Theory for a Sequence-Specific “Fuzzy” Binding Mechanism Between a Pair of Intrinsically Disordered Proteins

Alan N. Amin,†§ Yi-Hsuan Lin,†,‡§ Suman Das,† and Hue Sun Chan*,†,¶

†Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada
‡Molecular Medicine, Hospital for Sick Children, Toronto, Ontario, Canada
¶Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
§Contributed equally to this work

E-mail: huesun.chan@utoronto.ca
Abstract

Many intrinsically disordered proteins (IDPs), proteins that do not fold into a unique tertiary structure in isolation because their depletion in hydrophobicity whereas enrichment in charge and aromaticity, retain their structural disorderness when participating in IDP-IDP binding and IDP phase separation via a sequence-specific “fuzzy” molecular recognition mechanism. We implement the cluster expansion method in statistical mechanics to develop an analytical theory for elucidating the nature of the fuzzy Coulomb interaction between two charged IDPs. We apply the theory to calculate binding affinities between various pairs of IDP sequences, and extract a parameter \( j_{SCD} \) encompassing the sequence-specific IDP-IDP interaction. Our theory provides a general framework to investigate IDP-IDP interaction.
Intrinsically disordered proteins (IDPs) do not fold into a unique structure in isolation because they are lacking hydrophobicity but enriched in charged, polar, and aromatic residues. While many intrinsically disordered proteins (IDPs) fold upon binding their target, many instead interact with their folded partner in a complex with large conformational heterogeneity. Recently, two strongly charged IDPs, ProTα and H1, have been observed to retain disordered without forming any significant, stable tertiary structures when forming a heterodimer complex with dissociation constant reported to be in sub-micromolar to nanomolar level. These IDPs interact with their partners via a so-called “fuzzy” mechanism that an IDP exploits its whole structural ensemble to facilitate the interaction, instead of relying on solid complementary conformations as the folded globular proteins do.

Electrostatic interaction often play a critical role in such fuzzy molecular mechanism, involved in not only IDP binding but also IDP-driven liquid-liquid phase separation. It has been noted that changing charge sequence while retaining amino acid composition of an IDP can result in dramatically different radii of gyration. Two measures have been proposed to quantify the different charge patterns: κ is an intuition-based parameter measuring the local charge blockiness along an IDP primary sequence, and sequence charge decoration (SCD) is a parameter derived from an analytical theory, given by

$$SCD(\{\sigma\}) = \frac{1}{2N} \sum_{s,t=1}^{N} \sigma_s \sigma_t \sqrt{|s-t|}$$

for charge sequence $\{\sigma\} = \{\sigma_1, \sigma_2, ... \sigma_N\}$. Both of these measures were recently found to be evolutionarily conserved among IDPs in a bioinformatic screen, demonstrating the significance of charge pattern on modulating protein interactions in the cell. Such sequence-specific mechanism has also been observed and investigated in IDP phase separation systems. There are efforts on disclosing the correlation between the single-chain structural properties and the multiple-chain phase behaviors. The interaction between individual pairs of IDPs, which can be quantified by the osmotic second virial coefficient, $B_2$, is also...
proposed to be a measure for IDP phase separation propensity. However, to obtain knowledge about IDP-IDP interaction, a wide variety of transient, fast-switching structures in the two IDP structural ensembles have to be well sampled and investigated, which is costly by means of both experiments and computer simulation citation needed. A concise measure for IDP-IDP binding propensity and coacervation between same or different species of IDP sequences will greatly help understand the fuzzy molecular mechanism and, moreover, facilitate high-throughput IDP bioinformatics screen.

Here, we present an analytic theory for evaluating pairwise IDP-IDP interaction. We consider two charged IDPs $A$ and $B$ with lengths $N_A$ and $N_B$, charge sequences $\{\sigma^A\} = \{\sigma^A_1, \sigma^A_2...\sigma^A_{N_A}\}$ and $\{\sigma^B\} = \{\sigma^B_1, \sigma^B_2...\sigma^B_{N_B}\}$, and residue coordinates $\{R^A\} = \{R^A_1, R^A_2...R^A_{N_A}\}$ and $\{R^B\} = \{R^B_1, R^B_2...R^B_{N_B}\}$, respectively. We allow both $A = B$ and $A \neq B$ for homotypic and heterotypic interactions, respectively. For these two polymers with internal degrees of freedom, the second virial coefficient $B_2$ is given by

$$B_2 = \int dR^A_{CM} \left( 1 - e^{-\beta U^{AB}(R^A_{CM}, \{R^A\}, \{R^B\}))} \right)_{A,B}, \quad (2)$$

where $\beta = 1/(k_B T)$, $R^A_{CM}$ is the distance between the centers of mass of $A$ and $B$, $U^{AB}$ is the overall interaction between $A$ and $B$, and $\langle \cdots \rangle_{A,B}$ is an average over the conformational ensembles of $A$ and $B$. As derived in Supporting Information, Eq. 2 is equivalent to a representation that takes the average of $U^{AB}$ reweighted by probability distributions of the IDPs, namely

$$B_2 = V \int \mathcal{D}[R^A] \mathcal{D}[R^B] \mathcal{P}^A[R^A] \mathcal{P}^B[R^B] \left( 1 - e^{-U^{AB}(R^A, R^B)} \right), \quad (3)$$

where $V$ is system volume and $\mathcal{D}[R^i] = \int \prod_{s=1}^{N_i} dR^i_s$ with $i = A, B$. The probability integral can also be presented as the ratio between the conformational partition functions of $A$ and $B$, denoted as $Q_A$ and $Q_B$, respectively, and the partition function of the $A-B$ complex $Q_{AB}$,
namely

\[ B_2 = V - \frac{Q_{AB}}{Q_A Q_B}. \tag{4} \]

We have provided the proof for the equivalence of Eqs. 2, 3, and 4 in Supporting Information. We note that all three equations retard to the common formula \( B_2 = \int d^3r \{ 1 - \exp[-U_{AB}(r)/k_B T] \} \) when \( N_A = N_B = 1 \).

We continue our derivation from the probability distribution representation in Eq. 3. We consider the intermolecular interaction \( U^{AB} \) a summation of pairwise interactions between residues in different polymers, namely

\[ U^{AB}[\mathbf{R}^A, \mathbf{R}^B] = \sum_{s=1}^{N_A} \sum_{t=1}^{N_B} \mathcal{V}^{AB}_{st}(\mathbf{R}^{AB}_{st}), \tag{5} \]

where \( \mathbf{R}^{AB}_{st} \equiv \mathbf{R}^A_s - \mathbf{R}^B_t \) and \( \mathcal{V}^{AB}_{st} \) is the interaction between the residue pair \((s^A, t^B)\) denoting the \( s \)th residue in protein \( A \) and the \( t \)th residue in protein \( B \). The integrand in Eq. 3 can then be evaluated by the cluster expansion of \( \exp(-U^{AB}) - 1 \), which is given by

\[
e^{-U^{AB}[\mathbf{R}^A, \mathbf{R}^B]} - 1 = \prod_{s=1}^{N_A} \prod_{t=1}^{N_B} \left( e^{-\mathcal{V}^{AB}_{st}(\mathbf{R}^{AB}_{st})} - 1 \right)
= \sum_{s=1}^{N_A} \sum_{t=1}^{N_B} f_{st} + \sum_{s \geq t=1}^{N_A} \sum_{l \geq m=1}^{N_B} f_{st} f_{tm} - \sum_{s=1}^{N_A} \sum_{t=1}^{N_B} f_{st}^2 + O(f^3), \tag{6} \]

where \( f_{st} = \exp[-\mathcal{V}^{AB}_{st}(\mathbf{R}^{AB}_{st})] - 1 \) is the Mayer f-function for the pair of residues \((s^A, t^B)\). We note that the first term of \( f_{st} \) and the third term of \( f_{st}^2 \) include only the interaction between an intermolecular residue pair, and thus the single-polymer conformational probabilities \( \mathcal{P}^A, \mathcal{P}^B \) have no influence in the integral; the second term of \( f_{st} f_{tm} \), however, includes two different intermolecular residue pairs, which are correlated with each other via the conformational ensembles. We show the details of the derivation in Supporting Information, and simply summarize the results here: by Fourier transformation, we define \( \mathbf{k} \)-space matrices for the
intramolecular residue-residue correlation function,

\[ \hat{P}^i(k)_{st} \equiv \int D[R^i] D[R^j] e^{i\mathbf{k}(\mathbf{R}^i - \mathbf{R}^j)} , \quad i = A, B, \] (7)

and the Mayer f-function

\[ \hat{f}(k)_{st} \equiv \int dr f_{st}(r) e^{i\mathbf{k} \cdot \mathbf{r}}, \] (8)

and obtain the \( O(f^2) \) cluster expansion representation of \( B_2 \) as

\[
B_2 = - \sum_{s=1}^{N_A} \sum_{t=1}^{N_B} \hat{f}(0)_{st} - \frac{1}{2} \int \frac{d^3k}{(2\pi)^3} \text{Tr} [\hat{f}(k) \hat{P}^B(-k) \hat{f}^T(k) \hat{P}^A(-k) - \hat{f}(k) \hat{f}^T(-k)] + O(f^3),
\] (9)

where the “T” superscript denotes transpose of the matrices.

In this work we focus on the electrostatic interaction in charged IDPs; we consider a screened Coulomb potential, \( V_{st}^{ij}(r) = l_b \sigma_s^i \sigma_t^j \exp(-\kappa r)/r \), which in \( k \)-space results in interaction matrix \( \hat{V}(k) \) with elements defined by

\[
\hat{V}(k)_{st} = \frac{4\pi l_b}{k^2 + \kappa^2} \sigma_s^i \sigma_t^j, \] (10)

where \( l_b = e^2/(4\pi\varepsilon k_BT) \) is Bjerrum length and \( \kappa \) is Debye screening wave number. For simplification and tractability, we take two approximations from the general representation in Eq. 9. First, we approximate the IDPs to Gaussian chains with Kuhn length \( b \), which results in correlation functions

\[
[\hat{P}^i(k)]_{st} \approx [\hat{G}(k)]_{st} = e^{-\frac{1}{2}(kb)^2 |s-t|};
\] (11)

with \( k \equiv |k| \). Second, we apply high-temperature expansion to Mayer f-functions, namely

\[
[\hat{f}(k)]_{st} = -\frac{4\pi l_b}{k^2 + \kappa^2} \sigma_s^i \sigma_t^j + \frac{8\pi^2 l_b^2}{(k^2 + \kappa^2)^2} \left( \sigma_s^i \right)^2 \left( \sigma_t^j \right)^2 + O(l_b^3).
\] (12)
With the two approximations, we obtain a $B_2$ formula up to $O(l_b^2)$ as

$$B_2 \approx 4\pi l_b^2 q^A q^B - 4l_b^2 \int \frac{dk k^2}{(k^2 + \kappa^2)^2} \sum_{s,t=1}^{N_A} \sum_{l,m=1}^{N_B} \sigma_s^A \sigma_t^A \sigma_l^B \sigma_m^B e^{-\frac{1}{\kappa}(k(k)|s-t|+l-m)|}$$

where $q^i = \sum_{s=1}^{N_i} \sigma^i$ is the net charge of protein $i$. The first term in Eq. S24 is the mean-field Coulomb interaction between net charges of the two proteins, and the second term accounts for sequence specificity.

