Deciphering how naturally occurring sequence features impact the phase behaviors of disordered prion-like domains

Supplementary Information

Includes Supplementary Methods and details regarding the analyses as well as Supplementary Tables 1 and 2 and Supplementary Figures 1 – 7

Supplementary Methods

Details of constructs used in the current study: All A1-LCD variants were based on the LCD (residues 186-320) from human hnRNPA1 (UniProt: P09651; Isoform A1-A). The coding sequences for the variants were synthesized (by Thermo Fisher or Genscript) 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. In the expressed protein, the N-terminal 6xHis-tag was cleaved using the TEV protease cleavage site, leaving only an additional GS sequence at the N-terminus of each of the 38 constructs (underlined in Table S1). Amino acid sequence details for each of the constructs are shown in Table S1 below.

Supplementary Table 1: Amino acid sequences of A1-LCD and designed variants

Construct	Amino acid sequence
A1-LCD^{-NLS}	GSMASASSSQ RGRSGSGNFG GGRGGGFGGN DNFGRGGNFS GRGGFGGSRG GGGYGGSGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF GPMKGGNFGG RSSGGSGGGG QYFAKPRNQG GYGGSSSSSS YGSGRRF
A1-LCD^{+NLS}	GSMASASSSQ RGRSGSGNFG GGRGGGFGGN DNFGRGGNFS GRGGFGGSRG GGGYGGSGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF GPMKGGNFGG RSSGPYGGGG QYFAKPRNQG GYGGSSSSSS YGSGRRF
A1-LCD$^{-12F+12Y}$	GSMASASSSQ RGRSGSGNYG GGRGGGYGGN DNYGRGGNYS GRGGYGGSRG GGGYGGSGDG YNGYGNDGSN YGGGGSYNDY GNYNNQSSNY GPMKGGNYGG RSSGGSGGGG QYYAKPRNQG GYGGSSSSSS YGSGRRY
A1-LCD^{+7F-7Y}	GSMASASSSQ RGRSGSGNFG GGRGGGFGGN DNFGRGGNFS GRGGFGGSRG GGGFGGSGDG FNGFGNDGSN FGGGGSFNDF GNFNNQSSNF GPMKGGNFGG RSSGGSGGGG QFFAKPRNQG GFGGSSSSSS FGSGRRF
A1-LCD^{-9F+6Y}	GSMASASSSQ RGRSGSGNFG GGRGGGYGGN DNYGRGGNYS GRGGFGGSRG GGGYGGSGDG YNGGGNDGSN YGGGGSYNDS GNYNNQSSNF GPMKGGNYGG RSSGGSGGGG QYGAKPRNQG GYGGSSSSSS YGSGRRY
A1-LCD^{-8F+4Y}	GSMASASSSQ RGRSGSGNFG GGRGGGYGGN DNGGRGGNYS GRGGFGGSRG GGGYGGSGDG YNGGGNDGSN YGGGGSYNDS GNYNNQSSNF GPMKGGNYGG RSSGGSGGGG QYGAKPRNQG GYGGSSSSSS YGSGRRF
A1-LCD^{-9F+3Y}	GSMASASSSQ RGRSGSGNFG GGRGGGYGGN DNGGRGGNYS GRGGFGGSRG GGGYGGSGDG YNGGGNDGSN YGGGGSYNDS GNGNNQSSNF GPMKGGNYGG RSSGGSGGGG QYGAKPRNQG GYGGSSSSSS YGSGRRS
A1-LCD^{-10R}	GSMASASSSQ GGSSGSGNFG GGGGGGFGGN DNFGGGGNFS GSGGFGGSGG GGGYGGSGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF GPMKGGNFGG SSSGPYGGGG QYFAKPGNQG GYGGSSSSSS YGSGGGF
A1-LCD^{-6R}	GSMASASSSQ GGRSGSGNFG GGRGGGFGGN DNFGGGGNFS GSGGFGGSRG GGGYGGSGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF GPMKGGNFGG SSSGPYGGGG QYFAKPGNQG GYGGSSSSSS YGSGGRF
A1-LCD^{+2R}	GSMASASSSQ RGRSGSGNFG GGRGGGFGGN DNFGRGGNFS GRGGFGGSRG GGGYGGSGDG YNGFRNDGSN FGGGGRYNDF GNYNNQSSNF GPMKGGNFGG RSSGPYGGGG QYFAKPRNQG GYGGSSSSSS YGSGRRF
A1-LCD^{+7R}	GSMASASSSQ RGRSGRGNFG GGRGGGFGGN DNFGRGGNFS GRGGFGGSRG GGRYGGSGDR YNGFGNDGRN FGGGGSYNDF GNYNNQSSNF GPMKGGNFRG RSSGPYGRGG QYFAKPRNQG GYGGSSSSRS YGSGRRF
A1-LCD^{-2K}	GSMASASSSQ RGRSGSGNFG GGRGGGFGGN DNFGRGGNFS GRGGFGGSRG GGGYGGSGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF GPMGGGNFGG RSSGPYGGGG QYFAGPRNQG

	GYGGSSSSSS YGSGRRF
A1-LCD^{-3R+3K}	GSMASASSSQ RGKSGSGNFG GGRGGGFGGN DNFGRGGNFS GRGGFGGSKG GGGYGGSGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF GPMKGGNFGG RSSGGSGGGG QYFAKPRNQG GYGGSSSSSS YGSGRKF
A1-LCD^{-6R+6K}	GSMASASSSQ KGKSGSGNFG GGRGGGFGGN DNFGKGGNFS GRGGFGGSKG GGGYGGSGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF GPMKGGNFGG KSSGGSGGGG QYFAKPRNQG GYGGSSSSSS YGSGRKF
A1-LCD$^{-10R+10K}$	GSMASASSSQ KGKSGSGNFG GGKGGGFGGN DNFGKGGNFS GKGGFGGSKG GGGYGGSGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF GPMKGGNFGG KSSGGSGGGG QYFAKPKNQG GYGGSSSSSS YGSGKKF
A1-LCD^{-4D}	GSMASASSSQ RGRSGSGNFG GGRGGGFGGN GNFGRGGNFS GRGGFGGSRG GGGYGGSGGG YNGFGNSGSN FGGGGSYNGF GNYNNQSSNF GPMKGGNFGG RSSGPYGGGG QYFAKPRNQG GYGGSSSSSS YGSGRRF
A1-LCD^{+4D}	GSMASASSSQ RDRSGSGNFG GGRGGGFGGN DNFGRGGNFS GRGDFGGSRG GGGYGGSGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF GPMKGGNFGG RSSDPYGGGG QYFAKPRNQG GYGGSSSSSS YDSGRRF
A1-LCD^{+8D}	GSMASASSSQ RDRSGSGNFG GGRDGGFGGN DNFGRGDNFS GRGDFGGSRD GGGYGGSGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF GPMKGGNFGG RSSDPYGGGG QYFAKPRNQD GYGGSSSSSS YDSGRRF
A1-LCD^{+12D}	GSMASADSSQ RDRDDSGNFG DGRGGGFGGN DNFGRGGNFS DRGGFGGSRG DGGYGGDGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF DPMKGGNFGD RSSGPYDGGG QYFAKPRNQG GYGGSSSSSS YGSDRRF
A1-LCD^{+12E}	GSMASAESSQ REREESGNFG EGRGGGFGGN DNFGRGGNFS ERGGFGGSRG EGGYGGEGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF EPMKGGNFGE RSSGPYEGGG QYFAKPRNQG GYGGSSSSSS YGSERRF
A1-LCD$^{+7R+10D}$	GSMASADSSQ RDRDGRGNFG DGRGGGFGGN DNFGRGGNFS DRGGFGGSRG GGRYGGDGDR YNGFGNDGRN FGGGGSYNDF GNYNNQSSNF DPMKGGNFRD RSSGPYDRGG QYFAKPRNQG GYGGSSSSRS YGSDRRF
A1-LCD$^{+7R+12D}$	GSMASADSSQ RDRDDRGNFG DGRGGGFGGN DNFGRGGNFS DRGGFGGSRG DGRYGGDGDR YNGFGNDGRN FGGGGSYNDF GNYNNQSSNF DPMKGGNFRD RSSGPYDRGG QYFAKPRNQG GYGGSSSSRS YGSDRRF
A1-LCD$^{+7K+12D}$	GSMASADSSQ RDRDDKGNFG DGRGGGFGGN DNFGRGGNFS DRGGFGGSRG DGKYGGDGDK YNGFGNDGKN FGGGGSYNDF GNYNNQSSNF DPMKGGNFKD RSSGPYDKGG QYFAKPRNQG GYGGSSSSKS YGSDRRF
A1-LCD$^{-2R-2K+3D}$	GSMASASSSQ DGRSGSGNFG GGRGGGFGGN DNFGRGGNFS GRGGFGGSRG GGGYGGSGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF GPMDGGNFGG RSSGPYGGGG QYFADPRNQG GYGGSSSSSS YGSGGRF
A1-LCD$^{-4R-2K+5D}$	GSMASASSSQ DGRSGSGNFG GGDGGGFGGN DNFGRGGNFS GGGGFGGSRG GGGYGGSGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF GPMDGGNFGG RSSGPYGGGG QYFADPRNQG GYGGSSSSSS YGSGDRF
A1-LCD$^{-10G+10S}$	GSMASASSSQ RSRSGSGNFG GGRSGGFGGN DNFGRSGNFS GRGGFGGSRG GGGYGGSGDS YNGFGNDGSN FGGSGSYNDF GNYNNQSSNF GPMKSGNFGG RSSGSSGGSG QYFAKPRNQG SYSGSSSSSS YGSGRRF
A1-LCD$^{-20G+20S}$	GSMASASSSQ RSRSGSGNFS GSRSGSFSGN DNFGRSGNFS GRSGFGGSRS GGGYSGSGDS YNSFGNDGSN FSGSGSYNDF GNYNNQSSNF GPMKSGNFGG RSSGSSGGSG QYFAKPRNQG SYSGSSSSSS YGSSRRF
A1-LCD$^{-30G+30S}$	GSMASASSSQ RSRSSSGNFS GSRSGSFSGN DNFGRSGNFS GRSGFSGSRS GSGYSGSSDS YNSFGNDSSN FSGSSSYNDF GNYNNQSSNF GPMKSGNFSG RSSSSSGSSG QYFAKPRNQG SYSGSSSSSS YSSSRRF
A1-LCD$^{+23G-23S}$	GSMAGAGGGQ RGRGGGGNFG GGRGGGFGGN DNFGRGGNFG GRGGFGGGRG GGGYGGGGDG YNGFGNDGGN FGGGGGYNDF GNYNNQGGNF GPMKGGNFGG RGGGGGGGGG QYFAKPRNQG GYGGGGGGGG YGGGRRF
A1-LCD$^{-30G+30S+7F-7Y}$	GSMASASSSQ RSRSSSGNFS GSRSGSFSGN DNFGRSGNFS GRSGFSGSRS GSGFSGSSDS FNSFGNDSSN FSGSSSFNDF GNFNNQSSNF GPMKSGNFSG RSSSSSGSSG QFFAKPRNQG SFSGSSSSSS FSSSRRF
A1-LCD$^{-30G+30S-12F+12Y}$	GSMASASSSQ RSRSSSGNYS GSRSGSYSGN DNYGRSGNYS GRSGYSGSRS GSGYSGSSDS YNSYGNDSSN YSGSSSYNDY GNYNNQSSNY GPMKSGNYSG RSSSSSGSSG QYYAKPRNQG SYSGSSSSSS YSSSRRY

