Science Advances
AAAS

# Supplementary Materials for

## Design of intrinsically disordered protein variants with diverse structural properties

Francesco Pesce *et al.*

Corresponding author: Francesco Pesce, francesco.pesce@bio.ku.dk; Kresten Lindorff-Larsen, lindorff@bio.ku.dk

*Sci. Adv.* **10**, eadm9926 (2024)
DOI: 10.1126/sciadv.adm9926

**This PDF file includes:**

Supplementary Materials and Methods
Figs. S1 to S14
Tables S1 to S3
References

# Supplementary experimental materials and methods

## Protein constructs

Sequences of wild-type A1-LCD and variants are based on the low complexity domain (residues 186-320) of the human hnRNPA1 (UniProt: P09651; Isoform A1-A). The coding sequences for the variants were synthesized (Thermo Fisher) including a coding sequence for an N-terminal ENLYFQGS TEV protease cleavage site and 5' and 3' attB sites for Gateway cloning. The sequences were recombined via LR reactions into the pDEST17 vector (Thermo Fisher), which includes an N-terminal 6xHis-tag coding sequence. After expression, we cleaved of the N-terminal 6xHis-tag using TEV protease, leaving only an additional GS sequence at the N-terminus (Table S1).

## Protein expression and purification

A1-LCD variants were expressed and purified as previously reported for similar constructs (*39, 41*). The *E. coli* BL21 (DE3) pLysS strain was used for expression and grown in ZYM5052 auto induction media at 37°C for 24 hours. Cell pellets were recovered by centrifugation and resuspended in 50 mM MES pH 6.0, 500 mM NaCl, 20 mM 2-mercaptoethanol. Cell lysis was achieved via sonication. Cell lysates were centrifuged to collect inclusion bodies, that were resuspended in 6 M GdmHCl, 20 mM Tris pH 7.5, 15 mM imidazole overnight at 4°C. The solutions containing the solubilized inclusion bodies were cleared from cell debris by centrifugation, and supernatants were loaded onto self-packed columns of chelating Sepharose fast flow beads (GE Healthcare) charged with nickel sulfate. The columns were washed with four column volumes of 4 M urea, 20 mM Tris pH 7.5, 15 mM imidazole. Proteins were eluted from the Ni-NTA resin with 4 M urea, 20 mM Tris pH 7.5, 500 mM imidazole. TEV cleavage of the 6xHis-tag was done in 2 M urea, 20 mM Tris pH 7.5, 50 mM NaCl, 0.5 mM EDTA, 1 mM DTT overnight at 4°C. Cleaved protein solutions were loaded onto Ni-NTA columns. The flow-through and wash fractions were collected and concentrated using a 3000 MWCO Amicon centrifugal filter. Finally, samples were transfered in 2 M GdmHCl, 20 mM MES pH 5.5 over a S75 Superdex size exclusion column (GE Healthcare). The molecular weight of the proteins and the purity of samples were confirmed via intact mass spectrometry and SDS-PAGE. Samples were stored in 6 M GdmHCl, 20 mM MES pH 5.5 at 4°C.

We attempted to express and purify 15 variants of A1-LCD. Five of them (Table S1) expressed in *E. coli* with the protocol described above. The other ten (Table S2) expressed either at very low yield (X3) or with no detectable protein (X1–X2 and X4–X10).

## SDS-PAGE

Gel electrophoresis was carried out using NuPAGE 4–12% Bis-Tris gradient gels (Invitrogen). 1x NuPAGE MES SDS Running buffer (Invitrogen) was used to run gels. After the run, gels

were washed with water and stained with SimplyBlue SafeStain (Thermo Fisher Scientific) before destaining with water. PageRuler Plus Prestained protein ladder (Thermo Fisher Scientific) was used as a molecular weight reference.

### Buffer exchange

To remove the denaturant buffer used for storage and transfer the protein to 20 mM HEPES (pH 7.0) we used Zeba™ Spin Desalting Columns (Thermo Fisher Scientific) with 7k MWCO and 0.5 mL volume. After removal of storage solution from the column by centrifugation at 1000 $\times g$ for 1 min, columns were washed three times with 300 $\mu$L of 20 mM HEPES (by centrifugation at 1000 $\times g$ for 1 min). Finally, protein sample is applied to the column and recovered in 20 mM HEPES after a centrifugation. Additional washing steps (3–5) were carried out in Amicon Ultra-0.5 Centrifugal Filter Units to remove residual denaturant.

### Determination of saturation concentrations

Phase separation of protein samples was induced by adding NaCl to a final concentration of 150 mM. The dilute and dense phase were separated via centrifugation (*101*). The $c_{\text{sat}}$ was determined by the absorbance of the dilute phase at 280 nm.

### DIC microscopy

Differential interference contrast microscopy (DIC) images were obtained at room temperature using a Nikon Eclipse Ni Widefield microscope with a 20X objective. Samples were at concentrations slightly above their $c_{\text{sat}}$ at room temperature. Phase separation was induced by adding NaCl to the protein stock solution to reach a concentration of 150 mM. 2 $\mu$L of the protein solution were positioned in between two glass coverslips held toghether by 3M 300 LSE high-temperature double-sided tape (0.34 mm) with a window for microscopy cut out.

## Supplementary computational methods

The $R_{\text{h}}$ for protein conformations was calculated using HullRadSAS (*75, 102*). The ensemble-averaged $R_{\text{h}}$ was calculated as $1/n^{-1}\sum_i^n(1/R_{\text{h},i})$ (*103, 104*), from each conformer $i$ of an ensemble. Sequence clustering was performed with a 65% sequence identity threshold using the CD-HIT software (*105, 106*). Calculations of $\omega_{aro}$ and $\kappa$ from sequences were performed using the localCIDER python package (`https://github.com/Pappulab/localCIDER`), while the $z(\delta_{+-})$ scores for the IDRome sequences and the A1-LCD swap variants were calculated using a modified version of the NARDINI software which allowed us to define a custom threshold for the largest fraction of negatively and positively charged residues below which the program

sets $z(\delta_{+-})$ to zero (*65*). We set this threshold to 2.5% to obtain a non-zero $z(\delta_{+-})$ score for A1-LCD and sequences in the IDRome with fraction of charged residues similar to A1-LCD. For the NARDINI analysis of IDRome sequences, we generated $10^5$ randomly shuffled sequences, while for the wild type and variants of A1-LCD, we used $5\times10^5$ randomly shuffled sequences.

We calculated error bars on averages calculated from MD simulations using block averaging (`https://github.com/fpesceKU/BLOCKING`). Calculation of SAXS data from conformations was performed with Pepsi-SAXS (v3.0) (*107*), using fixed parameters for the contrast of the hydration layer and the effective atomic radius (respectively 3.34 e/nm$^3$ and 1.025 $\times$ $r_m$, where $r_m$ is the average atomic radius of the protein) (*73*). Prior to calculating the $\chi^2_r$, experimental SAXS curves are rebinned to 158 scattering angles and experimental error bars are rescaled using the Bayesian indirect Fourier transform (BIFT) (*108*). Both rebinning and error correction were carried out with the BayesApp webserver (`https://somo.chem.utk.edu/bayesapp/`) (*109*).

# Supplementary figures and tables

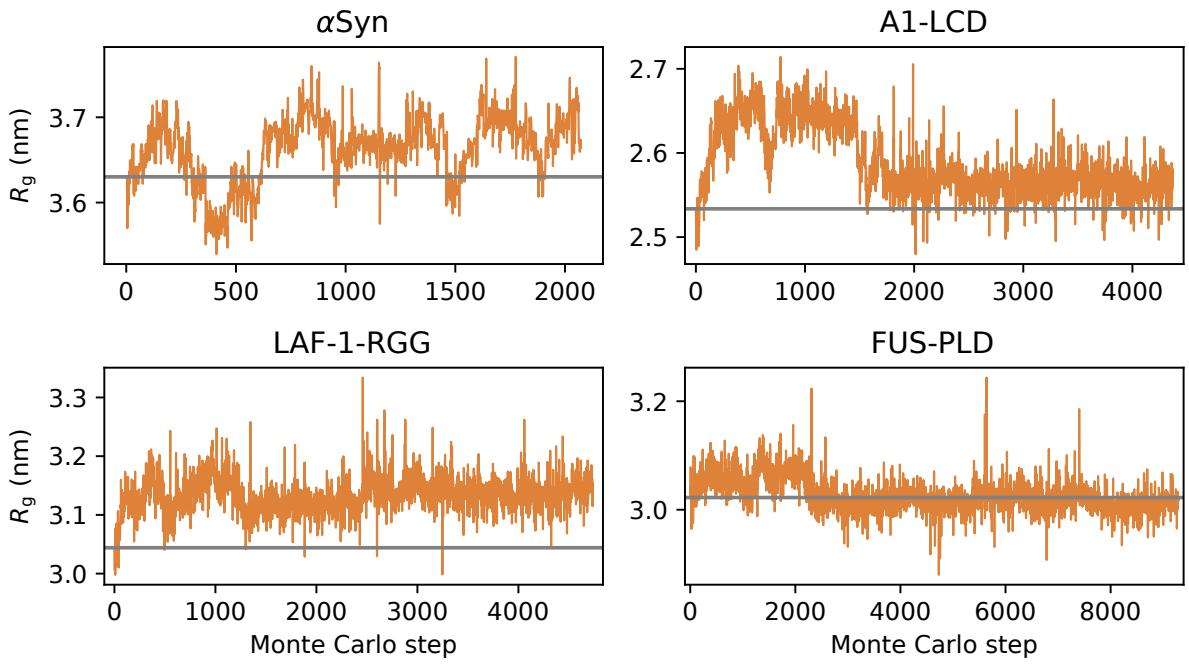


Figure S1: Design of more expanded variants for $\alpha$Syn, A1-LCD, LAF-1-RGG and FUS-PLD, starting from the wild-type sequences.