**Binding affinity.** We first apply the $B_2$ derived from Eq. S24 to calculate binding affinity of H1 and ProTα. We consider a solution including protein A and B with the same concentration $[c]$. To calculate $B_2$, the volume $V$ is defined as the space including only a pair of A and B, which is given by $[c] = 1/V$. Defining binding probability $\theta$, the A-B unbinding probability is given by

$$1 - \theta \equiv \frac{V Q_A Q_B}{Q_{AB}} = \frac{1}{1 - B_2/V},$$

where $V$ accounts for the degrees of freedom of the distance between the two independent proteins in the unbound state. The dissociation constant in terms of $\theta$ is given by

$$\frac{1}{K_D} = \frac{\theta[c]}{(1 - \theta)^2[c]^2} = -B_2(1 - B_2/V)$$

$$\approx -B_2,$$

We first applied our theory to the binding of H1 and ProTα. The theoretical results are compared with the published experiments of isothermal calorimetry (ITC) and single molecular FRET (smFRET) in Fig. 1. As shown in the figure, the theory and the two experiments show the same trend that adding salt depletes the binding reaction.

Quantitatively, our theoretical $K_D$’s agree better with ITC results as they are within an order of magnitude, showing merely a 3.5% difference from experimental data at 350mM.
NaCl; however, at 165mM NaCl, theoretical $K_D$ is about ten times that of the ITC result (3.41 v.s. 0.46µM). This discrepancy may be a result of the higher order attractive terms in the cluster expansion in Eq. S12 which are not taken into account in our formula of $B_2$. When salt concentration is high and screening effect is strong, contributions from higher order terms decay rapidly, and our formula of second-order expansion works well.

Compared to smFRET results, however, our theory gives much larger $K_D$ and much weaker salt-concentration dependence. The discrepancy may be attributed to the role of ion condensation in H1-ProTα binding, which is not taken into account in our theory whereas this effect was deemed to be the major driving force leading to the extremely strong salt concentration dependency ($K_D$ increases 1,000,000 folds as NaCl concentrations increases from 150 to 350mM) inferred by smFRET. Possible ion condensation effects on H1-ProTα binding remain to be elucidated by further experimental as well as theoretical investigations.

In Table 1 we also list the $K_D$ calculated by only considering the net charges on H1 and ProTα, which greatly underestimates H1-ProTα binding affinity, yielding $K_D$’s that are two orders of magnitude larger than ITC data. Clearly, the dynamic conformational ensembles of H1 and ProTα, which is taken into account by the $F_2$ in Eq. S24 is indispensable in their binding reaction. Considering that our Gaussian-chain approximation might not be a good model for the N-terminal globular domain of H1, we also apply our theory to the binding of full-length ProTα and the fully disordered C-terminal domain of H1, termed H1-CTD. The resulting $K_D$’s are about 1.2–1.5 times higher than those of full-length H1, as shown in Table 1. The difference between full-length and C-terminal H1 can be attributed to the subtraction of the +18 charges in its N-terminal domain.

The consistency of our theory with experiment for H1-ProTα complex inspires us to derive a general measure for sequence-specificity in IDP-IDP interaction. We consider a salt-free solution of two proteins $A$ and $B$ such that at least one of them is neutral, namely $κ → 0$ and $q^Aq^B = 0$. In this case, according to the derivation in Supporting Information,
Table 1: $K_D$ (µM) of H1-ProTα complex at different NaCl concentrations

| [NaCl] (mM) | Theory H1-CTR | Net Charge | ITC     | [NaCl] (mM) | smFRET |
|-------------|---------------|------------|---------|-------------|--------|
| 165         | 3.41          | 4.59       | 142     | 0.46±0.05   | 160    |
| 220         | 5.09          | 6.77       | 189     | 0.72±0.03   | 180    |
| 260         | 6.46          | 8.55       | 223     | 2.0±0.1     | 205    |
| 300         | 7.94          | 10.46      | 257     | 6.1±0.1     | 240    |
| 350         | 9.95          | 13.06      | 300     | 9.6±0.7     | 290    |

Figure 1: Theoretical and experimental dissociation constants at different salt concentrations. Depiction of the data in Table 1.

Eq. S24 can be rewritten as

$$B_2|_{\kappa \to 0, \epsilon (q^A,q^B)} = -8\sqrt{\frac{\pi}{6}} \frac{b b_{e N_A N_B} \times jSCD(\sigma^A, \sigma^B)}{1}, \quad (16)$$

where jSCD is the newly defined “joint sequence charge decoration”,

$$jSCD(\sigma^A, \sigma^B) \equiv -\frac{1}{2N_A N_B} \sum_{s,t=1}^{N_A} \sum_{l,m=1}^{N_B} \sigma^A_s \sigma^A_t \sigma^B_l \sigma^B_m \sqrt{|s-t| + |l-m|}, \quad (17)$$

We note that even though the jSCD is derived from the condition $q^A q^B = 0$, it can also be applied to compare phase separation propensity among a set of IDP pairs with the same charge composition of arbitrary net charges, as can be seen from Eq. S33 in Supporting
Information.

To generate predictions from our model and also elucidate the mathematical characteristics of jSCD, we considered the 30 model charge neutral polymers shown in Fig. 2a. The 30 sequences, labeled as sv1 to sv30, are proposed by Das and Pappu, each consisting of the same charge composition, 25 lysines (K) with +1 charges and 25 glutamates (E) with -1 charges, but different charge patterns\textsuperscript{11} quantified by parameters $\kappa$\textsuperscript{11} and SCD\textsuperscript{19} As all the 30 sequences are neutral, Eq. 16 is applicable. Calculating the predicted $K_D$ for all pairs of these sequences reveals dramatic binding constants ranging from under 5 $\mu$M to over 2 mM, which indicates that a wide range of different IDP-IDP binding affinities can be obtained via changing sequence pattern without interfering amino acid composition. Generally, $K_D$ decreases and binding affinity increases with the charge segregation of the interacting polymers as measured by both SCD and $\kappa$, as shown in Figs. 2b and S1 respectively.

![Figure 2](image)

Figure 2: (a) The 30 Das-Pappu “sv” sequences ordered by their SCD values; the number after the “sv” on the right hand side indicates their ranking by $\kappa$ parameter. (b) Heatmap of binding affinities of all 30×30 pairs of Das-Pappu sequences. Sequences with higher -SCD bind together stronger.

**Phase separation.** The parameter jSCD also appears in the RPA theory for charge neutral sequences. Substituting a salt-free electrostatic interaction without short-range cutoff into Eqs. 39 and 40 in Ref.\textsuperscript{22} we obtain an electrostatic free energy that can be expanded
to \( O(l_b^3) \) as

\[
f_{el} = \int_0^\infty \frac{dk k^2}{4\pi^2} \left\{ \ln \left[ 1 + \frac{\phi}{k^2 T^* N} (\sigma | \hat{G}(k) | \sigma) \right] - \frac{\phi}{k^2 T^* N} (\sigma | \hat{G}(k) | \sigma) \right\}
\]

\[
= - \frac{2\phi^2}{T^* N^2} \int_0^\infty \frac{dk}{k^2} (\sigma | \hat{G}(k) | \sigma)^2 + O(l_b^3)
\]

\[
= - \frac{\phi^2}{T^*} \sqrt{\frac{8\pi}{3}} \text{jSCD}(\sigma, \sigma) + O(l_b^3),
\]

where \( N, \phi \) are length and volume fraction of the protein, respectively, \( T^* = b/l_b \) is the reduced temperature, and \( \langle \sigma | \hat{G}(k) | \sigma \rangle = \sum_{s,t=1}^N \sigma_s \sigma_t \exp(-k^2 s - t/6) \) is charge structure factor. The \( \phi^2 \) formula of Eq. 18 indicates an effective Flory-Huggins \( \chi \) parameter defined as

\[
\chi(\sigma, \sigma) \equiv \sqrt{\frac{8\pi}{3} \text{jSCD}(\sigma, \sigma)} / T^*.
\]

In a solution of two IDP components A and B, one similarly arrives at

\[
f_{el} = -\chi(\sigma^A, \sigma^A) \phi_A^2 - 2\chi(\sigma^A, \sigma^B) \phi_A \phi_B - \chi(\sigma^B, \sigma^B) \phi_B^2 + O(l_b^3),
\]

which is just the Eq. 27 in Ref. 8 that includes both homotypic and heterotypic interactions.

Combining with the fixed Flory-Huggins critical point at \( \chi_c = (\sqrt{N} + 1)^2 / (2N) \), Eq. 19 suggests the critical temperature of an \( N = 50 \) Das-Pappu-type sequence is given by

\[
T^*_c(\sigma) \approx 2.11 \times \text{jSCD}^{1/2}(\sigma, \sigma).
\]

Motivated by the strong correlation between jSCD and the products of two SCD shown in Fig. 2, we calculate the SCD and homotypic jSCD of the 30 Das-Pappu sequences plus 1,000 randomly generated partially-charged, overall neutral 50-residue KE sequences (see Supporting Information for sequence sampling method). The jSCD v.s. SCD correlation is shown in Fig. 3a, where a power-law regression suggests \( \text{jSCD}(\sigma, \sigma) = 0.293 \times \text{SCD}(\sigma)^{1.77} \) (\( R^2 = 0.983 \)). We also calculate the heterotypic jSCD(\( \sigma^A, \sigma^B \)) and SCD(\( \sigma^A \)SCD(\( \sigma^B \)) of all \( 30 \times 30 \) pairs of
Das-Pappu sequences and 1,000 pairs randomly selected from our 1,000 partially-charged sequences, and a similar power-law correlation $\text{jSCD}(\sigma^A, \sigma^B) = 0.313[\text{SCD}(\sigma^A)\text{SCD}(\sigma^B)]^{0.920}$ $(R^2=0.967)$ is shown in Fig. 3b. Intriguingly, in both Fig. 3a and 3b, those sequences with SCD$^2 \sim 1$, especially the sv1 that has the lowest SCD among Das-Pappu sequence, deviates the most from the linear relation, which may be attributed to its “unnatural” strictly alternating charge pattern.

