SUPPLEMENTARY INFORMATION

Novel fold and wing structure of Forkhead transcription factor facilitate DNA binding

George L. Wang[1,†], Yibei Jiang[1,†], Yuying Sun[2,†], Fariborz Nasertorabi[3,†], Jesse A. Weller[1,6], Raktim Mitra[1,‡], Alexander Batyuk[4], Oscar M. Aparicio[2], Vadim Cherezov[3,5], and Remo Rohs[1,5,6,7,8,9,*]

[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA

[2]Molecular and Computational Biology Section, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA

[3]Structural Biology Center, Bridge Institute, Michelson Center for Convergent Bioscience, University of Southern California, Los Angeles, CA 90089, USA

[4]Linac Coherent Light Source, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA

[5]Department of Chemistry, University of Southern California, Los Angeles, CA 90089, USA

[6]Department of Physics & Astronomy, University of Southern California, Los Angeles, CA 90089, USA

[7]Thomas Lord Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA

[8]Division of Medical Oncology, Department of Medicine, University of Southern California, Los Angeles, CA 90033, USA

[9]Alfred E. Mann Department of Biomedical Engineering, 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

[†]The first four authors should be regarded as Joint First Authors.

[‡]Present address: Institute for Protein Design, University of Washington, Seattle, WA 98195, USA

Supplementary Section S-I: Supplementary experimental methods and results

Experimental details for in-vivo functional testing

	Sequence
Forward sequence for 19-bp DNA binding oligo	5'-CGAAATGTAAACATACCGC-3'
Reverse sequence for 19-bp DNA binding oligo	5'- GCGGTATGTTTACATTTCG-3'
Forward primer for amplification in yeast experiments	5'-atacgactcactatagggcgaattgggtaccgggccccccCGCCCTTATCAATTATAATATGCGCT-3'
Reverse primer for amplification in yeast experiments	5'- caagcgcgcaattaaccctcactaaagggaacaaaagctgTTGCCGCTAGCAGTCTCTTC-3'

Supplementary Table S1. DNA sequences for DNA binding oligos for Fkh1–DNA complex formation as well as primers for amplification in yeast experiments. For the forward and reverse primers, upper case letters anneal to the targeted DNA while lowercase letters are additions to the primer.

Experimental details for X-ray structure determination

X-ray data collection	
PDB ID	9EFW
Wavelength [Å]	0.97946
Space group	*P1*
Unit cell dimensions	
a, *b*, *c* [Å]	48.16, 50.25, 59.04
α, β, γ [°]	110.8, 100.4, 82.6
Resolution range [Å]	29.51-2.10 (2.17-2.10)[a]
No. of observed reflections	78,431 (7,894)
No. of unique reflections	27,067 (2,687)
Multiplicity	2.9 (2.9)
Completeness [%]	90.47 (90.59)
<I/σI>	10.3 (1.09)
Rmerge [%]	4.2 (105.2)
Rpim [%]	2.9 (73.6)
CC1/2	0.999 (0.649)
Wilson B-factor [Å^2]	56.0
Refinement	
Resolution range [Å]	29.51-2.20 (2.3-2.2)
Reflections used in refinement	23,291 (2,983)
Reflections used for R-free	1,168 (150)
Rwork (%)/ Rfree (%)	18.89 / 23.75
No. of atoms	
Protein	A: 938 B: 950
DNA	C: 386 D: 387 E: 386 F: 387
Ions	G: 2
Water	S: 99
Average B-factor [Å^2]	
Protein	A: 72.5 B: 65.1
DNA	C: 81.3 D: 82.6 E: 74.1 F: 75.4
Ions	G: 57.4
Water	S: 68.5
RMSD	
Bond lengths [Å]	0.006
Bond angles [°]	0.597
Ramachandran statistics [%]	
Favored	99.56
Outliers	0.00
Molprobity score	0.82

[a] Values in parentheses are for the highest-resolution shell.

Supplementary Table S2. Crystallographic statistics of the X-ray structure of the Fkh1-DBD–DNA complex for data collection and refinement.

Flanking sequence constructs used in molecular dynamics (MD) simulations

We selected four flanking regions of various binding affinities located 5' and 3' of the core region. We added 6-bp 5' and 3' G/C caps with a maximized number of intra-bp hydrogen bonds on both oligo end regions to reduce end effects in MD simulations (Supplementary Table S3).

The high-affinity sequence ('High') has the highest relative binding affinity derived from SELEX-seq experiments (1) for the GTAAACA core. The medium-to-high affinity ('Med-High') sequence is also the sequence observed in the co-crystal structure. The low-affinity sequence ('Low') is a weak binder and does not bind to the Fkh1 protein in the SELEX-seq experiment. Because the lowest affinity sequence derived from SELEX-seq experiments has a relative binding affinity of 0.02, the relative binding affinity of the low-affinity sequence ('Low') is denoted as <0.02.

Binding affinity	Relative binding affinity from SELEX-seq	Sequence
Low	<0.02	5' CGCGCG **GGCG** GTAAACA **GG** CCGCGCGC 3'
Med-Low	0.35	5' CGCGCG **CTTA** GTAAACA **CA** CCGCGCGC 3'
Med-High	0.74	5' CGCGCG **AAAT** GTAAACA **TA** CCGCGCGC 3'
High	0.92	5' CGCGCG **AAAT** GTAAACA **TT** CCGCGCGC 3'

Supplementary Table S3. DNA sequences with various binding affinities were used in MD simulations. The 7-bp Fkh1 consensus core motif was used for all sequences. We varied the 4-bp flanking region on the 5' end and 2-bp flanking region on the 3' end. This selection was made based on our previous work studying Fkh1 binding to DNA in SELEX-seq experiments (1), which indicated a different role of the 5' and 3' proximal flanks. The relative binding affinity of each sequence obtained from the SELEX-seq experiments is listed in the second column.

Average DNA curvature of MD constructs used in MD simulations

System	Average bend angle [°] (bound DNA)	Average bend angle [°] (unbound DNA)
WT	28±7	21±10
R401A	30±8	21±10
R400A, R401A	22±7	21±10

Supplementary Table S4. DNA curvature of various MD systems averaged over five replicas. Standard deviation is indicated following the ± sign. Values for unbound DNA are identical because the protein that differentiates the three systems is not present.

Supplementary Section S-II: Supplementary computational methods and results

Protocol for MD simulations

The MD simulation protocol was previously described (2). Simulations utilized GROMACS 2020.3 software running on an NVIDIA Tesla A40/100 GPU, with the AMBER ff14SB (3) force field for proteins and the Parmbsc1 (4) force field for DNA. Each complex was solvated with the explicit TIP3P water model. To neutralize the negative net charge of the Fkh1–DNA complex, Na^+ and Cl^- counterions were added, bringing the NaCl concentration to 150 mM, approximating physiological conditions. The distribution of counterions throughout the simulation box was achieved using the GROMACS 2020.3 *genion* program.

Initially, all systems underwent 2,000 steps of steepest descent energy minimization to ensure proper solvent distribution around the solute. Following this, systems underwent three rounds of NVT equilibration at 300 K, each lasting 10 ps, to gradually raise the temperature. Subsequently, one round of NPT equilibration was performed for 700 ps to equilibrate pressure using a v-rescale thermostat and Parrinello-Rahman barostat with coupling constants τ_T = 0.1 ps and τ_P = 1 ps with a reference temperature of 300 K and reference pressure of 1 bar.

After equilibration, production simulations were performed for 300 ns in the isobaric-isothermal ensemble (*P*=1 bar, *T*=300 K) using a 2-fs integration time step. The Verlet cutoff scheme was employed for all calculations, with long-range electrostatic interactions computed using the Particle Mesh Ewald method (5) with a 12 Å cutoff. Nonbonded van der Waals interactions were calculated with a 12 Å cutoff. The LINCS (6) algorithm was used to constrain all bonds throughout simulations. To assess simulation convergence, we employed the GROMOS algorithm (7, 8) implemented in GROMACS (*cluster* package), tracking structural clusters over time (Supplementary Figure S1). Convergence was observed after around 200 ns, and each system was simulated with five replicas using distinct random seeds.

Root-mean-square fluctuation (RMSF) calculation

RMSF values were calculated using the GROMACS 2020.3 *rmsf* program. RMSF was computed for C_α atoms of protein residues for the last 100 ns of the trajectory. A least-squares superposition to the equilibrated structure (after NVT and NPT) was performed before computing RMSF.

DNA curvature calculation

To calculate DNA curvature, we first identified the center of mass (COM) of the 5'-end nucleotide pair and then calculated a vector connecting it to the COM of the central nucleotide pair, $\overrightarrow{v_1}$. Next, we obtained the vector formed by the 3'-end nucleotide pair to the center of DNA, $\overrightarrow{v_2}$. The curvature was derived by subtracting 180° from the angle between $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$.

Mutation of R400 and R401 residues

To study the contribution of residues R400 and R401 in Fkh1–DNA binding (Supplementary Figure S19A), we conducted 300 ns MD simulations for the double-mutant system (R400A, R401A). Given the role of these arginine residues in modulating MGW, we compared MGW values across these systems. Mutations at the Wing 2 contact region led to wider and less stable MGW profiles (Figures 4B and Supplementary Figure S22A). In the Wing 1 region, we observed a slight increase in MGW in both mutant systems compared to WT. The DNA RMSF analysis highlights that residue K373 in Wing 1 can make the DNA of the mutant systems more constrained when interacting with the 3'-end flanking regions, which may indicate more stable binding in Wing 1 despite mutations (Supplementary Figure S23).

After analyzing the most representative structures of the WT, single-mutant, and double-mutant systems, we observed various degrees of DNA curvature (Supplementary Figure S24). We hypothesized that this bending could be due to interactions with the R400 and R401 residues. After mutation of these arginine residues, we calculated and compared DNA curvature among the WT and mutant systems. We first

compared the bound DNA curvatures with those of free DNA. The result shows that protein binding results in a more curved DNA in the studied systems. The two arginine residues can be important in DNA bending to facilitate DNA recognition. Interestingly, the R401A system shows more bending (Supplementary Table S4). This can be attributed to R400 compensation or its role in stabilizing the DNA backbone in the R401A system (Supplementary Figure S19), as the average DNA bend angle among replicas 1–3 is ~5.1° greater than that of replicas 4 and 5.

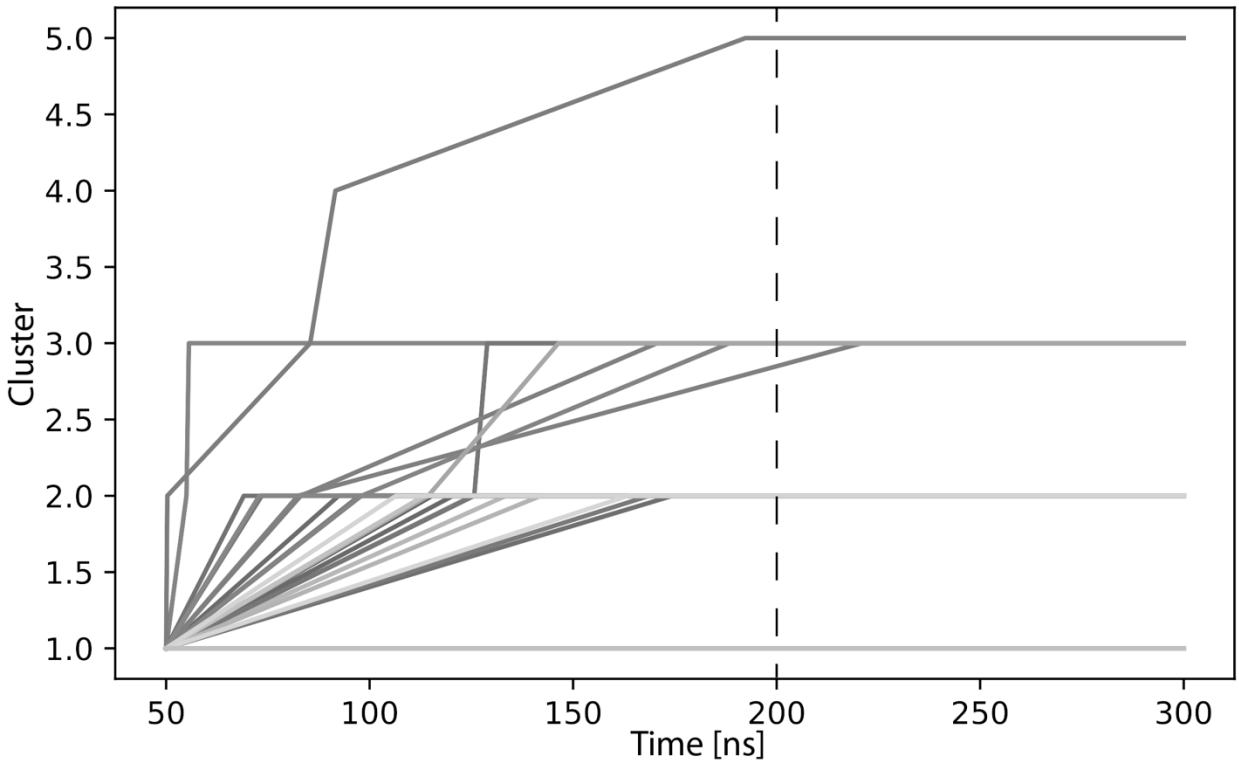
Our simulations revealed that the interaction between DNA and positively charged amino acids stabilizes MG narrowing, accompanied by a correlated bending of the DNA towards the positively charged residue in a complex (Supplementary Section S-XIV and Supplementary Figure S24). This indicates that an electrostatic mechanism can play a crucial role in enabling Fkh1 to recognize and bind its DNA targets. Furthermore, in the AlphaFold 3 (9) (AF3) -predicted structure (Figure 4D), we observed a loss of contact in the Wing 2 region compared to the other two systems. These results underscore the critical role of the R400 and R401 residues in stabilizing the entire complex.

