Supporting Information: Towards Stable Biologics: Understanding Co-Excipient Effects on Hydrophobic Interactions and Solvent Network Integrity

Jonathan W. P. Zajac,[†,‡] Praveen Muralikrishnan,[¶,‡] Caryn L. Heldt,[§] Sarah L. Perry,[‖] and Sapna Sarupria[*,†,‡]

†*Department of Chemistry, University of Minnesota, Minneapolis, MN 55455, USA*

‡*Chemical Theory Center, University of Minnesota, Minneapolis, MN 55455, USA*

¶*Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN 55455, USA*

§*Department of Chemical Engineering, Michigan Technological University, Houghton, MI 49931, USA*

‖*Department of Chemical Engineering, University of Massachusetts Amherst, MA 01003, USA*

E-mail: sarupria@umn.edu

PMF Convergence Checks

In Figure S1, the Kolmogorov-Smirnov (K-S) statistics is used as a measure of similarity between two PMFs, taken at different time points. In column 3, the K-S statistics represent the similarity between a PMF taken after 5t ns versus 5(t+1) ns. Plotting the data in this manner reflects how the PMF evolves over time. It is expected that as the simulations are run longer, the K-S value between consequent windows will decrease suggesting convergence. In column 4, the K-S statistics represent the similarity between a PMF taken after 5t ns and the PMF obtained at the end of the simulation. This comparison shows how and when a given system converges. Simulations were stopped once the deviation between replicate PMFs was observed to fluctuate around an average K-S value of less than 0.2, resulting in total simulation times per replica of between 100-250 ns. Across 3 replicate simulations, each with 12 windows, an aggregate simulation time of 3.6-9.0 μs was carried out for each system.

Table S1: Setup of simulated systems. A range of used N_{wat} is provided, given the differences in excipient sizes.

System	Simulation Time (ns)	Concentration (M)	N_{Exc}	N_{wat}
Excipient	20	0.25	47-48	9700-10000
Excipient	20	0.50	93-94	9100-9700
Excipient	20	1.0	185-186	8000-9100
Polymer	100 x 12	0.00	0	10188
Polymer + Excipient	3 x 100 x 12	0.25	47-48	9700-10100
Polymer + Excipient	3 x 100 x 12	0.50	93-94	9100-9600
Polymer + Excipient	3 x 250 x 12	1.0	185-186	8000-8600