To test the robustness of the correlation between jSCD and SCD, we apply a Coulomb potential with short-range cutoff, $U(r) = l_b[1-\exp(-r/b)]/r$, which we have used in our RPA theory for polyampholyte phase separation,18, 19, 22, 23 and derive a modified jSCD as

$$\text{jSCD}_{\text{cutoff}}(\sigma^A, \sigma^B) := \frac{1}{NA NB} \sqrt{\frac{3}{2\pi}} \int_0^\infty \frac{dk}{k^2(1 + k^2)^2} \sum_{s,t=1}^{NA} \sum_{l,m=1}^{NB} \sigma^A_s \sigma^A_t \sigma^B_l \sigma^B_m e^{-\frac{k(kb)^2}{2}[|s-t|+|l-m|]}.$$  \hspace{1cm} (22)

We calculate the homotypic and heterotypic jSCD$_{\text{cutoff}}$ and SCD of the same set sequences above and plot their correlation in Fig. 3c and 3d, which show $\text{jSCD}_{\text{cutoff}}(\sigma, \sigma) = 0.118 \times \text{SCD}(\sigma)^{2.007}$ $(R^2=0.997)$ and $\text{jSCD}_{\text{cutoff}}(\sigma^A, \sigma^B) = 0.109[\text{SCD}(\sigma^A)\text{SCD}(\sigma^B)]^{1.003}$ $(R^2=0.994)$, respectively. Combining with Eqs. 19 and 21, the jSCD$_{\text{cutoff}}$ correlations rationalize the linear relation $T^*_c \propto \text{SCD}$ in Ref. 19 and the SCD($\sigma^A$) $\times$ SCD($\sigma^B$) in Ref. 8. We note that, although it looks quite robust that a power-law correlation is between jSCD and SCD, the exact exponent seems to be sensitive to details in interaction potential.

**Simulation.** To compare with our approximate theory, we performed molecular dynamics simulations of a box containing two Das-Pappu sequences, one of which is sv28 and the other is selected among sv1, sv10, sv15, sv20, sv24, and sv25. Technical details of the simulations are described in Supporting Information. As the conformational ensemble of our IDPs are highly flexible and dynamics, multiple configurations, for example those shown in Fig. 4a, contribute to the binding reaction, and it is inappropriate to identify the bound state of the dimer complex by merely measuring the distance between the centers of mass. Thus, we collect the protein configurations in 1,000,000 simulation snapshots, and plot a
Figure 3: jSCD v.s. SCD plots of the 30 Das-Pappu sequences plus 1,000 partially-charged 50-residue KE sequences: (a) homotypic and (b) heterotypic interaction with simple Coulomb potential, and (c) homotypic and (d) heterotypic interaction with a Coulomb potential with short-range cutoff. Black lines indicate the result of power-law regression.

histogram of all intermolecular residue-residue distances. As shown in Fig. 4b, a pair of proteins with binding reaction results in a histogram that shows a bimodal distribution, one peak at roughly the length scale of amino acid and the other at the length scale of the simulation space, corresponding to the formation of A-B complex and the state that A and B are separated from each other, respectively. We also calculate the probability distribution of the center of mass distance of two non-interacting particles in a box as large as our simulation system and plot it in Fig. 4b, which clearly does not show the same bimodal distribution.

We then normalize the histogram in Fig. 4b and define the binding probability $\theta$ as the area of the small-distance peak. For comparison, we subtract the probability of free-particle collisions and define an effective binding probability $\tilde{\theta} = \theta - 4\pi r_{\text{cut}}^3/(3V)$, where $r_{\text{cut}}$ is the cutoff distance of the small-distance peak. In Fig. 4(c) we compare the trends of theoretical and simulated binding affinities, namely $K_D^{-1}$ and $\tilde{\theta}$, respectively, of the six Das-Pappu sequences. Most notably, binding affinity of the pair (sv25, sv28) is weaker
than that of (sv24, sv28), contradicting with the trend of \( \kappa \) that sv25 > sv24, whereas consistent with the opposite trend of their SCD’s. However, the theory and simulation disagree on the binding affinity of (sv20, sv28) and (sv15, sv28), as theory predicts the former stronger than the later whereas simulation shows opposite result. Besides the trend, we note that, if we directly substitute the simulated \( \theta \) into Eq. [13] we will obtain \( K_D \)'s that are about two orders of magnitude smaller than theoretical predictions. This difference can be attributed to the different models in analytical theory and explicit chain simulation, for example, in simulation, the residues have excluded volumes (Eq. S47), and the Coulomb interaction is screened (Eq. S45), whereas the theoretical \( K_D \)'s of Das-Pappu sequences are calculated without considering excluded volume and electrostatic screening. Similar difference between theory and simulation has also occurred in phase separation research on the Das-Pappu sequences.

Figure 4: Simulation at \( T^* = 0.35 \) (a) Three disordered complexes of two sv sequences. The “surface touch” structure is of sv24 and sv28, “entangled” structure is of sv25 and sv28, and the “extended” structure is of sv1 and sv28. The residues of the chain of sv28 in all three structures are colored as cyan (K) and orange (E), and its binding partners are colored by blue (K) and red (E). (b) Histogram of interchain residue-residue distance between sv24 and sv28 over 1,000,000 snapshots. The small-\( r \) peak in the green frame shows the bound state. Black curve indicates the probability distribution of the center of mass distance between two non-interacting particles in the same size box. (c) Comparison of the theoretical \( K_D \) and the simulated \( \tilde{\theta} = \theta - \theta_0 \), where \( \theta_0 = 4\pi \times 10^3/(3 \times 100^3) \) is the probability that two non-interacting particles get closer than \( 10a \) to each other. The negative \( \tilde{\theta} \) for (sv1,sv28) indicates that their net interaction is repulsive.

The discrepancy between the trends of theory and simulation on the sequences with intermediate charge segregation level imply influences from geometric and energetic factors
that are not taken into account in our Gaussian-chain model in Eq. S24. Back to the
general formula of $B_2$ in Eq. 9, the intramolecular residue-residue correlation function $\hat{P}$ is
influenced by intramolecular Coulomb interaction and thus can deviate from the Gaussian-
chain function given in Eq. 11. Empirically, the intramolecular interaction can be taken into
account in $\hat{P}$ by introducing the idea of effective Kuhn length, $b_{st}^{\text{eff}} \equiv x_{st}b$, which results in

$$\left[ \hat{P}^i \right]_{st} \rightarrow e^{-\frac{x_{st}^i}{\sigma_s}(kb)^2|s-t|}. \quad (23)$$

The $B_2$ for neutral polyampholytes is then modified from Eq. 16 and becomes

$$B_2^{\text{eff}} \big|_{k=0,0 \epsilon \{a^A,a^B\}} = 4\sqrt{\frac{\pi}{6}}l_b^2 b \sum_{s,t=1}^{N_A} \sum_{l,m=1}^{N_B} \sigma_s^A \sigma_t^A \sigma_l^B \sigma_m^B \sqrt{x_{st}^A |s-t| + x_{lm}^B |l-m|}. \quad (24)$$

The values of the $x_{st}^i$ may be calculated by a variational theory that minimizing the difference
between the correlation function calculated by the effective Gaussian chain and the complete
single-chain Hamiltonian, which, however, will result in $k$-dependent $x_{st}^i$ that cannot be
included in the empirical equation. Taking small-$k$ approximation, constant $x_{st}^i$ can be
obtained from a variational theory with respect to the squared of residue-residue distance,
$\langle R_{ij}^2 \rangle$, which has been proposed by Sawle and Ghosh and is applied to the Das-Pappu
sequences in Supporting Information. However, the IDP-IDP interaction is highly sensitive
to the structural details in small length scale, and thus the loss of large-$k$ information in the
variational theory for $\langle R_{ij}^2 \rangle$ can result in unpredictable ramification. In summary, further
improvement beyond the current framework of second-order cluster expansion of $B_2$ requires
much intense investigations via experiment or simulation, which are beyond the scope of this
manuscript.
Supporting Information

Derivation for $B_2$ representations

We start from the partition function representation in Eq. 4,

$$B_2 = V - \frac{Q_{AB}}{Q_A Q_B}.$$  

First, we define the single-molecule Hamiltonians $\mathcal{H}^A[R^A]$, $\mathcal{H}^B[R^B]$ for $A$, $B$, respectively. The conformational partition functions are then expressed as

$$Q_i = \frac{1}{V} \int \mathcal{D}[R^i] e^{-\mathcal{H}[R^i]} , \ i = A, B \quad (S1a)$$

$$Q_{AB} = \frac{1}{V} \int \mathcal{D}[R^A] \mathcal{D}[R^B] e^{-\mathcal{H}^A[R^A]-\mathcal{H}^B[R^B]-U^{AB}[R^A,R^B]}, \quad (S1b)$$

where $1/V$ in each formula cancels the degeneracy resulting from translational invariance. The ratio $Q_{AB}/(Q_A Q_B)$ can then be rewritten as

$$\frac{Q_{AB}}{Q_A Q_B} = V \int \frac{\mathcal{D}[R^A] \mathcal{D}[R^B] e^{-\mathcal{H}^A[R^A]-\mathcal{H}^B[R^B]-U^{AB}[R^A,R^B]}}{\mathcal{D}[R^A] e^{-\mathcal{H}^A[R^A]} \mathcal{D}[R^B] e^{-\mathcal{H}^B[R^B]} e^{-U^{AB}[R^A,R^B]}}$$

$$= V \int \mathcal{D}[R^A] \mathcal{D}[R^B] \mathcal{P}^A[R^A] \mathcal{P}^B[R^B] e^{-U^{AB}[R^A,R^B]} \quad (S2)$$

where

$$\mathcal{P}^i[R^i] \equiv \frac{e^{-\mathcal{H}^i[R^i]}}{\int \mathcal{D}[R^i] e^{-\mathcal{H}^i[R^i]}}, \ i = A, B. \quad (S3)$$

Notice that $\int \mathcal{D}[R^i] \mathcal{P}^i[R^i] = 1$. Substituting Eq. S2 for the $Q_{AB}/(Q_A Q_B)$ in Eq. 4 results in the probability representation of $B_2$ in Eq. 3

$$B_2 = V \int \mathcal{D}[R^A] \mathcal{D}[R^B] \mathcal{P}^A[R^A] \mathcal{P}^B[R^B] \left(1 - e^{-U^{AB}[R^A,R^B]} \right).$$
Next, we decouple translational invariance from the integral of all residue coordinates. We perform a change of coordinates

\[ \{ R^i_1, R^i_2, \ldots, R^i_N \} \rightarrow \{ R^i_1, z_1^i, z_2^i, \ldots, z_{N-1}^i \}, \quad z^i_s = R^i_{s+1} - R^i_s, \]  

(S4)

which allows us to present all intramolecular residue-residue distances by

\[ R^i_s - R^i_t = \sum_{\tau=t}^{s-1} z^i_\tau \quad (s > t). \]  

(S5)

