Supporting Information
Sequence-dependent Conformational Transitions of Disordered Proteins During Condensation

Jiahui Wang,[1] Dinesh Sundaravadivelu Devarajan,[1] Keerthivasan Muthukumar, [1]

Young C. Kim,[2,*] Arash Nikoubashman,[3,4,5,*] and Jeetain Mittal[1,6,7,*]

[1]Artie McFerrin Department of Chemical Engineering, Texas A&M University, College Station, TX 77843, United States

[2]Center for Materials Physics and Technology, Naval Research Laboratory, Washington, DC 20375, United States

[3]Leibniz-Institut für Polymerforschung Dresden e.V., Hohe Straße 6, 01069 Dresden, Germany

[4]Institut für Theoretische Physik, Technische Universität Dresden, 01069 Dresden, Germany

[5]Cluster of Excellence Physics of Life, Technische Universität Dresden, 01062 Dresden, Germany

[6]Department of Chemistry, Texas A&M University, College Station, TX 77843, United States

[7]Interdisciplinary Graduate Program in Genetics and Genomics, Texas A&M University, College Station, TX 77843, United States

[*]Corresponding author email:
youngchan.kim@nrl.navy.mil
anikouba@ipfdd.de
jeetain@tamu.edu

Simulation details

We have used a coarse-grained model for the intrinsically disordered proteins (IDPs), in which each monomer (residue) is represented by a single bead. Beads are bonded *via* a harmonic spring:

$$U_{\mathrm{b}}(r) = \frac{k_{\mathrm{b}}}{2}(r - r_0)^2, \qquad (1)$$

where r is the distance between two bonded beads, k_{b}= 20 kcal/(mol Å^2) is the spring constant, and r_0 = 3.8 Å is the equilibrium bond length. The van der Waals interactions between nonbonded beads were modeled using a modified Lennard-Jones (LJ) potential[1]

$$U_{\mathrm{vdW}}(r) = \begin{cases} U_{\mathrm{LJ}}(r) + (1-\lambda)\varepsilon, & r \le 2^{1/6}\sigma \\ \lambda U_{\mathrm{LJ}}(r), & otherwise' \end{cases} \qquad (2)$$

where U_{LJ} is the standard LJ potential

$$U_{\mathrm{LJ}}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right], \qquad (3)$$

with average hydropathy $\lambda = (\lambda_i + \lambda_j)/2$, average diameter $\sigma = (\sigma_i + \sigma_j)/2$ for residues *i* and *j*. The hydropathy values for E and K were $\lambda_{\mathrm{E}} = 0.460$ and $\lambda_{\mathrm{K}} = 0.514$, and diameters of the E and K residues were $\sigma_{\mathrm{E}} = 5.92$Å and $\sigma_{\mathrm{K}} = 6.36$Å, respectively[2,3]. For natural proteins, λ_i values were set according to the HPS-Urry model[4]. The interaction strength was fixed to $\varepsilon = 0.2\,\mathrm{kcal/mol}$ in all simulations. For computational efficiency, the pair potential U_{vdW} and its associated forces were truncated to zero at a distance of 4 σ. The electrostatic interactions between nonbonded residues were modeled using a Coulombic potential with Debye–Hückel electrostatic screening[5]

$$U_{\mathrm{e}}(r) = \frac{q_i q_j}{4\pi D \epsilon_0 r} e^{-r/\ell}, \quad (4)$$

where D = 80 was the dielectric constant of the medium, ϵ_0 was the permittivity of vacuum, and ℓ=10 Å was the Debye screening length. The electrostatic potential and its forces were truncated to zero at a distance of 35 Å.

Single chain simulations were performed by placing one chain into a cubic box of edge length 160 Å with periodic boundary conditions in all directions (the simulation box was large enough to prevent unphysical self-interactions). Langevin dynamics simulations were performed at constant temperature (T = 300 K for EKVs, FUS LC, TDP-43 and hnRNPA2, T = 260 K for LAF-1 RGG). The friction coefficient was set as $\gamma_i = m_i/t_{damp}$, where m_i is the mass of residues and t_{damp} = 1000 fs. Simulations were performed for 1.0 μs with a time step of 10 fs using HOOMD-blue[6] (ver. 2.9.3) with features extended using azplugins[7] (ver. 0.10.2). The simulation trajectories were saved every 1000 fs.

Dense phase simulations were performed in the *NPT* ensemble with an external isotropic pressure of P = 0 atm. We permitted the systems to equilibrate at their preferred concentrations for 0.5 μs. After the chains achieved their preferred dense-phase concentration, the simulations were performed using Langevin dynamics at constant volume for 1.0 μs. All the simulations were done using t_{damp} = 1000 ps and at fixed temperatures. Other simulation settings were also the same as the single-chain simulations.

For concentration scan simulations, we first compact a linear IDP chain by using the LJ potential, then duplicate for the monomer-monomer interactions. Then, we duplicate and arrange

the compacted chains into a droplet-like structure (fixed number of chains, *N*=500). From this point on, we replace the LJ potential with the appropriate non-bonded interactions, as provided above. For concentrations ≤ 40 mg/ml, the size of the cubic simulation box is adjusted instantaneously to achieve the desired concentrations, and then simulated for 1 ns. The configuration after this 1 ns run serves as the initial configuration for our production runs. For concentrations > 40 mg/ml, the initial droplet-like structure is first placed in a cubic box at a concentration of 40 mg/ml and simulated for 1 ns. The simulation box is then linearly compressed in 1 ps to reach the desired concentration, serving as the initial configuration for our production runs. Then use Langevin dynamics (t_{damp} = 1000 ps) at specific constant volumes for 2 μs. For concentrations where more than one big droplet formed, the simulation time was extended until the number of clusters remained constant for at least 1 μs. Other simulation settings were the same as those used in the single-chain simulations. Other simulation settings were the same as those used in the single-chain simulations.

Droplet initial configurations were obtained from equilibrated dense-phase droplets formed at high concentrations, where all chains are within the droplet. The droplet configuration was then placed in simulation boxes of varying sizes to achieve different concentrations. The other simulation settings were the same as those used in the concentration scan simulations.

Slab simulations were conducted by first preparing the dense phase in a cubic box of edge length 150 Å with periodic boundary conditions in all directions for 100 ns. Then, the z-dimension of the box was extended to 1200 Å, and simulations were performed for 3 μs using Langevin

dynamics (t_{damp} = 1000 ps). Other simulation settings were the same as the single chain simulations.

The chosen γ values aim to establish a weak coupling with the thermostat. This approach is designed primarily for temperature regulation rather than to simulate the hydrodynamic forces acting on the residues from the solvent[8]. Different γ values are used for single- and multi-chain simulations to ensure effective temperature control. Compared to multi-chain simulations, single-chain simulations use higher γ values due to the reduced number of beads. We verified that the velocity distribution of the particles follows the expected Maxwell-Boltzmann distribution in both the single- and multi-chain simulations.

