Supplementary information for: Folded domain charge properties influence the conformational behavior of disordered tails

Ishan Taneja[1,2], Alex S. Holehouse[1,2]

[1]**Department of Biochemistry and Molecular Biophysics**, Washington University School of Medicine, 660 S. Euclid Ave., St. Louis, MO, 63110, USA
[2]**Center for Science and Engineering of Living Systems (CSELS)**, Washington University in St. Louis, St. Louis, MO 63130, USA

Supplementary Methods

Additional details regarding $\hat{\varphi}$

While calculating the net charge of a protein can be thought of as a way to quantify its electrostatic nature, such a metric ignores the spatial orientation of the protein's atoms, the partial charges induced by individual atoms, and the protein's interactions with its neighboring solvent and ions. By using Poisson-Boltzmann electrostatics calculations, one can take these factors into account and calculate the electrostatic potential at a given point on the surface of the protein [1]. Averaging the electrostatic potential across all points on the surface protein then provides a measure of the overall electrostatic potential generated by the surface of the protein in a solvated environment.

1

Supplementary Tables

Parameter	Group Comparison	Effect Size / P-Value
FCR	All IDRs vs. All Tails	0.004 / 0.38
FCR	All Tails vs. Tails with PDB Structure	0.009 / 0.29
NCPR	All IDRs vs. All Tails	0.03 / <1e-5
NCPR	All Tails vs. Tails with PDB Structure	0.01 / 0.13
FP	All IDRs vs. All Tails	0.02 / 0.0003
FP	All Tails vs. Tails with PDB Structure	0.02 / 0.03
FN	All IDRs vs. All Tails	0.03 / <1e-5
FN	All Tails vs. Tails with PDB Structure	0.0001 / 0.99
Kappa	All IDRs vs. All Tails	0.04 / <1e-5
Kappa	All Tails vs. Tails with PDB Structure	0.02 / 0.02
% of Phosphorylatable Residues	All IDRs vs. All Tails	0.05 / <1e-5
% of Phosphorylatable Residues	All Tails vs. Tails with PDB Structure	0.02 / 0.04
% of Hydrophobic Residues	All IDRs vs. All Tails	0.08 / <1e-5
% of Hydrophobic Residues	All Tails vs. Tails with PDB Structure	0.009 / 0.32
% of Aromatic Residues	All IDRs vs. All Tails	0.03 / <1e-5
% of Aromatic Residues	All Tails vs. Tails with PDB Structure	0.02 / 0.08
# of Residues	All IDRs vs. All Tails	0.05 / <1e-5
# of Residues	All Tails vs. Tails with PDB Structure	0.02 / 0.09

Supplementary Table 1. Comparisons of distributions for a given parameter between two groups were performed using the Wilcoxon rank-sum test (also known as the Mann-Whitney U test). The corresponding effect size (*r*) can range between 0 and 1, and values below 0.1 are considered to be insignificant.

Tail	Length	NCPR	FCR	Kappa	Disorder Promoting
$(GSK)_4$	12	0.33	0.33	0.053	1.00
$(GSK)_6$	18	0.33	0.33	0.031	1.00
$(GSK)_8$	24	0.33	0.33	0.027	1.00
$(GSK)_{10}$	30	0.33	0.33	0.025	1.00
$(GSK)_{12}$	36	0.33	0.33	0.024	1.00
$(GSE)_4$	12	−0.33	0.33	0.053	1.00
$(GSE)_6$	18	−0.33	0.33	0.031	1.00
$(GSE)_8$	24	−0.33	0.33	.027	1.00
$(GSE)_{10}$	30	−0.33	0.33	0.025	1.00
$(GSE)_{12}$	36	−0.33	0.33	0.024	1.00

Supplementary Table 2. Characteristics for each IDR. GSX_n refers to a block of GSX repeating *n* times.

GFP variant	Net Charge	$\widehat{\varphi}$	p
GFP^{+15}	15	6.04	0.60
GFP^{+5}	5	2.84	0.61
GFP^{-15}	−15	−2.11	0.63

Supplementary Table 4. Characteristics of each GFP variant. $\widehat{\varphi}$ and p respectively refer to the mean electrostatic potential and patchiness among all residues.

GFP variant	Sequence
GFP^{+15}	VPILVKLNGNVNGHKFSVSGEGEGDATYGELTLEFICT TGKLPVPWPTLVTTLGGGVQCFSQYPNHMKQHDFFESA MPEGYVQKRTISFKNNGNYKTRAKVEFEGDTLVNRIKL KGINFKKNGNILGHKLKYNYNSHNVYITANKQKNGIKA NFKIRHNIKNGSVQLANHYQQNTPIGNGPVLLPNNHYL STQSALSKNPNKKRNHMVLLEFVTA
GFP^{+5}	VPILVELNGNVNGHKFSVSGEGEGDATYGELTLEFICT TGKLPVPWPTLVTTLGGGVQCFSQYPDHMEQHDFFESA MPEGYVQKQTISFKNNGNYKTRAEVEFEGDTLVNQIKL KGINFKKNGNILGHKLKYNYNSHNVYITANKQKNGIKA NFKIRHNIKNGSVQLANHYQQNTPIGNGPVLLPNNHYL STQSALSKDPNKKRNHMVLLEFVTA
GFP^{-15}	VPILVELDGNVNGHKFSVSGEGEGDATYGELTLEFICT TGKLPVPWPTLVTTLGGGVQCFSQYPDHMEQHDFFESA MPEGYVQEQTISFKNNGNYKTQAEVEFEGDTLVNQIEL KGINFKKNGNILGHKLKYNYNSHNVYITANKQENGIEA NFKIRHNIKNGSVQLANHYQQNTPIGDGPVLLPDNHYL STQSALSEDPNEEQDHMVLLEFVTA

Supplementary Table 5. Sequence for each GFP variant.

System	Droplet radius (Å)	Total # of replicas	Steps per simulation	Equilibration steps	Conformations per simulation	Total ensemble size
(GSE)$_4$-GFP^X	233	30	20e6	2e6	2e3	6e4
(GSE)$_6$-GFP$^{+15/+5}$	236	30	20e6	2e6	2e3	6e4

$(GSE)_8\text{-}GFP^{+15/+5}$	240	30	20e6	2e6	2e3	6e4
$(GSE)_6\text{-}GFP^{-15}$	236	60	20e6	2e6	2e3	12e4
$(GSE)_8\text{-}GFP^{-15}$	240	60	20e6	2e6	2e3	12e4
$(GSE)_{10}\text{-}GFP^{X}$	244	30	20e6	2e6	2e3	6e4
$(GSE)_{12}\text{-}GFP^{X}$	124	30	20e6	2e6	2e3	6e4
$(GSK)_4\text{-}GFP^{X}$	233	30	20e6	2e6	2e3	6e4
$(GSE)_6\text{-}GFP^{X}$	236	30	20e6	2e6	2e3	6e4
$(GSK)_8\text{-}GFP^{X}$	240	30	20e6	2e6	2e3	6e4
$(GSK)_{10}\text{-}GFP^{X}$	244	30	20e6	2e6	2e3	6e4
$(GSK)_{12}\text{-}GFP^{X}$	124	30	20e6	2e6	2e3	6e4
$(GSE)_4$	39	10	200e6	2e6	2e4	2e5
$(GSE)_6$	50	10	200e6	2e6	2e4	2e5
$(GSE)_8$	60	10	200e6	2e6	2e4	2e5
$(GSE)_{10}$	69	20	100e6	2e6	2e4	2e5
$(GSE)_{12}$	77	20	100e6	2e6	2e4	2e5
$(GSK)_4$	39	10	200e6	2e6	2e4	2e5
$(GSK)_6$	50	10	200e6	2e6	2e4	2e5
$(GSK)_8$	60	10	200e6	2e6	2e4	2e5
$(GSK)_{10}$	69	20	100e6	2e6	2e4	2e5
$(GSK)_{12}$	77	20	100e6	2e6	2e4	2e5
$HCoV\text{-}OC43_{NTD\text{-}IDR}\text{-}HCoV\text{-}OC43_{RBD}$	174	50	24e6	8e6	4e2	2e4
$HCoV\text{-}OC43_{NTD\text{-}IDR}\text{-}SARS\text{-}CoV\text{-}2_{RBD}$	174	50	24e6	8e6	4e2	2e4
$SARS\text{-}CoV\text{-}2_{NTD\text{-}IDR}\text{-}HCoV\text{-}OC43_{RBD}$	159	50	24e6	8e6	4e2	2e4
$SARS\text{-}CoV\text{-}2_{NTD\text{-}IDR}\text{-}SARS\text{-}CoV\text{-}2_{RBD}$	159	50	24e6	8e6	4e2	2e4
$HCoV\text{-}OC43_{NTD\text{-}IDR}$	108	20	100e6	2e6	2e4	2e5
$SARS\text{-}CoV\text{-}2_{NTD\text{-}IDR}$	94	20	100e6	2e6	2e4	2e5

Supplementary Table 6. Overview of the simulation input details for each system.

Parameter 1	Parameter 2	Pearson Correlation	p-value
IDR NCPR	FD NCPR	.12	.08
IDR NCPR	FD Total Patch Size	-.02	.73
IDR NCPR	FD Net Patch Size*Potential	.08	.24
IDR FCR	FD FCR	.21	.002
IDR FCR	FD Total Patch Size	.17	.01
IDR FCR	FD Net Patch Size*Potential	.11	.12

Supplementary Table 7. Correlations between relevant IDR and FD charge properties. Total patch size for a given protein is calculated as $\sum_i s_i$ where s_i refers to the relative size of patch i. Net patch size*potential for a given protein is calculated as $\sum_i s_i * \varphi_i$ where s_i refers to the relative size of patch i and $\widehat{\varphi_i}$ refers to the mean electrostatic potential of patch i.

Sequence	Length	Kappa	FCR	NCPR	Hydropathy
HCoV-OC43$_{NTD-IDR}$	50	.37	.20	.08	2.85
SARS-CoV-2$_{NTD-IDR}$	63	.26	.18	.11	3.32

Supplementary Table 8. Sequence properties for HCoV-OC43$_{NTD-IDR}$ and SARS-CoV-2$_{NTD-IDR}$.

Sequence	Length	FCR	NCPR	Hydropathy
HCoV-OC43$_{RBD}$	126	.20	-.02	3.80
SARS-CoV-2$_{RBD}$	124	.20	.04	3.86

Supplementary Table 9. Sequence properties for HCoV-OC43$_{RBD}$ and SARS-CoV-2$_{RBD}$.