As single-molecule Hamiltonian in a homogeneous solution should be a function of the relative positions of the residues and be irrelevant to the location of the molecule, it should simply be a function of \( z \)'s and independent of \( R^i_1 \); namely, the partition functions \( Q_A, Q_B \) can be rewritten as

\[ Q_i = \frac{1}{V} \int dR^i_1 \mathcal{D}[z^i] e^{-\mathcal{H}[z^i]} \]

\[ = \int \mathcal{D}[z^i] e^{-\mathcal{H}[z^i]} \quad i = A, B, \]  

(S6)

where \( \mathcal{D}[z^i] \equiv \prod_{s=1}^{N-1} dz^i_s \) and \( \int dR^i_1/V = 1 \). For the distances between an intermolecular residue pair, we have

\[ R^A_s - R^B_t = \sum_{\tau=1}^{s} z^A_\tau + \sum_{\mu=1}^{t} z^B_\mu + R^{AB}_{11}, \]  

(S7)

where \( R^{AB}_{11} \equiv R^A_1 - R^B_1 \). Thus, the intermolecular interaction \( \mathcal{U}^{AB} \) is a function of \( z^A, z^B, \) and \( R^{AB}_{11} \). The partition function of A-B complex is then given by

\[ Q_{AB} = \frac{1}{V} \int dR^A_1 dR^B_1 \mathcal{D}[z^A] \mathcal{D}[z^B] e^{-\mathcal{H}^{A}[z^A] - \mathcal{H}^{B}[z^B] - \mathcal{U}^{AB}[z^A, z^B, R^{AB}_{11}]} \]

\[ = \int dR^{AB}_{11} \mathcal{D}[z^A] \mathcal{D}[z^B] e^{-\mathcal{H}^{A}[z^A] - \mathcal{H}^{B}[z^B] - \mathcal{U}^{AB}[z^A, z^B, R^{AB}_{11}]}, \]  

(S8)

where the second equality is obtained by a change of variable \( \{ R^A_1, R^B_1 \} \rightarrow \{ R^{AB}_{11}, R^{AB}_1 \} \) and the spatial integral \( \int dR^B_1/V = 1 \). We note that, in terms of \( \{ z^i \} \), the probability distributions
of polymer conformations are given by

$$P^i[\mathbf{s}^i] = \frac{e^{-H^i[\mathbf{s}^i]}}{\mathcal{D}[\mathbf{s}^i]e^{-H^i[\mathbf{s}^i]}}, \quad i = A, B. \quad (S9)$$

Finally, we perform one more change of variable that linearly shifts $\mathbf{R}_{11}^{AB}$ to center of mass distance $\mathbf{R}_{CM}^{AB}$ by

$$\mathbf{R}_{CM}^{AB} = \sum_{s=1}^{N_A} \frac{M_s^A \mathbf{R}_s^A}{\sum_{s=1}^{N_A} M_s^A} - \frac{\sum_{t=1}^{N_B} M_t^B \mathbf{R}_t^B}{\sum_{t=1}^{N_B} M_t^B} = \mathbf{R}_{11}^{AB} + \frac{\sum_{s=1}^{N_A} M_s^A \sum_{t=1}^{s-1} \mathbf{R}_t^A}{\sum_{s=1}^{N_A} M_s^A} - \frac{\sum_{t=1}^{N_B} M_t^B \sum_{t=1}^{s-1} \mathbf{R}_t^B}{\sum_{t=1}^{N_B} M_t^B}, \quad (S10)$$

where $M_s^i$ is the mass of the $s$th residue in protein $i$. Eq. (S10) allows us to shift the integral by $d\mathbf{R}_{11}^{AB} \to d\mathbf{R}_{CM}^{AB}$ to obtain

$$\frac{Q_{AB}}{Q_A Q_B} = \int d\mathbf{R}_{CM}^{AB} \mathcal{D}[\mathbf{s}^A] \mathcal{D}[\mathbf{s}^B] \mathcal{D}[\mathbf{z}^A] \mathcal{D}[\mathbf{z}^B] e^{-U_{AB}[\mathbf{z}^A, \mathbf{z}^B, \mathbf{R}_{CM}^{AB}]} \equiv \int d\mathbf{R}_{CM}^{AB} \left( e^{-U_{AB}[\mathbf{R}_{CM}^{AB}; \mathbf{z}^A, \mathbf{z}^B]} \right)_{A,B}, \quad (S11)$$

which can then be substituted for the $Q_{AB}/(Q_A Q_B)$ in Eq. (4) and result in the classical center-of-mass representation of $B_2$ in Eq. (2) viz.

$$B_2 = \int d\mathbf{R}_{CM}^{AB} \left( 1 - e^{-\beta U_{AB}[\mathbf{R}_{CM}^{AB}; \mathbf{R}^A, \mathbf{R}^B]} \right)_{A,B}.$$ 

### Derivation for $B_2$ in terms of Mayer f-functions

We substitute the cluster expansion in Eq. (S12)

$$e^{-U_{AB}} - 1 \approx \sum_{s=1}^{N_A} \sum_{t=1}^{N_B} f_{st} + \sum_{s1=1}^{N_A} \sum_{t1=1}^{N_B} \sum_{l1=1}^{N_P} \sum_{m1=1}^{N_P} f_{st} f_{tm} - \sum_{s=1}^{N_A} \sum_{t=1}^{N_B} f_{st}^2 \quad (S12)$$
into the $B_2$ formula in Eq. 3

$$B_2 = -V \int \mathcal{D}[\mathbf{R}^A] \mathcal{D}[\mathbf{R}^B] \mathcal{P}^A[\mathbf{R}^A] \mathcal{P}^B[\mathbf{R}^B] \left( e^{-\mathcal{U}^{AB}[\mathbf{R}^A,\mathbf{R}^B]} - 1 \right)$$

and individually integrate each term in the expansion. Before evaluating the integrations, we perform change of variables $\{\mathbf{R}^i\} \rightarrow \{\mathbf{r}^i\} \cup \{\mathbf{R}^i_1\}$ and substitute the $\mathcal{P}^i[\mathbf{r}^i]$ in Eq. S9 for $\mathcal{P}^i[\mathbf{R}^i]$ to rewrite Eq. 3 as

$$B_2 = -\int d\mathbf{R}_{11}^{AB} \mathcal{D}[\mathbf{z}^A] \mathcal{D}[\mathbf{z}^B] \mathcal{P}^A[\mathbf{z}^A] \mathcal{P}^B[\mathbf{z}^B] \left( e^{-\mathcal{U}^{AB}[\mathbf{z}^A,\mathbf{z}^B,\mathbf{R}_{11}^{AB}]} - 1 \right) \quad (S13)$$

in which $\mathbf{R}_{11}^{AB} = \mathbf{R}^A_1 - \mathbf{R}^B_1$. We note the intermolecular interaction $\mathcal{U}^{AB}[\mathbf{R}^A,\mathbf{R}^B]$ has the form of $\mathcal{U}^{AB}[(\mathbf{R}_s^A - \mathbf{R}_t^B)]$ so that the transformation $\mathcal{U}^{AB}[\mathbf{R}^A,\mathbf{R}^B] \rightarrow \mathcal{U}^{AB}[\mathbf{z}^A,\mathbf{z}^B,\mathbf{R}_{11}^{AB}]$ is legitimate. In addition, we recall the matrix of the Fourier transformation of Mayer f-function $f_{st}$ defined in Eq. 8

$$f_{st}(\mathbf{r}) = \int \frac{d^3k}{(2\pi)^3} \left[ \hat{f}(\mathbf{k}) \right]_{st} e^{i\mathbf{k} \cdot \mathbf{r}}$$

The cluster expansion formula of $B_2$ is then given by substituting Eq. S12 into Eq. S13 which yields

$$B_2 \approx -\sum_{s=1}^{N_A} \sum_{t=1}^{N_B} \int d\mathbf{R}_{11}^{AB} \mathcal{D}[\mathbf{z}^A] \mathcal{D}[\mathbf{z}^B] \mathcal{P}^A[\mathbf{z}^A] \mathcal{P}^B[\mathbf{z}^B] f_{st}(\mathbf{R}_{st}^{AB})$$

$$+ \sum_{s=1}^{N_A} \sum_{t=1}^{N_B} \int d\mathbf{R}_{11}^{AB} \mathcal{D}[\mathbf{z}^A] \mathcal{D}[\mathbf{z}^B] \mathcal{P}^A[\mathbf{z}^A] \mathcal{P}^B[\mathbf{z}^B] f_{st}^2(\mathbf{R}_{st}^{AB})$$

$$- \sum_{s=1}^{N_A} \sum_{t=1}^{N_B} \int d\mathbf{R}_{11}^{AB} \mathcal{D}[\mathbf{z}^A] \mathcal{D}[\mathbf{z}^B] \mathcal{P}^A[\mathbf{z}^A] \mathcal{P}^B[\mathbf{z}^B] f_{sl}(\mathbf{R}_{sl}^{AB}) f_{tm}(\mathbf{R}_{tm}^{AB})$$

$$\equiv B_2^{**} + B_2^{**^2} + B_2^{****}.$$  

We then evaluate $B_2^{**}$, $B_2^{**^2}$, and $B_2^{****}$ separately. Each of the integrands $f_{st}$ in $B_2^{**}$ is
evaluated by

\[
\int dR^{AB}_{11} \mathcal{D}[\tau^A] \mathcal{D}[\tau^B] \mathcal{P}^A[\tau^A] \mathcal{P}^B[\tau^B] f_{st}(R^{AB}_{st}) = \int dR^{AB}_{st} \mathcal{D}[\tau^A] \mathcal{D}[\tau^B] \mathcal{P}^A[\tau^A] \mathcal{P}^B[\tau^B] \int \frac{d^3k}{(2\pi)^3} \left[ \hat{f}(k) \right]_{st} e^{ikR^{AB}_{st}}
\]

\[
= \int \frac{d^3k}{(2\pi)^3} \left[ \hat{f}(k) \right]_{st} \int dR^{AB}_{st} e^{ikR^{AB}_{st}}
\]

\[= \int \frac{d^3k}{(2\pi)^3} \left[ \hat{f}(k) \right]_{st} (2\pi)^3 \delta^3(k).
\]

in which we shift the integral \(dR^{AB}_{11} \rightarrow dR^{AB}_{st}\) for the residue pair \((s^A, t^B)\) with no influences on \(\mathcal{P}[t^t]\). Thus, the \(B^{+2}_{2^{-}}\) is given by

\[B^{+2}_{2^{-}} = - \sum_{s=1}^{N} \sum_{t=1}^{N} \left[ \hat{f}(0) \right]_{st}
\]

Each of the integrands in \(B^{+2}_{2^{-}}\) can be integrated by the same change of variable, which yields

\[
\int dR^{AB}_{11} \mathcal{D}[\tau^A] \mathcal{D}[\tau^B] \mathcal{P}^A[\tau^A] \mathcal{P}^B[\tau^B] f_{st}^2(R^{AB}_{st}) = \int dR^{AB}_{st} \mathcal{D}[\tau^A] \mathcal{D}[\tau^B] \mathcal{P}^A[\tau^A] \mathcal{P}^B[\tau^B] \int \frac{d^3k}{(2\pi)^3} \left[ \hat{f}(k) \right]_{st} e^{ikR^{AB}_{st}} \int \frac{d^3k'}{(2\pi)^3} \left[ \hat{f}(k') \right]_{st} e^{i(k + k')R^{AB}_{st}}
\]

\[
= \int \frac{d^3k}{(2\pi)^3} \frac{d^3k'}{(2\pi)^3} \left[ \hat{f}(k) \right]_{st} \left[ \hat{f}(k') \right]_{st} \int dR^{AB}_{st} e^{i(k + k')R^{AB}_{st}}
\]

\[
= \int \frac{d^3k}{(2\pi)^3} \frac{d^3k'}{(2\pi)^3} \left[ \hat{f}(k) \right]_{st} \left[ \hat{f}(k') \right]_{st} (2\pi)^3 \delta^3(k + k')
\]

\[= \int \frac{d^3k}{(2\pi)^3} \left[ \hat{f}(k) \right]_{st} \left[ \hat{f}(-k) \right]_{st},
\]

and the summation is given by

\[B^{+2}_{2^{-}} = \sum_{s=1}^{N} \sum_{t=1}^{N} \left[ \hat{f}(k) \right]_{st} \left[ \hat{f}(-k) \right]_{st} = \text{Tr} \left[ \hat{f}(k) \hat{f}^T(-k) \right],
\]

where the “T” superscript denotes transpose of the matrix.