Martini simulation details: We initially distributed 80 chains within a slab-shaped simulation box, placing them at random positions near the center of the box, with each chain in the same extended conformation. Subsequently, we solvated the system with roughly 50,000 water molecules and introduced ions to establish a concentration of 100 mM. The system's energy was minimized using the steepest descent algorithm over a duration of 0.3 ns, employing a timestep of 30 fs. Next, we equilibrated the system through *NPT* simulations, maintaining a constant temperature (*T*) of 300 K and a constant pressure (*P*) of 1 bar for 20 ns with a finer timestep of 20 fs. To accurately maintain the specified temperature and pressure conditions, we utilized a velocity-rescaling thermostat with a time constant of 1 ps and a Parrinello–Rahman barostat with a time constant of 12 ps, respectively. Following the equilibration, a production run was executed for an extended period of 18 μs. For the single-chain simulations, we conducted 5 independent replicas using

approximately 30,000 water molecules and a 100 mM salt concentration. These simulations were performed within a cubic box with an edge length of 150 Å for a duration of 1 μs each.

For all analyses presented in this study, the error bars indicate the standard error of the mean calculated by dividing analyzed trajectories into five blocks.

Normalized sequence charge decoration parameter (nSCD):

$$\mathrm{SCD} = \frac{1}{N}\sum_{i=2}^{N}\sum_{j=1}^{i-1} q_i q_j (i-j)^{1/2}, \quad (5)$$

$$\mathrm{nSCD} = \frac{\mathrm{SCD}-\mathrm{SCD}_{\mathrm{max}}}{\mathrm{SCD}_{\mathrm{min}}-\mathrm{SCD}_{\mathrm{max}}}, \quad (6)$$

where q_i and q_j are the charges of residues *i* and *j*, respectively. $\mathrm{SCD}_{\mathrm{max}}$ is the SCD value for the perfectly alternating sequence and $\mathrm{SCD}_{\mathrm{min}}$ is for charge segregated sequence.

Natural protein sequences used in this work

FUS LC (length = 163)

MASNDYTQQATQSYGAYPTQPGQGYSQQSSQPYGQQSYSGYSQSTDTSGYGQSSYSSY
GQSQNTGYGTQSTPQGYGSTGGYGSSQSSQSSYGQQSSYPGYGQQPAPSSTSGSYGSSS
QSSSYGQPQSGSYSQQPSYGGQQQSYGQQQSYNPPQGYGQQNQYNS

TDP-43 (length = 141)

GRFGGNPGGFGNQGGFGNSRGGGAGLGNNQGSNMGGGMNFGAFSINPAMMAAAQAA
LQSSWGMMGMLASQQNQSGPSGNNQNQGNMQREPNQAFGSGNNSYSGSNSGAAIGW
GSASNAGSGSGFNGGFGSSMDSKSSGWGM

hnRNPA2 (length = 152)

GRGGNFGFGDSRGGGGNFGPGPGSNFRGGSDGYGSGRGFGDGYNGYGGGPGGGNFGG
SPGYGGGRGGYGGGGPGYGNQGGGYGGGYDNYGGGNYGSGNYNDFGNYNQQPSNY
GPMKSGNFGGSRNMGGPYGGGNYGPGGSGGSGGYGGRSRY

LAF-1 RGG (length = 168)

MESNQSNNGGSGNAALNRGGRYVPPHLRGGDGGAAAAASAGGDDRRGGAGGGGYRR
GGGNSGGGGGGGYDRGYNDNRDDRDNRGGSGGYGRDRNYEDRGYNGGGGGGGNRG

YNNNRGGGGGGYNRQDRGDGGSSNFSRGGYNNRDEGSDNRGSGRSYNNDRRDNGGDG

Cluster analysis

We defined two chains as part of the same cluster if the distance between their monomers was less than 1.5σ (equivalent to 9.54 Å). Subsequently, we calculated the probability $P(N_C)$ of a chain being part of a cluster of size N_C and the number of clusters formed in the system.

Supplementary Figures

Fig. S1 Selected E-K variants (EKVs) for chain length N=50 with their identifying number and normalized SCD parameter.

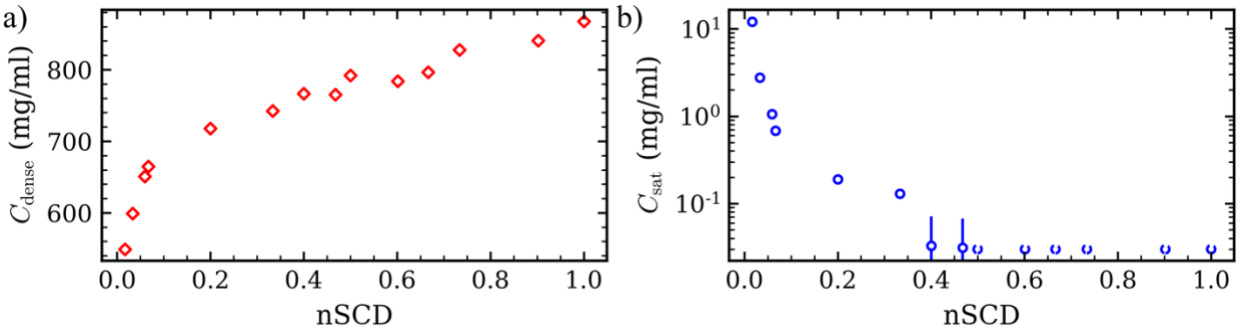

Fig. S2 (a) Dense phase concentrations for EKVs as a function of nSCD. (b) Saturation concentrations for EKVs as a function of nSCD. The open symbols represent the extrapolated concentrations for EKV10 to EKV15. The mean values are obtained by dividing the trajectory into 5 independent blocks. Error bars in all panels indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

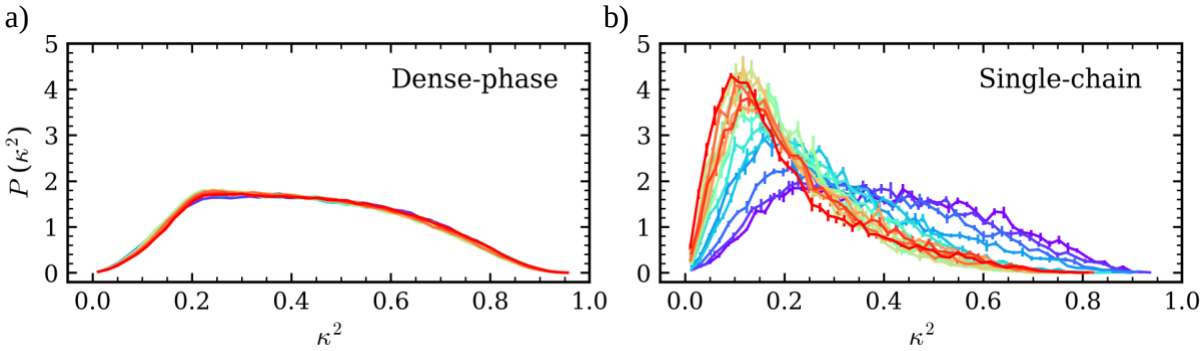

Fig. S3 Probability distribution of relative shape anisotropy κ^2 for EKVs in the (a) dense phase and as (b) a single chain. Line colors ranging from purple to red, indicate increasing nSCD. The mean values are obtained by dividing the trajectory into 5 independent blocks. Error bars in all panels indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

