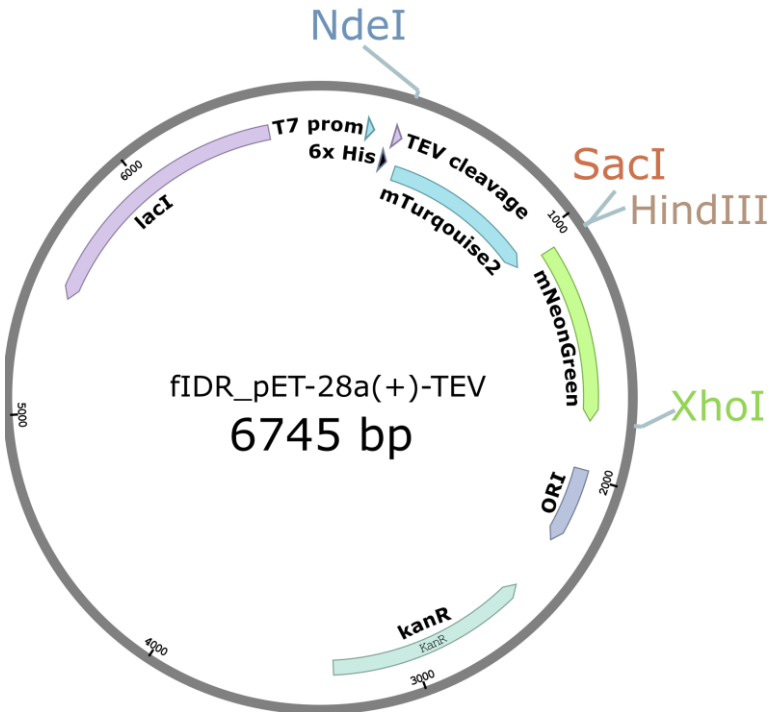

Figure A1.9. Plasmid map for FRET construct bacterial expression vector. Disordered sequences are inserted between 5' SacI and 3' HindIII restriction sites.

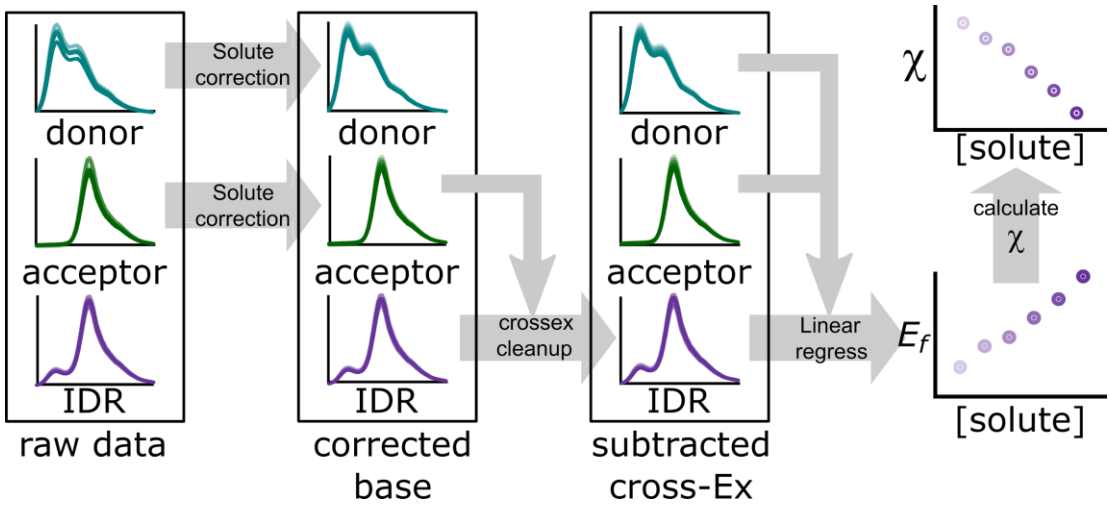

Figure A1.10. Visual summary of the data processing procedure detailed in **Section A1.1.5**. All panels show intensity vs. wavelength data for solutions containing donor-only, acceptor-only, and IDR construct (unless specified otherwise). Spectra are arranged from light to dark going from buffer to 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 **Section A1.1.5**, and E_f is used to calculate χ as described in **Section A1.1.7**.

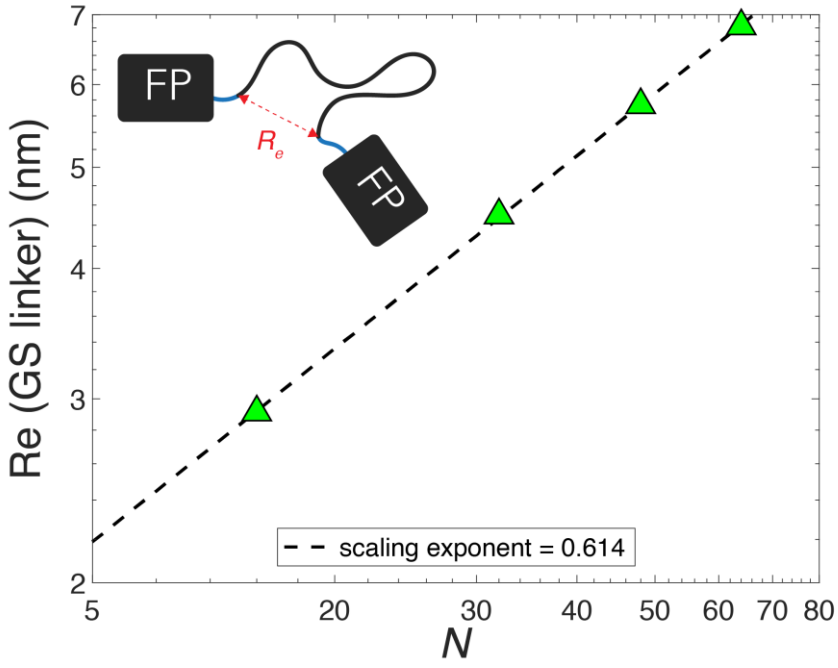

Figure A1.11. The intra-chain distance of glycine-serine (GS) linkers connecting two fluorescent proteins in a system that rigorously behaves as a self-avoiding random coil. GS linker end-to-end distance is measured between the first and last residue in the GS repeat region. Note that short (3-7 residue) 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.

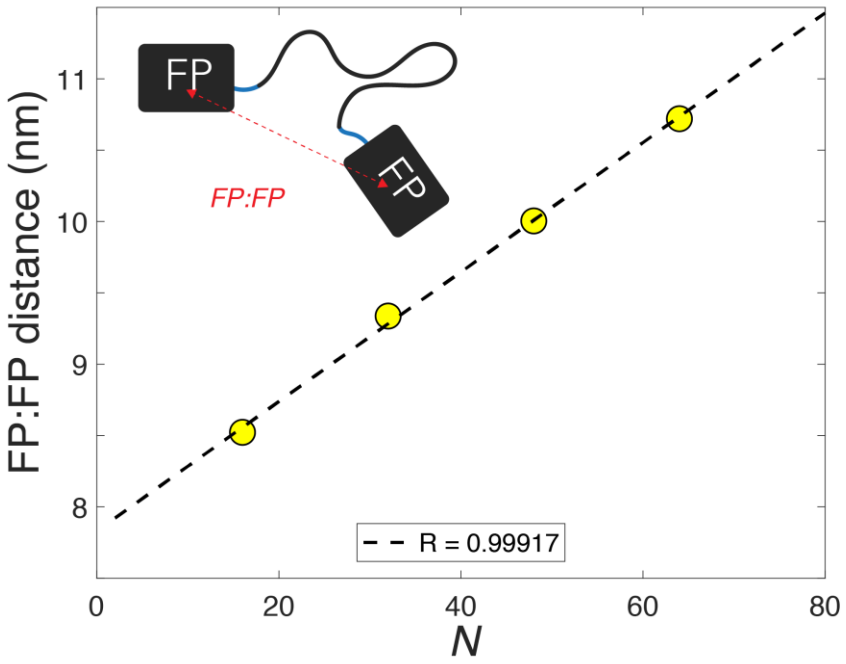

Figure A1.12. Intra-fluorescent protein distance for the same system as in **Fig. A.1.11**. The distance here is measured between the two chromophore centers in each of the two fluorescent proteins. Note that when intra-fluorescent protein distances are measured, we obtain a linear relationship (as opposed to a power law relationship as in **Figs. A1.11** and **A1.2**).

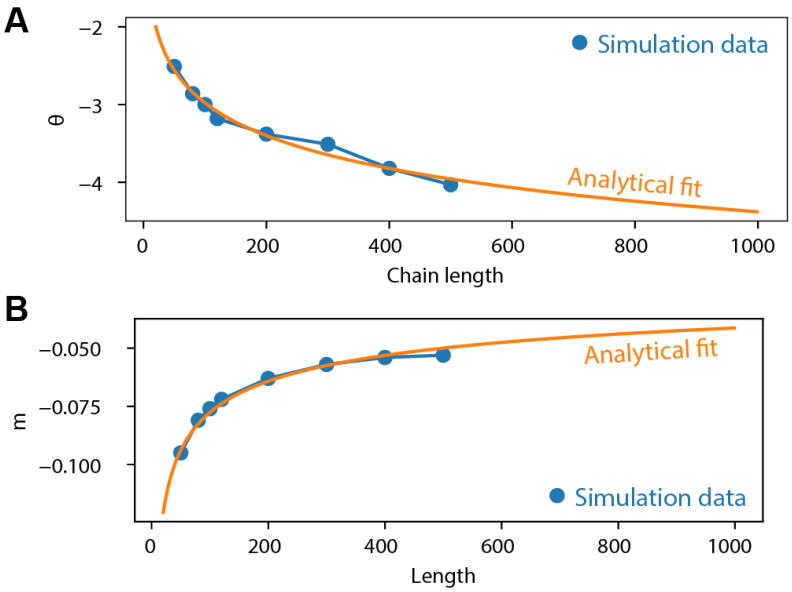

Figure A1.13. Fit of length-dependent model parameters to match PIMMS homopolymer simulations. The orange curves represent the analytical expressions defined in **Section A1.2** using the best fit parameters to fit to the experimentally measured values. **(A)** The fitting of the parameters c and d to 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.

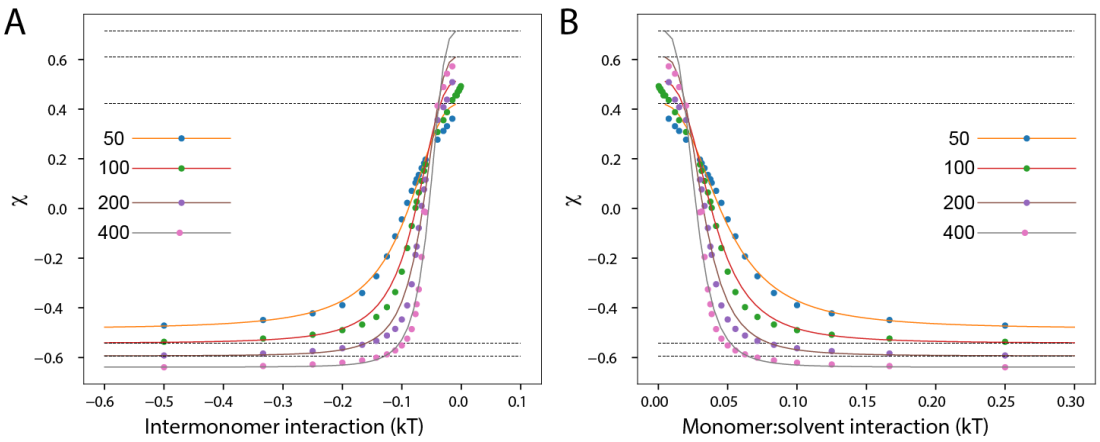

Figure A1.14. Best fit of floating parameters for analytical model (line) to PIMMS simulations (filled circles). (**A)** Data plotted in terms of inter-monomer interaction strength (assuming neutral chain-solvent interactions). **(B)** Same data plotted in terms of chain-solvent interaction strength (assuming neutral inter-monomer interactions). Colors denote chain lengths as specified in the legends.

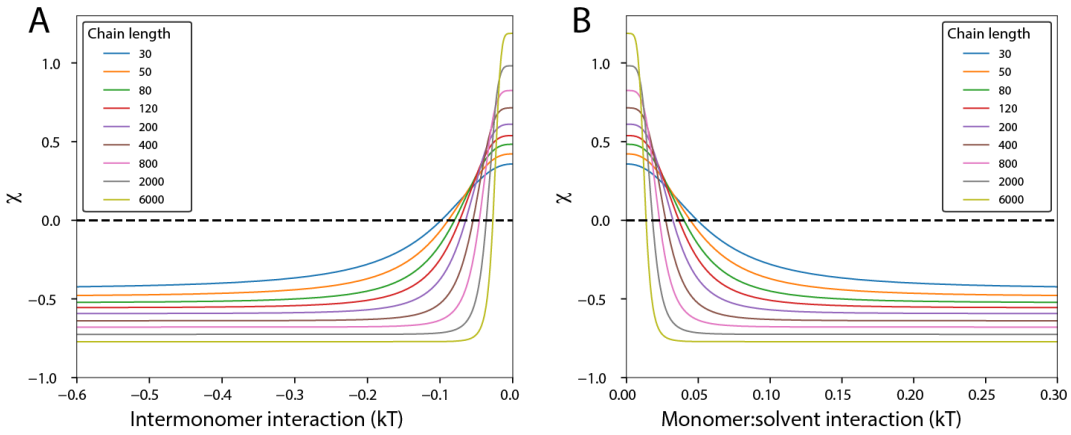

Figure A1.15. Relationship between inter-monomer interaction strength and χ. As chain length increases, cooperativity of the coil-to-globule transition increases. Note that the maximum and minimum χ values show a modest but well-defined length dependence. (**A)** Data plotted in terms of inter-monomer interaction strength (assuming neutral chain-solvent interactions). **(B)** Same data plotted in terms of chain-solvent interaction strength (assuming neutral inter-monomer interactions).

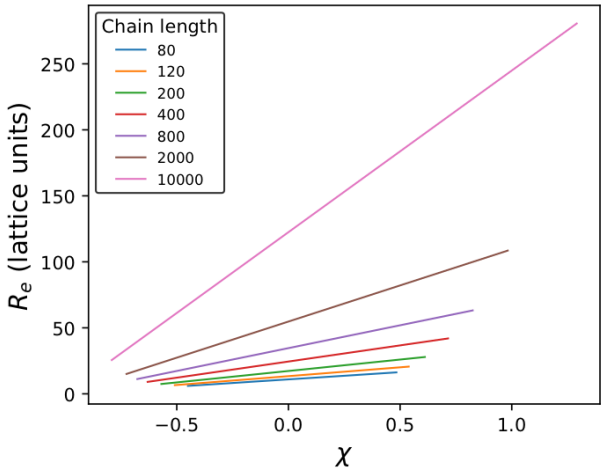

Figure A1.16. Dependence of the end-to-end distance (R_e) on χ. As the chain becomes longer, both the maximum and the steepness of the R_e-dependence on χ become larger.

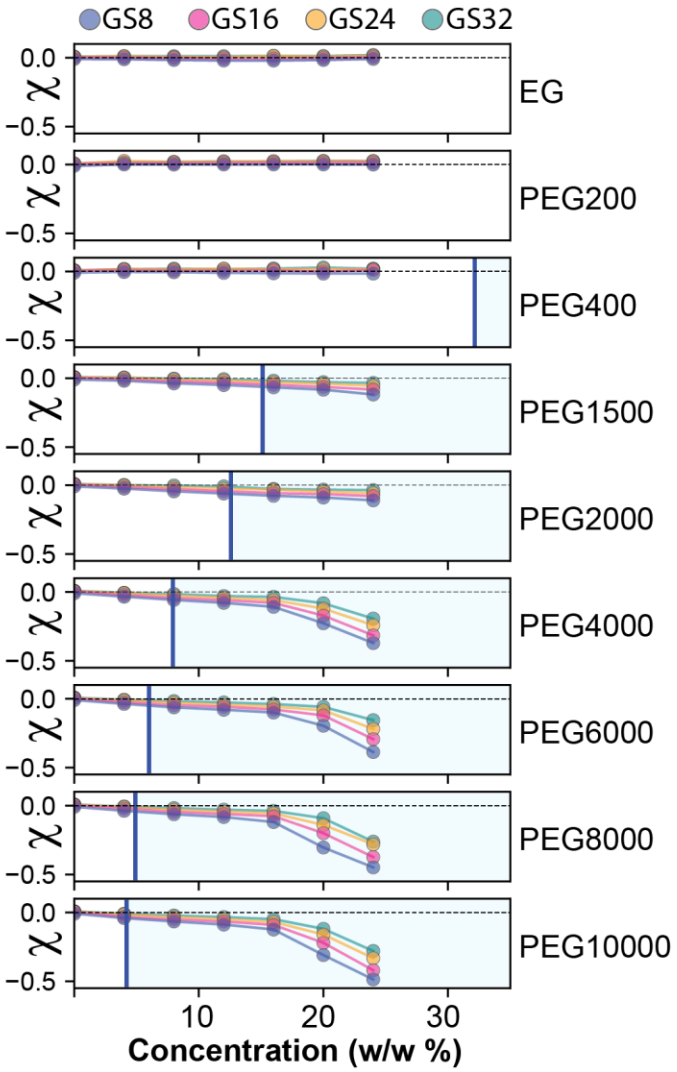

Figure A1.17: PEG-dependence of GS linkers plotted on same axes. Blue line represents overlap concentration (c^*), with concentrations higher than c^* identified by the light-blue shaded regime.

