Biophysical Journal, Volume *123*

Supplemental information

Probing the role of the protonation state of a minor groove-linker histidine in Exd-Hox–DNA binding

Yibei Jiang, Tsu-Pei Chiu, Raktim Mitra, and Remo Rohs

Supplemental Information

Probing the role of the protonation state of a minor groove-linker histidine in Exd-Hox–DNA binding

Yibei Jiang[1], Tsu-Pei Chiu[1], Raktim Mitra[1], and Remo Rohs[1,2,3,4,*]

[1]Department of Quantitative and Computational Biology, [2]Department of Chemistry, [3]Department of Physics and Astronomy, and [4]Thomas Lord Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA

*To whom correspondence should be addressed. Tel: +1 (213) 740-0552; Fax: +1 (213) 821-4257; Email: rohs@usc.edu

S-I SELEX-SEQ SEQUENCE CONSTRUCT

S-I.1 Flanking sequence preference among Scr core motifs

We investigated the DNA shape for flanks with different affinities using Top-Down Crawl (1). However, the high-affinity Sex combs reduced (Scr) core motif is a palindromic sequence and cannot be analyzed directly by this approach. Therefore, we looked for similar flanking preferences of other Scr cores (Fig. S1). Here, we fixed the 5' end to a high-affinity flank, AGAA. We selected four flanks of various affinities for the 3' end, which is in contact with Scr protein, to focus on the effect of His-12. We also included a 4-base pair (bp) A-tract sequence due to its known influence on DNA shape (2). Sequences used in the simulation are provided in Table S1. A GC cap was used on both the 5' and 3' ends in the molecular dynamics (MD) simulations (Table S1).

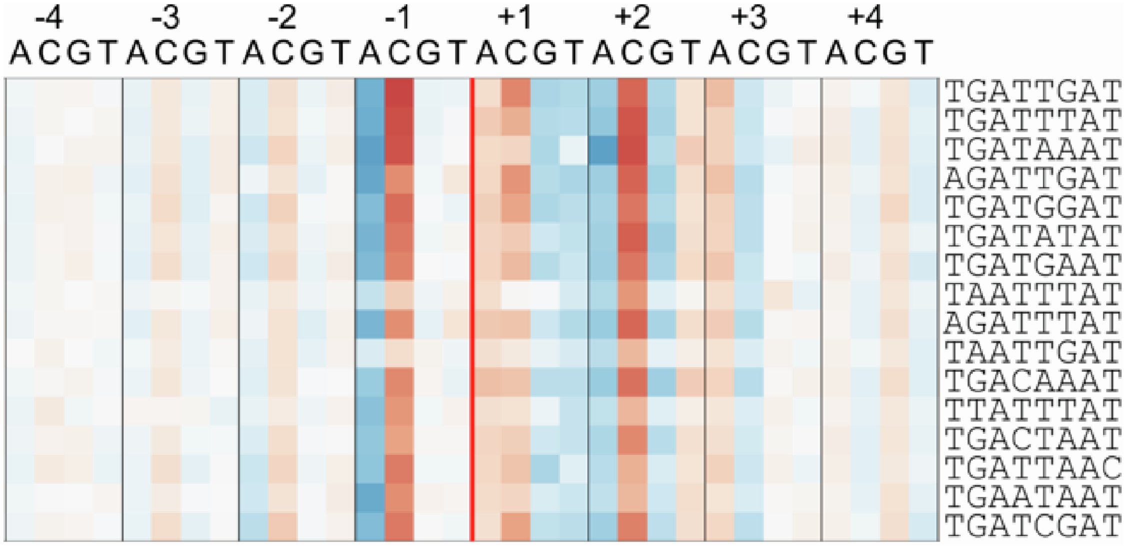

Figure S1. Flanking sequence preference table for various Scr preferred cores. Blue and red indicate higher and lower binding affinity, respectively. -X (e.g., -4, -3, etc.) or +X (+4, +3, etc.) indicate X nucleotide positions to the left or right, respectively, of the core sequence.

S-I.2 Flanking sequence constructs used in this MD study

Affinity	**Sequence**
Very-Low	5’ GC **AGAA** $T_1G_2A_3T_4T_5A_6A_7T_8$ **AAAA** GC 3’
Low	5’ GC **AGAA** $T_1G_2A_3T_4T_5A_6A_7T_8$ **CCAG** GC 3’
Med-Low	5’ GC **AGAA** $T_1G_2A_3T_4T_5A_6A_7T_8$ **ATTA** GC 3’
Med-High	5’ GC **AGAA** $T_1G_2A_3T_4T_5A_6A_7T_8$ **TGGC** GC 3’
High	5’ GC **AGAA** $T_1G_2A_3T_4T_5A_6A_7T_8$ **GACT** GC 3’

Table S1. DNA sequences with various affinities were used in the MD simulations. The 8-bp Scr-preferred core motif was used. The 5’ flank was fixed (green), while the 3’ flanks had different affinities. Blue and red indicate higher and lower affinity, respectively.

S-II. Simulation convergence evaluation

To assess simulation convergence, we clustered the simulations using the Gromos (3) method with different root mean squared deviation (RMSD) cutoffs. The RMSD between any two structures is calculated on the protein and DNA heavy atoms. The simulation was considered to converge when no more new clusters were discovered. For clustering, 300-ns trajectories were divided into 3x10^{4} frames, each 10-ps apart. Three RMSD cutoffs, 2.0 Å, 2.5 Å, and 3.0 Å, were chosen (as in (4)) to ensure that the convergence assessment was not dependent on the choice of cutoff. The number of new clusters over time is shown in Fig. S2. The rate of discovery of new clusters approached 0 after 150 ns for most systems, showing that the simulations converged after 150 ns.