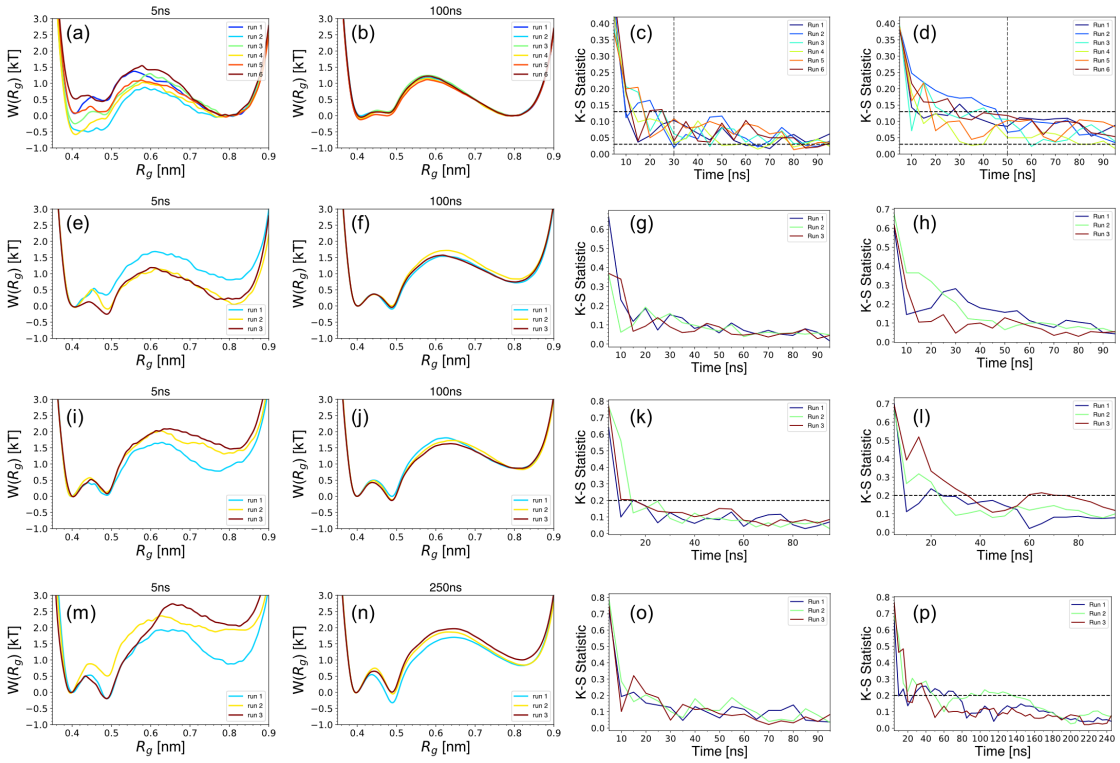


Figure S1: PMF convergence checks for REUS simulations. Hydrophobic polymer in (a-d) water, (e-h) 0.25 M arginine, (i-l) 0.5 M arginine, and (m-p) 1.0 M arginine. In the first column, PMFs obtained after 5 ns per replica are shown. In the second column, PMFs obtained after 100 or 250 ns are shown, highlighting converged replicate simulations. In column 3, Kolmogorov-Smirnov (K-S) statistics are plotted as a comparison for PMFs obtained in 5 ns blocks ($5t$ vs $5(t+1)$). In column 4, K-S statistics are plotted as a comparison between PMFs at time $5t$ vs the PMF obtained at either 100 ns (d,h,l) or 250 ns (p). t goes from 1 through 19 for 100 ns, and 1 through 49 for 250 ns.

Error Calculations

The errors for PMF were calculated through the propagation of uncertainty using 3 replicate simulations (N = 3). The derivation of uncertainty in the free energy of unfolding is shown below. σ represents the standard deviation, exp represents the exponential term, ln represents the logarithmic term and int represents the integral.

$$\Delta G_{\text{unfold}} = -k_B T \ln \frac{\int_{R_g^{\text{cut}}}^{R_g^{\max}} \exp\left(-\frac{W(R_g)}{k_B T}\right) dR_g}{\int_{R_g^{\min}}^{R_g^{\text{cut}}} \exp\left(-\frac{W(R_g)}{k_B T}\right) dR_g} \tag{1}$$

The integral is approximated as a sum and divided into discrete bins in the Rg coordinate. The Rg space (from 0.3 to 0.9 nm) is divided into 600 bins, giving a $\Delta R_g = 0.001$ nm.

$$\sigma_{W(R_g)} = \sqrt{\frac{\sum \left(W\left(R_g\right)_i - \mu_{W(R_g)}\right)^2}{N}} \tag{2}$$

$$\sigma_{\exp} = \left|\exp\left(-\frac{W\left(R_g\right)}{k_B T}\right)\right| * \left|\frac{1}{k_B T} * \sigma_{W(R_g)}\right| \tag{3}$$

$$\sigma_{\text{int}} = \Delta R_g * \sqrt{\sum \sigma_{\exp}^2} \tag{4}$$

$$\sigma_{ln} = \frac{\sigma_{\text{int}}}{\text{int}} \tag{5}$$

$$\sigma_{\Delta G} = k_B T * \sqrt{(\sigma_{ln})_{num}^2 + (\sigma_{ln})_{den}^2} \tag{6}$$

The errors in PMF decomposition were calculated using error propagation rules. An example of error calculation for ΔE_{unfold} is shown below:

$$\Delta E_{unfold} = \langle E \rangle_u - \langle E \rangle_f \tag{7}$$

$$\langle E\rangle_f = \frac{\sum_{r_{\min}}^{r_{\text{cut}}} E\left(R_g\right) P\left(R_g\right)}{\sum_{r_{\min}}^{r_{\text{cut}}} P\left(R_g\right)}, \quad \langle E\rangle_u = \frac{\sum_{r_{\text{cut}}}^{r_{\max}} E\left(R_g\right) P\left(R_g\right)}{\sum_{r_{\text{cut}}}^{r_{\max}} P\left(R_g\right)} \tag{8}$$

$$\sigma_{E(R_g)} = \sqrt{\frac{\sum \left(E\left(R_g\right)_i - \mu_{E(R_g)}\right)^2}{N}} \tag{9}$$

$$\sigma_{\langle E\rangle} = \frac{\sqrt{\sum_{r_{\min}}^{r_{\text{cut}}} \sigma_{E(R_g)}^2 P\left(R_g\right)^2}}{\sum_{r_{\min}}^{r_{\text{cut}}} P\left(R_g\right)} \tag{10}$$

$$\sigma_{\Delta E} = \sqrt{\sigma_{int,f}^2 + \sigma_{int,u}^2} \tag{11}$$

Water Dynamics Analysis

Clustering was achieved via the leaf algorithm of HDBSCAN.[S1] The minimum cluster size parameter was set to 100, while the minimum samples parameter was set to 50. Clustering was carried out on the principal moments of the gyration tensor of the hydrophobic polymer. Data were obtained from the final 100 ns in each window, saving coordinates every 100 ps. Data points not belonging to clusters were removed, for clarity. Clusters identified in principal moment space were projected onto end-to-end vs radius of gyration space (Figs. S2 and S3). Representative configurations corresponding to the highest cluster membership probability were then used to start the unbiased simulations. The unbiased simulations were performed for 300 ps in the NPT ensemble and configurations were stored every 0.1 ps. Water reorientation times were then computed as described in the Methods and Scheme S1.