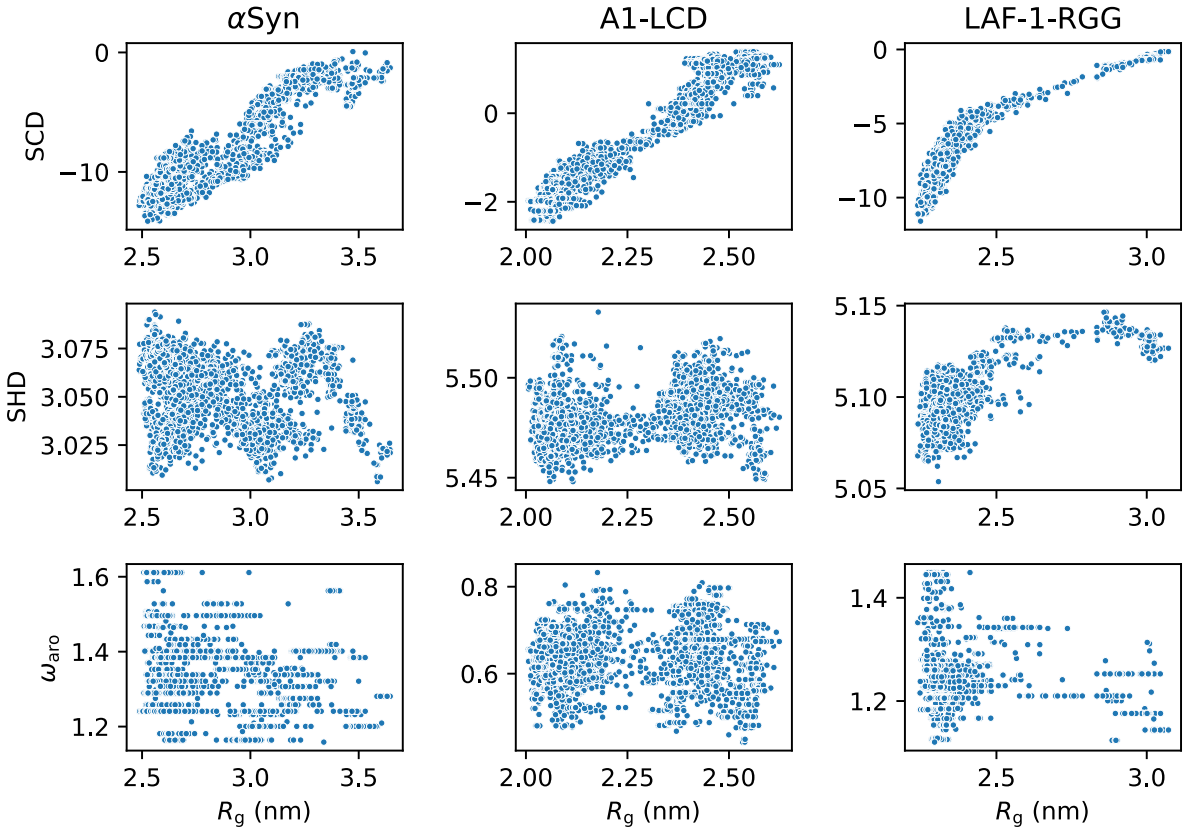


Figure S2: Multiple sequence features were calculated from the variant sequences of $\alpha$Syn, A1-LCD and LAF-1-RGG and correlated with the $R_g$. SCD, similarly to $\kappa$, is related to the patterning of charged residues. SHD (sequence hydropathy decoration) quantifies the patterning of hydrophobic residues. $\omega_{aro}$ quantifies the patterning of aromatic residues.

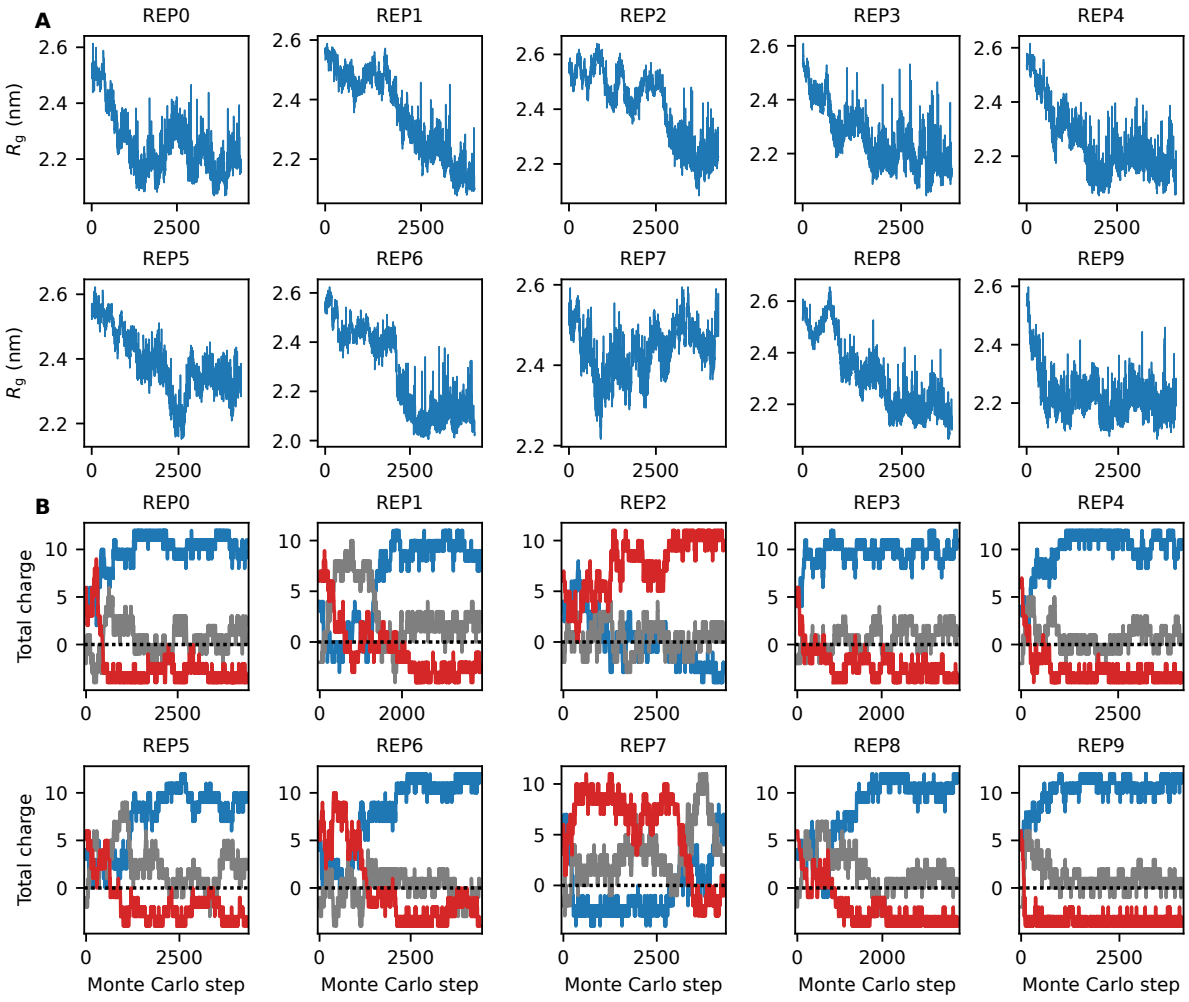


Figure S3: We performed ten runs for generating compact variants of A1-LCD. For each replica we show (a) the evolution of the $R_g$ from the generated sequences and (b) the total charge for the N-terminal third (blue), the middle third (grey), and the C-terminal third (red) of each sequence.

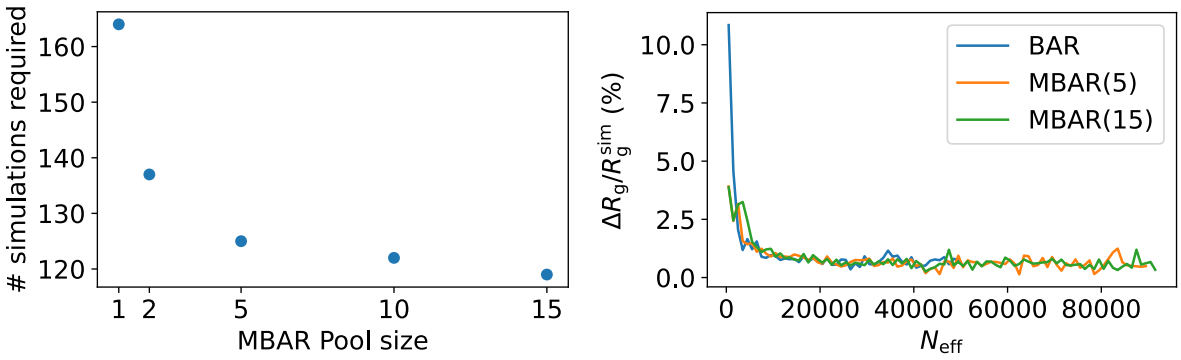


Figure S4: To test the accuracy and efficiency of MBAR reweighting, we generated a random sequence of 140 residues and performed 1000 position swaps between two randomly selected residues. We simulated all 1000 sequences and calculate their $R_g$. Then we iterate through the 1000 sequences trying to predict their $R_g$ by reweighting simulations from previous iterations. We vary the maximum size of the MBAR pool and add a new simulation to the pool when the $N_{eff}$ drops below 10000. Then we compare the reweighted $R_g$ from MBAR with the simulated $R_g$. The left panel shows the number of simulations required by varying the maximum MBAR pool size. The right panel shows the relative absolute difference between reweighted and simulated $R_g$ ($|\Delta R_g|/R_g^{sim}$) as a function of $N_{eff}$. For better visualization, we binned the data on the $N_{eff}$ coordinate (with a bin width of 1000) and plot the average in the bins.