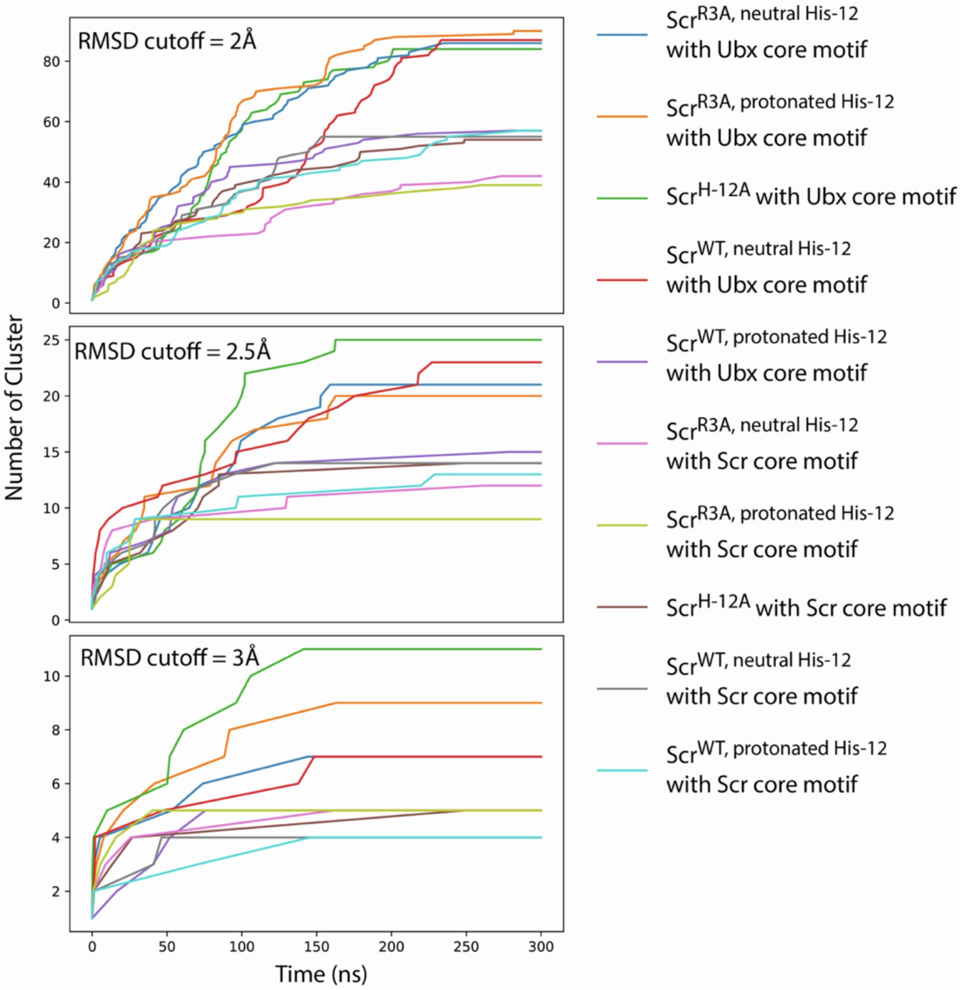

Figure S2. Number of structural clusters as a function of time. Cutoffs of 2 Å, 2.5 Å, and 3 Å were used when assigning structural clusters.

S-III 21-bp MGW shape profile

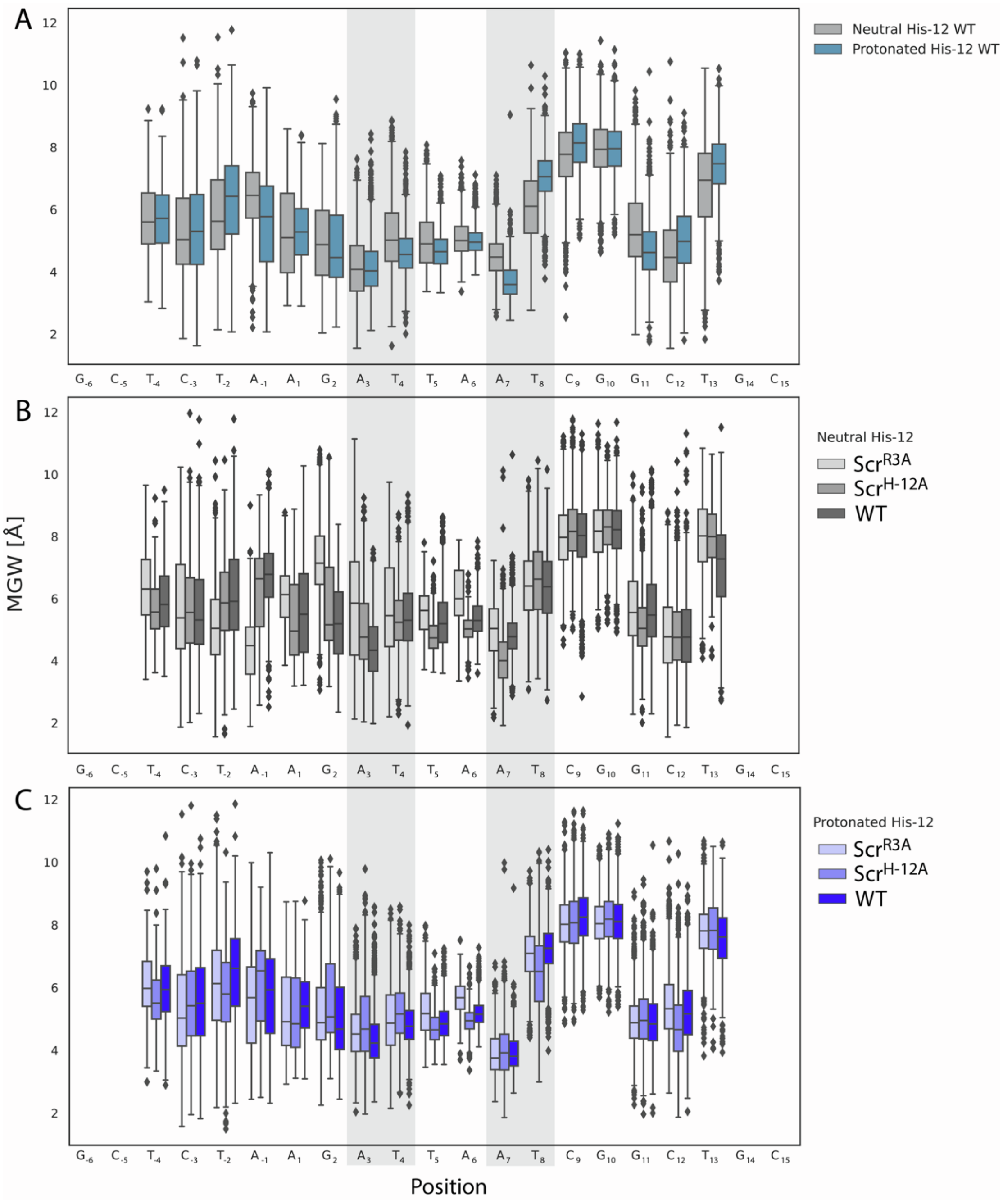

Figure S3. Distribution of MGW for the entire 21-bp sequence aggregated from three replicas. MGWs for terminal padding GCs are missing (G_{-6}, C_{-5}, G_{14}, C_{15}) because high fluctuations existed during the trajectory, and MGW calculations failed. MGW values for the remaining 17 bp are plotted for **(A)** the DNA sequence in PDB ID 2R5Z, **(B)** protein mutant–DNA complexes with neutral His-12, and **(C)** protein mutant–DNA complexes with protonated His-12. Gray regions indicate the locations of the two MGW minima.

S-IV RMSF of wild-type (WT) systems

To comprehend the effect of histidine protonation on Scr protein residues, the root mean square fluctuation (RMSF) was calculated for Cα atoms using the last 150 ns of the trajectory. The RMSF showed reductions of ~0.3 Å in the linker region (Fig. S4A), and the distribution of the A7 MGW was narrowed (Fig. S9A), in the protonated His-12–containing protein with Scr core motif. Therefore, not only did the second minimum have less fluctuation, but the Scr protein containing protonated His-12 exhibited less motion. The distribution of the RMSF deviation in Fig. S4B shows that most residue has less than 0.2 Å deviations among replicas.