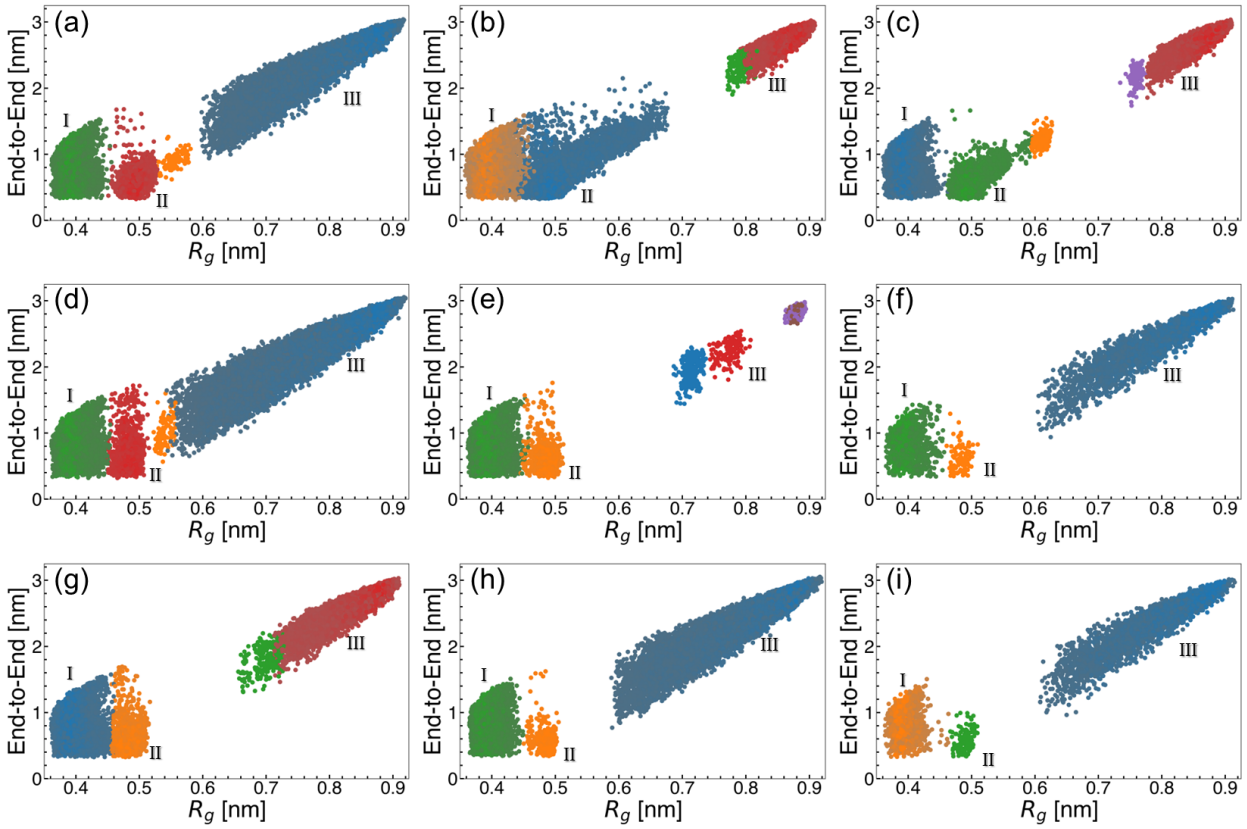


Figure S2: HDBSCAN clustering of polymer configurations for (a-c) Arg, (d-f) Glu, and (g-i) Lys solutions.