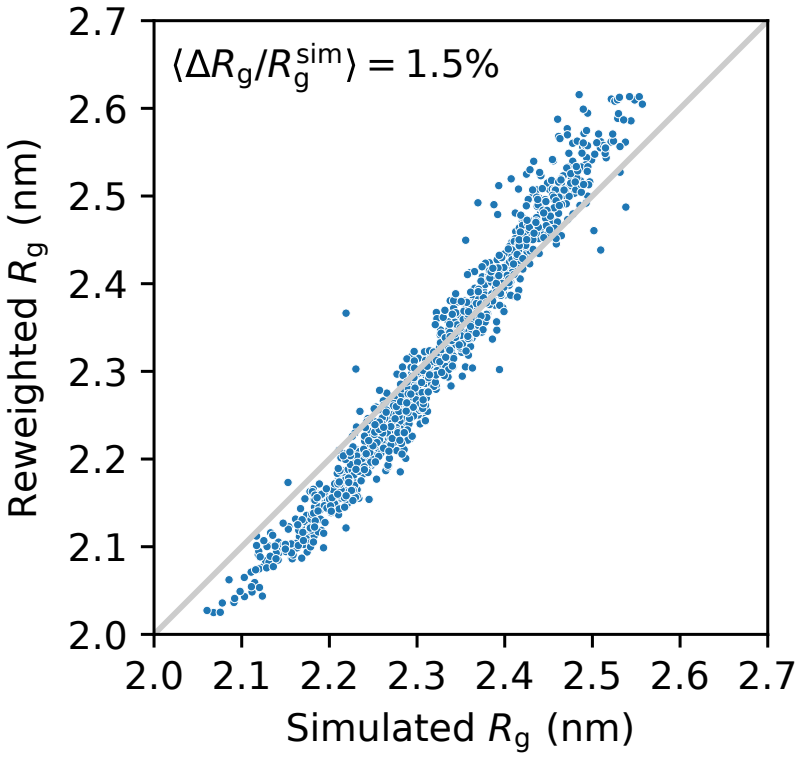


Figure S5: For some of the centroids selected from the sequence clustering of the A1-LCD variants the $R_g$ values had been obtained by reweighting. We simulated each of these for 1 $\mu$s to assess the accuracy of the reweighting. The reweighted and simulated $R_g$ values are compared. We observe an average error of 1.5% on the reweighted $R_g$, with a slight bias for the most compact and expanded chains.

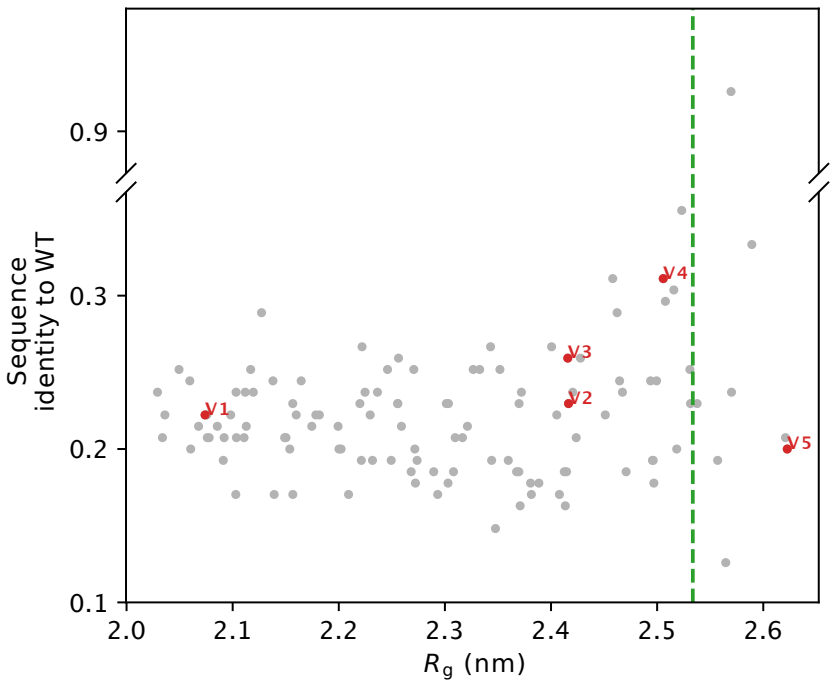


Figure S6: Sequence identity to wild-type A1-LCD for the 119 designed A1-LCD variants. Green vertical line correspond to the $R_g$ of wild-type A1-LCD.

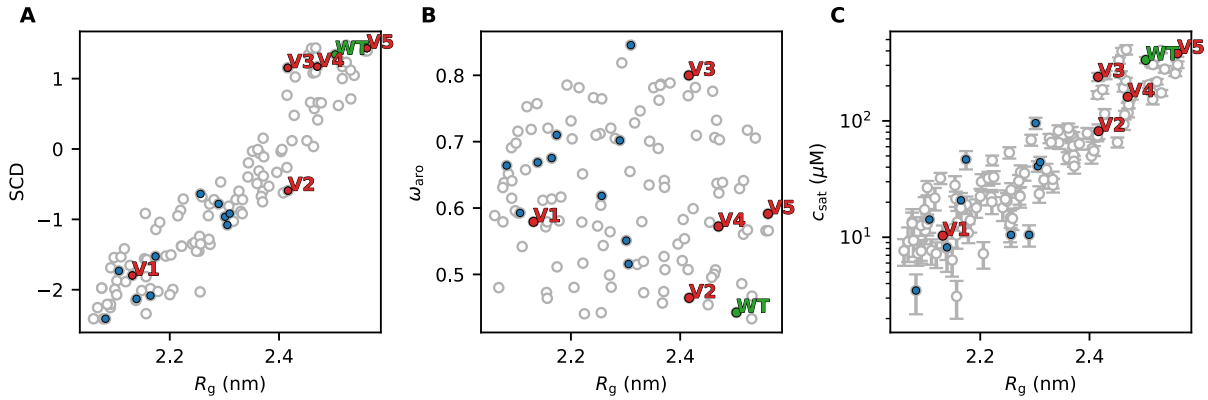


Figure S7: Characterization of the 120 variants of A1-LCD. We show the relationship between $R_g$ and (a) SCD, (b) $\omega_{aro}$ (patterning of aromatic residues) and (c) the $c_{sat}$ calculated from simulations of 100 chains in slab geometry. We highlight the wild-type sequence of A1-LCD in green, the five variants that we characterized experimentally in red, and ten variants that did not express in *E. coli* in blue.

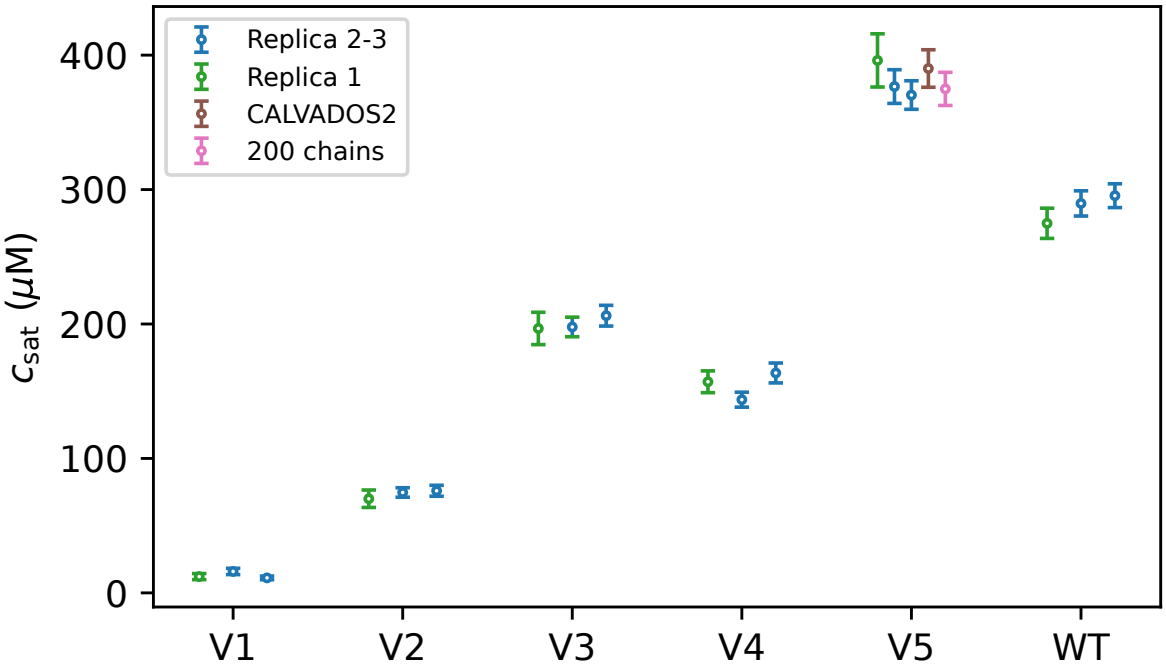


Figure S8: $c_{sat}$ values calculated by slab simulations of experimental constructs. Replicas 1,2 and 3 were performed with CALVADOS M1 (*49*) and 100 chains in the simulation box. Replica 1 (green) is 20-$\mu$s long, while replicas 2 and 3 (blues) are 50-$\mu$s long. For V5, we also performed a 20-$\mu$s long simulation with CALVADOS M1 but using 200 chains (pink), and a 20-$\mu$s long simulation with 100 chains but the CALVADOS 2 parameters (brown) (*53*).

