SUPPLEMENTARY INFORMATION:

Sequence-ensemble-function relationships for disordered proteins in live cells

Ryan J. Emenecker[1,2,*], Karina Guadalupe[3,4,*], Nora M. Shamoon[4,5], Shahar Sukenik[3,4,5,6,✉], Alex S. Holehouse[1,2,✉]

1. Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO
2. Center for Biomolecular Condensates (CBC), Washington University in St. Louis, St. Louis, MO
3. Department of Chemistry and Biochemistry, University of California, Merced, CA
4. Center for Cellular and Biomolecular Machines, University of California, Merced, CA
5. Quantitative Systems Biology Program, University of California, Merced, CA
6. Health Sciences Research Institute, University of California, Merced, CA

* These authors contributed equally to the work

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

ONLINE METHODS

Fluorescence reporter constructs

Our fluorescence reporter construct places the disordered protein sequences from our library (**Table S1**) between an N-terminal mTurquoise2 FRET donor and a C-terminal mNeonGreen acceptor. Genes for each IDR were obtained from GenScript and ligated between the two fluorescent proteins using 5' SacI and 3' HindIII restriction sites in a pcDNA3.1(+) backbone, as described previously[1].

Mammalian Cell culture

U2-OS cells were cultured in Corning-treated flasks with Dulbecco's modified Eagle medium (Gibco Advanced DMEM:F12 1X) supplemented with 10% FBS (Gibco) and 1% penicillin/streptomycin (Gibco). For live-cell microscopy experiments, 8,000 cells were plated in a µ-Plate 96 Well Black treated imaging plate (Ibidi) 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) was used to transfect FRET construct plasmids into U2-OS cells per the manufacturer's protocol. Cells were incubated at 37°C and 5% CO_2 for 48 hours. NaCl stock solution of 5 M was prepared by dissolving the corresponding amount of NaCl (Fisher Bioreagents) in 1X PBS (Gibco) and filtering using a 0.2 µm filter. The solutions used for perturbations were obtained by diluting 1X PBS with autoclaved DI water to achieve hypoosmotic conditions or by adding NaCl stock solution to achieve hyperosmotic conditions.

Live-cell Microscopy

Imaging was done on a Zeiss epifluorescent microscope using a 10X 0.3 NA dry objective. Excitation was done with a Colibri LED excitation module, and data was collected on a duocam setup with two linked Hamamatsu flash v3 sCMOS cameras. The cells were imaged at room temperature before and after perturbation with 150 ms exposure times. Imaging was done by exciting mTurquoise2 at 430 nm (donor and acceptor channels) or mNeonGreen at 511 nm (direct acceptor channel). Emitted light was passed on to the camera using a triple bandpass dichroic (467/24, 555/25, 687/145). When measuring FRET, emitted light was split into two channels using a downstream beamsplitter with a 520 nm cutoff. For each perturbation, the cells were focused using the acceptor channel and imaged before manually adding water (hypoosmotic conditions), PBS (isosmotic condition), or NaCl solution (hyperosmotic conditions) with a pipette and pipetting up and down 10 times to ensure mixing. The final osmolarities that were used for the perturbations were: 100 mOsm (hypo-osomotic), 300 mOsm (iso-osmotic), and 750 mOsm (hyper-osmotic), with NaCl as the osmotic agent. Imaging was typically completed within ~30 seconds of osmotic change. Cells used for localization measurements were imaged using a 20X 0.8 dry objective.

Image Analysis

Images were analyzed using ImageJ[2]. Images collected before and after osmotic challenge, containing three channels each, were stacked and aligned using the StackReg plugin with rigid transformation. The aligned image was segmented based on the donor channel before perturbation. Segmentation was done using a fixed threshold that selected only pixels with an intensity of 1,500 - 40,000. The resulting mask was corrected using the Open and Watershed binary algorithms. Cells were selected using the Analyze Particles option of ImageJ, selecting only those that were 100-2,000 μm² in size and with a circularity of 0.1 to 0.8. The resulting ROIs were averaged in each channel at each time point. Bleedthrough and cross-excitation corrections were the same as described previously[1]. All constructs displayed similar average cell properties (**Fig. S15).** Cell FRET efficiency before and after perturbation ($E^{cell}_{f,\,before}$ and $E^{cell}_{f,\,after}$ respectively) was calculated by $E^{cell}_{f} = \frac{F_A}{F_D + F_A}$. Here F_D is the fluorescence of the donor and F_A is the fluorescence of the acceptor following bleedthrough and cross-excitation corrections. Localization measurements were obtained as described previously[1]. The acceptor emission under acceptor excitation was used as a proxy to measure protein concentration. The localization ratio was presented as $log_2\left(\frac{nucleus}{cytoplasm}\right)$.

GS-linker reference

As an internal standard, we also used a glycine-serine repeat linker, $(GS)_{32}$ (red line in **Fig. S3A**). Previous work by us and by others have shown that GS in vitro behaves as a Gaussian chain[1,3]. All $(GS)_{32}$ measurements used for comparisons are shown in **Fig. S16.**

Statistical Analysis

The statistical analysis for all of the experimental data was performed using the SciPy library in Python[4]. Experiments were done on 96-well plates, across multiple cell passage numbers and multiple days, and each well was plated and transfected individually. We therefore considered each well a biological repeat of the experiment. Therefore, the median E_f and ΔE_f values for all of the cells measured per well were used to generate a single violin plot (**Fig. 1F**, **S17**). We excluded wells that contained under 60 cells. The standard deviation and average values were calculated from the medians of all wells from each experimental condition (**Fig. 1F**, **S17**). To assess significance of the differences between two constructs, a double-sided Student's t-test was performed between all medians of the two constructs.

Correlation analysis for live cell imaging

Correlation between sequence parameters and ΔE_{FRET} on hyperosmotic shock and hyperosmotic shock (**Fig. S5**) involves sequences where sufficient statistics exist to assess changes in FRET efficiency. Specifically for changes upon hyperosmotic shock (**Fig. 4G**, **Fig. S10**), this means 6/32 sequences were excluded (6, 7, 10, 19, 23, 24). For changes upon hypo-osmotic shock (**Fig. S10**), this means 4/32 (10,19,23,24) were excluded. For correlations with radii of gyration (R_g) and end-to-end distances (R_e) from coarse-grained simulations (**Fig. 4G**, **Fig. S10**), six highly charged sequences [9,10, 17, 18, 24, 32] were excluded.

Limitations, drawbacks, and caveats of live cell imaging experiments

As with any study, our work is not without limitations, drawbacks, and caveats.

A potential critique of our work is the size of our library. At 32 sequences, the number of unique sequences we have compared here is much smaller than alternative approaches that leverage fluorescence-activated cell sorting (FACS) and/or sequencing-based readouts for assay sequence-function relationships. While this is true, a major confounding factor in screen-based experimental setups such as ours is sequence-dependent changes in expression, mRNA stability, protein degradation, and subcellular localization. Our live-cell approach, while medium throughput, allows us to systematically and rigorously assess all these factors and ensure our conclusions are based on protein-dependent effects corrected for abundance and subcellular localization.

While we interpret our positively charged sequences as interacting with intracellular polyanions, we are unable at this stage to identify the specific identities of what these anions may be. Based on prior work, we anticipate these anions to be RNA[5]. Future work – likely mass spectrometry-based – will be required to elucidate the specific components that engage with synthetic IDRs. This is an area of active interest and ongoing work. We also note that pioneering experimental[3] and computational[6] work has shown lysine vs. arginine in IDRs leads to distinctive ensemble properties, yet our work here does not directly compare that difference and largely maintains a relatively consistent argenine vs. lysine ratio. This again would be an area of interest moving forward.

One alternative explanation for why positively charged IDRs are more compact is in the experimental setup. Although using FPs in our FRET assay allows for rapid characterization of ensemble dimensions in live cells, the presence of N- and C-terminal folded domains could perturb IDR dimensions compared to IDRs without adjacent FPs. We reason that, at least for some sequences, the FP:IDR interaction could overrule the intra-IDR interactions in determining ensemble dimensions. To minimize the probability of this impacting our overall results, we designed our experiments and analysis to focus on pairs or triplets of sequences with similar features. That said, the majority (~95%) of IDRs are found directly adjacent to folded domains, such that even if FP-mediated interactions influence our trends, that perturbation is biologically relevant[7,8].

Another alternative explanation is that the residual secondary structure in our IDRs underlies some of the behavior observed. However, our computational analysis provides no strong evidence of this (**Fig. S12**) for the overall basal FRET efficiencies or distinct response profiles to hyper or hypo-osmotic shock.

Finally, whether or not our conclusions here hold across all cell types remains an open question. Our focus here on U2OS cells reflects their convenience for imaging and broad use in biomedical research. While we anticipate the general conclusions drawn here to hold in different cell types, this should be explicitly tested.

GOOSE: a software package for the design of disordered sequences

The sequences used in this manuscript were designed using the Python (version 3.7+) package GOOSE (https://github.com/idptools/goose). GOOSE (Generate disOrdered prOtiens Specifying propErties) continues our goal of pushing the frontiers of acronym technology but also implements a novel software package developed as part of this manuscript for the rational design of intrinsically disordered protein regions with bespoke sequence properties.

GOOSE uses sparrow (https://github.com/idptools/sparrow/) to calculate sequence properties. Ensemble predictions used for the design of IDRs with a desired radius of gyration or end-to-end distance use ALBATROSS, as implemented in sparrow[9]. ALBATROSS is a deep-learning tool for predicting ensemble-average IDR dimensions directly from sequence and was parameterized based on coarse-grained simulations performed with a modified variant of the Mpipi model[10].

GOOSE enables the design of fully synthetic sequences based on requested design constraints, as well as systematic perturbations to existing sequence variants. In this way, GOOSE is poised to facilitate the rational design of small numbers of sequences but can also be used to create libraries of thousands of sequences for systematic investigation of sequence-ensemble and sequence-function properties. A key feature of GOOSE is that it takes advantage of the metapredict (V2-FF) backend to ensure rapid and accurate assessment of disorder propensity for designed sequences. Developing a fast and accurate disorder predictor (1000s seconds/sequence with state-of-the-art accuracy[11]) was essential to enable

high-throughput library design. Based on this, GOOSE also uses a set of default parameters for sequence design, all of which can be overridden by the user should they choose (**Table S6**).

GOOSE is open source and can be used as a Python library or within a Google Colab notebook (https://colab.research.google.com/drive/1U9B-TfoNEZbbjhPUG5lrMPS0JL0nDB3o?usp=sharing). We provided extensive documentation (https://goose.readthedocs.io/en/latest/index.html), which is not reproduced in this supplementary information due to length but can be readily accessed through the web.

Functionally, GOOSE relies on a stochastic design algorithm, which enables GOOSE to generate unique sequences, even if numerous sequence properties are specified. Sequence generation starts with the creation of a 'base sequence' that comes close to satisfying user-specified input parameters. From here, various functions are used to fine-tune the sequence such that the sequence parameters match the input parameters. Then, optimization functions are employed to optimize for sequence disorder while maintaining any sequence parameter constraints. Finally, the sequence is checked for predicted disorder using Metapredict V2-FF[9,12]. GOOSE includes functionality to generate sequences by specifying sequence properties, fractions of amino acids, radius of gyration, or end-to-end distance. GOOSE can also generate sequence variants with specific design constraints from a starting IDR sequence of interest. **Table S2** summarizes the types of sequence designs enabled via GOOSE.

GOOSE documentation is provided through ReadTheDocs (https://about.readthedocs.com/) and Sphinx (https://www.sphinx-doc.org/en/master/), with unit testing provided by PyTest (https://docs.pytest.org/). Version control is done via Git (https://git-scm.com/) and GitHub (https://github.com/). GOOSE uses metapredict[12] (V2-FF) for disorder prediction, as well as Numpy, SciPy, and PyTorch (https://pytorch.org/) for various internal functions[4,13]. GOOSE continues to be in active development, and new features will be added regularly. The version associated with this manuscript is version 0.1.2 at the time of submission.

Sequence designs in this paper

This paper used sequence generation by specifying "sequence properties" functionality for sequence design. In particular, sequences were designed with the following quantized sequence properties: NCPR of –0.6, –0.3, 0.0, +0.3, +0.6, FRC of 0.0, 0.3, or 0.6, Kyte-Doolittle hydrophobicity of 1.0 or 3.0 (on a 0-to-9 scale), and kappa [κ] (a measure of charge patterning, see **Fig. S4**) was set to be between 0.05 and 0.22 (low-to-average, depending on sequence composition) and then above 0.5 for highly clustered sequences. The quantization of charged residues was selected to match specific regions on the Das-Pappu diagram of states, enabling the exploration of IDRs with distinct charge properties (**Fig. S18**)[14,15]. The quantization of hydrophobicity (and 1.0 or 3.0) was selected for two reasons. Firstly, keeping hydrophobicity low minimizes the risk of our synthetic IDRs triggering the unfolded protein response. Secondly, because hydrophobicity is intrinsically coupled with FRC, enabling the FCR and hydrophobicity to be independently varied required lower hydrophobicity scores to accommodate highly charged sequences. Finally (and expected), all designed sequences are strongly predicted to be disordered (**Fig. S19**).

Given the scope of sequence space for 60-residue disordered proteins (a conservative lower bound of 60^{10}) and the relatively low-throughput experimental characterization employed here to ensure high-quality data is reported, we opted to approach our design problem in terms of designing sets of pairs of sequences (**Table S3**). Each pair enables the specific comparison of one sequence parameter by holding others fixed while varying one specific parameter (e.g., net charge, hydrophobicity, etc). By designing our library to multiplex distinct hypotheses, the same sequences could be members of multiple pairs, enabling us to systematically test a collection of hypotheses with a relatively low number of sequences.

Applications of GOOSE

While the backend of GOOSE is a relatively large software package, the user-facing functionality was designed to provide a minimalist interface that makes systematic titration of specific sequence properties straightforward, abstracting the complexities of sequence design entirely from the user.

We have previously used GOOSE to design libraries of thousands of sequences, which provided input data for deep learning models when used in conjunction with molecular dynamics simulations[9]. While it is commonplace to use natural sequences when performing high-throughput computational or experimental studies, natural sequences only explore small slithers of the potential sequence space available to polypeptides. As such, we have found that combining biological sequences (which effectively biases a library towards biologically relevant sequences) with fully synthetic sequences enables a much more comprehensive exploration of sequence space.

Helicity prediction

Helicity prediction (**Fig. S12**) was performed using JPred4 in batch mode[16].