Supplementary Figures

Supplementary Figure 1. Bioinformatic workflow for IDRs with or without structural information. The proportion of N and C-terminal tails is roughly equivalent.
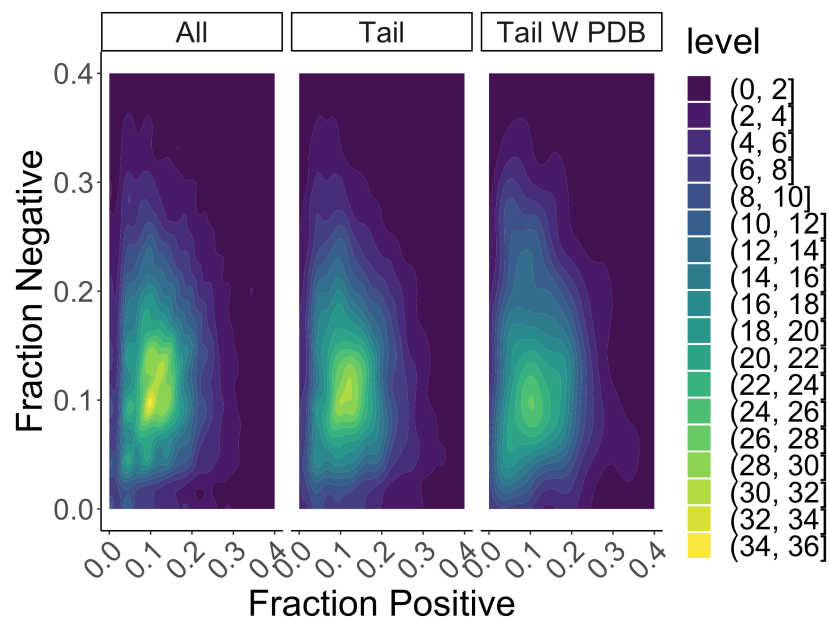
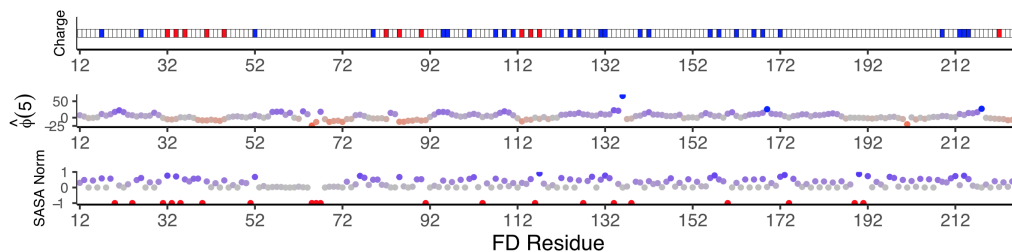**Supplementary Figure 2.** Bioinformatic workflow for disordered tails with structural information.

Supplementary Figure 3. The charge properties of disordered tails are similar to those of all IDRs. 4 sets of IDRs were included in this analysis: all IDRs with or without structural data (N = 34,095), all tails with or without structural data (N = 11,254), and all tails with structural data (N = 619). **(A,B,C,D)** Violin/box plots of fraction of positive residues (A), fraction of negative residues (B), kappa (C), percentage of phosphorylatable sites (D), percentage of hydrophobic residues (E), and percentage of aromatic residues (F) comparing all IDRs with or without structural data, all tails with or without structural data, and all tails with structural data.