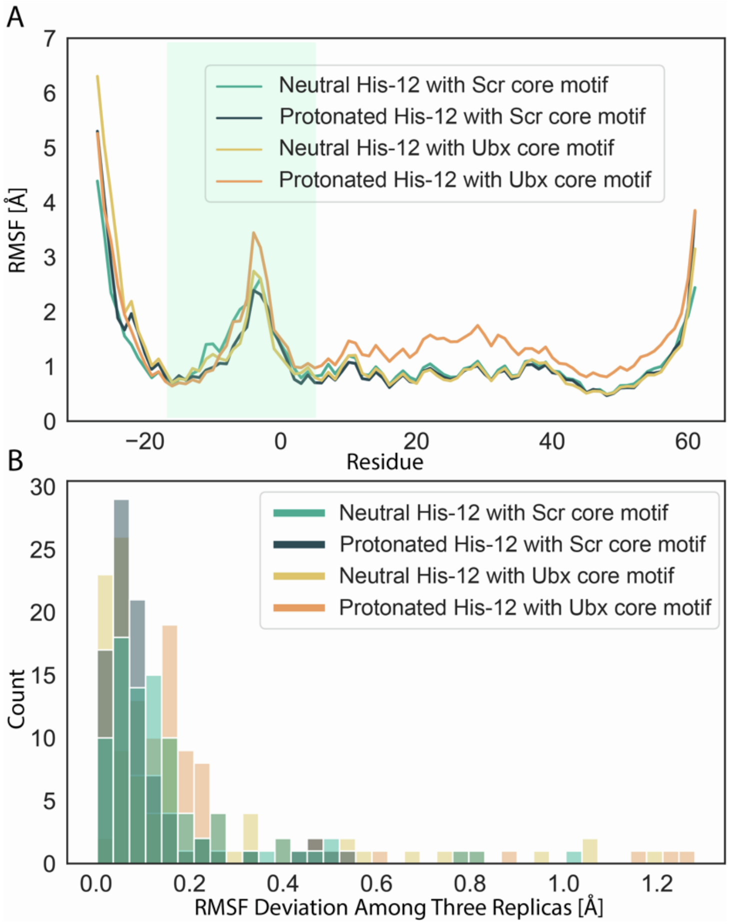

Figure S4. RMSF values for wildtype systems with the Scr- or Ubx-preferred core. (A) RMSF values are shown for systems with Scr target site in complex with either neutral His-12 (dashed green line) or protonated His-12 (solid green line). RMSF values are shown for systems with Ubx preferred target in complex with either neutral His-12 (solid yellow line) or protonated His-12 (solid orange line). Green highlighted region represents the Scr linker region. (B) Histogram of the RMSF deviation among three replicas.

S-V Mechanism of MGW narrowing

We investigated how the MGW narrows by calculating distances from the centers of mass (CMs) of the sugar rings in T_6(**II**) and T_8(**I**) to the imidazole ring of His-12. T_6(**II**) and T_8(**I**) were chosen because they are the closest DNA bases that are directly adjacent to His-12 (Fig. 1A). Distances from T_6(**II**) and T_8(**I**) to His-12 showed a high positive correlation, revealing a synchronized 'pinching' from both backbone strands toward His-12 (Figs. S5A and S5B). This concerted backbone narrowing was also correlated with MGW at the second minimum, with a Pearson correlation of 0.50, in the neutral His-12 system (Fig. S5A). We observed an elevated correlation of 0.70 in the protonated His-12 system (Fig. S5B). The enhanced correlation with MGW and higher stability in calculated distances further validated the groove-narrowing effect of the His-12 protonation. This was also confirmed in other replicas.

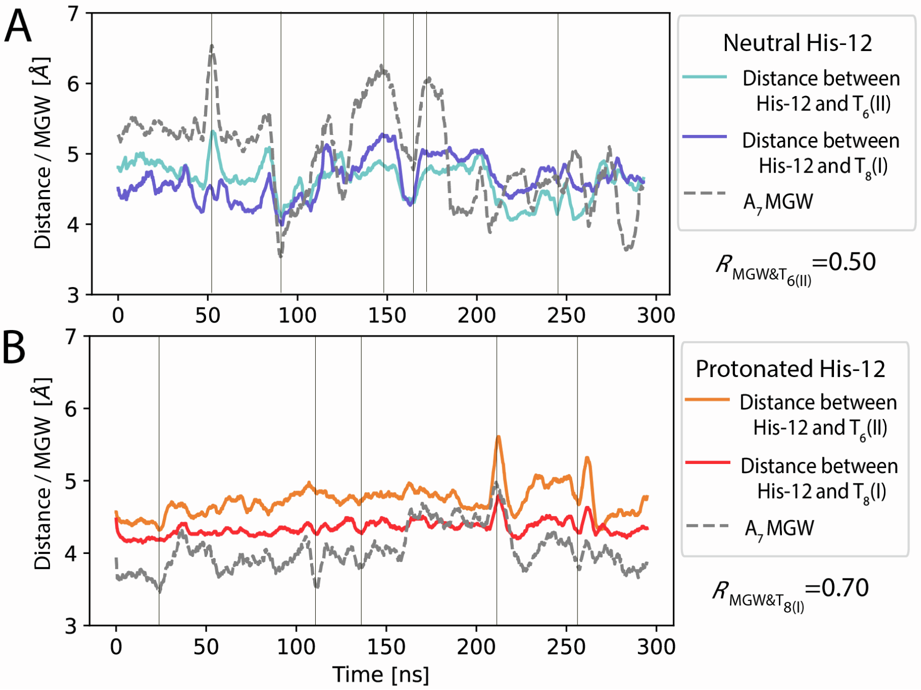

Figure S5. (A) Time-series data for calculated distance between neutral His-12 and T_6(**II**) (light blue) or T_8(**I**) (dark blue) for replica 1. **(B)** Distance between protonated His-12 and T_6(**II**) (orange) or T_8(**I**) (red). In (A) and (B), MGW at the A_7 bp over 300 ns is plotted in a gray dashed line. The Pearson correlation coefficient between the movement of thymine and MGW is shown. Areas with relative movement are indicated by vertical lines. **I** indicates DNA strand 1 and **II** indicates DNA strand 2.

S-VI Binding free energy difference between WT and mutated systems

We calculated binding free energy using the MM/PBSA protocol implemented by g_mmpbsa (5). Briefly, the binding free energy of the protein–DNA complex in solution was calculated by using the relationship of the thermodynamic cycle and separating the calculation of solvation energy (polar and apolar) and the binding energy in a vacuum (molecular mechanics energy) (Fig. S6A). Next, we decomposed the total binding energy to visualize the average contribution of each residue to binding. The energy contribution per residue was calculated with *g_mmpbsa* (with *MmPbSaDecomp.py script*) (5). Residues were colored by their influence on binding from -50 to 50 kJ/mol, with blue, red, or white color indicating a positive, negative, or neutral contribution to binding, respectively (Fig. S6B). His-12 was clearly indispensable to binding of the complex in the protonated His-12 system but had less of an influence in the neutral case. This finding is consistent regardless of the dielectric constant.

To ascertain agreement between the calculated binding free energy and experimental data, we obtained the dissociation constant (K_d) from published data (6). In (6), the authors performed the same mutation experiments in vivo with Exd-Scr heterodimer binding to an Scr-preferred DNA sequence fkh250 and a consensus Hox-Exd site fkh250^{con}. The fkh250 sequence corresponds to our simulation with the Scr site, and fkh250^{con} corresponds to the Ubx site. From the reported dissociation constant K_d, we calculated the free energy, $\Delta G = RTln(K_d)$, for fkh250 and fkh250^{con}, where R is the gas constant and T is temperature.

As both His-12 protonation states exist in vivo, we averaged the binding free energy values between either state in each frame and calculated the mean binding free energy for each mutation system. We then compared the binding free energy values from the MD trajectory and binding free energy results calculated from experimental data in Figs. S6D, S6E, S6G and S6H for WT, Scr$^{H\text{-}12A}$, and Scr^{R3A}. We plotted both the binding free energy results and the MM/PBSA binding free energy results on a dual-axis plot.