Bioinformatic analysis

Bioinformatic sequence analysis was performed using localCIDER[14] and sparrow (https://github.com/idptools/sparrow/). Disorder prediction shown in **Fig. 1** and **Fig. S18** was performed using metapredict (V2-FF)[9,12]. Protein sequences were obtained from UniProt[17] and download September 2023, and reflect the following proteomes: UP000000589 (*Mus musculus*, TaxonID: 10090), UP000000803 (*Drosophila melanogaster*, TaxonID:7227), UP000001805 (*Neurospora crassa* OR74A, TaxonID: 367110), UP000001940 (*Caenorhabditis elegans*, TaxonID:6239), UP000002311 (*Saccharomyces cerevisiae* S288C, TaxonID:559292), UP000005640 (*Homo sapiens*, TaxonID:9606), and UP000006548 (*Arabidopsis thaliana*, TaxonID:3702). All IDRs for all organisms are precomputed and provided in the shared GitHub directory.

Coarse-grained simulations

Coarse-grained molecular dynamics simulations were performed using the LAMMPS simulation engine[18] using a modified version of the Mpipi[10] parameters, Mpipi-GG[9]. Starting positions for IDRs were generated by assembling beads as a random coil in the excluded volume limit (i.e.,

where beads do not overlap). From this position, an energy minimization protocol was carried out with a maximum of 1,000 iterations. Simulations were then carried out with an implicit salt concentration of 150 mM and a temperature of 300 K. Simulation analysis was performed using MDTraj[19] and SOURSOP[20].

For simulations of each of the 32 primary sequences (60 residues, **Fig. S11**), all sequences were run in triplicate for 50,000,000 steps with a ten femtosecond timestep for a total of 1.5 μs per sequence. The first 1,000,000 steps for each simulation were discarded as equilibration. After equilibration, output coordinate positions for each trajectory were recorded at intervals of 10,000 steps, for a total of 4,900 recorded steps per individual simulation. This simulation length was chosen based on prior work to benchmark appropriate simulation lengths to obtain robust conformational sampling[9]. Error bars (shown in **Fig. S11**) show minimal variability between independent replicas (on the order of the marker size in the figure), confirming that simulations sufficiently sample the conformational landscape. The simulation used a 500 $Å^3$ box with periodic boundary conditions.

For simulations of IDR designed to match specific radii of gyration or end-to-end distances (200 residues, **Fig. S14**), all sequences were run in triplicate for 200,000,000 steps with a 20 femtoseconds timestep for a total of 12 μs per sequence. The first 1,000,000 steps for each simulation were discarded as equilibration. After equilibration, output coordinate positions for each trajectory were recorded at intervals of 100,000 steps for 1,990 recorded steps per simulation. This simulation length was chosen based on prior work to benchmark appropriate simulation lengths to obtain robust conformational sampling[9]. Error bars (shown in **Fig. S14**) show that the variability between independent replicas is minimal (on the order of the marker size in the figure), again confirming that simulations sufficiently sample the conformational landscape. The simulation used a 500 $Å^3$ box with periodic boundary conditions.

Limitations, drawbacks, and caveats of GOOSE

GOOSE was designed to generate fully synthetic IDR sequences. In the current version of GOOSE (0.1.2), we do not constrain the predicted secondary structure, assuming that sequences with a strong tendency towards disorder prediction will – in isolation – be largely disordered. That said, GOOSE does offer the ability to check for predicted helicity if this is a possible confounding factor of concern.

Secondly, rationally designed sequences may possess motifs or sequence features that make them good targets for phosphorylation, degradation, or unexpected interaction with cellular components. This is not a “limitation” in as much as our goal in GOOSE is to generate fully synthetic sequences and variants, but it is a factor that should be considered when designing libraries.

Sequences designed to match specific ensemble properties (i.e., the radius of gyration or end-to-end distance) use ALBATROSS, our deep-learning tool for sequence-ensemble prediction. ALBATROSS enables the rapid prediction of ensemble dimensions from sequence[9]. While ALBATROSS is reasonably accurate, it is certainly not perfect. ALBATROSS may be less accurate for sequences with substantial secondary structure or that are extreme in terms of

composition or sequence patterning. With this in mind, we encourage scrutiny and skepticism when designing sequences with extreme values in terms of both sequence properties and predicted dimensions.

GOOSE does not currently offer the ability to optimize nucleotide sequences to minimize repetitive sequences at the DNA level and/or codon optimization for a specific organism. Given the repetitive nature of some IDRs, we plan to introduce this feature going forward, but for now, nucleotide sequence optimization must be done independently of protein sequence design.

Finally, recent complementary work by Strome et al. offers the ability to design IDRs to match bulk sequence properties against IDRs of a specific biological class or group[21]. This approach is conceptually distinct from ours and enables a different set of questions to be asked (i.e., designing synthetic IDRs to 'mimic' a large set of sequence features). We see this as highly complementary to our work. In parallel, work by Pesce et al. has recently shown the ability to design IDRs with specific ensemble properties[22]. GOOSE does enable a similar feature, but unlike the work by Pesce et al., GOOSE is limited to designing sequences with specified ensemble-average properties only, whereas, in principle, the Pesce *et al.* approach could be used to create sequences with specific local and global conformational biases. Again, this work is highly complementary to ours, highlighting the growing importance of IDR design as an approach to synthetic biology and basic science. Finally, these two approaches build on prior work, notably the ability to couple simulations and a genetic algorithm to design IDRs with bespoke helicity profiles (GADIS)[23], as well as early work enabling the design of IDR sequences of a specified length [24]

Data availability

GOOSE source code is available at https://github.com/idptools/goose/

GOOSE documentation is available at https://goose.readthedocs.io/

Data and analysis scripts used for figures and analysis in this paper are available at https://github.com/holehouse-lab/supportingdata/tree/master/2023/emenecker_guadalupe_2023 and also at https://github.com/sukeniklab/emenecker_guadalupe_2023.

SUPPLEMENTARY TABLES

Table S1: Sequences and parameters. Columns reported here include the Fraction of Charged Residues (FCR), Net Charge Per Residue (NCPR), charge-distribution metric kappa (κ), Kyte-Doolittle hydrophobicity (H), and average disorder prediction (Disorder). Note that a sequence must have both positive and negative residues for κ to be calculated, otherwise is reported as -1. Disorder here is calculated as the average disorder score as predicted by metapredict[12].

#	Sequence	FCR	NCPR	κ	H	Disorder
1	QNNNQQQNQQNQNNQNNNQNNNNQNQNNQQQQNQQQNN NQQQNNQQNNNNQQQNNNQNNN	0	0	-1	1	0.88
2	THNHHSTPGTPGHHHPGSPHSPHPTHTTPSHHGTGGGH GGSTTQSHSNGSATGQHGSSGP	0	0	-1	3	0.89
3	THNHHSTGTGHHHGSHSHTHTTSHHGTGGGHGGSTTQS HSNGSTGQHGSSGPPPPPPPAP	0	0	-1	3	0.91
4	THNHHGPSTGTGPHHHGSPHSHTHPTTSHHPGTGGGPH GGSTPTQSHSANGSTGPQHGSS	0	0	-1	3	0.9
5	HNNQQQQQNNQNNQNQQNNQNNNNNQQQQNQQEDEQQD DENQDDDQQDEDQNEDEQQDDE	0.3	-0.3	-1	1	0.94
6	EDDEHNNQDQQQQENNQNDNQNQDQNNQDNNNNNQQQE QNQQDQQNQDQQQNDQQEEDED	0.3	-0.3	-1	1	0.92
7	KKKRSGQNRNQNNRQQQQKNQNQRQNQNKNNQQNQNNE NNNNDNNQQENNQQENQDDDDD	0.3	0	0.53	1	0.92
8	REKRSGQNKNQNNRQQQQDNQNQEQNQNDNNQQNQNND NNNNKNNQQDNNQQRNQEKDKD	0.3	0	0.13	1	0.95
9	KKRSGKRRHHKKRHQKRRQHRKKQQRKKQNQNNNNNQN QQQQHHNQQNNQNNPNNNNHHP	0.3	+0.3	-1	1	0.85
10	KRKKSGHHKHQQHRQQQNKQNNNRNNQNKQQQQHHNQK QNNQRNNPNRNNNHRHPKKRKR	0.3	+0.3	-1	1	0.9
11	GQSTSTSWGGWGTSGSGGGTSSGGWSSTTSGGEDDGTD DEGTDDETADDDSTDEEGSDED	0.3	-0.3	-1	3	0.94
12	EDDDGQSTESTSWDGGWGDTSGSDGGGTDSSGGWSSTD TSGGEGTGTDTASTDGSEEDED	0.3	-0.3	-1	3	0.94

13	KKKKGHTGRTGTGRGTTGRSSGARGGSARTTTTSSSSE SGSSDSSSAETASSDSSDEEEE	0.3	0	0.53	3	0.95
14	KEDKGHTGDTGTGRGTTGRSSGAKGGSAETTTTSSSSE SGSSRSSSAKTASSESSERRED	0.3	0	0.13	3	0.95
15	GHTGTGTGGTTGSSGGGSTTTTSSSSSGSSSSSTSSSS DKEKDREAKERDARERERKEAAR	0.3	0	0.01	3	0.89
16	GHDTGKTGETGKGTDTGRSSEGGAGSKTTETTRSSDSS ASGRSSESSRSTESSKSSEAAR	0.3	0	0.06	3	0.95
17	KRRNQKRRGTRKRSGKRKGSRKRGARRKSTATATGSTT SSATASASSSSSSSGAGATGGS	0.3	+0.3	-1	3	0.94
18	KRRKNQGTKSGGSKGASTRATATRGSTTKSSATASASR SSSSRSSGARGATGKGSRRKRR	0.3	+0.3	-1	3	0.94
19	KKRKAHHNKNNQNRQNNNRNQNQRNQNQKNQQQDDEEE DDDDDDDEEDDDDDDDDDDEE	0.6	-0.3	0.84	1	0.95
20	DKDAHDKRHNKEDNNKDDQNEDEQNDDDNNNQDEKNQR DDNQERENQDEDNQDDDQQRDD	0.6	-0.3	0.19	1	0.97
21	KRKSNRRRPPRRKNNKKRPNRRRQQRKRNNPQDDENPE DDNNDDEQQDEEQHEDEPQEEE	0.6	0	0.64	1	0.96
22	DDRSNERRPPERENNEKKPNEREQQDKKNNPQKRRNPD DRNNDRKQQEREQHRDEPQDRE	0.6	0	0.06	1	0.96
23	RRRKKKKTKRRKMRKRRPRKKRNKRKRPRRRKPQQNNE QQNNDNQNNENQNSDQSEDDDD	0.6	+0.3	0.72	1	0.94
24	DDRTMRRKPNKKEPPRRRQQKKDNNRRKQQNNRRENQR RDNNRKKNQKKRNSKKDQSRDE	0.6	+0.3	0.22	1	0.91
25	KKKKMVVSRSVVGRAGMARSVVAKSVVARPVSGDDDEE EEEDDDEDDDDDEEDEDDDDDD	0.6	-0.3	0.84	3	0.86
26	DDKEMVDKDVSEEKSVDEDVGKRDAGDDDMASVVDDAS VDEVAPREVSGERDRDKDEDED	0.6	-0.3	0.16	3	0.92
27	KRRAVKKKVAKRKSSRKKVVKKRGARKKSVSVDEDVPE DDVVDDDGSDDDVSDDEAVDEE	0.6	0	0.64	3	0.88
28	EKDAVDEKVADKKSSKDKVVDDRGAKRDSVSVDRDVPD EKVVRDKGSDREVSEKKAVKRD	0.6	0	0.06	3	0.9

29	SSGSSGSSEKDAVDEKVADKKKDKVVDDRAKRDVVDRD VPDEKVVRDKDREVEKKAVKRD	0.6	0	0.05	3	0.84
30	SESKGDSASVGDSESKVADKKKDKVVDDRAKRDVVDRD VPDEKVVRDKDREVEKKAVKRD	0.6	0	0.05	3	0.86
31	KRRKRKRVRKRRVKRKKAKRRRAKRRKVKKKRSAAVGE AVAADSAVADAVVVDSADDDDD	0.6	0.3	0.72	3	0.76
32	KRKVVDRDAARRRVSDRKAAKDRVGKRKAVAARKKSAR KDVADEKAVRDRVVKRKSADRK	0.6	0.3	0.18	3	0.61

Table S2: Summary of GOOSE design options

Goose function	Description
sequence()	***De novo sequence generation.*** Function that allows disordered sequence generation by specifying length, hydropathy (optional), FCR (optional), NCPR (optional), and kappa (optional). Any properties not specified will be unconstrained during sequence generation and will change from sequence to sequence if multiple sequences are generated.
seq_fractions()	***De novo sequence generation.*** Function that allows disordered sequence generation by specifying the fractions of amino acids. Multiple fractions can be specified. Any amino acids not specified will be unconstrained during sequence generation and will change from sequence to sequence if multiple sequences are generated.
seq_re()	***De novo sequence generation.*** Function that allows disordered sequence generation by specifying sequence length and end-to-end distance.
seq_rg()	***De novo sequence generation.*** Function that allows disordered sequence generation by specifying sequence length and the radius of gyration.
constant_class_var()	***Variant generator.*** Function that allows the creation of variants with the same overall bulk properties (FCR, NCPR, hydropathy, kappa) as the input variant as well as the same order and number of amino acids, as grouped by class (see **Table S4**). Variants will have different amino acid identities while keeping everything else constant.

new_seq_constant_class_var()	***Variant generator.*** Function that allows the creation of variants where the sequence composition is new, but the numbers of each residue from each class and the overall properties (FCR, NCPR, hydrophobicity) are the same. Unlike variants generated by the constant_class_var() function, the order of the amino acids (in terms of class) is not preserved.
constant_properties_var()	***Variant generator.*** Function that allows the creation of variants where only the sequence properties (FCR, NCPR, hydrophobicity, kappa) are constrained. There are no constraints on classes of amino acids.
constant_residue_var()	***Variant generator.*** Function that allows the creation of variants where specific (user-specified) residues are held constant by position and number. The variant will have the same overall bulk properties (FCR, NCPR, hydrophobicity) as the original sequence.
shuffle_var()	***Variant generator.*** Function that allows the creation of variants where specific regions of an IDR are shuffled. Multiple regions can be specified simultaneously.
excluded_shuffle_var()	***Variant generator.*** Function that allows the creation of variants where an entire sequence is shuffled except for user-specific residues. Note this is *not* the reciprocal of shuffle_var(), which operates in terms of regions instead of residues.
targeted_shuffle_var()	***Variant generator.*** Function that allows the creation of variants where only user-specified residues are shuffled. Any residues not specified will not be shuffled.
asymmetry_var()	***Variant generator.*** Function that allows the creation of variants where a class of residues (or a user-specified list of residue identities) is changed to become more asymmetrically or less asymmetrically distributed throughout the sequence. Does NOT change sequence composition.
hydro_class_var()	***Variant generator.*** Function that allows the creation of variants where, like the constant_class_var() function, the overall sequence properties (FCR, NCPR, kappa), the order, and the number of amino acids according to each class is held constant, however, the hydropathy can be increased or decreased (within the inherent constraints imposed by the class constraint).
fcr_class_var()	***Variant generator.*** Function that allows the creation of variants where variants adjust the FCR while minimizing changes to the position and number of amino acids by

	class.
ncpr_class_var()	***Variant generator.*** Function that allows the creation of variants where variants adjust the NCPR while minimizing changes to the position and number of amino acids by class.
kappa_var()	***Variant generator.*** Function that allows the creation of variants where the charge asymmetry is altered by changing the sequence's kappa value. Requires that both positively charged and negatively charged residues are found in the original sequence.
all_props_class_var()	***Variant generator.*** Function that allows the creation of sequence variants that adjust the FCR, NCPR, hydropathy, and kappa values while minimizing changes to the position and number of amino acids by class.
re_var()	***Variant generator.*** Function that allows the creation of sequence variants that adjust the sequence radius of gyration (R_g).
rg_var()	***Variant generator.*** Function that allows the creation of sequence variants that adjust the sequence end-to-end distance (R_e).
seq_property_library()	***Library generator.*** Function that generates a library of sequences that span a range of user-specified properties including length, FCR, NCPR, hydropathy, and kappa.
seq_fractions_library()	***Library generator.*** Function that generates a library of sequences that span a range of user-specified fractions of amino acids. Multiple fractions can be specified simultaneously.