The integrands in \(B^{+2}_{2^{-}}\), which includes two residue sets \((s^A, t^B)\) and \((t^A, m^B)\) such that
\( s > t, l > m \), can also be integrated by the same change of variable by noticing

\[
R_{sl}^{AB} = R_{11}^{AB} + \sum_{\tau=1}^{s-1} Z_\tau^{A} - \sum_{\mu=1}^{l-1} Z_\mu^{B} \\
= R_{11}^{AB} + \left( \sum_{\tau=1}^{s-1} + \sum_{\tau=1}^{l-1} \right) Z_\tau^{A} - \left( \sum_{\mu=1}^{s-1} + \sum_{\mu=1}^{m} \right) Z_\mu^{B} \\
= R_{tm}^{AB} + \sum_{\tau=t}^{s-1} Z_\tau^{A} - \sum_{\mu=m}^{l-1} Z_\mu^{B}. 
\] (S19)

Combining with \( dR_{11}^{AB} \rightarrow dR_{tm}^{AB} \), we obtain

\[
\int dR_{11}^{AB} \mathcal{D}[Z^{A}] \mathcal{D}[Z^{B}] \mathcal{P}^{A}[Z^{A}] \mathcal{P}^{B}[Z^{B}] f_{sl}(R_{sl}^{AB}) f_{tm}(R_{tm}^{AB}) \\
= \int dR_{tm}^{AB} \mathcal{D}[Z^{A}] \mathcal{D}[Z^{B}] \mathcal{P}^{A}[Z^{A}] \mathcal{P}^{B}[Z^{B}] \int \frac{d^3k}{(2\pi)^3} \left[ \hat{f}(\mathbf{k}) \right]_{sl} \int \frac{d^3k'}{(2\pi)^3} \left[ \hat{f}(\mathbf{k}') \right]_{tm} e^{ikR_{sl}^{AB}} \int e^{ik'R_{tm}^{AB}} \\
= \frac{d^3k}{(2\pi)^3} \frac{d^3k'}{(2\pi)^3} \left[ \hat{f}(\mathbf{k}) \right]_{sl} \left[ \hat{f}(\mathbf{k}') \right]_{tm} (2\pi)^3 \delta^3(\mathbf{k} + \mathbf{k}') \left( e^{ik(R_{A}^{A} - R_{A}^{A})} \right)_A \left( e^{-ik(R_{B}^{B} - R_{B}^{B})} \right)_B \\
\equiv \int \frac{d^3k}{(2\pi)^3} \left[ \hat{f}(\mathbf{k}) \right]_{sl} \left[ \hat{f}(\mathbf{k}) \right]_{tm} \left[ \hat{P}^{A}(\mathbf{k}) \right]_{sl} \left[ \hat{P}^{B}(-\mathbf{k}) \right]_{tm}. 
\] (S20)

where \( \left[ \hat{P}^{i}(\mathbf{k}) \right]_{sl} \) is the Fourier transformation of the intramolecular residue-residue correlation function defined in Eq. [7] viz.

\[
\left[ \hat{P}^{i}(\mathbf{k}) \right]_{sl} = \int \mathcal{D}[Z^{A}] \mathcal{P}^{A}[Z^{A}] e^{i\mathbf{k}\Sigma_{\tau=1}^{s-1} Z_\tau^{A}} \\
= \int \mathcal{D}[R^{i}] \mathcal{P}[R^{i}] e^{i\mathbf{k}(R_{A}^{i} - R_{i}^{i})} , \quad i = A, B. 
\] (S21)
The summation of the integrals of \( f_{st} f_{tm} \) then yield \( B_2 \) as

\[
B_2 = - \sum_{s \geq l \geq m = 1}^{N_A} \sum_{N_R} f_{st} f_{tm} \\
= - \frac{1}{2} \sum_{s,i = 1}^{N_A} \sum_{l,m = 1}^{N_R} f_{st} f_{tm} - \frac{1}{2} \sum_{s = 1}^{N_A} \sum_{l,m = 1}^{N_R} f_{sl}^2 \\
= - \frac{1}{2} \int \frac{d^3 k}{(2\pi)^3} \left\{ \text{Tr} \left[ \hat{f}(k) \hat{P}^A(-k) \hat{f}^T(-k) \hat{P}^A(-k) \right] + \text{Tr} \left[ \hat{f}(k) \hat{f}^T(-k) \right] \right\}.
\]

Notice that we have applied \( \hat{P}^i(k)_{st} = [\hat{P}^i(-k)]_{ts} \). Combining Eqs. S16, S18, and S22, we obtain \( B_2 \) up to \( O(f^2) \) as

\[
B_2 \approx B_2^* + B_2^{****} + B_2^{-2} \\
= - \sum_{s = 1}^{N_A} \sum_{l,m = 1}^{N_R} [\hat{f}(0)]_{st} - \frac{1}{2} \int \frac{d^3 k}{(2\pi)^3} \text{Tr} \left[ \hat{f}(k) \hat{P}^B(-k) \hat{f}^T(-k) \hat{P}^A(-k) - \hat{f}(k) \hat{f}^T(-k) \right],
\]

which is just Eq. [9]

**Derivation of \( B_2 \) at salt-free limit**

We start from the \( B_2 \) in Eq. S24,

\[
B_2 = \frac{4\pi l_b}{k^2} q^A q^B - 4l_b^2 \int \frac{d k d^2 \theta}{(k^2 + \kappa^2)^2} \sum_{s=1}^{N_A} \sum_{l,m=1}^{N_R} \sigma_s^A \sigma_t^A \sigma_i^B \sigma_m^B e^{-\frac{1}{2}(k-b)^2 \sigma^2} |s-t| \sigma^2 |l-m|
\]

\[
= F_1 + F_2,
\]

in which \( F_1 \) is the interaction of net charges in \( O(l_b) \), and \( F_2 \) accounts for sequence specificity.

We rewrite \( F_2 \) as

\[
F_2 = - 4l_b^2 \int_0^\infty \frac{d k d^2 \theta}{(k^2 + \kappa^2)^2} \sum_{s=1}^{N_A} \sum_{l,m=1}^{N_R} \sigma_s^A \sigma_t^A \sigma_i^B \sigma_m^B e^{-\frac{1}{2}(k-b)^2 \sigma^2} |s-t| \sigma^2 |l-m|
\]

\[
= \frac{4l_b^2}{\sqrt{6}} \sum_{s,t=1}^{N_A} \sum_{l,m=1}^{N_R} \sigma_s^A \sigma_t^A \sigma_i^B \sigma_m^B I(|s-t| \sigma^2 |l-m|),
\]

22
where $I$ is the $k$-integral

$$I_X = \int_0^\infty \frac{d\bar{k} \bar{k}^2}{(k^2 + \bar{k}^2)^2} e^{-X\bar{k}^2}, \quad (S26)$$

where $\bar{k} \equiv kb/\sqrt{6}$, $\bar{\kappa} \equiv \kappa b/\sqrt{6}$, and $X \equiv |s - t| + |l - m|$. We then calculate the integral as

$$I_X = -\frac{1}{2} \left[ \bar{k} e^{-X\bar{k}^2} \right]_0^\infty + \frac{1}{2} \int_0^\infty d\bar{k} \frac{1 - 2X\bar{k}^2}{k^2 + \bar{k}^2} e^{-X\bar{k}^2}$$

$$= \left( \frac{1}{2} + X\bar{\kappa}^2 \right) \int_0^\infty d\bar{k} \frac{e^{-X\bar{k}^2}}{k^2 + \bar{k}^2} - X \int_0^\infty d\bar{k} e^{-X\bar{k}^2}$$

$$= \left( \frac{\pi}{4\bar{\kappa}} + \frac{\pi X\bar{\kappa}}{2} \right) e^{X\bar{\kappa}^2} \text{erfc} \left( \bar{\kappa}\sqrt{X} \right) - \frac{\sqrt{\pi X}}{2}, \quad (S27)$$

where $\text{erfc}(x)$ is the complementary error function. We consider $\bar{\kappa} \ll 1$ and substitute the Taylor series

$$e^x \text{erfc}(x) = 1 - \frac{2x}{\sqrt{\pi}} + x^2 + O(x^3), \quad (S28)$$

which obtains

$$I_X = \frac{\pi}{4\bar{\kappa}} - \sqrt{\pi X} + \frac{3}{4} \pi \bar{\kappa} X + O(\bar{\kappa}^2). \quad (S29)$$

The $F_2$ in Eq. (S24) then becomes

$$F_2 = -\frac{4l_b^2}{\sqrt{6}} \left[ \frac{\pi \sqrt{6}}{4\kappa b} (q^A)^2 (q^B)^2 - \sqrt{\pi} \sum_{s,t=1}^{N_A} \sum_{l,m=1}^{N_B} \sigma_s^A \sigma_t^A \sigma_l^B \sigma_m^B \sqrt{|s - t| + |l - m|} \right] + O(\bar{\kappa}) \quad (S30)$$

where the first term is the net charge interaction in $O(l_b^2)$, and the second term accounts for the lowest-order sequence specificity. We then define the measure “joint sequence charge decoration” as

$$\text{jSCD}(\sigma^A, \sigma^B) \equiv -\frac{1}{2N_A N_B} \sum_{s,t=1}^{N_A} \sum_{l,m=1}^{N_B} \sigma_s^A \sigma_t^A \sigma_l^B \sigma_m^B \sqrt{|s - t| + |l - m|}. \quad (S31)$$