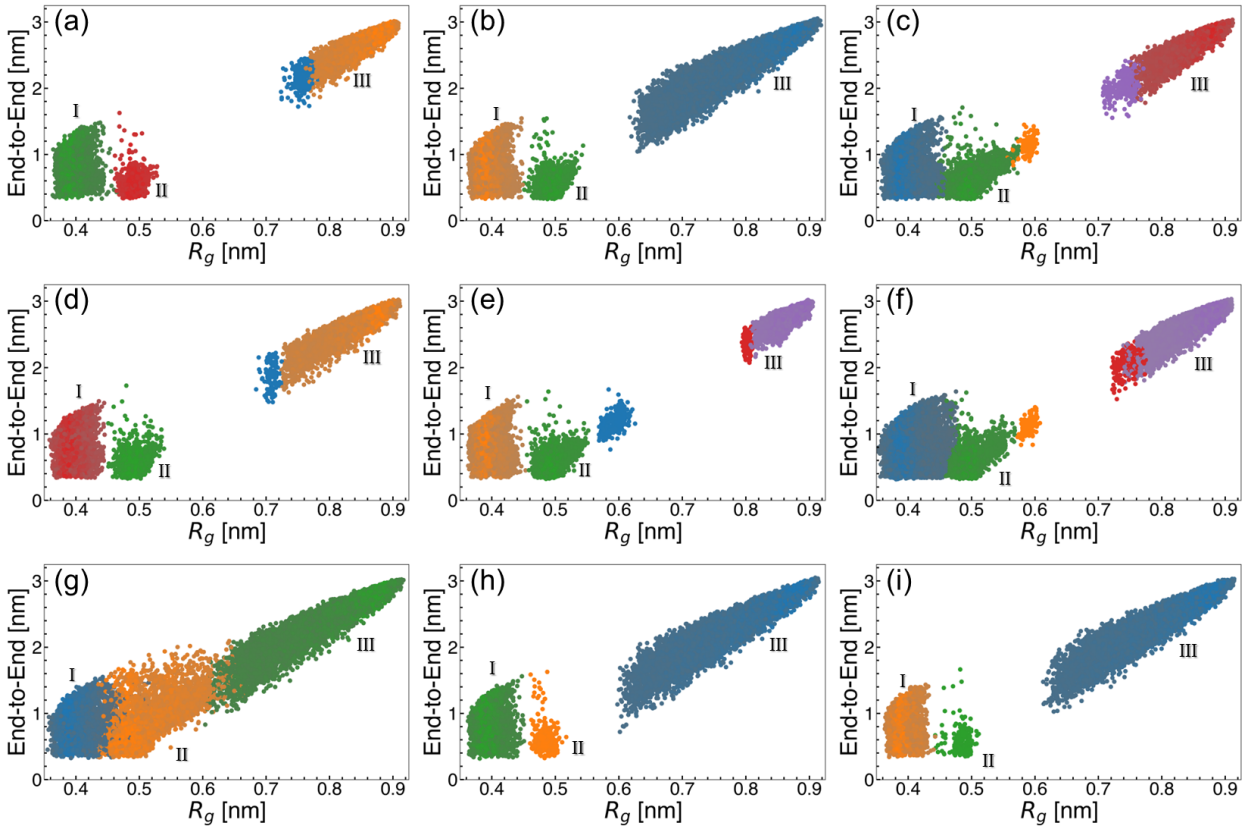
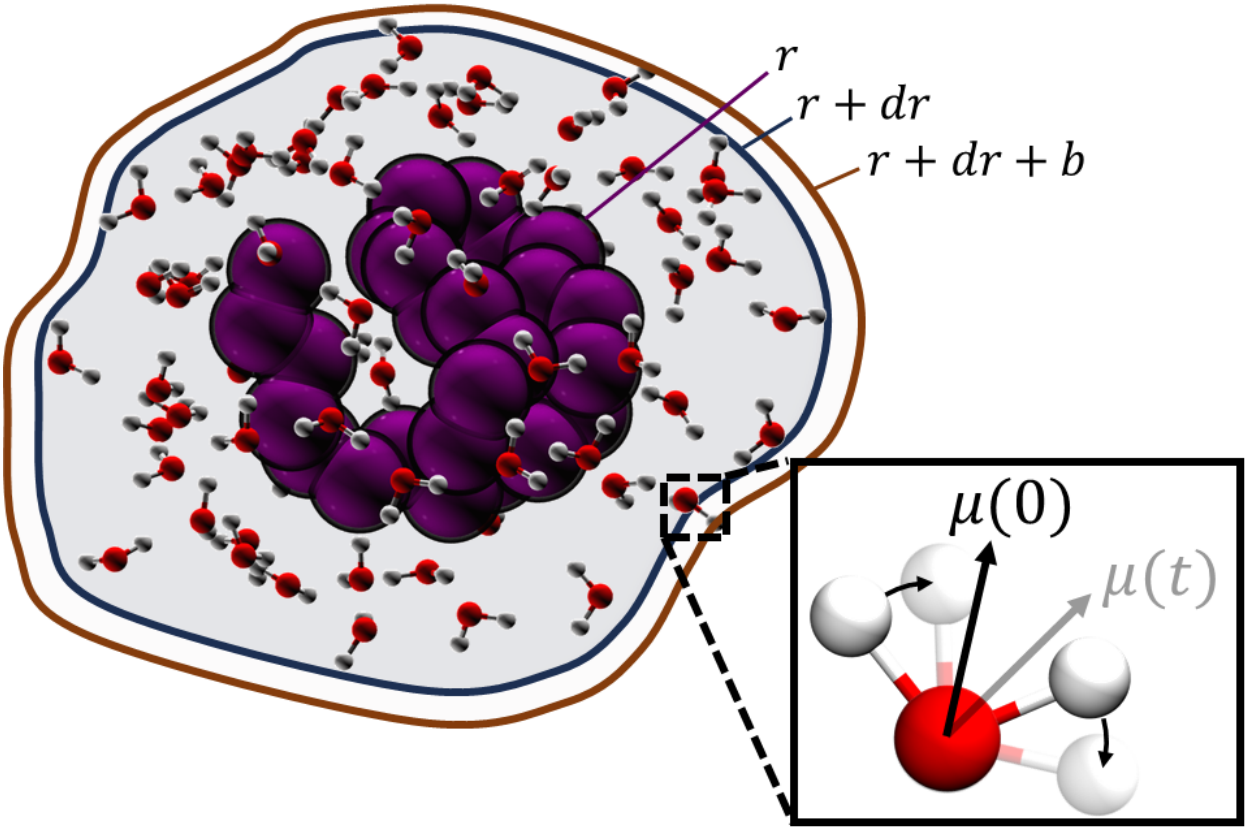


Figure S3: HDBSCAN clustering of polymer configurations for (a-c) Arg/Glu, (d-f) Arg/Lys, and (g-i) Lys/Glu solutions.