Table S3: All sequence pairs compared in **Fig. 2** and **Fig. 3**

2C	6 EDDEHNNQDQQQQENNQNDNQNQDQNNQDNNNNNQQQEQNQQDQQNQDQQQNDQQEEDED 20 DKDAHDKRHNKEDNNKDDQNEDEQNDDDNNNQDEKNQRDDNQERENQDEDNQDDDQQRDD 12 EDDDGQSTESTSWDGGWGDTSGSDGGGTDSSGGWSSTDTSGGEGTGTDTASTDGSEEDED 26 DDKEMVDKDVSEEKSVDEDVGKRDAGDDDMASVVDDASVDEVAPREVSGERDRDKDEDED 10 KRKKSGHHKHQQHRQQQNKQNNNRNNQNKQQQQHHNQKQNNQRNNPNRNNNHRHPKKRKR 24 DDRTMRRKPNKKEPPRRRQQKKDNNRRKQQNNRRENQRRDNNRKKNQKKRNSKKDQSRDE 18 KRRKNQGTKSGGSKGASTRATATRGSTTKSSATASASRSSSSRSSGARGATGKGSRRKRR 32 KRKVVDRDAARRRVSDRKAAKDRVGKRKAVAARKKSARKDVADEKAVRDRVVKRKSADRK

2D	5 HNNQQQQQNNQNNQNQQNNQNNNNNQQQQNQQEDEQQDDENQDDDQQDEDQNEDEQQDDE 19 KKRKAHHNKNNQNRQNNNRNQNQRNQNQKNQQQDDEEEDDDDDDDEEDDDDDDDDDDEE 11 GQSTSTSWGGWGTSGSGGGTSSGGWSSTTSGGEDDGTDDEGTDDETADDDSTDEEGSDED 25 KKKKMVVSRSVVGRAGMARSVVAKSVVARPVSGDDDEEEEDDDEDDDDDEEDEDDDDDD 9 KKRSGKRRHHKKRHQKRRQHRKKQQRKKQNQNNNNNQNQQQQHHNQQNNQNNPNNNNHHP 23 RRRKKKKTKRRKMRKRRPRKKRNKRKRPRRRKPQQNNEQQNNDNQNNENQNSDQSEDDDD
2E	8 REKRSGQNKNQNNRQQQQDNQNQEQNQNDNNQQNQNNDNNNNKNNQQDNNQQRNQEKDKD 22 DDRSNERRPPERENNEKKPNEREQQDKKNNPQKRRNPDDRNNDRKQQEREQHRDEPQDRE 14 KEDKGHTGDTGTGRGTTGRSSGAKGGSAETTTTSSSSESGSSRSSSAKTASSESSERRED 28 EKDAVDEKVADKKSSKDKVVDDRGAKRDSVSVDRDVPDEKVVRDKGSDREVSEKKAVKRD 7 KKKRSGQNRNQNNRQQQQKNQNQRQNQNKNNQQNQNNENNNNDNNQQENNQQENQDDDDD 21 KRKSNRRRPPRRKNNKKRPNRRRQQRKRNNPQDDENPEDDNNDDEQQDEEQHEDEPQEEE 9 KKRSGKRRHHKKRHQKRRQHRKKQQRKKQNQNNNNNQNQQQQHHNQQNNQNNPNNNNHHP 23 RRRKKKKTKRRKMRKRRPRKKRNKRKRPRRRKPQQNNEQQNNDNQNNENQNSDQSEDDDD
2F	6 EDDEHNNQDQQQQENNQNDNQNQDQNNQDNNNNNQQQEQNQQDQQNQDQQQNDQQEEDED 8 REKRSGQNKNQNNRQQQQDNQNQEQNQNDNNQQNQNNDNNNNKNNQQDNNQQRNQEKDKD 10 KRKKSGHHKHQQHRQQQNKQNNNRNNQNKQQQQHHNQKQNNQRNNPNRNNNHRHPKKRKR 12 EDDDGQSTESTSWDGGWGDTSGSDGGGTDSSGGWSSTDTSGGEGTGTDTASTDGSEEDED 14 KEDKGHTGDTGTGRGTTGRSSGAKGGSAETTTTSSSSESGSSRSSSAKTASSESSERRED 18 KRRKNQGTKSGGSKGASTRATATRGSTTKSSATASASRSSSSRSSGARGATGKGSRRKRR 5 HNNQQQQQNNQNNQNQQNNQNNNNNQQQQNQQEDEQQDDENQDDDQQDEDQNEDEQQDDE 7 KKKRSGQNRNQNNRQQQQKNQNQRQNQNKNNQQNQNNENNNNDNNQQENNQQENQDDDDD 9 KKRSGKRRHHKKRHQKRRQHRKKQQRKKQNQNNNNNQNQQQQHHNQQNNQNNPNNNNHHP
2G	20 DKDAHDKRHNKEDNNKDDQNEDEQNDDDNNNQDEKNQRDDNQERENQDEDNQDDDQQRDD 22 DDRSNERRPPERENNEKKPNEREQQDKKNNPQKRRNPDDRNNDRKQQEREQHRDEPQDRE 24 DDRTMRRKPNKKEPPRRRQQKKDNNRRKQQNNRRENQRRDNNRKKNQKKRNSKKDQSRDE 26 DDKEMVDKDVSEEKSVDEDVGKRDAGDDDMASVVDDASVDEVAPREVSGERDRDKDEDED 28 EKDAVDEKVADKKSSKDKVVDDRGAKRDSVSVDRDVPDEKVVRDKGSDREVSEKKAVKRD 32 KRKVVDRDAARRRVSDRKAAKDRVGKRKAVAARKKSARKDVADEKAVRDRVVKRKSADRK
3B	6 EDDEHNNQDQQQQENNQNDNQNQDQNNQDNNNNNQQQEQNQQDQQNQDQQQNDQQEEDED 12 EDDDGQSTESTSWDGGWGDTSGSDGGGTDSSGGWSSTDTSGGEGTGTDTASTDGSEEDED

	20 DKDAHDKRHNKEDNNKDDQNEDEQNDDDNNNQDEKNQRDDNQERENQDEDNQDDDQQRDD 26 DDKEMVDKDVSEEKSVDEDVGKRDAGDDDMASVVDDASVDEVAPREVSGERDRDKDEDED 19 KKRKAHHNKNNQNRQNNNRNQNQRNQNQKNQQQDDEEEDDDDDDDEEDDDDDDDDDDEE 25 KKKKMVVSRSVVGRAGMARSVVAKSVVARPVSGDDDEEEEEDDDEDDDDDEEDEDDDDDD 1 QNNNQQQNQQNQNNQNNNQNNNNQNQNNQQQQNQQQNNNQQQNNQQNNNNQQQNNNQNNN 2 THNHHSTPGTPGHHHPGSPHSPHPTHTTPSHHGTGGGHGGSTTQSHSNGSATGQHGSSGP 7 KKKRSGQNRNQNNRQQQQKNQNQRQNQNKNNQQNQNNENNNNDNNQQENNQQENQDDDDD 13 KKKKGHTGRTGTGRGTTGRSSGARGGSARTTTTSSSSESGSSDSSSAETASSDSSDEEEE 10 KRKKSGHHKHQQHRQQQNKQNNNRNNQNKQQQQHHNQKQNNQRNNPNRNNNHRHPKKRKR 18 KRRKNQGTKSGGSKGASTRATATRGSTTKSSATASASRSSSSRSSGARGATGKGSRRKRR 9 REKRSGQNKNQNNRQQQQDNQNQEQNQNDNNQQNQNNDNNNNKNNQQDNNQQRNQEKDKD 17 KRRNQKRRGTRKRSGKRKGSRKRGARRKSTATATGSTTSSATASASSSSSSSGAGATGGS 24 DDRTMRRKPNKKEPPRRRQQKKDNNRRKQQNNRRENQRRDNNRKKNQKKRNSKKDQSRDE 32 KRKVVDRDAARRRVSDRKAAKDRVGKRKAVAARKKSARKDVADEKAVRDRVVKRKSADRK
3C	5 HNNQQQQQNNQNNQNQQNNQNNNNNQQQQNQQEDEQQDDENQDDDQQDEDQNEDEQQDDE 11 GQSTSTSWGGWGTSGSGGGTSSGGWSSTTSGGEDDGTDDEGTDDETADDDSTDEEGSDED 21 KRKSNRRRPPRRKNNKKRPNRRRQQRKRNNPQDDENPEDDNNDDEQQDEEQHEDEPQEEE 27 KRRAVKKKVAKRKSSRKKVVKKRGARKKSVSVDEDVPEDDVVDDDGSDDDVSDDEAVDEE 8 REKRSGQNKNQNNRQQQQDNQNQEQNQNDNNQQNQNNDNNNNKNNQQDNNQQRNQEKDKD 14 KEDKGHTGDTGTGRGTTGRSSGAKGGSAETTTTSSSSESGSSRSSSAKTASSESSERRED 22 DDRSNERRPPERENNEKKPNEREQQDKKNNPQKRRNPDDRNNDRKQQEREQHRDEPQDRE 28 EKDAVDEKVADKKSSKDKVVDDRGAKRDSVSVDRDVPDEKVVRDKGSDREVSEKKAVKRD

Table S4
Default amino acid classes used in GOOSE

Class name	Amino acids
Aromatic	F, W, Y
Polar	Q, N, S, T
Positive	K, R
Negative	D, E
Hydrophobic	I, V, L, A, M
Polar	C, P, G, H

Note that in other contexts, G and H might be considered polar, and H may also be considered positive under depressed pH regimes.

Table S5 N for all violin plots. Basal corresponds to all measurements before any perturbation. Columns 100, 300, and 750 correspond to hypo-osmotic, iso-osmotic and hyper-osmotic shock.

Sequence ID	Basal counts	100 mOsm counts	300 mOsm counts	750 mOsm counts
1	16	3	3	3
2	20	4	4	4
3	21	7	5	4
4	19	4	4	4
5	18	6	3	4
6	15	3	5	1
7	9	2	2	1
8	9	3	2	2
9	22	5	6	3
10	7	1	1	1
11	21	7	4	4
12	20	6	4	3
13	19	5	4	4

14	19	5	5	5
15	9	3	3	2
16	21	7	4	5
17	23	7	3	4
18	17	3	5	3
19	9	1	1	1
20	20	5	5	6
21	21	5	3	4
22	9	3	3	3
23	8	1	1	1
24	7	1	1	1
25	20	3	4	3
26	13	2	5	4
27	14	5	5	3
28	14	5	3	3
29	24	5	6	6
30	15	3	4	3
31	18	2	4	3
32	19	4	3	3

Table S6
Default GOOSE parameters

Parameter	Default Value
Minimum Length	10
Maximum Length	10,000
Maximum Hydropathy (Kyte-Doolittle Scale shifted scale of 0 to 9)	6.1
Disorder Threshold (metapredict V2)	0.5
Max deviation from user-input hydropathy	0.07
Max deviation from user-input kappa	0.03
Number of attempts to make sequence	100
Max Fraction A	0.95
Max Fraction C	1.0
Max Fraction D	1.0
Max Fraction E	1.0
Max Fraction F	1.0
Max Fraction G	1.0
Max Fraction H	1.0
Max Fraction I	0.53
Max Fraction K	1.0
Max Fraction L	0.42
Max Fraction M	0.62
Max Fraction N	1.0
Max Fraction P	1.0
Max Fraction Q	1.0
Max Fraction R	1.0
Max Fraction S	1.0

Max Fraction T	1.0
Max Fraction V	0.71
Max Fraction W	0.55
Max Fraction Y	0.99

Note: Maximum fractions were determined by attempting to generate a sequence of 100 amino acids in length at each fraction for every amino acid between the decimal fraction values of 0.01 to 1.00. For each fraction value, the sequence was populated with the necessary number of the amino acids of interest, and then the rest of the sequence was generated by populating the sequence with any amino acid other than the amino acid that had its maximum fraction determined. 500,000 sequences were attempted at each fractional value and then checked to be disordered using metapredict V2 with a cutoff of 0.5.