Conformational flexibility due to mutation

To investigate effects of the aa 417–423 region on Wing 2 of Fkh1, we used BioEmu (10) to generate an ensemble of protein structures, which represent conformational flexibility. Unlike AF3 (9), which is not very sensitive to small changes, BioEmu can capture a wide variety of possible structures and can identify conformational changes of key residues. BioEmu predictions of the WT sequence used in the crystal structure demonstrated a high degree of consistency and stability. Mutations of I419 and P420 to alanine resulted in a significant increase in the variability of Wing 2 structures in the ensemble, while the core domain retained a consistent conformation, indicating the importance of the side chain interactions of I419 and P420. C-terminal deletion of residues 419–421 resulted in an even larger conformational variability, likely due to the loss of interactions with the protein backbone of P420. This result suggests that these residues are important for Wing 2 stability. To verify that these changes are not an artifact of introducing mutations, the adjacent G417 and E418 residues were mutated into alanine in separate BioEmu analyses. The resulting protein ensemble produced by BioEmu had variability comparable to the crystal structure (Supplementary Figure S26). This result seems to suggest that I419 and P420 are critical residues in the aa 417–423 region (10), and that the difference from previously observed Fkh1 function and locus binding efficiency may be due to the change in Wing 2 stability.

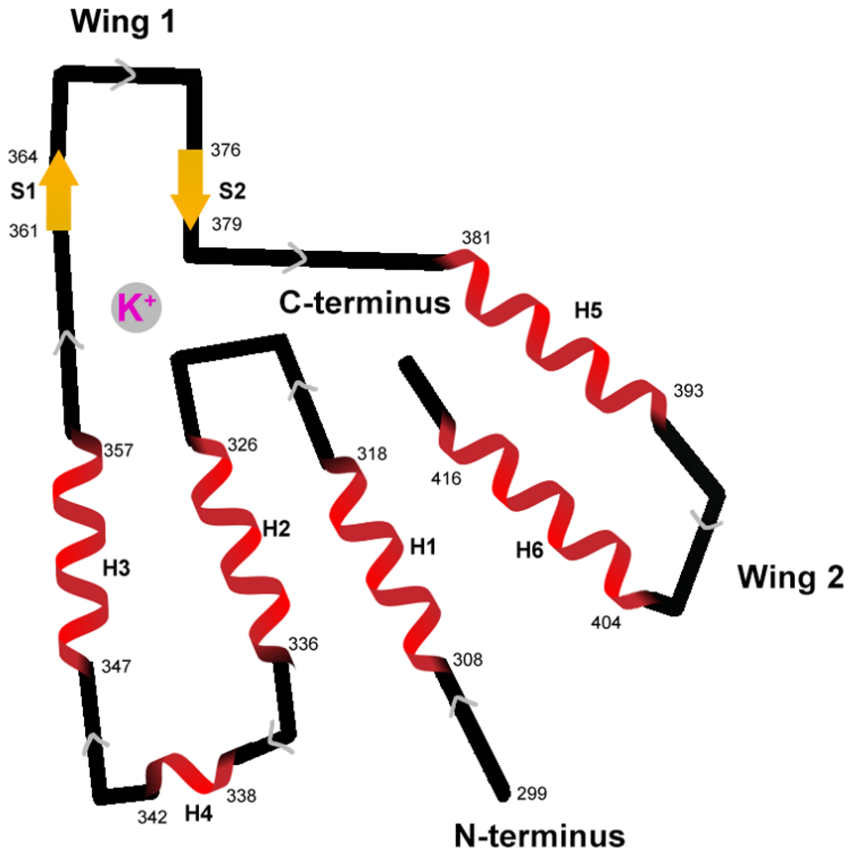
MD simulation convergence evaluation

To assess MD simulation convergence, we clustered simulations using the GROMACS package (11) with a root-mean-squared deviation (RMSD) cutoff. RMSD calculations were based on heavy atoms of Fkh1–DNA structures. To evaluate convergence, we divided 300-ns trajectories into 30,000 frames, spaced by 10 ps. We chose a small RMSD cutoff of 1.0 Å because an even smaller cutoff cannot distinguish structures effectively, and a larger cutoff would cause clusters to converge quickly. We determined that simulations converged when no new clusters were identified. Analysis showed that the rate of new cluster discovery approached zero after 200 ns for most systems. Therefore, we concluded that the simulations reached convergence after 200 ns, where the structural dynamics of Fkh1–DNA complexes stabilized.



Supplementary Figure S1. Number of structural clusters as a function of time using an RMSD cutoff of 1.0 Å for all simulated systems. Dashed line indicates the 200-ns time mark.

Supplementary Section S-II. Fkh1-DBD schematic fold diagram



Supplementary Figure S2. Schematic fold diagram of Fkh1-DBD (aa 299–421). While the core structure (aa 299–379) is similar to DNA binding domains of other Forkhead proteins, the H5 and H6 helices are unique to the protein fold first described in this study for Fkh1 and have not been observed for other Forkhead proteins with known structure. Helices H5 and H6 wrap around and back onto the core structure by interacting with Helix H1, stabilizing the Wing 2 loop.

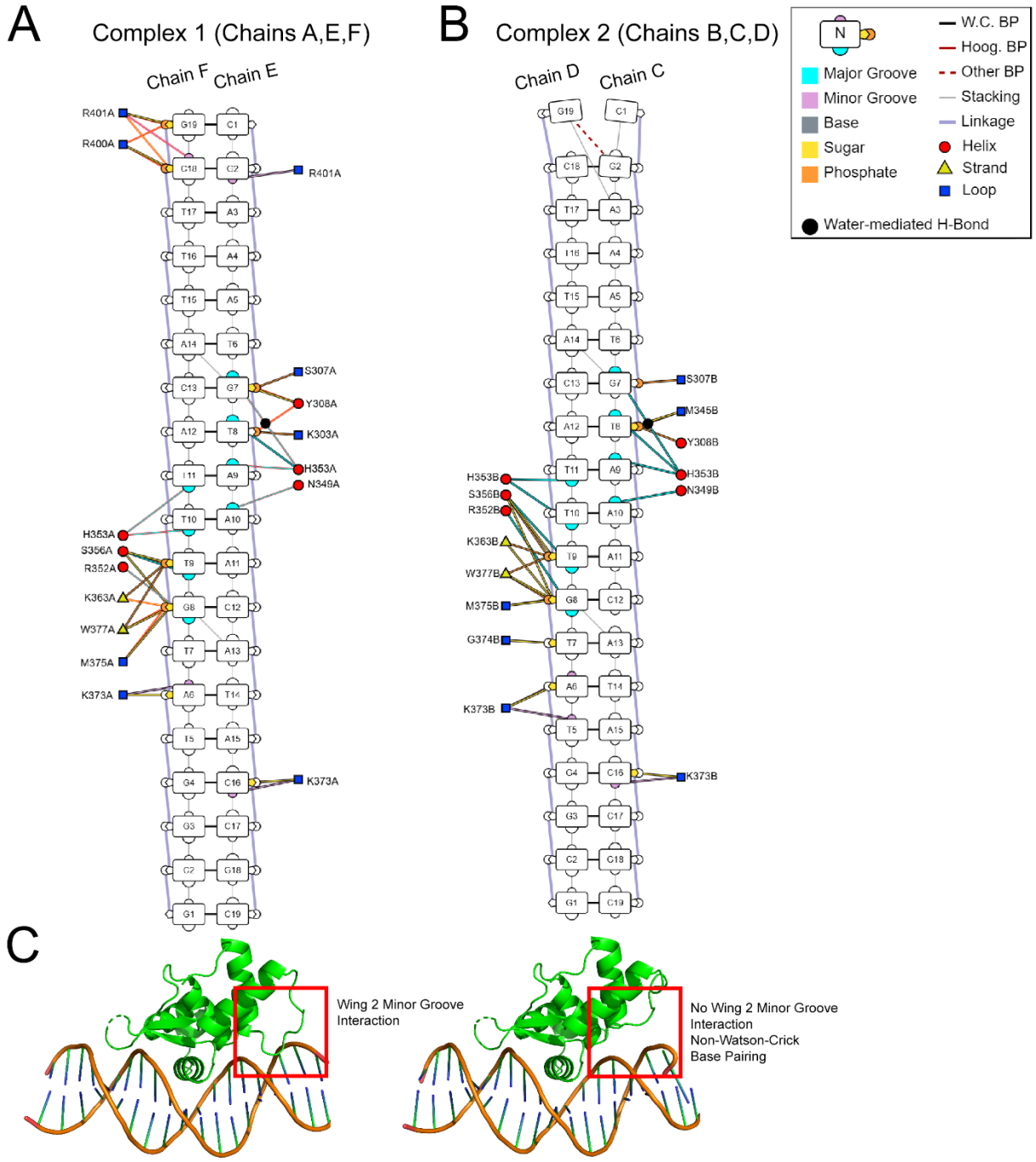
Supplementary Section S-III. Differences between Complex 1 (Chains AEF) and Complex 2 (Chains BCD) in asymmetric unit

The two Fkh1-DBD–DNA complexes in the asymmetric unit have similar structures, with an RMSD of 0.61 Å. Structural differences in the DNA and protein–DNA interactions between the complexes contribute to the difference in RMSD. While Complex 1 described by chains AEF includes a fully canonical 19-bp DNA oligo, Complex 2 described by chains BCD contains a non-Watson-Crick trans sugar edge/sugar edge base pair between base G19 on chain C and base G2' on chain D, as visualized using DNAproDB (12) in Supplementary Figure S3. This difference in DNA corresponds to a difference in binding interactions between the two reported Fkh1-DBD–DNA complexes, causing Wing 2 of Complex 2 to move away from the minor groove (MG) and lose contact. There are a few interactions between the complexes, mainly consisting of protein interactions with the DNA backbone of the other complex (Supplementary Figure S4), that are not atypical for crystal packing contacts.

To explore this difference, we conducted 300-ns MD simulations starting from the co-crystal structures of each complex as shown in Supplementary Figure S5A. We first assessed protein structural similarities in the wing region using Ramachandran maps, revealing comparable folding patterns (Supplementary Figure S5B, upper panel) with differences mostly in two alanine residues (Supplementary Figure S5B, center panel) and the R401 residue (Supplementary Figure S5B, lower panel). Additionally, RMSF analysis indicated the presence of minor differences, particularly in the positions of arginine residues R400 and R401 in the Wing 2 region (Supplementary Figures S5A and S6). These minor differences were potentially influenced by interactions of the arginine residues with DNA, as previously discussed. To analyze the influence of DNA, we examined DNA structural properties by analyzing DNA shape profiles for each complex. Notably, the two complexes exhibited similar mean DNA shape values in the core and Wing 1 regions (Supplementary Figure S5D). However, a notable difference was observed in the Wing 2 region (gray bar in Supplementary Figure S5D); specifically, Complex 1 displayed a narrower MG at the A4 nucleotide compared to Complex 2. This discrepancy correlates with the interactions involving the R400 and R401 residues. In Complex 1, the distance distribution between R401 and the A4 base centered around 5.5 Å; whereas in Complex 2, the distance between R400 and the A4 base centered around 7.3 Å and occasionally exceeded 10 Å (Supplementary Figure S5C), suggesting a lack of direct arginine contact at this position leading to a wider MG. However, this lack of contact could be due to the non-Watson-Crick base pairing due to crystal packing. The spatial arrangement of the R400 and R401 residues is depicted in Figure 4A. The distance between the A4 base to either R400 (Complex 1) or R401 (Complex 2) is more than 10 Å away and will have little influence on MG width (MGW) at this position.