In the case of $q^A = 0$ and/or $q^B = 0$, $F_1$ and the first term of $F_2$ in Eq. (S30) both become
zero, and thus $B_2$ is proportional to jSCD as

$$B_{2|\kappa \to 0, \epsilon (q^A, q^B)} = -8\sqrt{\frac{\pi}{6}} b^2 N_A N_B \times jSCD(\sigma^A, \sigma^B). \quad (S32)$$

For the case that both two proteins are not neutral, namely $q^A \neq 0$ and $q^B \neq 0$, we note that the $q^A q^B$ terms in Eqs. $[S24]$ and $[S30]$ are part of the Taylor series of the Mayer f-function of the mean-field net charge interaction, as can be shown from

$$B^{MF}_2 = \int d^3r \left( 1 - e^{-l_b q^A q^B e^{-\kappa r}/r} \right) \quad (S33)$$

$$= 4\pi \int_0^{\infty} dr r^2 \left( l_b^2 \frac{q^A q^B e^{-\kappa r}}{r^2} \right) + O(l_b^3). \quad (S33)$$

Thus, these $q^A q^B$ terms include no sequence specificity, and jSCD is still the lowest-order term taking into account sequence specificity for non-neutral IDPs. It is worth noting that the divergence of these net charge terms in the limit of $\kappa \to 0$ is a result of the ultraviolet divergence in calculating the electric energy of point charges, which can be resolved by introducing short-range cutoff into Coulomb potential.\[16,23]\]

**Generation of random sequences**

For each integer $i$ between 1 and 25, 40 random neutral sequences containing $i$ positively charged residues (carry +1 charge), $i$ negatively charged residues (carry -1 charges), and $50 - 2i$ neutral residues were generated; this was done by randomly permuting the array $(+1, \ldots, +1, 0, \ldots, 0, -1, \ldots, -1)$ with +1 and -1 each repeated $i$ times and 0 repeated $50 - 2i$ times. 1,000 random pairs of the sequences in this pool of 1,000 polymers were then selected for investigation of the correlation between heterotypic jSCD and SCD.
Modification from intrachain interaction

The influence from intrachain interactions on IDP conformation can be taken into account by treating the IDP as a Gaussian chain with an effective Kuhn length, $l_1$. Here, we apply the variational theory in Ref. [12] to derive such a Gaussian chain.

The correlation function of a Gaussian chain given in Eq. (11) is a result of a single-molecule Hamiltonian including only elastic chain connectivity, namely

$$\mathcal{H}^i[R^i] = \frac{3}{2b^2} \sum_{s=1}^{N^i-1} (R^i_{s+1} - R^i_s)^2.$$ (S34)

Now we consider an intrachain interaction potential $U^i[R^i]$ including electrostatic interaction and excluded volumes, namely

$$U^i[R^i] = \sum_{s>t=1}^{N^i} \left[ l_0 \sigma_t \sigma_i e^{-\kappa |R^i_s - R^i_t|} + w^i \delta^3(R^i_s - R^i_t) \right],$$ (S35)

where $w^i$ is the two-body excluded volume for protein $i$; for Das-Pappu sequences $w^i$ can be obtained from Table 1 in Ref. [12]. A new Hamiltonian is then given by

$$\mathcal{H}^i_{\text{new}}[R^i] = \mathcal{H}^i[R^i] + U^i[R^i].$$ (S36)

We then assume that this new Hamiltonian can be approximated to a new Gaussian chain with an effective Kuhn length $b_1 = bx$, namely

$$\mathcal{H}^i_{\text{new}}[R^i] \approx \mathcal{T}^i_x[R^i] = \frac{3}{2bx^2} \sum_{s=1}^{N^i-1} (R^i_{s+1} - R^i_s)^2,$$ (S37)

where $x$ is determined by the variational approach in Ref. [12]. Here we briefly summarize the concept and result, and refer the readers to the original paper for details of the method. We separate the new Hamiltonian to a “principal” component and the remaining “pertur-
ations”, namely

\[ \mathcal{H}_{\text{new}}^i[\mathbf{R}^i] = T_x^i[\mathbf{R}^i] + \Delta \mathcal{H}^i_x[\mathbf{R}^i], \tag{S38} \]

where \( \Delta \mathcal{H}^i_x[\mathbf{R}^i] \) is given by

\[ \Delta \mathcal{H}^i_x[\mathbf{R}^i] = \frac{3}{2b^2} \left( 1 - \frac{1}{x} \right) \sum_{s=1}^{N_i-1} (R_{s+1}^i - R_s^i)^2 + \sum_{s>t}^{N_i} \left[ l_{is} \sigma_s^i \sigma_t^i e^{-x|R_s^i - R_t^i|} + w^i \delta^3(R_s^i - R_t^i) \right]. \tag{S39} \]

In applying the variational approach, for a physical observable \( A \), its thermodynamic average with respect to the total Hamiltonian can be expanded with respect to the perturbation Hamiltonian as

\[ \langle A \rangle_{\text{all}} = \langle A \rangle_x + \langle A \rangle_x \langle \Delta \mathcal{H}^i \rangle_x - \langle A \Delta \mathcal{H}^i \rangle_x + O\left( (\Delta \mathcal{H}^i)^2 \right), \tag{S40} \]

where the averages \( \langle \cdots \rangle \) are defined as

\[ \langle A \rangle_{\text{all}} = \frac{\int \mathcal{D}[\mathbf{R}] A[\mathbf{R}] e^{\mathcal{H}_{\text{new}}^i[\mathbf{R}^i]}}{\int \mathcal{D}[\mathbf{R}] e^{\mathcal{H}_{\text{new}}^i[\mathbf{R}^i]}} \tag{S41a} \]

\[ \langle A \rangle_x = \frac{\int \mathcal{D}[\mathbf{R}] A[\mathbf{R}] e^{T_x^i[\mathbf{R}^i]}}{\int \mathcal{D}[\mathbf{R}] e^{T_x^i[\mathbf{R}^i]}}. \tag{S41b} \]

To minimize the difference between the thermodynamic averages obtained from \( \mathcal{H}_{\text{new}}^i \) and \( T_x^i \), we have to find the \( x \) that eliminate the \( O((\Delta \mathcal{H}^i)) \) components in Eq. \( \text{S40} \). We pick up the square of intrachain residue-residue distance \( |R_s^i - R_t^i|^2 \) as the characteristic observable for finding the best-fit \( x = x_{st}^i \) for the pair of residues \( (s^i, t^i) \), and derive the \( x_{st}^i \) by solving

\[ \langle |R_s^i - R_t^i|^2 \rangle_{x_{st}^i} \langle \Delta \mathcal{H}^i \rangle_{x_{st}^i} = \langle |R_s^i - R_t^i|^2 \Delta \mathcal{H}^i \rangle_{x_{st}^i}. \tag{S42} \]

The exact formulas for solving Eq. \( \text{S42} \) are given in Eqs. 6–10 in Ref. 12. With the solved
for $x_{st}$, we rewrite the equation of $B_2$ as

$$B_2 = \frac{4\pi l_b}{k^2} q_A q_B - 4l_b^2 \int \frac{dk k^2}{(k^2 + \kappa^2)^2} \sum_{s,t=1}^{N_A} \sum_{l,m=1}^{N_B} \sigma_A^s \sigma_t^A \sigma_l^B \sigma_m^B e^{-\frac{1}{k}(kb)^2 [x_{st}|s-t| + x_{lm}|l-m|]}, \quad (S43)$$

which in salt-free solution of charge neutral polymers becomes

$$B_{2e}^{\text{eff}} \xrightarrow{\kappa \to 0} 0 = 4\sqrt{\frac{\pi}{6}} l_b^2 b \sum_{s,t=1}^{N_A} \sum_{l,m=1}^{N_B} \sigma_A^s \sigma_t^A \sigma_l^B \sigma_m^B \sqrt{x_{st}|s-t| + x_{lm}|l-m|}, \quad (S44)$$

which is just Eq. S44.

A heatmap of the $K_D$ calculated by the $B_2$ in Eq. S44 is plotted in Fig. S2a. Unlike the results of bare-Gaussian approximation, the theory of effective Kuhn length predicts that some sequences will not bind, as can be seen from the white regions in the figure. As well, instead of binding propensity being monotonic with charge segregation as predicted by the base theory, some sequence pairs diverge from the trend: highly charge segregated sequences seem to avoid interactions with sequences with only a medium charge segregation. In Fig. S2b we compare the difference between the binding propensities predicted by the base theory and this theory of effective Kuhn length, where one can see that taking into account intramolecular interactions imparts a penalty to the binding of highly charge-segregated polymers. The $K_D$ of the two theories and the binding probability obtained from simulation (method described in the next section) are summarized in Fig. S3.

**Simulation Methods**

Molecular dynamics simulations have been performed with 6 separate pairs of fully-charged polyampholytes of length $N=50$. Each of these sequences contains 25 lysines (K) and 25 glutamic acids (E). These sequences are popularly referred as “sv” sequences and have been studied in great details using a variety of simulation techniques\(^{11,17,18,24}\) and also by random phase approximation (RPA) theory.\(^{19}\) Initially the 30 sequences were introduced by Das.
and Pappu\textsuperscript{11} covering a wide range of charge patterns as characterised by two well-known charge pattern parameters, and SCD.\textsuperscript{11,12} Here we select 6 pairs, (sv1, sv28), (sv10, sv28), (sv15, sv28), (sv20, sv28), (sv24, sv25), and (sv25, sv28), spanning almost the full range of sv sequences.