Supplementary Figure 4. The charge diagram of states of disordered tails is similar to those of all IDRs. 2D density plot of fraction of positive charges and fraction of negative charges for all IDRs with or without structural data, all tails with or without structural data, and all tails with structural data.
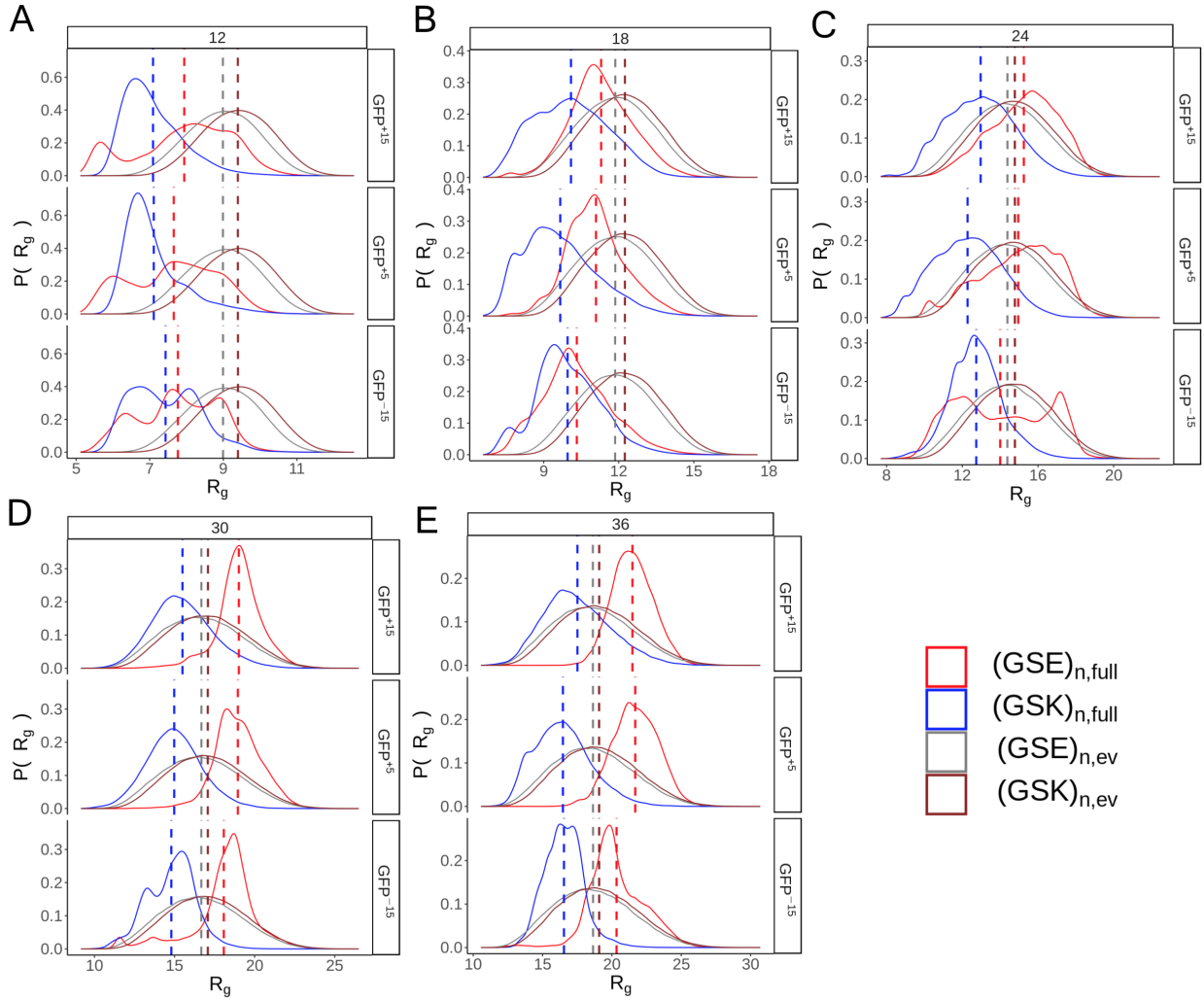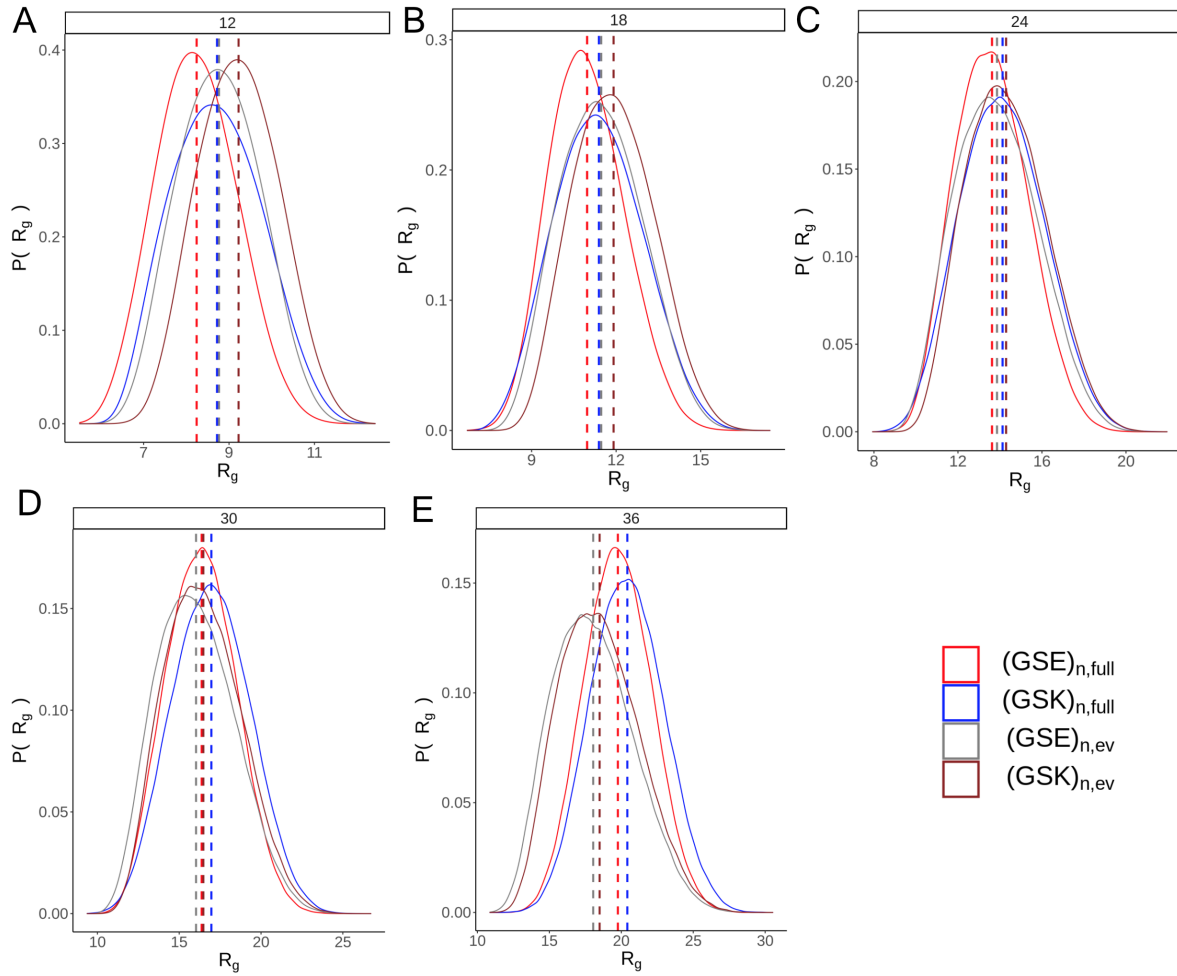
Supplementary Figure 5. Sequence characteristics for each GFP variant. For each GFP variant, we characterized the charge of each residue (ASP/GLU = −1 [Red], LYS/ARG = +1 [Blue]). In addition, we characterized the mean electrostatic potential within a 5 Å radius radius to assess the local electrostatic potential around that region. Finally, we showed the normalized solvent accessibility of each residue, where 1 corresponds to being fully accessible and 0 corresponds to being fully inaccessible.
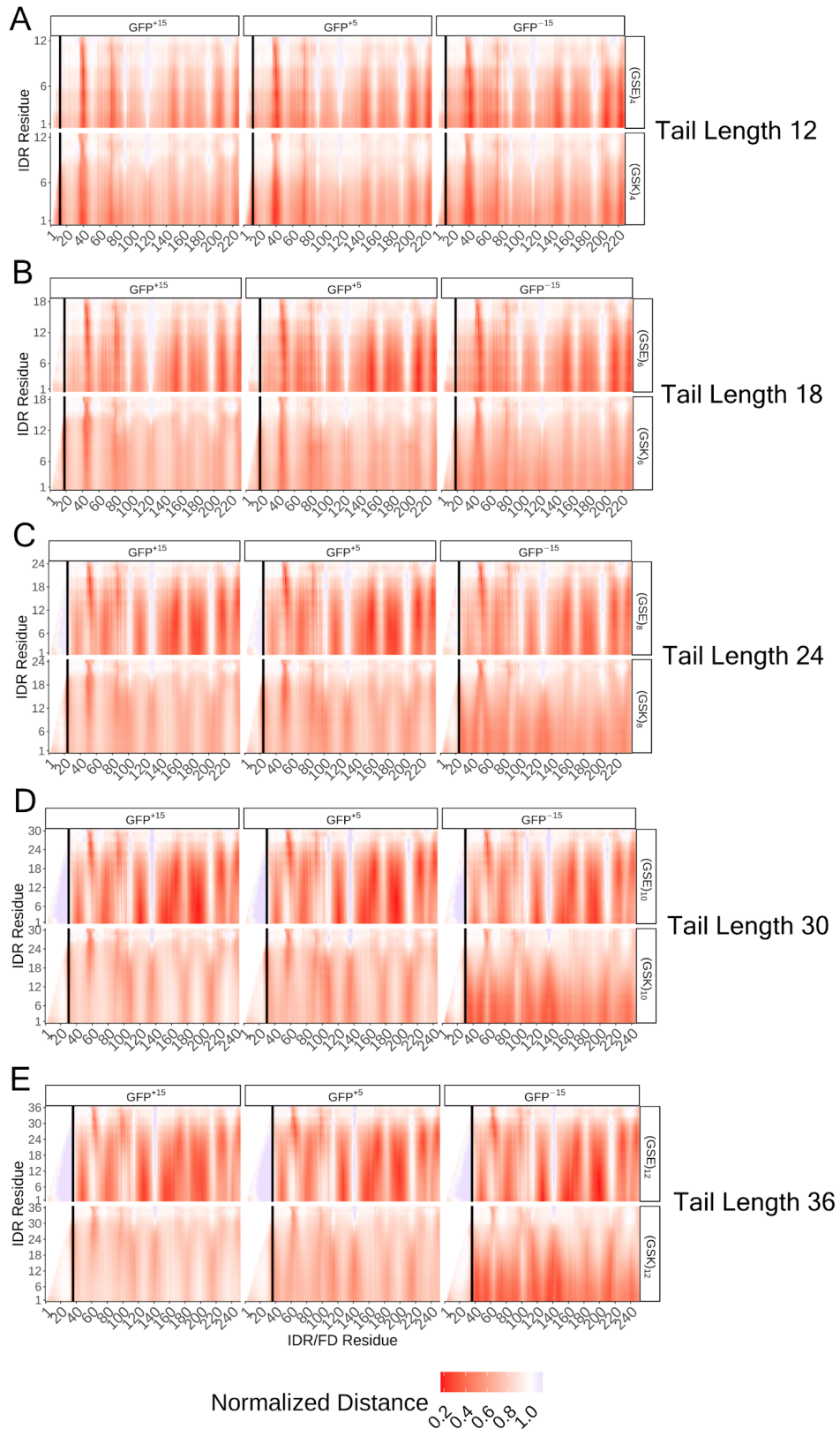
Supplementary Figure 6. The radius of gyration of $(GSE)_n$-GFP^X and $(GSK)_n$-GFP^X is dependent on tail length and the charge properties of both tails and the folded domain surface. Each panel refers to the radius of gyration distributions for systems consisting of a given tail length and GFP variant for both the full Hamiltonian and excluded volume simulations. For instance, at a tail length of 36, the GFP^{+15} subpanel consists of the radius of gyration distributions for $(GSE)_{12}$-GFP^{+15} and $(GSK)_{12}$-GFP^{+15} in the full Hamiltonian and excluded volume simulations. **(A)** R_g distributions for $(GSE/K)_4$-GFP^X systems. **(B)** R_g distributions for $(GSE/K)_6$-GFP^X systems. **(C)** R_g distributions for $(GSE/K)_8$-GFP^X systems. **(D)** R_g distributions for $(GSE/K)_{10}$-GFP^X systems. **E)** R_g distributions for $(GSE/K)_{12}$-GFP^X systems.
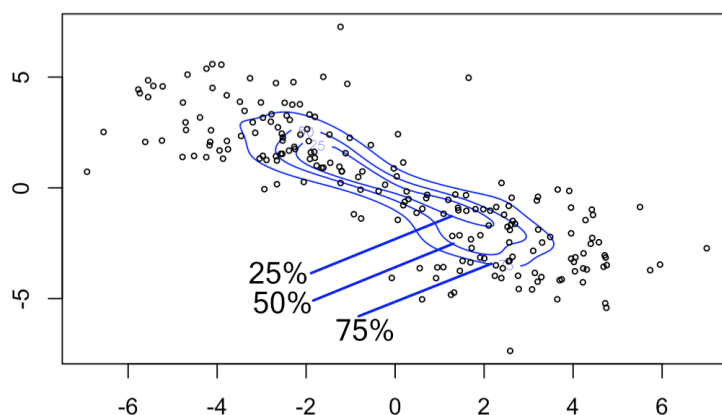
Supplementary Figure 7. Radius of gyration distributions for (GSE)$_n$ and (GSK)$_n$. Each panel refers to the radius of gyration distributions for each tail in isolation for both the full Hamiltonian and excluded volume simulations. **(A)** R$_g$ distributions for (GSE/K)$_4$. **(B)** R$_g$ distributions for (GSE/K)$_6$. **(C)** R$_g$ distributions for (GSE/K)$_8$. **(D)** R$_g$ distributions for (GSE/K)$_{10}$. **(E)** R$_g$ distributions for (GSE/K)$_{12}$.
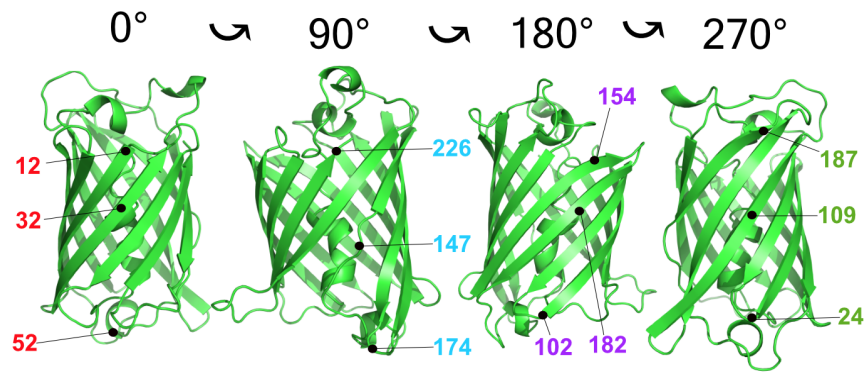
Supplementary Figure 8. The conformational behavior of (GSE)$_{4/6/8/10/12}$-GFP^X and (GSK)$_{4/6/8/10/12}$-GFP^X is dependent on the tail's net charge. In addition, the conformational behavior of (GSK)$_{8/10/12}$-GFP^X is