A1-LCD$^{-20G+20S+7F-7Y}$	GSMASASSSQ RSRSGSGNFS GSRSGSFSGN DNFGRSGNFS GRSGFGGSRS GGGFSGSGDS FNSFGNDGSN FSGSGSFNDF GNFNNQSSNF GPMKSGNFGG RSSGSSGGSG QFFAKPRNQG SFSGSSSSSS FGSSRRF
A1-LCD$^{-20G+20S-12F+12Y}$	GSMASASSSQ RSRSGSGNYS GSRSGSYSGN DNYGRSGNYS GRSGYGGSRS GGGYSGSGDS YNSYGNDGSN YSGSGSYNDY GNYNNQSSNY GPMKSGNYGG RSSGSSGGSG QYYAKPRNQG SYSGSSSSSS YGSSRRY
A1-LCD$^{+23G-23S+7F-7Y}$	GSMAGAGGGQ RGRGGGGNFG GGRGGGFGGN DNFGRGGNFG GRGGFGGGRG GGGFGGGGDG FNGFGNDGGN FGGGGGFNDF GNFNNQGGNF GPMKGGNFGG RGGGGGGGGG QFFAKPRNQG GFGGGGGGGG FGGGRRF
A1-LCD$^{+23G-23S-12F+12Y}$	GSMAGAGGGQ RGRGGGGNYG GGRGGGYGGN DNYGRGGNYG GRGGYGGGRG GGGYGGGGDG YNGYGNDGGN YGGGGGYNDY GNYNNQGGNY GPMKGGNYGG RGGGGGGGGG QYYAKPRNQG GYGGGGGGGG YGGGRRY
A1-LCD$^{-14N-4Q+18G}$	GSMASASSSG RGRSGSGGFG GGRGGGFGGG DGFGRGGGFS GRGGFGGSRG GGGYGGSGDG YGGFGGDGSG FGGGGSYGDF GGYGGGSSGF GPMKGGGFGG RSSGGSGGGG GYFAKPRGGG GYGGSSSSSS YGSGRRF
A1-LCD$^{-14N+14Q}$	GSMASASSSQ RGRSGSGQFG GGRGGGFGGQ DQFGRGGQFS GRGGFGGSRG GGGYGGSGDG YQGFGQDGSQ FGGGGSYQDF GQYQQQSSQF GPMKGGQFGG RSSGGSGGGG QYFAKPRQQG GYGGSSSSSS YGSGRRF
A1-LCD$^{-23S+23T}$	GSMATATTTQ RGRTGTGNFG GGRGGGFGGN DNFGRGGNFT GRGGFGGTRG GGGYGGTGDG YNGFGNDGTN FGGGGTYNDF GNYNNQTTNF GPMKGGNFGG RTTGGTGGGG QYFAKPRNQG GYGGTTTTTT YGTGRRF

Protein expression and purification: All hnRNPA1-LCD variants were expressed in *E. coli* BL21-Gold (DE3) strain in ZYM5052 auto induction media at 37°C for 24 hours. For NMR samples, cultures were grown in isotopically labeled M9 media, induced at OD_{600}=0.8 with 1 mM IPTG and cultured at 37°C for an additional 6 hours. Cell pellets were resuspended in 50 mM MES pH 6.0, 500 mM NaCl, 20 mM 2-mercaptoethanol and lysed via sonication. Cell lysates were centrifuged, and the variants were purified from insoluble inclusion bodies as previously described [1]. The inclusion bodies were resuspended in 6 M GdmHCl, 20 mM Tris pH 7.5, 15 mM imidazole overnight at 4°C. Solutions of solubilized inclusion bodies were cleared 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 4 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. As a final purification step, the samples were passed in 2 M GdmHCl, 20 mM MES pH 5.5 over a S75 Superdex size exclusion column (GE Healthcare). The identity of each protein was confirmed via intact mass spectrometry. All proteins were stored in 4 M GdmHCl, 20 mM MES pH 5.5 at 4°C. For the -2R-2K+3D construct, the procedure was modified as follows. The sample was cleaved in a minimum of 30 mL of buffer per 1 L of culture. Following the post-cleavage nickel column, the sample was rapidly exchanged into 20 mM MES pH 5.5 and 6 M GdmHCl using a 10K MWCO 15 mL Amicon centrifugal filter prior to size exclusion by a Superdex 75 column to avoid the protein being concentrated at a pH near its theoretical pI. For the -4R-2K+5D construct the procedure was modified such that the protein was cleaved in a minimum of 30 mL of buffer per 1 L of culture. The sample pH was then rapidly increased after the post-cleave nickel column by adding 1/10 volume of 1 M CAPS pH 10.5 before concentrating with a 10K MWCO 15 mL Amicon centrifugal filter. The sample was then subjected to size exclusion chromatography using a Superdex 75 column equilibrated with 20 mM CAPS pH 10.5, 2 M GdmHCl.

Buffer exchange to remove denaturant: Buffer exchange was achieved in two-steps. First, the protein in 4 M GdmHCl, 20 mM MES pH 5.5 storage buffer was exchanged into 1 M MES pH 5.5 by multiple dilution and concentration steps using a 3K MWCO Amicon centrifugal filter as previously described [1]. The protein was then dialyzed overnight against 20 mM HEPES pH 7.0 (without excess salt) at room temperature. The pH of the buffer was adjusted using ammonium hydroxide to prevent the

introduction of excess salt into the sample. The protein was filtered through a 0.22 mm Millex-GV filter (Merck) to remove potential aggregates from the solution, which might have formed during dialysis.

SDS-PAGE: All gel electrophoresis was carried out using NuPAGE 4-12% Bis-Tris gradient gels (Invitrogen). The gels were run using NuPAGE MES SDS Running buffer (Invitrogen) diluted to 1x until the dye-front had traveled a suitable distance. The 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.

Measurements of saturation concentrations for specific variants that required special handling: Experiments on variant +8D was carried out in 20 mM HEPES, 150 mM NaCl pH 8.0 because of its net neutral charge. Because +7F-7Y lacks Tyr residues, its protein concentration was determined at 205 nm using a Cary 300 UV-Vis spectrophotometer (Agilent). For determination of low protein concentrations, a 10 mm pathlength quartz cuvette was used.