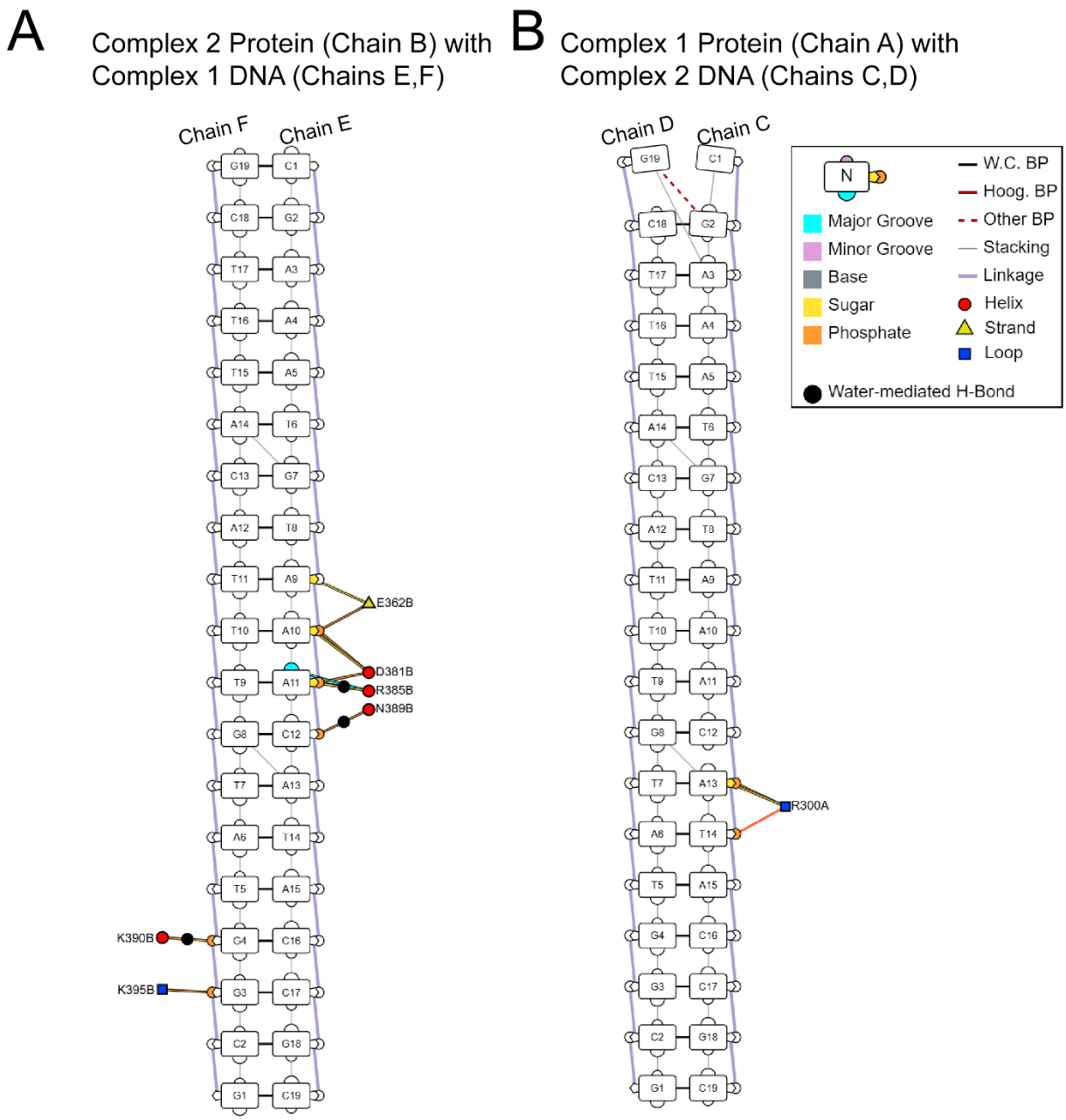
Evidence from the co-crystal structure and MD simulations suggests that in general, the two complexes observed in the asymmetric unit are highly similar. Differences in protein–DNA interactions seem to be caused by crystallization conditions, crystal packing interactions, and the finite length of the DNA oligo used, which do not occur naturally. The Wing 2 interactions with Fkh1-DBD in Complex 2 are influenced by non-Watson-Crick base pairing at one end region of the DNA oligo, which would not occur when bound to a naturally occurring genomic sequence. The asymmetric unit containing two complexes seems to occur only in the context of crystallization, given that size-exclusion chromatography (SEC) and a Native-PAGE gel did not reveal any dimer formation (Supplementary Figures S7 and S8). Therefore, in this study, we will further describe the DNA binding interactions of Fkh1-DBD–DNA using Complex 1, which characterizes the protein–DNA interactions in a more realistic context.

Additionally, an analysis of the residues participating in crystal packing contacts showed that although individual protein–DNA complexes in the asymmetric unit are stable in solution, the dimer of two protein–DNA complexes found in the asymmetric unit is unstable (Supplementary Figure S9 and S10). Looking into the crystal packing diagram, it appears that the non-Watson-Crick base pairing in Complex 2 seems to optimize end-to-end stacking with the DNA oligo of Complex 1 in the adjacent asymmetric unit cell. There are two potential hydrogen bonds as well as additional weaker interactions. It is conceivable that this crystal packing interaction across asymmetric unit cells could contribute to the conformational variance between the two complexes (Supplementary Figures S11A and S11B)

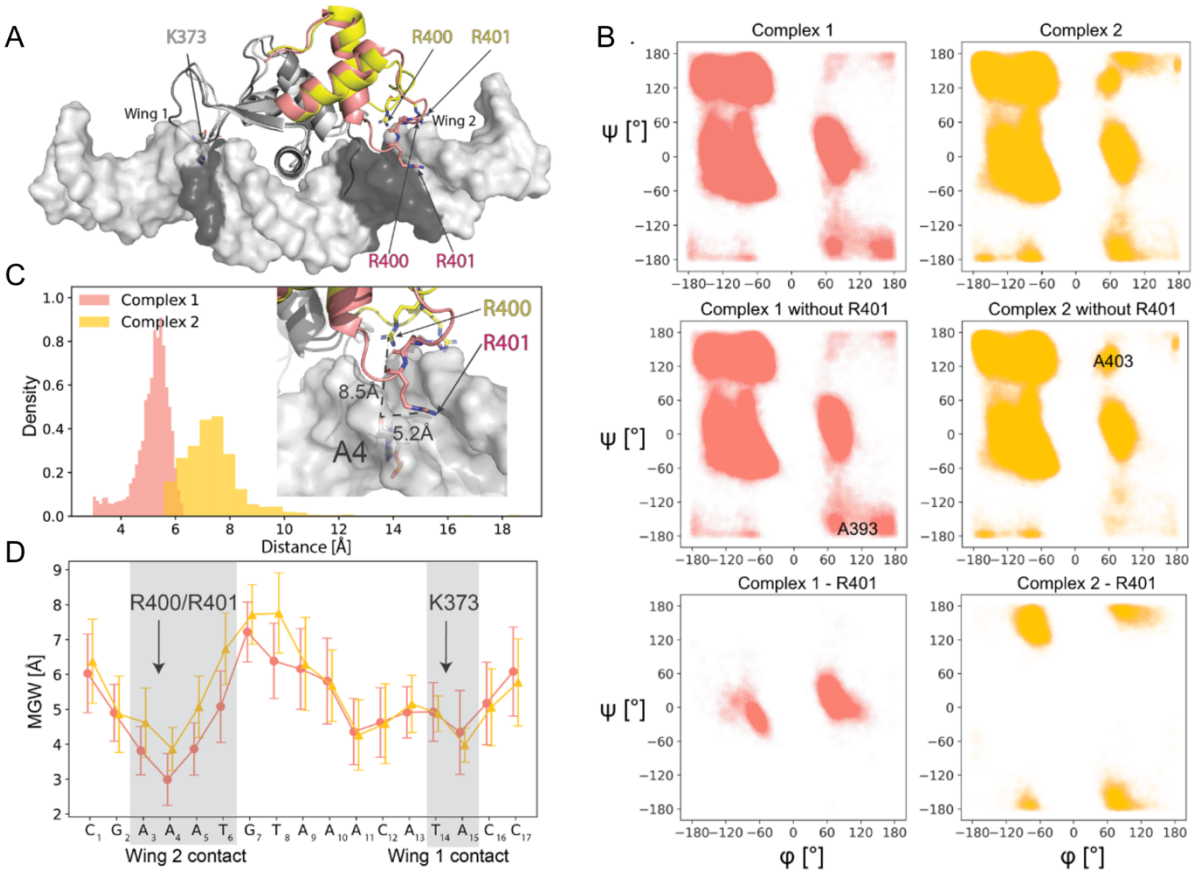


Supplementary Figure S3. Structural analysis of Fkh1–DNA binding sites for both Complex 1 and Complex 2 of the asymmetric unit cell of the co-crystal structure. Contact plot was generated using DNAproDB (12). Hydrogen bonds are highlighted in red. **(A)** Schematic diagram of protein–DNA interactions in Fkh1-DBD Complex 1 with DNA, illustrating hydrogen bonds, van der Waals interactions, and water-mediated hydrogen bonds with both the DNA major groove and MG, as well as backbone interactions with sugar and phosphate moieties. **(B)** Schematic diagram of protein–DNA interactions in Fkh1-DBD Complex 2 with the same DNA sequence. A notable difference between Complex 2 and Complex 1 is that Complex 2 has a non-Watson-Crick base pair between bases G2 and G19', causing a change in the DNA structure at that location. Residues R400 and R401 no longer form interactions with

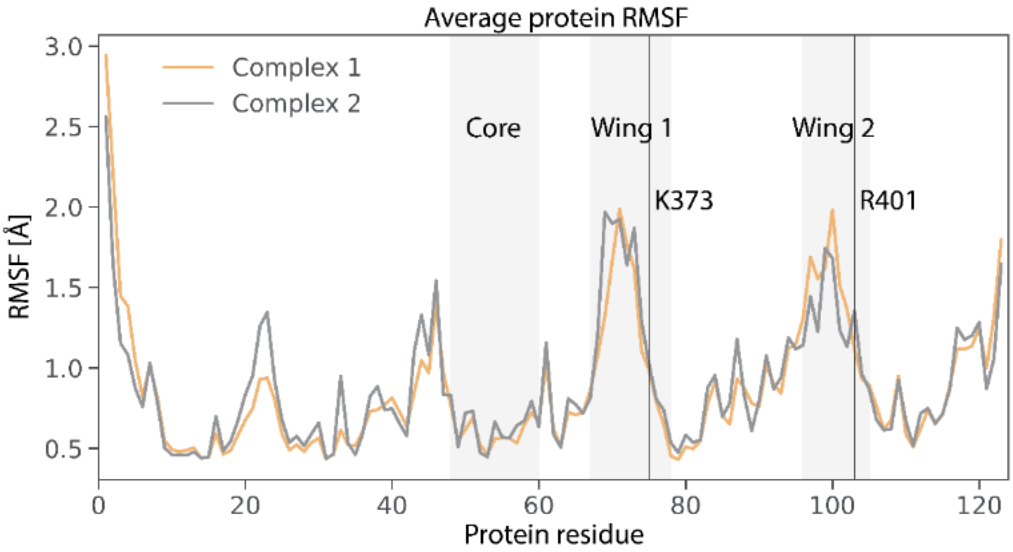
DNA in this end region. **(C)** Comparison between the overall structures of Complex 1 and Complex 2. Because of the non-Watson-Crick base pairing in Complex 2, there is a change in the local structure of the DNA, as well as a change in the interaction of Wing 2 with the nearby MG. Complex 2 lacks the MG interaction with R400 and R401 that is observed in Complex 1.



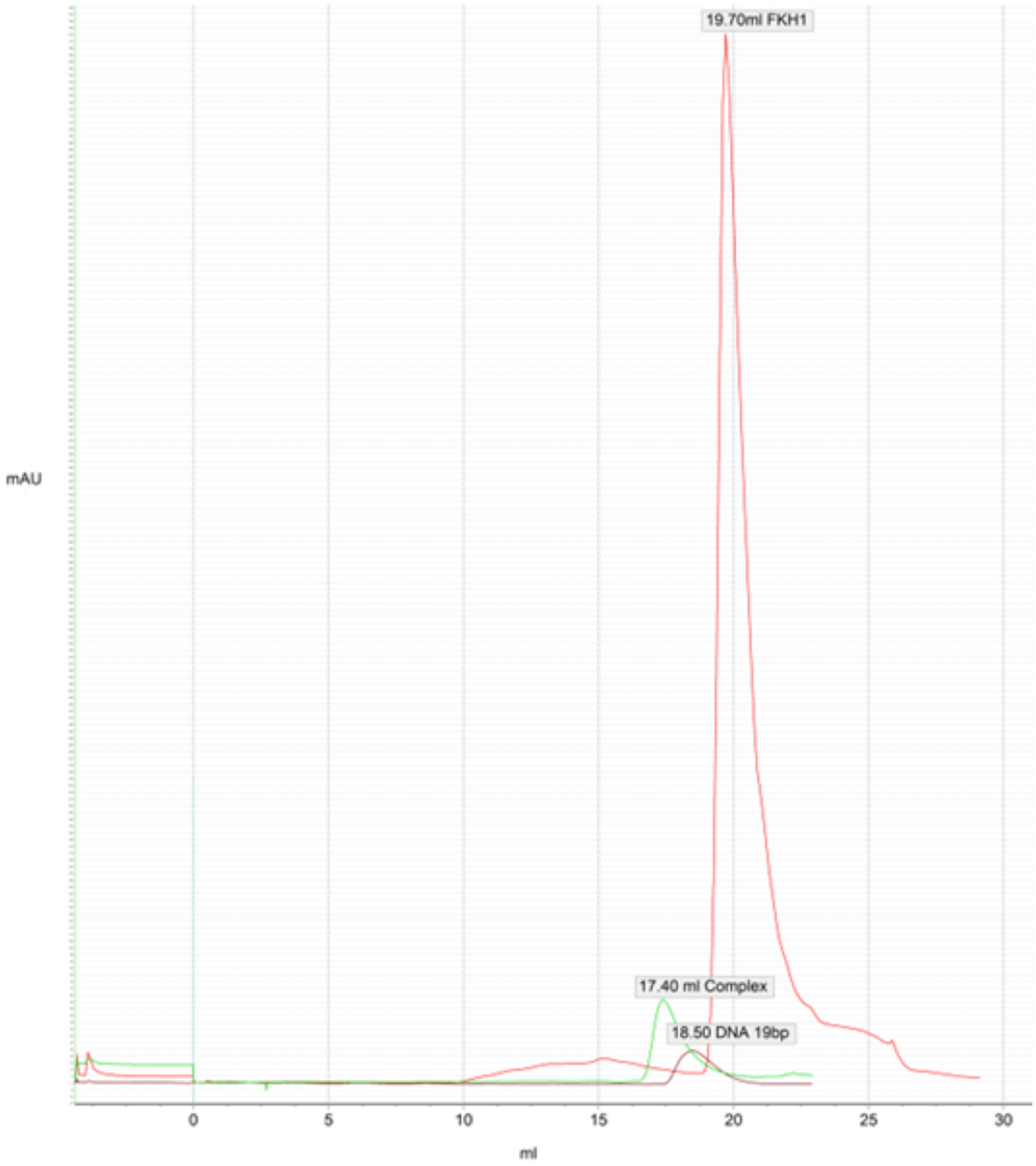
Supplementary Figure S4. Structural analysis of cross-complex interactions of protein and DNA between Complex 1 and Complex 2 of the asymmetric unit cell of the co-crystal structure. Contact plot was generated using DNAproDB (12). Interactions between the two complexes are overall minimal and limited to the DNA backbone. It is unlikely that Fkh1-DBD–DNA forms a dimer in solution, as observed from SDS-PAGE and chromatography analysis obtained during the preparation process (Supplementary Figure S8).