Table S7 Sequences that show naive response (expand) under hypo-osmotic shock (100 mOsm)

Sequence ID	**Sequence**
4	THNHHGPSTGTGPHHHGSPHSHTHPTTSHHPGTGGGPHGGSTPTQSHSANGSTGPQHGSS
15	GHTGTGTGGTTGSSGGGSTTTTSSSSSGSSSSSTSSSSDKEKDREAKERDARERERKEAAR
26	DDKEMVDKDVSEEKSVDEDVGKRDAGDDDMASVVDDASVDEVAPREVSGERDRDKDEDED
29	SSGSSGSSEKDAVDEKVADKKKDKVVDDRAKRDVVDRDVPDEKVVRDKDREVEKKAVKRD
30	SESKGDSASVGDSESKVADKKKDKVVDDRAKRDVVDRDVPDEKVVRDKDREVEKKAVKRD

Table S8 Sequences that show insensitive response under hypo-osmotic shock (100 mOsm)

Sequence ID	Sequence
1	QNNNQQQNQQNQNNQNNNQNNNNQNQNNQQQQNQQQNNNQQQNNQQNNNNQQQNNNQNNN
2	THNHHSTPGTPGHHHPGSPHSPHPTHTTPSHHGTGGGHGGSTTQSHSNGSATGQHGSSGP
3	THNHHSTGTGHHHGSHSHTHTTSHHGTGGGHGGSTTQSHSNGSTGQHGSSGPPPPPPPAP
5	HNNQQQQQNNQNNQNQQNNQNNNNNQQQQNQQEDEQQDDENQDDDQQDEDQNEDEQQDDE
6	EDDEHNNQDQQQQENNQNDNQNQDQNNQDNNNNNQQQEQNQQDQQNQDQQQNDQQEEDED
7	KKKRSGQNRNQNNRQQQQKNQNQRQNQNKNNQQNQNNENNNNDNNQQENNQQENQDDDDD
8	REKRSGQNKNQNNRQQQQDNQNQEQNQNDNNQQNQNNDNNNNKNNQQDNNQQRNQEKDKD
11	GQSTSTSWGGWGTSGSGGGTSSGGWSSTTSGGEDDGTDDEGTDDETADDDSTDEEGSDED
12	EDDDGQSTESTSWDGGWGDTSGSDGGGTDSSGGWSSTDTSGGEGTGTDTASTDGSEEDED
13	KKKKGHTGRTGTGRGTTGRSSGARGGSARTTTTSSSSESGSSDSSSAETASSDSSDEEEE
14	KEDKGHTGDTGTGRGTTGRSSGAKGGSAETTTTSSSSESGSSRSSSAKTASSESSERRED
16	GHDTGKTGETGKGTDTGRSSEGGAGSKTTETTRSSDSSASGRSSESSRSTESSKSSEAA
20	DKDAHDKRHNKEDNNKDDQNEDEQNDDDNNNQDEKNQRDDNQERENQDEDNQDDDQQRDD
21	KRKSNRRRPPRRKNNKKRPNRRRQQRKRNNPQDDENPEDDNNDDEQQDEEQHEDEPQEEE
22	DDRSNERRPPERENNEKKPNEREQQDKKNNPQKRRNPDDRNNDRKQQEREQHRDEPQDRE
27	KRRAVKKKVAKRKSSRKKVVKKRGARKKSVSVDEDVPEDDVVDDDGSDDDVSDDEAVDEE
28	EKDAVDEKVADKKSSKDKVVDDRGAKRDSVSVDRDVPDEKVVRDKGSDREVSEKKAVKRD
31	KRRKRKRVRKRRVKRKKAKRRRAKRRKVKKKRSAAVGEAVAADSAVADAVVVDSADDDDD
32	KRKVVDRDAARRRVSDRKAAKDRVGKRKAVAARKKSARKDVADEKAVRDRVVKRKSADRK

Table S9 Sequences that show inverse response (compact) under hypo-osmotic shock (100 mOsm)

Sequence ID	Sequence
9	KKRSGKRRHHKKRHQKRRQHRKKQQRKKQNQNNNNNQNQQQQHHNQQNNQNNPNNNNHHP
17	KRRNQKRRGTRKRSGKRKGSRKRGARRKSTATATGSTTSSATASASSSSSSSGAGATGGS
18	KRRKNQGTKSGGSKGASTRATATRGSTTKSSATASASRSSSSRSSGARGATGKGSRRKRR
25	KKKKMVVSRSVVGRAGMARSVVAKSVVARPVSGDDDEEEEEDDDEDDDDDEEDEDDDDDD

Table S10 Sequences that show naive response (compact) under hyper-osmotic shock (750 mOsm)

Sequence ID	Sequence
1	QNNNQQQNQQNQNNQNNNQNNNNQNQNNQQQQNQQQNNNQQQNNQQNNNNQQQNNNQNNN
2	THNHHSTPGTPGHHHPGSPHSPHPTHTTPSHHGTGGGHGGSTTQSHSNGSATGQHGSSGP
3	THNHHSTGTGHHHGSHSHTHTTSHHGTGGGHGGSTTQSHSNGSTGQHGSSGPPPPPPPAP
4	THNHHGPSTGTGPHHHGSPHSHTHPTTSHHPGTGGGPHGGSTPTQSHSANGSTGPQHGSS
11	GQSTSTSWGGWGTSGSGGGTSSGGWSSTTSGGEDDGTDDEGTDDETADDDSTDEEGSDED
14	KEDKGHTGDTGTGRGTTGRSSGAKGGSAETTTTSSSSESGSSRSSSAKTASSESSERRED
15	GHTGTGTGGTTGSSGGGSTTTTSSSSSGSSSSSTSSSSDKEKDREAKERDARERERKEAAR
16	GHDTGKTGETGKGTDTGRSSEGGAGSKTTETTRSSDSSASGRSSESSRSTESSKSSEAAR
20	DKDAHDKRHNKEDNNKDDQNEDEQNDDDNNNQDEKNQRDDNQERENQDEDNQDDDQQRDD
29	SSGSSGSSEKDAVDEKVADKKKDKVVDDRAKRDVVDRDVPDEKVVRDKDREVEKKAVKRD

Table S11 Sequences that show insensitive response under hyper-osmotic shock (750 mOsm)

Sequence ID	Sequence
5	HNNQQQQQNNQNNQNQQNNQNNNNNQQQQNQQEDEQQDDENQDDDQQDEDQNEDEQQDDE
8	REKRSGQNKNQNNRQQQQDNQNQEQNQNDNNQQNQNNDNNNNKNNQQDNNQQRNQEKDKD
12	EDDDGQSTESTSWDGGWGDTSGSDGGGTDSSGGWSSTDTSGGEGTGTDTASTDGSEEDED
13	KKKKGHTGRTGTGRGTTGRSSGARGGSARTTTTSSSSESGSSDSSSAETASSDSSDEEEE
21	KRKSNRRRPPRRKNNKKRPNRRRQQRKRNNPQDDENPEDDNNDDEQQDEEQHEDEPQEEE
22	DDRSNERRPPERENNEKKPNEREQQDKKNNPQKRRNPDDRNNDRKQQEREQHRDEPQDRE
25	KKKKMVVSRSVVGRAGMARSVVAKSVVARPVSGDDDEEEEEDDDEDDDDDEEDEDDDDDD
26	DDKEMVDKDVSEEKSVDEDVGKRDAGDDDMASVVDDASVDEVAPREVSGERDRDKDEDED
27	KRRAVKKKVAKRKSSRKKVVKKRGARKKSVSVDEDVPEDDVVDDDGSDDDVSDDEAVDEE
28	EKDAVDEKVADKKSSKDKVVDDRGAKRDSVSVDRDVPDEKVVRDKGSDREVSEKKAVKRD
30	SESKGDSASVGDSESKVADKKKDKVVDDRAKRDVVDRDVPDEKVVRDKDREVEKKAVKRD
31	KRRKRKRVRKRRVKRKKAKRRRAKRRKVKKKRSAAVGEAVAADSAVADAVVVDSADDDDD
32	KRKVVDRDAARRRVSDRKAAKDRVGKRKAVAARKKSARKDVADEKAVRDRVVKRKSADRK

Table S12 Sequences that show inverse response (expand) under hyper-osmotic shock (750 mOsm)

Sequence ID	Sequence
9	KKRSGKRRHHKKRHQKRRQHRKKQQRKKQNQNNNNNQNQQQQHHNQQNNQNNPNNNNHHP
17	KRRNQKRRGTRKRSGKRKGSRKRGARRKSTATATGSTTSSATASASSSSSSSGAGATGGS
18	KRRKNQGTKSGGSKGASTRATATRGSTTKSSATASASRSSSSRSSGARGATGKGSRRKRR

SUPPLEMENTARY FIGURES

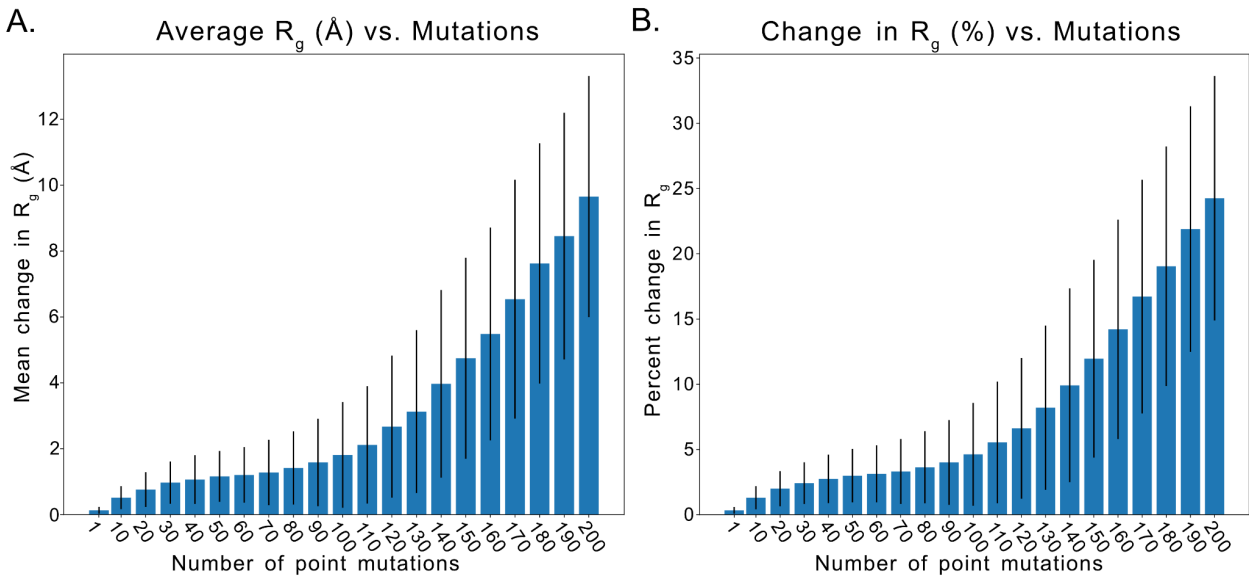

Figure S1. Global dimensions for disordered regions are relatively insensitive to individual point mutations. To assess IDR sensitivity to point mutations, we compared changes in the predicted radius of gyration (R_g) for 2000 randomly generated 200-residue disordered sequences in response to specific numbers of mutations. Specifically, for each sequence, we determine how the radius of gyration changes in response to 1, 10, 20, ..., 200 individual point mutations. Radii of gyration are predicted using ALBATROSS[9]. In general, 1-10 mutations lead to relatively small changes in the overall dimensions. **(A)** The average change in R_g as compared to the starting sequence. Error bars show the standard deviation of the change in Å. **(B)** Percentage change in R_g from the starting sequence. Error bars show the standard deviation of the change by % difference.

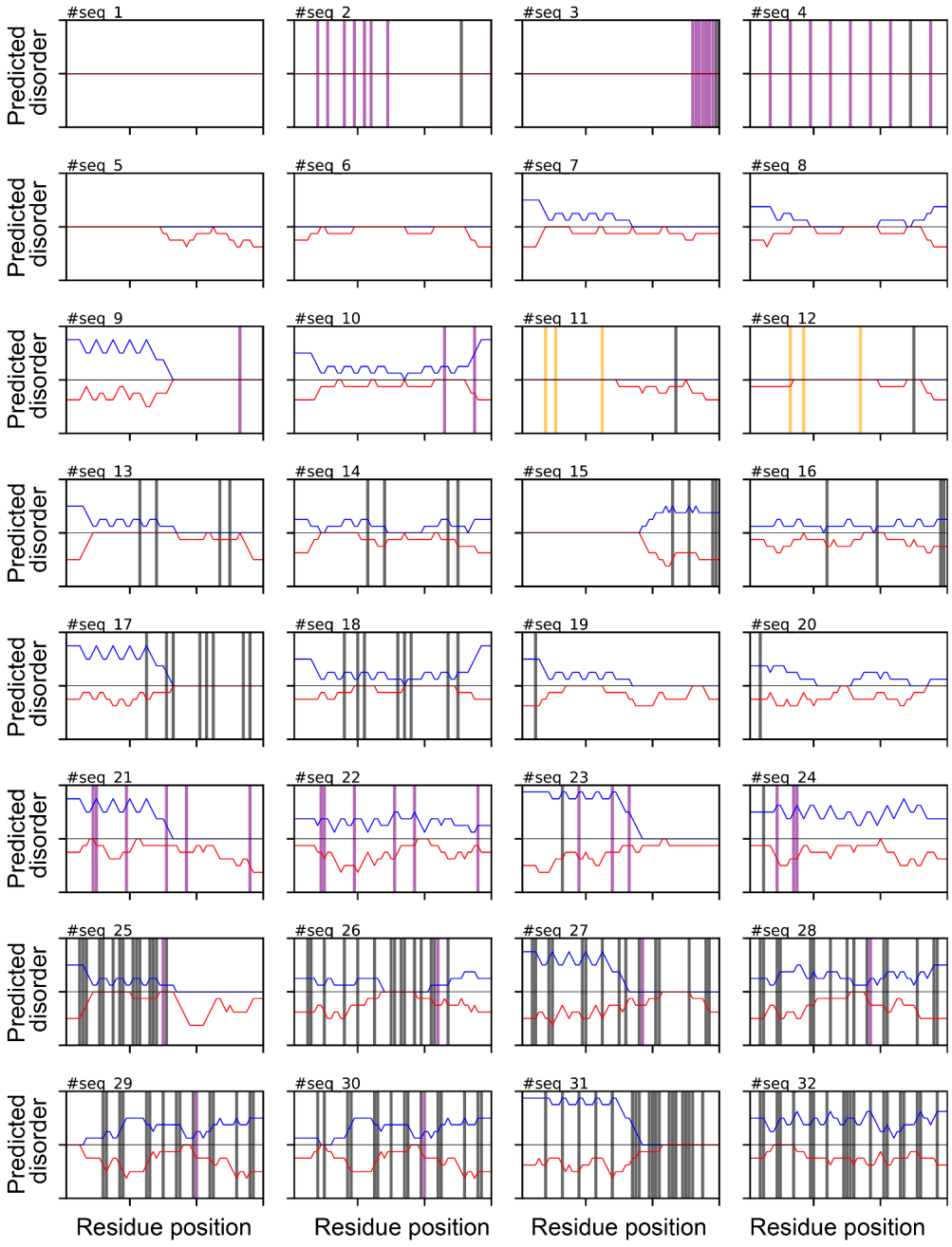

Figure S2. Overview of designed IDR library. The amino acid composition of all sequences is shown in terms of per-residue local chemistry. Red (negative) and Blue (positive) lines track local smoothed charge profiles for negatively charged (E/D) and positively charged (R/K) residues using a window size of 15 residues. Purple bars report on the location of proline residues, orange bars on the location of aromatic (Y/F/W) residues, and back bars on the location of aliphatic (I/L/V/M/A) residues These sequences are also provided in **Table S1**.