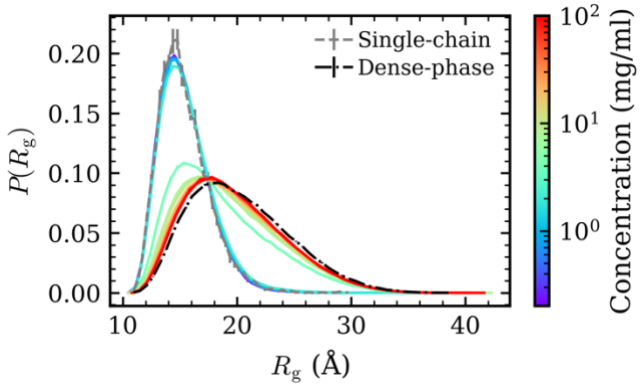

Fig. S4 Probability distribution of R_g for EKV5. Line colors ranging from purple to red, indicate increasing concentrations from 0.2 mg/ml to 100 mg/ml. The gray dashed line represents the R_g distribution for a single chain and the black dashed line represents the distribution within the dense

phase. The mean values are obtained by dividing the trajectory into 5 independent blocks. Error bars in all panels indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

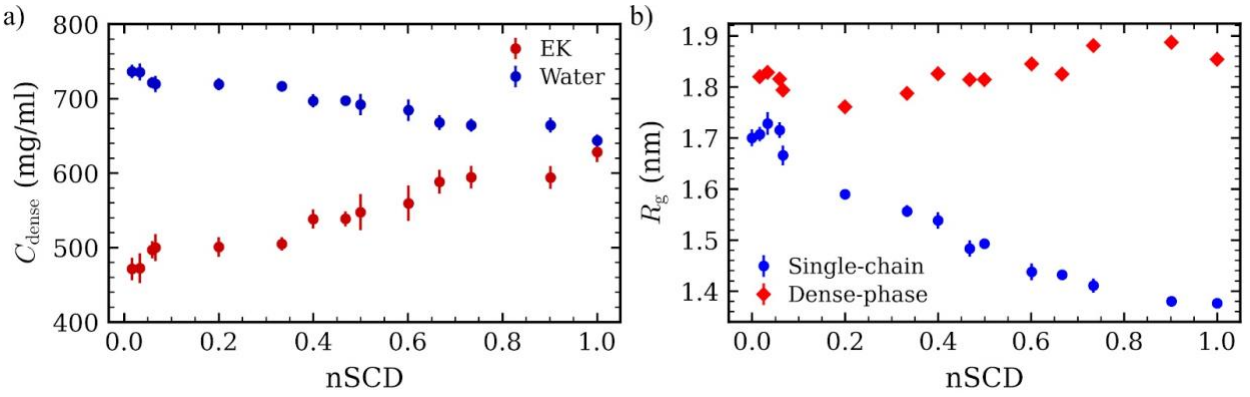

Fig. S5 Martini simulation results. (a) Dense phase concentrations of EKVs (red circles) and water (blue circles) as a function of nSCD. (b) R_g of EKVs in the dense phase (red diamonds) and as a single chain (blue circles). The mean values are obtained by dividing the trajectory into 5 independent blocks or by each replica for single-chain. Error bars in all panels indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

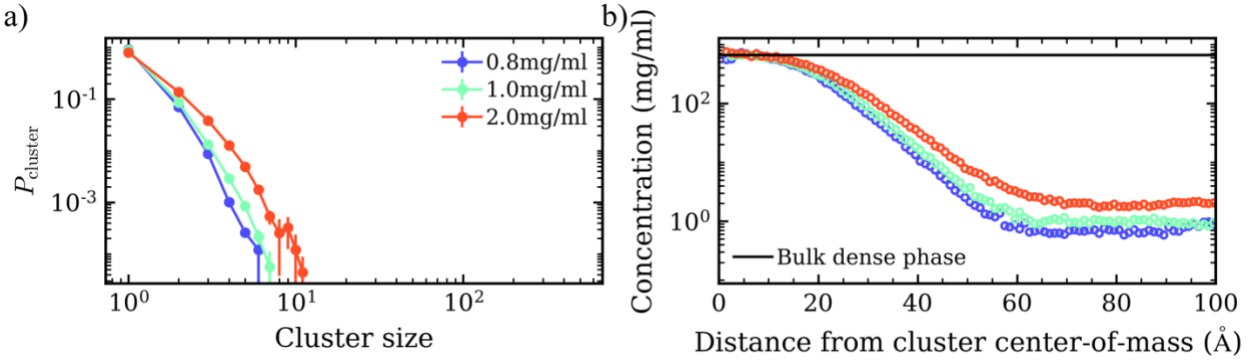

Fig. S6 (a) Probability distribution of cluster size for EKV5 at the concentrations of 0.8 mg/ml, 1.0 mg/ml, and 2.0 mg/ml. (b) Radial density profile of monomers with respect to the distance from the center-of-mass of the largest cluster. The mean values are obtained by dividing the trajectory into 5 independent blocks. Error bars in all panels indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

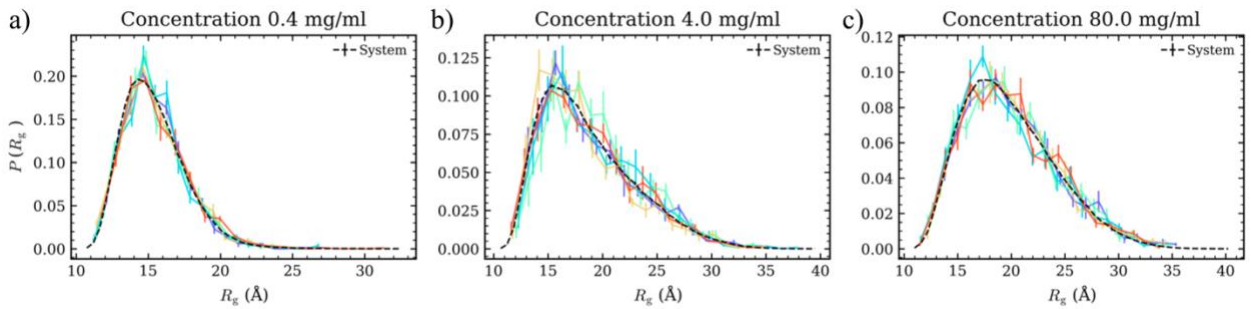

Fig. S7 (a, b, c) R_g distribution for five randomly selected chains and the distribution for the entire system (black dashed line) at concentrations of (a) 0.4 mg/ml, (b) 4.0 mg/ml, and (c) 80.0 mg/ml of EKV5. The mean values are obtained by dividing the trajectory into 5 independent blocks. Error bars in all panels indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