Scheme S1: Schematic for water reorientation time calculation. The hydrophobic polymer is shown as purple spheres. r denotes the van der Waals surface of the hydrophobic polymer, $r + dr$ denotes the region included in the calculation, and $r + dr + b$ denotes the furthest edge of the buffer region.

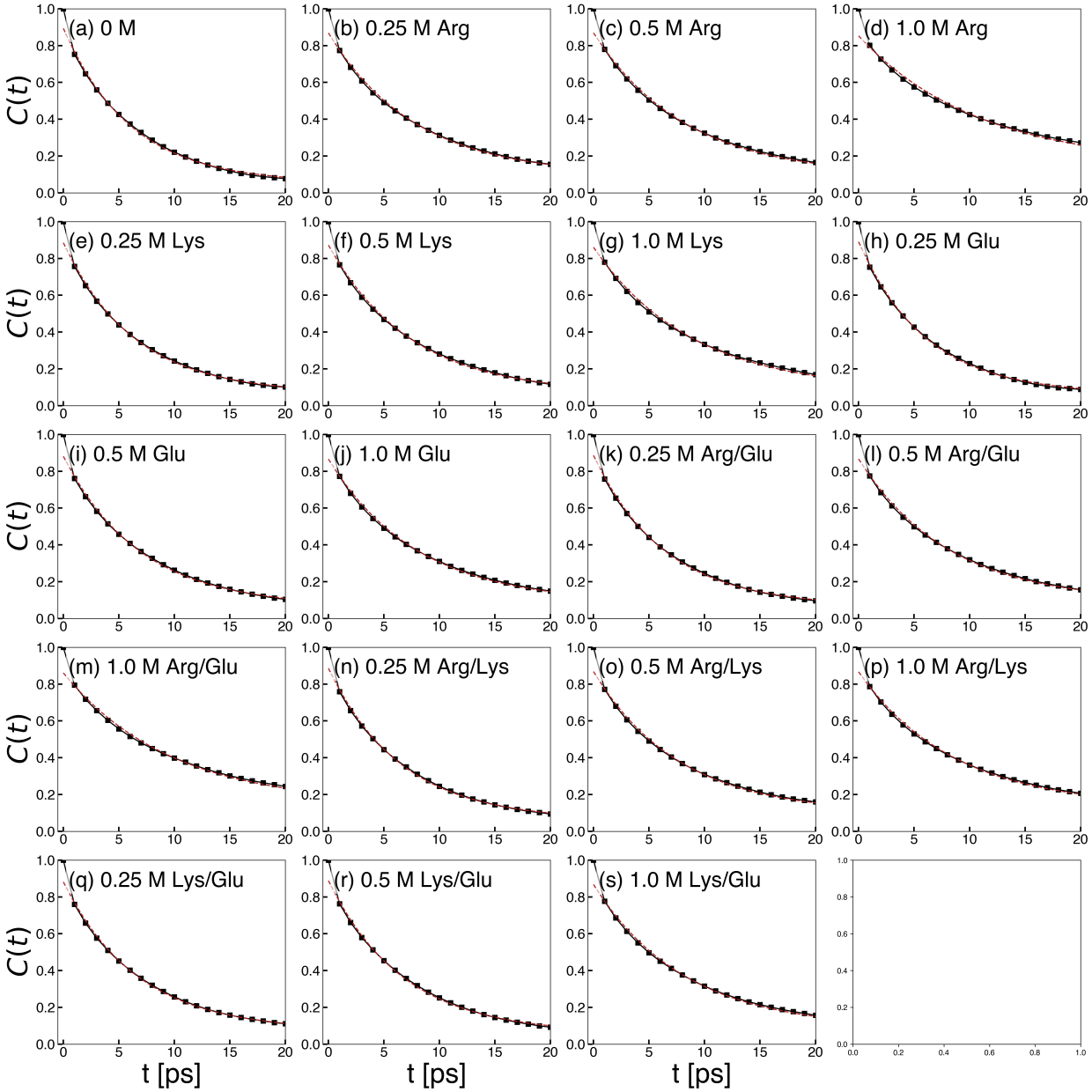


Figure S4: Water dipole vector correlation function (C(t); black) plotted against the exponential fit (red) used in τ estimation. (a) Water, (b-d) Arg/water, (e-g) Lys/water, (h-j) Glu/water, (k-m) Arg/Glu, (n-p) Arg/Lys, and (q-s) Lys/Glu solutions.