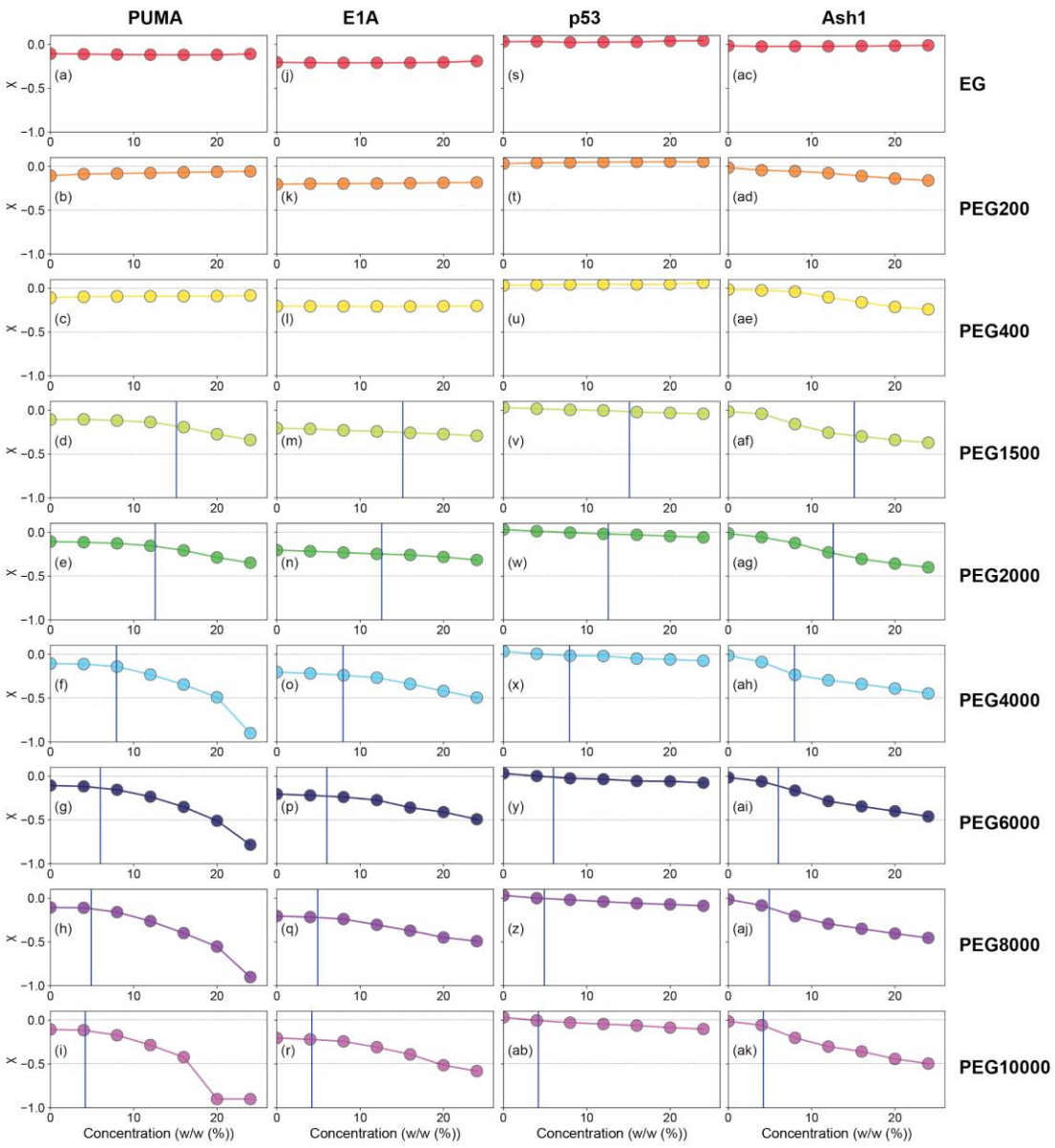

Figure A1.18: PEG-dependence of χ from naturally occurring IDRs as a function of PEG concentration and PEG molecular weight. Where present, horizontal lines are PEG-specific overlap concentration (c^*). Each of the four columns represents a distinct IDR and each row is a distinct PEG solution. **(a-i)** 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.

Supplementary Tables

Table A1.1. IDR sequences used in this study.

IDR	Number of residues	Amino acid sequence
GS0	0	
GS8	16	GSGSGSGSGSGSGSGS
GS16	32	GSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGS
GS24	48	GSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGS
GS32	64	GSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGS
GS48	96	GSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGS
PUMA	34	VEEEEWAREIGAQLRRIADDLNAQYERRRQEEQH
p53	61	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPD
Ash1	83	GASASSSPSPSTPTKSGKMRSRSSSPVRPKAYTPSPRSPNYHRFALDSPPQSPRRSSNSSITKKGSRRSSGSSPTRHTTRVCV
E1A	40	MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEP

Appendix 2

A2.1 Materials and Methods

A2.1.1 FRET construct design and cloning

The monomeric fluorescent proteins mTurquoise2 and mNeonGreen have high quantum yields, fast maturation and are highly photostable[73,75]. Compared to other FRET pairs, mTurquoise2 and mNeonGreen have a higher FRET efficiency with little cell to cell variation and were therefore selected as our FRET pair in our experiments[75]. The FRET backbone for bacterial expression (fIDR_pET-28a(+)-TEV) or for mammalian expression (fIDR_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 IDR 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 IDR inserts used in this study are shown in **Table A2.1**.

A2.1.2 FRET construct expression and purification

BL21 (DE3) cells were transformed with fIDR_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[171]. 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.

A2.1.3 Preparation of solutions for solution-space scanning

Solutes were purchased from Alfa Aesar (Sarcosine, PEG400, PEG2000), GE Healthcare (Ficoll), Thermo Scientific (Guanidine Hydrochloride), and Fisher (Ethylene Glycol, Glycine, Potassium Chloride, Sodium Chloride, Sucrose, 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 or KCl was varied, which began free of additional salt. The same buffer was used for all dilutions.

A2.1.4 *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[171,172].

A2.1.5 Calculation of FRET efficiencies

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[190]:

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

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 mTurquoise2, and Q_a = 0.8 is the quantum yield of mNeonGreen[75,172].

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[19].
(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.

A2.1.6 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. A2.2**) 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[191–193]. 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. A2.18**)[194,195] 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. A2.3**), as well as Kratky transformations (**Fig. A2.4**) and deconvolution of peaks using EFA (**Fig. A2.18**), 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[196]:

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

A2.1.7 Mammalian cell culture

HEK293T and U-2 OS 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 HEK293T cells or 10,000 U-2 OS 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% CO_2. 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 or U-2 OS cells per manufacturer's protocol. Cells were incubated at 37 °C and 5% CO_2 for 48 hours post transfection, ensuring that maturation of the FRET pairs (which occurs in 10-30 minutes[74,197]) does not alter our results. 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. Isosmotic (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.

A2.1.8 Live-cell microscopy

Imaging was done on a Zeiss epifluorescent microscope using a 10X 0.3 NA dry objective for whole-cell experiments or a 40X 0.9 NA dry objective for localization experiments. 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, **Fig. 3.1E**) or mNeonGreen at 511 nm (direct acceptor channel, **Fig. 3.1E**). 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.

A2.1.9 Image analysis

Images were analyzed using ImageJ[198]. Images collected before and after osmotic challenge, containing three channels each, were stacked and aligned using the StackReg plugin with rigid transformation (**Fig. A2.19**).[199] 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 μm², 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. A2.20**) 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. A2.21**). 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 ($E_{f,before}^{cell}$ and $E_{f,after}^{cell}$ respectively) was calculated by:

$$E_f^{cell} = \frac{F_A}{F_D + F_A}.$$

Images for localization experiments contained three channels that were stacked and aligned using the StackReg plugin with rigid transformation. The multipoint tool was used to manually select one 10 μm² circle in the cytoplasm and a second in the nucleus for each cell. The resulting measurements were filtered to remove cells with an intensity over 10,000 (to correlate with *in vitro* experiment concentrations). Cell FRET efficiency was calculated as previously stated.

A2.1.10 *In vitro* concentration dependence experiments

Protein aliquot samples were diluted into a series of varying concentrations using 20 mM sodium phosphate, 100 mM NaCl, pH 7.4 buffer. Samples were prepared on a μ-Plate 96-well black treated imaging plate (Ibidi Cat. No. 89626). Fluorescent beads (Phosphorex Cat. No. 2225) were added to the prepared aliquots to ensure focus on the bottom of the well. Imaging parameters were the same parameters as were used for the live-cell microscopy experiments. Images were also analyzed using ImageJ[198]. Instead of segmentation, the center of the images were selected and the average pixel intensities were measured. In order to correlate emission with concentration, we plotted protein concentration against direct acceptor emission (**Fig. A2.20**).

A2.1.11 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% H_2O, 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 x 30 mm, 5 µm, 100 Å) (**Fig. A2.22**). 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. A2.23**, **Table A2.2**).

A2.1.12 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.

A2.1.13 All-atom simulations of constructs with fluorescent proteins

To verify that our *in vitro* SAXS and FRET results report on the same conformational ensemble, we performed all-atom simulations of full-length constructs that include both fluorescent proteins using an identical amino acid sequence to the experimental constructs. Fluorescent protein models were constructed from PDB files 4AR7 (mTurquoise2)[200] and 5LTR (mNeonGreen)[74]. Simulations were performed using the ABSINTH implicit solvent model and CAMPARI Monte Carlo simulation engine[83].

Considering the size of these proteins, simulating them at full-length and all-atom resolution raises a number of challenges. Given that our objective here was to determine whether SAXS and FRET were in agreement in the context of a simple homopolymeric linker, we took advantage of the ABSINTH implicit forcefield's ability to tune specific components of the Hamiltonian. Specifically, we performed simulations in which all excluded-volume interactions were present (i.e., the repulsive component of the Lennard-Jones potential was turned on). However, the attractive component of the Lennard-Jones potential was only turned on for residues within the glycine-serine (GS) linker, and limited only to intra-linker interactions by varying the inherent Lennard-Jones parameters of all atoms outside of the GS linker. Beyond these two components, all additional non-bonded Hamiltonian terms (i.e., long and short-range electrostatics and solvation effects) were turned off, dramatically lowering the computational cost of simulations. By systematically tuning the overall strength of the attractive GS repeat intramolecular interactions, we in effect performed simulations for GS homopolymers for all relevant homopolymer interaction strengths and GS-repeat lengths from 0 to 48 (i.e., 0 residues to 96 residues).

We initially performed simulations using a GS0 construct, where the only backbone degrees of freedom available were associated with the set of flexible residues that connect the two beta-barrels. Specifically, all backbone dihedral angles for amino acids within the two beta-barrels were switched off, but all sidechain degrees of freedom were accessible. The residues between the two beta-barrels that had their backbone degrees of freedom sampled consist of amino acids 227 to 255 (GITLGMDELYKEGLSKLMVSKGEEDNMAS) in the GS0 construct[19]. After running thousands of short independent simulations in which these twenty-nine amino acids were sampled with variable intramolecular interaction strengths, we subselected an ensemble of 1000 distinct conformations which, on average, reproduced the experimentally measured SAXS scattering data for the GS0 construct (**Fig. A2.5A**). This GS0 ensemble was then used to define the starting configurations of the FPs and the 'handles' (non-GS component of the construct) for all other GS simulations.

For each of the other GS-repeat lengths (8, 16, 24, 32, 48), we performed simulations in which the attractive Lennard-Jones potential was scaled from 0.30 to 0.62 in steps of 0.02. This range straddles intramolecular interaction strengths that cause the longer GS chains to behave as a self-avoiding random coil (attractive LJ scaling parameter = 0.3) and a compact globule (attractive LJ scaling parameter = 0.62). For each combination of GS length and LJ strength, we

performed 1000 independent simulations in which the fluorescent proteins and associated handles defined in the GS0 simulations were also fixed in place. As such, in total we performed 17,000 independent simulations for each separate GS length (i.e., 85,000 independent simulations in total). This approach enables to in effect construct a large collection of ensembles with (defined GS lengths but variable GS intramolecular interaction strengths) from which we will ultimately subselect and concatenate many individual simulation trajectories based on the agreement between the simulated scattering profile and the real scattering profile. These data will be used to construct a sub-ensemble that recapitulates the experimental scattering data – i.e., an unbiased, data-driven approach to construct an ensemble consistent with the experimental measurements.

Each simulation was run in a spherical droplet with a radius of 500 Å to avoid any possible finite size effects. Given the absence of any electrostatic components, no ions were included in the simulations. Each simulation was run for 100,000 Monte Carlo steps. The first 50,000 steps were discarded as equilibration, and conformations were then sampled every 5000 steps. As such, each independent simulation generated 10 conformations, such that each GS/LJ combination generated a 10,000 conformer ensemble (1000 independent simulations with 10 conformations per simulation). Other than the repulsive component of the Lennard-Jones potential and (for some atoms) the attractive component of the Lennard-Jones potential, all other modes of nonbonded interactions were switched off. As such, each individual simulation takes on the order of 10 minutes.

Having performed this set of simulations we calculated predicted scattering profiles for each independent simulation using FoXS software, as described previously[117,201]. To assess the agreement between each short simulation and the experimental scattering data we computed χ^2_{free} , a parameter explicitly developed to assess the goodness-of-fit for scattering data[202].We calculated χ^2_{free} for GS-length matching simulations to assess how well each length-matched sub-ensemble compared to the experimentally measured scattering data. Using a χ^2_{free} of 3.2 (a large value that reflects the relatively small error in the experimentally measured SAXS data), we generated sub-ensembles with scattering curves that quantitatively reproduced the experimental data at each of the GS-repeat lengths (**Fig. A2.5A**).

Finally, using the SAXS-matched sub-ensembles, we computed the inter-barrel distance based on the distance between two residues in the center of the

beta-barrel (**Fig. A2.5B**). Distances were calculated between alpha-carbon atoms, such that we subtracted a 6 Å offset to approximately account for the distance between the alpha-carbon atoms and the anticipated chromophore centers. The resulting inter-beta-barrel distances are in excellent agreement with distances measured from ensemble FRET experiments. For **Fig. 3.2B**, these end-to-end distances (R_e) were converted to E_f values using the equation $E_f = R_0^6 / (R_0^6 + R_e^6)$, assuming R_0, the Förster distance for the mTurquoise2-mNeonGreen FRET pair, to be 62 Å[172]. Taken together, this approach shows that the ensembles that best describe the SAXS data also correctly describe the distances inferred from FRET, confirming that these orthogonal methods are reporting on the same underlying conformational ensemble. The final sub-ensembles for each GS-repeat length and the associated data are provided in https://github.com/sukeniklab/IDP_structural_bias. Simulation analysis was performed with SOURSOP (https://soursop.readthedocs.io/)[203].

A2.1.14 All-atom simulations of IDRs

Simulations of label-free IDR sequences used in this study were done using the CAMPARI simulation suite and the ABSINTH forcefield[83,177]. For each sequence, five independent simulations were run at 310 K using 8x10^7 Monte Carlo steps (following 1x10^7 steps of equilibration) starting from random conformations to ensure proper sampling. Protein conformations were written out every 12,500 steps. The end-to-end distance and the helicity of the simulated conformation ensembles were determined using the MDTraj python library[178].

Supplementary Figures

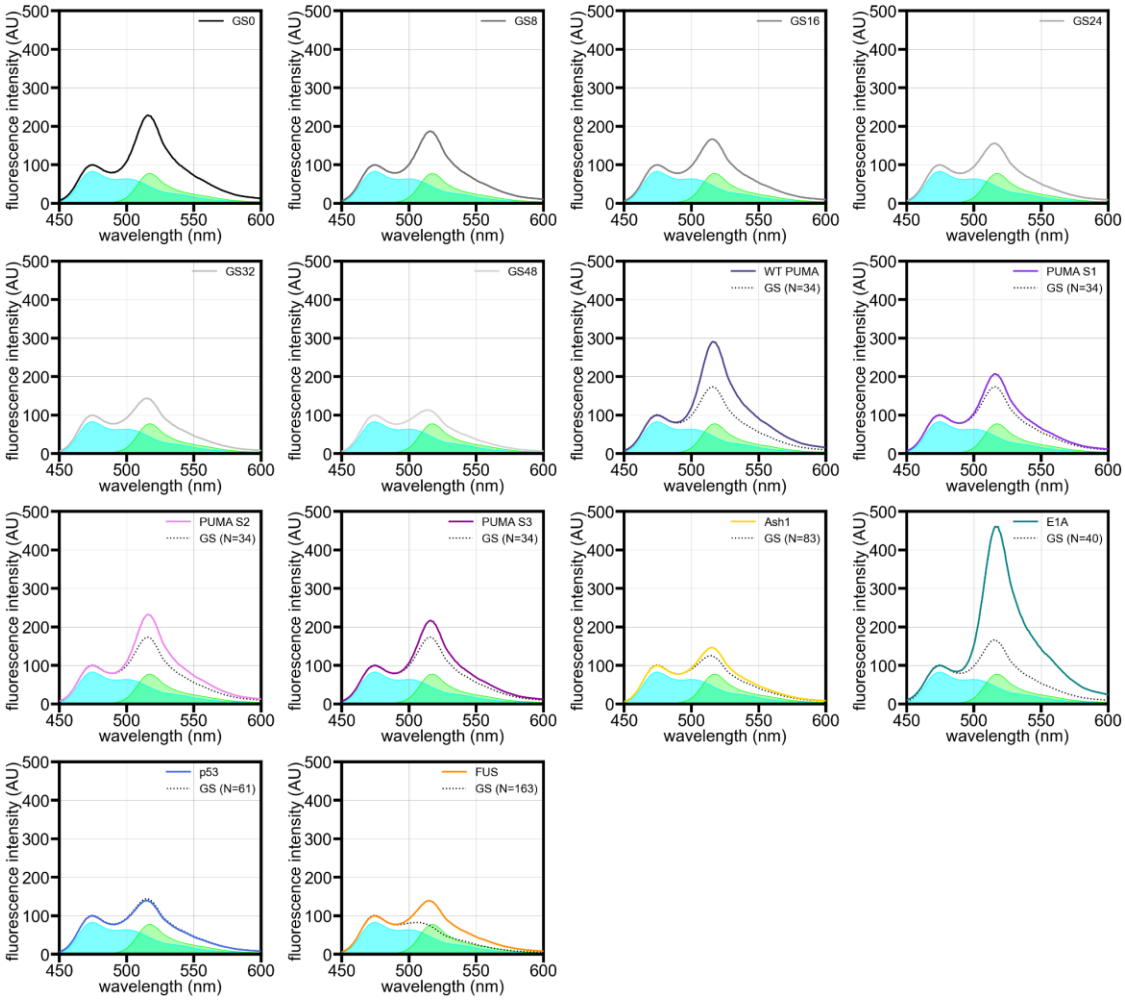

Figure A2.1. Fluorescence spectra from measurements of FRET constructs in a dilute phosphate buffer solution. Spectra of constructs incorporating IDRs that are not GS-repeat sequences are compared with expected fluorescence spectra of FRET constructs incorporating GS-repeat sequences of equal length (black dotted curves), where N refers to the number of amino acids. Blue and green shaded areas are the base spectra for mTurquoise2 and mNeonGreen, respectively, in the same buffer solution.