dependent on surface charge properties. Finally, certain systems (*e.g.,* $(GSE)_n\text{-}GFP^{-15}$ and $(GSK)_n\text{-}GFP^{-15}$) exhibit a dependence between tail length and the strength of the IDR-FD interactions. Each panel refers to a scaling map for all systems at a given tail length: **(A)** $(GSE/K)_4\text{-}GFP^X$. **(B)** $(GSE/K)_6\text{-}GFP^X$. **(C)** $(GSE/K)_8\text{-}GFP^X$. **(D)** $(GSE/K)_{10}\text{-}GFP^X$. **(E)** $(GSE/K)_{12}\text{-}GFP^X$. Each entry in the scaling map refers to the average distance between residues *i* and *j* in the full Hamiltonian simulation divided by the average distance between residues *i* and *j* in the corresponding excluded volume simulation. Inter-residue distances were calculated between all residues of the tail (residues 1-36) and the entire protein.
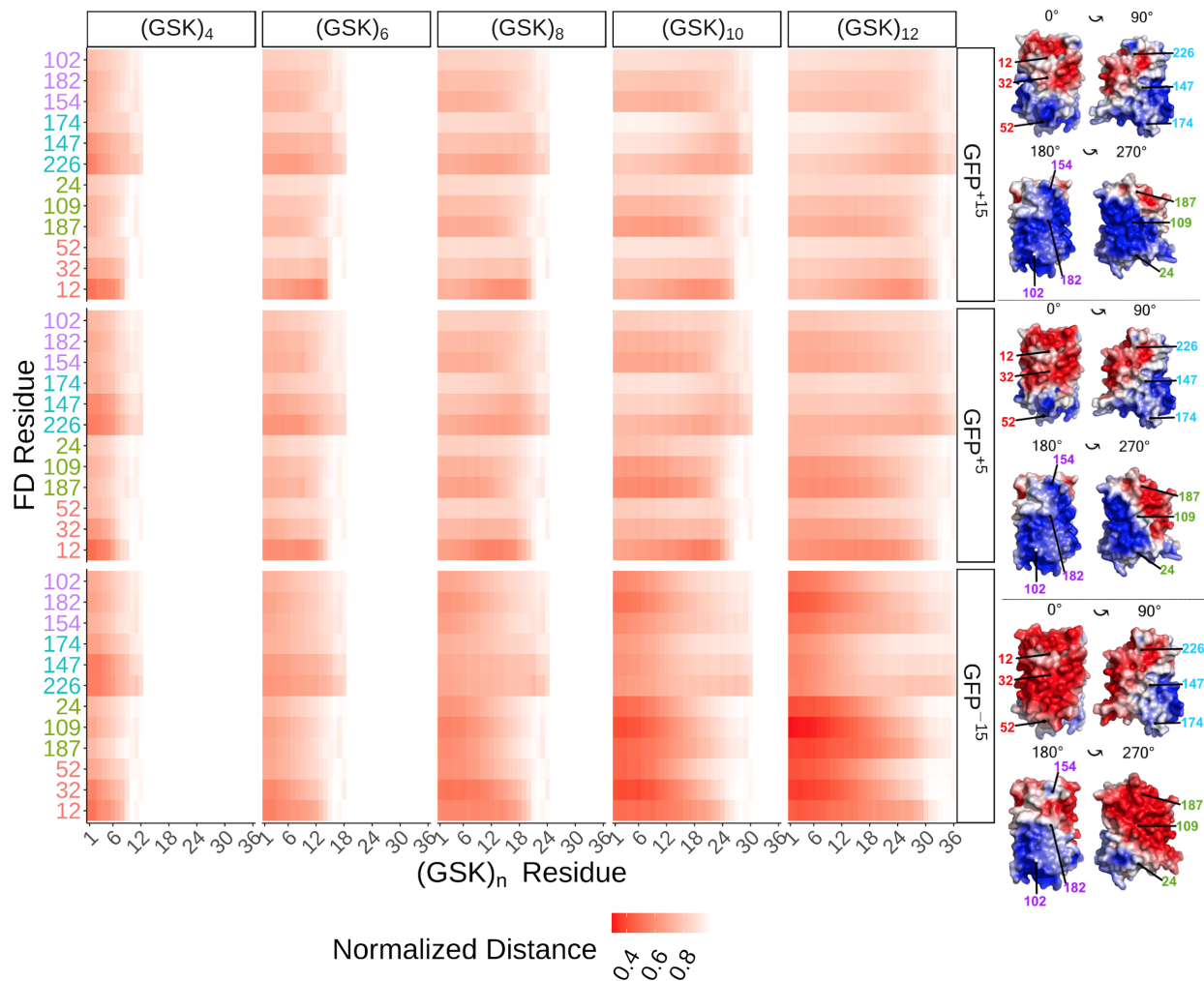

Supplementary Figure 9. 2D contour area plot for 200 data points generated from a mixture of three normal distributions centered at (-2,2), (0,0), and (2,-2) respectively. Each boundary contains either 25, 50, or 75 percent of the area of a probability density distribution. On average, the x% area contour contains x% of the points that were used to generate the kernel density estimate. For pedagogic references, see [2] and http://www.spatialecology.com/htools/pctvolcontour.php.
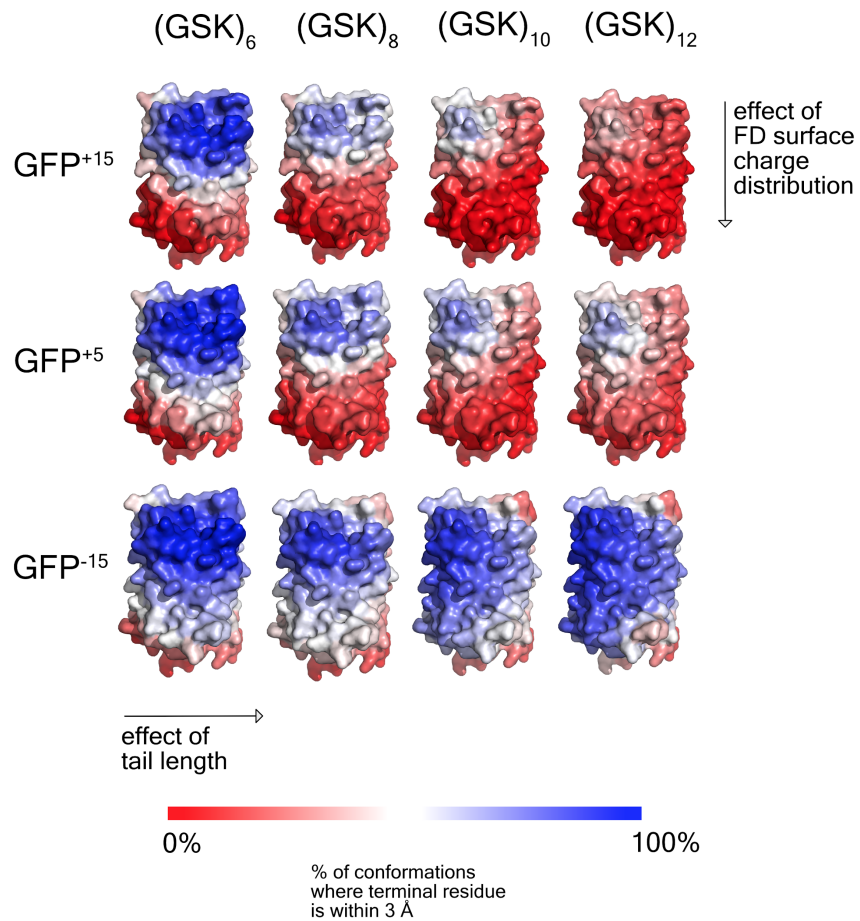
16

Orientation	Residues Used
0°	12 (top), 32 (middle), 52 (bottom)
90° counterclockwise	226 (top), 147 (middle), 174 (bottom)
180° counterclockwise	154 (top), 182 (middle), 102 (bottom)
270° counterclockwise	187 (top), 109 (middle), 24 (bottom)