Preferential Interaction Coefficients

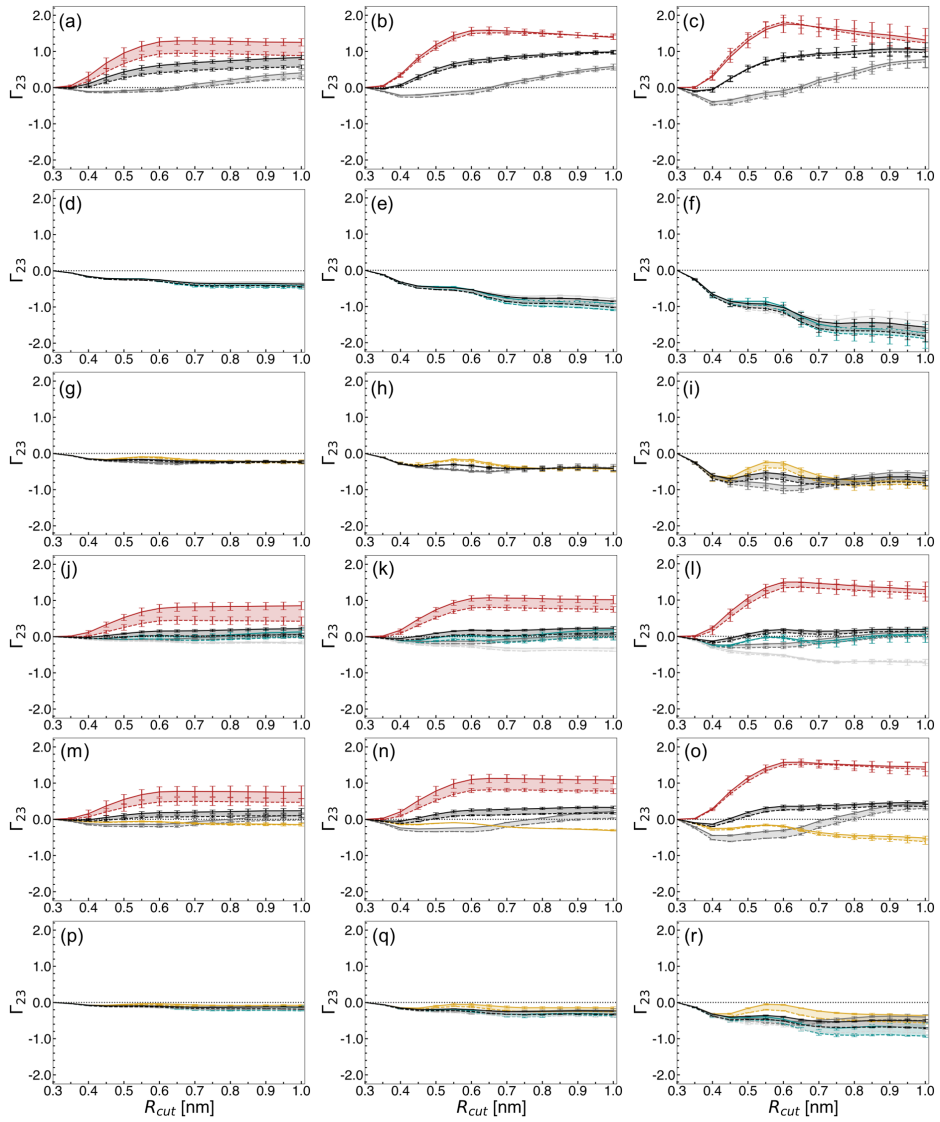


Figure S5: Preferential interaction coefficients for (a-c) Arg/water, (d-f) Glu/water, (g-i) Lys/water, (j-l) Arg/Glu, (m-o) Arg/Lys, and (p-r) Lys/Glu solutions. Arginine is colored in red, glutamate in blue, lysine in yellow, sodium in light gray, and chloride in dark gray. In all plots, the net preferential interaction coefficient is colored in black. Dashed lines indicate values for the unfolded state, while solid lines denote the folded state. Concentration increases from left to right in the order 0.25 M, 0.5 M, and 1.0 M. Mean values are reported from three replicate REUS simulations. Error bars were estimated as standard deviations from three replicate simulations.