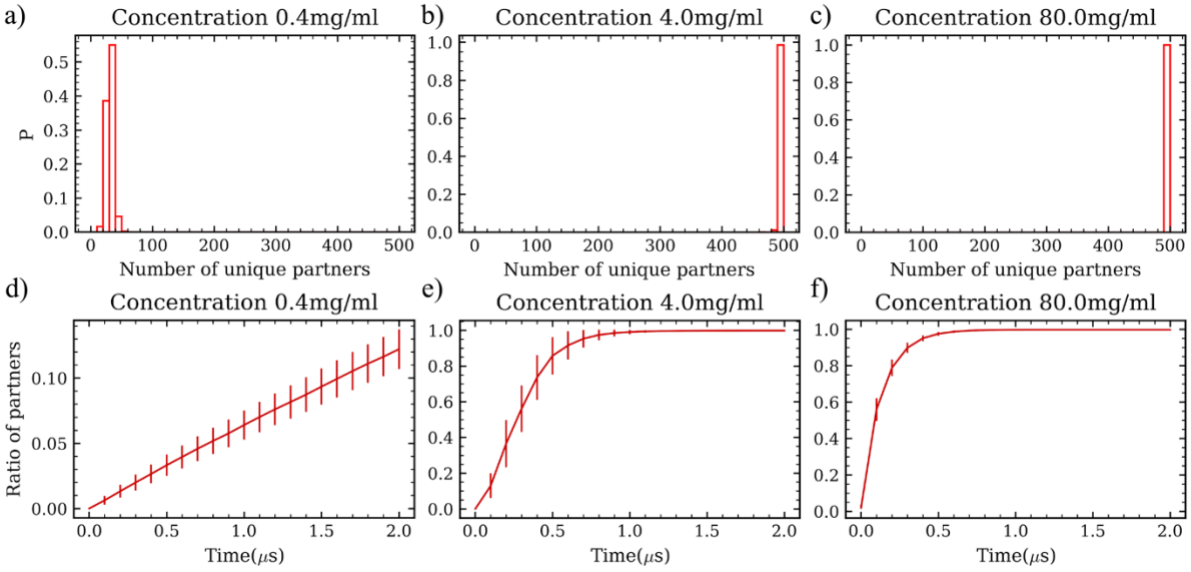

Fig. S8 (a, b, c) Distribution of each chain's number of unique partners in the last 1 μs at 0.4 mg/ml, 4.0 mg/ml, and 80.0 mg/ml of EKV5. (d, e, f) Ratio of average unique partners over time at 0.4 mg/ml, 4.0 mg/ml, and 80.0 mg/ml of EKV5 over 2 μs. The mean values are obtained by dividing the trajectory into 5 independent blocks. Error bars in panels (d,e,f) indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

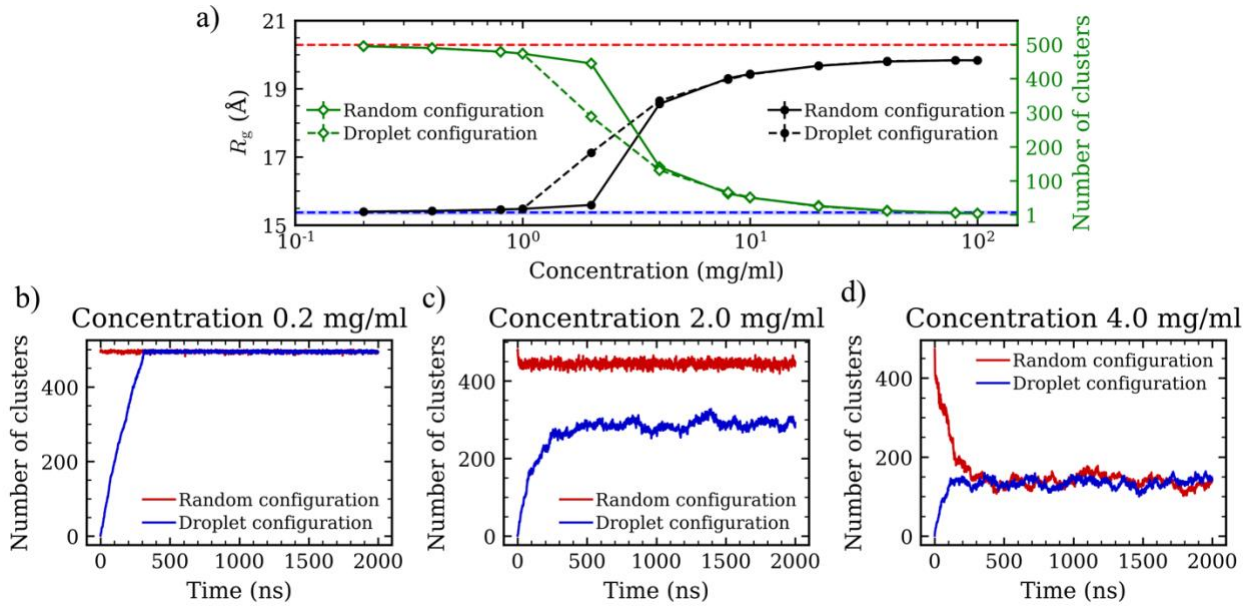

Fig. S9 (a) R_g (black, left *y*-axis) and number of clusters (green, right *y*-axis) of EKV5 as functions of concentration, comparing different initial configurations. Solid lines represent simulations starting from random initial configurations, while dashed lines represent simulations starting from droplet initial configurations. The red and blue horizontal dashed lines indicate the R_g in the bulk dense phase and of a single chain, respectively, with shaded horizontal areas indicating the corresponding error bars. Note that the error bars are smaller than the line width and may not be visible. (b–d) Time evolution of the number of clusters for EKV5 with different initial configurations at concentrations of (b) 0.2 mg/ml, (c) 2.0 mg/ml, and (d) 4.0 mg/ml. The mean values are obtained by dividing the trajectory into 5 independent blocks. Error bars in panel (a) indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