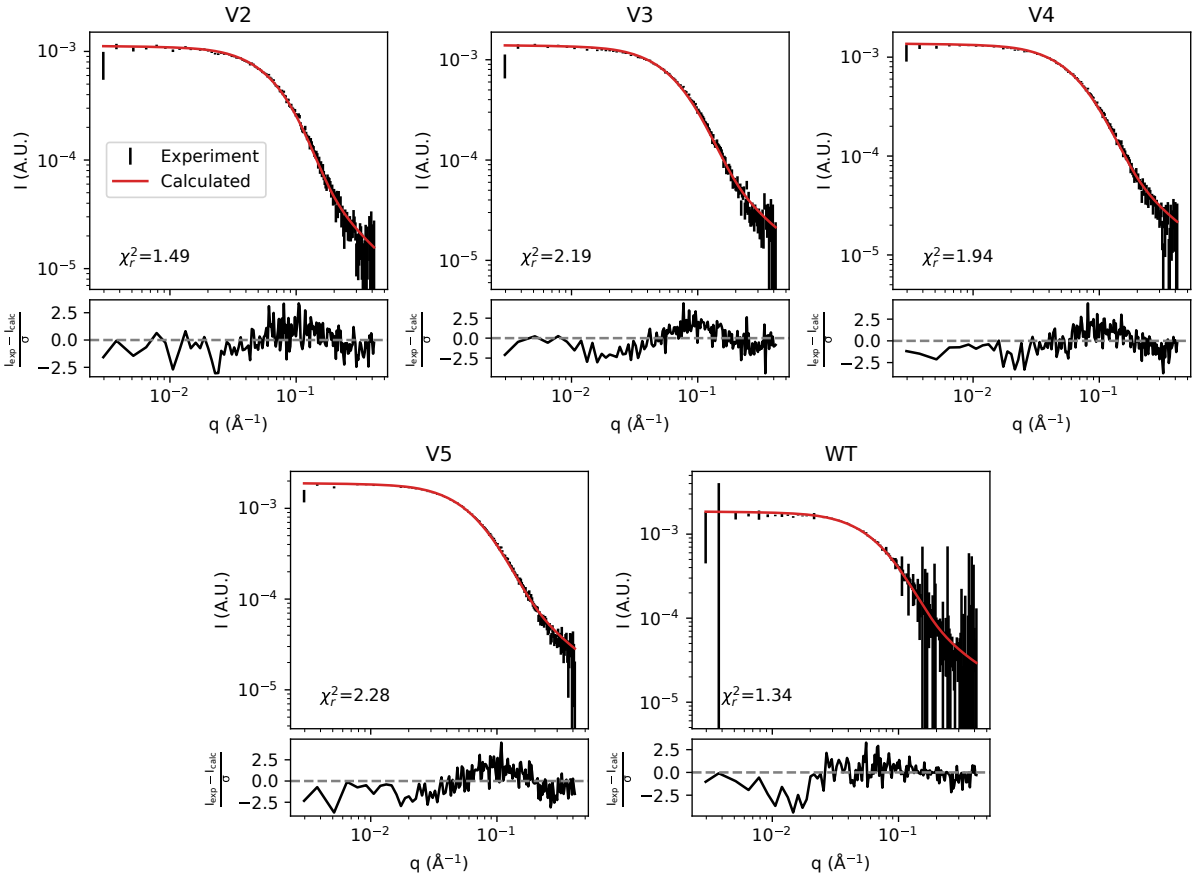


Figure S9: Rebinned experimental SAXS data with corrected error bars (black) compared to SAXS curves calculated from simulations.

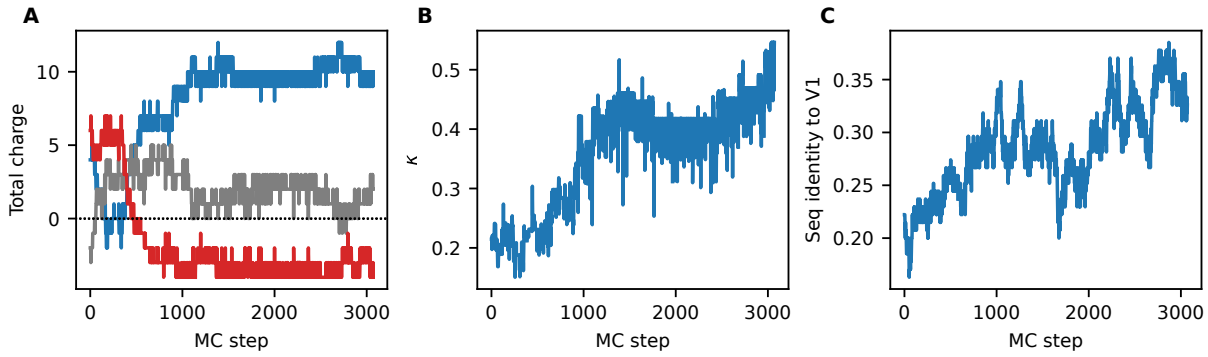


Figure S10: Design of swap variants starting from A1-LCD to target the V1 contact map. (A) Total charge in the N-terminal third (blue), middle third (grey) and C-terminal third (red) of the variants proposed during the design. (B) Charge segregation (as quantified by $\kappa$) of the sequences proposed during the design. (C) Sequence identity to V1 of the variants proposed during the design.

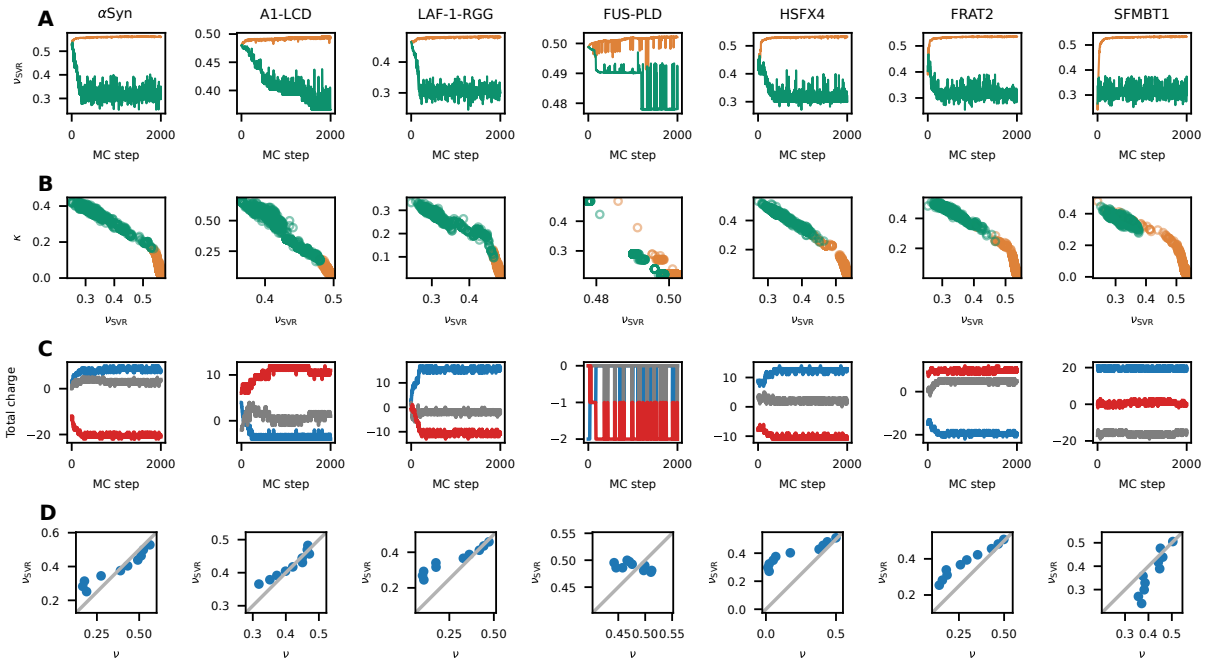


Figure S11: Design of swap variants with varying compaction using a support vector regression machine learning model to predict the scaling exponent ($\nu_{SVR}$) from sequences. We applied this to the seven IDPs that we also studied using the simulation-based algorithm (main text Fig. 2 and 7) . For each sequence during design we show (A) $\nu_{SVR}$ (B) targeting either $\nu_{SVR}$=0.3 (green lines) or $\nu_{SVR}$=0.7 (orange lines), $\kappa$, (C) total charge in the N-terminal third (blue), middle third (grey) and C-terminal third (red). Starting $c$ value for Monte Carlo was set to $2\times10^{-5}$. (D) We ran molecular dynamics simulations of 10 variants of each protein and compared the simulation-derived values of $\nu$ with $\nu_{SVR}$.

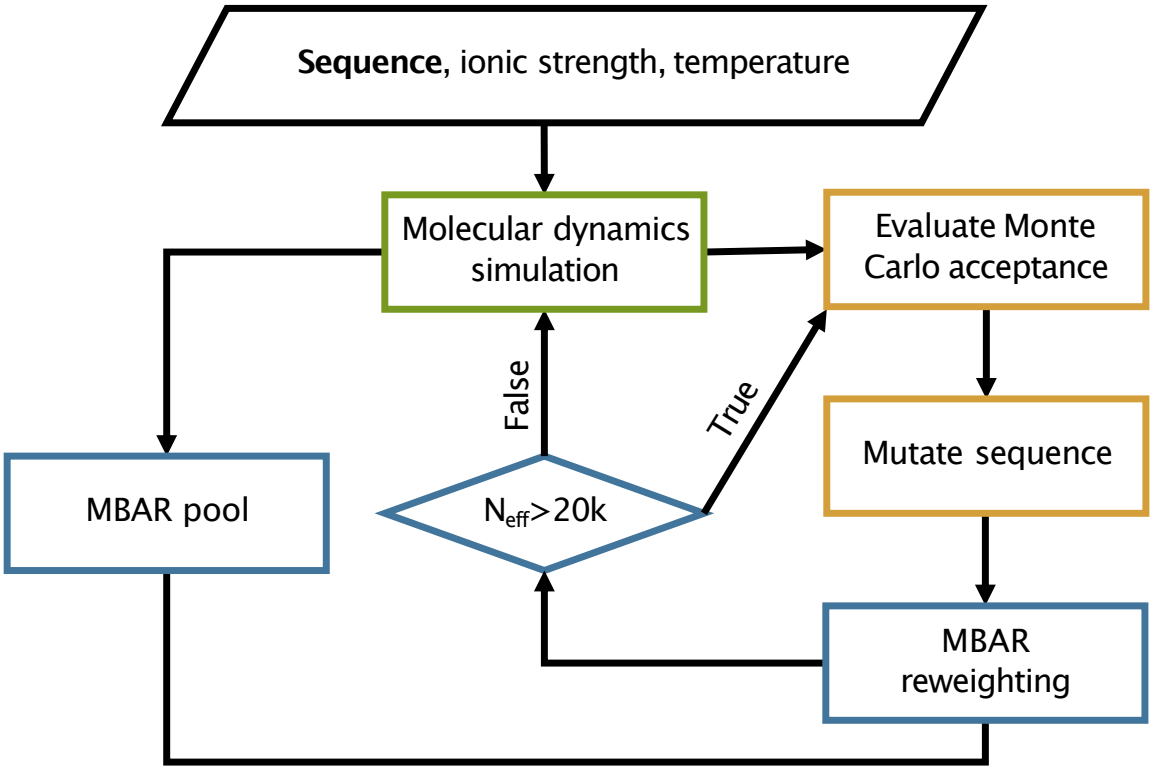


Figure S12: Schematic outline of the design algorithm.