Supplementary Figure 10. The 12 residues of the FD used in the scaling maps of Figure 7 and Figure 9. Each black dot represents the approximate location of one of the 12 residues.
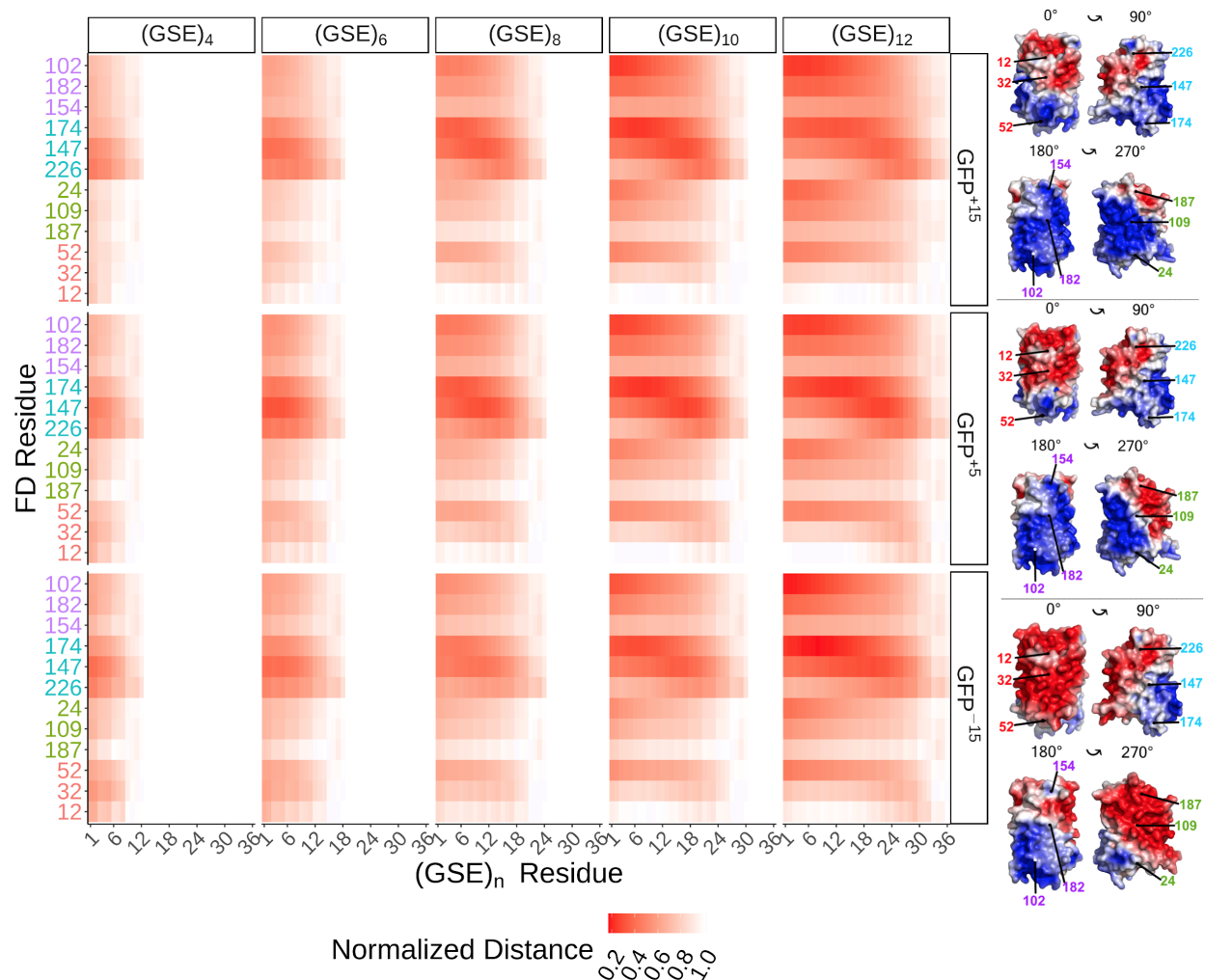
Supplementary Figure 11. Scaling maps for (GSK)ₙ-GFP⁺¹⁵, (GSK)ₙ-GFP⁺⁵, and (GSK)ₙ-GFP⁻¹⁵ at tail lengths 18, 24, 30, and 36 illustrate that conformations tend to cluster closer to the negative patch of the GFP in a manner dependent on tail length and surface charge. Each entry in the scaling map refers to the average distance between residues *i* and *j* in the full Hamiltonian simulation divided by the average distance between residues *i* and *j* in the corresponding excluded volume simulation. Inter-residue distances were calculated between all residues of a given tail and 12 specific residues on the FD. To the right of the scaling map, we show these residues in context of the electrostatic surface potential maps for each GFP variant.
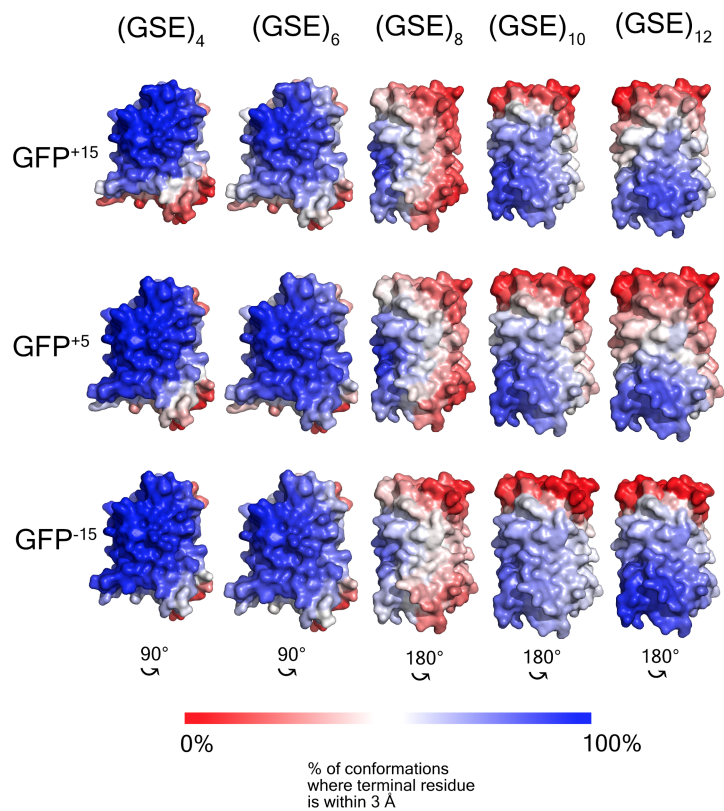
Supplementary Figure 12. Contact maps overlaid for each (GSK)$_{6/8/10/12}$-GFP^x system illustrate that the degree of interaction with the surface around the N-terminus increases as the size/charge of the anchoring negative patch increases. In addition, the degree of interaction with the surface around the N-terminus is correlated with tail length in a manner dependent on the surface charge distribution of the FD. For each residue on the FD, we calculated the percentage of conformations within 3 Å of the terminal residue of the tail and then superimposed this information on each GFP variant.
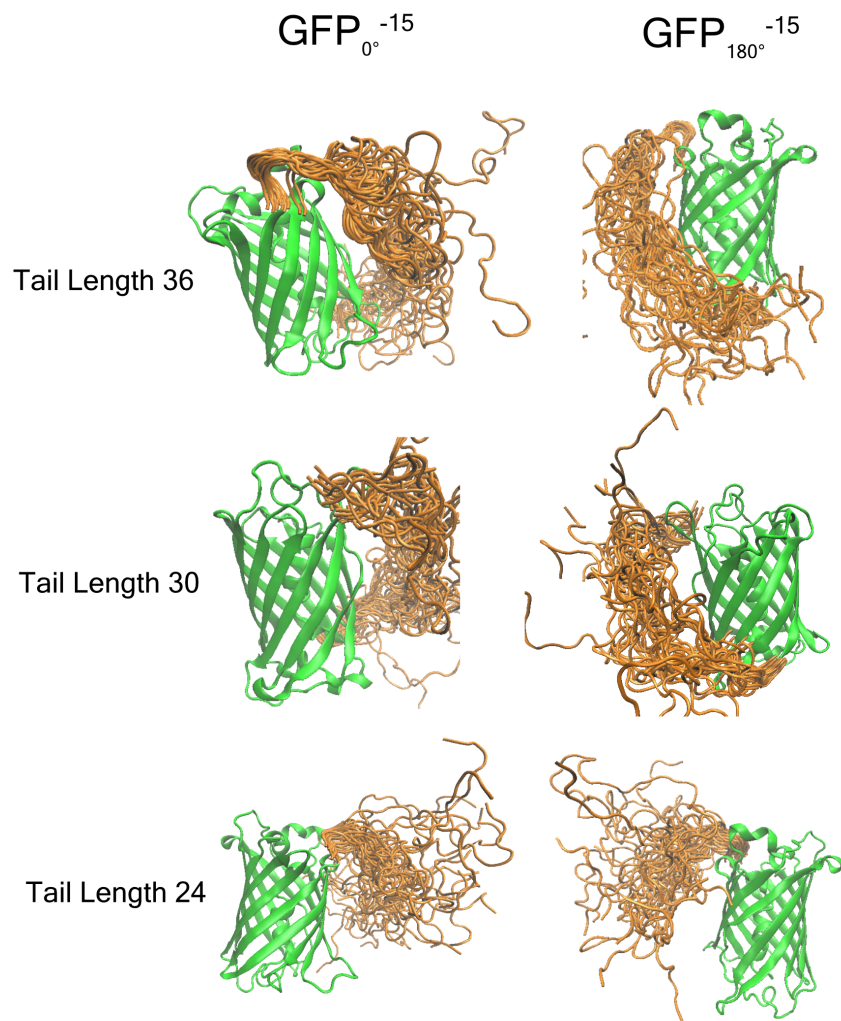
Supplementary Figure 13. Scaling maps of (GSE)$_n$-GFP^{+15}, (GSE)$_n$-GFP^{+5}, and (GSE)$_n$-GFP^{-15} at tail lengths 18, 24, 30, and 36 illustrate that conformations cluster towards or near the back of the GFP as tail length increases. Each entry in the scaling map refers to the average distance between residues i and j in the full Hamiltonian simulation divided by the average distance between residues i and j in the corresponding excluded volume simulation. Inter-residue distances were calculated between all residues of a given tail and 12 specific residues on the FD. To the right of the scaling map, we show these residues in context of the electrostatic surface potential maps for each GFP variant.
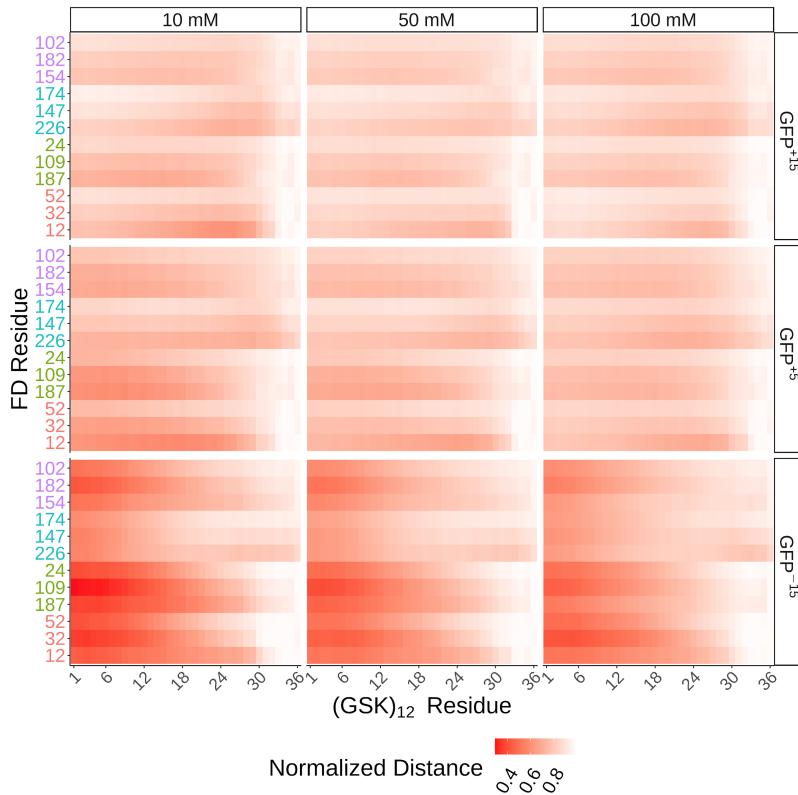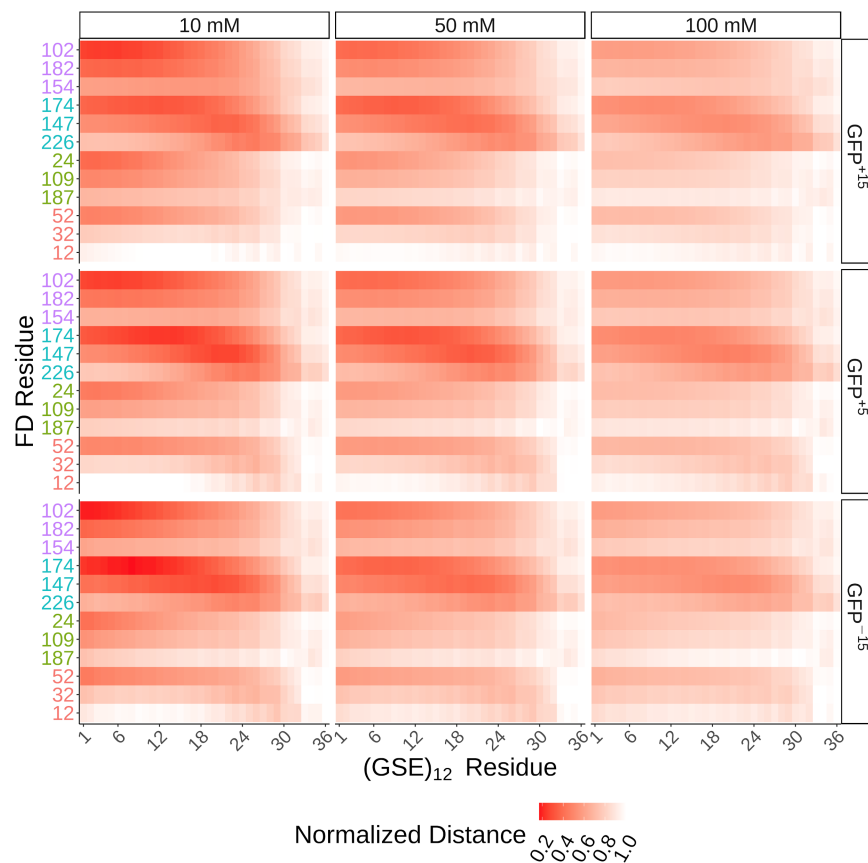
Supplementary Figure 14. Contact maps overlaid on each GFP variant for each $(GSE)_n$-GFP^x system illustrate that the negatively charged tail either tends to orient towards or be in close contact with the positive patch localized around $GFP_{180°}^x$. For each residue on the FD, we calculated the percentage of conformations within 3 Å of the terminal residue of the tail and then superimposed this information on each GFP variant.
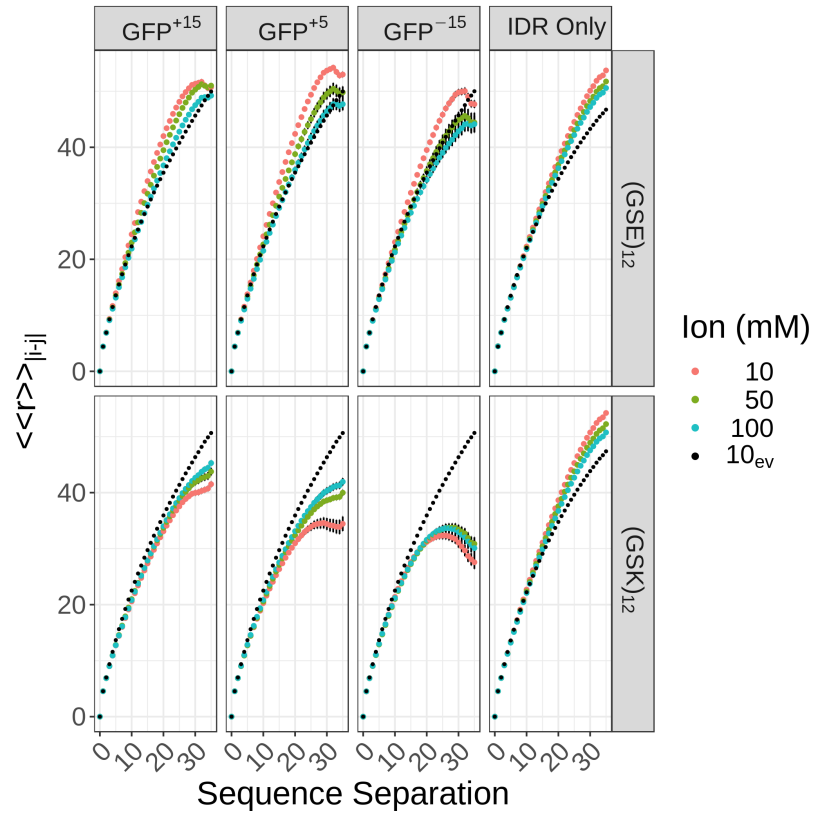
$$\text{GFP}_{0°}^{-15} \qquad \text{GFP}_{180°}^{-15}$$