We also measured saturation concentrations for variant +7K+12D as a function of pH. The sample of variant +7K+12D in denaturing buffer was rapidly exchanged into non-denaturing buffers using Zeba spin columns (Thermo Fischer) following standard procedures. The columns were equilibrated with buffers prepared at room temperature to contain 150 mM NaCl and 20 mM of one of the following buffering agents, MES at pH 5.5, and 6.5, HEPES at pH 6.5, 7, and 8, Tris at pH 8, and 9, and HEPBS at pH 8, 8.3, 8.7, and 9. The pH of each buffering condition was measured at 4ºC to account for temperature dependent pK_a shifts of the buffer. The saturation concentration at 4ºC was then measured by separating dilute and dense phase by centrifugation as described in the Methods. All measurements were done as at least 3 replicates. The theoretical protein net charge at each pH was calculated using protpi.ch. The protonated state of the lysine side chain at each pH was calculated using the estimated net charge of the protein and accounting for the charge state of the amino terminal at the given pH, assuming the remainder of the charge difference results from deprotonated Lys residues.

Far UV-CD spectra: CD spectra were recorded with a J-1500 spectrophotometer (Jasco). ~0.5mg/mL protein solutions were measured in a 0.1 mm pathlength cuvette (Hellma). The spectra were accumulated with a response time of 4 s, 1 nm data pitch, 1 nm band width from 195 to 260 nm. The CD spectra were collected at 25ºC and at least two replicates were measured and averaged for each condition.

Small Angle X-ray Scattering (*SAXS*) *measurements*: All measurements were performed at BioCat (beamline 18ID at the Advanced Photon Source, Chicago) with in-line size exclusion chromatography (SEC-SAXS) as previously reported [1, 2]. Experiments were conducted at room temperature in 20 mM HEPES, 150 mM NaCl, pH 7.0. Protein samples stored in 4 M GdmHCl, 20 mM MES pH 5.5 were loaded onto either a Superdex 75 5/150 GL or a Superdex 75 Increase 10/300 column (GE Life Science) with a flow rate of 0.4 mL/min. The column eluent passed through the UV monitor and proceeded through the SAXS sheath flow capillary in the coflow system [3]. Scattering intensity was recorded using a Pilatus3 1 M (Dectris) detector placed 3.5 m from the sample providing a q-range of 0.004-0.4 $Å^{-1}$. Exposure time was 0.5 sec. Raw SAXS data was reduced at the beamline using BioXTAS RAW 1.6.3 and 2.0.2 [4]. Buffer subtraction, Guinier fits, and Kratky transformations were performed using the BioXTAS Raw software [4]. Raw data were additionally fit using an empirically derived molecular form factor (MFF) developed by Riback et al. [5].

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 prepared by adding NaCl to 150 mM to the protein stock solution. Protein concentrations were selected such that they were slightly above their corresponding c_{sat} at 20°C. 2 μL of the protein solution was sandwiched between two coverslips sandwiched with 3M 300 LSE high-temperature double-sided tape (0.34 mm) with a window for microscopy cut out.

NMR spectroscopy: NOESY and TOCSY experiments for A1-LCD Δhexa were acquired on either a Bruker Avance 1.1 GHz or 850 MHz spectrometer equipped with TCI triple-resonance cryogenic probes and pulse-field gradient units. A ^{13}C-resolved $^{1}H^{aromatic}$-$^{1}H^{aliphatic}$ NOESY spectrum (64 scans,

2048 (^{1}H) × 64 (^{13}C) × 80 (^{1}H) complex data points, with 14.2 ppm, 16.0 ppm, and 1.5 ppm as ^{1}H, ^{13}C and ^{1}H sweep width, respectively) and a mixing time of 250 ms were measured at 1.1 MHz and 313 K in 20 mM HEPES, 200 mM NaCl, 5 mM TCEP, 150 μM DSS, and 5% D_2O at pH 6.3. ^{13}C editing was not necessary for aromatic protons as they are well separated, but all post-NOE protons were ^{13}C resolved. The concentration of the Δhexa LCD was ~800 μM. 2D planes from the 3D spectra corresponding to the arginine δ-position were assessed for NOEs between the arginine δ-proton and the aromatic protons. A triple resonance (H)CC(CO)NH spectrum (16 scans, 2048 (^{1}H) × 80 (^{13}C) × 200 (^{1}H) complex data points, with 16.3 ppm, 22.0 ppm, and 54.0 ppm as the ^{1}H, ^{13}C and ^{1}H sweep width, respectively) was used to assign the arginine sidechain frequencies. Data were processed using BRUKER Topspin version 4.0, NMRPipe version 10.4 [6] and analyzed using NMRfam SPARKY [7]. All spectra were referenced directly using DSS for the ^{1}H dimension; ^{13}C and ^{15}N frequencies were referenced indirectly.

NMR data for A1-LCD +7K+12D were acquired on Bruker Avance 600 and 800 MHz spectrometers equipped with TCI triple-resonance cryogenic probes and pulsed-field gradient units. All samples were prepared in a buffer consisting of 20 mM HEPES pH 6.8, 0.5 mM EDTA and 10% D_2O at 20°C. For assignment, samples of ^{15}N,^{13}C +7K+12D with concentrations between 85 and 220 μM were used to acquire standard triple-resonance backbone assignment experiments based on a sensitivity enhanced ^{1}H-^{15}N HSQC (32 scans, 2048 x 512 complex data points, with 12 ppm and 20 ppm as ^{1}H and ^{15}N sweep widths). These included HNCACB and CBCA(CO)NH (16 scans, 2048 (^{1}H) × 64 (^{15}N) × 128 (^{13}C) complex data points, with 12 ppm, 20 ppm, and 72 ppm as ^{1}H, ^{15}N and ^{13}C sweep width, respectively), HN(CA)CO (16 scans, 2048 (^{1}H) × 64 (^{15}N) × 128 (^{13}C) complex data points, with 10 ppm, 20 ppm, and 72 ppm as ^{1}H, ^{15}N and ^{13}C sweep widths, respectively), HNCO (8 scans, 2048 (^{1}H) × 64 (^{15}N) × 80 (^{13}C) complex data points, with 10 ppm, 20 ppm, and 14 ppm as ^{1}H, ^{15}N and ^{13}C sweep widths, respectively), and NH(CA)NNH (32 scans, 2048 (^{1}H) × 50 (^{15}N) × 100 (^{15}N) complex data points, with 10 ppm, 20 ppm, and 22 ppm as ^{1}H, ^{15}N F1 and ^{15}N F2 sweep widths, respectively) spectra.

Data were processed using BRUKER Topspin version 3.2, or NMRPipe (v.7.9) [6] and analyzed using NMRviewJ [8]. All spectra were referenced directly using DSS for the ^{1}H dimension, ^{13}C and ^{15}N frequencies were referenced indirectly.

^{15}N R_2 relaxation experiments were acquired at 600 MHz at 293 K using a standard Carr-Purcell-Meiboom-Gill (CPMG)-based Bruker pulse program (32 scans, 2048 (^{1}H) x 256 (^{15}N) complex data points) with the following delays of 16.8, 33.5, 67, 100.5, 134.1, 167.6, 251.4, and 335.2 ms with a recovery delay of 3 s. Due to significant overlap in the spectra Peakipy was used to attempt to deconvolute the peak intensities. The relaxation rates for residues 13, 15, 17, 19, 20, 21, 23, 25, 26, 27, 29, 31, 36, 37, 38, 39, 40, 42, 47, 51, 52, 53, 56, 57, 58, 59, 60, 61, 62, 66, 67, 68, 71, 74, 77, 78, 83, 86, 88, 89, 90, 93, 96, 97, 98, 100, 103, 110, 121, 122, 124, 125, 126, 127, 130, 131, and 132 were omitted as the overlap was too great for deconvolution. Fitting of the R_2 rate profile to an analytical model was done as previously described [1]. To account for gaps in experimental data, the maximum cluster height was limited to 9 s^{-1}.

Bioinformatics analysis: Homologous sequences were identified using version 5.0 of the EggNOG database [9]. Sequences of homologs were aligned using the EMBL-EBI Clustal Omega tool [10]. The sequences were then trimmed to include only intrinsically disordered regions using the UniProt annotation [11] of the canonical human isoform. Sequence analyses were performed using the localCIDER tool [12]. The sequences culled for LCDs from homologs of hnRNPA1 and FUS / FET family proteins are included in separate Supplementary spreadsheets.