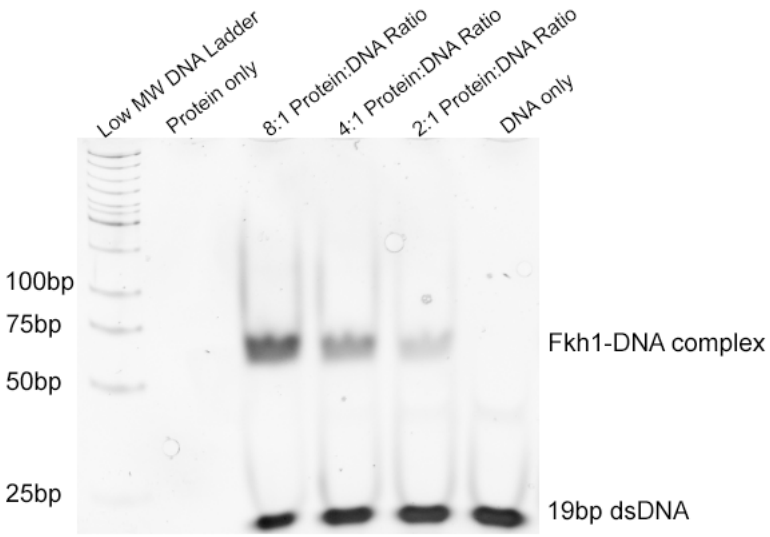
Supplementary Figure S5. Two complexes reveal similar global structural properties. **(A)** Structural overview of Complex 1 (salmon color) and Complex 2 (yellow) obtained from the mean structure of MD simulations aligned by their Wing 1 and core domain regions (gray). The R400 and R401 residues are shown as sticks. **(B)** Ramachandran maps for Complex 1 and Complex 2 (upper panel). Ramachandran plots for Complex 1 and Complex 2 without the respective R401 residues (center panel). Residues resulting in (ψ, ϕ) regions that are apparently different are labeled. Ramachandran plots for the R401 residue in each complex (lower panel). **(C)** Distance distribution of MG-interacting arginine residue in Wing 2 regions of the two complexes. Specific distance measured is illustrated in the zoomed-in structure from the mean structure shown in panel A. **(D)** Minor groove width (MGW) as function of nucleotide position to compare DNA shape profiles. Wing regions are labeled in gray.



Supplementary Figure S6. Protein root-mean-square fluctuation (RMSF) for Fkh1 protein in Complex 1 (orange color) and Complex 2 (gray). Core, Wing 1, and Wing 2 regions are shaded in light gray. Vertical lines indicate positions of the K373 and R401 residues.



Supplementary Figure S7. Size-exclusion chromatography (SEC) chromatogram demonstrating that Fkh1 is monomeric in solution. According to the BioRad Gel Filtration Standard (Catalog # 1511901), the expected size of the Fkh1-DBD–DNA complex should be visible at ~17.4 ml. We observe the peak associated with Fkh1-DBD–DNA at ~17.4 ml, confirming the monomeric state of the complex. According to the reference, a dimer of this complex (~60 kDa) would have appeared at ~16.0 ml, which was not observed, indicating that no dimer of the Fkh1-DBD–DNA complex was present.



Supplementary Figure S8. Native PAGE gel showing various protein:DNA ratios used to test binding prior to SEC purification. This was an 8% Native PAGE gel run at 4°C, 120 V for 60 minutes, after letting the binding proceed at room temperature for 60 minutes. In this gel, we evaluated several protein:DNA ratios (using 8:1, 4:1, 2:1 ratios) for formation of the Fkh1-DBD–DNA complex in two different buffer conditions. For reference, a 5:1 protein:DNA ratio was used before SEC. More binding occurred at higher protein:DNA ratios for both buffers. Only a monomer complex was observed, and no dimer complex was visible in this gel.

The following quaternary structures appear to be stable in solution

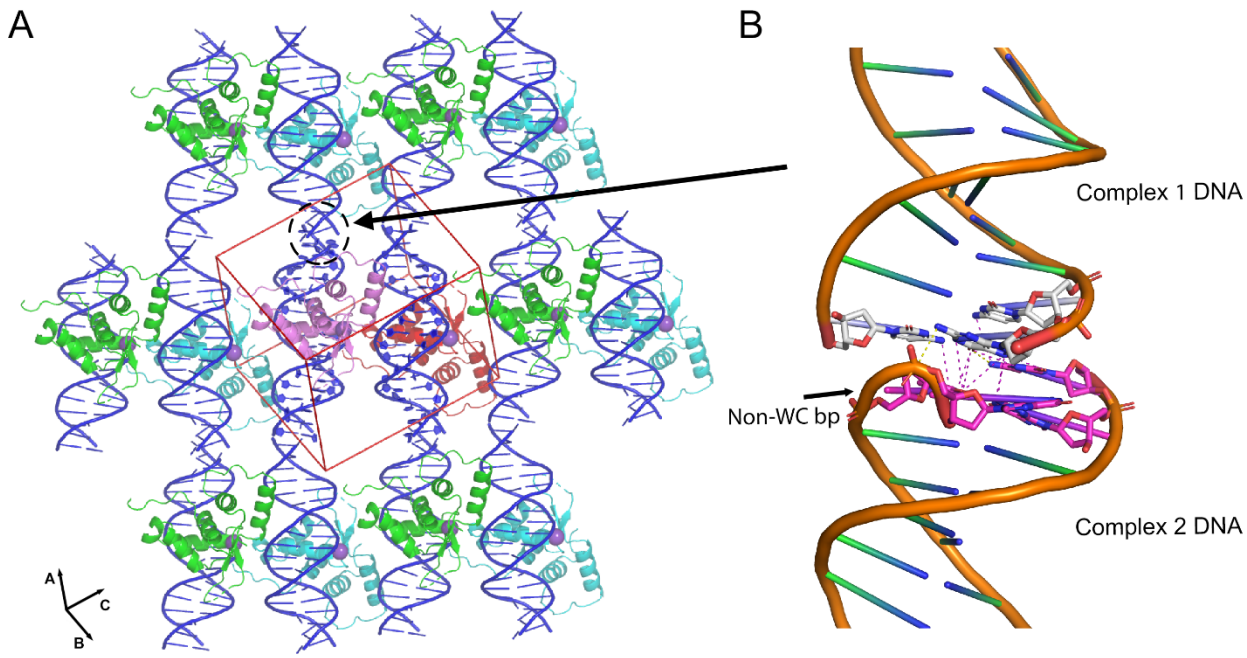
	Split No.	Size	Type	ASA	BSA	dG_diss	dG0	Formula	Composition
1	1	3	1	12260.3	4235.0	14.1	11.9	ABC	AFE
2		3	1	12288.7	3895.4	13.7	9.9	ABC	BDC
3	2	2	2	7108.8	1721.1	21.7	10.8	BC	FE
4		2	2	6975.8	1898.5	16.0	8.0	BC	DC
5		1	3	7309.8	0.0	0.0	0.0	A	B
6		1	3	7665.4	0.0	0.0	0.0	A	A

Supplementary Figure S9. Assembly analysis using QtPISA from ccp4 (13). PISA predicted that the Complex 1 (chains AEF) and Complex 2 (chains BCD) are stable in solution, while the dimer (chains AEFBCD) is not stable in solution.

Residue contacts, 40 residues						
residue1	ncs1	residue2	ncs2	sym2	cell2	distance
ARG 300 A	0	ILE 399 B	0	0	-1 0 0	3
TYR 301 A	0	SER 397 B	0	0	-1 0 0	3.27
TYR 301 A	0	LYS 398 B	0	0	-1 0 0	3.45
ILE 302 A	0	SER 397 B	0	0	-1 0 0	2.77
PRO 305 A	0	TRP 391 B	0	0	-1 0 0	3.1
GLN 306 A	0	ALA 393 B	0	0	-1 0 0	3.17
LYS 390 A	0	DG 4 D	0	0	0 1 1	2.93
TRP 391 A	0	ARG 300 B	0	0	-1 1 1	2.75
GLY 394 A	0	ARG 300 B	0	0	-1 1 1	3.32
LYS 395 A	0	TYR 301 B	0	0	-1 1 1	3.24
LYS 415 A	0	DG 3 F	0	0	-1 0 0	3.21
PHE 416 A	0	DG 1 F	0	0	-1 0 0	3.43
ASN 299 B	0	DA 15 E	0	0	0 -1 -1	2.34
ARG 300 B	0	TRP 391 A	0	0	1 -1 -1	2.75
ARG 300 B	0	GLY 394 A	0	0	1 -1 -1	3.32
ARG 300 B	0	DT 14 E	0	0	0 -1 -1	3.27
TYR 301 B	0	LYS 395 A	0	0	1 -1 -1	3.24
TRP 391 B	0	PRO 305 A	0	0	1 0 0	3.1
ALA 393 B	0	GLN 306 A	0	0	1 0 0	3.17
SER 397 B	0	ILE 302 A	0	0	1 0 0	2.77
SER 397 B	0	TYR 301 A	0	0	1 0 0	3.27
LYS 398 B	0	TYR 301 A	0	0	1 0 0	3.45
ILE 399 B	0	ARG 300 A	0	0	1 0 0	3
DC 1 C	0	DG 19 F	0	0	1 -1 0	2.86
DC 1 C	0	DC 18 F	0	0	1 -1 0	3.25
DC 1 C	0	HOH 86 S	0	0	1 -1 0	2.89
DG 2 C	0	DG 19 F	0	0	1 -1 0	3.29
DC 19 C	0	DG 1 F	0	0	-1 0 -1	3.29
DC 19 C	0	DC 19 E	0	0	-1 0 -1	3.45
DG 1 D	0	DC 19 E	0	0	-1 0 -1	3.28
DG 4 D	0	LYS 390 A	0	0	0 -1 -1	2.93
DC 18 D	0	DG 19 F	0	0	1 -1 0	3.24
DC 18 D	0	DC 1 E	0	0	1 -1 0	3.31
DC 18 D	0	HOH 30 S	0	0	1 -1 0	3.11
DG 19 D	0	DG 19 F	0	0	1 -1 0	3.13
DC 1 E	0	DC 18 D	0	0	-1 1 0	3.31
DT 14 E	0	ARG 300 B	0	0	0 1 1	3.27
DA 15 E	0	ASN 299 B	0	0	0 1 1	2.34
DC 19 E	0	DC 19 C	0	0	1 0 1	3.45
DC 19 E	0	DG 1 D	0	0	1 0 1	3.28
DG 1 F	0	PHE 416 A	0	0	1 0 0	3.43
DG 1 F	0	DC 19 C	0	0	1 0 1	3.29
DG 3 F	0	LYS 415 A	0	0	1 0 0	3.21
DC 18 F	0	DC 1 C	0	0	-1 1 0	3.25
DG 19 F	0	DG 19 D	0	0	-1 1 0	3.13
DG 19 F	0	DG 2 C	0	0	-1 1 0	3.29
DG 19 F	0	DC 1 C	0	0	-1 1 0	2.86
DG 19 F	0	DC 18 D	0	0	-1 1 0	3.24
HOH 30 S	0	DC 18 D	0	0	-1 1 0	3.11
HOH 49 S	0	HOH 84 S	0	0	-1 1 0	2.41
HOH 84 S	0	HOH 49 S	0	0	1 -1 0	2.41
HOH 86 S	0	DC 1 C	0	0	-1 1 0	2.89
HOH 98 S	0	HOH 99 S	0	0	0 1 0	2.56
HOH 99 S	0	HOH 98 S	0	0	0 -1 0	2.56

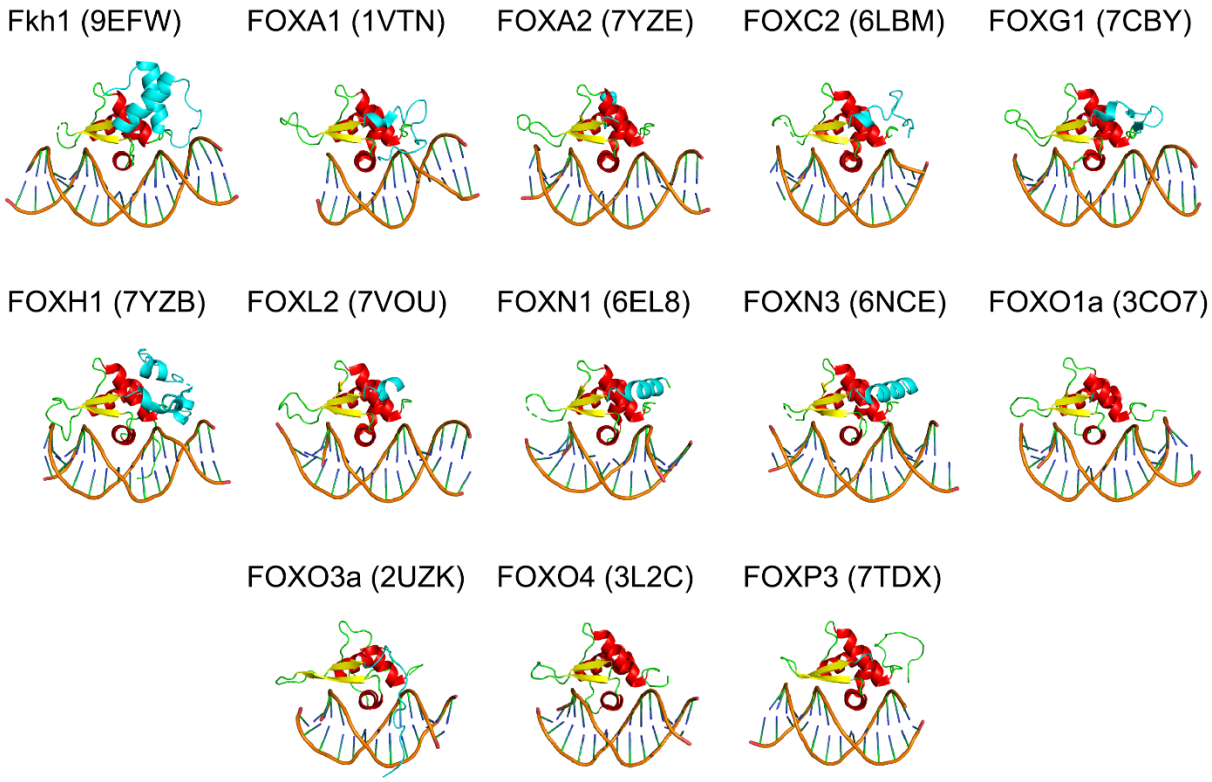
Residue buried areas, 40		
residue	ncs	buried
ARG 300 A	0	0.306
TYR 301 A	0	1.480
ILE 302 A	0	2.340
PRO 305 A	0	1.470
GLN 306 A	0	0.000
LYS 390 A	0	0.622
TRP 391 A	0	4.160
GLY 394 A	0	0.000
LYS 395 A	0	0.497
LYS 415 A	0	0.000
PHE 416 A	0	1.210
ASN 299 B	0	3.670
ARG 300 B	0	4.290
TYR 301 B	0	0.501
TRP 391 B	0	1.520
ALA 393 B	0	0.000
SER 397 B	0	2.500
LYS 398 B	0	1.330
ILE 399 B	0	0.305
DC 1 C	0	2.500
DG 2 C	0	0.000
DC 19 C	0	0.000
DG 1 D	0	0.658
DG 4 D	0	0.626
DC 18 D	0	0.000
DG 19 D	0	3.070
DC 1 E	0	0.000
DT 14 E	0	0.164
DA 15 E	0	3.820
DC 19 E	0	0.656
DG 1 F	0	1.210
DG 3 F	0	0.000
DC 18 F	0	0.244
DG 19 F	0	5.260
HOH 30 S	0	0.000
HOH 49 S	0	2.320
HOH 84 S	0	2.320
HOH 86 S	0	0.143
HOH 98 S	0	1.630
HOH 99 S	0	1.630