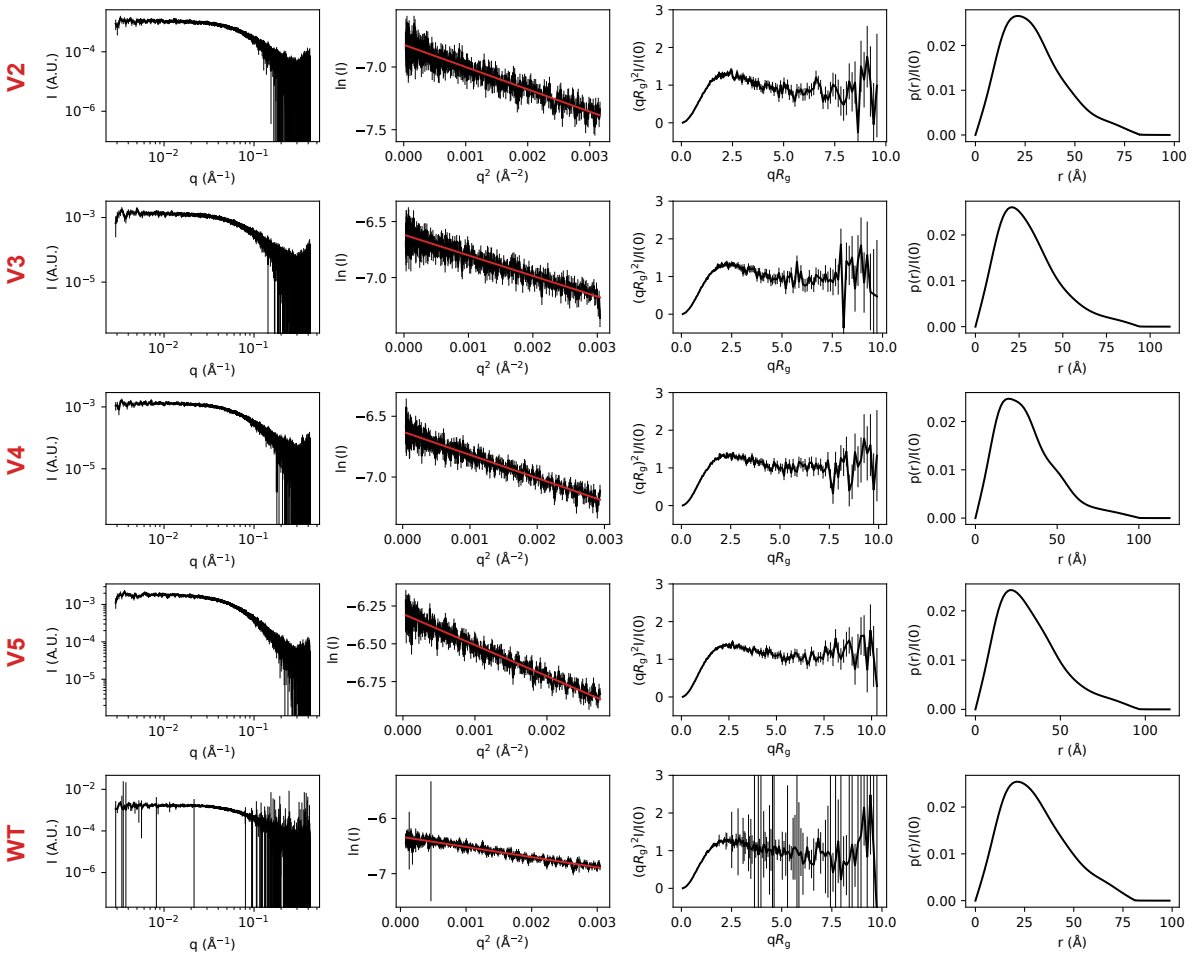


Figure S13: SAXS data collected on samples of (from top to bottom rows) V2, V3, V4, V5 and wild-type A1-LCD. From the left to the right column, SAXS profiles are shown on logarithmic scales, as a Guinier plot in the range used for the linear fit (in red) to derive the the Guinier $R_g$, the dimensionless Kratky plot with rebinned SAXS data, and the normalized pair distance distribution function (calculated using BIFT (*108*)).

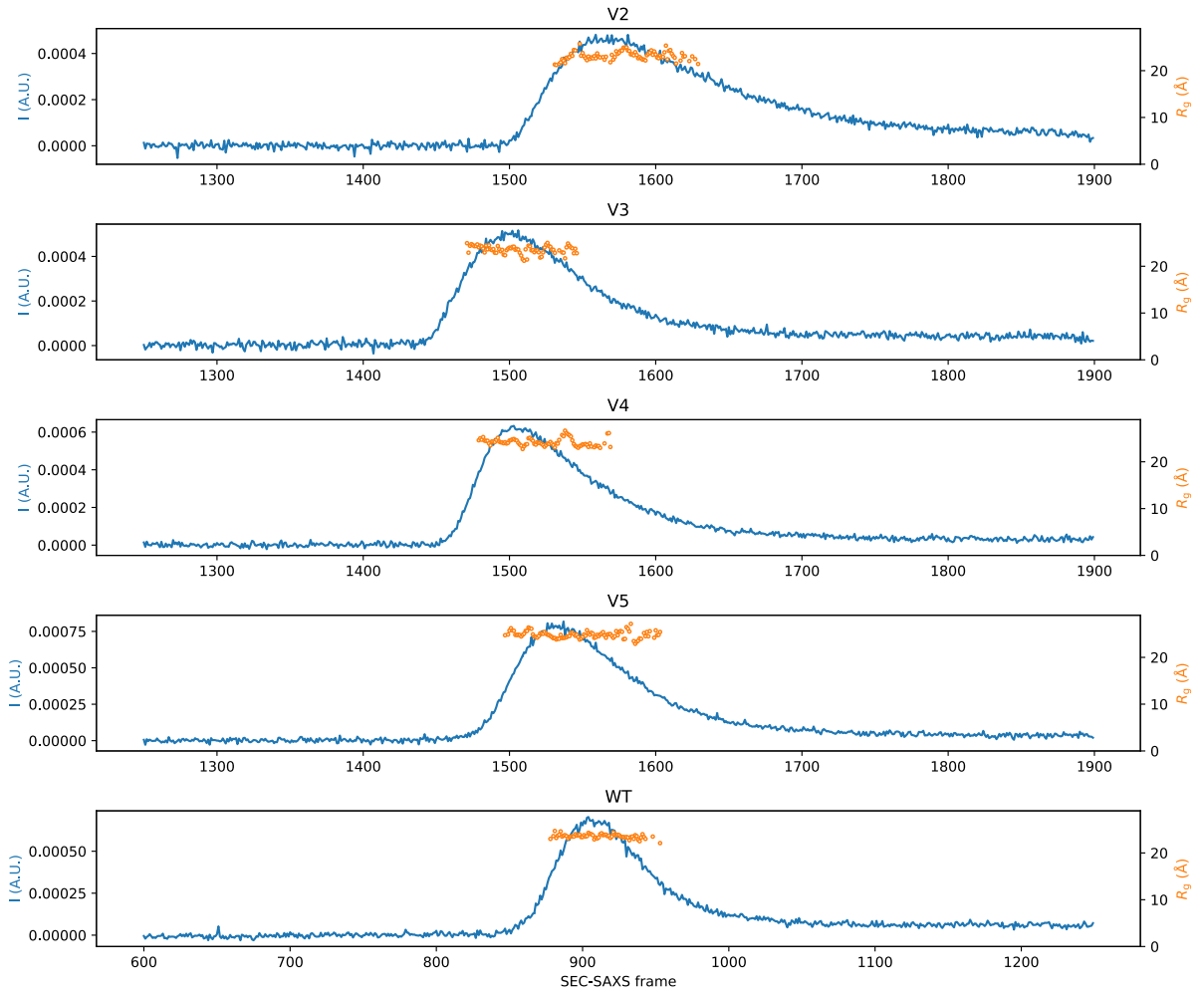


Figure S14: In line SEC-SAXS data collection. In blue, we show the mean solvent-subtracted intensity for each SAXS frame collected during sample elution from the SEC column. In orange, we show the Guinier $R_g$ calculated for each SAXS frame.

Table S1: Sequences of the wild type A1-LCD and the five designed variants that we characterized experimentally. The first two residues (GS) are left over by the TEV protease cleavage of the 6xHis-tag.

| Label | Sequence |
|---|---|
| **WT** | GS—MASASSSQRGRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFG GSRGGGGYGGSGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFGP MKGGNFGGRSSGGSGGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF |
| **V1** | GS—GSGSGGSRGGNKRRRKRRGGSGGYRYSRRGGGFNQGGGFNSSGF FGGMGSGGGSGGGFGNGPSFAGSNNFNGGGGGSAGNFGQYGGRGGPY SGSGGSGSGSNSGQNGGSGNYMGSGYDAFYNSSFNNQSFFGDDD |
| **V2** | GS—GGYGSSQGGFFGGGDAGGNGDGSDFGGGYPSGSNQNSGGFSGYG NDSFQGSAGMFNGFKSASKFSNSGGYGGGGQGNNNGSGGGSSFRNRR RRSNYSGGGSGRGRRYGSNFGGMYGGRSGFGGNGPGRSGFGGSN |
| **V3** | GS—KQGGRGGNRSGSGNGNASGAGGGGRDGGSDGGFDGFDYQFSGG GNPSSQYYGSRGGSGRNSAGGYYFFRNSSGGNGSSGNMNPGNGYFGFS RSGGRGQNRGFFFGGMGGGGFGRSSNFGSYNSSNKSGSGGGGGG |
| **V4** | GS—GSNGGGSQSSGQGYGKSGGNRRRGRGGAGGGFGMGDGSNQYGY GPFRRGSGFNGNGDYANYGGNGDSNNFSNYRGGNSANGNFQSGGGGG FDNGGGSGFGGSFSMSGGSSSGKRRGSGGFFSGRSGSGFGGFYPS |
| **V5** | GS—GFSNMGNGFGGRFGGGRGFSRYSQQFSYYDGGQSSGGNGSSGGF NSYGGYNNGRNGSSFGGAGGGGRSSFGFSGGGGFGADGGYNRFSSGD RNNNGPSKGGGGGNGSGSRGFAGNGSMSDRGNSYGGGPGRQKGS |

Table S2: Sequences of the computationally designed variants that we could not express in *E. coli*. The first two residues (GS) are part of the TEV protease cleavage site.