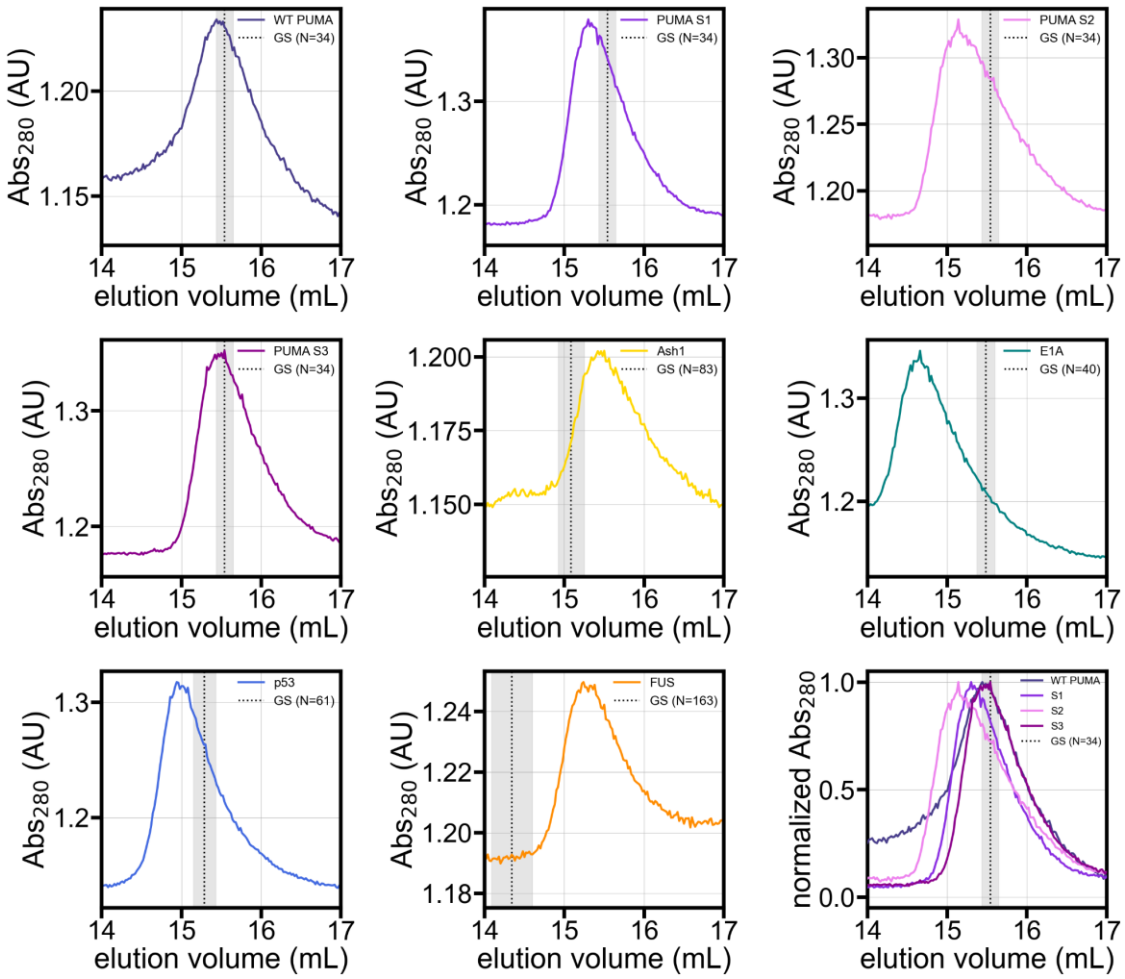

Figure A2.2. Chromatograms from SEC-SAXS experiments in which the samples were donor-IDR-acceptor FRET constructs in a dilute phosphate buffer solution. Vertical dotted line labeled “GS” in each panel represents the expected elution peak position of a FRET construct containing a GS-repeat sequence equal in length to the IDR, where N refers to the number of amino acids. Shaded region in each panel represents the standard error of the expected GS peak position.

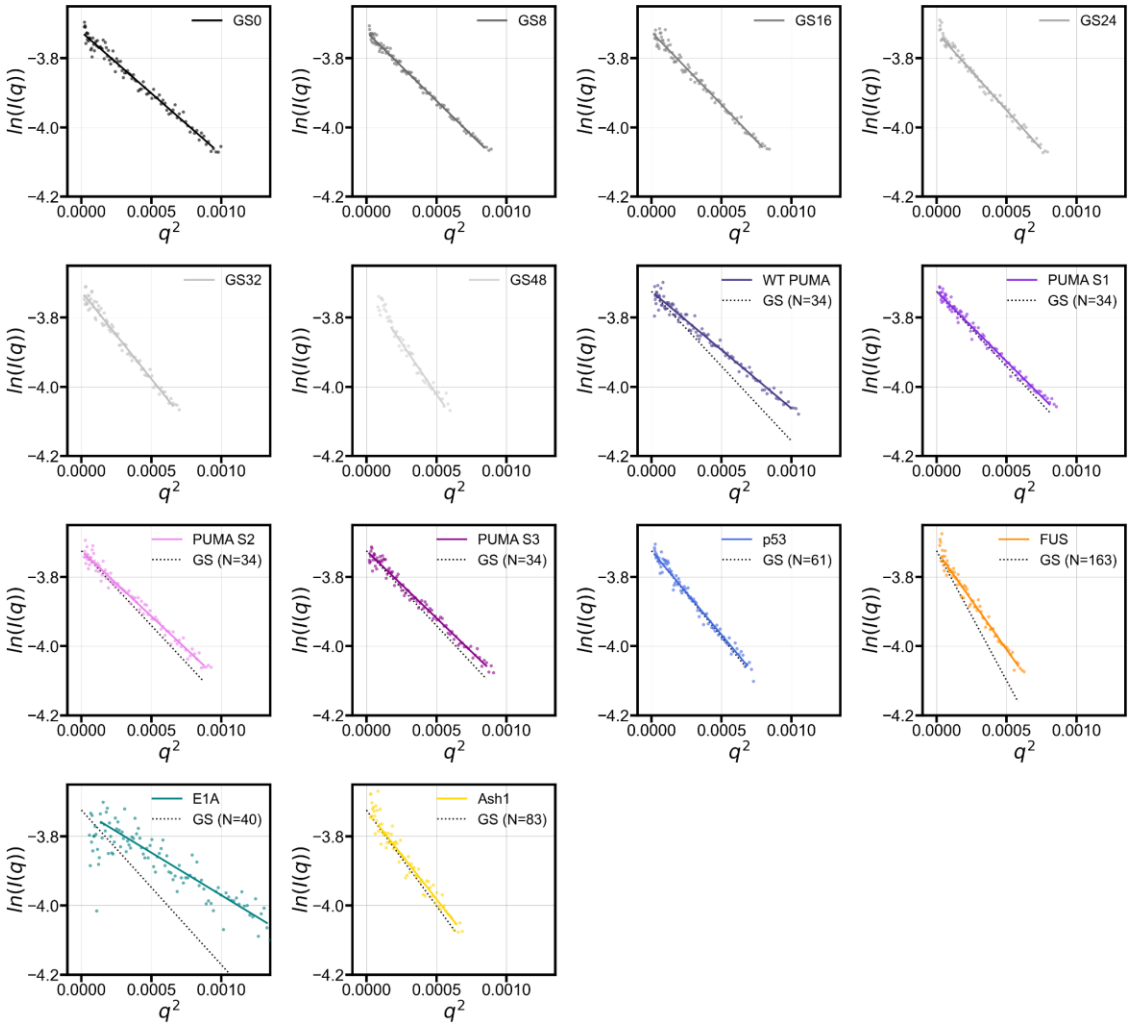

Figure A2.3. Guinier plots for donor-IDR-acceptor FRET constructs from SEC-SAXS experiments. For IDRs that are not GS-repeat sequences, the fitted line is compared with the expected fitted line for a construct containing a GS-repeat sequence of the same length (black dotted lines), where N refers to the number of amino acids. Lines are fitted to a maximum $q * R_g$ value of 1.

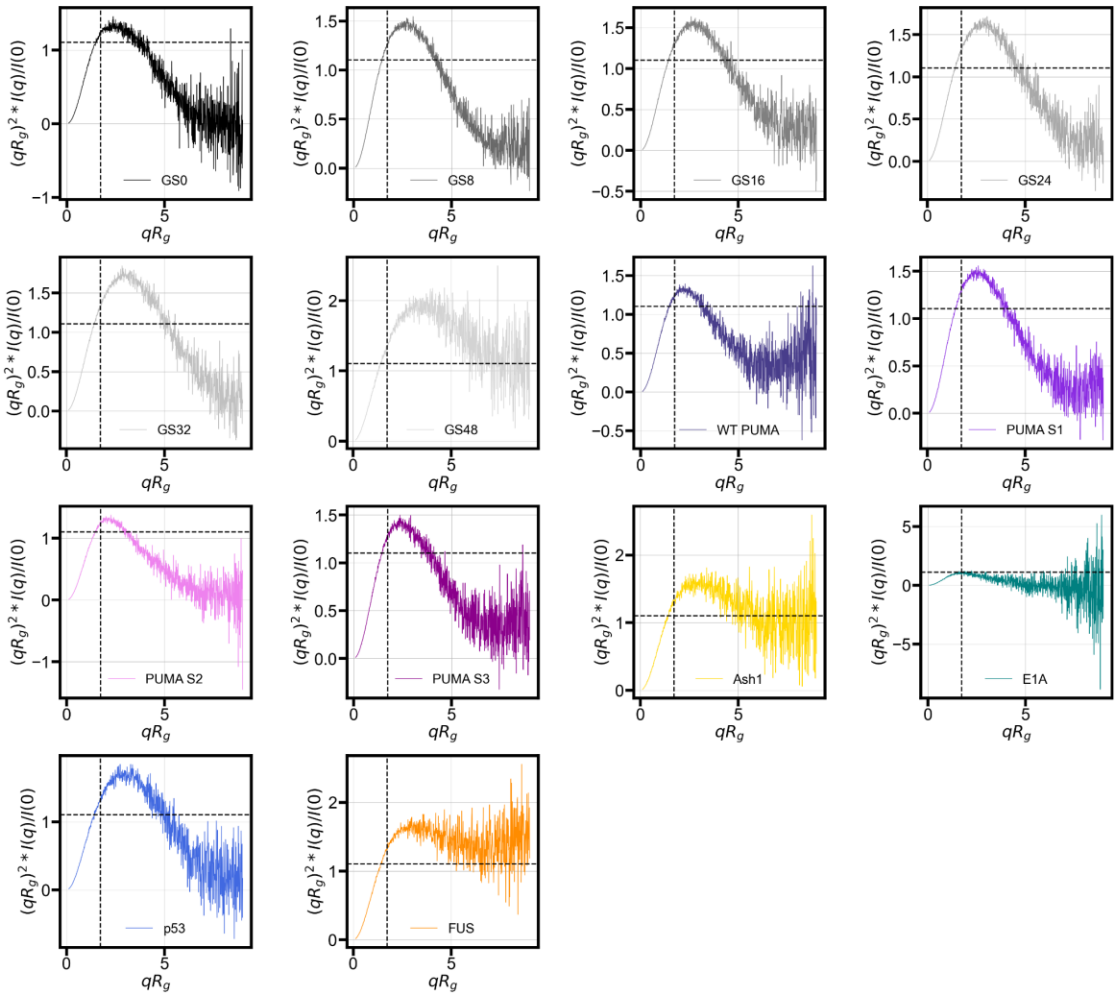

Figure A2.4. Dimensionless Kratky plots derived by transforming the scattering profiles from which the R_g values reported in the main text were calculated. For a globular protein, the peak position should be at $qR_g = \sqrt{3} \sim 1.73$ (shown by vertical dashed line) and the peak height should be $(qR_g)^2 * I(q)/I(0) = 3/e \sim 1.1$ (shown by horizontal dashed line).

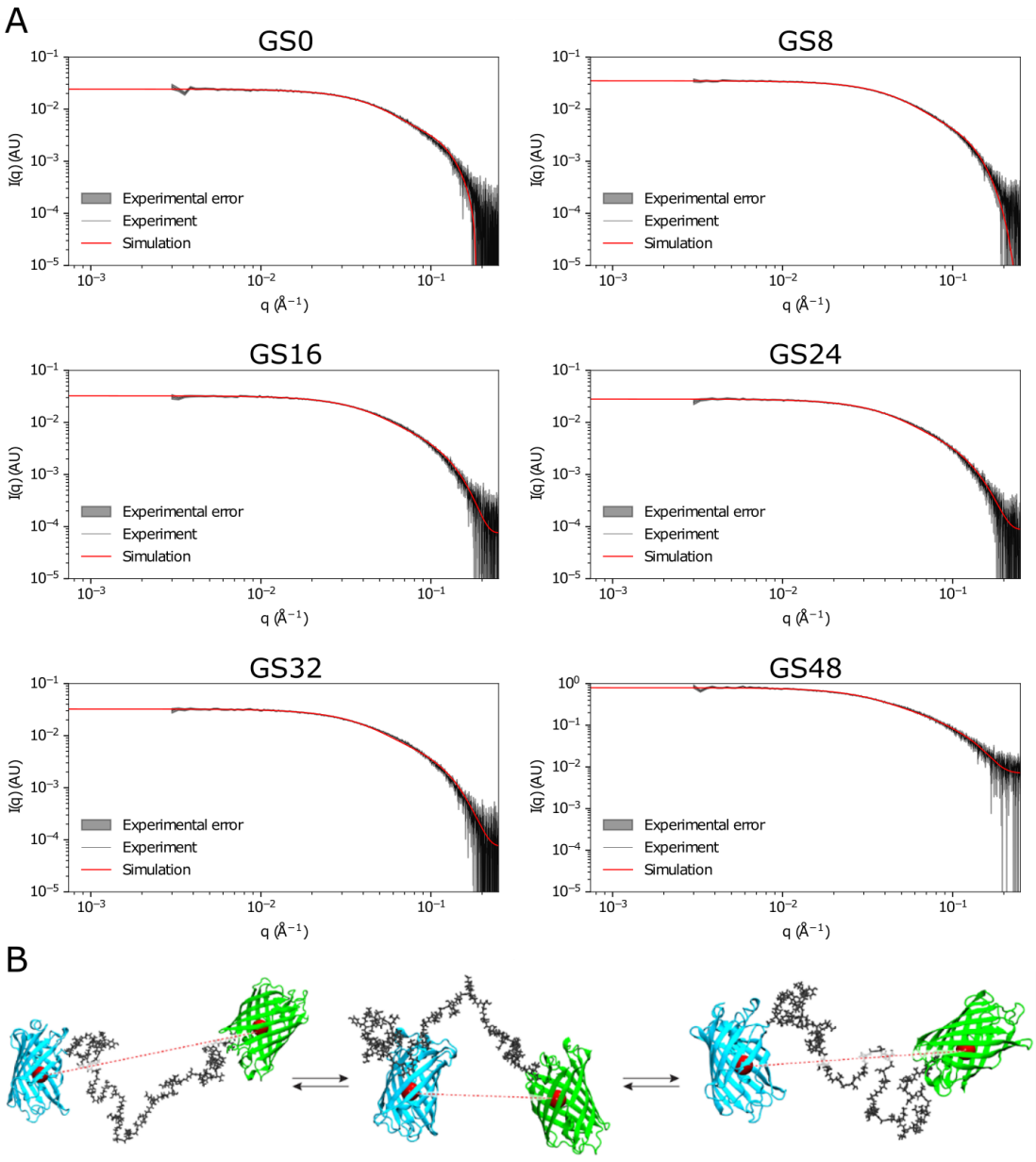

Figure A2.5. (**A**) Comparison of experimentally measured small-angle X-ray scattering profiles with simulation-derived synthetic for GS-repeat sequences of different lengths. (**B**) Example snapshots from simulations. R_e^{app} is calculated based on the distance between residues from the center of the two beta-barrels (dashed red line).

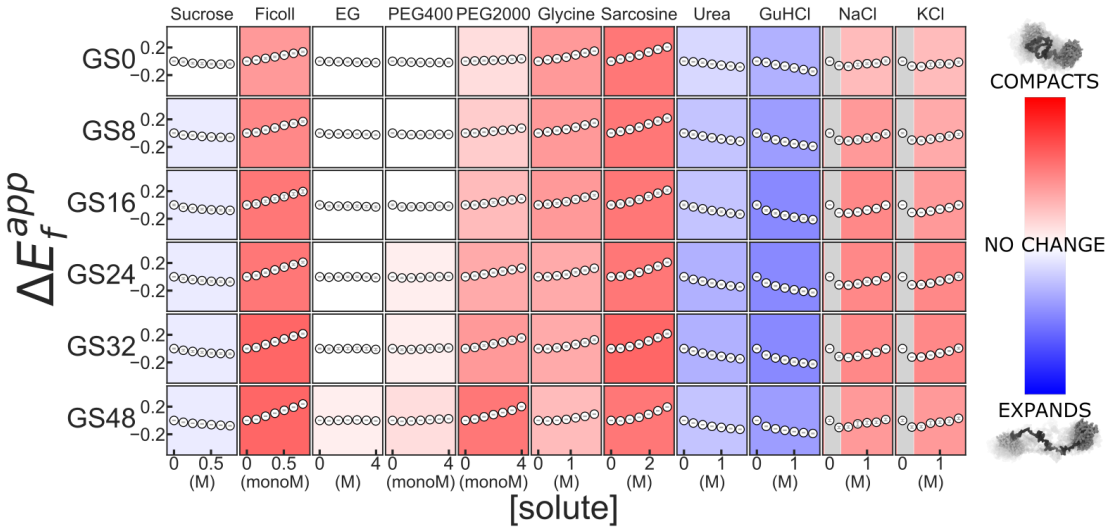

Figure A2.6. Solution space scans of GS-repeat homopolymers. Each cell shows ΔE_f^{app} as a function of increasing solute concentration. Blue background indicates expansion and red indicates compaction, with deeper colors indicating more change. monoM: Concentration of a polymer expressed as a concentration of monomeric units. Light gray shaded regions on left side of cells for solutes NaCl and KCl: approximate range of concentrations within which electrostatic screening is the dominant effect; the leftmost two points of each series, since they are within that range, are not used in the assignment of background color.