We adopt coarse-grained model for the sequences which has been used before for studying the liquid-liquid phase separation of IDP polymer chains.\textsuperscript{17,24} Each amino acid is represented as a single bead of lysine with +1 charge or glutamic acid with -1 charge. We further assume that each amino acid has same diameter and same mass. The potential energy function used for the study consists of bonded, non-bonded Lennard-Jones (LJ) and screened electrostatic interaction. For any two amino acid residues \((i, s)\) and \((j, t)\), namely the \(s\)th residue of the \(i\)th protein and the \(t\)th residue of the \(j\)th protein that carry charges \(\sigma_i^s\) and \(\chi_j^t\) respectively, the residue-residue electrostatic interaction is given by

\[
U_{el} = \frac{\sigma_i^s \sigma_j^t e^2}{4\pi\varepsilon_0\varepsilon_r r_{is,jt}^2} \exp\left(-\kappa r_{is,jt}\right),
\]

where \(\kappa\) is the Debye screening wave number determined by salt concentration, \(e\) is elementary electric charge, \(\varepsilon_0\) is vacuum permittivity, \(\varepsilon_r\) is relative permittivity of the solution, and \(r_{is,jt}\) is the distance between amino acids \((i, s), (j, t)\). The non-bonded LJ interaction is given by

\[
U_{LJ} = 4\varepsilon_{LJ} \left[ \left( \frac{a}{r_{is,jt}} \right)^{12} - \left( \frac{a}{r_{is,jt}} \right)^{6} \right]
\]

where \(\varepsilon_{LJ}\) and \(a\) are the depth and range of the LJ potential, respectively. The non-bonded LJ interaction is truncated at the minimum of the LJ potential, i.e. \(r_{cut} = a^{1/6}a\) and then shifted to satisfy \(U_{LJ}(r_{cut}) = 0\), resulting a pure repulsive potential as

\[
U_{\text{cut}, LJ}^{\text{LJ}} = \begin{cases} 
4\varepsilon_{LJ} \left[ \left( \frac{a}{r_{is,jt}} \right)^{12} - \left( \frac{a}{r_{is,jt}} \right)^{6} + 1 \right], & r_{is,jt} \leq 2^{1/6}a \\
0, & r_{is,jt} > 2^{1/6}a
\end{cases}
\]

As we have learned from previous study, the interaction among the sv sequences can be
strongly influenced by the background non-electrostatic attraction. To make the system dominated by electrostatic interaction so as to be compared with our pure-Coulombic theory, we set $\varepsilon = e^2/(4\pi \varepsilon_0 \epsilon r_a)$ to be the unit of energy in our simulation, and apply $\varepsilon_{LJ} = \varepsilon/48$ to make the short-range repulsion much weaker than electrostatic interaction. The bonded interaction among the amino acid monomers is modelled by the harmonic potential

$$U_{\text{bond}} = \frac{K_{\text{bond}}}{2} (r_{is,is+1} - a)^2$$  \hspace{1cm} (S48)

where $K_{\text{bond}} = 75000\varepsilon/a^2$ is the bond force constant we have used in our previous study, which is also consistent with the TraPPE force field.

All the simulations are performed with the GPU version of HOOMD-blue simulation package \cite {29,30} at 10 different temperatures between 0.05$T^*$ and 0.5$T^*$ with an interval of 0.05$T^*$ using a timestep of 0.001$\tau$, where $T^* \equiv k_B T/\varepsilon = l_b/a$ is the reduced temperature and $\tau = \sqrt{ma^2/\varepsilon}$ is the reduced time defined by amino acid mass $m$. We initialize the simulation by randomly place the selected pair of sv sequences in a large cubic box of dimension $100a \times 100a \times 100a$ and then run $500\tau$ for energy minimization. The electrostatic interaction among the residues is treated with PPPM method using a real space cutoff distance of 15$a$ and a fixed Debye screening length of 3$a$. After energy minimization the system is heated to its desired temperature in $2500\tau$ via a Langevin dynamics with a weak friction coefficient of $0.1m/\tau$. The motion of the particles are integrated using velocity-Verlet scheme with periodic boundary condition. After achieving desired temperature, we conduct a production run of $500,000\tau$ and save the trajectories per 0.5$\tau$ for further analysis.

**Simulation result analysis**

As mentioned, we conduct simulations for each pair of sequences at 10 different temperatures between 0.05$T^*$ and 0.5$T^*$. For each simulation, we look into the trajectory of the distance between centers of mass of the two sequences, and notice that, for simulations conducted at
$T^* < 0.35$, their trajectories only switch between bound state (small center-of-mass distance) and unbound state (large center-of-mass distance) less than 5 times among the 1,000,000 fetched snapshots, implying that the systems are often, due to the large friction at low temperature, stuck at local traps and do not perform complete sampling through all polymer conformations. Thus, we discard the low-temperature simulation results, and analyze only simulations at $T^* \geq 0.35$.

The binding probabilities $\theta$ of the six pairs of sv sequences at the remaining four temperatures, $T^* = 0.35, 0.4, 0.45, 0.5$, are calculated by the method described in the main text. As described in main text, we subtract the collision probability of two non-interacting monomer $\theta_0 = \frac{4 \pi 10^3}{100}$ from the experimentally measured bound-state ratio, using $\tilde{\theta} = \theta - \theta_0$ as the measure for binding probability. To aggregate the simulation results from temperatures $T^* = 0.35, 0.4, 0.45, 0.5$ we calculate the enthalpy and entropy of binding for each polymer pair by performing a linear regression according to

$$\Delta H \left( \frac{1}{T^*} \right) - \Delta S = \log(\theta^{-1} - 1).$$  \hfill (S49)

The $\theta$ measured in simulation and the regression results of the six sequence pairs are exhibited in Table S1. We then interpolate $\theta$ from our regression analysis and compare $\tilde{\theta}$ to the $K_D$’s calculated from our theory in Fig. S3, which shows the same trend of that in Fig. 4(c).

Table S1: Simulation Binding data and regression parameters

| Sequence | $\theta$ | $\theta$ | $\theta$ | $\theta$ | $\Delta H$ | $\Delta S$ | $R^2$ |
|----------|----------|----------|----------|----------|------------|------------|-------|
| sv1      | 0.362 %  | 0.432 %  | 0.420 %  | 0.252 %  | -0.295     | -6.32      | 0.225 |
| sv10     | 0.736 %  | 0.743 %  | 0.591 %  | 0.351 %  | -0.810     | -7.08      | 0.720 |
| sv15     | 1.01 %   | 1.64 %   | 0.923 %  | 0.803 %  | -0.383     | -5.46      | 0.202 |
| sv20     | 0.812 %  | 1.34 %   | 0.381 %  | 0.700 %  | -0.594     | -6.33      | 0.178 |
| sv24     | 4.04 %   | 1.83 %   | 2.56 %   | 0.976 %  | -1.39      | -7.17      | 0.703 |
| sv25     | 3.00 %   | 0.590 %  | 0.912 %  | 0.228 %  | -2.59      | -11.1      | 0.787 |
Mathematical principles for negative SCD

Here we present a quick testing method, though not a rigorous proof, for whether all charge neutral sequences have a negative SCD. For a polymer of $N+1$ charges $q_0, \ldots, q_N$, $\text{SCD}(q) := \sum_{i=0}^{N} \sum_{j=i+1}^{N} q_i q_j \sqrt{|i-j|}$. If we define the matrix $(A_{N+1})_{ij} := \sqrt{|i-j|}$, $\text{SCD}(q) = \frac{1}{2} q^T A_{N+1} q$. Say $q$ is a charge pattern such that $\sum_{i=0}^{N} q_i = 0$, so, $q_0 = -\sum_{i=1}^{N} q_i$. Now, defining $\bar{q} = (q_1, \ldots, q_N)$ and the matrix $(B_N)_{ij} := \sqrt{|i-j|} - \sqrt{i} - \sqrt{j}$, one can see that, $\text{SCD}(q) = \frac{1}{2} q^T A_{N+1} q = \frac{1}{2} \bar{q}^T B_N \bar{q}$.

Thus the requirement that $\text{SCD}(q) \leq 0$ for every $q$ with $\sum_{i=0}^{N} q_i = 0$ is equivalent to the requirement that $v^T B_N v \leq 0$ for any $N$-dimensional vector $v$. It is a standard result of linear algebra that, since $B_N$ is self-adjoint, this is in turn equivalent to $B_N$ being a so-called "negative matrix": all of $B_N$’s eigenvalues being negative. Notice as well that for $M < N$, $B_M$ is the top left $M \times M$ submatrix of $B_N$, so, should $B_N$ be negative, $B_M$ would also be negative. For $N = 1000$, the maximum (least-negative) calculated eigenvalue was about -0.760, confirming that SCD is negative for neutral polymers at or under 1001 monomers. The distribution of eigenvalues of $B_{1000}$ is shown in Fig. S4(a).

**Most charge dispersed pattern (N=50).** Another value of interest would be the lowest possible -SCD possible for a neutral polyelectrolyte of some minimum charge (otherwise the 0 charge pattern would have the lowest -SCD at 0). Thus, it is of interest to determine the lowest possible $\frac{q^T A_N q}{q^T q}$ for charge neutral $q$ and the charge pattern that produces it. The minimal value of the above expression as produced using the method of gradient descent is about -0.761, produced by the eigenvector with the charge distribution shown in Fig. S4(b), compared to about -0.826 for the strictly alternating 50-residue polymer.

**Non-neutral sequences.** For a $N$-mer charge pattern $q$, we can define its average charge $\sigma := \frac{1}{N} \sum_{i=1}^{N} q_i$ and its pattern as $p_i = q_i - \sigma$, we can write $q = p + \sigma \mathbf{1}$ where $\mathbf{1}$ is the
vector with a 1 in every entry. Now we can write,

$$\text{SCD}(q) = \frac{1}{2} q^T A_N q$$

$$= \frac{1}{2} p^T A_N p + \sigma p^T A_N 1 + \frac{1}{2} \sigma^2 1^T A_N 1$$

$$= \text{SCD}(p) + \sigma \sum_{i=1}^{N} p_i (\sum_{j=1}^{N} \sqrt{|i-j|}) + \frac{1}{2} \sigma^2 \sum_{i} \sum_{j} \sqrt{|i-j|}$$

$$\approx \text{SCD}(p) + \frac{2}{3} \sigma \sum_{i=1}^{N} p_i (i^{3/2} + (N-i)^{3/2}) + \frac{4}{15} \sigma^2 N^{5/2} \quad \text{(S50)}$$

where the last approximation is by writing the sums as integrals. \(\text{SCD}(p)\) is negative as \(p\) is charge neutral while \(\frac{4}{15} \sigma^2 N^{5/2}\) is, of course, positive and seemingly the primary contributor in increasing \(\text{SCD}\) in the non-neutral case. As for the centre term, \((i^{3/2} + (N-i)^{3/2})\) is the greatest when \(i\) is low or high, i.e. when it represents monomers near the termini of the polymer. Thus, \(\sigma \sum_{i=1}^{N} p_i (i^{3/2} + (N-i)^{3/2})\) is positive if and only if the distribution of those monomers with the same charge as the average charge is biased towards the termini.
Figure S1: Heatmap of binding affinities of all $30 \times 30$ pairs of Das-Pappu sequences arranged in the order of increasing $\kappa$ in each axis. Sequences with higher $\kappa$ have higher binding affinities in agreement with Fig 2(b).