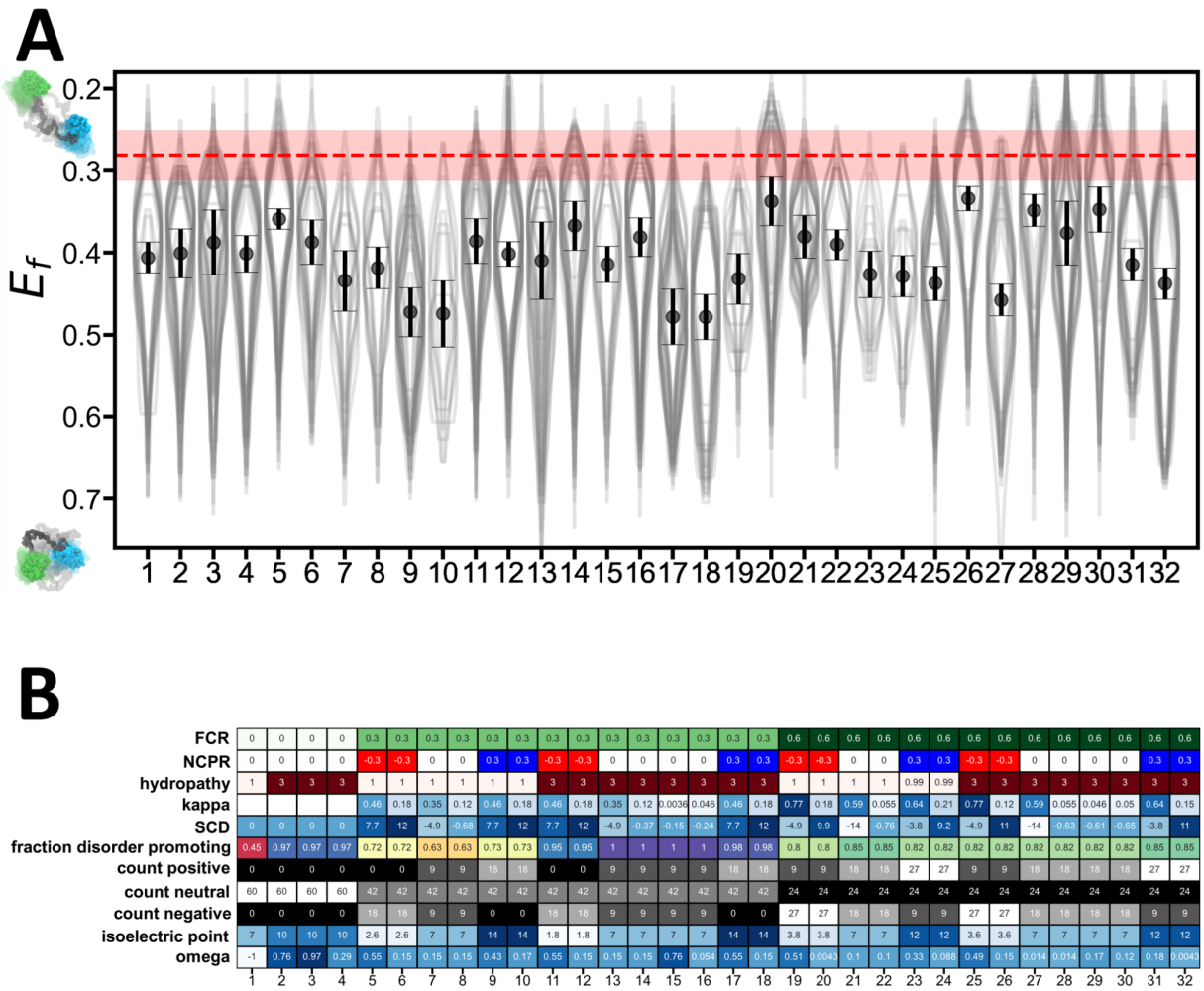

Figure S3. Summarized in-cell FRET data for the GOOSE library. (A) FRET efficiencies (E_f) of all constructs used in this work measured in U2-OS cells. Each violin outline represents the data distribution of one repeat, containing at least 60 cells. Circles represent the average of the medians of all violins, and the error bars represent the standard deviation of all the medians. The red line and shaded region represent the median and the median 50% of E_f for a glycine-serine repeat $(GS)_{32}$. **(B)** Sequence features obtained from localcider[14]. FCR is the fraction of charged residues. NCPR is the net charge per residue. Hydropathy describes the mean hydropathy calculated from the Kyte-Doolittle hydrophobicity scale[25]. Kappa describes the charge distribution[15]. SCD is the sequence charge decoration[26]. Fraction disorder promoting describes the sequence's fraction of residues which are considered disorder promoting[27]. Omega describes the patterning between charged/proline residues and all other residues[28].

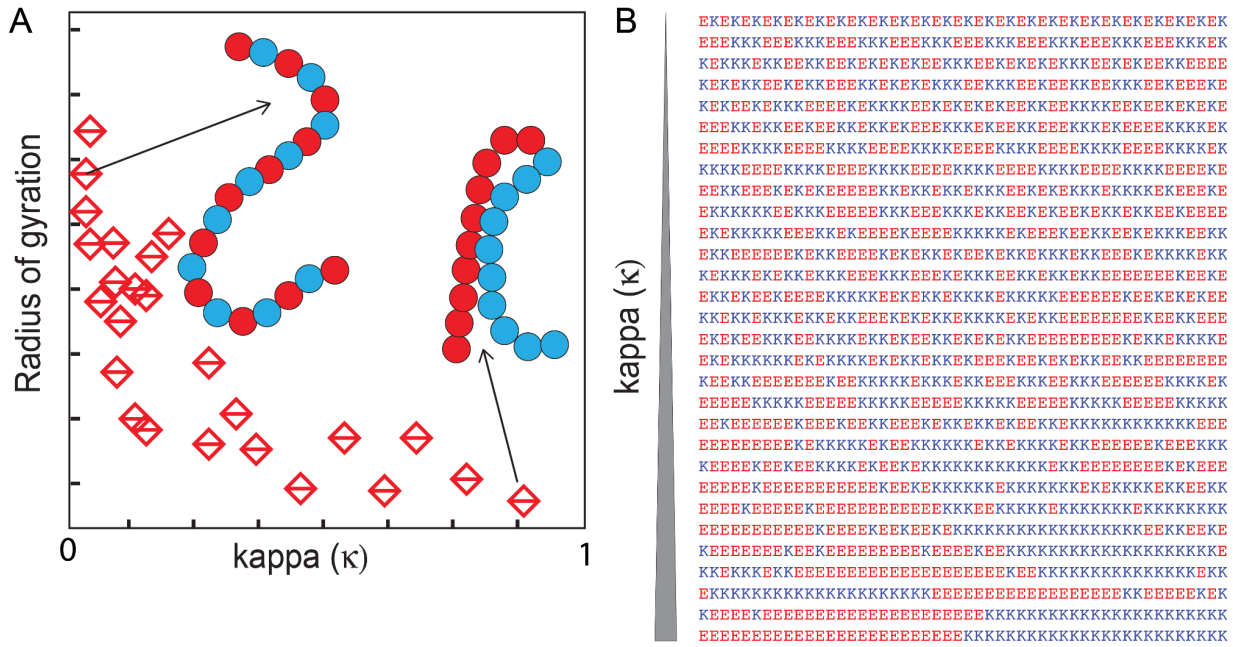

Fig. S4. Charge patterning is important for IDRs. Charge patterning can be quantified by kappa (κ), a parameter that quantifies the difference in local charge polarity compared to the overall average of the sequence, normalized by the most segregated possible sequence. **(A)** Schematized reproduction of the original dependence of the radius of gyration (R_g) on κ as described by Das & Pappu, as shown for a set of thirty strong polyampholytic sequences with the same composition but different charge patterning[15]. **(B)** Sequences examined in panel A are shown in order of κ value, illustrating how increasing κ relates to the patterning of oppositely charged residues.

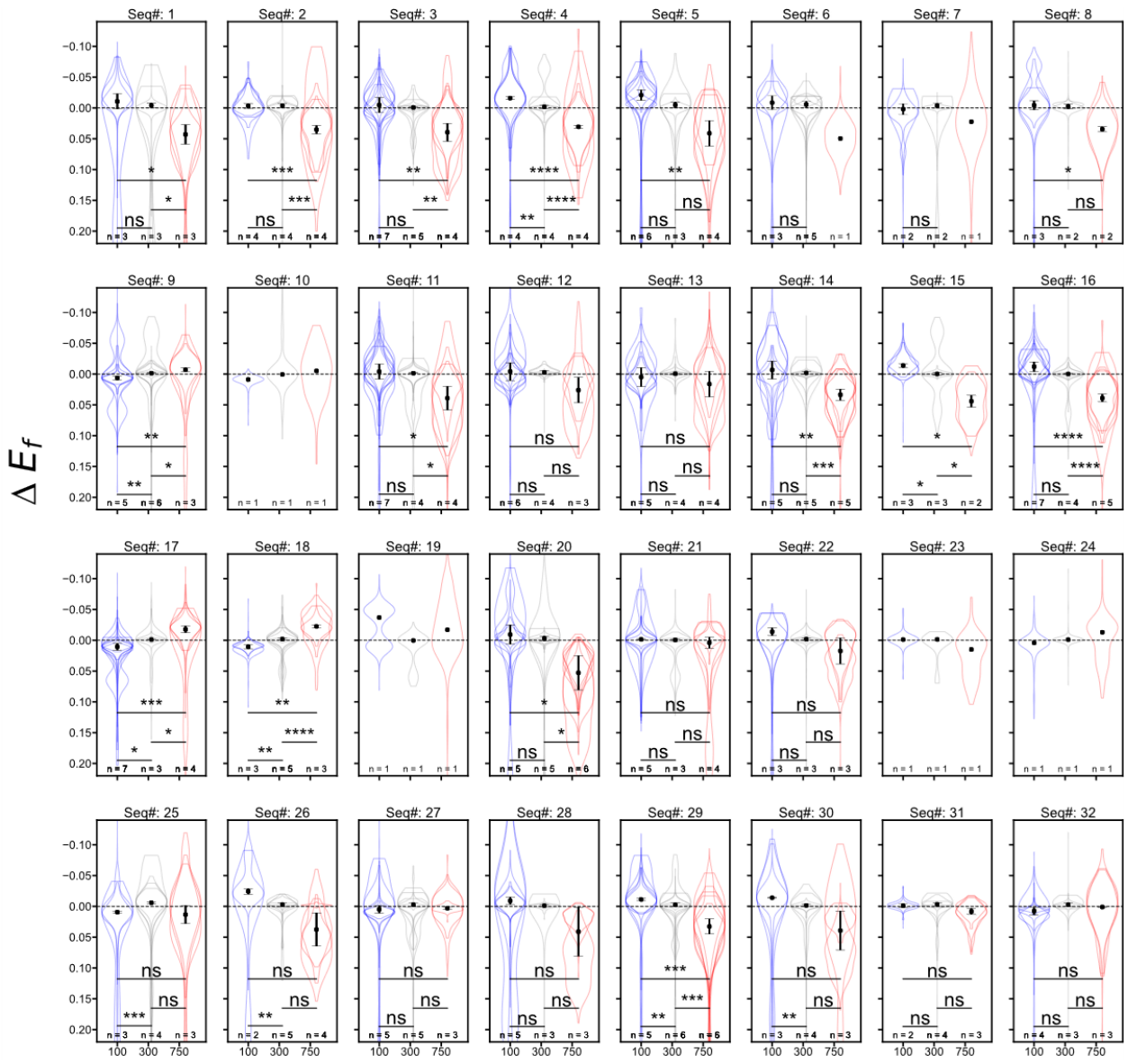

Figure S5. Sequences show variable responses to changes in cell volume. Change in FRET efficiencies following osmotic challenge ($\Delta E_f = E_f^{after} - E_f^{after}$) measured in U2-OS cells. The x-axis reports the final osmotic pressure following the challenge, reported in mOsm. Each violin outline represents the data distribution of one repeat for hypo (blue), iso (grey), and hyper (red) conditions and contains at least 60 cells. The circle represents the average of all medians for that construct, and the error bars represent the standard deviation of the medians. P-values were determined by Student's t-test where N's were sufficiently high (**** = P < 0.00001, *** = P < 0.0001, ** = P < 0.001, * = P < 0.01, ns = not significant). Sequences 10,19, 23, and 24 are excluded from the analysis in which change in FRET upon the change in cell volume is assessed. Furthermore, for sequences at 750 mOsm, we also excluded sequences 6 and 7 due to insufficient statistics.

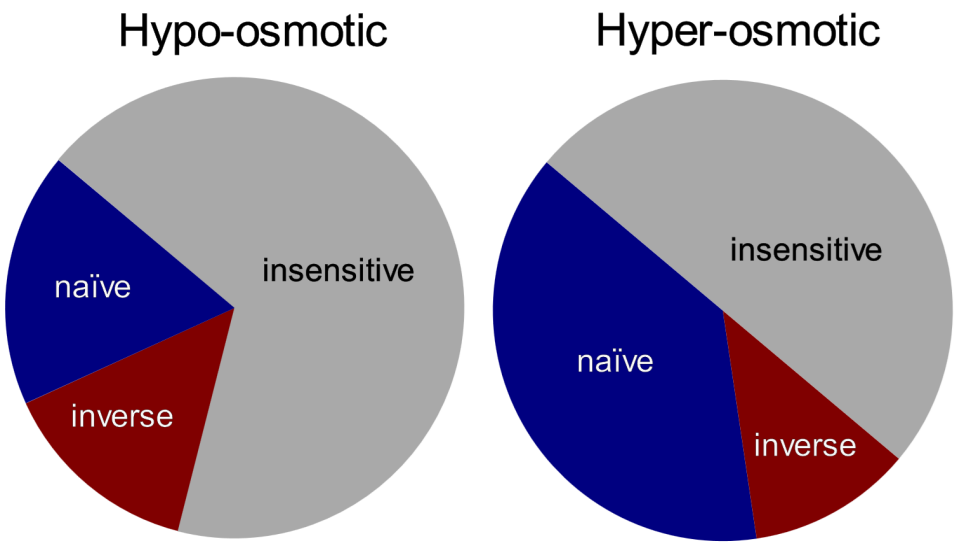

Figure S6. Pie charts summarizing Figure S5. Sequences were sorted into the following categories: naïve, insensitive, and inverse in response to hypo-osmotic (cell volume increase) and hyper-osmotic (cell volume decrease). The total number of sequences categorized under hypo-osmotic conditions as naïve, insensitive, and inverse were 5, 19, and 4 total sequences, respectively. The total number of sequences categorized under hyper-osmotic conditions as naïve, insensitive, and inverse were 10, 13, and 3, respectively. Specific sequence details for the categorized sequences are shown in **Tables S7 - S12**.