Supplementary Figure S10. Crystal contact analysis performed in ChimeraX (14). List of residues participating in crystal contacts within 3.5 Å and buried areas involving those contacts. Wing 2 in both Complex 1 (chain A) and Complex 2 (chain B) only has a small number of residues that participate in crystal contacts (residue 395 in Complex 1, residues 397–399 in Complex 2).



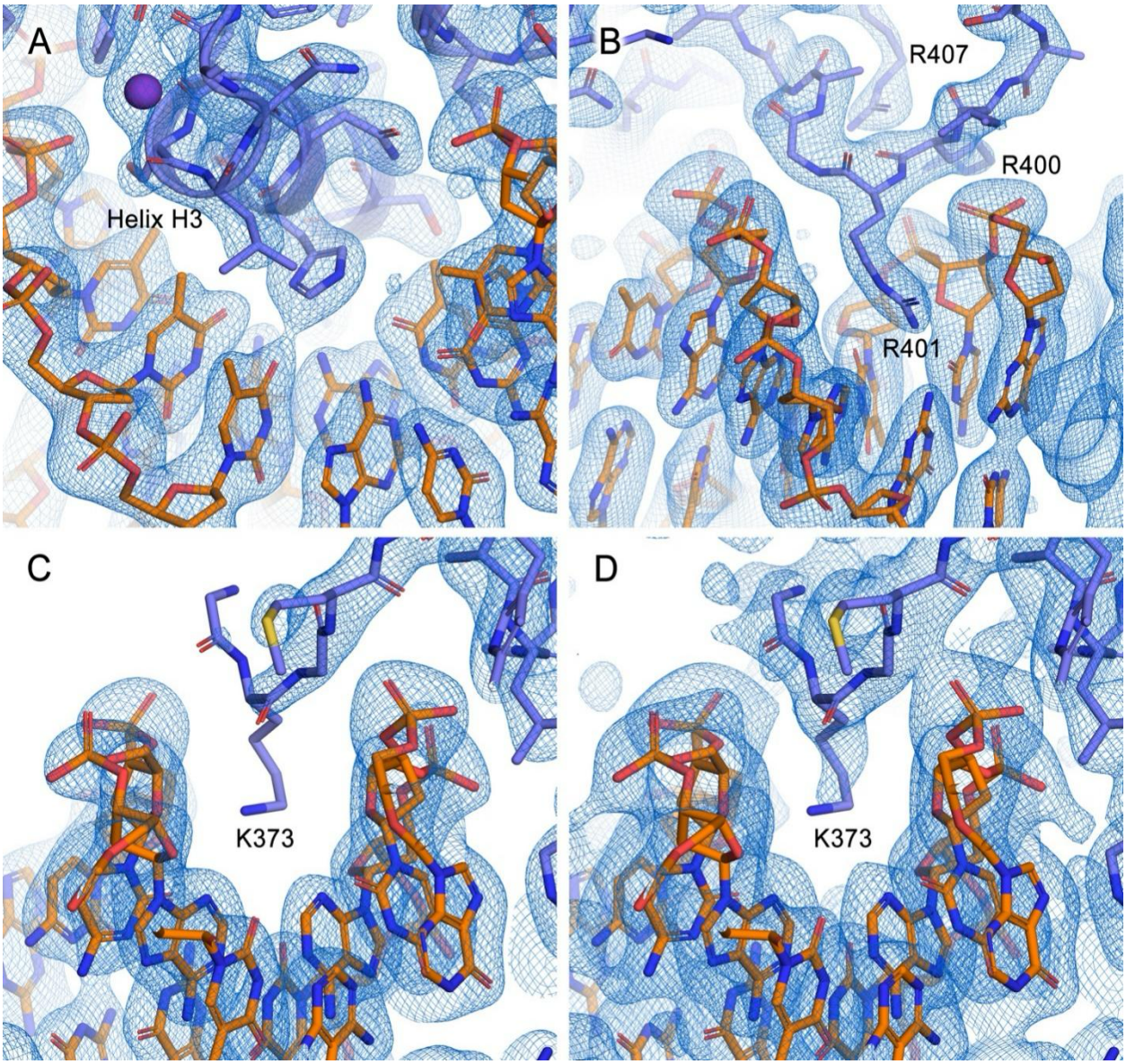
Supplementary Figure S11. Crystal packing diagram of Fkh1-DBD–DNA complexes. (**A**) The asymmetric unit cell is shown bounded in a red box, with Complex 1 colored in red and Complex 2 colored in pink. DNA is shown in blue, and in the other unit cells, Complex 1 is in cyan and Complex 2 is in green. Crystallographic axes are shown in the lower left. (**B**) A zoomed-in view of the region where a non-Watson-Crick base pair in Complex 2 is observed to interact with the DNA oligo of Complex 1 from another unit cell. Potential hydrogen bonds within 3.5 Å are shown in yellow, and less favorable interactions are shown in purple. It seems that DNA interactions across unit cells contribute to the irregular shape of the Complex 2 DNA which may be a cause of differences between Complex 1 and 2.

Supplementary Section S-IV. Comparison of Fkh1-DBD–DNA structure with other known Forkhead-DBD–DNA complexes



Supplementary Figure S12. Comparison between co-crystal structure of yeast Fkh1-DBD–DNA complex and known co-crystal structures of Forkhead-DBD–DNA complexes from other organisms. The Wing 2 region, located on the C-terminal end of β-sheet S2, is indicated in cyan for each complex. Some Forkhead proteins, such as FOXA2, FOXL2, FOXP3, FOXC2, and FOXO4, almost entirely lack a Wing 2 region. For others, such as FOXG1, FOXN3, and FOXC2, Wing 2 does not form interactions with DNA. For several complexes, such as FOXN1, FOXO3a, and FOXA3, Wing 2 interacts with DNA but tends to do so with a loop that is not stably connected to the core of the protein. FOXH1 has a stable Wing 2 that forms DNA interactions, although its structure is different from that of Fkh1. Compared to other Forkhead proteins, Fkh1 has a strikingly distinctive, well-structured Wing 2 structure with DNA interactions.

Supplementary Section S-V. Electron densities of key protein–DNA interactions



Supplementary Figure S13. 2mFo-Dfc electron density maps of important protein–DNA interactions of Fkh1-DBD–DNA Complex 1. (**A**) Pronounced electron density at σ=1 is visible for all residues comprising the core recognition helix H3, which interacts with the major groove. (**B**) Pronounced electron density at σ=1 is visible for residues R400 and R401, which are the residues of Wing 2 that form interactions with the nearby MG. (**C**) The side chain of the K373 residue is not fully visible at σ=1. (**D**) However, electron density at a lower contour level of σ=0.3 alongside other evidence indicates that this is the likely position and rotamer conformation of the K373 side chain.

Supplementary Section S-VI. Water-mediated contacts in the Fkh1–DNA system

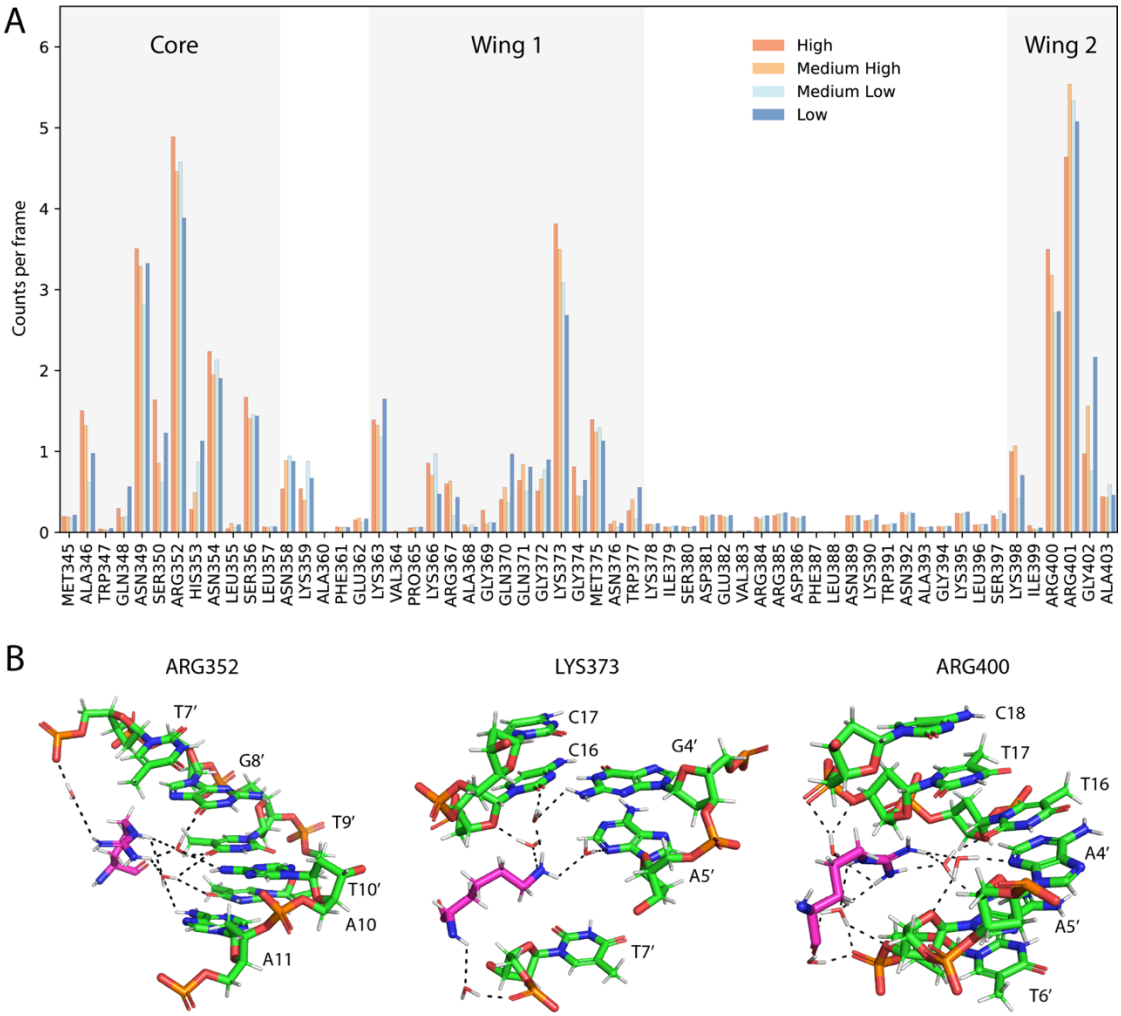
We investigated protein–DNA interactions mediated by water molecules in the final 100 ns of the MD trajectories. The final 100 ns of each trajectory were divided into 1,000 frames spaced 0.1 ns apart. Water-mediated contacts were quantified by counting hydrogen bonds formed concurrently by water molecules with proteins and DNA in each frame. Standard cutoffs of 120° and 3.5 Å between hydrogen bond donor and acceptor were used to identify hydrogen bonds. Average hydrogen bond counts were computed for each protein–DNA conformation across five replicas. In Supplementary Figure S14 below, we analyzed water-mediated contacts involving each protein residue. Note that each protein residue can establish hydrogen bonds with multiple nucleotides. The water-mediated contacts seen in the co-crystal structure (Supplementary Figure S4) were also confirmed in the MD simulations.

In the core motif region, we observed a correlation between binding affinity and hydrogen bond counts. Specifically, higher-affinity sequences exhibit higher water-mediated hydrogen bond counts involving residues A346, R352, and H353. Residue R352 forms extensive hydrogen bonds with nucleotides in the core motif region (Supplementary Figure S14B, left panel).

Some other important hydrogen bonds are formed with N349 and S356. The importance of these hydrogen bonding residues for DNA recognition was shown (Supplementary Figure S14).