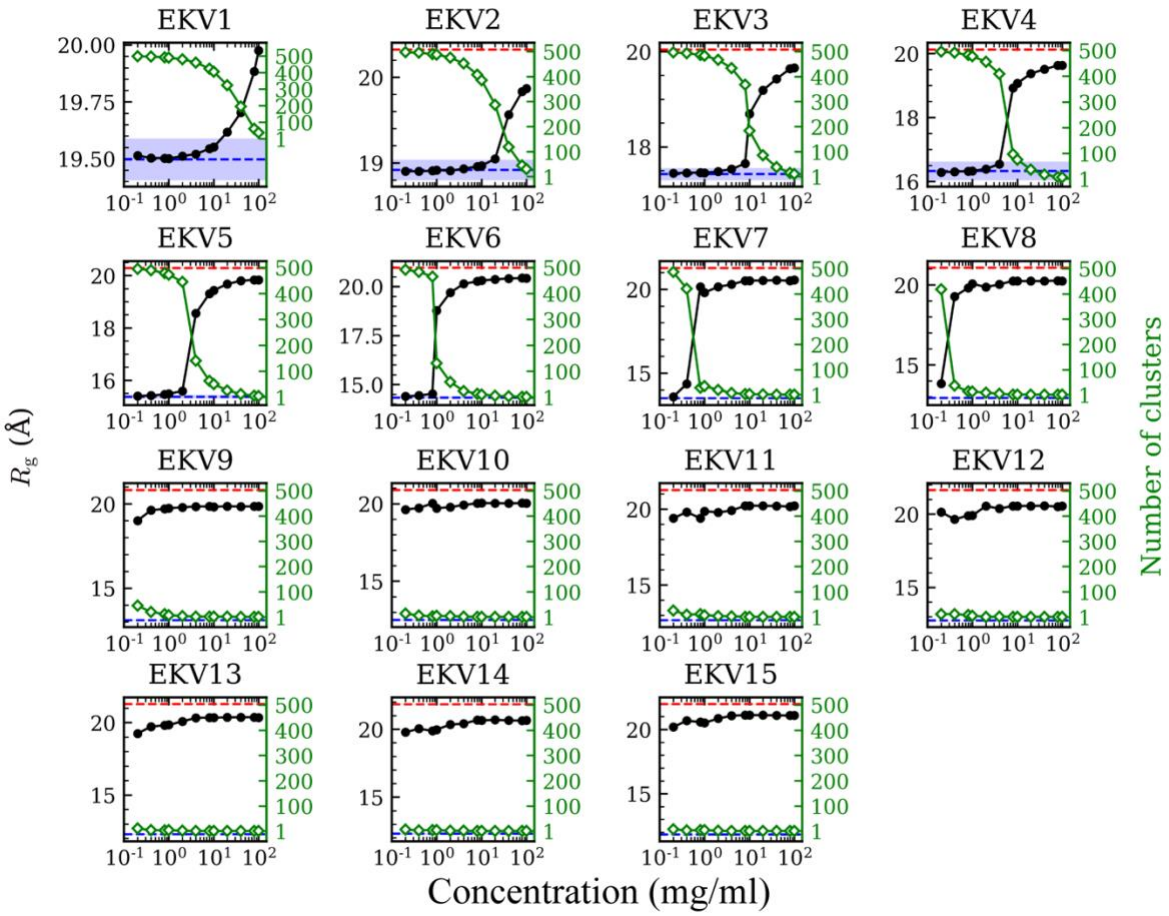

Fig. S10 R_g (black, left *y*-axis) and number of clusters (green, right *y*-axis) of EKVs as a function of concentration. The red and blue horizontal dashed lines indicate the R_g in the bulk dense phase and of a single chain, respectively, with shaded horizontal areas indicating the corresponding error bars. Note that the error bars are smaller than the line width and may not be visible. The mean values are obtained by dividing the trajectory into 5 independent blocks. Error bars in all panels indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

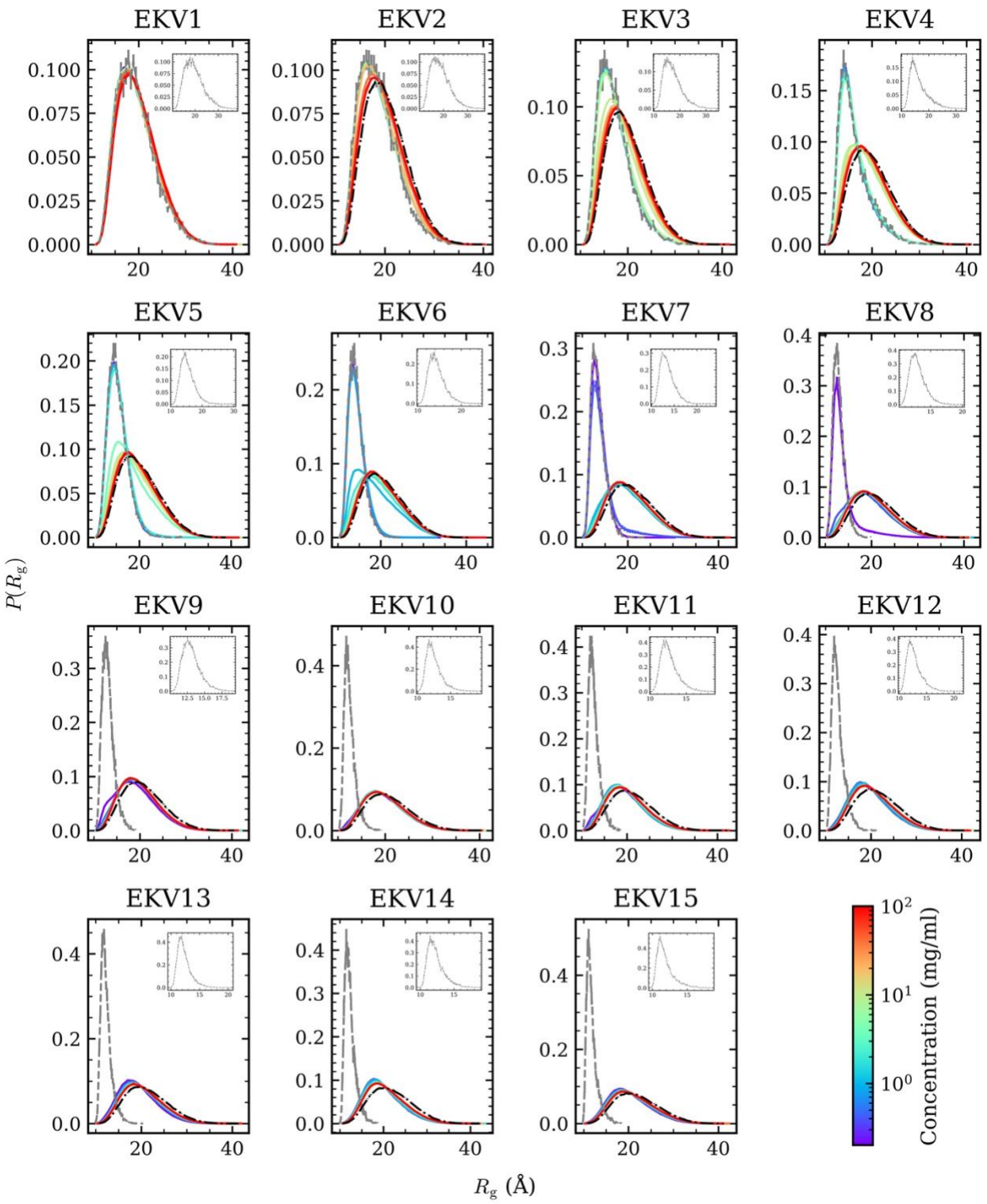

Fig. S11 Probability distribution of R_g for the EKVs. Line colors ranging from purple to red, indicate increasing concentrations. The gray dashed line represents the R_g distribution for a single chain and the black dashed line represents the distribution within the dense phase. The insets show the distributions of the single chain R_g. The mean values are obtained by dividing the trajectory into 5

independent blocks. Error bars in all panels indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