Figure S2: Heatmap of binding affinities of the Das-Pappu sequence pairs as calculated by Eq S15 with renormalized Kuhn lengths obtained using the variational theory of Sawle and Ghosh. White squares indicate an unfavourable (repulsive) interaction and grey squares indicate a weak $K_D$ of greater than 5 mM. The results is distinct from Fig. 2(b). (b) Heatmap of binding affinities of the Das-Pappu sequence pairs as calculated by Eq S15 subtracted from binding affinities as calculated by Eq 19. In general, more charge segregated polymers have a higher penalty on their binding as a result of accounting for intramolecular interactions.
Figure S3: Binding affinities of Das-Pappu sequences sv1, sv10, sv15, sv20, sv24, and sv25 with sv28 as calculated by Eq. 16 (dark blue) compared to that calculated by Eq. S44 (light blue) and \( \bar{\theta} \) calculated from simulation results by the linear interpolation at \( T^* = 0.35 \) in Eq. S49 with parameters as in Table S1 (red). Simulation data aggregated by linear interpolation has a similar profile to the data from the temperature \( T^* = 0.35 \) in Fig 5. Also, intramolecular effects could explain the discrepancy between theory and simulation in the binding of sv20 and sv28 but they are likely over-accounted for when Kuhn lengths are renormalized by the variational theory of Sawle and Ghosh.\(^{12}\)

Figure S4: (a) The eigenvalue distribution of the matrix \( B_{1000} \) for discussing mathematical principles of SCD; all eigenvalues, as shown, are negative, indicating that the SCD value of any charge neutral sequence with equal or fewer than 1001 residues is negative. (b) The charge distribution of a 50-residue, charge-neutral polyampholyte that results in the least-negative SCD value, produced numerically by gradient descent method.
Figure S5: Binary phase diagrams generated by the approximate Flory-Huggins (FH) interaction free energy given by Eq. 20 with the $\chi$ parameters given by Eq. 19 with $T^* = 10$. The sequence A is the Das-Pappu sequence sv28 and the B’s are sv24 in (a), sv25 in (b), and sv20 in (c). Blue dots are numerically solved phase separation states either $\alpha \equiv (\phi^\alpha_A, \phi^\alpha_B)$ or $\beta \equiv (\phi^\beta_A, \phi^\beta_B)$, and black dashed lines are tie lines connected a pair of states $(\alpha, \beta)$. Compared to the RPA phase diagrams in the Fig. 3 in Ref. [8], (a–c) show the same trend that sequences with similar SCDs coalesce whereas those with different SCDs exclude each other; however, the exclusion propensity is greatly overestimated. (d) The $A_{\alpha\beta}$ trends of RPA theory and the approximate FH interaction. $A_{\alpha\beta} \equiv (2/\pi)(|\tan^{-1}(\phi^\alpha_A/\phi^\alpha_B) - \tan^{-1}(\phi^\beta_A/\phi^\beta_B)|)$ is a measure defined in Eq. 26 in Ref. [8] for the exclusion propensity of the two sequences $A$ and $B$ phase separating into phases $\alpha$ and $\beta$, where $\langle \cdots \rangle$ averages the values of all $(\alpha, \beta)$ pairs in (a–c). The orange FH line is always higher than the green RPA line, indicating that the approximate FH interaction always overestimate the exclusion propensity. The last three FH dots are linked by dashed lines because $A_{\alpha\beta}$ is already saturated at the third dot (for sv28-sv20) and the remaining three points are given by the same saturated values.
Figure S6: Critical temperatures \( T^*_c \) of the 30 Das-Pappu sequences calculated by RPA (green dots) and its linear fitting function \( T^*_c = -0.314 \text{SCD} \) (blue line) obtained from Fig. 3b in Ref. [19] and the two linear functions obtained from the approximate FH interaction potential: \( T^*_c = 2.11\sqrt{j_{\text{SCD,cutoff}}} \) (orange line) obtained from Eqs. 21 and 22, and \( T^*_c = -2.11\sqrt{0.118 \times \text{SCD}} \) (red dashed line) obtained from replacing \( j_{\text{SCD,cutoff}} \) by the fitting result \( j_{\text{SCD,cutoff}} = 0.118 \times \text{SCD}^2 \) in Fig. 3c. The two functions obtained from the FH interaction both overestimate the phase separation propensities of all Das-Pappu sequences.

References

1. Kriwacki, R. W.; Hengst, L.; Tennant, L.; Reed, S. I.; Wright, P. E. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. Proc. Natl. Acad. Sci. U. S. A. 1996, 93, 11504–11509.

2. Wright, P. E.; Dyson, H. J. Linking folding and binding. Curr. Opin. Struct. Biol. 2009, 19, 31–38.

3. Bah, A.; Vernon, R. M.; Siddiqui, Z.; Krzeminski, M.; Muhandiram, R.; Zhao, C.; Sonenberg, N.; Kay, L. E.; Forman-Kay, J. D. Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. Nature 2014, 519.

4. Toth-Petroczy, A.; Palmedo, P.; Ingraham, J.; Hopf, T. A.; Berger, B.; Sander, C.; Marks, D. S. Structured states of disordered proteins from genomic sequences. Cell 2016, 167, 158–170.
(5) Mittag, T.; Orlicky, S.; Choy, W.-Y.; Tang, X.; Lin, H.; Sicheri, F.; Kay, L. E.; Tyers, M.; Forman-Kay, J. D. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 17772–17777.

(6) Feng, H.; Zhou, B.-R.; Bai, Y. Binding Affinity and Function of the Extremely Disordered Protein Complex Containing Human Linker Histone H1.0 and Its Chaperone ProT? *Biochemistry* **2018**, *57*, 6645?6648.

(7) Borgia, A.; Borgia, M. B.; Bugge, K.; Kissling, V. M.; Heidarsson, P. O.; Fernandes, C. B.; Sottini, A.; Soranno, A.; Buholzer, K. J.; Nettels, D. et al. Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **2018**, *555*, 61–66.

(8) Lin, Y.-H.; Brady, J. P.; Forman-Kay, J. D.; Chan, H. S. Charge pattern matching as a ‘fuzzy’ mode of molecular recognition for the functional phase separations of intrinsically disordered proteins. *New J. Phys.* **2017**, *19*, 115003.

(9) Wang, J.; Choi, J.-M.; Holehouse, A. S.; Lee, H. O.; Zhang, X.; Jahnel, M.; Maharana, S.; Lemaitre, R.; Pozniakovsky, A.; Drechsel, D. et al. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* **2018**, *174*, 688–699.e16.

(10) Tsang, B.; Arsenault, J.; Vernon, R. M.; Lin, H.; Sonenberg, N.; Wang, L.-Y.; Bah, A.; Forman-Kay, J. D. Phosphoregulated FMRP phase separation models activity-dependent translation through bidirectional control of mRNA granule formation. *Proc. Natl. Acad. Sci. U. S. A.* **2019**.

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

(12) Sawle, L.; Ghosh, K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J. Chem. Phys.* **2015**, *143*, 085101.
(13) Zarin, T.; Strome, B.; Nguyen Ba, A. N.; Alberti, S.; Forman-Kay, J. D.; Moses, A. M. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. bioRxiv 578716 2019.

(14) Nott, T. J.; Petsalaki, E.; Farber, P.; Jervis, D.; Fussner, E.; Plochowietz, A.; Craggs, T. D.; Bazett-Jones, D. P.; Pawson, T.; Forman-Kay, J. D. et al. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. Mol. Cell 2015, 57, 936–947.

(15) Pak, C. W.; Kosno, M.; Holehouse, A. S.; Padrick, S. B.; Mittal, A.; Ali, R.; Yunus, A. A.; Liu, D. R.; Pappu, R. V.; Rosen, M. K. Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein. Mol. Cell 2016, 63, 72–85.

(16) Lin, Y.-H.; Forman-Kay, J. D.; Chan, H. S. Sequence-specific polyampholyte phase separation in membraneless organelles. Phys. Rev. Lett. 2016, 117, 178101.

(17) Das, S.; Amin, A. N.; Lin, Y.-H.; Chan, H. S. Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters. Phys. Chem. Chem. Phys. 2018, 20, 28558–28574.

(18) McCarty, J.; Delaney, K. T.; Danielsen, S. P. O.; Fredrickson, G. H.; Shea, J.-E. Complete Phase Diagram for Liquid–Liquid Phase Separation of Intrinsically Disordered Proteins. J. Phys. Chem. Lett. 2019, 10, 1644–1652.

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

(20) Dignon, G. L.; Zheng, W.; Best, R. B.; Kim, Y. C.; Mittal, J. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. Proc. Natl. Acad. Sci. U. S. A. 2018, 115, 9929–9934.
(21) Pathria, R. K. *Statistical Mechanics, 2nd Ed.*; Elsevier, 2006.

(22) Lin, Y.-H.; Song, J.; Forman-Kay, J. D.; Chan, H. S. Random-phase-approximation theory for sequence-dependent, biologically functional liquid-liquid phase separation of intrinsically disordered proteins. *J. Mol. Liq.* 2017, 228, 176–193.

(23) Ermoshkin, A. V.; Olvera de la Cruz, M. Polyelectrolytes in the presence of multivalent ions: gelation versus segregation. *Phys. Rev. Lett.* 2003, 90, 125504.

(24) Das, S.; Eisen, A.; Lin, Y.-H.; Chan, H. S. A Lattice Model of Charge-Pattern-Dependent Polyampholyte Phase Separation. *J. Phys. Chem. B* 2018, 122, 5418–5431.

(25) Mundy, C. J.; Siepmann, J. I.; Klein, M. L. Calculation of the shear viscosity of decane using a reversible multiple timestep algorithm. *J. Chem. Phys.* 1995, 102, 3376–3380.

(26) Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes. *J. Phys. Chem. B* 1998, 102, 2569–2577.

(27) Nicolas, J. P.; Smit, B. Molecular dynamics simulations of the surface tension of n-hexane, n-decane and n-hexadecane. *Molecular Physics* 2002, 100, 2471–2475.

(28) Pàmies, J. C.; McCabe, C.; Cummings, P. T.; Vega, L. F. Coexistence Densities of Methane and Propane by Canonical Molecular Dynamics and Gibbs Ensemble Monte Carlo Simulations. *Molecular Simulation* 2003, 29, 463–470.

(29) Anderson, J. A.; Lorenz, C. D.; Travesset, A. General purpose molecular dynamics simulations fully implemented on graphics processing units. *J. Comput. Phys.* 2008, 227, 5342–5359.

(30) Glaser, J.; Nguyen, T. D.; Anderson, J. A.; Lui, P.; Spiga, F.; Millan, J. A.; Morse, D. C.; Glotzer, S. C. Strong scaling of general-purpose molecular dynamics simulations on GPUs. *Computer Physics Communications* 2015, 192, 97 – 107.
(31) LeBard, D. N.; Levine, B. G.; Mertmann, P.; Barr, S. A.; Jusufi, A.; Sanders, S.; Klein, M. L.; Panagiotopoulos, A. Z. Self-assembly of coarse-grained ionic surfactants accelerated by graphics processing units. *Soft Matter* **2012**, *8*, 2385–2397.

(32) Silmore, K. S.; Howard, M. P.; Panagiotopoulos, A. Z. Vapour-liquid phase equilibrium and surface tension of fully flexible Lennard-Jones chains. *Molecular Physics* **2017**, *115*, 320–327.