Tail Length 36

Tail Length 30

Tail Length 24

Supplementary Figure 15. Snapshots of frames illustrate that (GSE)$_{8/10/12}$ is able to orient away from its adjacent negative patch and then loop clockwise around the GFP to either orient towards or be in close contact with the positive patch. For each image, we display a snapshot of 40 frames (each separated by 500,000 steps) from a random replica of the simulations for that particular system in a specific orientation of the GFP.
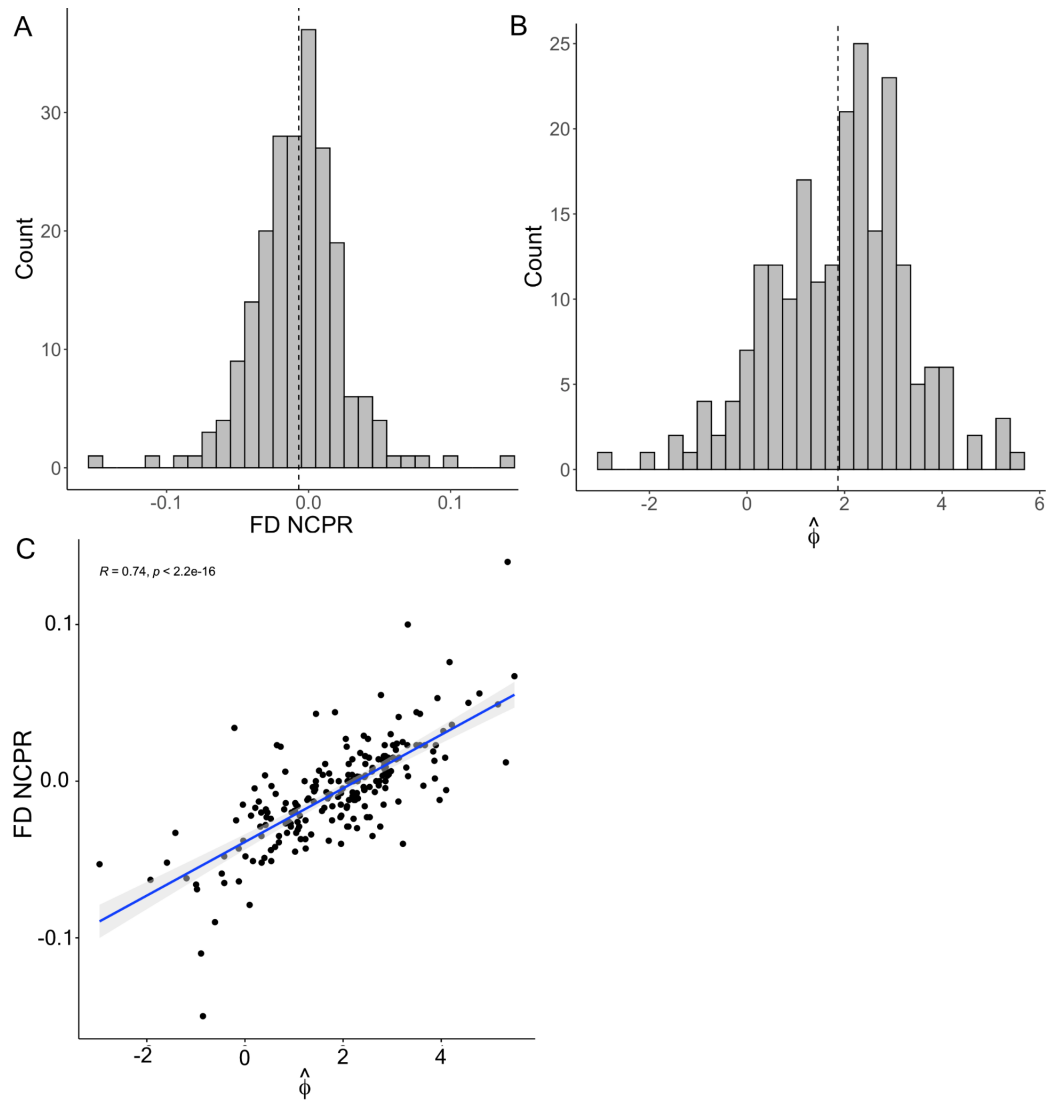
Supplementary Figure 16. Scaling maps of (GSK)$_{12}$-GFP^{+15}, (GSK)$_{12}$-GFP^{+5}, and (GSK)$_{12}$-GFP^{-15} at NaCl concentrations of 10 mM, 50 mM, and 100 mM. We observed that interactions between each tail residue and FD residue were consistently attenuated at higher ion concentrations compared to that of their corresponding 10 mM systems. The degree of reduction in interactions was roughly consistent at 50 mM and 100 mM, suggesting that the effect of increasing ion concentrations can be nuanced depending on the precise interplay between ion-IDR, ion-FD, IDR-IDR and IDR-FD interactions. Each entry in the scaling map refers to the average distance between residues i and j in the full Hamiltonian simulation divided by the average distance between residues i and j in the corresponding excluded volume simulation. Inter-residue distances were calculated between all residues of a given tail and 12 specific residues on the FD.