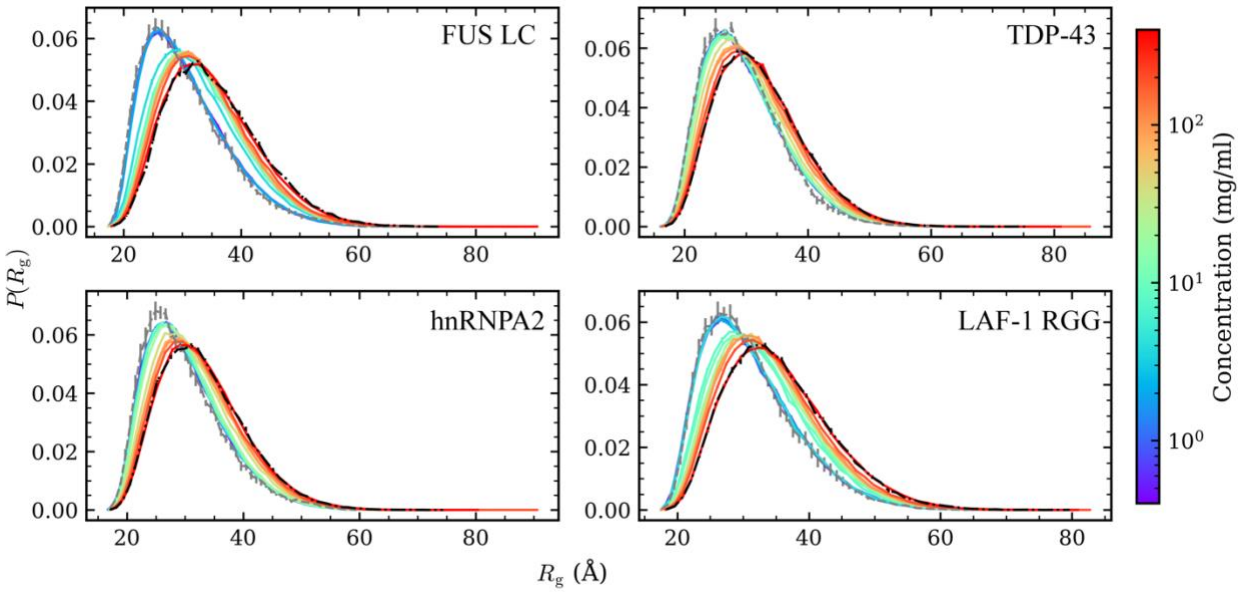

Fig. S12 Probability distribution of R_g for the disordered domains of natural proteins: FUS LC, TDP-43, hnRNPA2, and LAF-1 RGG. Line colors ranging from purple to red, indicate increasing concentrations for 0.2 mg/ml to 400 mg/ml. The gray dashed line represents the R_g distribution for a single chain and the black dashed line represents the distribution within the dense phase. The mean values are obtained by dividing the trajectory into 5 independent blocks. Error bars in all panels indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

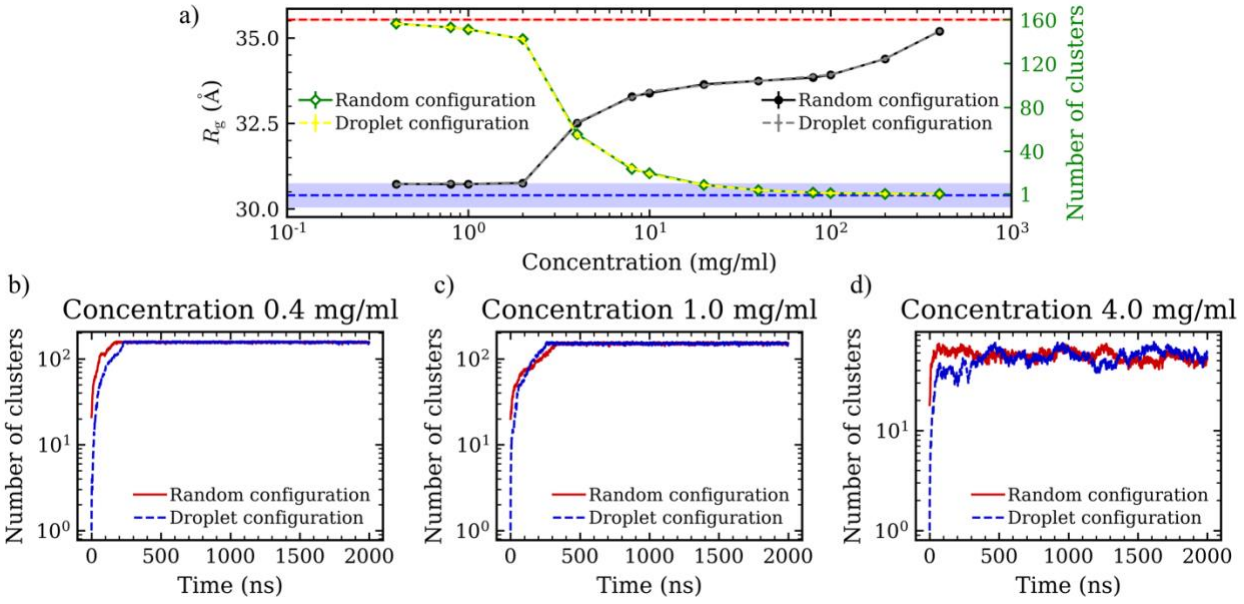

Fig. S13 (a) R_g—black lines represent random initial configurations and grey lines represent droplet initial configurations (left *y*-axis); number of clusters—green lines for random initial configurations and yellow lines for droplet initial configurations (right *y*-axis) of FUS LC as functions of concentration with different initial configurations. The red and blue horizontal dashed lines indicate the R_g in the bulk dense phase and of a single chain, respectively, with shaded horizontal areas indicating the corresponding error bars. Note that the error bars are smaller than the line width and may not be visible. (b–d) Time evolution of the number of clusters for FUS LC with different initial configurations at concentrations of (b) 0.4 mg/ml, (c) 1.0 mg/ml, and (d) 4.0 mg/ml. The mean values are obtained by dividing the trajectory into 5 independent blocks. Error bars in panel (a) indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

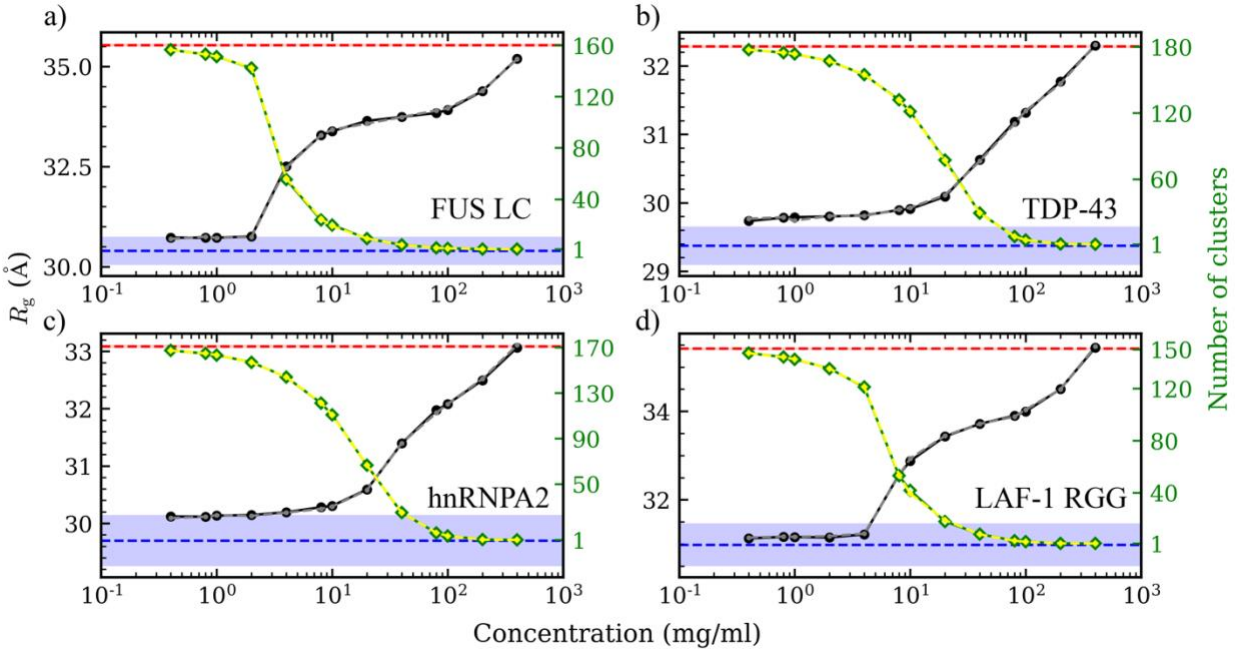

Fig. S14 Concentration scans for the disordered domains of natural proteins with different initial configurations: (a) FUS LC, (b) TDP-43, (c) hnRNPA2, and (d) LAF-1 RGG. R_g —black lines for random initial configurations and grey lines for droplet initial configurations (left *y*-axis); number of clusters—green lines for random initial configurations and yellow lines for droplet initial configurations (right *y*-axis) as functions of concentration. The red and blue horizontal dashed lines indicate the R_g in the bulk dense phase and of a single chain, respectively, with shaded horizontal areas indicating the corresponding error bars. Note that the error bars are smaller than the line width and may not be visible. The mean values are obtained by dividing the trajectory into 5 independent blocks. Error bars in all panels indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