Network Analysis

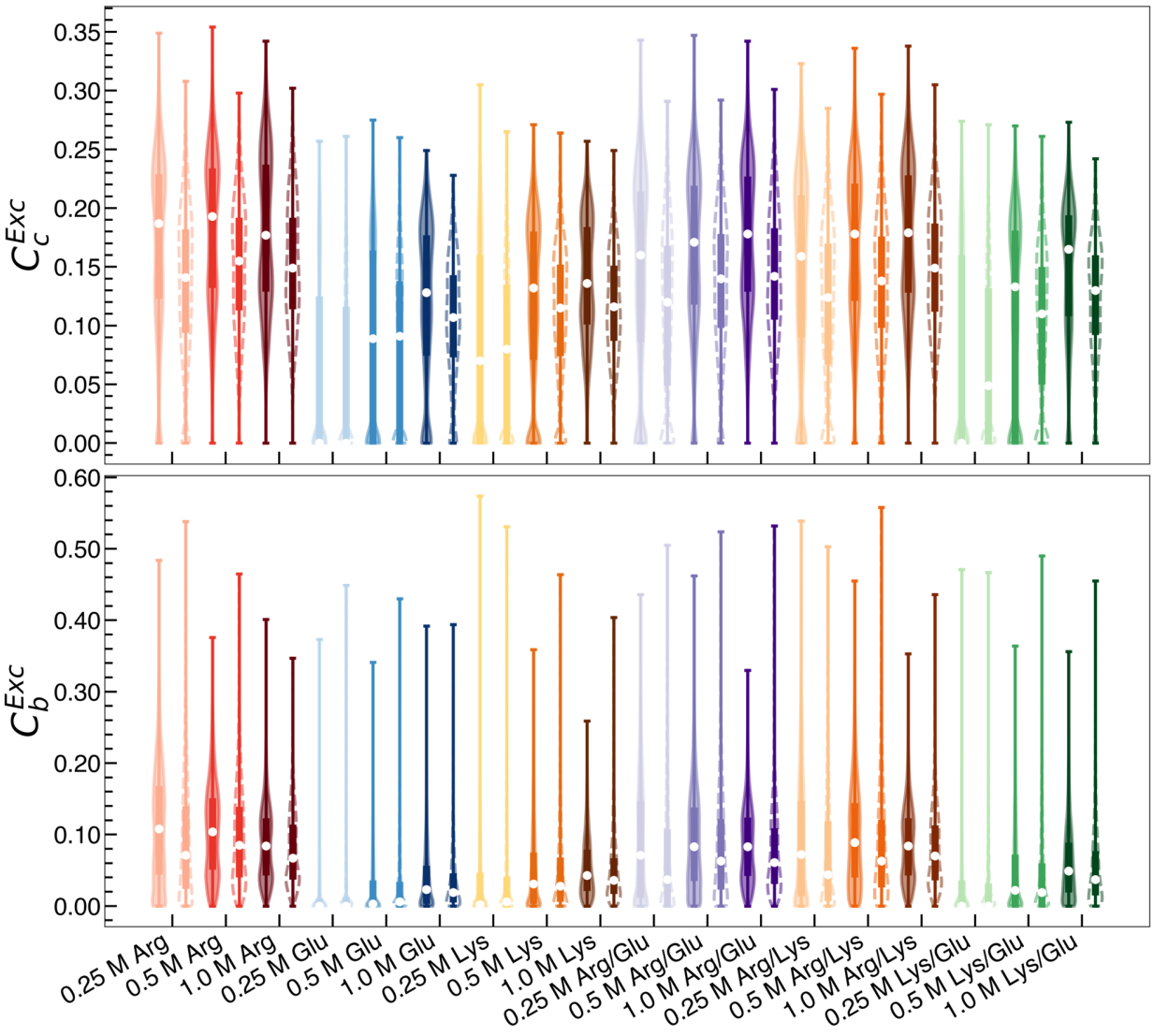


Figure S6: Excipient (top) closeness centrality and (bottom) betweenness centrality. Distributions with solid lines denote the folded ensemble, while dashed lines denote the unfolded ensemble.