Supplementary Figure 17. Scaling maps of (GSE)$_{12}$-GFP^{+15}, (GSE)$_{12}$-GFP^{+5}, and (GSE)$_{12}$-GFP^{-15} at NaCl concentrations of 10 mM, 50 mM, and 100 mM. We observed that interactions between each tail residue and FD residue were consistently attenuated at higher ion concentrations compared to that of their corresponding 10 mM systems. We also observed that the degree of reduction in interactions was proportional to the ion concentration. Each entry in the scaling map refers to the average distance between residues i and j in the full Hamiltonian simulation divided by the average distance between residues i and j in the corresponding excluded volume simulation. Inter-residue distances were calculated between all residues of a given tail and 12 specific residues on the FD.
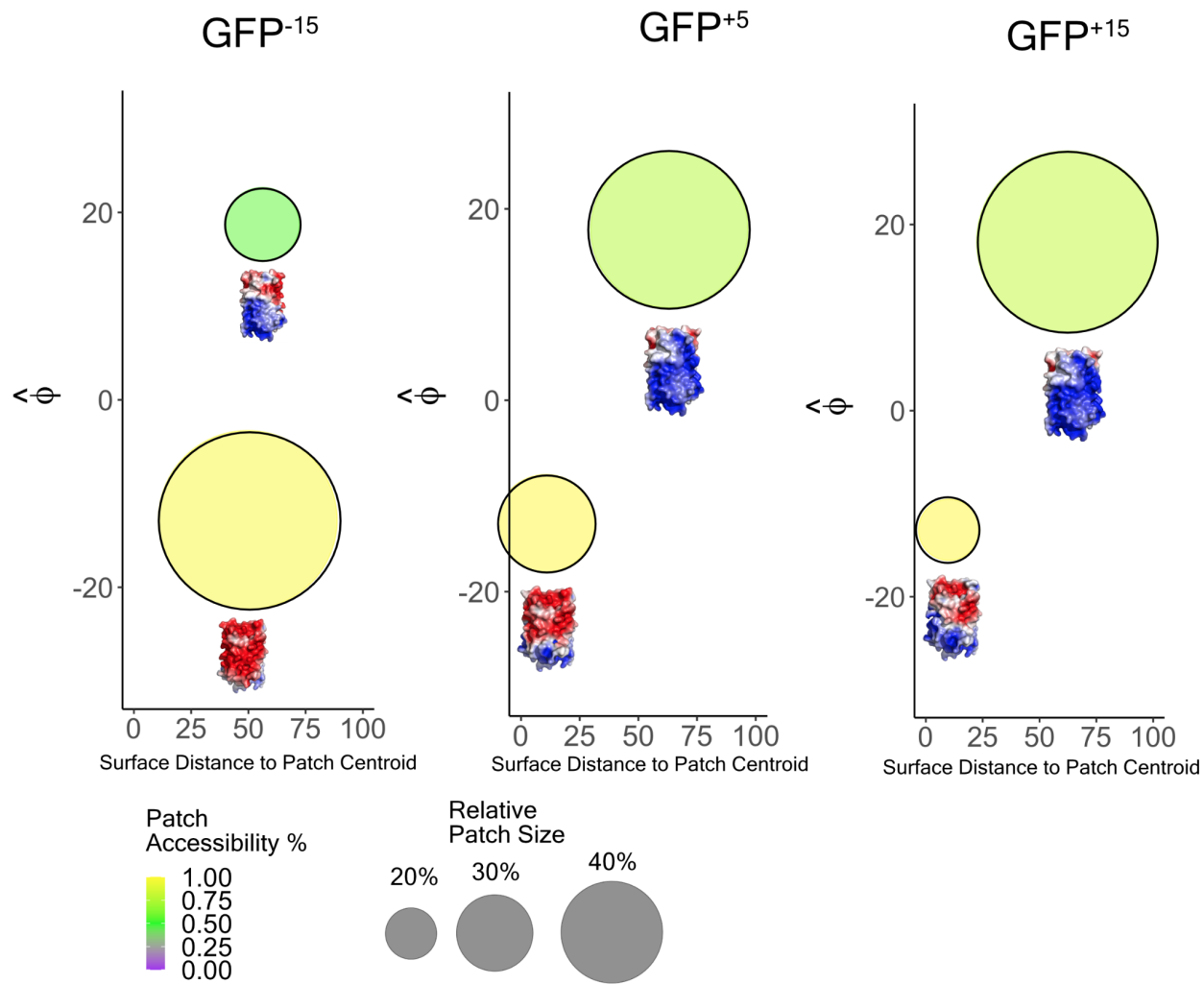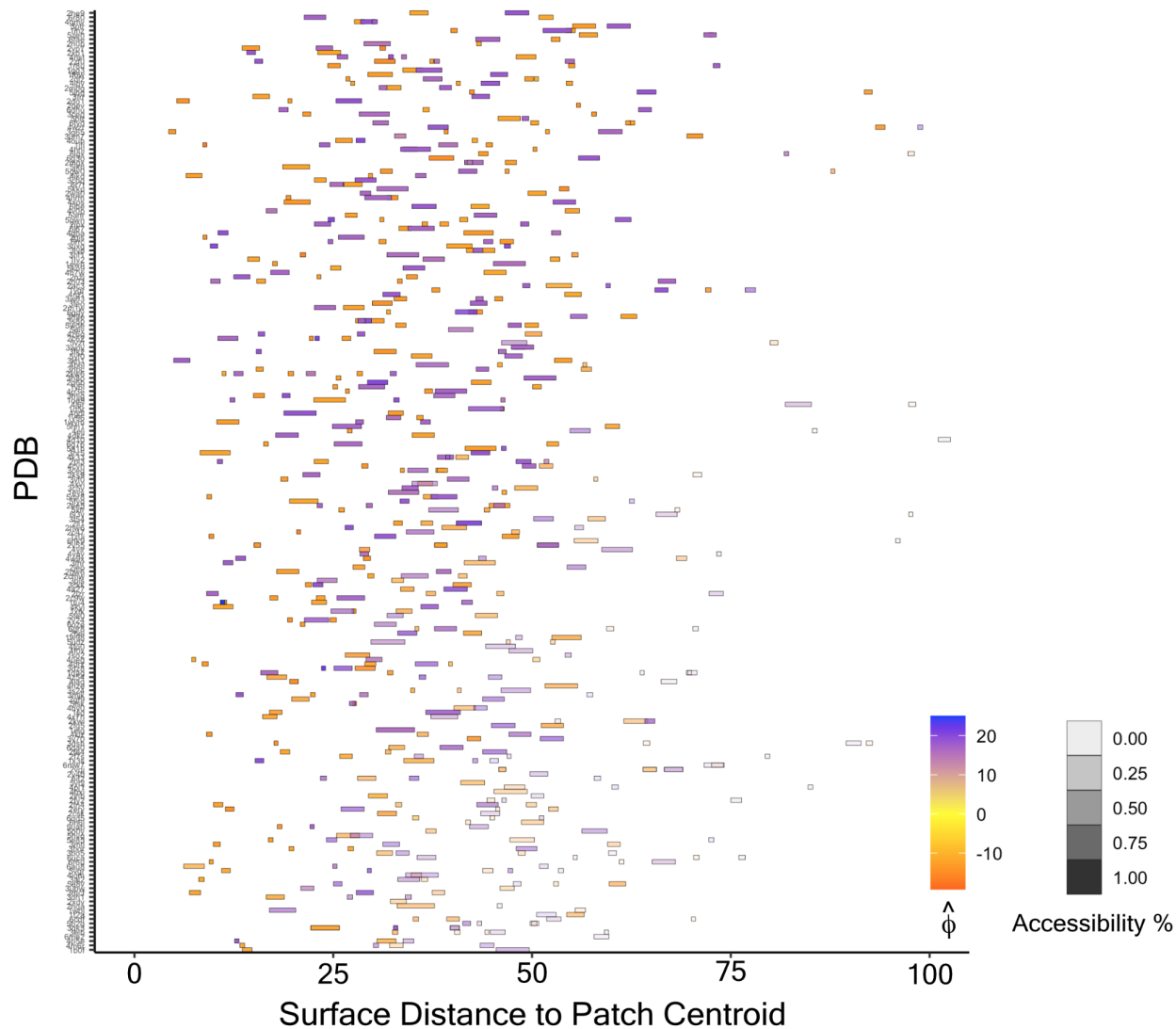
Supplementary Figure 18. Internal scaling profiles of (GSE)$_{12}$ and (GSK)$_{12}$ attached to each GFP variant (first, second, and third columns) and in isolation (fourth column) for the full Hamiltonian simulations as a function of NaCl concentration. The black dotted line represents the internal scaling profile for (GSE/K)$_{12}$ at 10 mM attached to each GFP variant and in isolation for the excluded volume simulations.The spatial separations for all pairs of IDR residues that are |j–i| apart in the linear sequence are calculated for each of the conformations in the relevant ensemble.The patterns of intra-IDR distances can be summarized in terms of so-called internal scaling profiles. The ensemble-averaged spatial distance for each sequence separation, denoted as $\langle\langle r_{i,j} \rangle\rangle_{|i-j|}$ is plotted against |j–i|.

Supplementary Figure 19. Though the NCPR and mean electrostatic potential of a protein are highly correlated, the mean electrostatic potential is capturing information that the NCPR is not. A) Histogram of the NCPR of the FDs across all 214 proteins. **B)** Histogram of the mean electrostatic potential of the FDs across all 214 proteins. **C)** Correlation between a FD's mean electrostatic potential and NCPR across all 214 proteins.
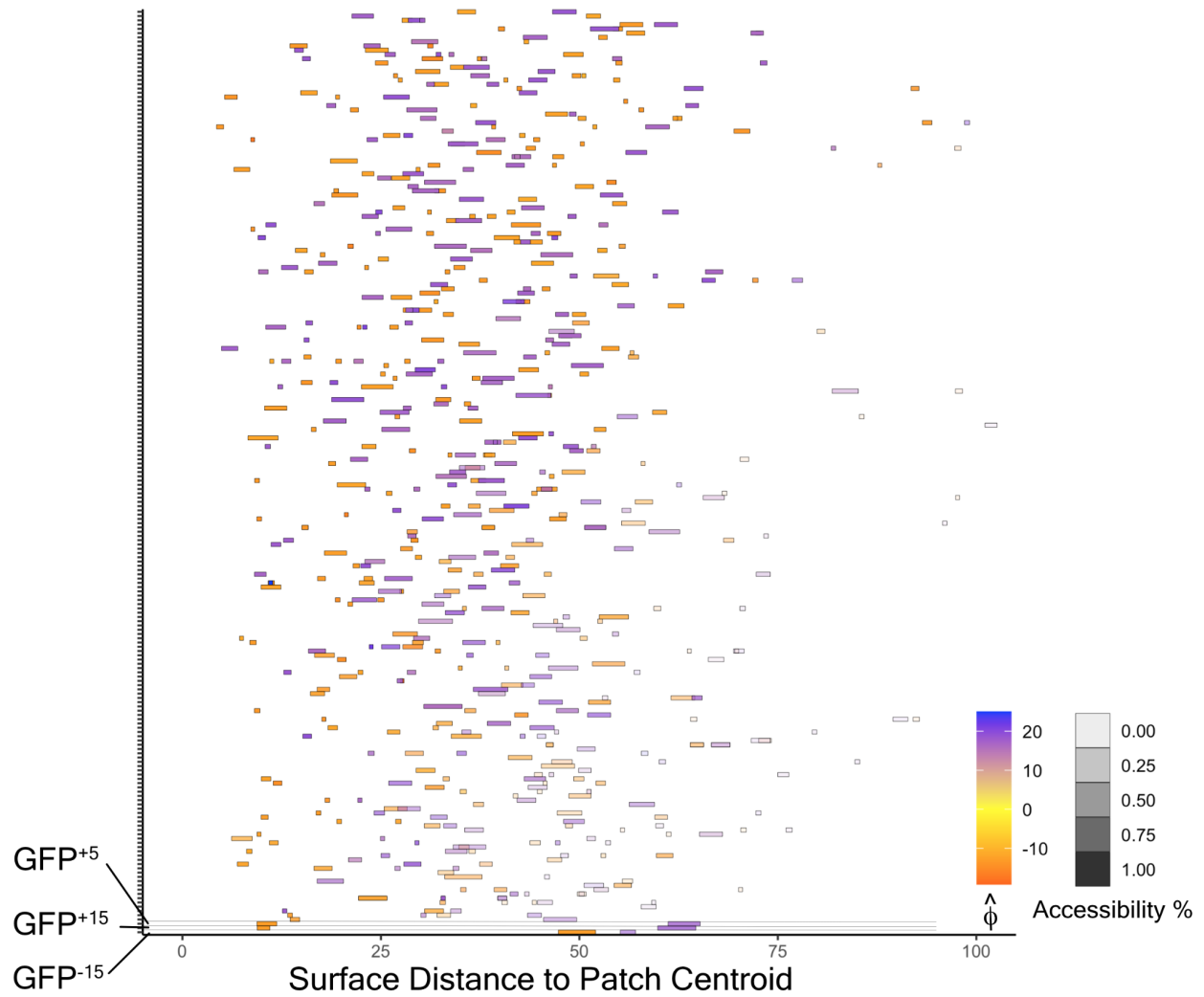
Supplementary Figure 20. Patch visualization for each GFP mutant. Each subplot visualizes the mean electrostatic potential, relative size, and accessibility fraction of each patch as a function of distance between the N-terminal junction and the patch centroid along the surface of the protein.
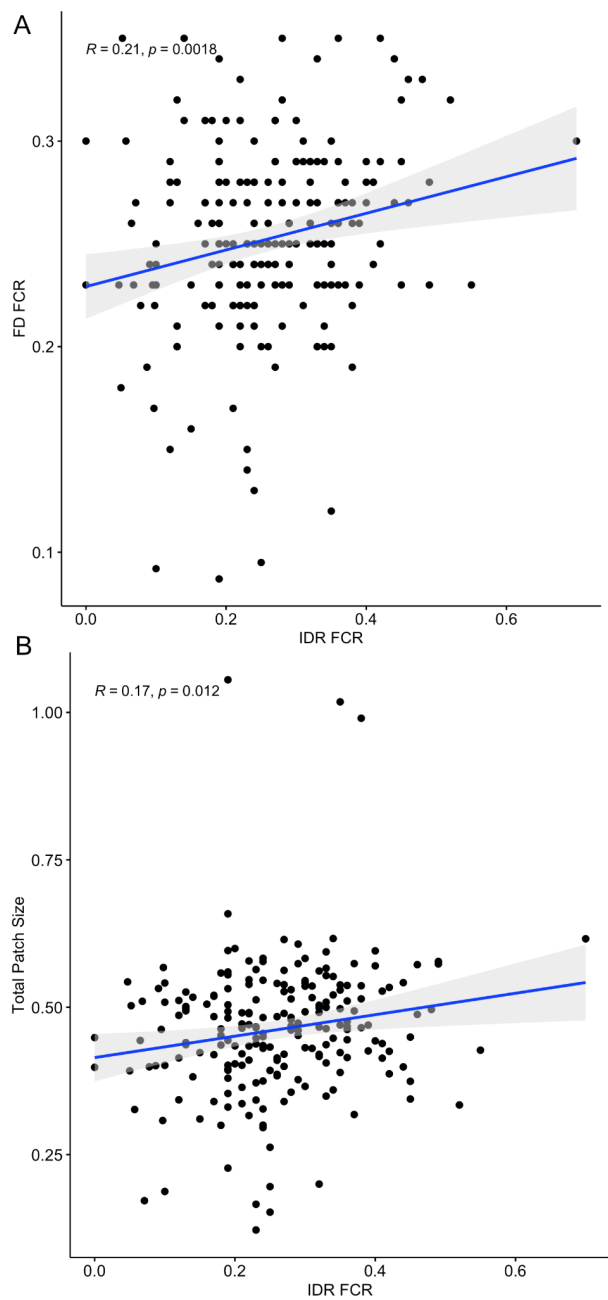
Supplementary Figure 21. Patch information for all 214 proteins illustrates the heterogeneity of the patchiness landscape. The y-axis corresponds to a given protein identified by its PDB ID. For each protein, we quantified its mean electrostatic potential, relative size, and accessibility fraction of each patch as a function of distance between the FD:IDR junction and the patch centroid along the surface of the protein. This data is represented in a tile plot, where each rectangle represents a patch and encodes information about the patch's mean electrostatic potential (color), relative size (width), and accessibility fraction (shade).