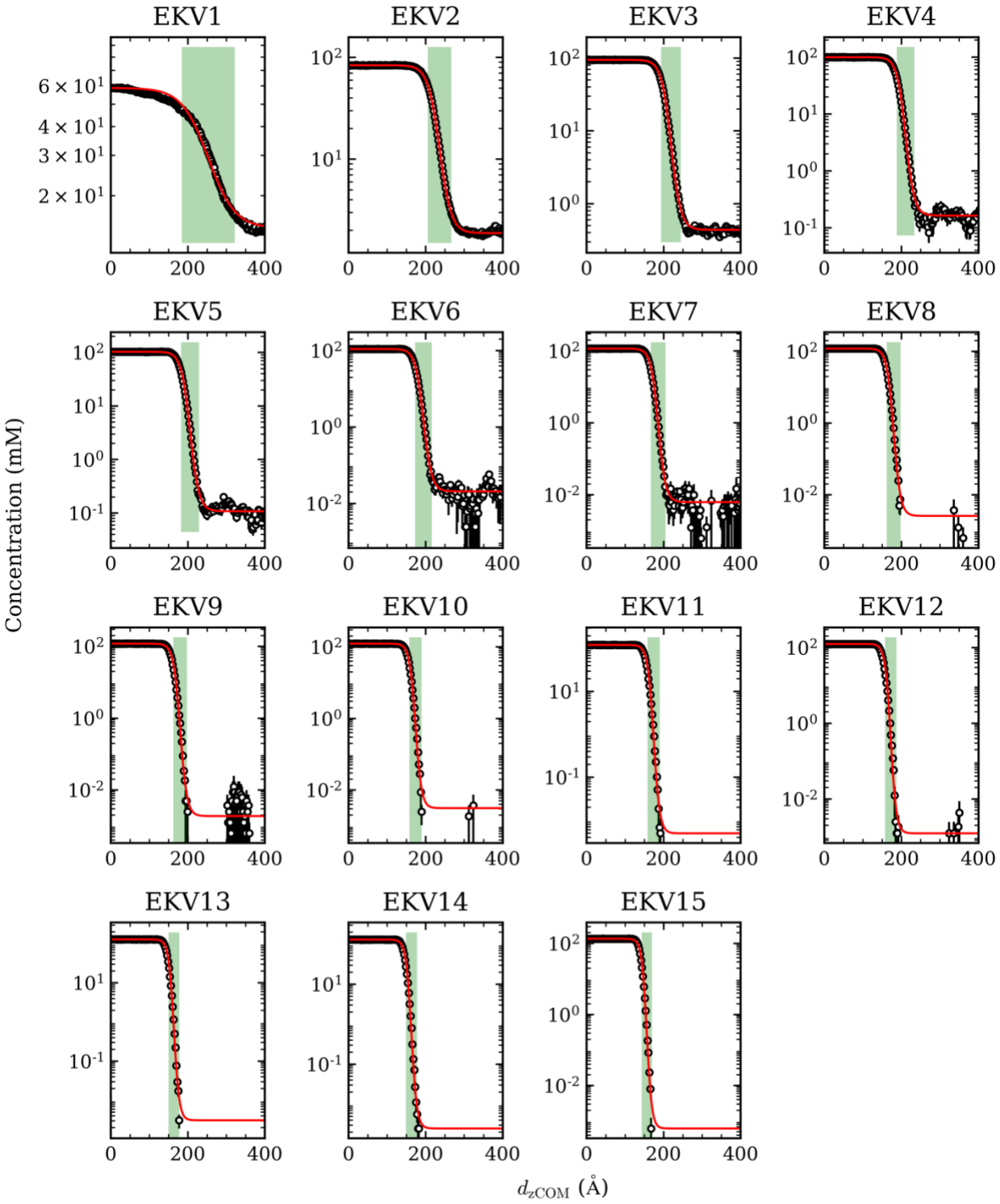

Fig. S15 Concentration profiles with respect to the distance from the condensate center-of-mass in the *z* direction. The red line is the fitted curve for the simulation data (symbols). The green shaded area indicates the interface region. The mean values are obtained by dividing the trajectory into 5 independent blocks. Error bars in all panels indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