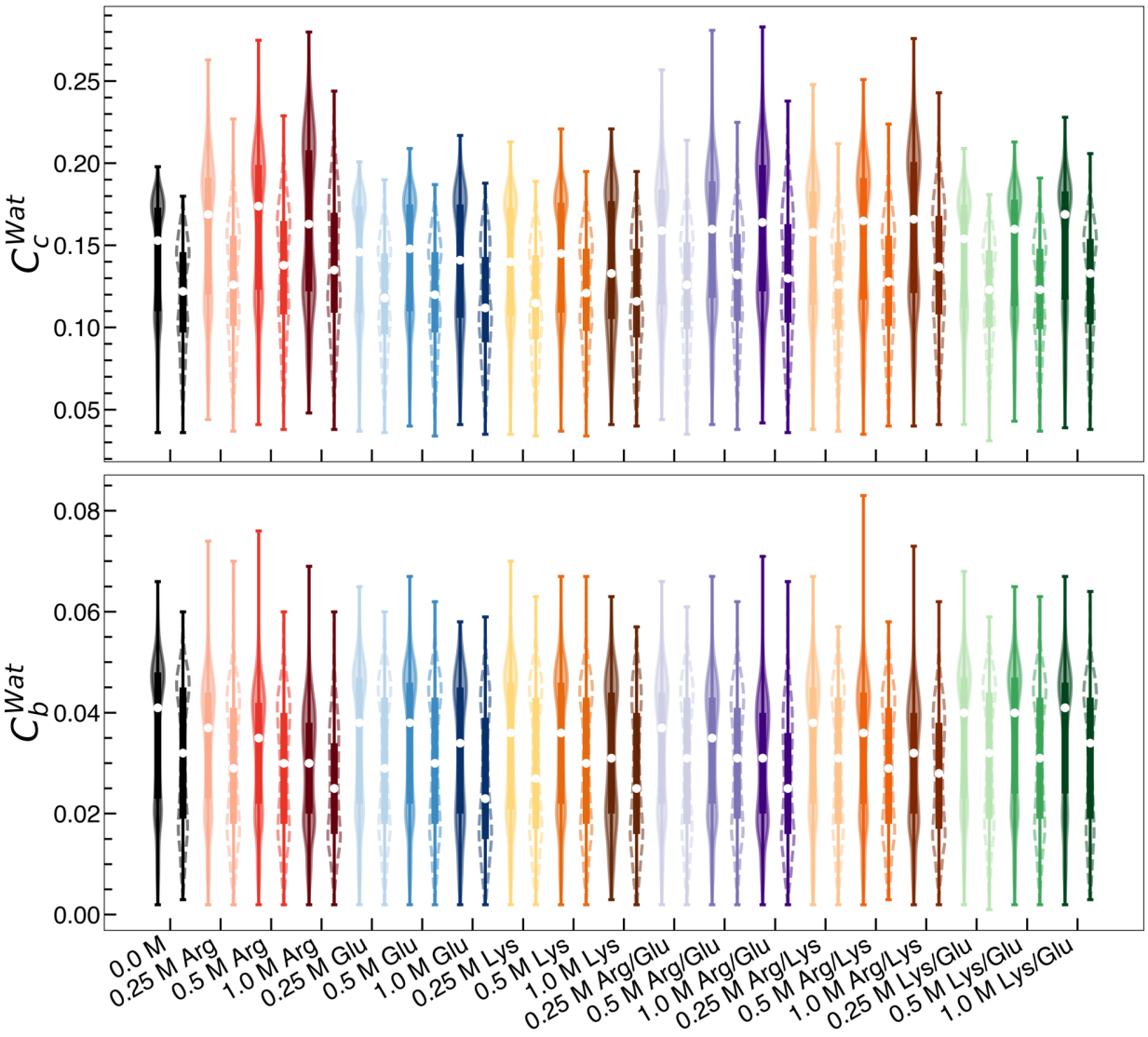


Figure S7: Water (top) closeness centrality and (bottom) betweenness centrality. Distributions with solid lines denote the folded ensemble, while dashed lines denote the unfolded ensemble.

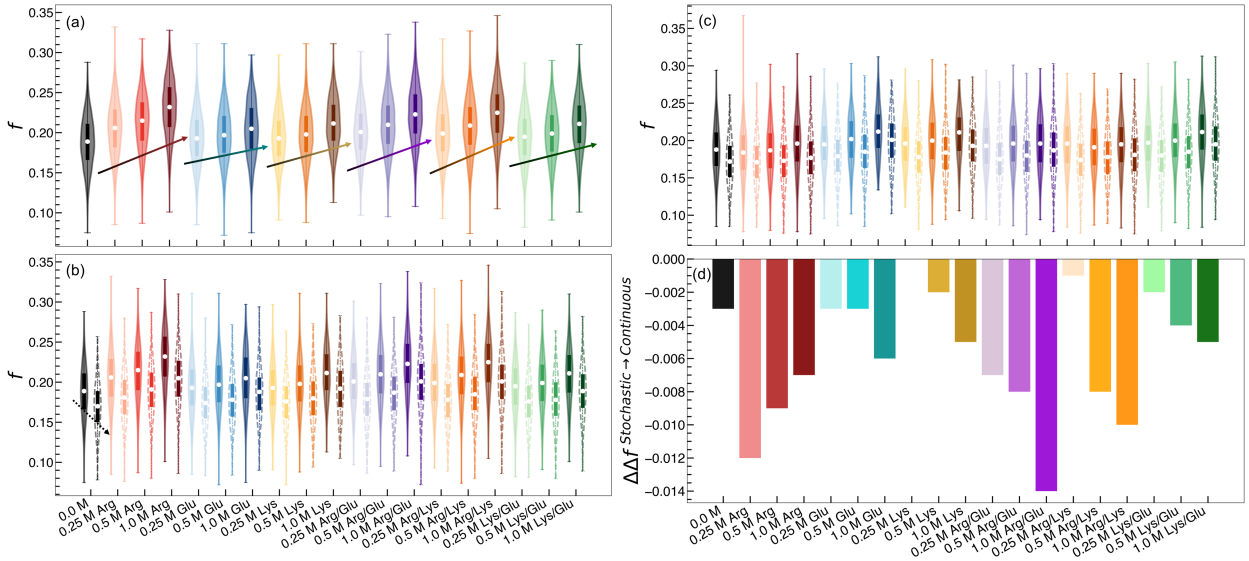


Figure S8: Network stability of the local polymer environment. (a) Fragmentation threshold, f, for all solutions. Distributions are taken over all folded state configurations. Arrows are drawn to guide changes with increasing concentration for a given excipient solution. (b) Fragmentation threshold distributions split between folded (solid) and unfolded (dashed) states. An arrow is drawn for the 0.0 M solution to guide changes in distribution upon polymer unfolding. (c) Fragmentation threshold distributions following sequential removal of continuously connected nodes. (d) Change in fragmentation threshold between folded and unfolded states (Δf) and between stochastic and continuously connected node removal ($\Delta\Delta f$).

References

(S1) McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* **2017**, *2*, 205.