Analysis of sequence compositions: Pairwise compositional similarities were determined in the following manner: (1) For each sequence, we create a 20×1 compositional vector. Each vector is of the form: (f_A, f_C, … f_Y). Here, each element quantifies the fraction of each amino acid within a sequence. (2) For a pair of compositional vectors, we compute the dot product of the two vectors. (3) Next, we divide the dot product by the product of the magnitudes of the vectors. This gives a value between 0 and 1, where values closer to 1 indicate higher compositional similarity. In Fig. 1b, Fig. 3a,b, Fig. 7a-c,

Extended Fig. 1b,c, and Extended Fig. 6, the sequences analyzed share a compositional similarity with the WT sequence of at least 0.8 (770 sequences) to limit the effects of strong outliers and/or partial homologs on the sequence analyses.

Analysis of c_{sat} data using a mean-field, stickers-and-spacers model: Wang et al. [13] adapted the mean-field stickers and spacers model of Semenov and Rubinstein [14] to a system with n_A stickers of type A and n_B stickers of type B. In this model, the saturation concentration c_{sat} was shown to be proportional to $(n_A n_B)^{-1}$ providing the heterotypic interactions among A and B stickers are the only determinants of c_{sat}. The model of Wang et al. [13] was generalized by Choi et al. [15] to account for the competing effects of homotypic A-A and B-B interactions. Additionally, Choi et al., accounted for cooperative effects, whereby the strengths of inter-sticker interactions can either be enhanced or weakened due to the influence of three-body interactions on the strengths of inter-sticker interactions. We adapted the approach of Choi et al. [15] to obtain rescaled values of c_{sat} that were analyzed as a function of NCPR. Here, we choose the aromatic residues (Tyr / Phe) as the primary stickers and Arg as auxiliary stickers.

In the generalized mean-field model of Choi et al. [15], c_{sat} is governed by the numbers of aromatic residues (n_a), the numbers of Arg residues (n_R), and the strengths of inter-aromatic (λ_{aa}) and aromatic-Arg (λ_{aR}) interactions. Note that the λ-values are dimensionless quantities. Accordingly, the functional form for c_{sat}, written in terms of a multiplicative constant is as shown in Equation (1).

$$c_{sat} = k\left(\lambda_{aa} n_a^2 + 2\lambda_{aR} n_a n_R\right); \quad (1)$$

Here, k is a constant that converts the right-hand side into units of concentrations. The value for k can be extracted by linear regression of measured c_{sat} values plotted against the quantity in the parenthesis on the right-hand side of Equation (1) [13]. In our analysis, we focus on a rescaling of c_{sat} and therefore we do not need an estimate for k. The first step in the rescaling, which accounts for the contributions of aromatic and Arg residues as stickers is written as:

$$c_{sc,1} = c_{sat}\left(\lambda_{aa} n_a^2 + 2\lambda_{aR} n_a n_R\right); \quad (2)$$

Note that we set λ_{aa} = 1 and hence the only free parameter in the regression analysis is the value of λ_{aR}. If the only determinants of c_{sat} were inter-sticker interactions, then the expectation would be that, with appropriate parameterization of λ_{aR}, the values of $c_{sc,1}$ would be similar to one another for all A1-LCD variants. Instead, we observe the emergence of a V-shaped profile for $c_{sc,1}$ plotted against NCPR, especially for the Arg and Asp/Glu variants (Fig. 5c).

We also notice that the Lys variants have considerably higher $c_{sc,1}$ values than would be expected based on the numbers of aromatic and Arg residues in these variants. The implication is that Lys plays a distinctive role, not just as a high-excluded volume spacer, but also in terms of its impact on the strengths of inter-sticker interactions. This model emerges from findings regarding the influence of positive and / or negative cooperativity on pi-pi and cation-pi interactions [16]. Here, we reason, based on the model of Choi et al. [15] that Lys residues appear to weaken inter-sticker interactions via three-body interactions. This effect is captured in Equation (3) as:

$$c_{sc,2} = c_{sat}\left[n_a^2\left(\lambda_{aa} + 3\lambda_K n_K\right) + n_a n_R\left(2\lambda_{aR} + 6\lambda_K n_K\right)\right]; \quad (3)$$

Here, λ_K quantifies the extent to which Lys residues impact the effective strengths of inter-sticker interactions and n_K is the number of protonated Lys residues. Notice that the effects of Lys residues are incorporated as contributions that affect c_{sat} via three-body interactions. The impact of accounting for the destabilizing effects of Lys residues is summarized in Fig. 5d. This requires parameterization of λ_{aR} and λ_K. This two-parameter fit of Equation (3) shows that the rescaled c_{sat} values at 4˚C collapse onto the V-shape profile that is plotted against NCPR (Fig. 5e).

Finally, since we have measurements of c_{sat} at a series of different temperatures, we account for the temperature dependence by noting that the dilute arms of the binodals are linear on a semi-log scale implying that the temperature dependence of c_{sat} may be written as:

$$c_{sat}(T) = c_{sat}(T_0)\exp\left[-\left(\frac{T-T_0}{m}\right)\right]; \quad (4)$$

Here, $c_{sat}(T_0)$ is the c_{sat} value at the reference temperature of 277 K and T is the actual temperature at which c_{sat} is measured. We combine Equations (3) and (4) to arrive at a final rescaled form for c_{sat} plotted against NCPR to assess the extent to which the data can be collapsed onto a master V-shaped profile. The final rescaled form of c_{sat} takes the form:

$$c_{sc,3} = c_{sat}\exp\left[-\left(\frac{T-T_0}{m}\right)\right]\left[n_a^2\left(\lambda_{aa}+3\lambda_K n_K\right)+n_a n_R\left(2\lambda_{aR}+6\lambda_K n_K\right)\right]; \quad (5)$$

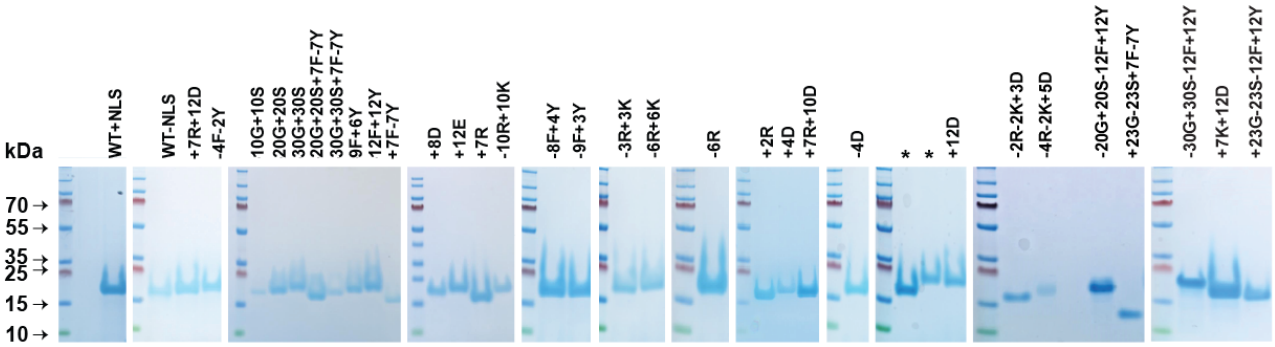
We applied Equation (5) to analyze the totality of variant-specific temperature dependent data for c_{sat}. The results are shown in Fig. 5e. Here, the dashed red lines show linear fits of each arm of the V-shaped plot. The associated Pearson *r*-values that quantify the linear correlation are also shown on the plot. In calculating the fits, the rescaled c_{sat} values for a given variant are averaged to one value so that each variant is weighted equally. The λ-values in Equations (2) – (5) were found by optimizing the Pearson correlation coefficients of the linear fits while keeping λ_{aa} fixed at unity. The m values in Equations (4) – (5) were determined by averaging the slopes of the dilute arms of the binodals of the relevant constructs. Accordingly, the fit of Equation (5) to all of the data has three free parameters *viz*., λ_{aR}, λ_K, and m. The parameters we obtain are: λ_{aR} = 1.69, λ_K = 0.0479, and m = 8.26 K.

We tested the accuracy of this model by overlaying $c_{sc,3}$ values for variants that were not used in the optimization. The results are shown in Fig. 5e. We find that the magnitudes of the Pearson *r*-values that quantify the strengths of linear correlations are still at least 0.95. The key message that is uncovered from the analysis in Fig. 5e and Equation (5) is that it helps us unmask the sticker and spacer determinants of the driving forces for phase separation. Importantly, it helps identify the contributions of NCPR to the driving forces for phase separation of PLCDs, even for PLCDs that are not enriched in charged residues.

Estimating c_{sat} values of A1-LCD homologs using our mean-field model: We applied our mean-field model to the set of A1-LCD homologs culled from our bioinformatics analysis to estimate their c_{sat} values at 4˚C. Specifically, we used the NCPR of a given homolog to estimate its $c_{sc,3}$ value based on the two linear fits in Fig. 5e. Next, we solved directly for c_{sat} in Equation (5) by inputting n_a, n_R, n_K, and the estimated $c_{sc,3}$ for the given homolog and setting T = 277 K. We restricted our analysis to homologs whose length ranged from 100 – 200 residues to limit the effects of length, since this is not directly accounted for in the mean-field model. The results are shown in Figure 5f. We find that the estimated c_{sat} values range over three orders of magnitude, demonstrating how similar sequences can display divergent phase separation behaviors based on small changes to sequence compositions.