| Label | Sequence |
|---|---|
| **X1** | GS—GYAGGRGKRRRRKRRNRRRGYQSGSGGGGGGNGSNGGSGGYGA<br>SNNNGSGFFGGYGGNSGFFSNSYFGGGQGANGSGNPGSFGGGGSGPSG<br>SGMSFYGGGSSGGFGQDDGDFDGQSNSSGNNSFSMFGGGNYSFG |
| **X2** | GS—GGFNSQQKRRRRKRRFRFRSYGGNGSSNGSGGGYFGNNSGSGSG<br>GGRGGYGFGRGNSGGSGNPGAGAGFPGYSSSGMAGYGQNGNSSGFGG<br>SNMSFGNGNGGFSGGSDDGDYDFSGSFGGGGFGQYGNSNGSGGG |
| **X3** | GS—GKGKRGGGRGSNGAFFSPRRRRFRRSASGRRAGSGGQGQGQGGG<br>YYYNGGSSFNSGMGSGNGNSPYGNGNGGGGGGGGFSYGNGSSMSGFN<br>SGSSGGGGNYGGGFGGSGNGSGSNQSSNFFSFDFDFGGDGYNGD |
| **X4** | GS—GGGGYGAGRGSGGNRSRRRRSRSRKRRKGYSGQSNFNGGSGFGG<br>FSQQGGPGSSGGGGNGGSYFAYFGGGGGQNGSSGSSGMYFNGSSNFGS<br>SGNNGDSGSGANMFSNGGSFPGGGGNDDGDNNFFFGGYYGGGG |
| **X5** | GS—SAGGNSKNGNGGRGFGGRRSSRMRRGFFRNFRRSFPGNNGSGGY<br>QGGFGGKFGGSSGSGGGRFYPYNNGGNNGSGFSGGSGGGSSGSGSSQG<br>SGGSNGASNGQGAGGGGGSNGGGFYGDSYYFYDNMDGGFSDQG |
| **X6** | GS—GGGGSMFKFSSSGRGGRRRKFRSRRGRGFGSGQGSNRGSSGFGGP<br>GSGGGYGPGGGSFNYSGSGAYGGSGNSQYGDNGSGQFYGNSQMNGY<br>GNGNGSGGGSNYSNGSGGGFRGFGGGNDDGDSNNFSFAGAGNGF |
| **X7** | GS—SSDFGNKRRRRRRFRRRFRGYGSFGGGSGGQANGGYMGAGGNG<br>GGSPGGNFMNSNGGGFGFFYGGSASGGGGNNGSNNSDYQGGDSNNGG<br>PFQGGSGGGNGSNFSFQYKGDYSSGYSSGGGGGSSGGSGGSFGSG |
| **X8** | GS—SPFFGYQGNNFRRKRGRRKGGGSYGSSRGSGNNSGGGGSRSSGG<br>GSGGGSGGFYYFRNNFSQRGNSGDGFSSGGGNYFGGSSGSAQGRGYG<br>GNFGAGNGGGSSNGGGNGGANNGFFFSGGQMGSGPMSYGGGDDD |
| **X9** | GS—QGGGSSSGNGGGGKGGGSNGNGRFFRFRGGFYGRRSGKNRRYN<br>GFGMSRYGGFSRSGNGGSMGFRGQSGGGPSSQSFAQYGGGSNFNGSA<br>GYNDGPFGGSSFGGSGAGDSYGDGGNNGFDGGSYSGGGGSNNGSN |
| **X10** | GS—SNFNGSGNRRQSRGSNRRGRRGYYSFRGNFYFRNGGGGGGGKNG<br>GSGGNNPGGMGGFQGSGGSGGGAGNSKRAGGMGGAGSFGYGSGNGG<br>DSGSGGQGSSNSGGNSGSSGQNGFFDSGFSPYGDDGFFGFSYYGS |

Table S3: SAXS sample, data-collection and analysis for the wild-type A1-LCD and its variants*.

| | V2 | V3 | V4 | V5 | WT |
|---|---|---|---|---|---|
| **(a) Sample details** | | | | | |
| Organism | Artificial | Artificial | Artificial | Artificial | Human |
| Source | *E. coli* BL21 (DE3) pLys recombinant expression | | | | |
| *Sample environment/configuration* | | | | | |
| Solvent composition | 20 mM HEPES pH 7.0, 150 mM NaCl, 2 mM DTT | | | | |
| Sample temperature (K) | 298 | | | | |
| In-beam sample cell | 1 mm quartz capillary flow cell | | | | |
| *Size exclusion chromatography* | | | | | |
| Sample injection concentration (mg/mL) | 2.6 | 2.6 | 2.6 | 2.6 | 2.6 |
| Sample injection volume (mL) | 250 | | | | |
| SEC column type | Superdex 75 Increase 10/300 GL column (Cytiva) | | | | |
| SEC flowrate (mL/min) | 0.6 | | | | |
| **(b) SAXS data collection** | | | | | |
| Data-acquisition/reduction software | BioXTAS RAW 2.1.4 | | | | |
| Source/instrument description | BioCAT (Sector 18, APS) | | | | |
| Measured q-range ($q_{\text{min}} - q_{\text{max}}$) (Å$^{-1}$) | 2.90e-03 – 4.17e-01 | | | | |
| Method for scaling intensities | Absolute scaling with glassy carbon | | | | |
| Exposure time (s) | 0.5 | | | | |
| **(c) SAS-derived structural parameters** | | | | | |
| *Guinier analysis* | | | | | |
| Method(s)/software | autorg (ATSAS 3.1.3) | | | | |
| I(0) | 0.0011 ± 5.4e-06 | 0.0013 ± 6.8e-06 | 0.0013 ± 5.3e-06 | 0.0018 ± 6.6e-06 | 0.0018 ± 8.7e-06 |
| $R_{\text{g}}$ (Å) | 23.1 ± 0.2 | 23.48 ± 0.21 | 23.95 ± 0.17 | 24.84 ± 0.16 | 23.55 ± 0.21 |
| $qR_{\text{g}}$ range | 0.13 – 1.3 | 0.12 – 1.3 | 0.16 – 1.3 | 0.15 – 1.3 | 0.21 – 1.3 |
| Linear fit assessment (*autorg* fidelity) | 1 | 1 | 1 | 0.98 | 0.01 |
| *Pair distance distribution function analysis* | | | | | |
| Method(s)/software | BIFT (BayesApp 1.1) | | | | |
| I(0) | 1.09e-03 | 1.35e-03 | 1.34e-03 | 1.85e-03 | 1.79e-03 |
| $R_{\text{g}}$ (Å) | 23.65 | 24.89 | 25.47 | 26.21 | 24.5 |
| $D_{\text{max}}$ (Å) | 82.56 | 93 | 98.89 | 95.35 | 80.32 |
| P(r) reciprocal-space fit: $\chi^2_r$, p-value | 0.80, 4.40e-04 | 0.77, 4.1e-05 | 0.87, 2.6e-02 | 0.84, 4.5e-03 | 0.73, 4.1e-07 |
| **(d) Scattering particle size** | | | | | |
| Porod volume (Å$^3$) | 16726 | 14254 | 14874 | 22680 | 15360 |
| Theoretical MW (kDA) | 13.1 | | | | |
| SAXS MW (DatBayes)** (kDA), probability | 15.475, 0.45 | 14.825, 0.48 | 14.825, 0.43 | 14.825, 0.39 | 15.475, 0.50 |
| **(e) Modelling (SAXS calculation from molecular simulations)** | | | | | |
| Software | Pepsi-SAXS 3.0 | | | | |
| q-range for calculation (Å$^{-1}$) | 2.90e-03 – 4.17e-01 | | | | |
| Number of frames used | 10000 | | | | |
| Scale factor and offset | Fixed to constant in Pepsi-SAXS, then globally fit to experiment by least square | | | | |
| $\delta\rho$ (e/nm$^3$) | 3.34 | | | | |
| Average atomic radius ($r_m$; Å) | 1.58 | | | | |
| $r_0/r_m$ | 1.025 | | | | |
| $\chi^2_r$ | 1.49 | 2.19 | 1.94 | 2.28 | 1.34 |
| **(f) Data deposition** | | | | | |
| SASDB ID | SASDTK2 | SASDTL2 | SASDTM2 | SASDTN2 | SASDTJ2 |

* Table in accordance with guidelines from (*110*) and (*111*).

** See (*112*).

**REFERENCES AND NOTES**

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

2. A. S. Holehouse, B. B. Kragelund, The molecular basis for cellular function of intrinsically disordered protein regions. *Nat. Rev. Mol. Cell Biol.* **25**, 187–211 (2024).

3. V. N. Uversky, J. R. Gillespie, A. L. Fink. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41**, 415–427 (2000).

4. T. Mittag, J. D. Forman-Kay, Atomic-level characterization of disordered protein ensembles. *Curr. Opin. Struct. Biol.* **17**, 3–14 (2007).

5. F. E. Thomasen, K. Lindorff-Larsen, Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. *Biochem. Soc. Trans.* **50**, 541–554 (2022).

6. M. Li, H. Cao, L. Lai, Z. Liu, Disordered linkers in multidomain allosteric proteins: Entropic effect to favor the open state or enhanced local concentration to favor the closed state? *Protein Sci.* **27**, 1600–1610 (2018).

7. A. A. Santner, C. H. Croy, F. H. Vasanwala, V. N. Uversky, Y.-Y. J. Van, A. K. Dunker, Sweeping away protein aggregation with entropic bristles: Intrinsically disordered protein fusions enhance soluble expression. *Biochemistry* **51**, 7250–7262 (2012).

8. D. Jamecna, J. Polidori, B. Mesmin, M. Dezi, D. Levy, J. Bigay, B. Antonny, An intrinsically disordered region in OSBP acts as an entropic barrier to control protein dynamics and orientation at membrane contact sites. *Dev. Cell* **49**, 220–234.e8 (2019).

9. N. E. Davey, K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, T. J. Gibson, Attributes of short linear motifs. *Mol. Biosyst.* **8**, 268–281 (2012).

10. B. A. Shoemaker, J. J. Portman, P. G. Wolynes, Speeding molecular recognition by using the folding funnel: The fly-casting mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 8868–8873 (2000).

11. P. E. Wright, H. J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).

12. J. Liu, N. B. Perumal, C. J. Oldfield, E. W. Su, V. N. Uversky, A. K. Dunker, Intrinsic disorder in transcription factors. *Biochemistry* **45**, 6873–6888 (2006).

13. S. F. Banani, H. O. Lee, A. A. Hyman, M. K. Rosen, Biomolecular condensates: Organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).