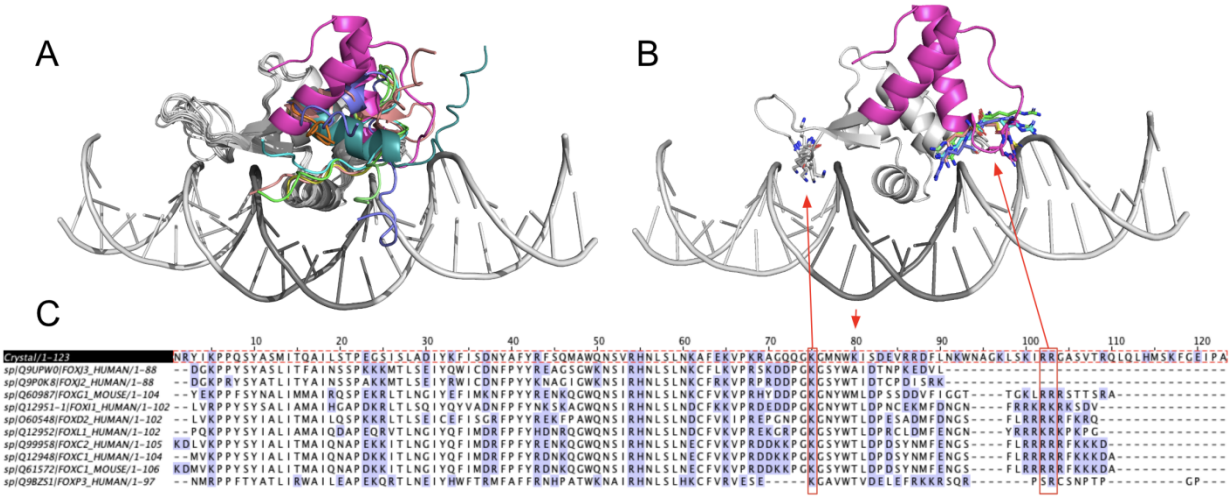
In the Wing 1 region, which is in close contact with DNA, residues K373 and M375 play crucial roles in forming water-mediated contacts between protein and DNA, which can result in increased binding affinity. Particularly, K373 appears to be crucial in influencing differences in hydrogen bond contacts across sequences with varying binding affinity in the 3’ flanking region (Supplementary Figure S14B, center panel).

Lastly, in the Wing 2 loop region, residues K398, R400, and R401 are notable for their important interactions with DNA across most sequences. K398 and R400 residues form more hydrogen bonds with higher-affinity sequences (Supplementary Figure S14B, right panel), possibly as an additional, albeit weaker but more specific contact.



Supplementary Figure S14. (A) Histogram of Fkh1–DNA water-mediated hydrogen bond counts for sequences of various binding affinities pooled over five replicas. The shaded regions in the plot indicate the residues in the core, Wing 1, and Wing 2 domains, respectively. **(B)** Visualization of water-mediated hydrogen bonds of R352, K373, and R400, indicates high hydrogen bond count in (A).

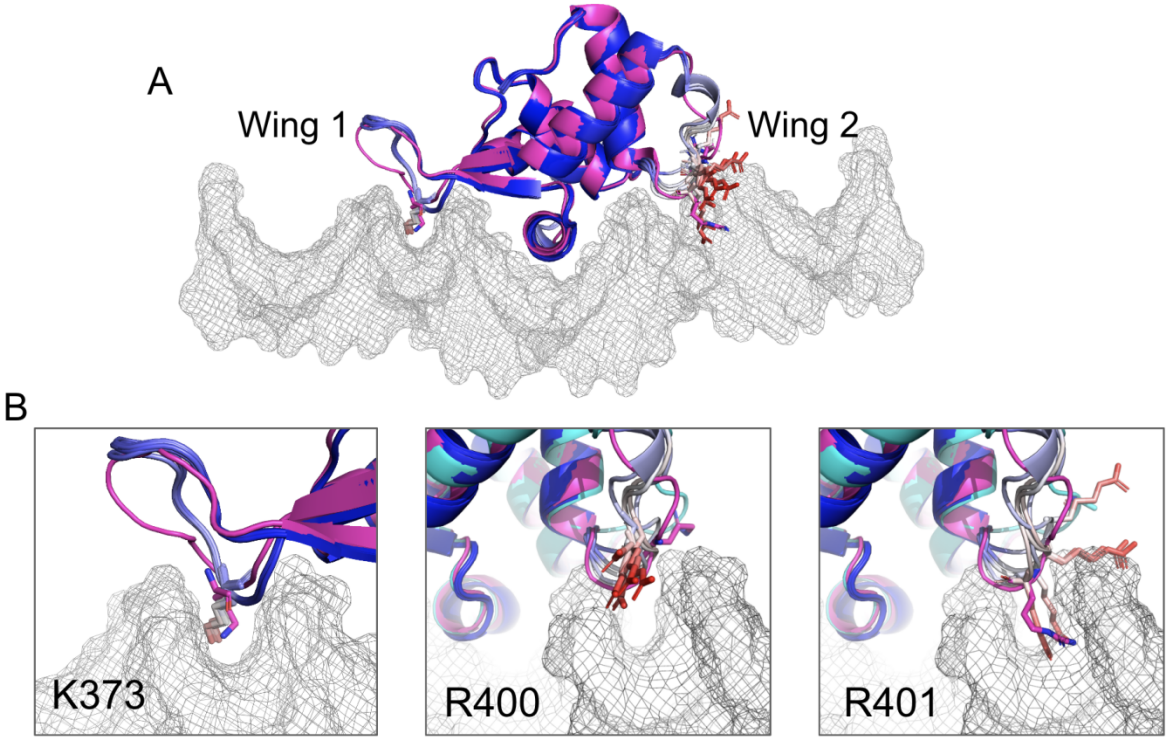
Supplementary Section S-VII. Sequence alignment of Fkh1 with other Forkhead proteins



Supplementary Figure S15. Fkh1 structure and sequence comparison for known Forkhead proteins with binding data. **(A)** Structural alignment of various Forkhead proteins aligned with the first 80 aa (in gray) indicated with red arrow in (C). The colored region shows the structure of the Wing 2 domain of different Forkhead proteins listed in (C). The co-crystal structure is colored in magenta. The 7-bp core motif region is colored in darker gray. **(B)** MG-contacting lysine (K373) and two arginine residues (R400 and R401) of Forkhead structures are shown in stick representation. **(C)** Sequence alignment of different Forkhead proteins. Alignment is colored by charged residues (purple).

Supplementary Section S-VIII. AlphaFold 3 predictions of the Fkh1–DNA complex

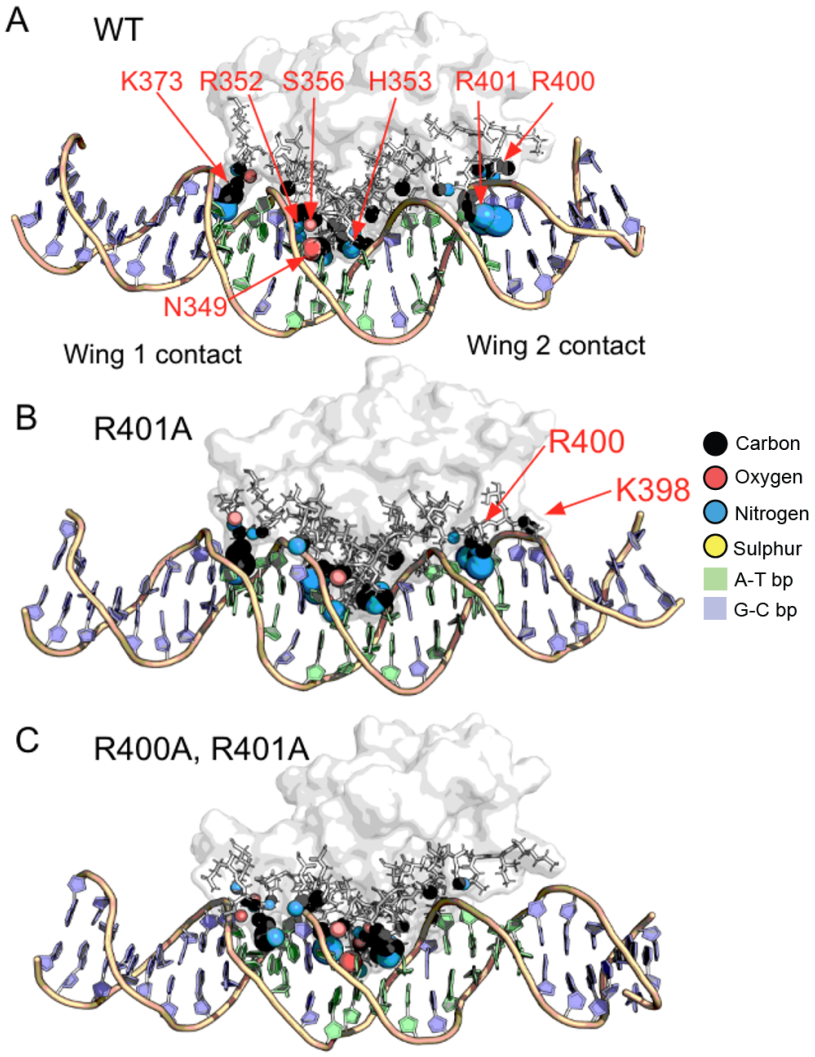
We used AF3 (9) to predict conformations of complexes that are equivalent to assemblies observed in our co-crystal structure. Prediction for the Fkh1–DNA complex shows that both the DNA and protein chains have high average *pLDDT* confidence scores of 81.31 and 88.83, respectively.



Supplementary Figure S16. (A) AF3-predicted structures for Fkh1–DNA complex using the same construct as the co-crystal structure. Residues K373, R400, and R401 are shown in stick representation. Complex 1 of the co-crystal structure is shown in magenta for comparison. **(B)** Zoomed-in views of amino acids K373 (left), R400 (center), and R401 (right) are shown in stick representation. Complex 1 is shown in magenta and Complex 2 in cyan. All structures, unless indicated, are colored by AF3 confidence score with the least confidence in red (*pLDDT* < 50), higher confidence in white (40 < *pLDDT* < 75), and greatest confidence in blue (*pLDDT* > 90).

Supplementary Section S-IX. DeepPBS residue-level importance score prediction for the most representative snapshots from MD trajectories

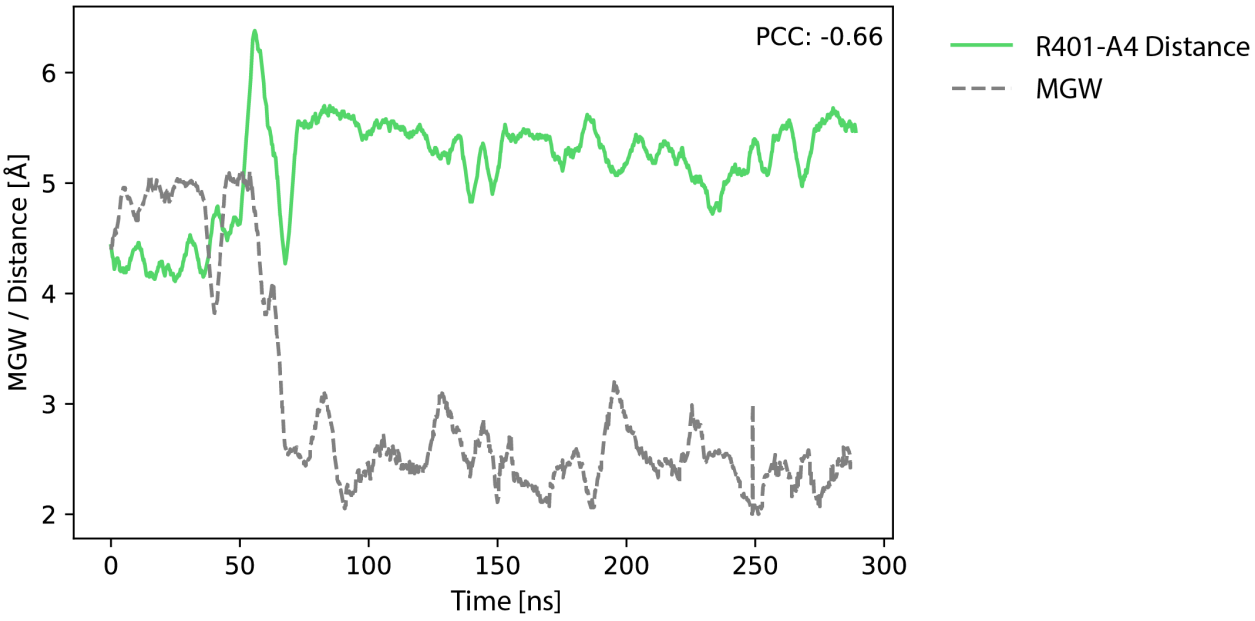
We used DeepPBS (15) to calculate heavy atom-level importance scores for binding specificity in the protein–DNA binding interface for (A) WT, (B) single-mutant, and (C) double-mutant systems.



Supplementary Figure S17. DeepPBS prediction and R401 impact on MGW. **(A)** DeepPBS prediction of residue-level importance scores for DNA binding specificity for average structure of WT system. **(B)** DeepPBS prediction of residue-level importance scores for DNA binding specificity for average structure of R401A system. **(C)** DeepPBS prediction of DNA binding specificity for average structure of double-mutant system. In (A), (B), and (C), the size of the spheres (indicating heavy atoms of the DNA-contacting amino acid side chains) represents their relative importance for the predicted DNA binding specificity. The color of the spheres indicates the electronegativity of the respective heavy atom, with blue representing electropositive atoms, red representing electronegative atoms, and black representing heavy atoms that are approximately neutral. Residues in the protein–DNA interface are shown in stick representation.