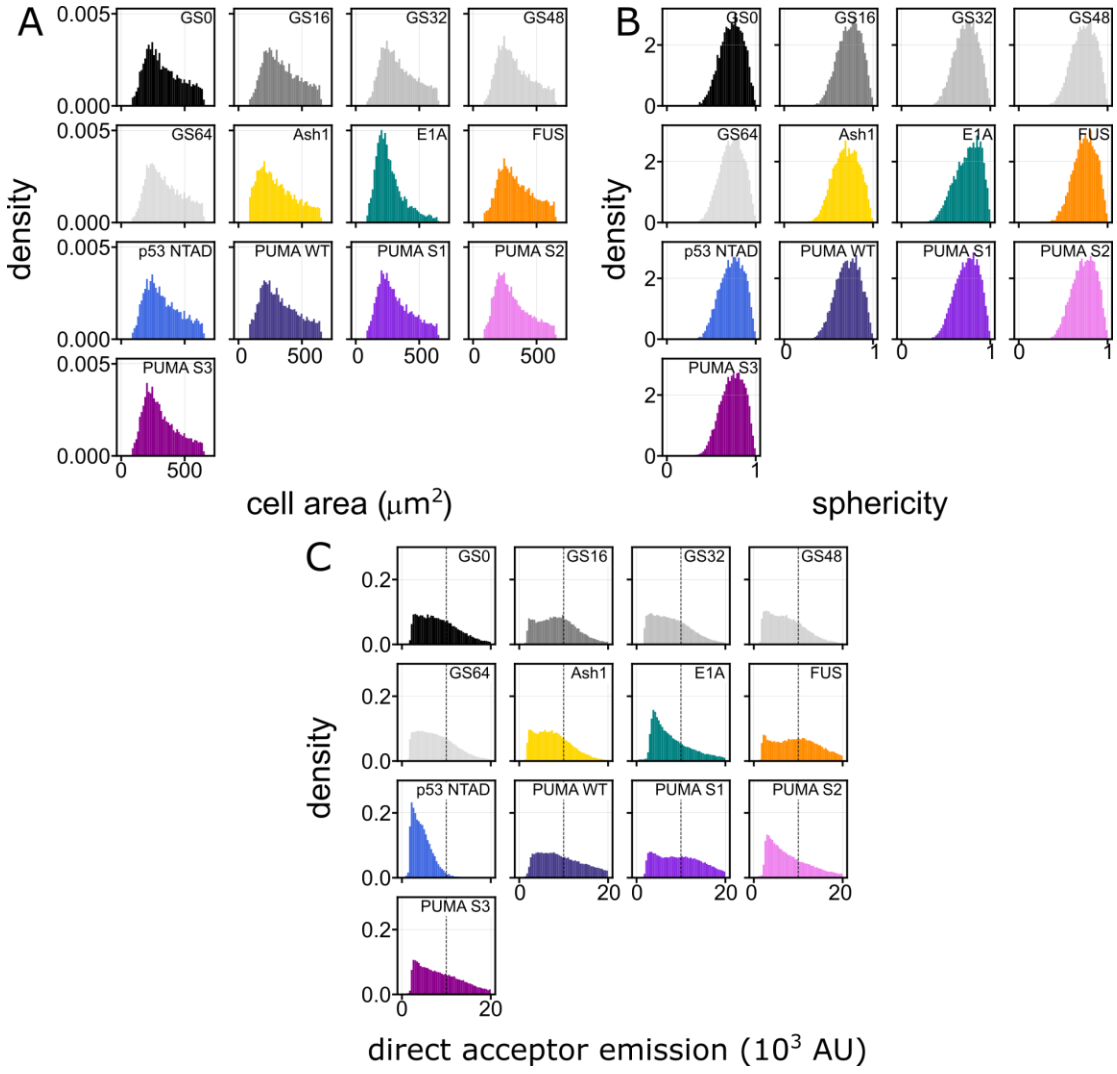

Figure A2.7. Probability density of cellular features across cells expressing different FRET constructs. Each histogram contains over 10^3 cells, with a total of over 2.1×10^5 cells. (**A**) Cell area (in μm²). (**B**) Cell circularity calculated as $4\pi(\text{area}/\text{perimeter}^2)$. A value of 1 is a perfect circle and as the value approaches 0 it indicates an increasingly elongated polygon. The HEK293T cells used in these experiments tend to be rounder than others, resulting in a sphericity close to 1. (**C**) Direct acceptor emission (cells excited at 511 nm) is a metric for in-cell construct concentration. To eliminate artifacts resulting from high overexpression, and to facilitate accurate comparison with *in vitro* measurements done at 1 uM, we left cells with a direct acceptor emission higher than 10,000 out of the analysis, indicated by the dashed vertical line (see also **Fig. A2.17**).

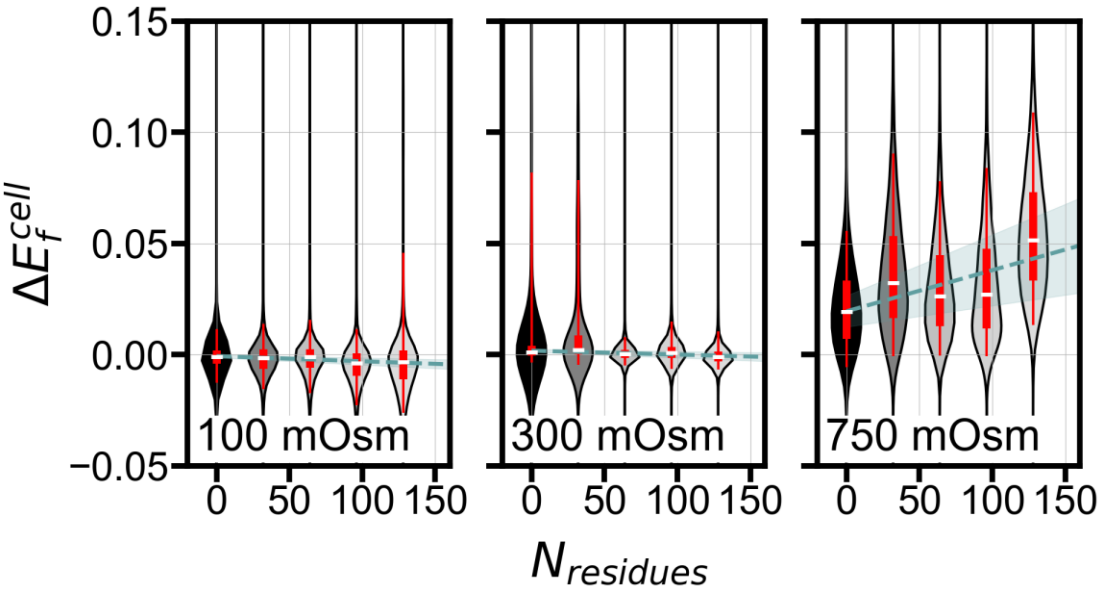

Figure A2.8. Linear fit of ΔE_f^{cell} as a function of GS linker length for hypo, iso, and hyperosmotic perturbations. Dashed green line is a linear fit of the medians, shown as white dashes, and shaded areas are the errors of the fit. Thick and thin red bars span median 50% and 90% of the data, respectively.

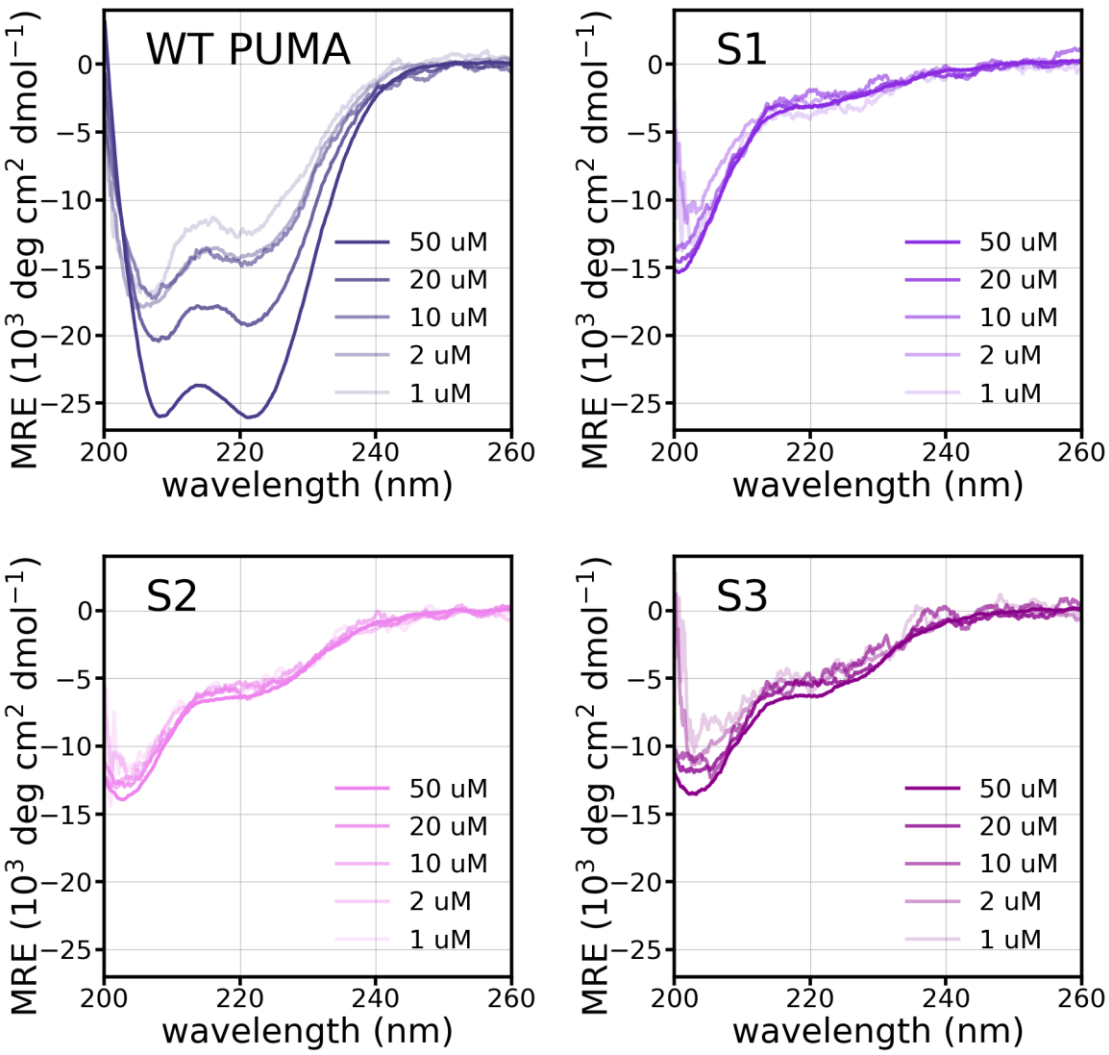

Figure A2.9. Concentration dependence of circular dichroism measurements of PUMA WT and sequence scrambles.

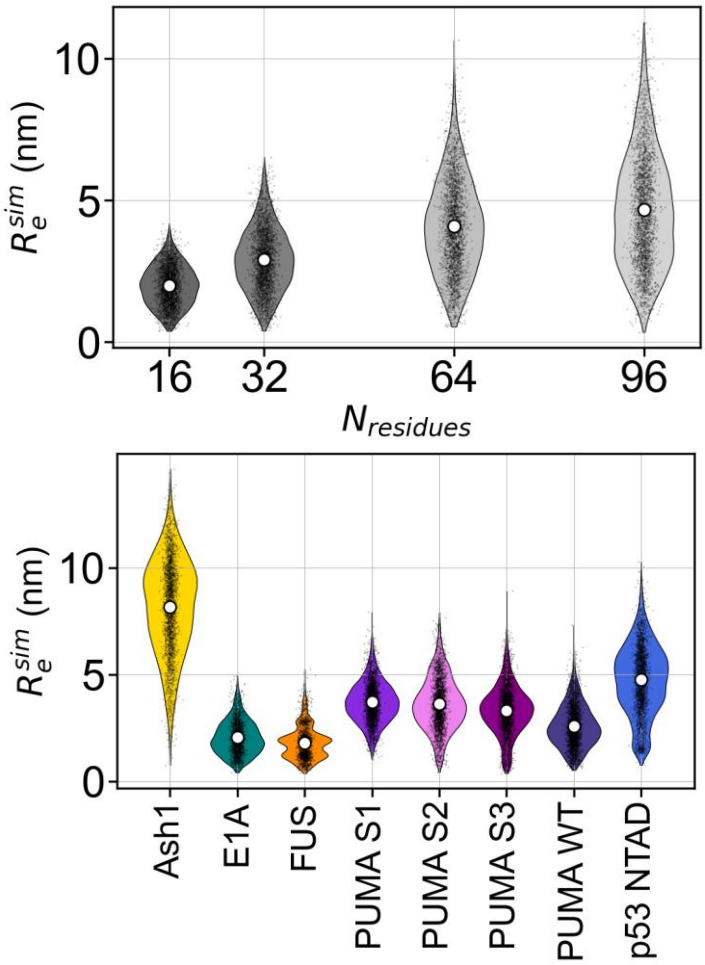

Figure A2.10. All-atom simulations, using the ABSINTH forcefield[83], of all sequences in this work. Simulations did not contain the fluorescent protein labels. Violin plots are obtained from random sampling of 2000 frames from ensembles containing at least 20,000 conformations. White circles represent the mean and black points are individual frames. FUS shows a multimodal, highly compact distribution since its ensemble is collapsed and poorly sampled, and is likely not indicative of the ensemble for this construct.

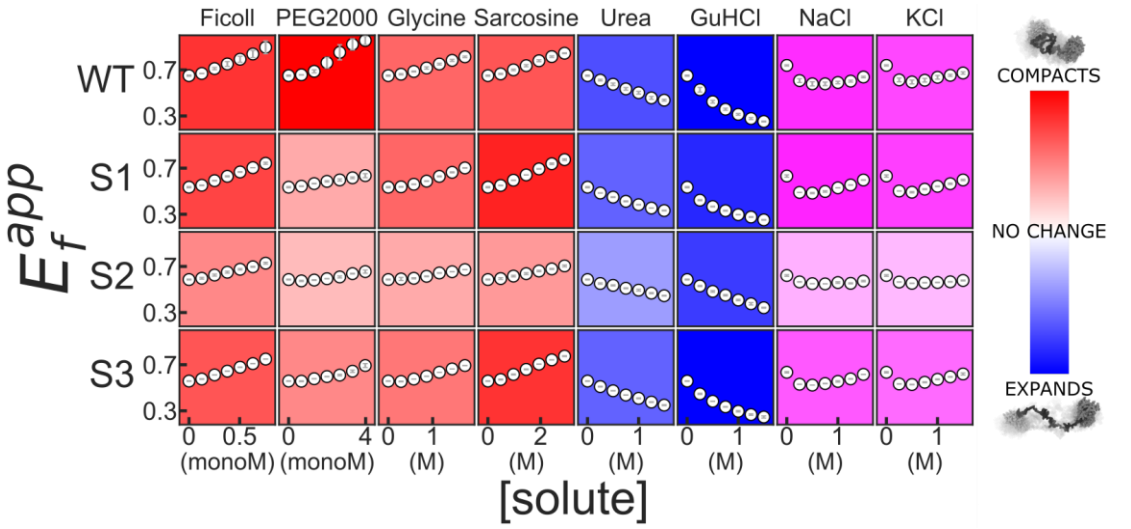

Figure A2.11. Solution space scans of WT PUMA and sequence scrambles. Each cell shows E_f^{app} as a function of increasing solute concentration. Blue background indicates expansion and red indicates compaction, with deeper shades indicating more change. Purple background indicates a non-monotonic response, with deeper shades representing more curvature. monoM: Concentration of a polymer expressed as a concentration of monomeric units.

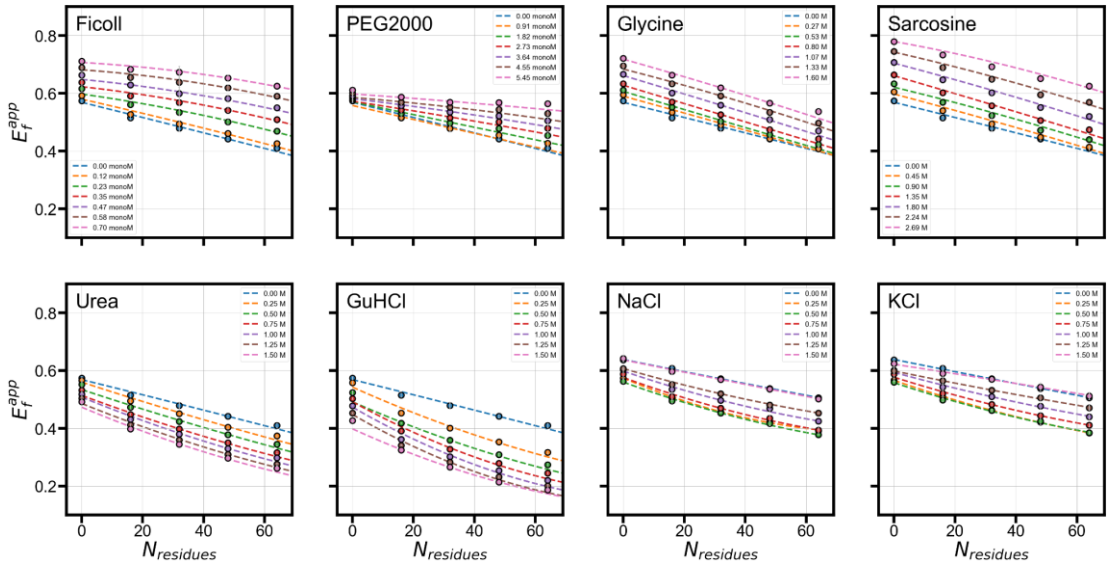

Figure A2.12. E_f^{app} vs. length of GS-repeat sequence in various solution conditions. A second degree polynomial fit, shown as dashed lines, is used for interpolation of ΔE_f^{app} for arbitrary sequence lengths in **Figs. 3H** and **4F**.