14. T. Mittag, R. V. Pappu, A conceptual framework for understanding phase separation and addressing open questions and challenges. *Mol. Cell* **82**, 2201–2214 (2022).

15. I. Alshareedah, W. M. Borcherds, S. R. Cohen, M. Farag, A. Singh, A. Bremer, R. V. Pappu, T. Mittag, P. R. Banerjee, Sequence-encoded grammars determine material properties and physical aging of protein condensates. *Nat. Phys.* 1–10 (2024).

16. I. Alshareedah, W. M. Borcherds, S. R. Cohen, M. Farag, A. Singh, A. Bremer, R.V. Pappu, T. Mittag, P. R. Banerjee, A sequence-encoded grammars determine material properties and physical aging of protein condensates. *Nat. Phys*. 1–10 (2024).

17. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein,

D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

18. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

19. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

20. J. A. Marsh, J. D. Forman-Kay. Sequence determinants of compaction in intrinsically disordered proteins. *Biophys. J.* **98**, 2383–2390 (2010).

21. A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, R. V. Pappu, Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 8183–8188 (2010).

22. S. Müller-Späth, A. Soranno, V. Hirschfeld, H. Hofmann, S. Rüegger, L. Reymond, D. Nettels, B. Schuler, Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 14609–14614 (2010).

23. R. K. Das, K. M. Ruff, R. V. Pappu, Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **32**, 102–112 (2015).

24. M. C. Cohan, K. M. Ruff, R. V. Pappu, Information theoretic measures for quantifying sequence–ensemble relationships of intrinsically disordered proteins. *Protein Eng. Des. Sel.* **32**, 191–202 (2019).

25. G. Tesei, A. I. Trolle, N. Jonsson, J. Betz, F. E. Knudsen, F. Pesce, K. E. Johansson, K. Lindorff-Larsen, Conformational ensembles of the human intrinsically disordered proteome. *Nature* **626**, 897–904 (2024).

26. J. M. Lotthammer, G. M. Ginell, D. Griffith, R. Emenecker, A. S. Holehouse, Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat. Methods* **21**, 465–476 (2024).

27. I. Pritišanac, T. R. Alderson, Đ. Kolarić, T. Zarin, S. Xie, A. X. Lu, A. Alam, A. Maqsood, J.-Y. Youn, J. D. Forman-Kay, A. M. Moses, A functional map of the human intrinsically disordered proteome. *bioRxiv* 2024.03.15.585291 [Preprint] (2024).

28. X. Pan, T. Kortemme. Recent advances in de novo protein design: Principles, methods, and applications. *J. Biol. Chem.* **296**, 100558 (2021).

29. D. N. Woolfson, A brief history of de novo protein design: Minimal, rational, and computational. *J. Mol. Biol.* **433**, 167160 (2021).

30. C. A. Goverde, B. Wolf, H. Khakzad, S. Rosset, B. E. Correia, De novo protein design by inversion of the AlphaFold structure prediction network. *Protein Sci.* **32**, e4653 (2023).

31. A. Garg, N. S. Gonzalez-Foutel, M. B. Gielnik, M. Kjaergaard, Design of functional intrinsically disordered proteins. *Protein Eng. Des. Sel.* **37**, gzae004 (2024).

32. M. Van Rosmalen, M. Krom, M. Merkx, Tuning the flexibility of glycine-serine linkers to allow rational design of multidomain proteins. *Biochemistry* **56**, 6565–6574 (2017).

33. M. Dzuricky, S. Roberts, A. Chilkoti, Convergence of artificial protein polymers and intrinsically disordered proteins. *Biochemistry* **57**, 2405–2414 (2018).

34. T. Lazar, E. Martínez-Pérez, F. Quaglia, A. Hatos, L. B. Chemes, J. A. Iserte, N. A. Méndez, N. A. Garrone, T. E. Saldaño, J. Marchetti, A. J. V. Rueda, P. Bernadó, M. Blackledge, T. N. Cordeiro, E. Fagerberg, J. D. Forman-Kay, M. S. Fornasari, T. J. Gibson, G. N. W. Gomes, C. C. Gradinaru, T. Head-Gordon, M. R. Jensen, E. A. Lemke, S. Longhi, C. Marino-Buslje, G. Minervini, T. Mittag, A. M. Monzon, R. V. Pappu, G. Parisi, S. Ricard-Blum, K. M. Ruff, E. Salladini, M. Skepö, D. Svergun, S. D. Vallet, M. Varadi, P. Tompa, S. C. E. Tosatto, D. Piovesan, PED in 2021: A major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.* **49**, D404–D411 (2021).

35. K. Lindorff-Larsen, B. B. Kragelund, On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins. *J. Mol. Biol.* **433**, 167196 (2021).

36. R. K. Das, R. V. Pappu, Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 13392–13397 (2013).

37. Y.-H. Lin, H. S. Chan, Phase separation and single-chain compactness of charged disordered proteins are strongly correlated. *Biophys. J.* **112**, 2043–2046 (2017).

38. B. S. Schuster, G. L. Dignon, W. S. Tang, F. M. Kelley, A. K. Ranganath, C. N. Jahnke, A. G. Simpkins, R. M. Regy, D. A. Hammer, M. C. Good, J. Mittal, Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11421–11431 (2020).

39. A. Bremer, M. Farag, W. M. Borcherds, I. Peran, E. W. Martin, R. V. Pappu, T. Mittag, Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat. Chem.* **14**, 196–207 (2022).

40. W. Zheng, G. Dignon, M. Brown, Y. C. Kim, J. Mittal, Hydropathy patterning complements charge patterning to describe conformational preferences of disordered proteins. *J. Phys. Chem. Lett.* **11**, 3408–3415 (2020).

41. E. W. Martin, A. S. Holehouse, I. Peran, M. Farag, J. J. Incicco, A. Bremer, C. R. Grace, A. Soranno, R. V. Pappu, T. Mittag, Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367**, 694–699 (2020).

42. A. S. Holehouse, G. M. Ginell, D. Griffith, E. Böke, Clustering of aromatic residues in prion-like domains can tune the formation, state, and organization of biomolecular condensates. *Biochemistry* **60**, 3566–3581 (2021).

43. R. K. Das, Y. Huang, A. H. Phillips, R. W. Kriwacki, R. V. Pappu, Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 5616–5621 (2016).

44. M. K. Shinn, M. C. Cohan, J. L. Bullock, K. M. Ruff, P. A. Levin, R. V. Pappu, Connecting sequence features within the disordered C-terminal linker of *Bacillus subtilis* FtsZ to functions and bacterial cell division. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2211178119 (2022).

45. R. J. Emenecker, K. Guadalupe, N. M. Shamoon, S. Sukenik, A. S. Holehouse, Sequence-ensemble-function relationships for disordered proteins in live cells. *bioRxiv* 2023.10.29.564547 [Preprint] (2023).

46. C. W. Pak, M. Kosno, A. S. Holehouse, S. B. Padrick, A. Mittal, R. Ali, A. A. Yunus, D. R. Liu, R. V. Pappu, M. K. Rosen. Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein, *Mol. Cell* **63**, 72–85 (2016).

47. J. A. Greig, T. A. Nguyen, M. Lee, A. S. Holehouse, A. E. Posey, R. V. Pappu, G. Jedd. Arginine-enriched mixed-charge domains provide cohesion for nuclear speckle condensation, *Mol. Cell* **77**, 1237–1250.e4 (2020).

48. J.-E. Shea, R. B. Best, J. Mittal. Physics-based computational and theoretical approaches to intrinsically disordered proteins, *Curr. Opin. Struct. Biol.* **67**, 219–225 (2021).

49. G. Tesei, T. K. Schulze, R. Crehuet, K. Lindorff-Larsen, Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2111696118 (2021).

50. T. Dannenhoffer-Lafage, R. B. Best, A data-driven hydrophobicity scale for predicting liquid–liquid phase separation of proteins. *J. Phys. Chem. B* **125**, 4046–4056 (2021).

51. R. M. Regy, J. Thompson, Y. C. Kim, J. Mittal, Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. *Protein Sci.* **30**, 1371–1379 (2021).

52. J. A. Joseph, A. Reinhardt, A. Aguirre, P. Y. Chew, K. O. Russell, J. R. Espinosa, A. Garaizar, R. Collepardo-Guevara, Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat. Comput. Sci.* **1**, 732–743 (2021).

53. G. Tesei, K. Lindorff-Larsen, Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Res. Eur.* **2**, 94 (2022).

54. J. Methorst, N. van Hilten, A. Hoti, K. S. Stroh, H. J. Risselada, When data are lacking: Physics-based inverse design of biopolymers interacting with complex, fluid phases. *J. Chem. Theory Comput.* **20**, 1763–1776 (2024).

55. T. S. Harmon, M. D. Crabtree, S. L. Shammas, A. E. Posey, J. Clarke, R. V. Pappu, GADIS: Algorithm for designing sequences to achieve target secondary structure profiles of intrinsically disordered proteins. *Protein Eng. Des. Sel.* **29**, 339–346 (2016).

56. X. Zeng, C. Liu, M. J. Fossat, P. Ren, A. Chilkoti, R. V. Pappu, Design of intrinsically disordered proteins that undergo phase transitions with lower critical solution temperatures. *APL Mater.* **9**, 021119 (2021).

57. S. M. Lichtinger, A. Garaizar, R. Collepardo-Guevara, A. Reinhardt, Targeted modulation of protein liquid–liquid phase separation by evolution of amino-acid sequence, *PLOS Comput. Biol.* **17**, e1009328 (2021).

58. N. van Hilten, J. Methorst, N. Verwei, H. J. Risselada, Physics-based generative model of curvature sensing peptides; distinguishing sensors from binders. *Sci. Adv.* **9**, eade8839 (2023).

59. A. B. Norgaard, J. Ferkinghoff-Borg, K. Lindorff-Larsen, Experimental parameterization of an energy function for the simulation of unfolded proteins. *Biophys. J.* **94**, 182–192 (2008).

60. S. Orioli, A. H. Larsen, S. Bottaro, K. Lindorff-Larsen, How to learn from inconsistencies: Integrating molecular simulations with experimental data. *Prog. Mol. Biol. Transl. Sci.* **170**, 123–176 (2020).

61. J. Köfinger, G. Hummer. Empirical optimization of molecular simulation force fields by Bayesian inference. *Eur. Phys. J. B* **94**, 245 (2021).