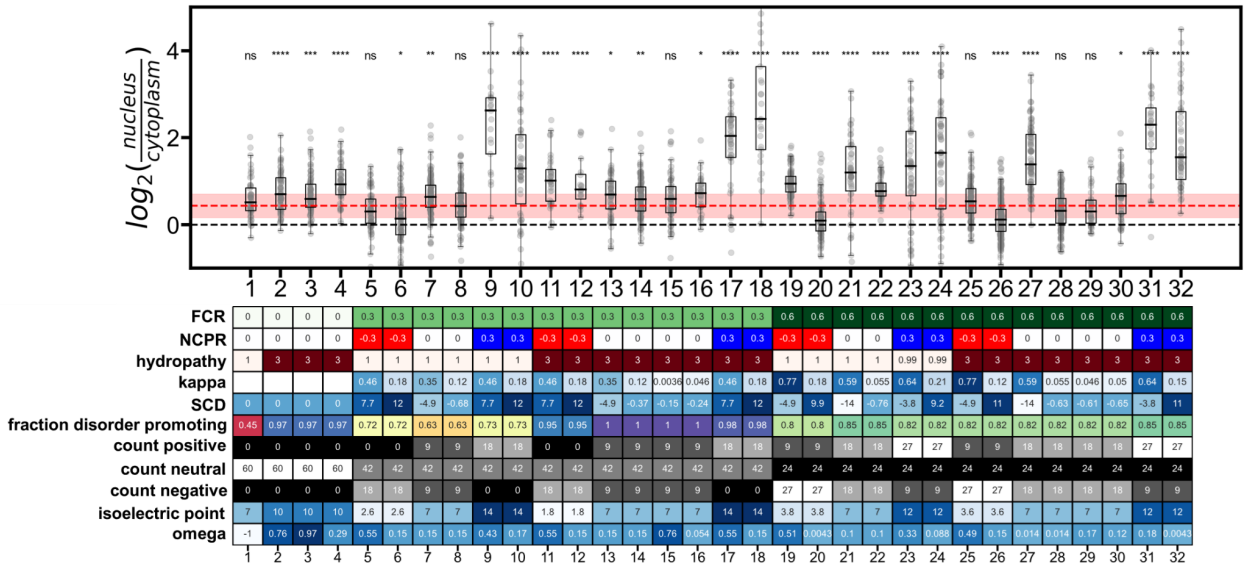

Figure S7. IDRs show sequence-specific subcellular localization preferences. Log fold change of the acceptor's fluorescence intensities between nucleus and cytoplasm ($log_2\left(\frac{nucleus}{cytoplasm}\right)$). U2-OS cells were imaged at 20x, and regions in the nucleus and cytoplasm for each cell were segmented and measured. Individual cells are shown as points, each box represents the 25th and 75th percentiles of the data, the whiskers show the minimum and maximum for each construct, and the median is shown as a black line. Box plots contain N > 20. Statistical significance is determined by a double-sided t-test against the subcellular localization ratio of $(GS)_{32}$ shown as the red dashed line. The median 50 for $(GS)_{32}$ is shown by the red shaded region. (**** = P < 0.00001, *** = P < 0.0001, ** = P < 0.001, * = P < 0.01, ns = not significant) (**see also Fig. S16**).

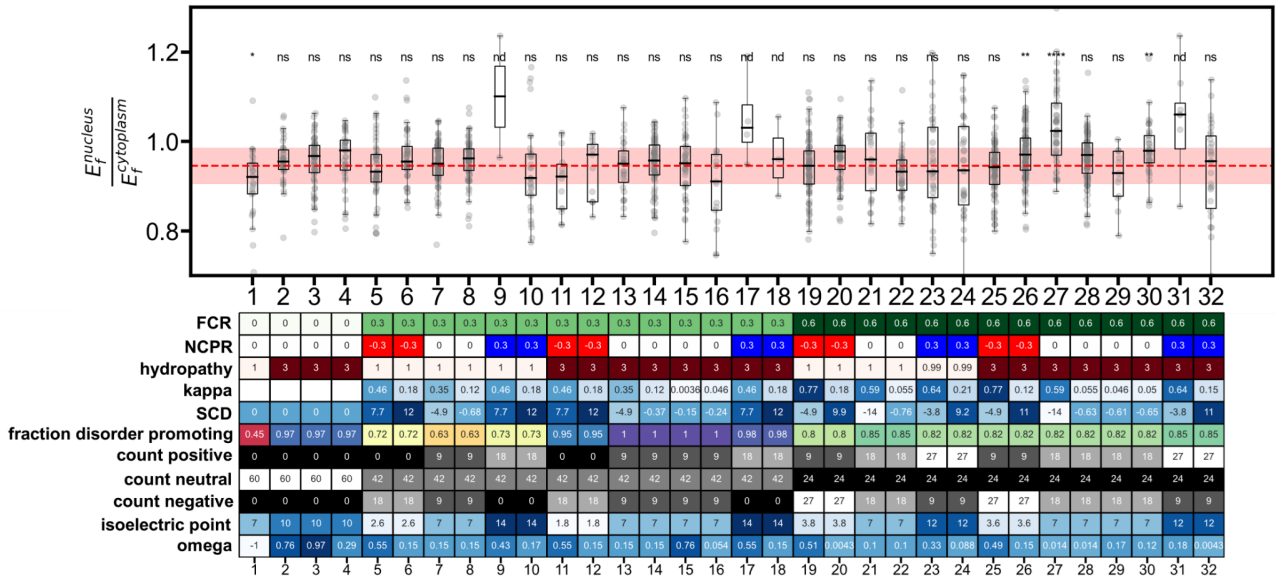

Figure S8. The ratio between nuclear and cytoplasmic E_f measurements. Figure features are as in **Fig. S7**, except the red line represents the corresponding E_f value of $(GS)_{32}$ shown as the red dashed line with the median 50 shown as the shaded region.

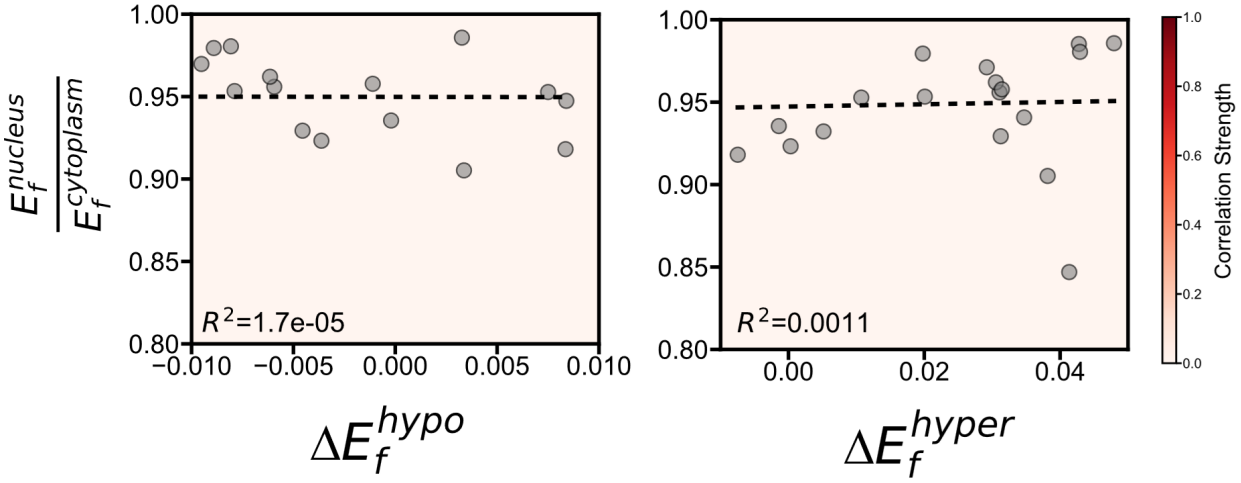

Figure S9. The nucleo-cytoplasmic FRET ratio shows no strong correlation with the change in FRET upon hypo- or hyper-osmotic conditions. Correlations between changes in cell volume (ΔE_f) with subcellular FRET measurements for hypo-osmotic (left) and hyper-osmotic (right) conditions. The Pearson's R^2 value is shown on the bottom right of each panel.

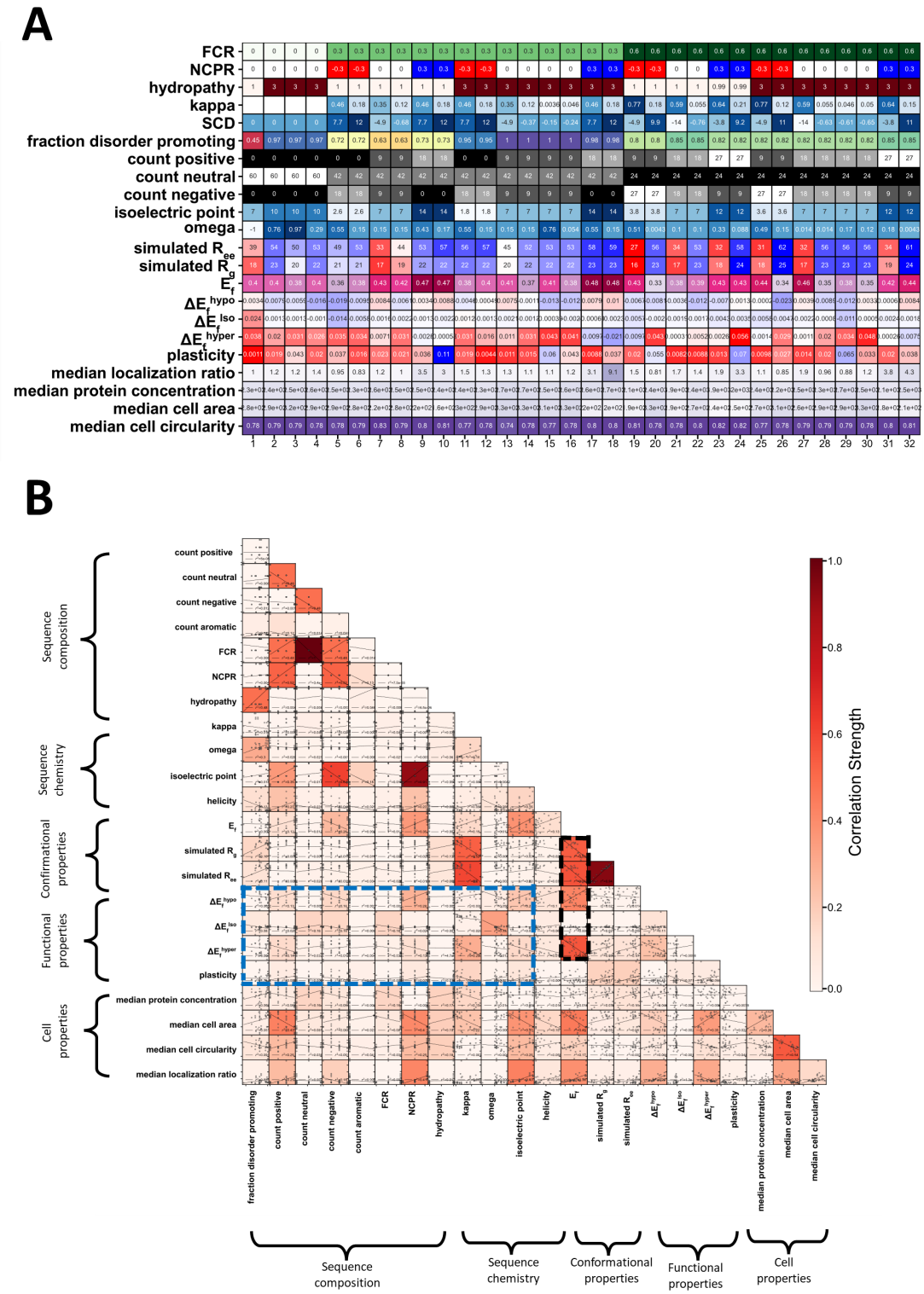

Figure S10. (A) Heatmap summarizing all sequence composition, sequence chemistry, conformational properties, functional properties, and cell properties for the complete library. **(B)** Ensemble dimension and sequence parameter correlations. The background of each plot corresponds to the correlation strength determined by Pearson's R^2.

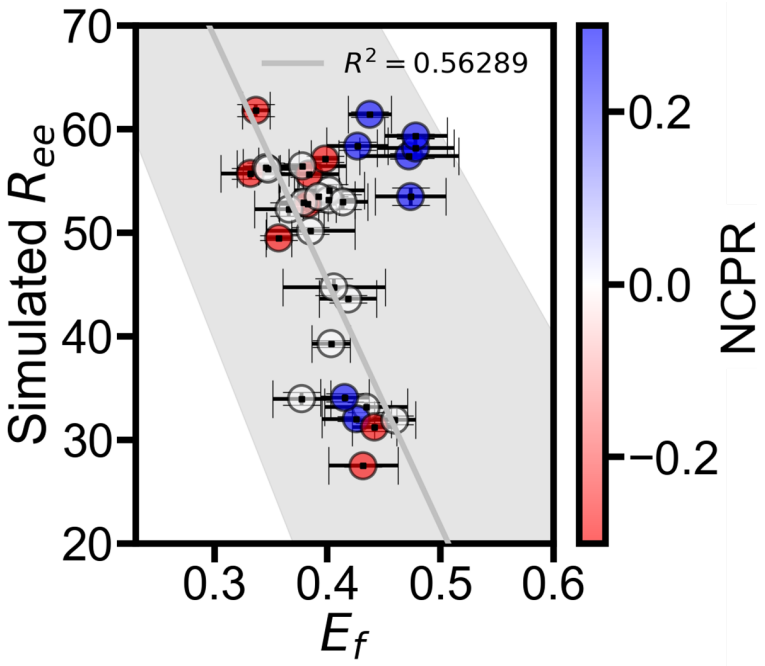

Figure S11. Linear fit of live cell FRET efficiencies (E_f) vs. the simulated end-to-end distance (R_{ee}) obtained from coarse-grained molecular dynamics simulations performed with the Mpipi forcefield[10]. Each scatter point is labeled with the sequence number used throughout the text. Simulation error bars are calculated as the standard error of the mean across three independent replicas. Experimental error bars are calculated as the standard deviation of the medians (see **Fig. S17**). The six major outliers (sequences #9, #10, #17, #18, #24, #32) are all highly positively charged (blue points) and show a higher basal FRET value, indicating they are more compact in cells than predicted by simulations. Outliers were not included in calculating the correlation coefficient.