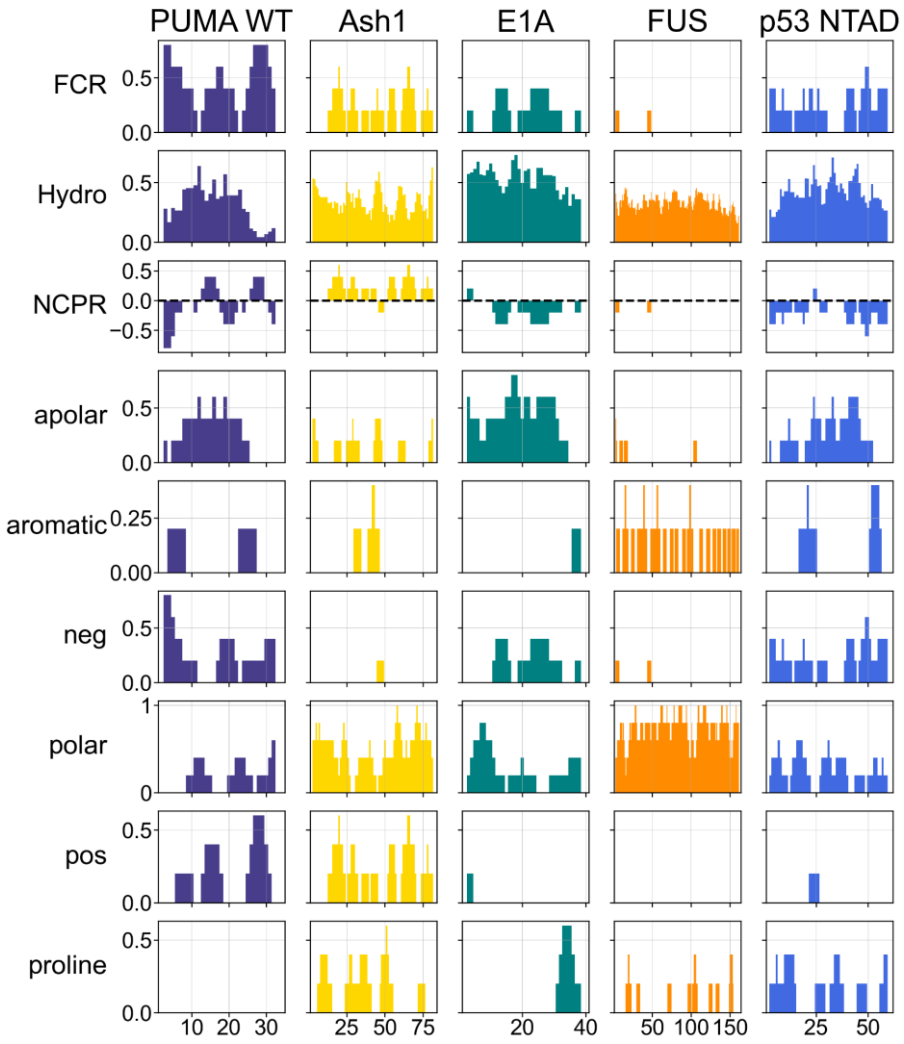

Figure A2.13. Sequence features of IDR sequences in **Fig. 3.4**, calculated using localcider[204]. All bars represent the average over a five-residue window centered at the specified residue number. FCR: Fraction of charged residues; Hydro: Kyte-Doolitle hydrophobicity scale; NCPR: net charge per residue; apolar: fraction of ALMIV residues; aromatic: fraction of FYW residues; neg: fraction of ED residues; pos: fraction of KR residues; polar: clusters of QNSTGHC residues; proline: fraction of P residues.

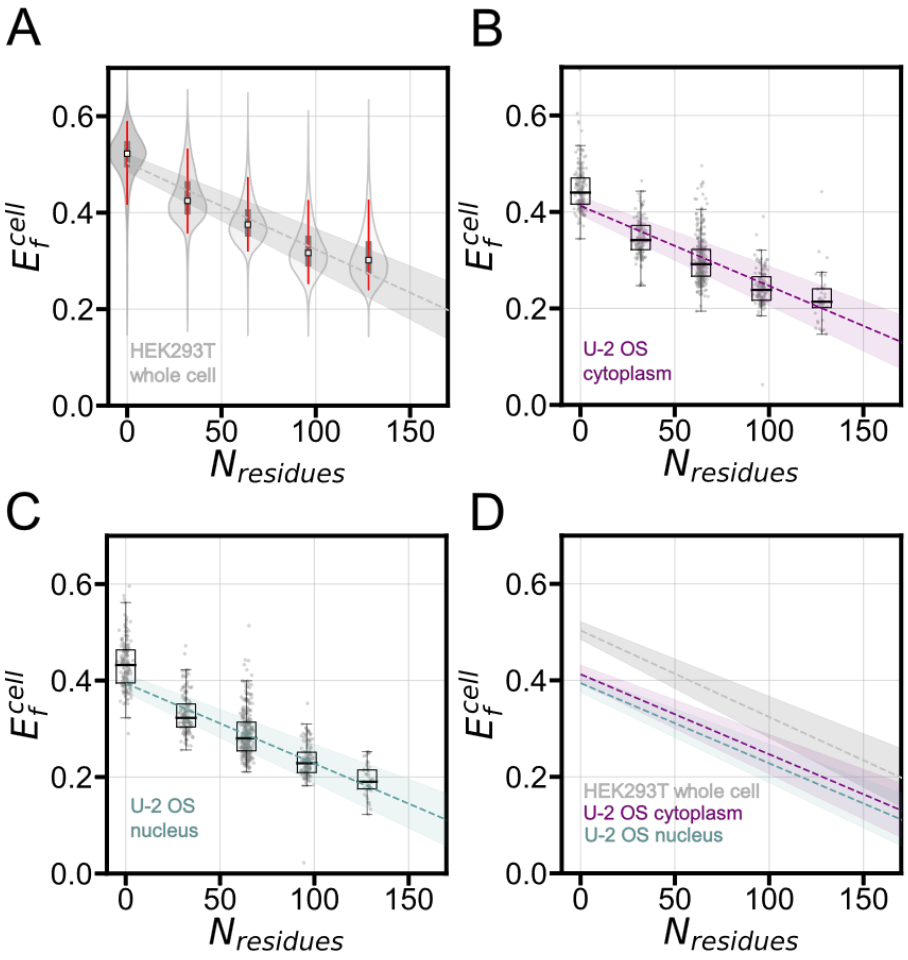

Figure A2.14. FRET efficiencies of GS-repeats measured in HEK293T and U-2 OS cells. (**A**) HEK293T live cell measurements - identical to Fig. 2G. The violin plots span the entire dataset and their thickness represents the probability. The median is shown as a white square, and the median 50% and 95% are shown as thick and thin lines at the center of the violin, respectively. The gray line is a linear fit of the medians, and fit error is shown by the shaded region. Images were taken at 10X with N > 5000 for each violin plot. (**B-C**) Cytoplasm and nucleus measurements taken in U-2 OS cells and at high magnification. The box represents the 25th and 75th percentiles, the whiskers showing the minimum and maximums for each GS-repeat construct, and the median is shown as a black line. The purple and blue lines are the linear fits of the medians and fit error is shown by the shaded regions. Points correspond to individual cells. Images were taken at 40X with N > 25 for each box plot. (**D**) The linear fits of the medians with fit error shown by the shaded regions for HEK293T (gray), U-2 OS cytoplasm (purple), and U-2 OS nucleus (blue).

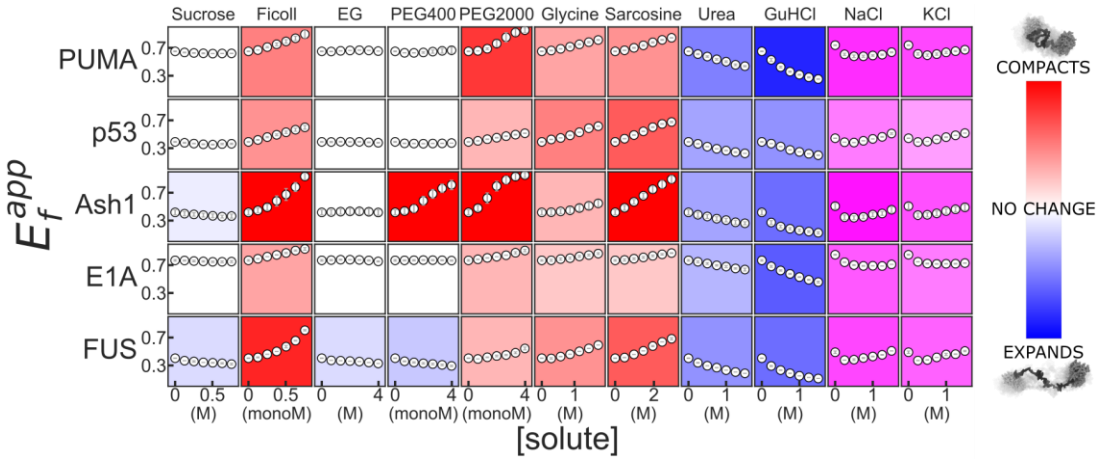

Figure A2.15. Solution space scans of naturally occurring IDRs. Features are as in **Fig. A2.11**.

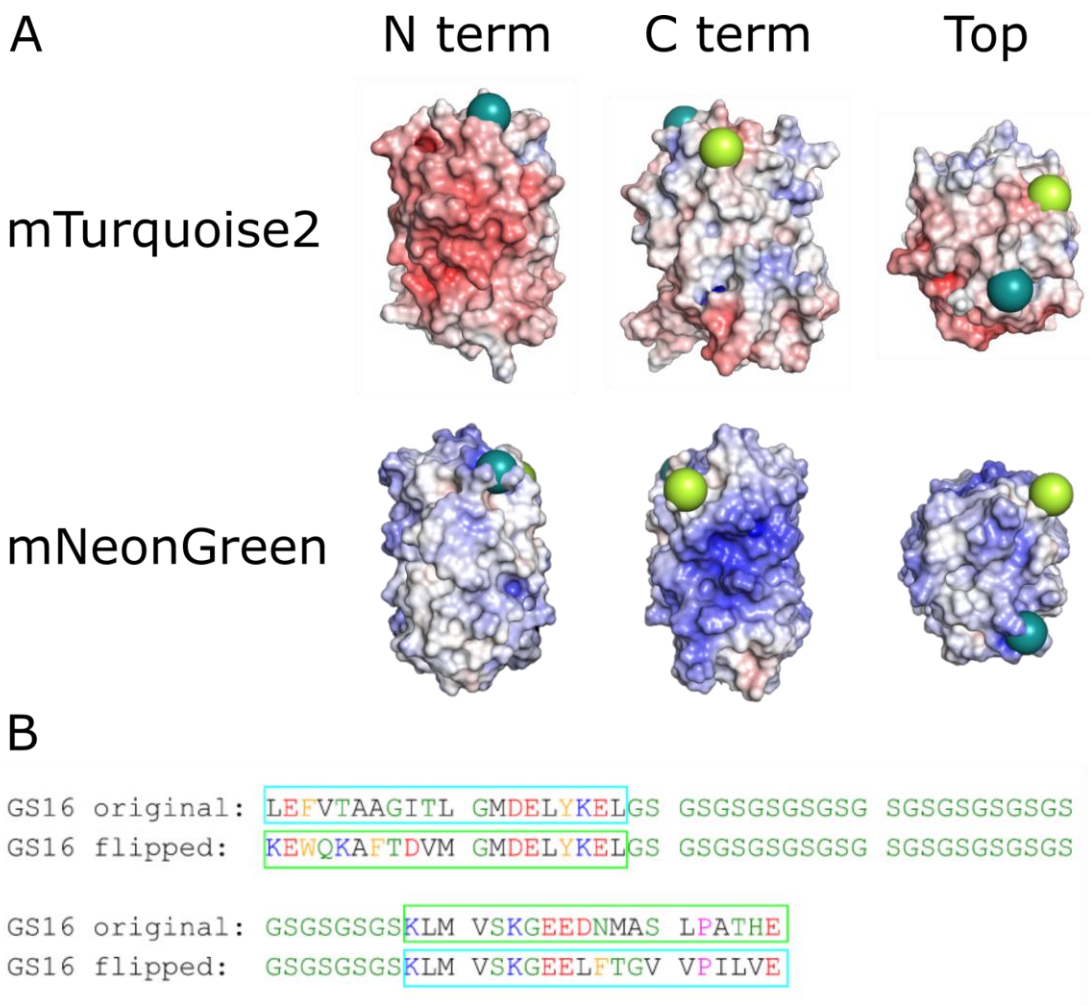

Figure A2.16. Original and flipped GS16 repeat constructs. (**A**) Surface electrostatic analysis using APBS[205] shows different surface charges for mNeonGreen and mTurquoise2. The N and C termini are labeled as yellow and cyan spheres, respectively. (**B**) Sequences of GS16 with the nearest 20 residues from the flanking fluorescent proteins, aligned using Clustal Omega[206,207]. Color codes are from CIDER analysis[208]. Red: negative charge; blue: positive charge; black: hydrophobic residues; green: polar residues; orange: aromatic residues. Blue and green boxes show residues at the terminals of mTurquoise2 and mNeonGreen, respectively, connected to the GS-repeat sequence.

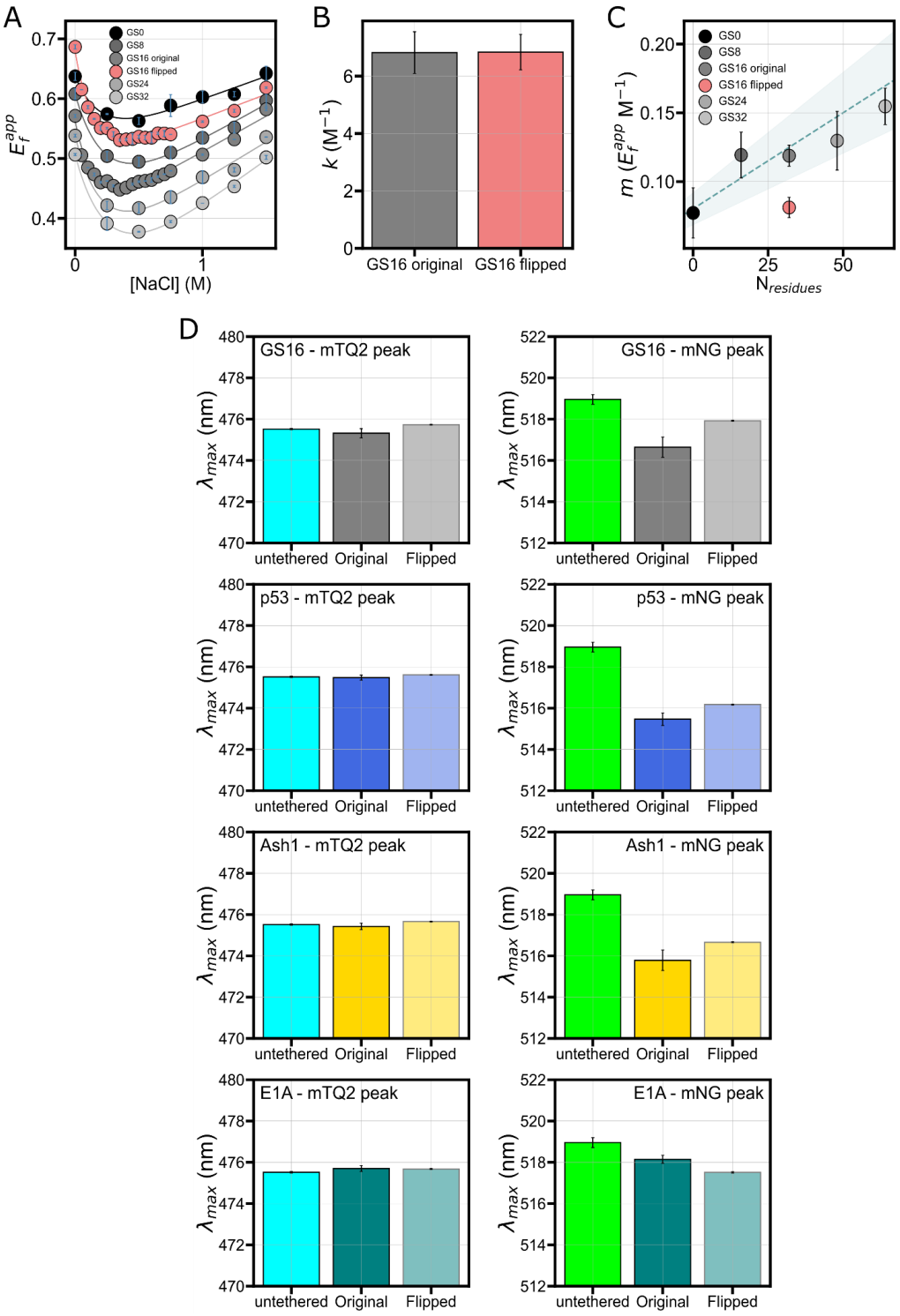

Figure A2.17. (**A**) FRET experiments showing effect of NaCl titration on original GS-repeat constructs and the GS16 construct with FP locations flipped. The experiments for GS0, GS8, GS24 and GS32 are the same as shown in **Fig. A2.6**, while measurements at a greater number of low NaCl concentrations were performed on original and flipped GS16 to facilitate a precise direct

comparison of electrostatic effects. Experimental data was fit to an exponential decay with a sloping baseline, $E_f^{app}([NaCl]) = Ae^{(-k[NaCl])} + m[NaCl] + b$. In this equation k is a decay constant that indicates the effect of screening of electrostatic interactions on ensemble structure, and m is a linear slope that accounts for the specific interactions of the ions at higher concentrations [19,82,209]. (**B**) Comparison of k of the original and flipped GS16 constructs. Identical k for the original (6.8 ± 0.7 M^{-1}) and the flipped construct (6.8 ± 0.6 M^{-1}) indicates that electrostatic interactions cannot explain the difference in E_f^{app} between the two constructs. (**C**) Slope m vs. the length of all GS-repeat sequences. All original GS repeats show a linear relationship between m and length. The flipped GS16 construct falls below this line, indicating a tighter packing of one or both of the FPs. (**D**) Comparison of the peak emission wavelengths for mTurquoise2 (left) and mNeonGreen (right) untethered vs. in the original and flipped constructs compared in **Fig. 3.5B** and **C**. For mNeonGreen, $P < 0.0001$ for untethered vs. each original construct and untethered vs. each flipped construct (n=24 for mNeonGreen untethered, n=24 for each original construct, and n=12 for each flipped construct). The shifts of the mNeonGreen peak indicate changes in mNeonGreen between the original and flipped construct as a result of IDR presence.