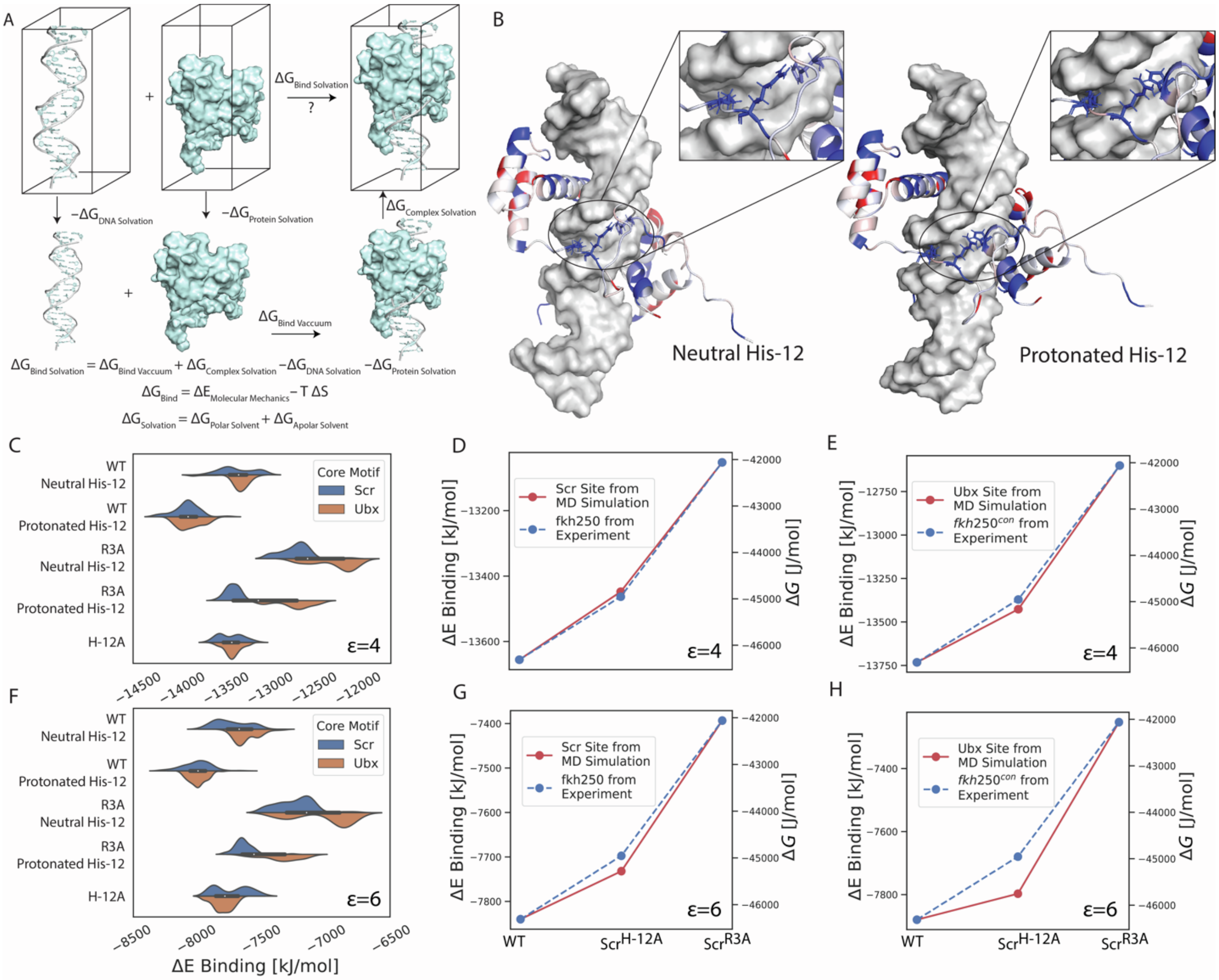

Figure S6. (A) Workflow for MM/PBSA calculation. **(B)** Decomposition of binding free energy of WT system with neutral His-12 (left) and binding free energy of WT system with protonated His-12 (right). Inset: Magnification showing contribution of His-12 to binding free energy. **(C)** and **(F)** Comparison of binding free energy values of WT and mutated systems with Scr-preferred site (blue) and Ubx-preferred site (orange) in both His-12 protonation states, aggregated over three replicas. **(D, E)** and **(G, H)** Comparison between simulated and experimental binding free energy values for Scr and Ubx sites, respectively. (C, D, E) shows binding free energy with dielectric constant 4. (F, G, H) shows binding free energy with dielectric constant 6, as indicated by ϵ.

S-VII Correlation in conformational differences among key residues in both protonation states

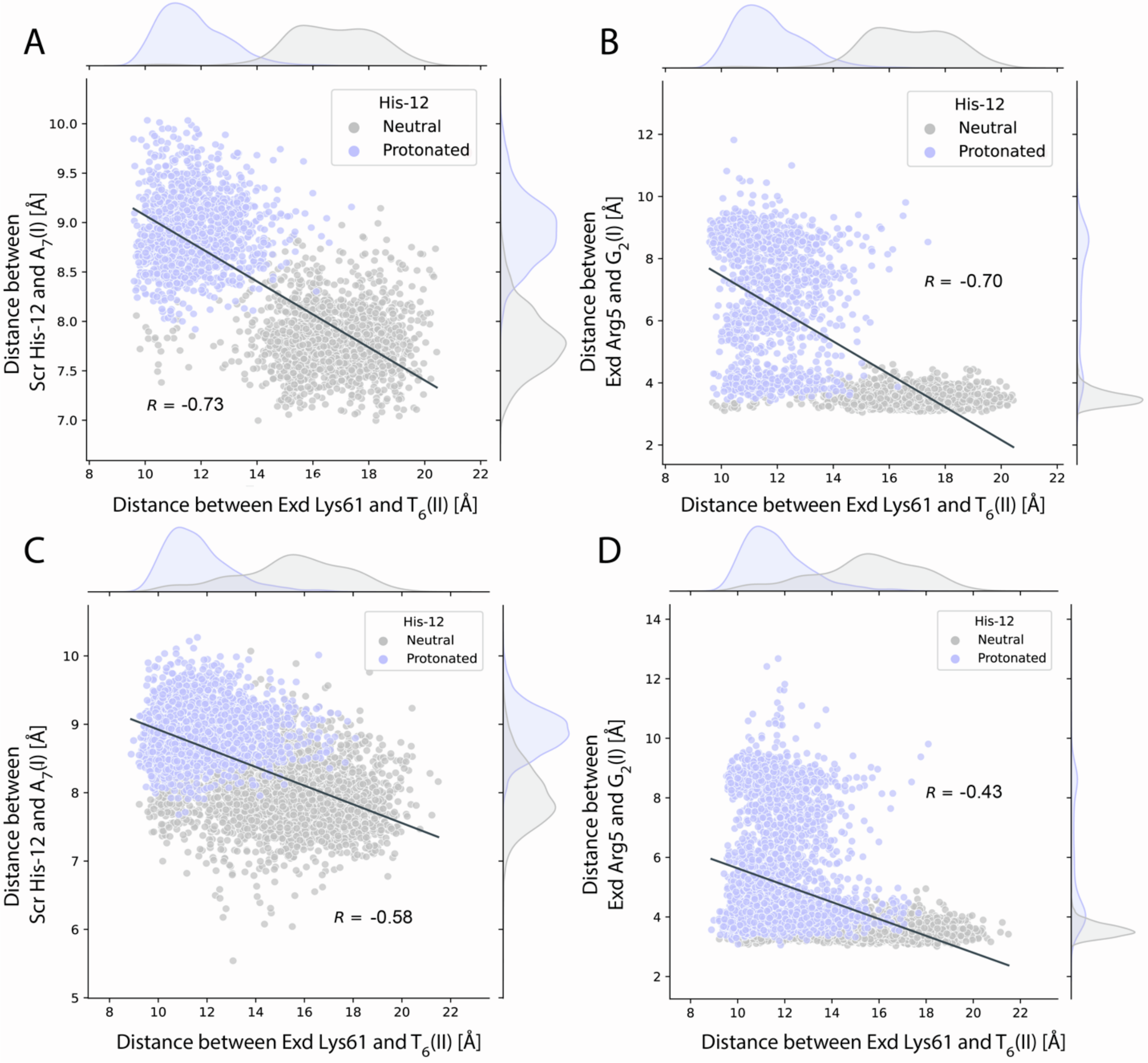

Figure S7. (A, B) Correlation of distance distributions between neutral His-12 (gray) and protonated His-12 (blue) for His-12 and Lys61 residue pairs and Arg5 and Lys61 residue pairs, respectively in replica 1. Fitted line through these distances was calculated, and Pearson correlation values are shown. **(C, D)** Untrimmed distances for (A, B) respectively.