Supplementary Figure 22. IDR per residue charge for all 214 proteins illustrates the heterogeneity of the IDR sequence landscape. The y-axis corresponds to a given protein identified by its PDB ID. For each protein, we characterized the charge of each residue (ASP/GLU = −1 [Red], LYS/ARG = +1 [Blue]) on its corresponding N/C terminal tail.

GFP^{+5}

GFP^{+15}

GFP^{-15}

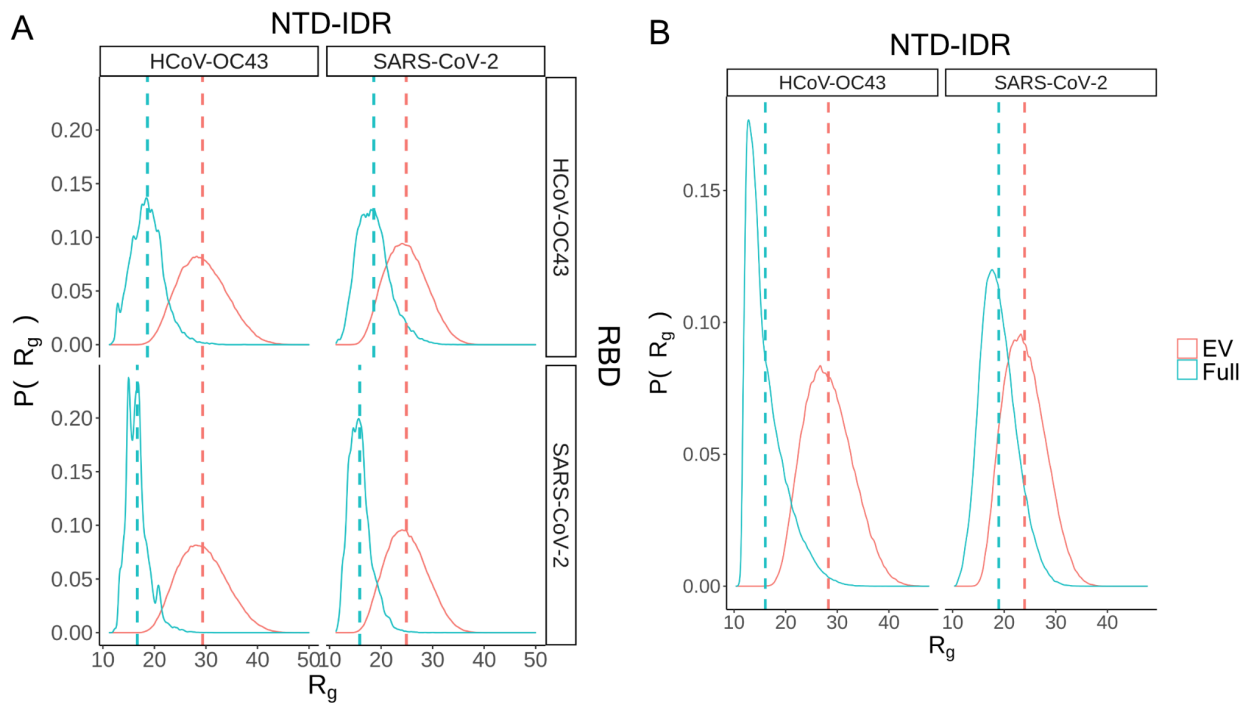Surface Distance to Patch Centroid

$\hat{\phi}$ Accessibility %

Supplementary Figure 23. While the three GFP mutants are arguably 'extreme' cases in terms of their surface charge distribution, it is not uncommon for real proteins to have complementary charged patches of varying sizes.
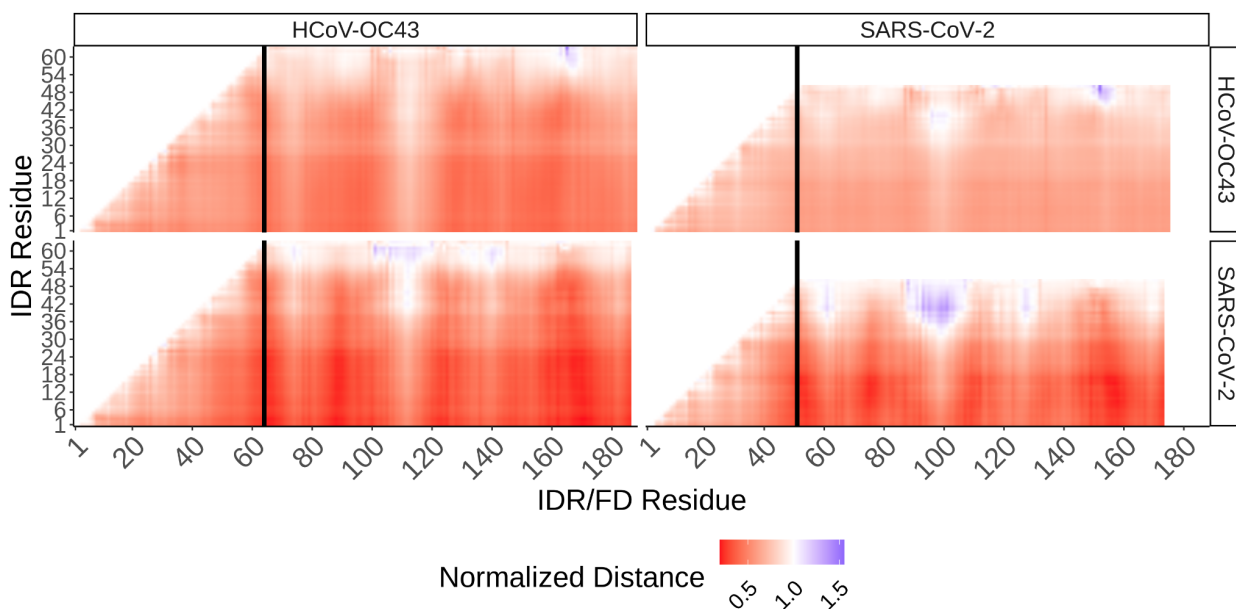
Supplementary Figure 24. Correlations between relevant IDR and FD charge properties. **A)** Correlation between IDR FCR and FD FCR. **B)** Correlation between IDR FCR and FD total patch size. Total patch size for a given protein is calculated as $\sum_i s_i$ where s_i refers to the relative size of patch i.

Supplementary Figure 25. The radius of gyration of HCoV-OC43$_{\text{NTD-IDR}}$ and SARS-CoV-2$_{\text{NTD-IDR}}$ is greater when attached to HCoV-OC43$_{\text{RBD}}$ compared to when it is attached to SARS-CoV-2$_{\text{RBD}}$. (A) R$_g$ distributions for HCoV-OC43$_{\text{NTD-IDR}}$-HCoV-OC43$_{\text{RBD}}$, HCoV-OC43$_{\text{NTD-IDR}}$-SARS-CoV-$_{\text{RBD}}$, SARS-CoV-2$_{\text{NTD-IDR}}$-HCoV-OC43$_{\text{RBD}}$, and SARS-CoV-2$_{\text{NTD-IDR}}$-SARS-CoV-2$_{\text{RBD}}$. **(B)**. R$_g$ distributions for HCoV-OC43$_{\text{NTD-IDR}}$ and SARS-CoV-2$_{\text{NTD-IDR}}$ in isolation.

Supplementary Figure 26. Scaling maps indicate that the SARS-CoV-2$_{RBD}$ yields preferential attractive and repulsive interactions with HCoV-OC43$_{NTD-IDR}$ and SARS-CoV-2$_{NTD-IDR}$ while HCoV-OC43$_{RBD}$ does not. Each entry in the scaling map refers to the average distance between residues *i* and *j* in the full Hamiltonian simulation divided by the average distance between residues *i* and *j* in the corresponding excluded volume simulation. Inter-residue distances were calculated between all residues of a given tail and the entire N-protein (i.e the NTD-IDR and RBD).

Supplementary References

[1]   N.A. Baker, D. Sept, S. Joseph, M.J. Holst, J.A. McCammon, Electrostatics of nanosystems: application to microtubules and the ribosome, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 10037–10041.

[2]   M. Gibin, P. Longley, P. Atkinson, Kernel Density Estimation and Percent Volume Contours in General Practice Catchment Area Analysis in Urban Areas, (n.d.). https://www.geos.ed.ac.uk/~gisteac/proceedingsonline/GISRUK2007/PDF/5A3.pdf.