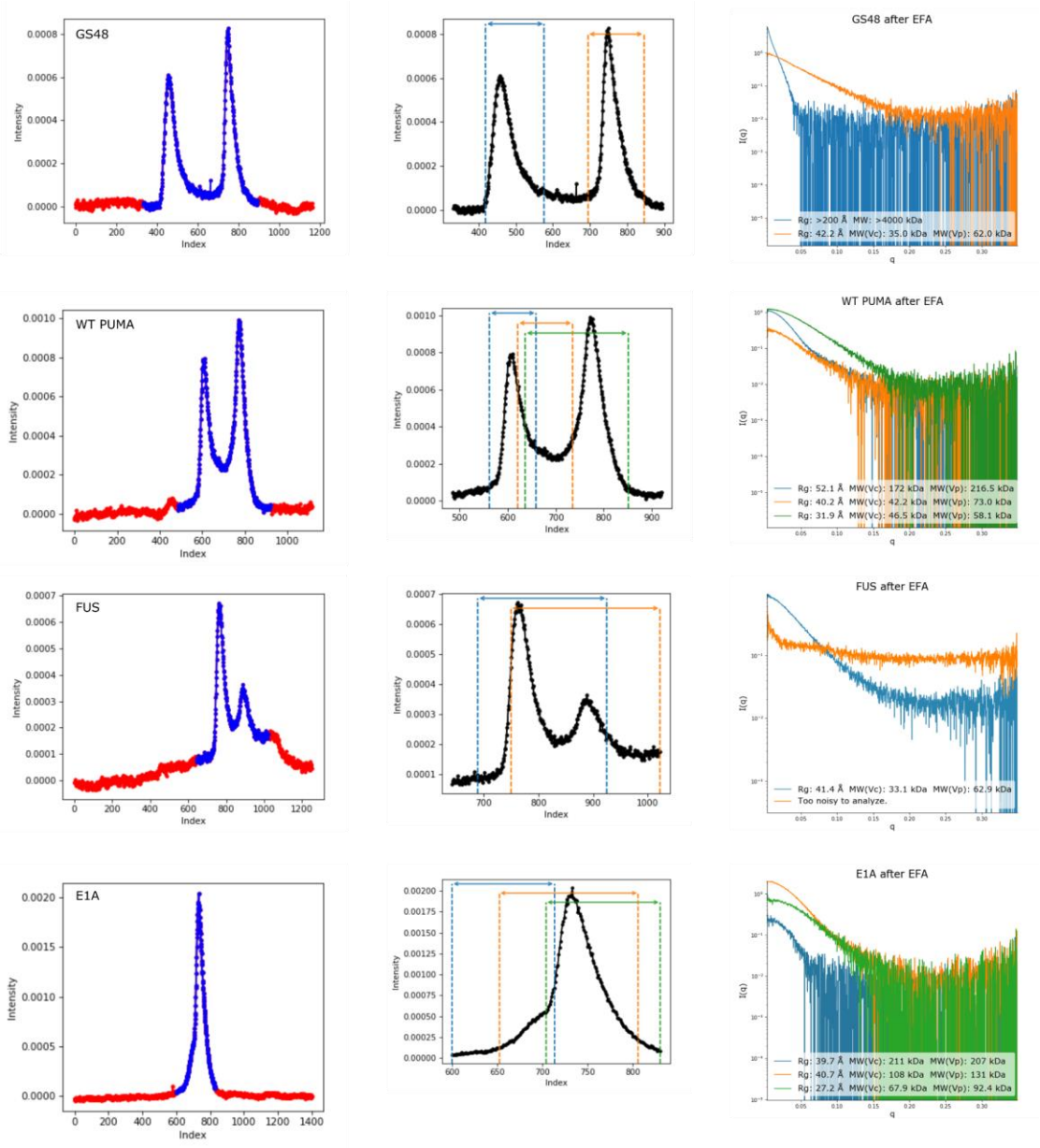

Figure A2.18. Screens from BioXTAS RAW software[193] showing process of deconvolution of SEC peaks using evolving factor analysis. Left: raw chromatograms. Center: ranges of deconvoluted peaks. Right: $I(q)$ *vs.* q series, calculated radius of gyration, and calculated molecular weight for each deconvoluted peak. Same colors in center and right panels represent the same deconvoluted peaks.

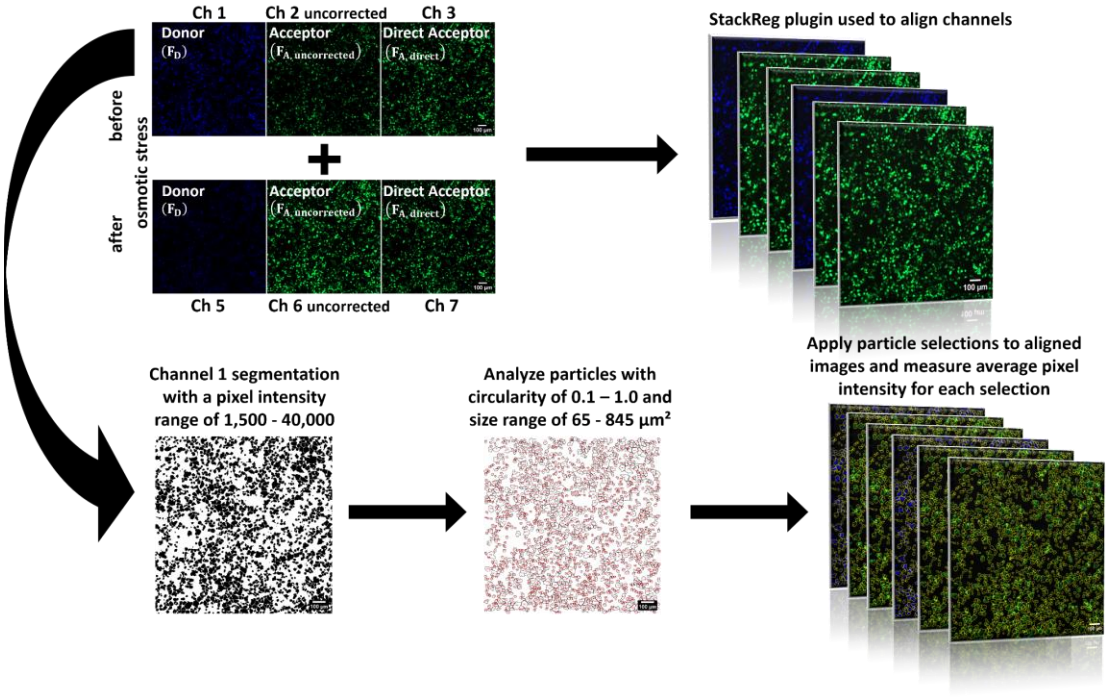

Figure A2.19. Analysis pipeline for live cell data. The donor channel before perturbation (Ch 1) was segmented using a fixed threshold to include any pixels with an intensity value between 1,500 - 40,000. The ImageJ “analyze particles” algorithm was used to select thresholded regions with a circularity 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. A2.21.**

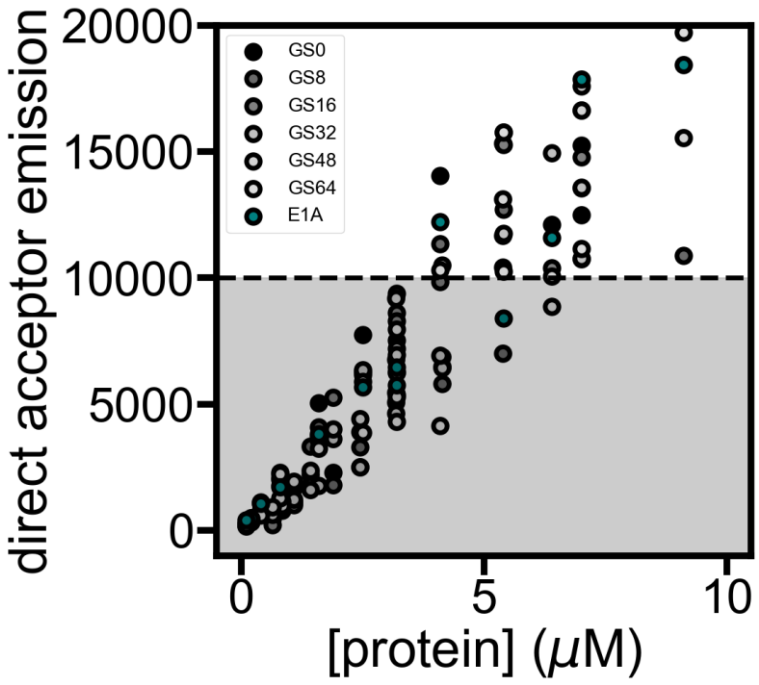

Figure A2.20. *In vitro* measurement of direct acceptor emission for known recombinant, purified proteins measured on the same setup as the live cells. Dashed line shows the emission cutoff used to select cells with a concentration range around 5 μM or lower to correlate with *in vitro* experiments.

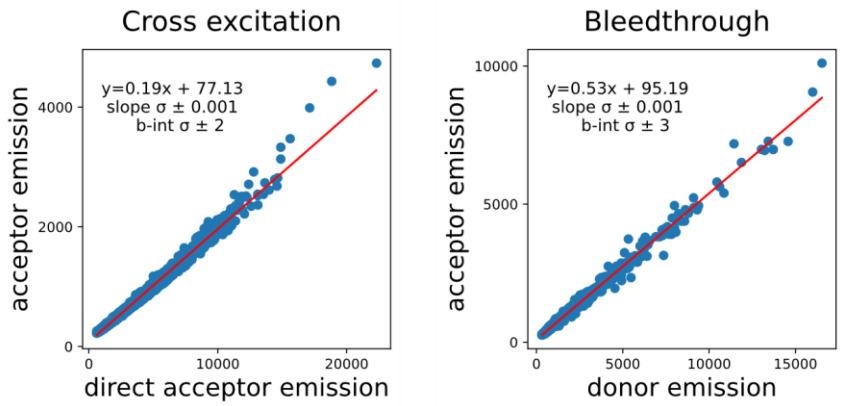

Figure A2.21. Measurements of cross-excitation (left) and bleedthrough (right) from donor to acceptor channel. To calculate cross-excitation, cells expressing mNeonGreen only were imaged. To calculate bleedthrough, cells expressing mTurquoise2 only were imaged. In both cases, the same imaging settings 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:

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

where F_A is used to calculate E_f^{cell}. The numbers 0.19 ± 0.001 and 0.53 ± 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).

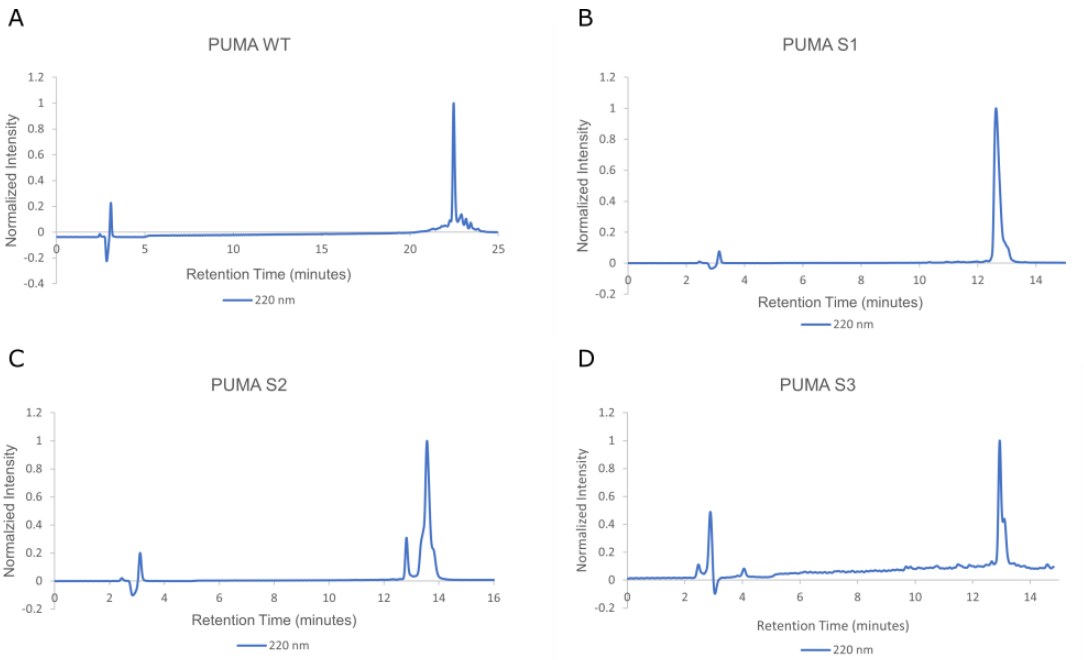

Figure A2.22. HPLC traces from purification of label-free peptides. **(A)** PUMA WT. **(B)** PUMA S1. **(C)** PUMA S2. **(D)** PUMA S3.

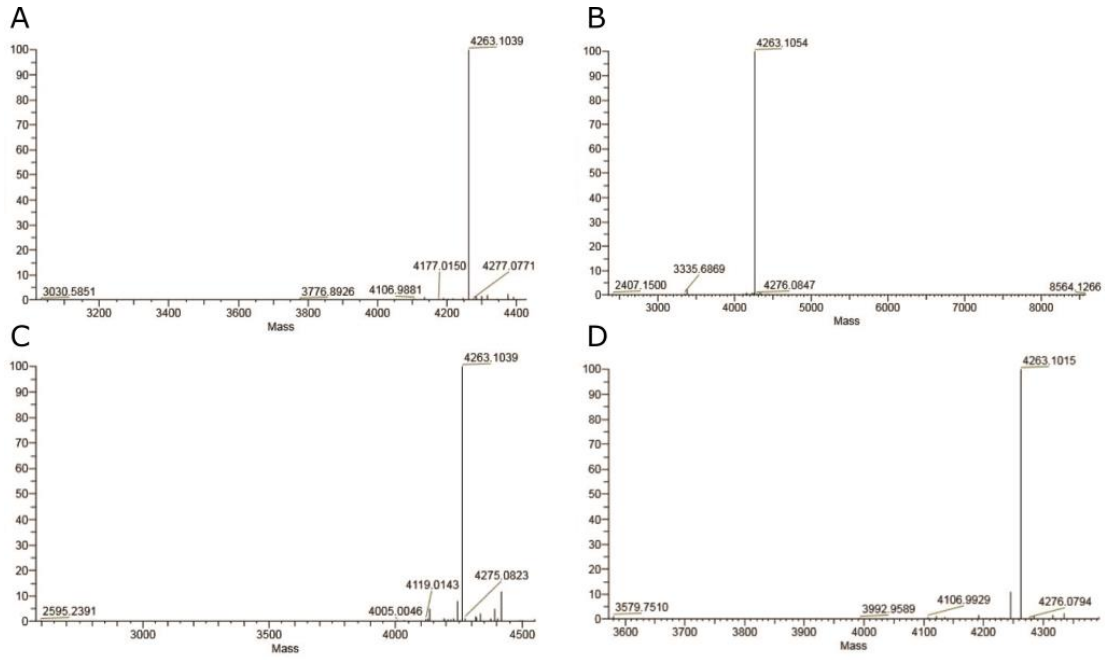

Figure A2.23. High-resolution ESI mass spectra of purified label-free peptides. (A) PUMA WT. (B) PUMA S1. (C) PUMA S2. (D) PUMA S3. Calculated and experimental masses are shown in **Table A2.2**.

Supplementary Tables

Table A2.1. IDR sequences used in this study.

IDR	Number of residues	Amino acid sequence
GS0	0	
GS8	16	GSGSGSGSGSGSGSGS
GS16	32	GSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGS
GS24	48	GSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGS
GS32	64	GSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGS
GS48	96	GSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGSGS
WT PUMA	34	VEEEEWAREIGAQLRRIADDLNAQYERRRQEEQH
PUMA S1	34	ELARQEERGIAVHYARQEQWANQLERERDEIERD
PUMA S2	34	LEWLERRRQEEVAGQEYIRDNRAEQDRIAAEEQH
PUMA S3	34	IHRAIDDQYERQLERQARLVWEEEGERERNEQAA
p53	61	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPD
Ash1	83	GASASSSPSPSTPTKSGKMRSRSSSPVRPKAYTPSPRSPNYHRFALDSPPQSPRRSSNSSITKKGSRRSSGSSPTRHTTRVCV
E1A	40	MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEP
FUS	163	MASNDYTQQATQSYGAYPTQPGQGYSQQSSQPYGQQSYSGYSQSTDTSGYGQSSYSSYGQSQNTGYGTQSTPQGYGSTGGYGSSQSSQSSYGQQSSYPGYGQQPAPSSTSGSYGSSSQSSSYGQPQSGSYSQQPSYGGQQQSYGQQQSYNPPQGYGQQNQYNS

Table A2.2. Calculated and experimental masses of label-free peptides.