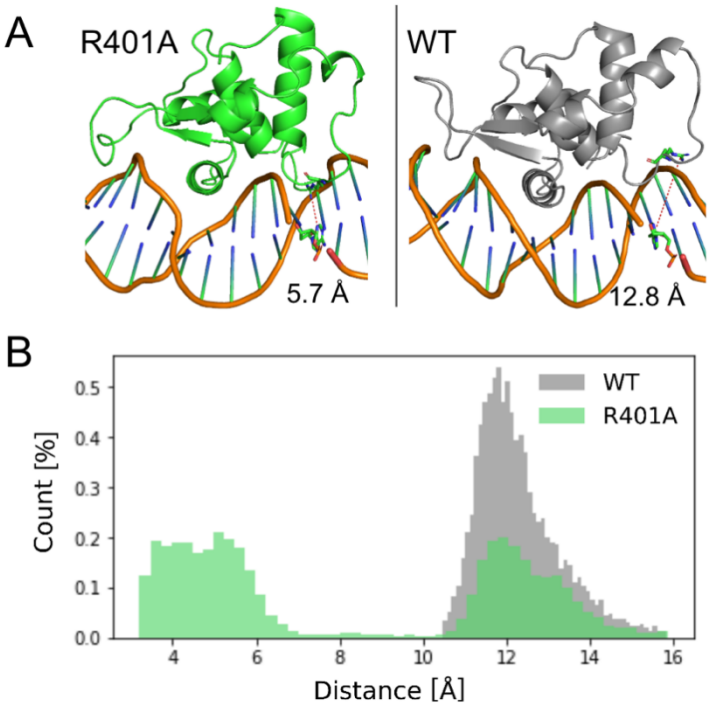
Supplementary Section S-X. Arginine residue R401 can modulate MGW in flanking regions

In the Complex 1 system, the correlation of MGW and the distance of R401 from the bases was calculated by comparing MGW and the distance between the guanidinium group of the R401 side chain and the N3 atom of the A4 base across the 300-ns trajectory while averaging over 500-ps sliding windows.



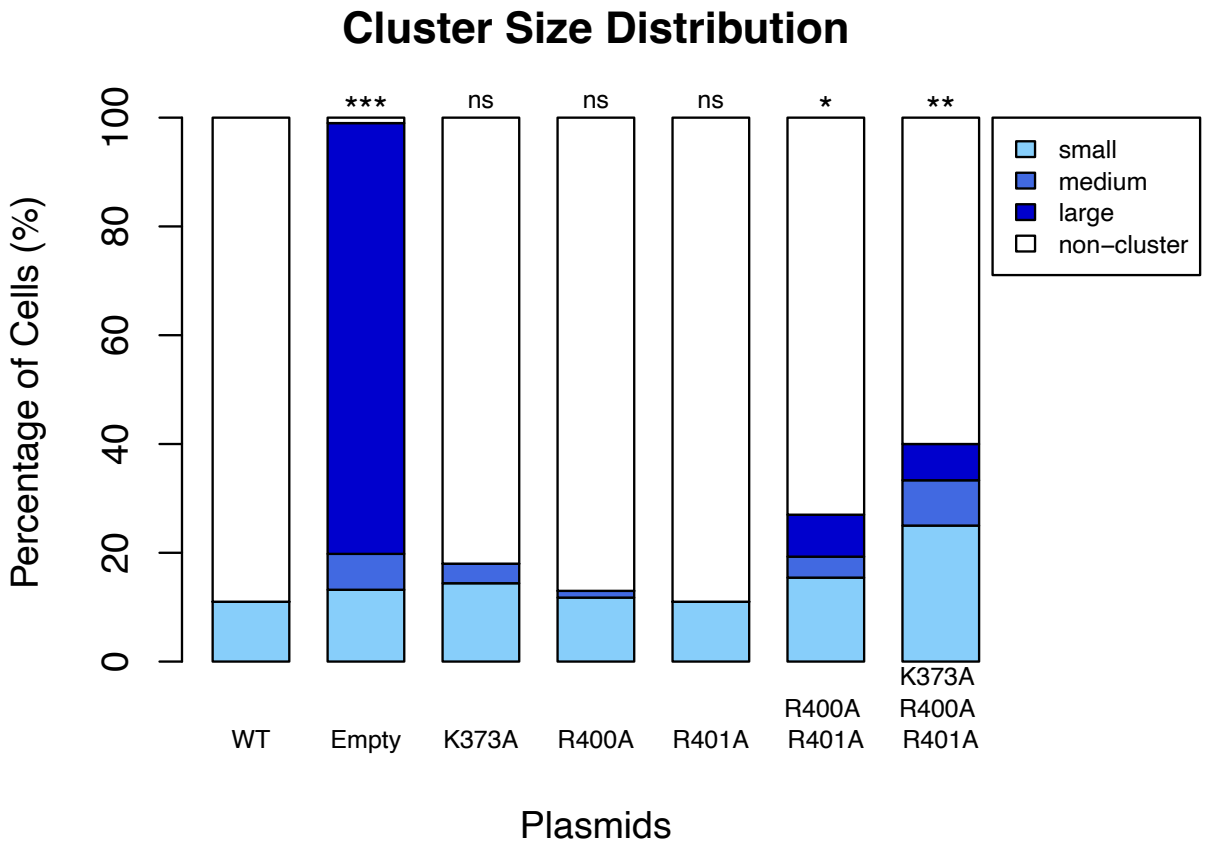
Supplementary Figure S18. Correlation between MGW at the A4 base (in gray dash line) and distance between R401 and A4 base in green in the WT system. Pearson correlation coefficient (PCC) is denoted in upper right corner.

Supplementary Section S-XI. Arginine residue R400 substitution

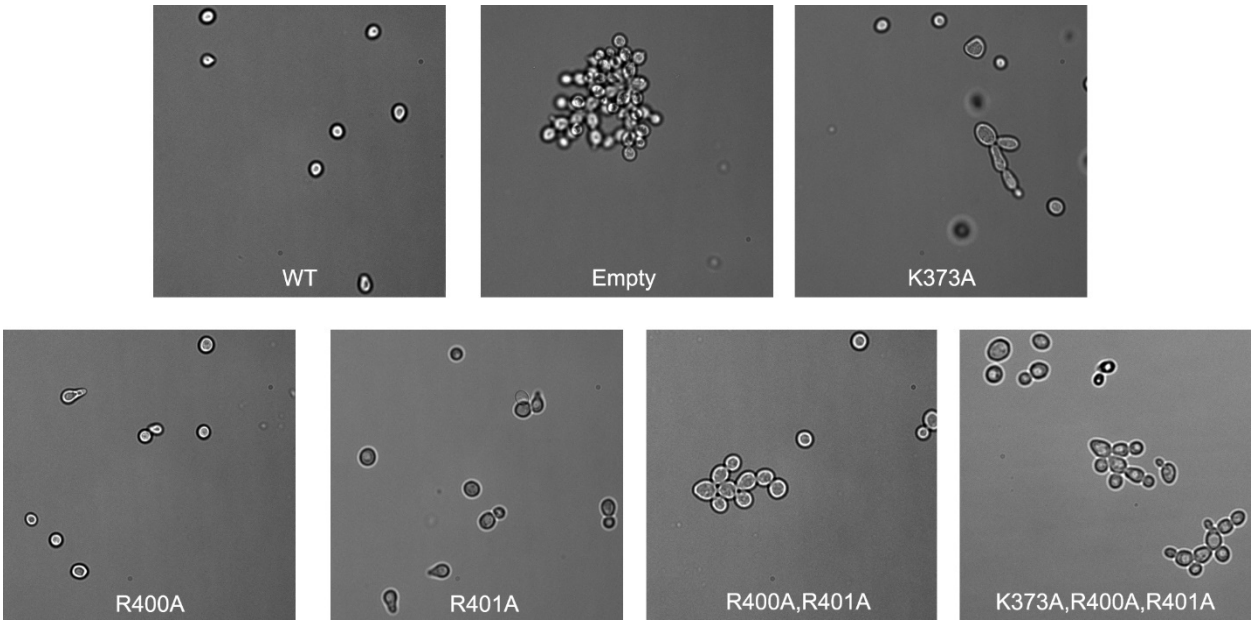


Supplementary Figure S19. R400 can enter the MG and substitute for R401. **(A)** Visualization when R400 is inserted into MG in the R401A mutant (left), and when it resides outside the MG in the WT system (right). Distance between R400 and A4 base is shown in bottom right corner. **(B)** Distance distribution for R401A mutant (left panel of A) in green, and that of the WT system (right panel in A) in gray among five replicas. The substitution mechanism is observed in three out of the five replicas.

Supplementary Section S-XII. Quantification of Fkh1 mutant yeast pseudohyphal morphology

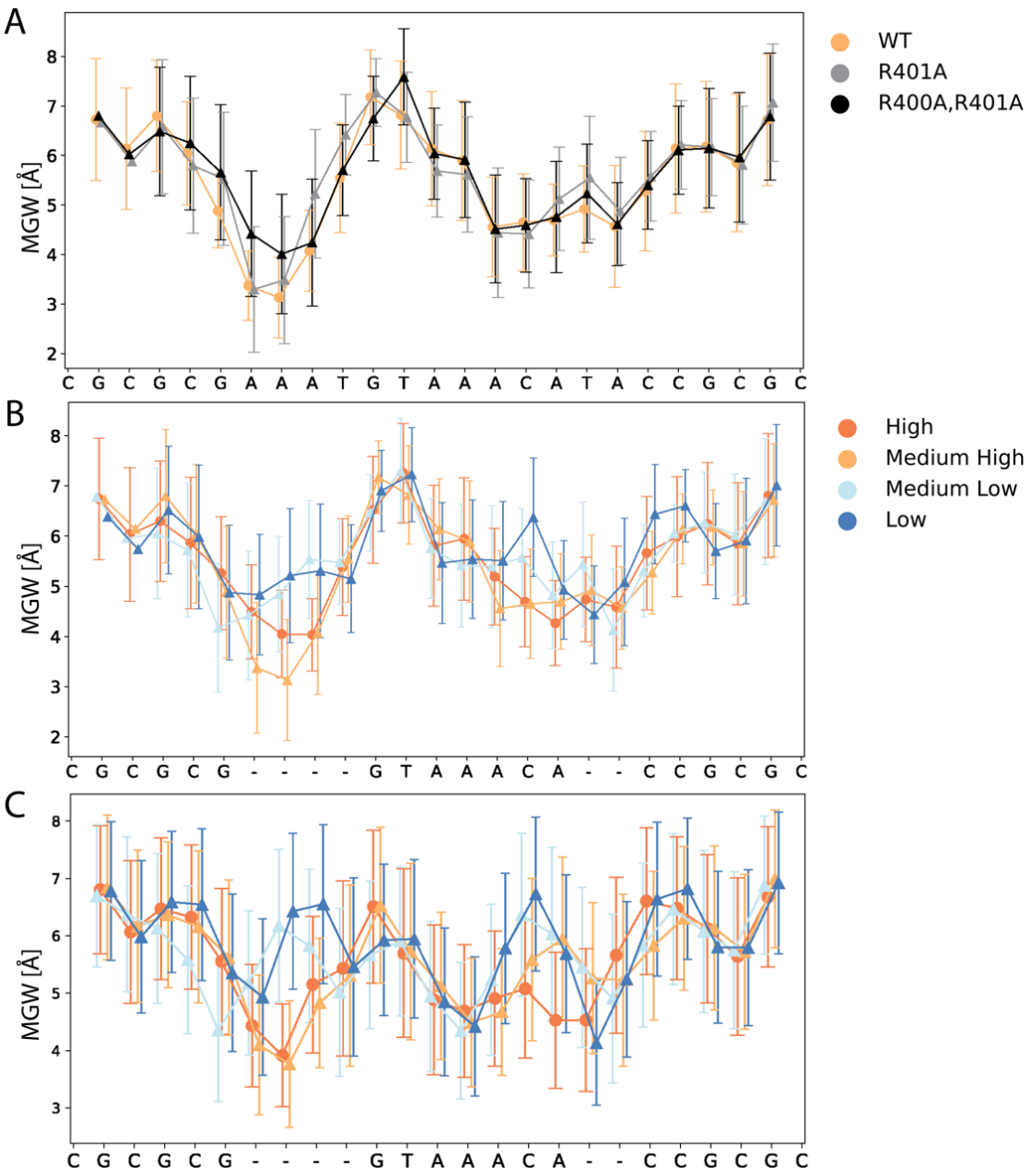


Supplementary Figure S20. Quantification of percentage of cells counted in clusters in *fkh1Δ fkh2Δ* strains carrying plasmids with WT *FKH1*, empty, or *FKH1* mutant alleles. Clusters were further stratified into three subsets: small (3 ≤ number of cells in cluster ≤ 4), medium (5 ≤ number of cells in cluster ≤ 8), and large (number of cells in cluster ≥ 9). Compositions of three cluster sizes are shown in the stacked bar plot. Asterisks indicate one-tailed t-test performed for each strain compared with WT.