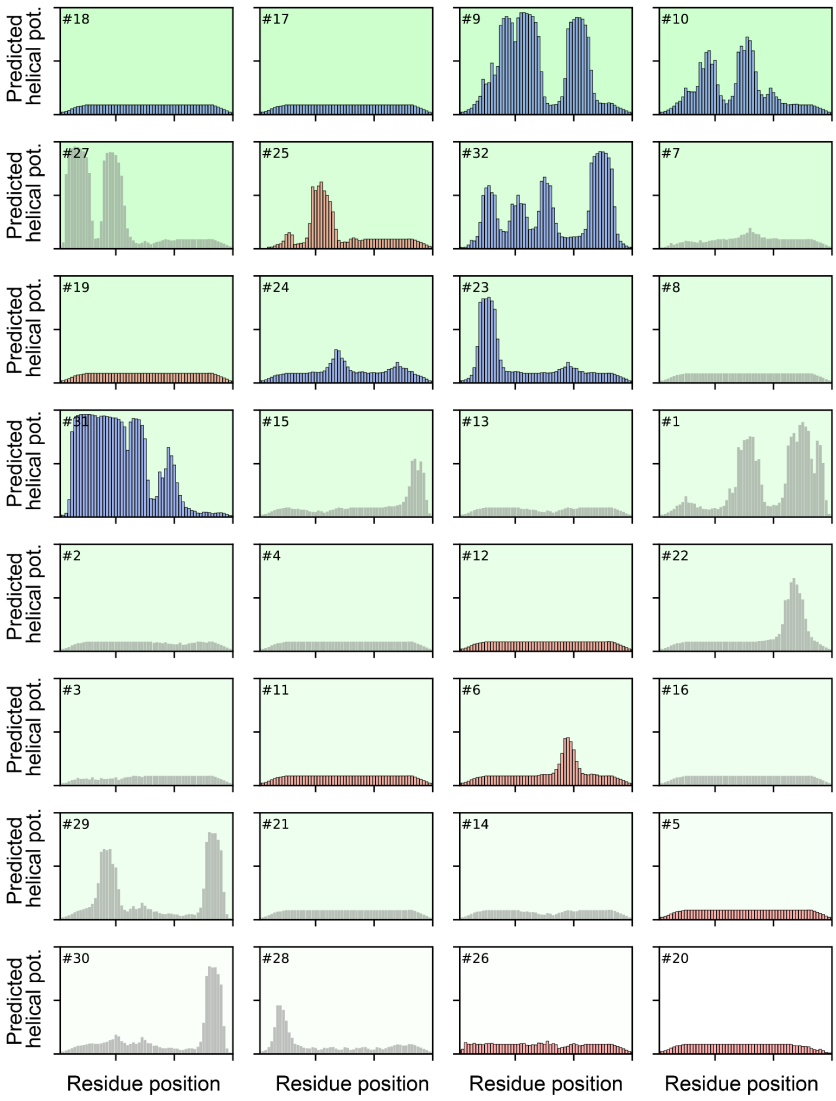

Figure S12. Predicted helicity potential for each sequence. We predicted per-residue helicity for each position using JPred4[16]. Despite several sequences possessing local helicity, all sequences are strongly predicted to be disordered (**Fig. S19**). Bar colors reflect sequence net charge (blue = positive, red = negative, grey = neutral), and the background color on each panel reflects the basal FRET efficiency. Sequences are rank-ordered by basal FRET efficiency (top-left to right, snaking around), such that the top left is the most compact and the bottom right is the most expanded. Predicted transient helicity does not explain compaction in positively charged proteins. In the top twelve most compact sequences, 50% possess none or minimal predicted helicity, while several are predicted to be more helical. Moreover, in many specific pair comparisons, a change in predicted helicity has no impact on dimensions (e.g., **Fig. 3B**: #10 vs. #18 and #9 vs. #17) or loss of helicity leads to compaction instead of expansion (e.g., **Fig. 2C**: #18 vs. #32). Taken together, while conclusive evidence would require systematic biophysical characterization of each IDR in the context of our fluorescence proteins, we see no evidence to support a model in which secondary structure is a major determinant of IDR global dimensions.

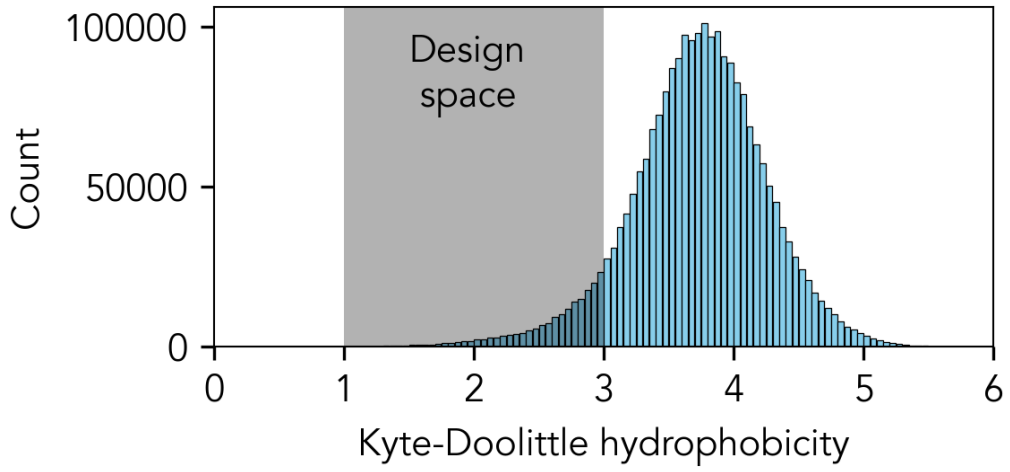

Figure S13. Proteome-wide hydrophobicity analysis. All IDRs longer than 60 amino acids were segmented into overlapping windows with a stepsize of 1, and the hydrophobicity within each window was calculated. The resulting histogram is plotted in blue. Designed sequences fall within the shaded region.

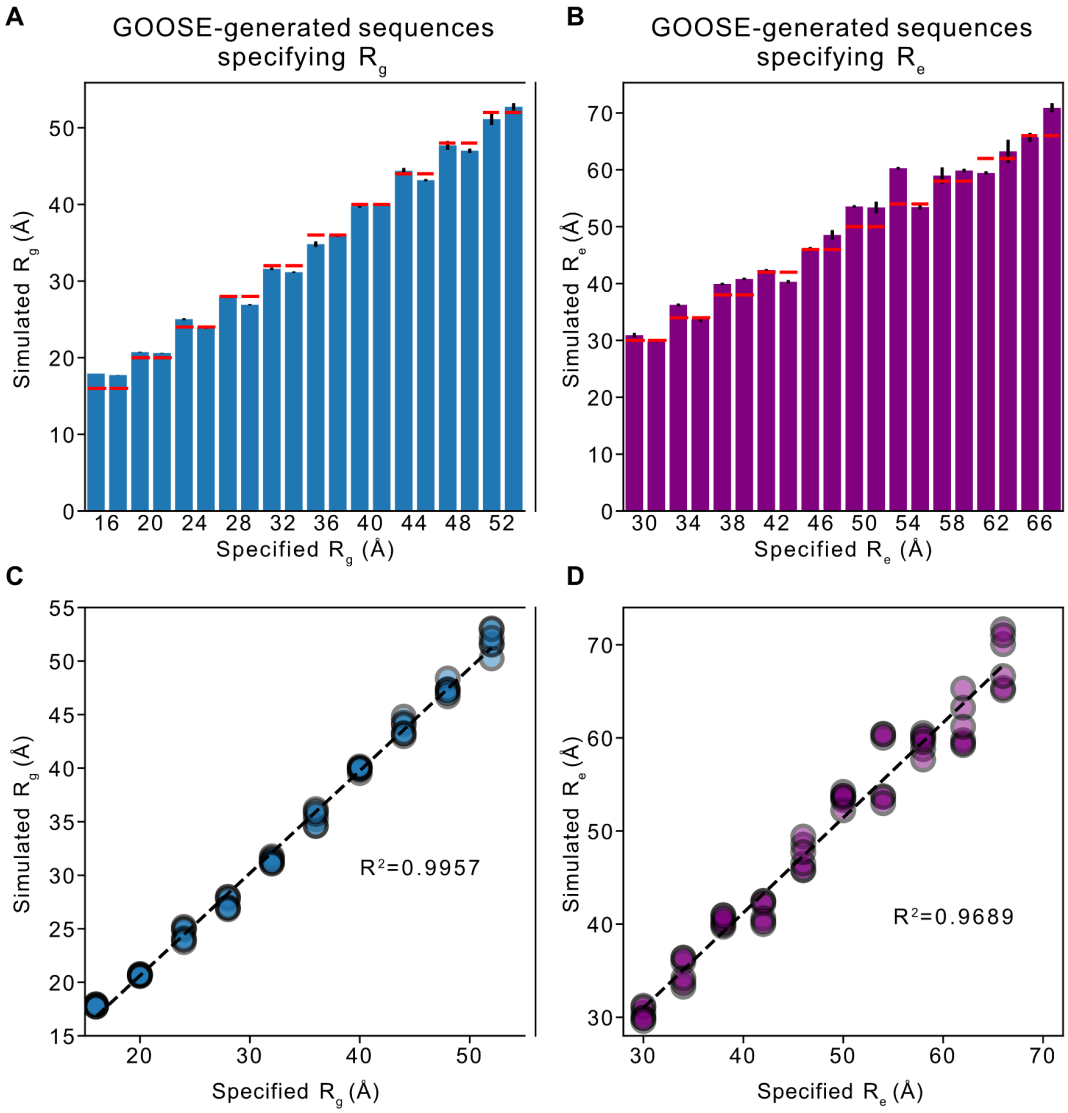

Figure S14. GOOSE can make sequences with specified length and radius of gyration (R_g) or end-to-end distance (R_e). All sequences generated were 200 amino acids in length. For R_g (A, C), two sequences with dimensions between 16 Å and 52 Å at intervals of 4Å (20 sequences total) were generated. A similar approach was used for specifying R_e, except a range of between 30Å and 66Å was used. After sequence generation, coarse-grain molecular dynamics simulations were run as described in the *Methods*. For bar plots (A, C), bars are equal to the mean of the average R_g or R_e of the triplicate for each sequence, error bars are the standard deviation between the means for each triplicate, and the x-axis labels denote the R_g or R_e specified for each sequence (two sequences per specified dimension). Red lines show the specified R_g or R_e for the sequence during sequence generation. Each point on the scatter plots (B, D) shows the average dimension for each simulation triplicate for both sequences for the desired R_g or R_e value (y-axis) with the specified dimension during sequence generation on the x-axis. The R^2 values were calculated using the mean value of the triplicate for each sequence vs. the specified R_g or R_e during sequence generation for each sequence.

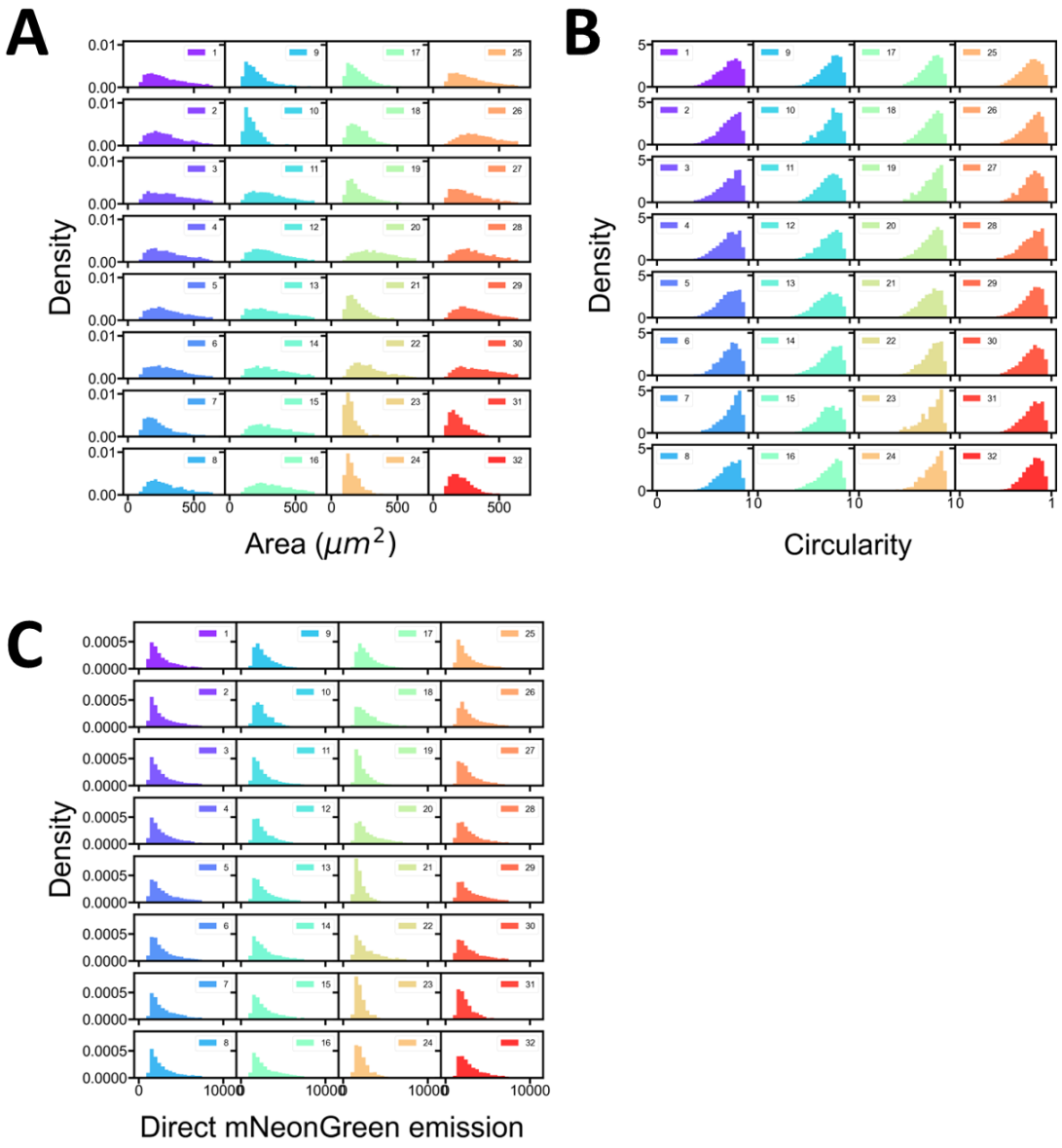

Figure S15. Histograms of the cell properties analyzed for this work for each of the 32 library constructs. (A) Cell area, measured following segmentation. (B) Cell circularity, measured from the area of each cell. A circularity value of 1 is a perfect circle. (**C**) Relative expression levels of the FRET construct in each cell, as assessed by mNeonGreen emission under mNeonGreen excitation.