Peptide	Calculated Mass (Da)	Experimental Mass from ESI-MS (Da)	Retention Time (minutes)
PUMA WT	4263.0918	4263.0974	22.47
PUMA S1	4263.0918	4263.1014	12.64
PUMA S2	4263.0918	4263.0923	13.59
PUMA S3	4263.0918	4263.0948	12.94

Appendix 3

A3.1 Materials and Methods

A3.1.1 Protein purification

2 L of LB supplemented with ampicillin (100 µg/mL) was inoculated with 20 mL/L of a saturated culture of cells containing pDEST-HisMBP-SED1. The culture was grown at 37°C, shaking at 225 rpm, to OD600 ~ 0.6. At this point, recombinant protein was induced with 1 mM IPTG (BIOSYNTH International) and the culture was transferred to 16°C and 250 rpm for 20 hours. Cells were collected and lysed in lysis buffer (50 mM NaH_2PO_4, pH 8, 0.5 M NaCl) by sonication on ice using a Q700 sonicator (Qsonica), and the extract was clarified by centrifugation. The His-tagged recombinant fusion protein was separated by affinity chromatography using Ni-NTA beads (Qiagen) and eluted with 50 mM NaH_2PO_4, pH 8, 0.5 M NaCl, 250 mM imidazole. To remove the His-MBP tag, the recombinant protein was incubated at 4°C overnight with TEV protease. Tag-free recombinant SED1 was separated by size-exclusion chromatography in an ÄKTA go purification system (Cytiva) in 20 mM sodium phosphate, pH 7.4, 100 mM NaCl. The purity of recombinant SED1 was confirmed by SDS-PAGE. The same strategy was followed for pDEST-HisMBP-CS.

A3.1.2 Solution preparation and specifics

Solutes were purchased from Alfa Aesar (Sarcosine, PEG200, PEG2000, PEG6000), VWR (D-Sorbitol), GE Healthcare (Ficoll), TCI (D-(+)-Trehalose Dihydrate, Trimethylamine N-Oxide Dihydrate (TMAO)), Thermo Scientific (Guanidine Hydrochloride), Acros Organics (Betaine Monohydrate), and Fisher BioReagents (Ethylene Glycol, Glycerol, Glycine, Magnesium Chloride Hexahydrate, Potassium Chloride, Sodium Chloride, Sucrose, 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 initially free of additional salt. The same buffer was used for all dilutions.

A3.1.3 Fluorescence analysis of purified recombinant proteins

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 0.8 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 construct, 24 replicates were performed in neat buffer containing NaCl, 12 replicates were performed in neat buffer not containing NaCl, and two repeats were performed in every other solution condition. Fluorescence spectra were obtained for each construct in each solution condition by exciting the sample in a 10-nm band centered at 433 nm, with a dichroic at 446.5 nm, and measuring fluorescence emission from 460 to 600 nm, averaging over a 10 nm window moved at intervals of 1 nm. The ratio of acceptor to donor intensity (DxAm/DxDm) was calculated by dividing the total measured fluorescence intensities from 500 to 600 nm by the total measured fluorescence intensities from 460 to 499 nm.

A3.1.4 All-atom simulations

Simulations of AtLEA4-5 protein, its scrambles, and other IDRs were done using solution space scanning[20], an all-atom Monte Carlo simulation method based on the ABSINTH force field[83,177] that has been previously described[20]. Briefly, the Hamitonian 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 $5x10^7$ Monte Carlo steps (following $1x10^7$ 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. 4.1B** was obtained using the same methods, is publicly available on https://github.com/sukeniklab/HiddenSensitivity, and has been previously described[19]. We analyzed the average radii of gyration of the simulated conformation ensembles using the MDTraj python library [178]. 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.

A3.1.5 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% CO_2 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.

A3.1.6 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% CO_2. 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 hours) 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 minutes at 37°C and 5% CO_2. After the incubation period, the cells were rinsed twice with PBS and 200 µL of PBS was added.

A3.1.7 U-2 OS fluorescence microscopy

Imaging was done on a Zeiss epifluorescent microscope using a 40X 0.9 NA dry objective. Excitation was done with a Colibri LED excitation module and data was collected on dual Hamamatsu Flash v3 sCMOS cameras. The cells were imaged at room temperature before and less than 1 minute 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.

A3.1.8 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.

References

1. van der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
2. 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.* **32**, 102–112 (2015).
3. Mao, A. H., Lyle, N. & Pappu, R. V. Describing sequence–ensemble relationships for intrinsically disordered proteins. *Biochem. J* **449**, 307–318 (2013).
4. Forman-Kay, J. D. & Mittag, T. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure* **21**, 1492–1499 (2013).
5. Mukhopadhyay, S., Krishnan, R., Lemke, E. A., Lindquist, S. & Deniz, A. A. A natively unfolded yeast prion monomer adopts an ensemble of collapsed and rapidly fluctuating structures. *Proceedings of the National Academy of Sciences* **104**, 2649–2654 (2007).
6. Bowman, M. A. *et al.* Properties of protein unfolded states suggest broad selection for expanded conformational ensembles. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 23356–23364 (2020).
7. Müller-Späth, S. *et al.* Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14609–14614 (2010).
8. Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L. & Pappu, R. V. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8183–8188 (2010).
9. Marsh, J. A. & Forman-Kay, J. D. Sequence Determinants of Compaction in Intrinsically Disordered Proteins. *Biophys. J.* **98**, 2383–2390 (2010).
10. Martin, E. W. *et al.* Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367**, 694–699 (2020).
11. Das, S., Lin, Y.-H., Vernon, R. M., Forman-Kay, J. D. & Chan, H. S. Comparative roles of charge, π, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 28795–28805 (2020).
12. Martin, E. W. *et al.* Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J. Am. Chem. Soc.* **138**, 15323–15335 (2016).
13. 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.* **110**, 13392–13397 (2013).

14. 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.* **113**, 5616–5621 (2016).
15. Beveridge, R. *et al.* Ion Mobility Mass Spectrometry Uncovers the Impact of the Patterning of Oppositely Charged Residues on the Conformational Distributions of Intrinsically Disordered Proteins. *J. Am. Chem. Soc.* **141**, 4908–4918 (2019).
16. Davey, N. E. The functional importance of structure in unstructured protein regions. *Curr. Opin. Struct. Biol.* **56**, 155–163 (2019).
17. Conicella, A. E., Zerze, G. H., Mittal, J. & Fawzi, N. L. ALS Mutations Disrupt Phase Separation Mediated by α-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain. *Structure* (2016) doi:10.1016/j.str.2016.07.007.
18. Borcherds, W. *et al.* Disorder and residual helicity alter p53-Mdm2 binding affinity and signaling in cells. *Nat. Chem. Biol.* **10**, 1000–1002 (2014).
19. Moses, D. *et al.* Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to their Chemical Environment. *J. Phys. Chem. Lett.* **11**, 10131–10136 (2020).
20. Holehouse, A. S. & Sukenik, S. Controlling Structural Bias in Intrinsically Disordered Proteins Using Solution Space Scanning. *J. Chem. Theory Comput.* **16**, 1794–1805 (2020).
21. Borgia, A. *et al.* Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *J. Am. Chem. Soc.* **138**, 11714–11726 (2016).
22. Hofmann, H. *et al.* Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 16155–16160 (2012).
23. Rogers, J. M. *et al.* Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 15420–15425 (2014).
24. Staller, M. V. *et al.* Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. *Cell Syst* **13**, 334-345.e5 (2022).
25. Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.* **33**, 2–8 (2008).
26. Tuttle, L. M. *et al.* Gcn4-Mediator Specificity Is Mediated by a Large and Dynamic Fuzzy Protein-Protein Complex. *Cell Rep.* **22**, 3251–3264 (2018).
27. Mateos, B. *et al.* Hyperphosphorylation of Human Osteopontin and Its Impact on Structural Dynamics and Molecular Recognition. *Biochemistry* (2021) doi:10.1021/acs.biochem.1c00050.

28. Zeno, W. F. *et al.* Molecular Mechanisms of Membrane Curvature Sensing by a Disordered Protein. *J. Am. Chem. Soc.* **141**, 10361–10371 (2019).
29. Keul, N. D. *et al.* The entropic force generated by intrinsically disordered segments tunes protein function. *Nature* **563**, 584–588 (2018).
30. 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).
31. Dyla, M. & Kjaergaard, M. Intrinsically disordered linkers control tethered kinases via effective concentration. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 21413–21419 (2020).
32. González-Foutel, N. S. *et al.* Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nat. Struct. Mol. Biol.* **29**, 781–790 (2022).
33. Cuevas-Velazquez, C. L. *et al.* Intrinsically disordered protein biosensor tracks the physical-chemical effects of osmotic stress on cells. *Nat. Commun.* **12**, 5438 (2021).
34. Moses, D. *et al.* Structural biases in disordered proteins are prevalent in the cell. *bioRxiv* (2021) doi:10.1101/2021.11.24.469609.
35. Soranno, A. *et al.* Integrated view of internal friction in unfolded proteins from single-molecule FRET, contact quenching, theory, and simulations. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1833–E1839 (2017).
36. Soranno, A. Physical basis of the disorder-order transition. *Arch. Biochem. Biophys.* **685**, 108305 (2020).
37. Martin, E. W. & Holehouse, A. S. Intrinsically disordered protein regions and phase separation: sequence determinants of assembly or lack thereof. *Emerg Top Life Sci* **4**, 307–329 (2020).
38. Elkjær, S. *et al.* Evolutionary fine-tuning of residual helix structure in disordered proteins manifests in complex structure and lifetime. *Commun Biol* **6**, 63 (2023).
39. 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.* **114**, 9882–9887 (2017).
40. 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**, 1012–1024 (2017).
41. Fuertes, G. *et al.* Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E6342–E6351 (2017).

42. Bennett, B. D. *et al.* Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. *Nat. Chem. Biol.* **5**, 593–599 (2009).
43. McManus, M. L., Churchwell, K. B. & Strange, K. Regulation of Cell Volume in Health and Disease. *New England Journal of Medicine* vol. 333 1260–1267 Preprint at https://doi.org/10.1056/nejm199511093331906 (1995).
44. Boron, W. F. Intracellular pH Regulation. *Physiology of Membrane Disorders* 423–435 Preprint at https://doi.org/10.1007/978-1-4613-2097-5_26 (1986).
45. Hoffmann, E. K., Lambert, I. H. & Pedersen, S. F. Physiology of Cell Volume Regulation in Vertebrates. *Physiological Reviews* vol. 89 193–277 Preprint at https://doi.org/10.1152/physrev.00037.2007 (2009).
46. Cook, B., Hardy, R. W., McConnaughey, W. B. & Zuker, C. S. Preserving cell shape under environmental stress. *Nature* **452**, 361–364 (2008).
47. Davis, C. M., Gruebele, M. & Sukenik, S. How does solvation in the cell affect protein folding and binding? *Curr. Opin. Struct. Biol.* **48**, 23–29 (2018).
48. Smoyer, C. J. & Jaspersen, S. L. Breaking down the wall: the nuclear envelope during mitosis. *Curr. Opin. Cell Biol.* **26**, 1–9 (2014).
49. Stewart, M. P. *et al.* Hydrostatic pressure and the actomyosin cortex drive mitotic cell rounding. *Nature* **469**, 226–230 (2011).
50. Brocker, C., Thompson, D. C. & Vasiliou, V. The role of hyperosmotic stress in inflammation and disease. *Biomol. Concepts* **3**, 345–364 (2012).
51. Son, S. *et al.* Resonant microchannel volume and mass measurements show that suspended cells swell during mitosis. *J. Cell Biol.* **211**, 757–763 (2015).
52. Guseman, A. J. & Pielak, G. J. Cosolute and Crowding Effects on a Side-By-Side Protein Dimer. *Biochemistry* **56**, 971–976 (2017).
53. Gorensek-Benitez, A. H., Smith, A. E., Stadmiller, S. S., Perez Goncalves, G. M. & Pielak, G. J. Cosolutes, Crowding, and Protein Folding Kinetics. *J. Phys. Chem. B* **121**, 6527–6537 (2017).
54. Sukenik, S., Salam, M., Wang, Y. & Gruebele, M. In-Cell Titration of Small Solutes Controls Protein Stability and Aggregation. *J. Am. Chem. Soc.* **140**, 10497–10503 (2018).
55. Sukenik, S., Ren, P. & Gruebele, M. Weak protein–protein interactions in live cells are quantified by cell-volume modulation. *Proceedings of the National Academy of Sciences* 201700818 Preprint at https://doi.org/10.1073/pnas.1700818114 (2017).
56. Minton, A. P. Excluded volume as a determinant of macromolecular structure and reactivity. *Biopolymers* vol. 20 2093–2120 Preprint at https://doi.org/10.1002/bip.1981.360201006 (1981).

57. Zimmerman, S. B. & Trach, S. O. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. *J. Mol. Biol.* **222**, 599–620 (1991).
58. Walter, H. & Brooks, D. E. Phase separation in cytoplasm, due to macromolecular crowding, is the basis for microcompartmentation. *FEBS Lett.* **361**, 135–139 (1995).
59. Soranno, A. *et al.* Single-molecule spectroscopy reveals polymer effects of disordered proteins in crowded environments. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 4874–4879 (2014).
60. Vancraenenbroeck, R., Harel, Y. S., Zheng, W. & Hofmann, H. Polymer effects modulate binding affinities in disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 19506–19512 (2019).
61. 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.* **117**, 13480–13489 (2020).
62. Riback, J. A. *et al.* Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. *Cell* **168**, 1028-1040.e19 (2017).
63. Ellis, R. J. Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Curr. Opin. Struct. Biol.* **11**, 114–119 (2001).
64. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).
65. 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.* **112**, 9614–9619 (2015).
66. Babu, M. M., van der Lee, R., de Groot, N. S. & Gsponer, J. Intrinsically disordered proteins: Regulation and disease. *Curr. Opin. Struct. Biol.* **21**, 432–440 (2011).
67. Tompa, P. The principle of conformational signaling. *Chem. Soc. Rev.* **45**, 4252–4284 (2016).
68. 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.* **114**, 1067–1079 (2018).
69. Hsu, P. P. & Sabatini, D. M. Cancer cell metabolism: Warburg and beyond. *Cell* **134**, 703–707 (2008).
70. Wu, D. & Zhou, H.-X. Designed Mutations Alter the Binding Pathways of an Intrinsically Disordered Protein. *Sci. Rep.* **9**, 6172 (2019).
71. Bah, A. & Forman-Kay, J. D. Modulation of intrinsically disordered protein function by post-translational modifications. *J. Biol. Chem.* **291**, 6696–6705 (2016).

72. Mansouri, A. L. *et al.* Folding propensity of intrinsically disordered proteins by osmotic stress. *Mol. Biosyst.* **12**, 3695–3701 (2016).
73. Goedhart, J. *et al.* Structure-guided evolution of cyan fluorescent proteins towards a quantum yield of 93%. *Nat. Commun.* **3**, 751 (2012).
74. Shaner, N. C. *et al.* A bright monomeric green fluorescent protein derived from Branchiostoma lanceolatum. *Nat. Methods* **10**, 407–409 (2013).
75. Mastop, M. *et al.* Characterization of a spectrally diverse set of fluorescent proteins as FRET acceptors for mTurquoise2. *Sci. Rep.* **7**, 11999 (2017).
76. Ferreon, A. C. M., Ferreon, J. C., Wright, P. E. & Deniz, A. A. Modulation of allostery by protein intrinsic disorder. *Nature* **498**, 390–394 (2013).
77. Sukenik, S., Sapir, L., Gilman-Politi, R. & Harries, D. Diversity in the mechanisms of cosolute action on biomolecular processes. *Faraday Discuss.* **160**, 225–37; discussion 311-27 (2013).
78. Senske, M. *et al.* Protein stabilization by macromolecular crowding through enthalpy rather than entropy. *J. Am. Chem. Soc.* **136**, 9036–9041 (2014).
79. 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).
80. Pegram, L. M. *et al.* Why Hofmeister effects of many salts favor protein folding but not DNA helix formation. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 7716–7721 (2010).
81. Sarkar, M., Li, C. & Pielak, G. J. Soft interactions and crowding. *Biophys. Rev.* **5**, 187–194 (2013).
82. Pegram, L. M. & Record, M. T., Jr. Thermodynamic origin of Hofmeister ion effects. *J. Phys. Chem. B* **112**, 9428–9436 (2008).
83. 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).
84. Piovesan, D. *et al.* DisProt 7.0: A major update of the database of disordered proteins. *Nucleic Acids Res.* **45**, D219–D227 (2017).
85. Liu, Y. & Bolen, D. W. The peptide backbone plays a dominant role in protein stabilization by naturally occurring osmolytes. *Biochemistry* **34**, 12884–12891 (1995).
86. Auton, M., Holthauzen, L. M. F. & Bolen, D. W. Anatomy of energetic changes accompanying urea-induced protein denaturation. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 15317–15322 (2007).
87. 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.* **105**, 13403–13408 (2008).