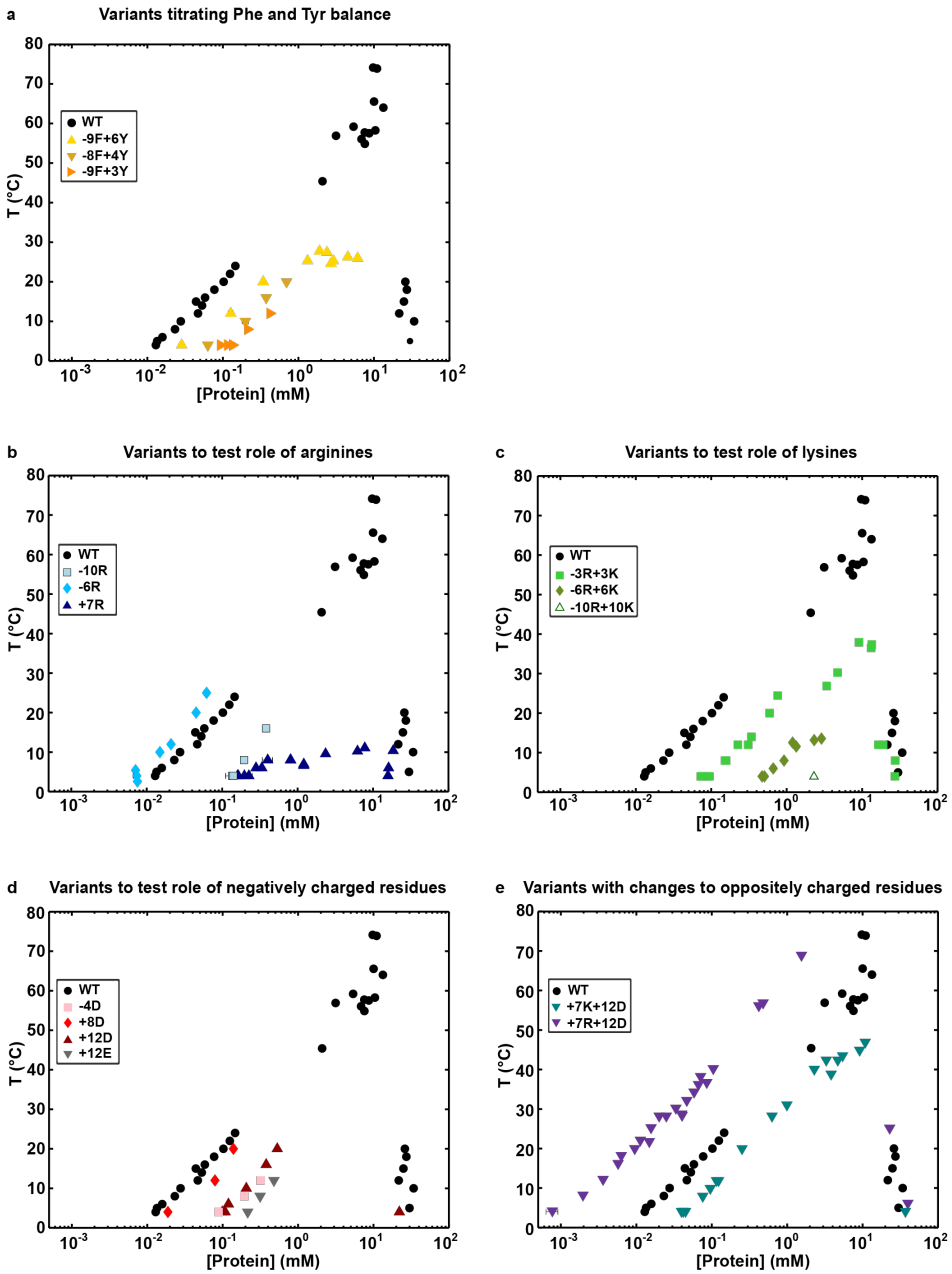
Supplementary Table 2: Estimated values for R_g and ν^{app} from analysis of SAXS data using an empirical molecular form factor.

Construct	R_g (Å)	R_g error	ν^{app}	Standard error in estimate of ν^{app}
A1-LCD^{-NLS}	27.60	0.16	0.442	0.006
A1-LCD^{+NLS}	25.83	0.11	0.430	0.004
A1-LCD$^{-12F+12Y}$	26.04	0.20	0.429	0.007
A1-LCD^{+7F-7Y}	27.18	0.13	0.454	0.006
A1-LCD^{-9F+6Y}	26.55	0.10	0.457	0.005
A1-LCD^{-8F+4Y}	27.07	0.07	0.461	0.003
A1-LCD^{-9F+3Y}	26.83	0.13	0.460	0.006
A1-LCD^{-10R}	26.71	0.07	0.468	0.004
A1-LCD^{-6R}	25.73	0.09	0.448	0.004
A1-LCD^{+2R}	26.23	0.23	0.440	0.009
A1-LCD^{+7R}	27.09	0.07	0.442	0.003
A1-LCD^{-3R+3K}	26.34	0.15	0.447	0.006
A1-LCD^{-6R+6K}	27.87	0.08	0.467	0.003
A1-LCD$^{-10R+10K}$	28.49	0.05	0.480	0.002
A1-LCD^{-4D}	26.42	0.12	0.446	0.005
A1-LCD^{+4D}	27.18	0.30	0.453	0.013
A1-LCD^{+8D}	26.85	0.07	0.437	0.003
A1-LCD^{+12D}	28.01	0.12	0.451	0.004
A1-LCD^{+12E}	28.52	0.05	0.457	0.002
A1-LCD$^{+7K+12D}$	29.21	0.08	0.467	0.003



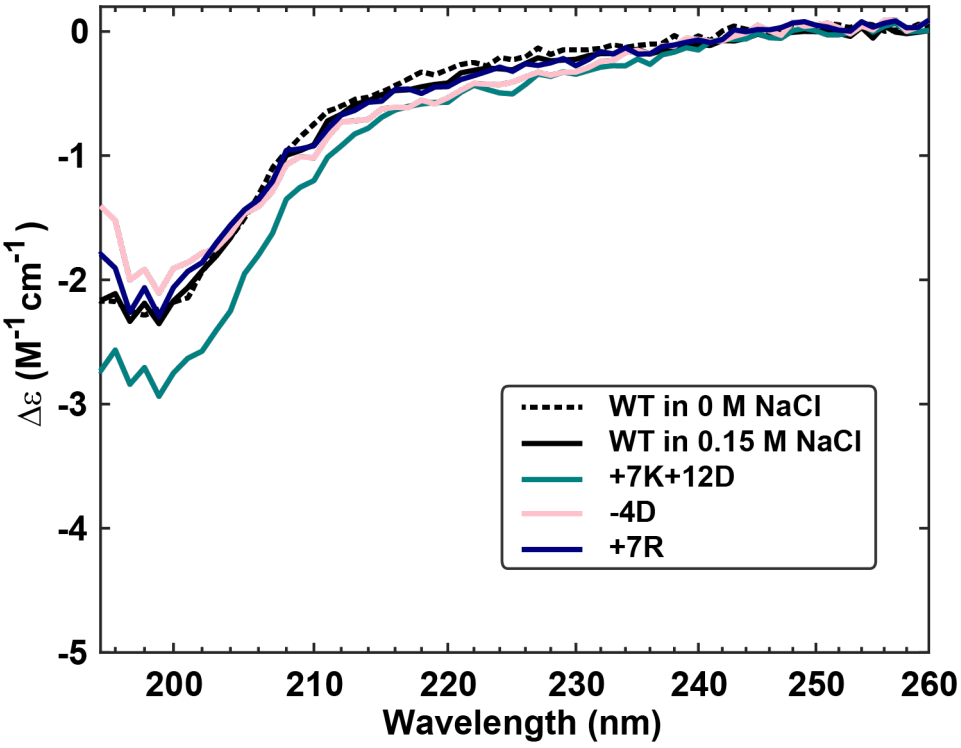
Supplementary Figure 1: SDS-PAGE analysis of purified A1-LCD variants used in this study.

Shifts in electrophoretic mobility are expected for certain mutations, particularly when multiple charged or bulky residues are substituted. This figure demonstrates the lack of contaminating proteins and that there is no evidence for proteolysis.