We sought to illustrate the correlation in conformational differences among key residues in both His-12 protonation states. Using frames extracted from the equilibration-stage 150 ns of the MD simulation, we calculated 1) the distance from the NZ atom of Lys61 (the last Nitrogen atom of the lysine side chain) to the CM of the T_6(**II**) sugar, and 2) the distance from the imidazole ring of His-12 to the CM of the A_7(**I**) sugar. We plotted the distance distribution for both the neutral and protonated His-12 states (Fig. S7A). Random motions from thermal fluctuations can occur in MD simulations, which can result in sudden movements and extreme values. To clearly identify the distance relationship between residues at most times, we removed these infrequent occurrences. To do this, we excluded extremely high or extremely low pairwise distances, retaining only pairwise distances between the 1% and 99% quantiles. The untrimmed data are presented in Figs. S7C and S7D.

Distance distributions between Lys61 and T_6(II) (shown in gray) ranged 14-20 Å in the system with neutral His-12. In the protonated His-12 system (blue), the distance was shorter, ranging from 10-14 Å (Fig. S7A). For both systems, the simulation began from the same initial structure, where Lys61 was far from T_6(II). Thus, the distribution captured the rotation of Lys61 toward the T_6(II) backbone in the protonated His-12 system. Similarly, the distance between His-12 and A_7(I) was 7.0-8.5 Å when His-12 was neutral, and was 8-10 Å when His-12 was protonated. Thus, neutral His-12 was closer to A_7(I), whereas protonated His-12 moved further away from it. In the protonated His-12 system, as His-12 moved away from the DNA residue A_7(I), the Lys61 residue rotated towards T_6(II), with a correlation coefficient of -0.73 (Fig. S7A). Protonated His-12 resulted in a difference in positioning compared to neutral His-12, which changed the positioning of the Lys61 residue.

Similarly, the high negative correlation (-0.70) between the distance distribution shift from Arg5 to the G_2(I) pair and from Lys61 to the T_6(II) pair suggests that the displacement of Arg5 away from G_2(I) is related to the positioning between Lys61 and T_6(II) (Fig. S7B). Distances between the carbon atom in the guanidinium group of Arg5 and the CM of G_2(I) are analyzed in Fig. S7B. The guanidinium group of Arg5 was far from G_2(I) and had large fluctuations in the complex with protonated His-12. By contrast, the group was much closer in the neutral His-12 system, where it also showed reduced variability. This correlation was also observed in other replicas.

S-VIII SCR vs. UBX

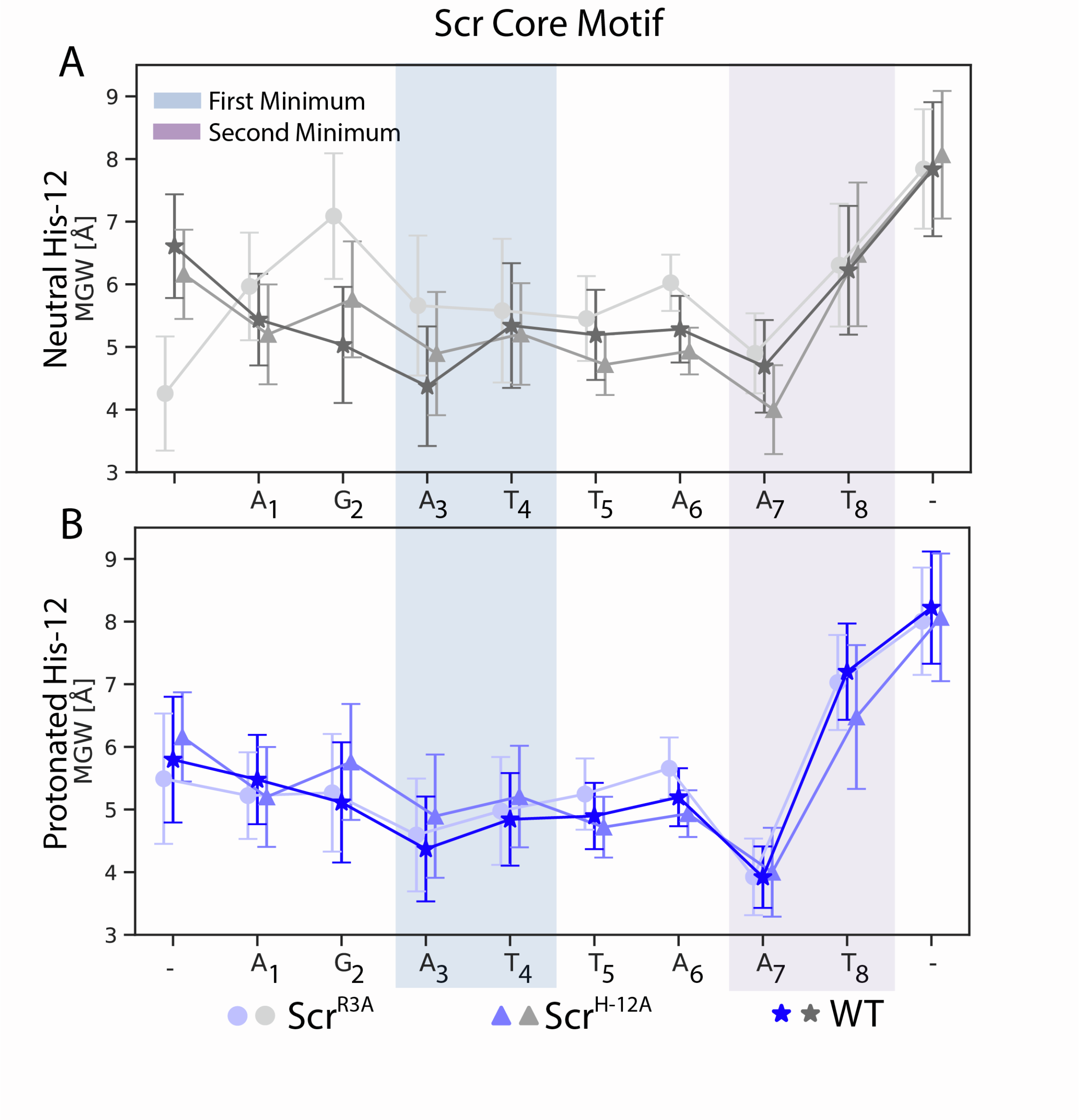

Figure S8. DNA shape profiles for wild-type (WT) and mutated systems with **(A)** neutral His-12 and **(B)** protonated His-12 averaged among three replicas. The error bars represent the pooled standard deviation of MGW from all replicas. Shaded region shows double minima for the Scr-preferred site. Blue and purple shaded regions: First and second minima, respectively. The individual replica results can be found in the provided Figshare link.