88. Harries, D. & Rösgen, J. A practical guide on how osmolytes modulate macromolecular properties. *Methods Cell Biol.* **84**, 679–735 (2008).
89. 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).
90. Riback, J. A. *et al.* Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **358**, 238–241 (2017).
91. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* **18**, 343–384 (2005).
92. Babu, M. M., Kriwacki, R. W. & Pappu, R. V. Structural biology. Versatility from protein disorder. *Science* **337**, 1460–1461 (2012).
93. Mittag, T., Kay, L. E. & Forman-Kay, J. D. Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recognit.* **23**, 105–116 (2010).
94. Conicella, A. E. *et al.* TDP-43 α-helical structure tunes liquid-liquid phase separation and function. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 5883–5894 (2020).
95. Bignon, C., Troilo, F., Gianni, S. & Longhi, S. Partner-Mediated Polymorphism of an Intrinsically Disordered Protein. *J. Mol. Biol.* **430**, 2493–2507 (2018).
96. Schrag, L. G. *et al.* Cancer-Associated Mutations Perturb the Disordered Ensemble and Interactions of the Intrinsically Disordered p53 Transactivation Domain. *J. Mol. Biol.* **433**, 167048 (2021).
97. Xu, H. *et al.* Cryptochrome 1 regulates the circadian clock through dynamic interactions with the BMAL1 C terminus. *Nat. Struct. Mol. Biol.* **22**, 476–484 (2015).
98. Panova, S. *et al.* Mapping Hidden Residual Structure within the Myc bHLH-LZ Domain Using Chemical Denaturant Titration. *Structure* **27**, 1537-1546.e4 (2019).
99. Ferrie, J. J., Karr, J. P., Tjian, R. & Darzacq, X. “Structure”-function relationships in eukaryotic transcription factors: The role of intrinsically disordered regions in gene regulation. *Mol. Cell* **82**, 3970–3984 (2022).
100. Wall, K. P. & Hough, L. E. In-Cell NMR within Budding Yeast Reveals Cytoplasmic Masking of Hydrophobic Residues of FG Repeats. *Biophys. J.* **115**, 1690–1695 (2018).
101. Theillet, F.-X. *et al.* Structural disorder of monomeric α-synuclein persists in mammalian cells. *Nature* **530**, 45–50 (2016).
102. König, I. *et al.* Single-molecule spectroscopy of protein conformational dynamics in live eukaryotic cells. *Nat. Methods* **12**, 773–779 (2015).

103. Rivas, G. & Minton, A. P. Macromolecular Crowding In Vitro, In Vivo, and In Between. *Trends Biochem. Sci.* **41**, 970–981 (2016).
104. Gnutt, D., Gao, M., Brylski, O., Heyden, M. & Ebbinghaus, S. Excluded-volume effects in living cells. *Angew. Chem. Int. Ed Engl.* **54**, 2548–2551 (2015).
105. Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* **324**, 1029–1033 (2009).
106. Monteith, W. B., Cohen, R. D., Smith, A. E., Guzman-Cisneros, E. & Pielak, G. J. Quinary structure modulates protein stability in cells. *Proceedings of the National Academy of Sciences* **112**, 1739–1742 (2015).
107. Gruebele, M., Dave, K. & Sukenik, S. Globular Protein Folding In Vitro and In Vivo. *Annu. Rev. Biophys.* **45**, 233–251 (2016).
108. Van Rosmalen, M., Krom, M. & Merkx, M. Tuning the Flexibility of Glycine-Serine Linkers to Allow Rational Design of Multidomain Proteins. *Biochemistry* **56**, 6565–6574 (2017).
109. Basak, S. *et al.* Probing Interdomain Linkers and Protein Supertertiary Structure In Vitro and in Live Cells with Fluorescent Protein Resonance Energy Transfer. *J. Mol. Biol.* **433**, 166793 (2021).
110. Sørensen, C. S. & Kjaergaard, M. Measuring Effective Concentrations Enforced by Intrinsically Disordered Linkers. *Methods Mol. Biol.* **2141**, 505–518 (2020).
111. Möglich, A., Joder, K. & Kiefhaber, T. End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12394–12399 (2006).
112. Evers, T. H., Van Dongen, E. M. W. M., Faesen, A. C., Meijer, E. W. & Merkx, M. Quantitative understanding of the energy transfer between fluorescent proteins connected via flexible peptide linkers. *Biochemistry* **45**, 13183–13192 (2006).
113. Ruff, K. M. & Holehouse, A. S. SAXS versus FRET: A Matter of Heterogeneity? *Biophysical journal* vol. 113 971–973 (2017).
114. 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* **125**, 6451–6478 (2021).
115. Gradinaru, C. C., Gomes, G. W., Mittag, T., Head-Gordon, T. L. & Forman-Kay, J. D. Integrating NMR, SAXS and Single-Molecule FRET Data to Infer Conformational Ensembles of the Yeast Sic1 Protein. *Biophysical Journal* vol. 120 30a Preprint at https://doi.org/10.1016/j.bpj.2020.11.436 (2021).

116. 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).
117. 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).
118. Wang, Y., Sukenik, S., Davis, C. M. & Gruebele, M. Cell Volume Controls Protein Stability and Compactness of the Unfolded State. *J. Phys. Chem. B* **122**, 11762–11770 (2018).
119. Rogers, J. M., Steward, A. & Clarke, J. Folding and binding of an intrinsically disordered protein: Fast, but not “diffusion-limited.” *J. Am. Chem. Soc.* **135**, 1415–1422 (2013).
120. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. G. & Pappu, R. V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys. J.* **112**, 16–21 (2017).
121. Uversky, V. N. Size-exclusion chromatography in structural analysis of intrinsically disordered proteins. *Methods Mol. Biol.* **896**, 179–194 (2012).
122. Harmon, T. S. *et al.* GADIS: Algorithm for designing sequences to achieve target secondary structure profiles of intrinsically disordered proteins. *Protein Eng. Des. Sel.* **29**, 339–346 (2016).
123. Burke, K. A., Janke, A. M., Rhine, C. L. & Fawzi, N. L. Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II. *Mol. Cell* **60**, 231–241 (2015).
124. Theillet, F.-X. *et al.* Physicochemical properties of cells and their effects on intrinsically disordered proteins (IDPs). *Chem. Rev.* **114**, 6661–6714 (2014).
125. Patel, A. *et al.* A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* **162**, 1066–1077 (2015).
126. Kato, M. *et al.* Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **149**, 753–767 (2012).
127. Tsvetkov, P., Reuven, N. & Shaul, Y. The nanny model for IDPs. *Nat. Chem. Biol.* **5**, 778–781 (2009).
128. Taneja, I. & Holehouse, A. S. Folded domain charge properties influence the conformational behavior of disordered tails. *Curr Res Struct Biol* **3**, 216–228 (2021).
129. 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) doi:10.1016/j.jmb.2018.05.012.

130. Martin, E. W. *et al.* Interplay of folded domains and the disordered low-complexity domain in mediating hnRNPA1 phase separation. *Nucleic Acids Res.* **49**, 2931–2945 (2021).
131. Zheng, T., Galagedera, S. K. K. & Castañeda, C. A. Previously uncharacterized interactions between the folded and intrinsically disordered domains impart asymmetric effects on UBQLN2 phase separation. *Protein Sci.* **30**, 1467–1481 (2021).
132. Zamboni, N., Saghatelian, A. & Patti, G. J. Defining the metabolome: size, flux, and regulation. *Mol. Cell* **58**, 699–706 (2015).
133. Son, S. *et al.* Direct observation of mammalian cell growth and size regulation. *Nat. Methods* **9**, 910–912 (2012).
134. Kim, D.-H. *et al.* Volume regulation and shape bifurcation in the cell nucleus. *J. Cell Sci.* **128**, 3375–3385 (2015).
135. Yoo, H., Triandafillou, C. & Drummond, D. A. Cellular sensing by phase separation: Using the process, not just the products. *J. Biol. Chem.* **294**, jbc.TM118.001191 (2019).
136. Murthy, A. C. *et al.* Molecular interactions underlying liquid-liquid phase separation of the FUS low-complexity domain. *Nat. Struct. Mol. Biol.* **26**, 637–648 (2019).
137. 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.* **291**, 10893–10903 (2016).
138. Wuttke, R. *et al.* Temperature-dependent solvation modulates the dimensions of disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 5213–5218 (2014).
139. Ortmayr, K., Dubuis, S. & Zampieri, M. Metabolic profiling of cancer cells reveals genome-wide crosstalk between transcriptional regulators and metabolism. *Nat. Commun.* **10**, 1841 (2019).
140. Sanchez, E. L. & Lagunoff, M. Viral activation of cellular metabolism. *Virology* **479–480**, 609–618 (2015).
141. Li, Y., Konstantopoulos, K., Zhao, R., Mori, Y. & Sun, S. X. The importance of water and hydraulic pressure in cell dynamics. *J. Cell Sci.* **133**, (2020).
142. Yancey, P. H. Organic osmolytes as compatible, metabolic and counteracting cytoprotectants in high osmolarity and other stresses. *J. Exp. Biol.* **208**, 2819–2830 (2005).
143. Hosseiniyan Khatibi, S. M. *et al.* Osmolytes resist against harsh osmolarity: Something old something new. *Biochimie* **158**, 156–164 (2019).

144. Record, M. T., Jr, Courtenay, E. S., Cayley, D. S. & Guttman, H. J. Responses of E. coli to osmotic stress: large changes in amounts of cytoplasmic solutes and water. *Trends Biochem. Sci.* **23**, 143–148 (1998).
145. Spitzer, J. & Poolman, B. The role of biomacromolecular crowding, ionic strength, and physicochemical gradients in the complexities of life's emergence. *Microbiol. Mol. Biol. Rev.* **73**, 371–388 (2009).
146. Persson, L. B., Ambati, V. S. & Brandman, O. Cellular Control of Viscosity Counters Changes in Temperature and Energy Availability. *Cell* **0**, (2020).
147. Haswell, E. S. & Verslues, P. E. The ongoing search for the molecular basis of plant osmosensing. *J. Gen. Physiol.* **145**, 389–394 (2015).
148. Bourque, C. W. Central mechanisms of osmosensation and systemic osmoregulation. *Nat. Rev. Neurosci.* **9**, 519–531 (2008).
149. Nongpiur, R. C., Singla-Pareek, S. L. & Pareek, A. The quest for osmosensors in plants. *J. Exp. Bot.* **71**, 595–607 (2020).
150. Scharwies, J. D. & Dinneny, J. R. Water transport, perception, and response in plants. *J. Plant Res.* **132**, 311–324 (2019).
151. Zhou, X., Mehta, S. & Zhang, J. Genetically Encodable Fluorescent and Bioluminescent Biosensors Light Up Signaling Networks. *Trends Biochem. Sci.* **45**, 889–905 (2020).
152. Okumoto, S., Jones, A. & Frommer, W. B. Quantitative imaging with fluorescent biosensors. *Annu. Rev. Plant Biol.* **63**, 663–706 (2012).
153. Greenwald, E. C., Mehta, S. & Zhang, J. Genetically Encoded Fluorescent Biosensors Illuminate the Spatiotemporal Regulation of Signaling Networks. *Chem. Rev.* **118**, 11707–11794 (2018).
154. Pittas, T., Zuo, W. & Boersma, A. J. Engineering crowding sensitivity into protein linkers. in *Methods in Enzymology* (Academic Press, 2020).
155. Liu, B., Poolman, B. & Boersma, A. J. Ionic Strength Sensing in Living Cells. *ACS Chem. Biol.* **12**, 2510–2514 (2017).
156. Mahon, M. J. pHluorin2: an enhanced, ratiometric, pH-sensitive green florescent protein. *Adv. Biosci. Biotechnol.* **2**, 132–137 (2011).
157. Boersma, A. J., Zuhorn, I. S. & Poolman, B. A sensor for quantification of macromolecular crowding in living cells. *Nat. Methods* **12**, 227–9, 1 p following 229 (2015).
158. Cuevas-Velazquez, C. L. & Dinneny, J. R. Organization out of disorder: liquid–liquid phase separation in plants. *Curr. Opin. Plant Biol.* (2018).
159. 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.* **154**, 373–390 (2010).

160. Kaper, T., Lager, I., Looger, L. L., Chermak, D. & Frommer, W. B. Fluorescence resonance energy transfer sensors for quantitative monitoring of pentose and disaccharide accumulation in bacteria. *Biotechnol. Biofuels* **1**, 11 (2008).
161. Mika, J. T. & Poolman, B. Macromolecule diffusion and confinement in prokaryotic cells. *Curr. Opin. Biotechnol.* **22**, 117–126 (2011).
162. Heo, J., Meng, F. & Hua, S. Z. Contribution of aquaporins to cellular water transport observed by a microfluidic cell volume sensor. *Anal. Chem.* **80**, 6974–6980 (2008).
163. Jalihal, A. P. *et al.* Multivalent Proteins Rapidly and Reversibly Phase-Separate upon Osmotic Cell Volume Change. *Mol. Cell* **79**, 978-990.e5 (2020).
164. Moses, D., Ginell, G. M., Holehouse, A. S. & Sukenik, S. Intrinsically disordered regions are poised to act as sensors of cellular chemistry. *Trends Biochem. Sci.* (2023) doi:10.1016/j.tibs.2023.08.001.
165. Lee, H.-J. *et al.* Proteomic and Metabolomic Characterization of a Mammalian Cellular Transition from Quiescence to Proliferation. *Cell Rep.* **20**, 721–736 (2017).
166. How Genetic Changes Lead to Cancer. *National Cancer Institute* https://www.cancer.gov/about-cancer/causes-prevention/genetics/genetic-changes-infographic.
167. Reznik, E. *et al.* A Landscape of Metabolic Variation across Tumor Types. *Cell Syst* **6**, 301-313.e3 (2018).
168. Pesce, F. *et al.* Design of intrinsically disordered protein variants with diverse structural properties. *bioRxiv* 2023.10.22.563461 (2023) doi:10.1101/2023.10.22.563461.
169. Strome, B., Elemam, K., Pritisanac, I., Forman-Kay, J. D. & Moses, A. M. Computational design of intrinsically disordered protein regions by matching bulk molecular properties. *bioRxiv* 2023.04.28.538739 (2023) doi:10.1101/2023.04.28.538739.
170. Lasker, K. *et al.* The material properties of a bacterial-derived biomolecular condensate tune biological function in natural and synthetic systems. *Nat. Commun.* **13**, 5643 (2022).
171. Cranfill, P. J. *et al.* Quantitative assessment of fluorescent proteins. *Nat. Methods* **13**, 557–562 (2016).
172. Lambert, T. J. FPbase: a community-editable fluorescent protein database. *Nat. Methods* **16**, 277–278 (2019).
173. Rubinstein, M. & Colby, R. H. *Polymer Physics*. (Oxford University Press, 2003).

174. 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).
175. 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).
176. Devanand, K. & Selser, J. C. Asymptotic behavior and long-range interactions in aqueous solutions of poly(ethylene oxide). *Macromolecules* **24**, 5943–5947 (1991).
177. 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).
178. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
179. 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).
180. 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).
181. Zheng, W. *et al.* Inferring properties of disordered chains from FRET transfer efficiencies. *J. Chem. Phys.* **148**, 123329 (2018).
182. Fuertes, G. *et al.* Comment on “Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water.” *Science* **361**, (2018).
183. Riback, J. A. *et al.* Response to Comment on “Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water.” *Science* vol. 361 (2018).
184. 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).
185. 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).
186. 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).

187. 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).
188. 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).
189. de Gennes, P. G. *Scaling concepts in polymer physics*. (Cornell University Press, 1979).
190. Wlodarczyk, J. *et al.* Analysis of FRET signals in the presence of free donors and acceptors. *Biophys. J.* **94**, 986–1000 (2008).
191. Nielsen, S. S. *et al.* BioXTAS RAW, a software program for high-throughput automated small-angle X-ray scattering data reduction and preliminary analysis. *Journal of Applied Crystallography* vol. 42 959–964 Preprint at https://doi.org/10.1107/s0021889809023863 (2009).
192. Hopkins, J. B., Gillilan, R. E. & Skou, S. : improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. *J. Appl. Crystallogr.* **50**, 1545–1553 (2017).
193. Hopkins, J. B., Gillilan, R. E. & Skou, S. BioXTAS RAW: a free open-source program for small-angle X-ray scattering data reduction. *Foundations of Crystallography* **74**, a219 (2018).
194. Maeder, M. & Zilian, A. Evolving factor analysis, a new multivariate technique in chromatography. *Chemometrics and Intelligent Laboratory Systems* vol. 3 205–213 Preprint at https://doi.org/10.1016/0169-7439(88)80051-0 (1988).
195. Meisburger, S. P. *et al.* Domain Movements upon Activation of Phenylalanine Hydroxylase Characterized by Crystallography and Chromatography-Coupled Small-Angle X-ray Scattering. *J. Am. Chem. Soc.* **138**, 6506–6516 (2016).
196. Martin, E. W., Hopkins, J. B. & Mittag, T. Small-angle X-ray scattering experiments of monodisperse intrinsically disordered protein samples close to the solubility limit. *Methods Enzymol.* **646**, 185–222 (2021).
197. Balleza, E., Kim, J. M. & Cluzel, P. Systematic characterization of maturation time of fluorescent proteins in living cells. *Nat. Methods* **15**, 47–51 (2018).
198. Abramoff, M. D., Magelhaes, P. J. & Ram, S. J. Image Processing with ImageJ. *Biophotonics International* **11**, 36–42 (2004).
199. Thevenaz, P., Ruttimann, U. E. & Unser, M. A pyramid approach to subpixel registration based on intensity. *IEEE Trans. Image Process.* **7**, 27–41 (1998).

200. Clavel, D. *et al.* Structural analysis of the bright monomeric yellow-green fluorescent protein mNeonGreen obtained by directed evolution. *Acta Crystallogr D Struct Biol* **72**, 1298–1307 (2016).
201. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **105**, 962–974 (2013).
202. Rambo, R. P. & Tainer, J. A. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature* **496**, 477–481 (2013).
203. Lalmansingh, J. M., Keeley, A. T., Ruff, K. M., Pappu, R. V. & Holehouse, A. S. SOURSOP: A Python package for the analysis of simulations of intrinsically disordered proteins. *bioRxiv* (2023) doi:10.1101/2023.02.16.528879.
204. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. G. & Pappu, R. V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys. J.* **112**, 16–21 (2017).
205. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10037–10041 (2001).
206. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
207. Goujon, M. *et al.* A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.* **38**, W695-9 (2010).
208. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. G. & Pappu, R. V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys. J.* **112**, 16–21 (2017).
209. Vancraenenbroeck, R., Harel, Y. S., Zheng, W. & Hofmann, H. Polymer effects modulate binding affinities in disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 19506–19512 (2019).