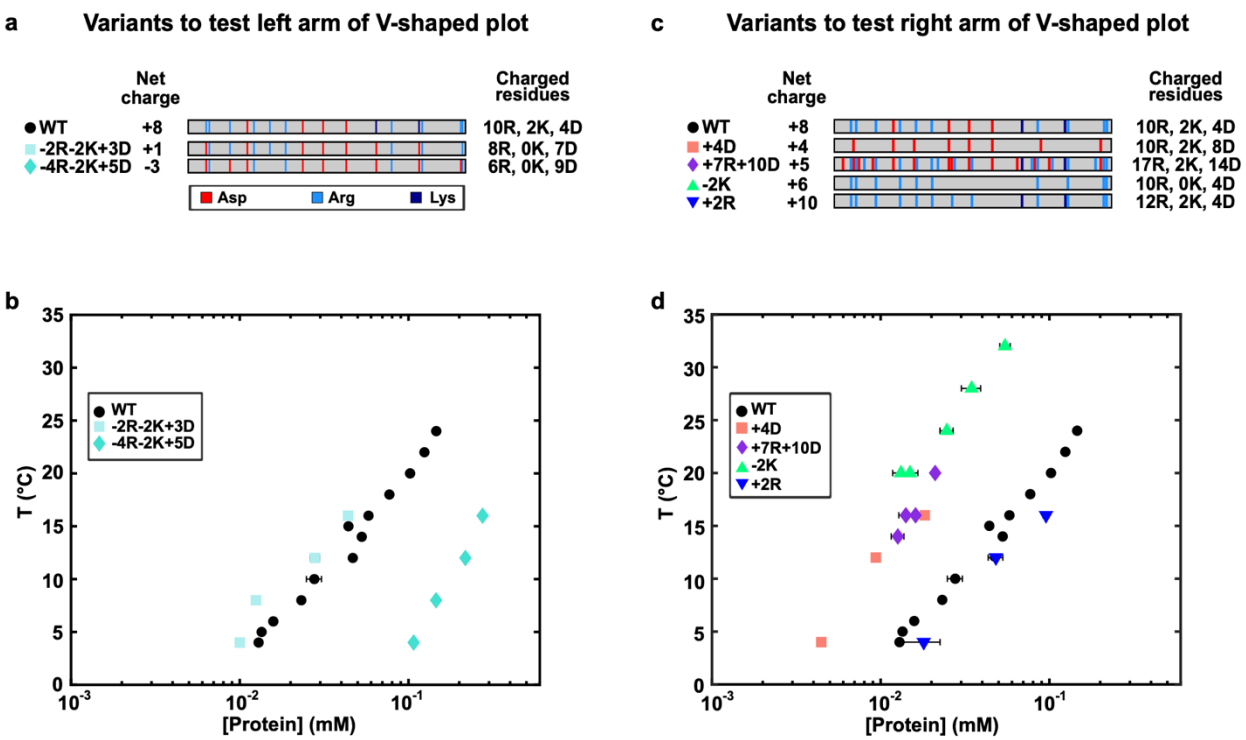
Supplementary Figure 2: Measured binodals of A1-LCD variants from Fig. 2, Fig. 3 and Fig. 4.

(a) Measured binodals of A1-LCD variants that titrate Phe and Tyr balance (Fig. 2a) as a function of temperature. **(b)** Measured binodals of A1-LCD variants that titrate Arg residues (Fig. 3c). **(c)** Measured binodals of A1-LCD variants that titrate Arg / Lys content (Fig. 4a). **(d)** Measured binodals of A1-LCD variants that titrate the content of negatively charged residues (Fig. 4c). **(e)** Measured binodals of A1-LCD variants that titrate the content of oppositely charged residues (Fig. 4e). The solution conditions for all experiments were 20 mM HEPES, 150 mM NaCl, pH 7.0.



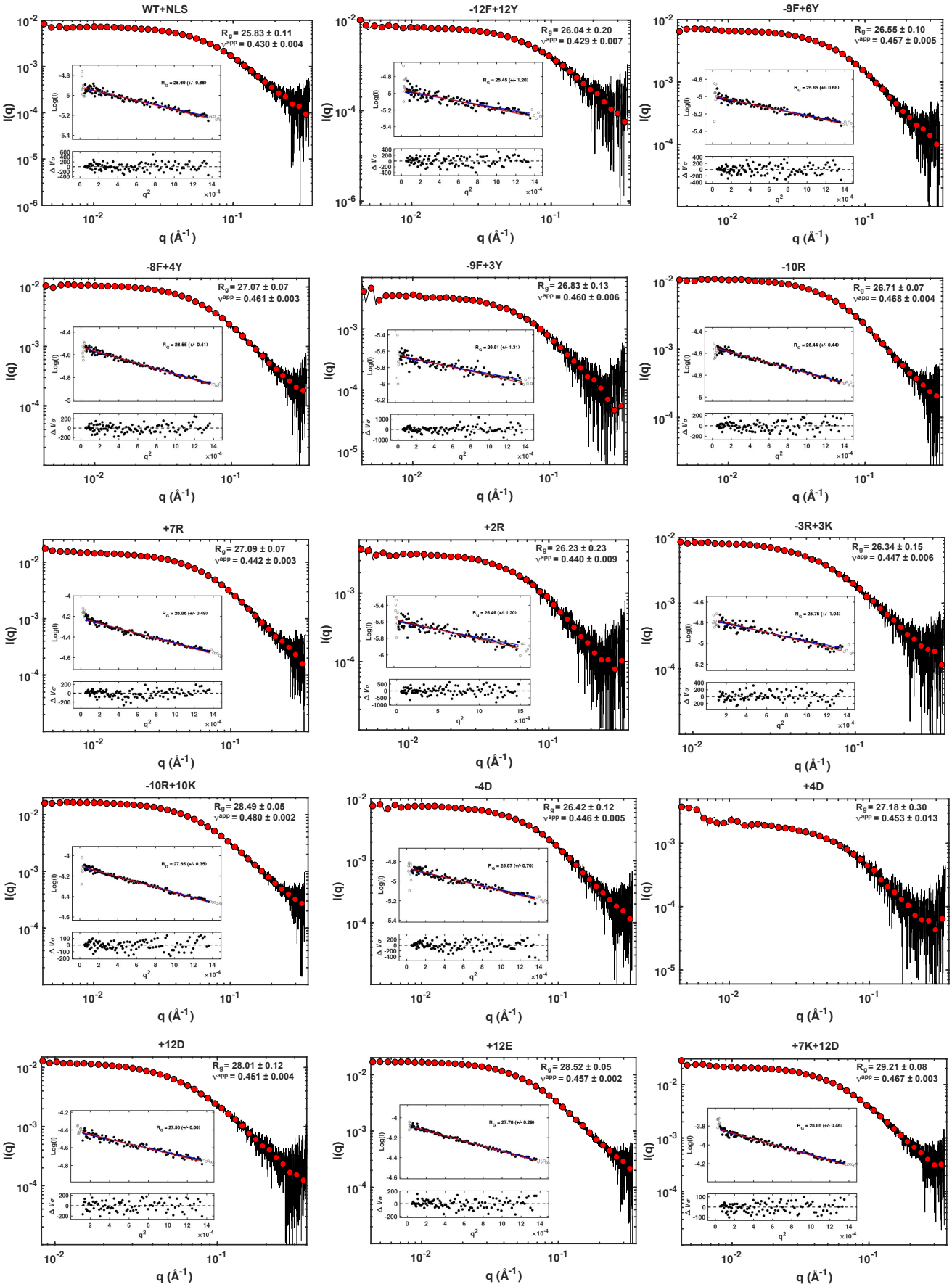
Supplementary Figure 3: Far-UV CD spectra of selected A1-LCD variants in 20 mM HEPES, pH 7.0 in the absence or presence of 150 mM NaCl.