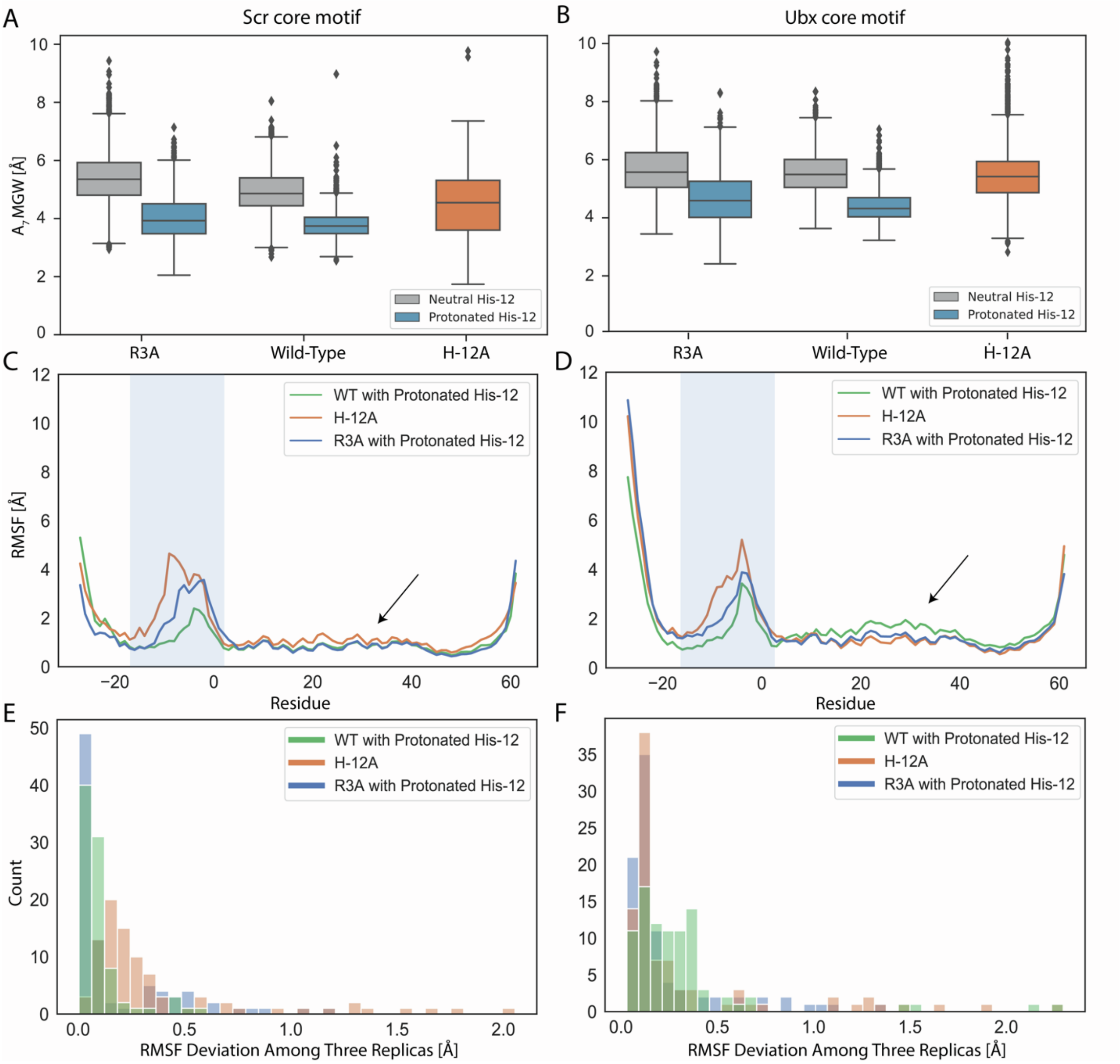

Figure S9. (A) MGW distribution at A_7 bp for systems with Scr-preferred site aggregated from three replicas, with neutral His-12 (gray), protonated His-12 (blue), and Scr$^{H\text{-}12A}$ (orange). **(B)** MGW distribution at A_7 bp for systems with Ubx-preferred site aggregated from three replicas, using same color scheme as in (A). **(C, D)** RMSF for systems with protonated His-12 and Scr$^{H\text{-}12A}$ from systems in (A) and (B), respectively, averaged over three replicas. Blue-shaded regions indicates the linker region, and the black arrow marks regions where Scr has higher fluctuation when bound to a Ubx-preferred site compared to an Scr-preferred site. **(E, F)** Histogram of the RMSF deviation among three replicas for systems in (A) and (B), respectively.

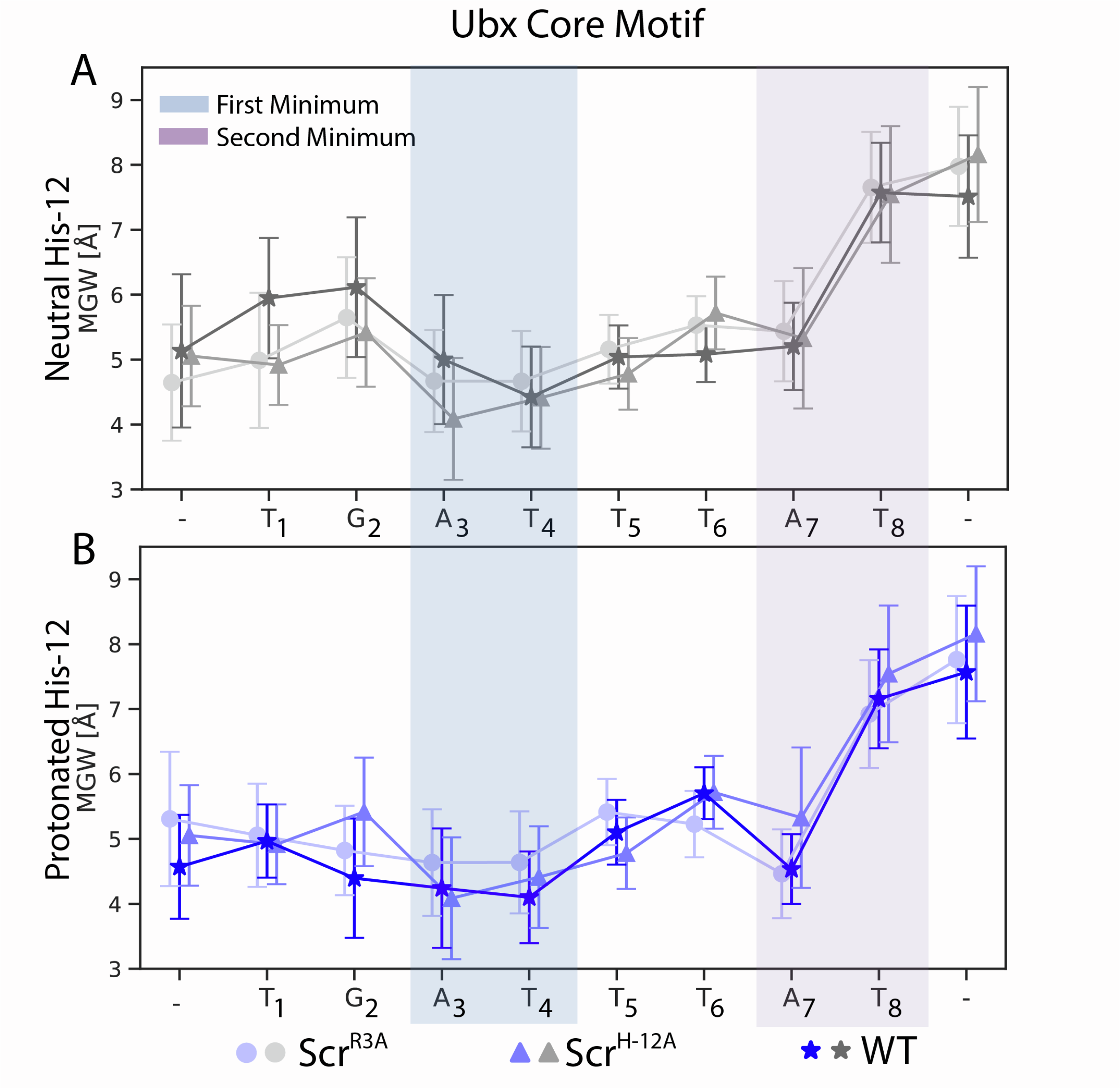

Figure S10. DNA shape profiles for wild-type (WT) and mutated systems with **(A)** neutral His-12 and **(B)** protonated His-12 averaged among three replicas. The error bars represent the pooled standard deviation of MGW from all replicas. Shaded region shows double minima for the Ubx-preferred site. Blue and purple shaded regions indicate first and second MGW minima, respectively.