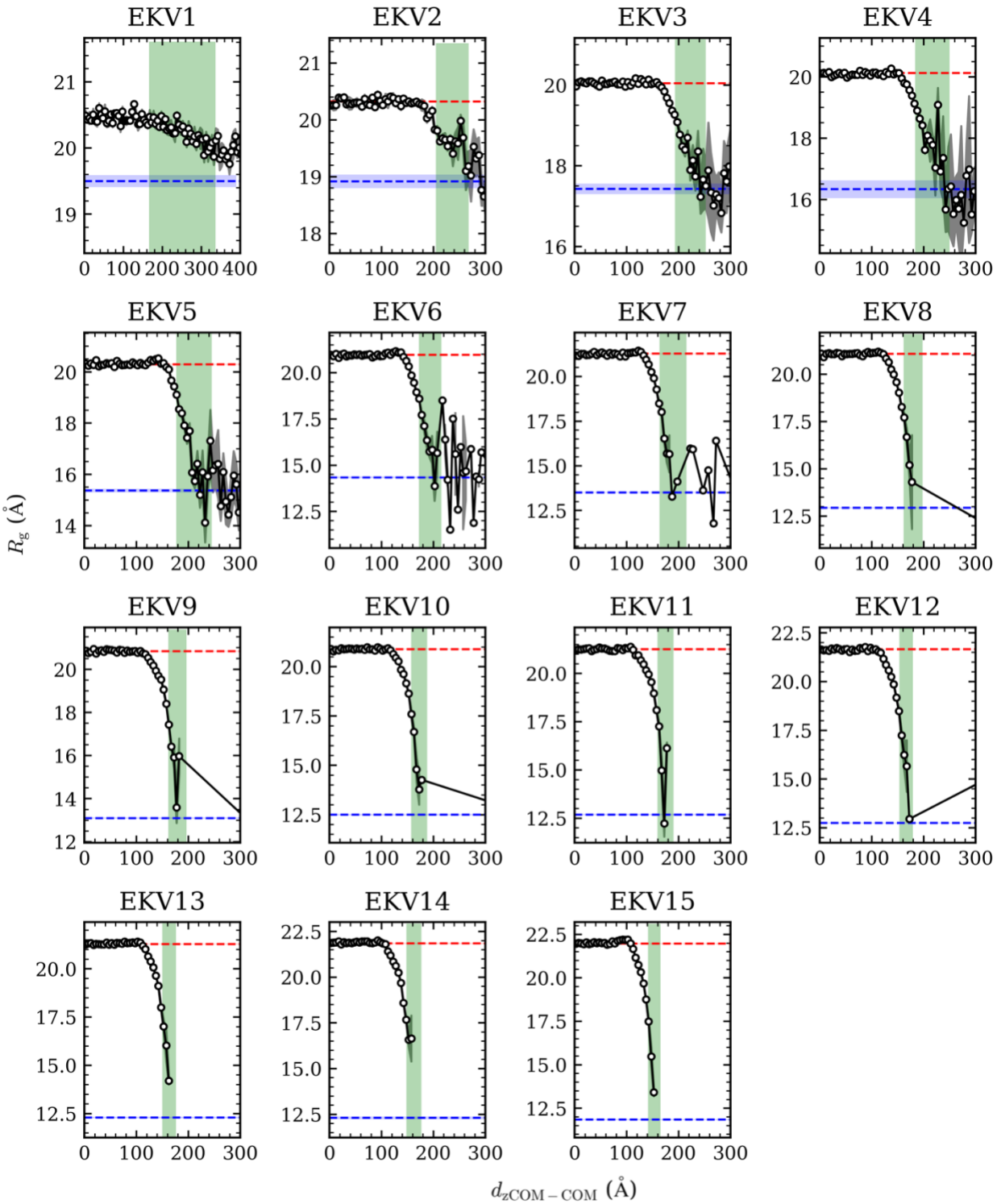

Fig. S16 Interface analysis for the EKVs *via* slab simulations. Average R_g with respect to distance from the condensate center-of-mass to chain's center-of-mass in the *z* direction, $d_{zCOM-COM}$. The red

and blue horizontal dashed lines indicate the R_g in the bulk dense phase and of a single chain, respectively, with shaded horizontal areas indicating the corresponding error bars. Note that the error bars are smaller than the line width and may not be visible. The gray-shaded polygon represents the standard error of the mean which was calculated based on the frames where chains are present at specific positions. In some cases, error bars are smaller than the symbol size.

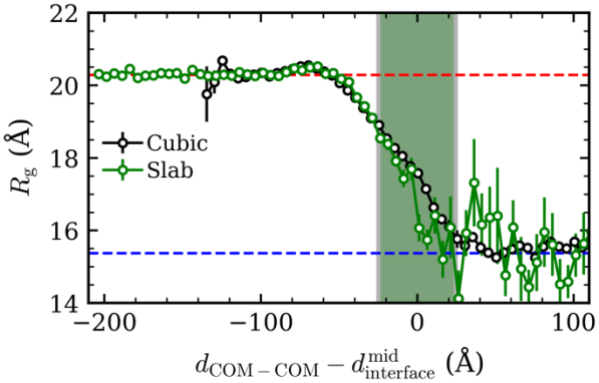

Fig. S17 (a) Average R_g for EKV5 in relation to the shifted distance. For droplet simulations (4 mg/ml, depicted in black), the distance is taken from the spherical droplet's center of mass to the chain's center of mass. For slab simulations (depicted in green), the distance is taken from the condensate's center of mass to the chain's center of mass in the z-direction. The distance was shifted by subtracting the midpoint of the interface. The shaded areas delineate the interfacial regions for the droplet (in gray) and slab (in green) simulations, respectively. The red and blue horizontal dashed lines indicate the R_g in the bulk dense phase and of a single chain, respectively, with shaded horizontal areas indicating the corresponding error bars. Note that the error bars are smaller than the line width and may not be visible. (b) Relative shape anisotropy (κ^2) for EKV5 droplet forming at a concentration of 4 mg/ml. The inset depicts a snapshot of the system at 1.2 μs. The standard error of the mean was calculated based on the frames where chains are present at specific positions. In some cases, error bars are smaller than the symbol size.

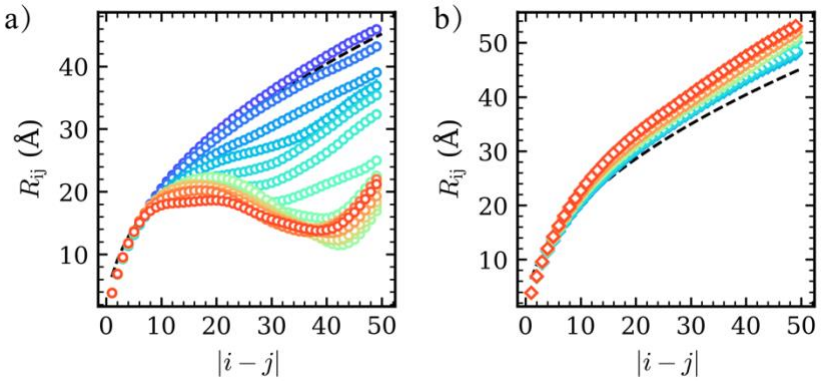

Fig. S18 Interresidue distance R_{ij} as a function of residue separation in the chain $|i-j|$ for the EKVs in the (a) single-state and the (b) dense phase. The dashed line corresponds to the ideal chain scaling $R_{ij} = b|i-j|^{1/2}$, where $b = 6.39\text{Å}$ was fitted for EKV1 using the theoretically expected end-to-end distance $R_e^2 = Nb^2$ of an ideal chain with $N = 50$. The symbol color, ranging from purple to red, indicates increasing nSCD. The mean values are obtained by dividing the trajectory into 5 independent blocks. Error bars in all panels indicate standard errors about the mean. In some cases, error bars are smaller than the symbol size.

References

(1) Ashbaugh, H. S.; Hatch, H. W. Natively Unfolded Protein Stability as a Coil-to-Globule Transition in Charge/Hydropathy Space. *J. Am. Chem. Soc.* **2008**, *130* (29), 9536–9542. https://doi.org/10.1021/ja802124e.

(2) Dignon, G. L.; Zheng, W.; Kim, Y. C.; Best, R. B.; Mittal, J. Sequence Determinants of Protein Phase Behavior from a Coarse-Grained Model. *PLOS Comput. Biol.* **2018**, *14* (1), e1005941. https://doi.org/10.1371/journal.pcbi.1005941.

(3) Kapcha, L. H.; Rossky, P. J. A Simple Atomic-Level Hydrophobicity Scale Reveals Protein Interfacial Structure. *J. Mol. Biol.* **2014**, *426* (2), 484–498. https://doi.org/10.1016/j.jmb.2013.09.039.

(4) Regy, R. M.; Thompson, J.; Kim, Y. C.; Mittal, J. Improved Coarse-Grained Model for Studying Sequence Dependent Phase Separation of Disordered Proteins. *Protein Sci.* **2021**, *30* (7), 1371–1379. https://doi.org/10.1002/pro.4094.

(5) Hückel, E.; Debye, P. The Theory of Electrolytes: I. Lowering of Freezing Point and Related Phenomena. *Phys Z* **1923**, *24* (185–206), 1.

(6) Anderson, J. A.; Glaser, J.; Glotzer, S. C. HOOMD-Blue: A Python Package for High-Performance Molecular Dynamics and Hard Particle Monte Carlo Simulations. *Comput. Mater. Sci.* **2020**, *173*, 109363. https://doi.org/10.1016/j.commatsci.2019.109363.

(7) https://github.com/mphowardlab/azplugins