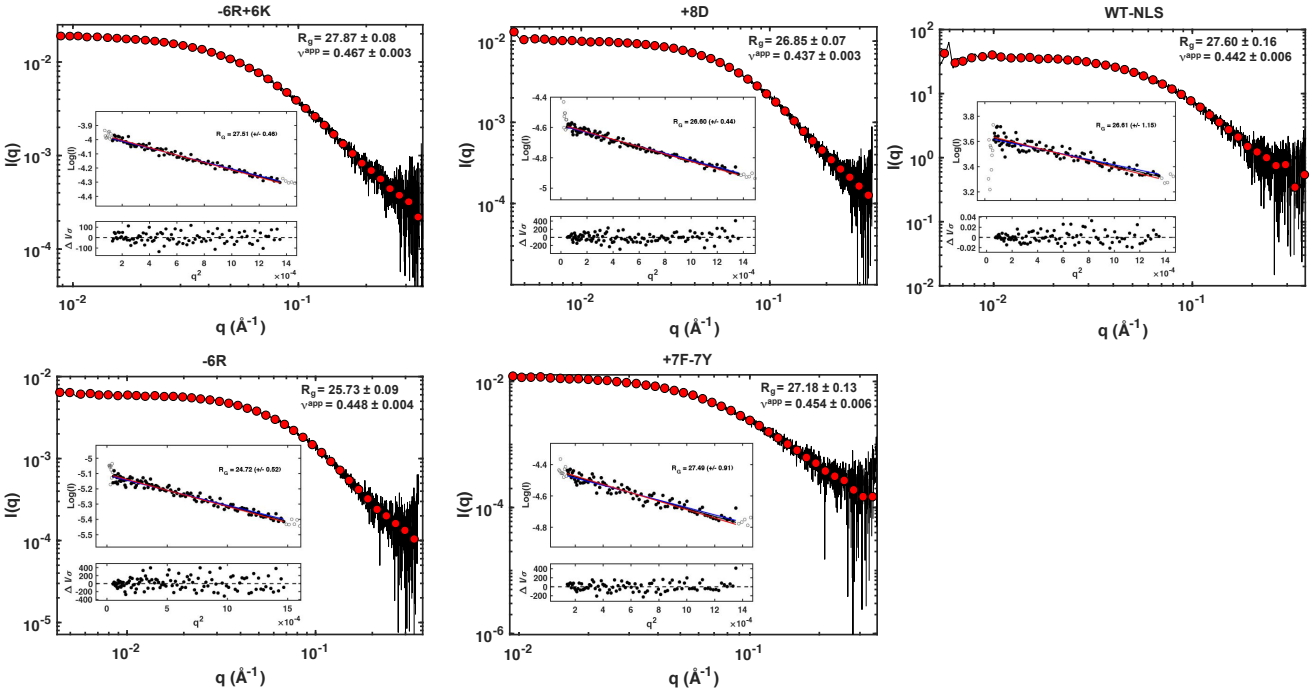
CD spectra of WT A1-LCD were collected using 0 M NaCl (dashed line) and 0.15 M NaCl (black line) solutions. All other spectra were collected using 0.15 M NaCl solutions.



Supplementary Figure 4: Measured binodals of A1-LCD variants designed to query the robustness of the master V-shaped plot in Fig. 5.

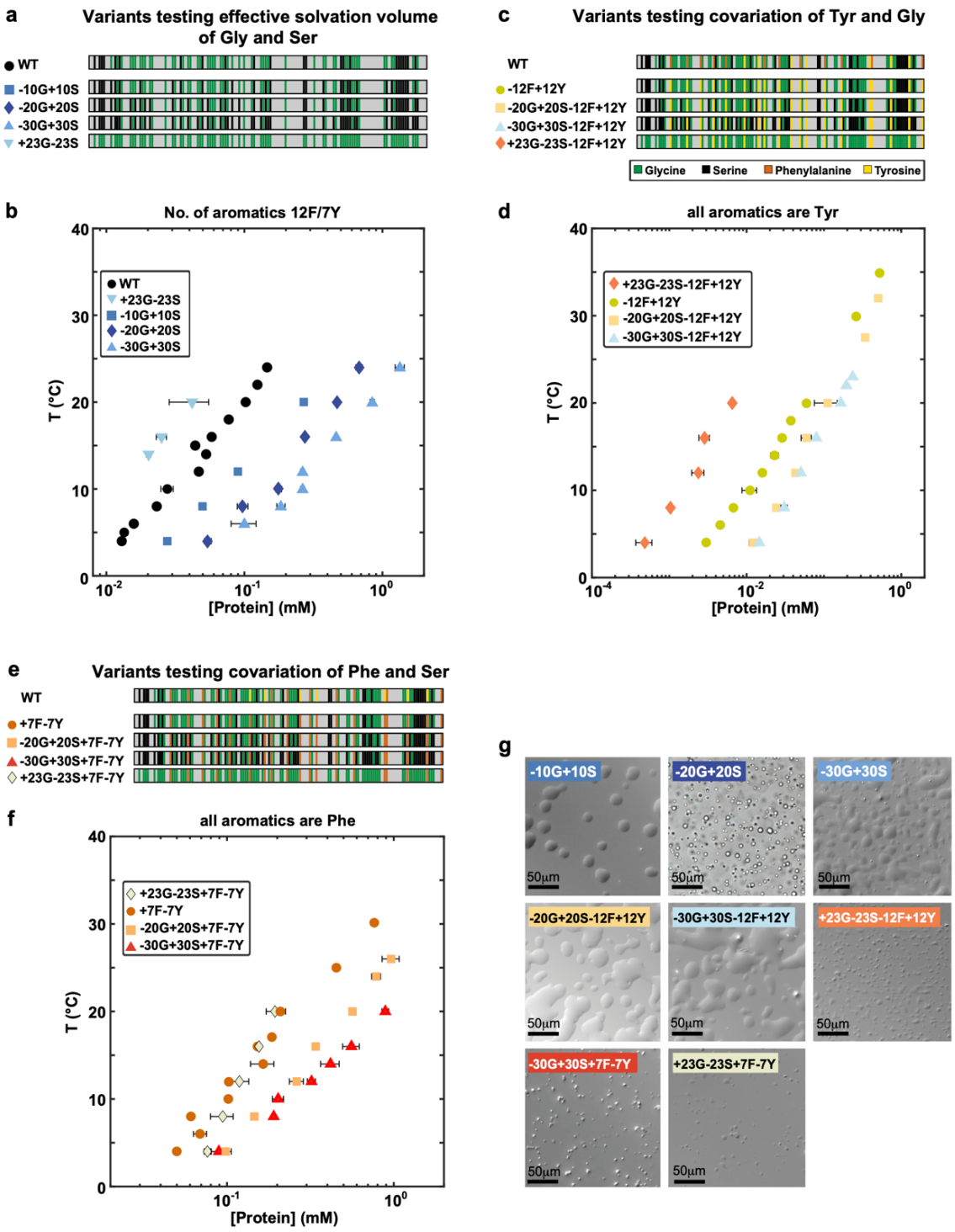
(a) Sequence designs to test if mean-field model holds true for the left arm of the V-shaped plot. **(b)** Measured saturation concentrations of variants in (a) as a function of temperature. **(c)** Sequence designs to test if mean-field model holds true for the right arm of the V-shaped plot. **(d)** Measured saturation concentrations of variants in (c) as a function of temperature. The solution conditions for all experiments were 20 mM HEPES, 150 mM NaCl, pH 7.0.


WT+NLS
R_g = 25.83 ± 0.11
ν^{app} = 0.430 ± 0.004
-12F+12Y
R_g = 26.04 ± 0.20
ν^{app} = 0.429 ± 0.007
-9F+6Y
R_g = 26.55 ± 0.10
ν^{app} = 0.457 ± 0.005
-8F+4Y
R_g = 27.07 ± 0.07
ν^{app} = 0.461 ± 0.003
-9F+3Y
R_g = 26.83 ± 0.13
ν^{app} = 0.460 ± 0.006
-10R
R_g = 26.71 ± 0.07
ν^{app} = 0.468 ± 0.004
+7R
R_g = 27.09 ± 0.07
ν^{app} = 0.442 ± 0.003
+2R
R_g = 26.23 ± 0.23
ν^{app} = 0.440 ± 0.009
-3R+3K
R_g = 26.34 ± 0.15
ν^{app} = 0.447 ± 0.006
-10R+10K
R_g = 28.49 ± 0.05
ν^{app} = 0.480 ± 0.002
-4D
R_g = 26.42 ± 0.12
ν^{app} = 0.446 ± 0.005
+4D
R_g = 27.18 ± 0.30
ν^{app} = 0.453 ± 0.013
+12D
R_g = 28.01 ± 0.12
ν^{app} = 0.451 ± 0.004
+12E
R_g = 28.52 ± 0.05
ν^{app} = 0.457 ± 0.002
+7K+12D
R_g = 29.21 ± 0.08
ν^{app} = 0.467 ± 0.003
I(q)
q (Å^{-1})
Log(I)
q^2
$\Delta V\sigma$



Supplementary Figure 5: Raw SAXS data of all A1-LCD variants.