62. E. W. Martin, A. S. Holehouse, C. R. Grace, A. Hughes, R. V. Pappu, T. Mittag. 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).

63. K. P. Sherry, R. K. Das, R. V. Pappu, D. Barrick, Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E9243–E9252 (2017).

64. R. Beveridge, L. G. Migas, R. K. Das, R. V. Pappu, R. W. Kriwacki, P. E. Barran, 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).

65. M. C. Cohan, M. K. Shinn, J. M. Lalmansingh, R. V. Pappu, Uncovering non-random binary patterns within sequences of intrinsically disordered proteins. *J. Mol. Biol.* 167373 (2021).

66. J. Wang, J.-M. Choi, A. S. Holehouse, H. O. Lee, X. Zhang, M. Jahnel, S. Maharana, R. Lemaitre, A. Pozniakovsky, D. Drechsel, I. Poser, R. V. Pappu, S. Alberti, A. A. Hyman, A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell* **174**, 688–699.e16 (2018).

67. T.-H. Chao, S. Rekhi, J. Mittal, D. P. Tabor, Data-driven models for predicting intrinsically disordered protein polymer physics directly from composition or sequence. *Mol. Syst. Des. Eng.* **8**, 1146–1155 (2023).

68. J.-M. Choi, A. S. Holehouse, R. V. Pappu, Physical principles underlying the complex biology of intracellular phase transitions. *Annu. Rev. Biophys.* **49**, 107–133 (2020).

69. M. J. Maristany, A. A. Gonzalez, R. Collepardo-Guevara, J. A. Joseph, Universal predictive scaling laws of phase separation of prion-like low complexity domains. *bioRxiv* 2023.06.14.543914 [Preprint] (2023).

70. G. L. Dignon, W. Zheng, R. B. Best, Y. C. Kim, J. Mittal, Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9929–9934 (2018).

71. E. W. Martin, J. B. Hopkins, T. Mittag, Small-angle X-ray scattering experiments of monodisperse intrinsically disordered protein samples close to the solubility limit. *Methods Enzymol.* **646**, 185–222 (2021).

72. J. Henriques, L. Arleth, K. Lindorff-Larsen, M. Skepö, On the calculation of SAXS profiles of folded and intrinsically disordered proteins from computer simulations. *J. Mol. Biol.* **430**, 2521–2539 (2018).

73. F. Pesce, K. Lindorff-Larsen, Refining conformational ensembles of flexible proteins against small-angle x-ray scattering data. *Biophys. J.* **120**, 5124–5135 (2021).

74. F. Pesce, E. A. Newcombe, P. Seiffert, E. E. Tranchant, J. G. Olsen, C. R. Grace, B. B. Kragelund, K. Lindorff-Larsen, Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins. *Biophys. J.* **122**, 310–321 (2022).

75. E. E. Tranchant, F. Pesce, N. L. Jacobsen, C. B. Fernandes, B. B. Kragelund, K. Lindorff-Larsen, Revisiting the use of dioxane as a reference compound for determination of the hydrodynamic radius of proteins by pulsed field gradient NMR spectroscopy. *bioRxiv* 2023.06.02.543514 [Preprint] (2023).

76. I. C. Sanchez, Phase transition behavior of the isolated polymer chain. *Macromolecules* **12**, 980–988 (1979).

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

78. M. R. Shirts, J. D. Chodera, Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129**, 124105 (2008).

79. J. Lu, C. Deutsch, Electrostatics in the ribosomal tunnel modulate chain elongation rates. *J. Mol. Biol.* **384**, 73–86 (2008).

80. J. C. Hansen, X. Lu, E. D. Ross, R. W. Woody, Intrinsic protein disorder, amino acid composition, and histone terminal domains. *J. Biol. Chem.* **281**, 1853–1856 (2006).

81. P. Tompa, M. Fuxreiter, Fuzzy complexes: Polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.* **33**, 2–8 (2008).

82. H. A. Moesa, S. Wakabayashi, K. Nakai, A. Patil, Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification. *Mol. Biosyst.* **8**, 3262–3273 (2012).

83. A. S. Holehouse, R. K. Das, J. N. Ahad, M. O. Richardson, R. V. Pappu, CIDER: Resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophys. J.* **112**, 16–21 (2017).

84. L. Sawle, K. Ghosh, A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J. Chem. Phys.* **143**, 085101 (2015).

85. J. Nilsson, M. Grahn, A. P. Wright, Proteome-wide evidence for enhanced positive darwinian selection within intrinsically disordered regions in proteins. *Genome Biol.* **12**, R65 (2011).

86. A. Schlessinger, C. Schaefer, E. Vicedo, M. Schmidberger, M. Punta, B. Rost, Protein disorder—A breakthrough invention of evolution? *Curr. Opin. Struct. Biol.* **21**, 412–418 (2011).

87. M. Pajkos, B. Mészáros, I. Simon, Z. Dosztányi, Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Mol. Biosyst.* **8**, 296–307 (2012).

88. J. D. Forman-Kay, T. Mittag. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure* **21**, 1492–1499 (2013).

89. C. Angermueller, D. Dohan, D. Belanger, R. Deshpande, K. Murphy, L. Colwell, Model-based reinforcement learning for biological sequence design, in *International Conference on Learning Representations (ICLR)*, A. Rush, Ed. (ICLR, 2020), pp. 1–16.

90. Y. Wang, H. Tang, L. Huang, L. Pan, L. Yang, H. Yang, F. Mu, M. Yang, Self-play reinforcement learning guides protein engineering. *Nat. Mach. Intell.* **5**, 845–860 (2023).

91. Z. Yang, K. A. Milas, A. D. White, Now what sequence? Pre-trained ensembles for Bayesian optimization of protein sequences. *bioRxiv* 2022.08.05.502972 [Preprint] (2022).

92. E. I. Shakhnovich, A. Gutin, A new approach to the design of stable proteins. *Protein Eng.* **6**, 793–800 (1993).

93. N. Kirby, N. Cowieson, A. M. Hawley, S. T. Mudie, D. J. McGillivray, M. Kusel, V. Samardzic-Boban, T. M. Ryan, Improved radiation dose efficiency in solution SAXS using a sheath flow sample environment. *Acta Crystallogr. D Struct. Biol.* **72**, 1254–1266 (2016).

94. J. B. Hopkins, R. E. Gillilan, S. Skou, *BioXTAS RAW*: Improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. *J. Appl. Crystallogr.* **50**, 1545–1553 (2017).

95. J. A. Riback, M. A. Bowman, A. M. Zmyslowski, C. R. Knoverek, J. M. Jumper, J. R. Hinshaw, E. B. Kaye, K. F. Freed, P. L. Clark, T. R. Sosnick, Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **358**, 238–241 (2017).

96. D. Wu, A. Chen, C. S. Johnson, An improved diffusion-ordered spectroscopy experiment incorporating bipolar-gradient pulses. *J. Magn. Reson. A* **115**, 260–264 (1995).

97. S. Leeb, J. Danielsson, Obtaining hydrodynamic radii of intrinsically disordered protein ensembles by pulsed field gradient NMR measurements. *Methods Mol. Biol.* **2141**, 285–302 (2020).

98. E. O. Stejskal, J. E. Tanner, Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. *J. Chem. Phys.* **42**, 288–292 (1965).

99. A. Prestel, K. Bugge, L. Staby, R. Hendus-Altenburger, B. B. Kragelund, Characterization of dynamic IDP complexes by NMR spectroscopy. *Methods Enzymol.* **611**, 193–226 (2018).

100. A. G. Kikhney, C. R. Borges, D. S. Molodenskiy, C. M. Jeffries, D. I. Svergun, SASBDB: Towards an automatically curated and validated repository for biological scattering data. *Protein Sci.* **29**, 66–75 (2020).

101. N. M. Milkovic, T. Mittag, Determination of protein phase diagrams by centrifugation. *Methods Mol. Biol.* **2141**, 685–702 (2020).

102. P. J. Fleming, J. J. Correia, K. G. Fleming, Revisiting macromolecular hydration with HullRadSAS. *Eur. Biophys. J.* **52**, 215–224 (2023).

103. W.-Y. Choy, F. A. Mulder, K. A. Crowhurst, D. Muhandiram, I. S. Millett, S. Doniach, J. D. Forman-Kay, L. E. Kay, Distribution of molecular size within an unfolded state ensemble using small-angle X-ray scattering and pulse field gradient NMR techniques. *J. Mol. Biol.* **316**, 101–112 (2002).

104. M. C. Ahmed, R. Crehuet, K. Lindorff-Larsen. Computing, analyzing, and comparing the radius of gyration and hydrodynamic radius in conformational ensembles of intrinsically disordered proteins. *Methods Mol. Biol.* **2141**, 429–445 (2020).

105. W. Li, A. Godzik. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

106. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

107. S. Grudinin, M. Garkavenko, A. Kazennov, Pepsi-SAXS: An adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. *Acta Crystallogr. D Struct. Biol.* **73**, 449–464 (2017).

108. A. H. Larsen, M. C. Pedersen, Experimental noise in small-angle scattering can be assessed using the Bayesian indirect Fourier transformation. *J. Appl. Crystallogr.* **54**, 1281–1289 (2021).

109. S. Hansen, BayesApp: A web site for indirect transformation of small-angle scattering data. *J. Appl. Crystallogr.* **45**, 566–567 (2012).

110. J. Trewhella, A. P. Duff, D. Durand, F. Gabel, J. M. Guss, W. A. Hendrickson, G. L. Hura, D. A. Jacques, N. M. Kirby, A. H. Kwan, J. Pérez, L. Pollack, T. M. Ryan, A. Sali, D. Schneidman-Duhovny, T. Schwede, D. I. Svergun, M. Sugiyama, J. A. Tainer, P. Vachette, J. Westbrook, A. E. Whitten, 2017 publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution: An update. *Acta Crystallogr. D Struct. Biol.* **73**, 710–728 (2017).

111. J. Trewhella, C. M. Jeffries, A. E. Whitten, 2023 update of template tables for reporting biomolecular structural modelling of small-angle scattering data. *Acta Crystallogr. D Struct. Biol.* **79**, 122–132 (2023).

112. N. R. Hajizadeh, D. Franke, C. M. Jeffries, D. I. Svergun, Consensus Bayesian assessment of protein molecular mass from solution X-ray scattering data. *Sci. Rep.* **8**, 7204 (2018).