S-IX Ubx-preferred core motif is more rigid

We analyzed the RMSF for the DNA heavy atoms (Fig. S11). Lower fluctuation of the DNA core residues was observed for the Ubx core compared to the Scr core motif, reflecting higher DNA rigidity of the Ubx core motif. Upon protonation, the DNA fluctuation for both cores was lower, showing that His-12 protonation can stabilize the DNA.

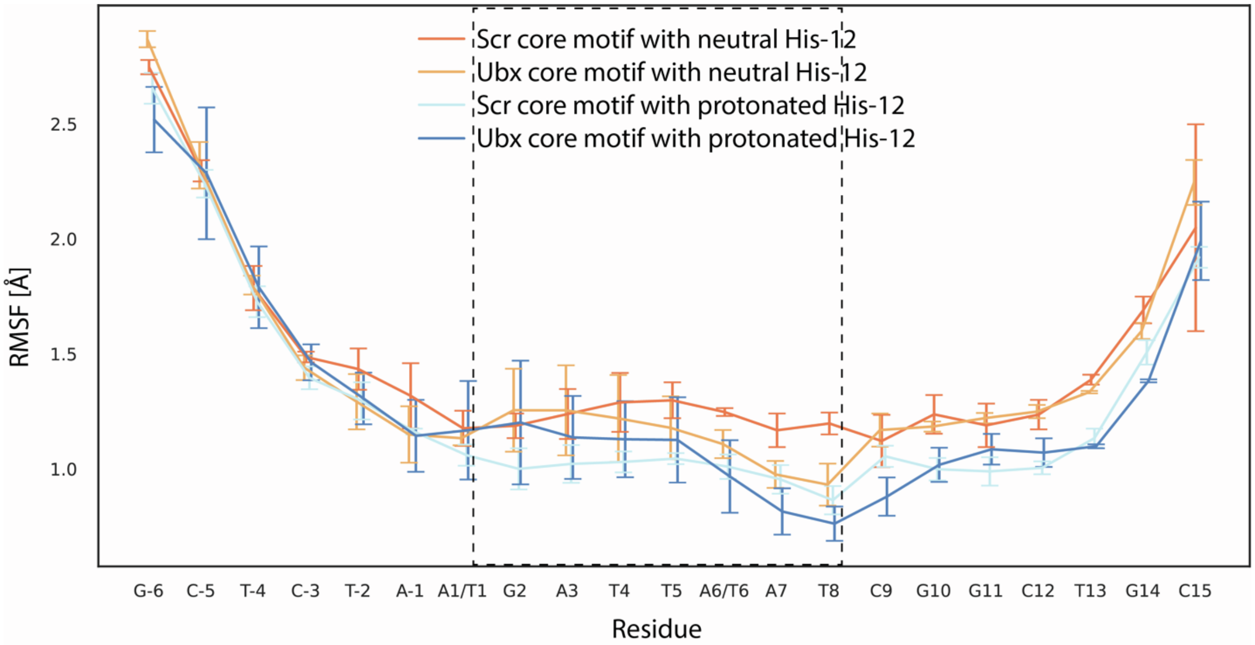

Figure S11. RMSF for DNA heavy atoms for each last 150 ns of the trajectory averaged among three replicas. The Scr core motif has an adenine at the 6th position, whereas the Ubx motif has a thymine at the 6th position. Dashed box indicates the core motif region. The error bars represent the standard deviation among all replicas.

S-X FLANKING SEQUENCE DYNAMICS

Comparing dynamics between different flanking constructs across the 300-ns simulations, we found that all low-affinity sequences in complex with neutral His-12 started with a not particularly narrow groove (4.5-5.0 Å) and remained conformationally stable over the trajectory. Sequences with high or medium-high affinity evolved to a narrower groove, and their MGW stabilized to ~3.5 Å at around 120 ns (Fig. S12A). In the protonated His-12 system, MGW at the A_7 step in high-affinity sequences narrowed to ~3.5 Å, similar to the neutral His-12 system. However, the MGW of low-affinity sequences decreased to 3.5-4.2 Å (Fig. S12B). The evolution of MGW over the trajectory revealed that His-12 protonation influences the MGW minimum in low-affinity sequences but has little effect on high-affinity sequences.

To quantify the "shape" of the second minimum, we considered MGW differences between the neighboring bases of A_7, A_6, and T_8 (Figs. S12C and S12D). This calculation can be described by the formula:

$$\Delta \mathrm{A}_7 MGW = MGW(A_6) + MGW(T_8) - 2 * MGW(A_7).$$

This calculation shows whether the MGW is similarly narrow among the A_6, A_7, and T_8 bases, or whether the A_7 base has a much narrower MGW than the two neighboring bases.