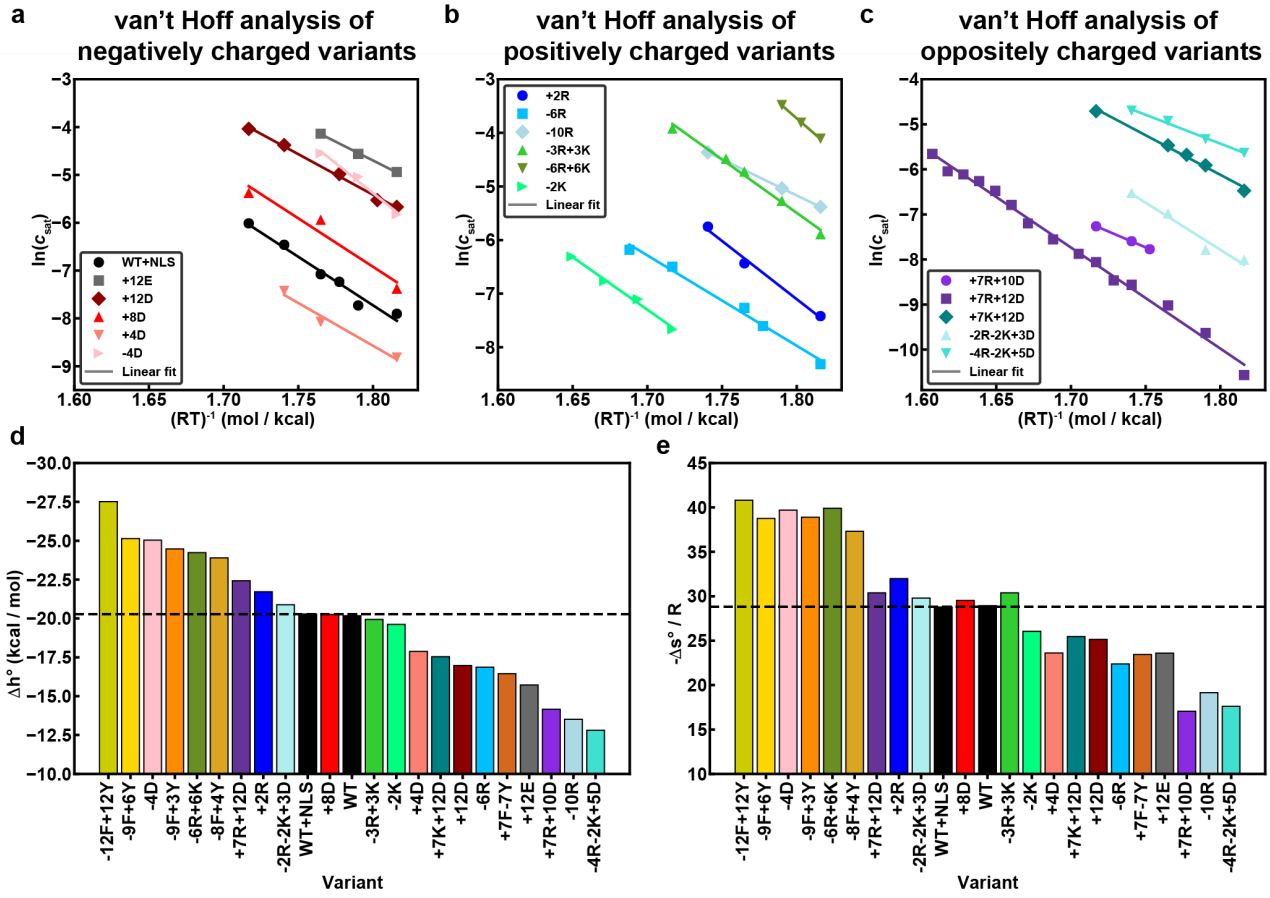
SAXS data for all A1-LCD variants that were analyzed in this manner presented as I(q) versus q normalized by the forward scattering. The raw data (black) is overlaid with logarithmically smoothed data for visualization (red circles). The results from the fit to the empirical MFF [5] are indicated in the upper right corner. Results are summarized in Table S2. The inset is the Guinier fit with the resulting R_g. Deviations from the linearity in the Guinier region prevented a Guinier fit for variant +4D; a fit to the MFF was possible, nonetheless.



Supplementary Figure 6: Examining the effects of Gly / Ser composition and their covariations with Tyr / Phe.

(a) Diagram of variants to understand the contributions of Gly and Ser to effective solvation volumes of A1-LCD. Vertical bars in the schematics indicate the position of residue types, namely Gly (green) and Ser (black). **(b)** Measured saturation concentrations of A1-LCD variants from (a) as a function of temperature. **(c)** Diagram of variants to understand the contributions of Gly and Ser to effective solvation volumes when all aromatics are Tyr. Vertical bars in the schematics indicate the position of residue types namely, Gly (green), Ser (black), Phe (brown), and Tyr (yellow). **(d)** Measured saturation concentrations of A1-LCD variants from (c) as a function of temperature. **(e)** Diagram of variants to

understand the contributions of Gly and Ser to effective solvation volumes when all aromatics are Phe. **(f)** Measured saturation concentrations of A1-LCD variants from (e) as a function of temperature. **(g)** DIC images showing dense liquid droplets.



Supplementary Figure 7: van't Hoff analysis for A1-LCD variants used in this study.

Panels **(a-c)** show plots of ln(c_{sat}) vs. $(RT)^{-1}$ for **(a)** Asp/Glu variants, **(b)** Arg/Lys variants, and **(c)** mixed charge variants. Panels **(d)** and **(e)** show variant-specific estimates for Δh° and $-\Delta s^{\circ}/R$ extracted from panels (a) – (c) using the van't Hoff analysis. Here, R = 1.98717×10^{-3} kcal/mol*K. The +7R variant is omitted due to the proximity of the measurements to the critical point restricting one from performing a van't Hoff analysis. Among the aromatic variants, Δh° increases as the Tyr:Phe ratio increases, indicating Tyr is a stronger sticker than Phe. In addition, among positively charged variants, Δh° increases as the number of Arg residues increases, which is attributable to the role of Arg as an auxiliary sticker.

References

1. Martin EW, Holehouse AS, Peran I, Farag M, Incicco JJ, Bremer A*, et al.* Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* 2020, **367**(6478)**:** 694-699.

2. Martin EW, Hopkins JB, Mittag T. Small angle x-ray scattering experiments of monodisperse samples close to the solubility limit. *arXiv* 2020.

3. Kirby N, Cowieson N, Hawley AM, Mudie ST, McGillivray DJ, Kusel M*, et al.* Improved radiation dose efficiency in solution SAXS using a sheath flow sample environment. *Acta Crystallogr D Struct Biol* 2016, **72**(Pt 12)**:** 1254-1266.

4. Hopkins JB, Gillilan RE, Skou S. BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. *J Appl Crystallogr* 2017, **50**(Pt 5)**:** 1545-1553.

5. Riback JA, Bowman MA, Zmyslowski AM, Knoverek CR, Jumper JM, Hinshaw JR*, et al.* Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* 2017, **358**(6360)**:** 238-241.

6. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. Nmrpipe - a Multidimensional Spectral Processing System Based on Unix Pipes. *J Biomol Nmr* 1995, **6**(3)**:** 277-293.

7. Lee W, Tonelli M, Markley JL. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* 2015, **31**(8)**:** 1325-1327.

8. Johnson BA. From Raw Data to Protein Backbone Chemical Shifts Using NMRFx Processing and NMRViewJ Analysis. *Methods Mol Biol* 2018, **1688:** 257-310.

9. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H*, et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* 2018, **47**(D1)**:** D309-D314.

10. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W*, et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 2011, **7**(1)**:** 539.

11. Consortium TU. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* 2018, **47**(D1)**:** D506-D515.

12. Holehouse AS, Das RK, Ahad JN, Richardson MOG, Pappu RV. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophysical Journal* 2017, **112**(1)**:** 16-21.

13. Wang J, Choi JM, Holehouse AS, Lee HO, Zhang X, Jahnel M*, et al.* A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* 2018, **174**(3)**:** 688-699 e616.

14. Semenov AN, Rubinstein M. Thermoreversible Gelation in Solutions of Associative Polymers. 1. Statics. *Macromolecules* 1998, **31**(4)**:** 1373-1385.

15. Choi J-M, Hyman AA, Pappu RV. Generalized models for bond percolation transitions of associative polymers. *Physical Review E* 2020, **102:** 042403.

16. Mahadevi AS, Sastry GN. Cooperativity in Noncovalent Interactions. *Chemical Reviews* 2016, **116**(5)**:** 2775-2825.

17. Choi J-M, Dar F, Pappu RV. LASSI: A lattice model for simulating phase transitions of multivalent proteins. *PLoS computational biology* 2019, **15**(10).

18. Meng W, Lyle N, Luan B, Raleigh DP, Pappu RV. Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proceedings of the National Academy of Sciences* 2013, **110**(6)**:** 2123.

19. Rasmussen CE, Williams CKI. *Gaussian Processes for Machine Learning*. MIT Press: Cambridge, MA, 2006.

20. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O*, et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011, **12**(null)**:** 2825–2830.

21. Ruff KM, Harmon TS, Pappu RV. CAMELOT: A machine learning approach for coarse-grained simulations of aggregation of block-copolymeric protein sequences. *The Journal of Chemical Physics* 2015, **143**(24)**:** 243123.

22. Wei MT, Elbaum-Garfinkle S, Holehouse AS, Chen CC, Feric M, Arnold CB*, et al.* Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. *Nature Chemistry* 2017, **9**(11)**:** 1118-1125.