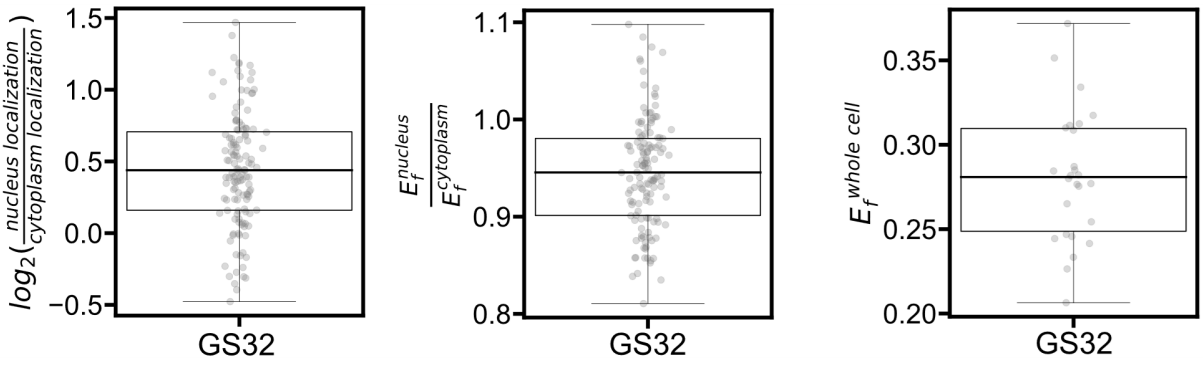

Figure S16. Localization and ensemble features for $(GS)_{32}$ reference in U2-OS cells. Glycine-serine repeat ($(GS)_{32}$) used for comparison. Boxplot features are as in **Fig. S7.** Points correspond to individual cells. Box plots contain N = 132 for the first and second box plots and N=26 for the last box plot.

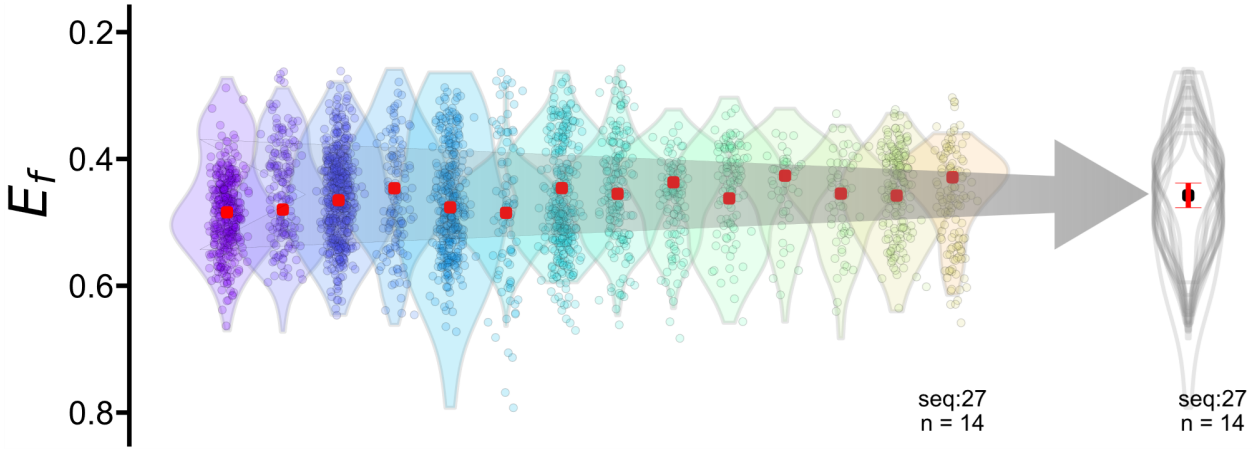

Figure S17. Visual explanation reporting how each violin plot was generated before performing statistical tests. Experiments were done on 96-well plates, and each well was considered one separate transfection (each colored violin here represents one well). Wells containing less than 60 cells were not included in the analysis. For each synthetic IDR sequence, the average and standard deviation of the medians from each well were used to obtain the average E_f and standard deviation of that specific IDR sequence. Student's t-test was performed between the calculated medians of the IDR sequences being compared (red points for both groups).

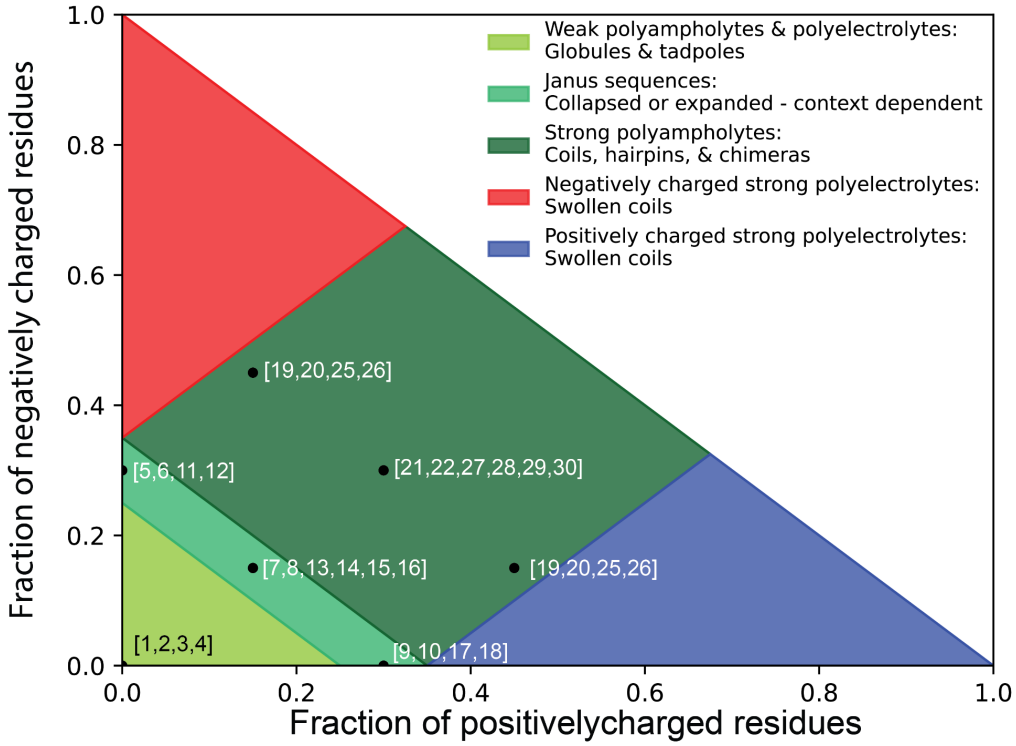

Figure S18. Sequences explore a broad set of sequence space. All sequences are placed on a Das-Pappu diagram of states.

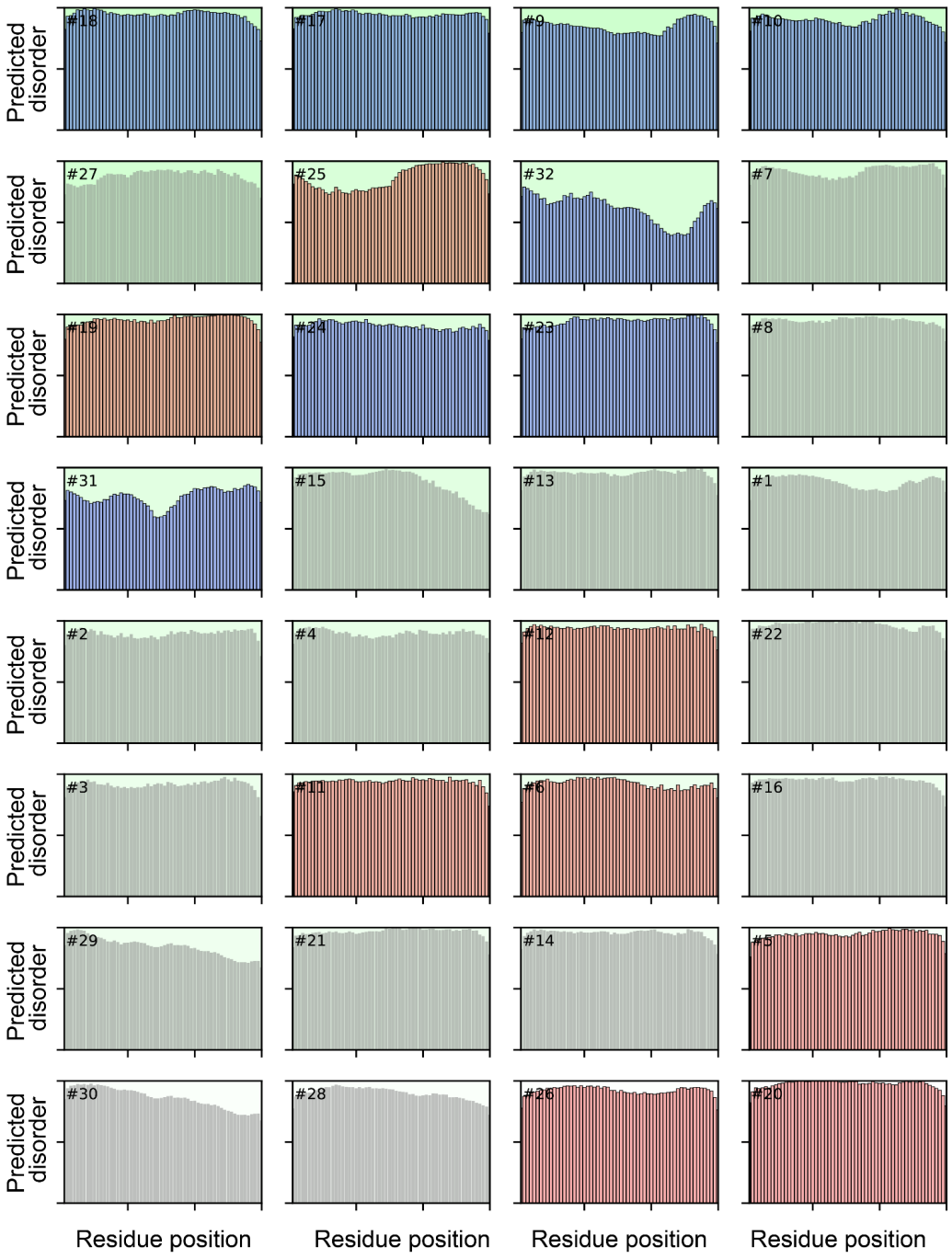

Figure S19. Predicted per-residue disorder scores using metapredict. Bar colors reflect sequence net charge (blue = positive, red = negative, grey = neutral), and the background color on each panel reflects the basal FRET efficiency. Sequences are rank-ordered by basal EFRET (top-left to right, snaking around). All sequences are strongly predicted to be fully disordered.

SUPPLEMENTARY REFERENCES

1. Moses, D., Guadalupe, K., Yu, F., Flores, E., Perez, A., McAnelly, R., Shamoon, N. M., Kaur, G., Cuevas-Zepeda, E., Merg, A. D., Martin, E. W., Holehouse, A. S. & Sukenik, S. Structural biases in disordered proteins are prevalent in the cell. *Nat. Struct. Mol. Biol.* **In Press,** (2023).
2. Abramoff, M. D., Magelhaes, P. J. & Ram, S. J. Image Processing with ImageJ. *Biophotonics International* **11,** 36–42 (2004).
3. 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).
4. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17,** 261–272 (2020).
5. Boeynaems, S., Rosa Ma, X., Yeong, V., Ginell, G. M., Chen, J.-H., Blum, J. A., Nakayama, L., Sanyal, A., Briner, A., Van Haver, D., Pauwels, J., Ekman, A., Broder Schmidt, H., Sundararajan, K., Porta, L., Lasker, K., Larabell, C., Hayashi, M. A. F., Kundaje, A., Impens, F., Obermeyer, A., Holehouse, A. S. & Gitler, A. D. Aberrant phase separation is a common killing strategy of positively charged peptides in biology and human disease. *bioRxiv* 2023.03.09.531820 (2023). doi:10.1101/2023.03.09.531820
6. Zeng, X., Ruff, K. M. & Pappu, R. V. Competing interactions give rise to two-state behavior and switch-like transitions in charge-rich intrinsically disordered proteins. *Proc. Natl. Acad.*

Sci. U. S. A. **119,** e2200559119 (2022).

7. Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered regions. *Nat. Rev. Mol. Cell Biol.* **(in press),** (2023).
8. Moses, D., Guadalupe, K., Yu, F., Flores, E., Perez, A., McAnelly, R., Shamoon, N. M., Cuevas-Zepeda, E., Merg, A. D., Martin, E. W., Holehouse, A. S. & Sukenik, S. Structural biases in disordered proteins are prevalent in the cell. *bioRxiv* 2021.11.24.469609 (2022). doi:10.1101/2021.11.24.469609
9. Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J. & Holehouse, A. S. Direct Prediction of Intrinsically Disordered Protein Conformational Properties From Sequence. *Nature Methods* **(in press),** (2023).
10. Joseph, J. A., Reinhardt, A., Aguirre, A., Chew, P. Y., Russell, K. O., Espinosa, J. R., Garaizar, A. & Collepardo-Guevara, R. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat Comput Sci* **1,** 732–743 (2021).
11. Conte, A. D., Mehdiabadi, M., Bouhraoua, A., Miguel Monzon, A., Tosatto, S. C. E. & Piovesan, D. Critical assessment of protein intrinsic disorder prediction (CAID) - Results of round 2. *Proteins* (2023). doi:10.1002/prot.26582
12. Emenecker, R. J., Griffith, D. & Holehouse, A. S. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys. J.* **120,** 4312–4319 (2021).
13. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. Array programming with NumPy. *Nature* **585,** 357–362 (2020).
14. 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).

15. 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).
16. Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43,** W389–94 (2015).
17. Acids Research, N. & 2023. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* **51,** D523–D531 (2023).
18. Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M., Crozier, P. S., in 't Veld, P. J., Kohlmeyer, A., Moore, S. G., Nguyen, T. D., Shan, R., Stevens, M. J., Tranchida, J., Trott, C. & Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271,** 108171 (2022).
19. McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L.-P., Lane, T. J. & Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109,** 1528–1532 (2015).
20. 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. *J. Chem. Theory Comput.* **19,** 5609–5620 (2023).
21. 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
22. Pesce, F., Bremer, A., Tesei, G., Hopkins, J. B., Grace, C. R., Mittag, T. & Lindorff-Larsen, K. Design of intrinsically disordered protein variants with diverse structural properties.

bioRxiv 2023.10.22.563461 (2023). doi:10.1101/2023.10.22.563461

23. Harmon, T. S., Crabtree, M. D., Shammas, S. L., Posey, A. E., Clarke, J. & Pappu, R. V. GADIS: Algorithm for designing sequences to achieve target secondary structure profiles of intrinsically disordered proteins. *Protein Eng. Des. Sel.* **29,** 339–346 (2016).
24. Schramm, A., Lieutaud, P., Gianni, S., Longhi, S. & Bignon, C. InSiDDe: A Server for Designing Artificial Disordered Proteins. *Int. J. Mol. Sci.* **19,** (2017).
25. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132 (1982).
26. Sawle, L. & Ghosh, K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J. Chem. Phys.* **143,** 085101 (2015).
27. Campen, A., Williams, R. M., Brown, C. J., Meng, J., Uversky, V. N. & Dunker, A. K. TOP-IDP-scale: A new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept. Lett.* **15,** 956–963 (2008).
28. Martin, E. W., Holehouse, A. S., Grace, C. R., Hughes, A., Pappu, R. V. & Mittag, T. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J. Am. Chem. Soc.* **138,** 15323–15335 (2016).