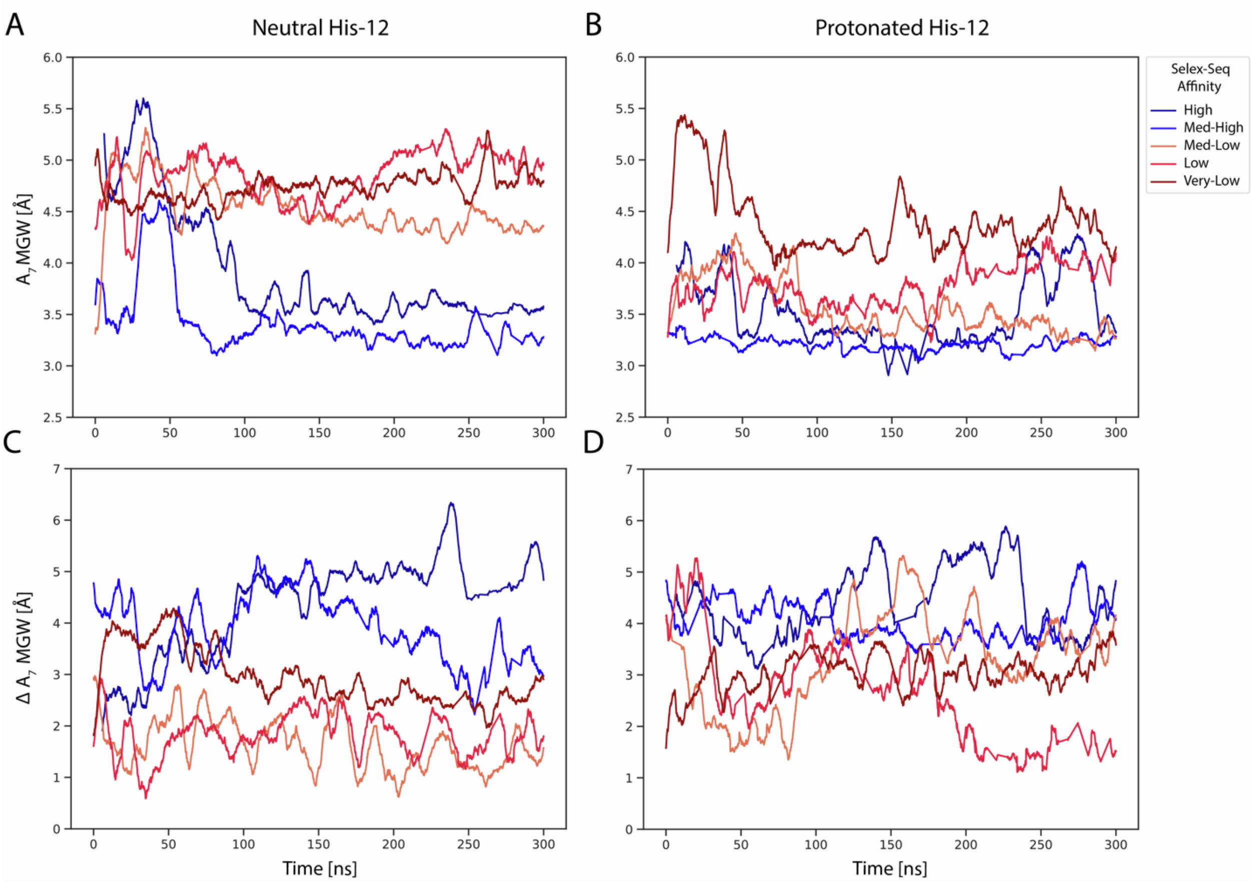

Figure S12. MGW at A_7 nucleotide in system with **(A)** neutral His-12 or **(B)** protonated His-12, plotted over 300 ns. Red and blue indicate lower and higher affinity sequences, respectively. MGW differences calculated between two neighboring bases of A_7 in system with **(C)** neutral His-12 or **(D)** protonated His-12.

S-XI Hydrogen donor-acceptor pairs breakdown

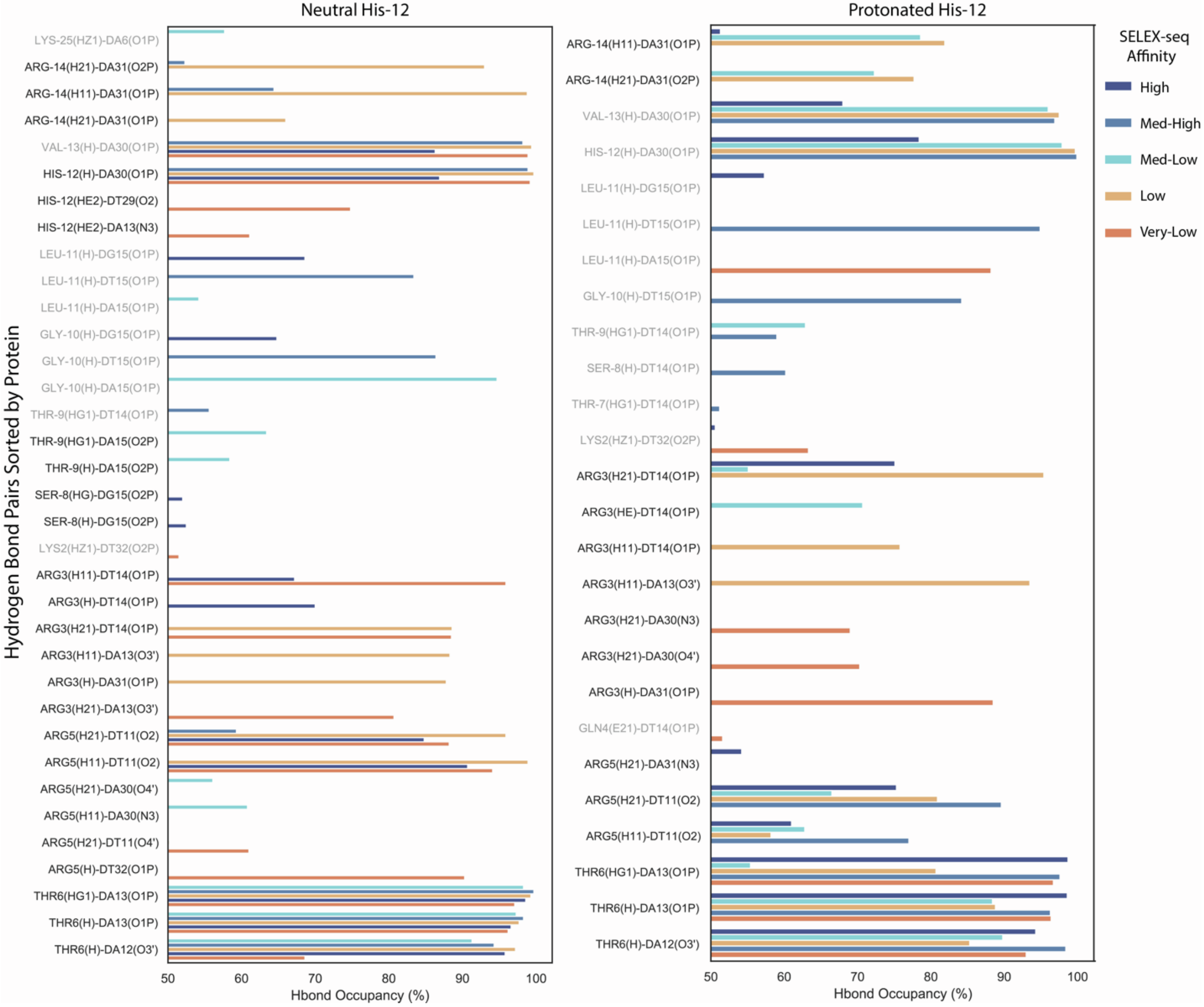

Figure S13. Hydrogen bond pairs with >50% occupancy, broken down by geometric or acceptor-donor species. Single hydrogen bonds (gray font) and bidentate or bifurcated hydrogen bonds (black font) are differentiated. Hydrogen bond pairs are sorted by residue ID of the protein. Colors reflect SELEX-seq-derived affinity of sequences that form the hydrogen bond, with orange and blue indicating lower and higher affinity, respectively.

Supplemental References

1. Cooper,B.H., Chiu,T.P. and Rohs,R. (2022) Top-Down Crawl: a method for the ultra-rapid and motif-free alignment of sequences with associated binding metrics. *Bioinformatics*, **38**, 5121–5123.
2. Haran,T.E. and Mohanty,U. (2009) The unique structure of A-tracts and intrinsic DNA bending. *Q. Rev. Biophys.*, **42**, 41–81.
3. Daura,X., van Gunsteren,W.F. and Mark,A.E. (1999) Folding–unfolding thermodynamics of a β-heptapeptide from equilibrium simulations. *Proteins: Struct. Funct. Bioinf.*, **34**, 269–280.
4. Lyman,E. and Zuckerman,D.M. (2006) Ensemble-based convergence analysis of biomolecular trajectories. *Biophys. J.*, **91**, 164–172.
5. Kumari,R., Kumar,R., Consortium,O.S.D.D. and Lynn,A. (2014) g_mmpbsa□ A GROMACS tool for high-throughput MM-PBSA calculations. *J. Chem. Inf. Model.*, **54**, 1951–1962.
6. Joshi,R., Passner,J.M., Rohs,R., Jain,R., Sosinsky,A., Crickmore,M.A., Jacob,V., Aggarwal,A.K., Honig,B. and Mann,R.S. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell*, **131**, 530–543.