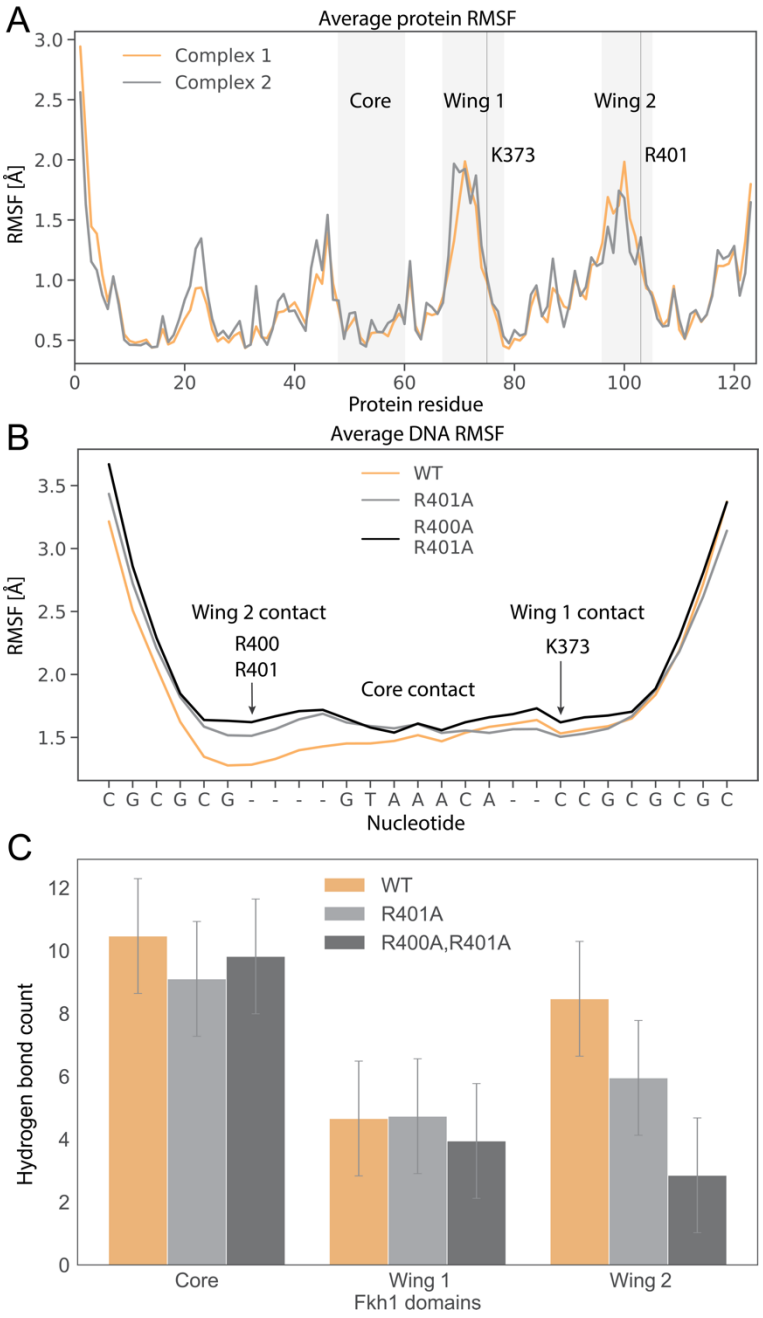


Supplementary Figure S21. Representative microscopic images showing morphology of *fkh1Δ fkh2Δ* strains containing plasmids expressing WT Fkh1 and different Fkh1 mutants.

Supplementary Section S-XIII. MGW shape profile with conformational fluctuations



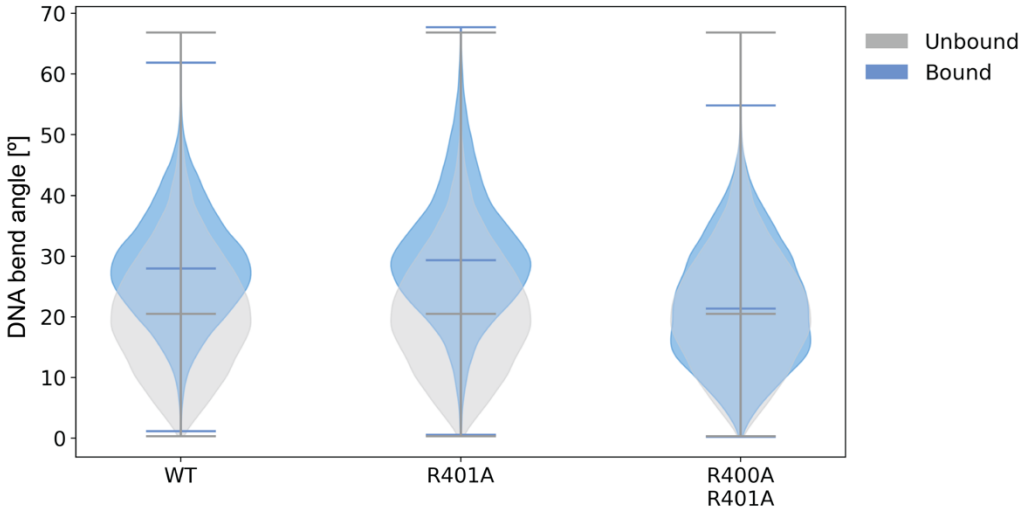
Supplementary Figure S22. Distribution of MGW for 27-bp oligos aggregated from five replicas. MGW values for terminal G/C bp are missing because of high fluctuations at these positions during simulation due to end effects. MGW values for central 17-bp regions are plotted for **(A)** DNA shape for WT and mutant protein–DNA complexes, **(B)** protein–DNA complexes with DNA sequences with various binding affinities, and **(C)** unbound DNA with various binding affinities. Dashes indicate locations of proximal DNA flanking regions.



Supplementary Figure S23. (A) Protein root-mean-square fluctuation (RMSF) for Fkh1 protein in Complex 1 (orange color) and Complex 2 (gray). Core, Wing 1, and Wing 2 regions are shaded in light gray. Vertical lines indicate position of residues K373 and R401. **(B)** DNA RMSF for WT and mutated protein–DNA complexes averaged over five replicas. Relative locations of R400, R401, and K373 residues are marked with arrows. Locations of Wing 1, Wing 2, and core domains are labeled with gray text. **(C)** Bar plot showing average hydrogen bond count in core, Wing 1, and Wing 2 regions across five replicas for both mutant and WT systems. Error bars represent variations among replicas. Whiskers represent standard deviations of hydrogen bond counts across all replicas.

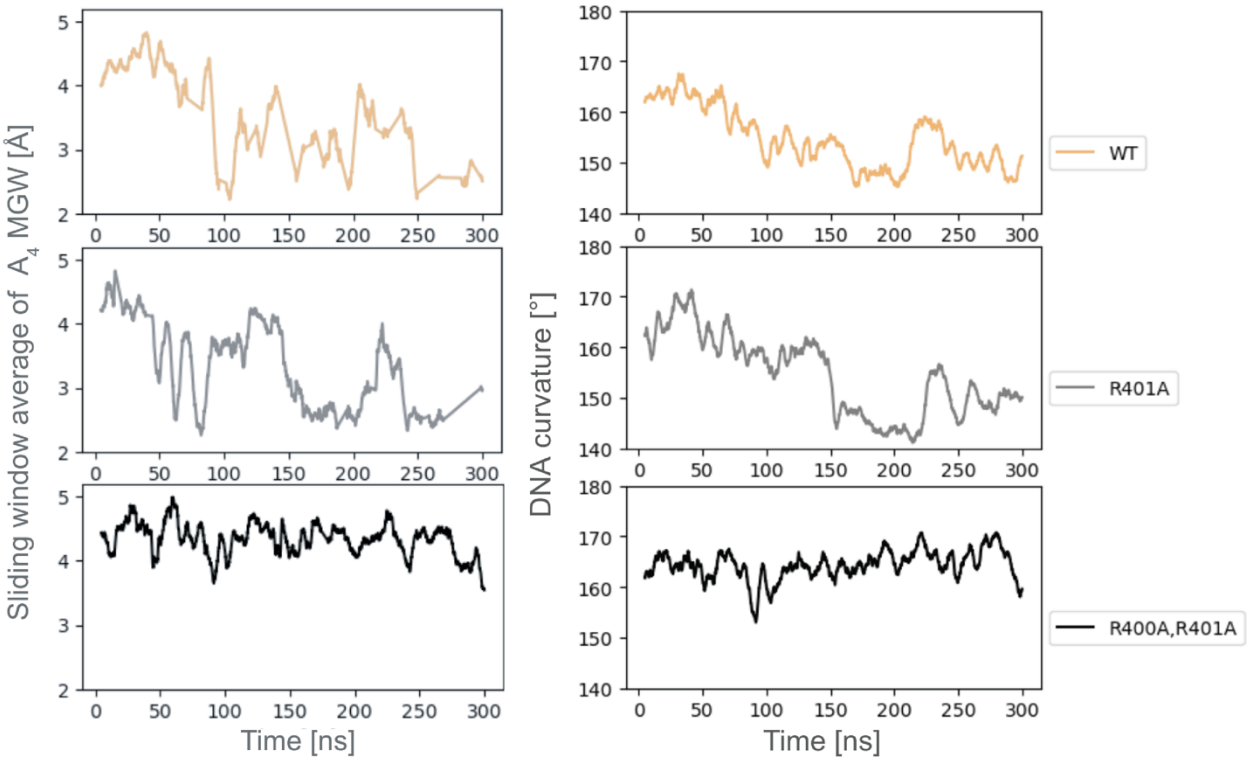
Supplementary Section S-XIV. DNA curvature

The WT protein results in about ~7° more bending than essentially straight free DNA, the most bent among the systems below. The double mutant is closest to free DNA (~1° more bent than free DNA) and has a highly similar distribution (Supplementary Table S4). DNA in the R401A system exhibits a curvature that is intermediate between WT and double mutant systems. This shows that these two arginine residues are important for forming the bound DNA shape by bending the DNA. Without these residues, the DNA contact in Wing 2 is abolished and the bidentate hydrogen bonding arginine releases the DNA, resulting in a DNA curvature that is similar to that of free DNA with the same nucleotide sequence.



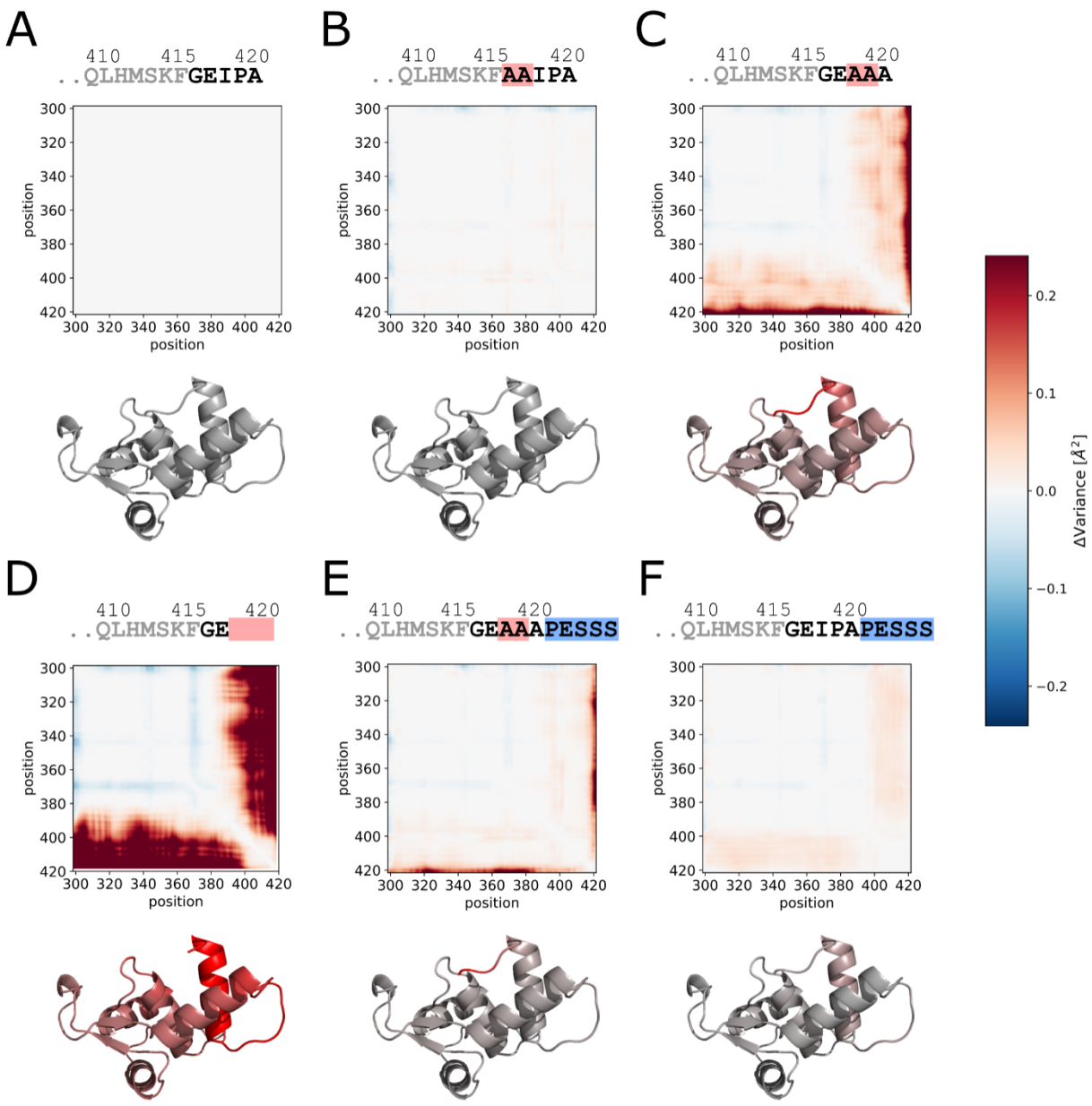
Supplementary Figure S24. DNA curvatures of mutant and WT systems calculated from MD trajectories. Whiskers represent min and max values. Center line represents the median.

Finally, we examined whether the observed bending is caused by asymmetric MG narrowing at the protein interaction site. In the WT and R401A systems, both bending and MG narrowing were detected, whereas neither phenomenon was observed in the double-mutant system during the MD simulations. Additionally, a time-series analysis revealed a correlation between MGW narrowing and DNA bending and that MG narrowing always occurs before DNA bending, suggesting that MGW narrowing can drive DNA bending (Supplementary Table S4).



Supplementary Figure S25. Time series comparing MGW (left) and DNA curvature (right) of the mutant and WT systems from MD trajectories for one replica. The DNA curvature here is calculated as 180 minus the DNA bend angle.

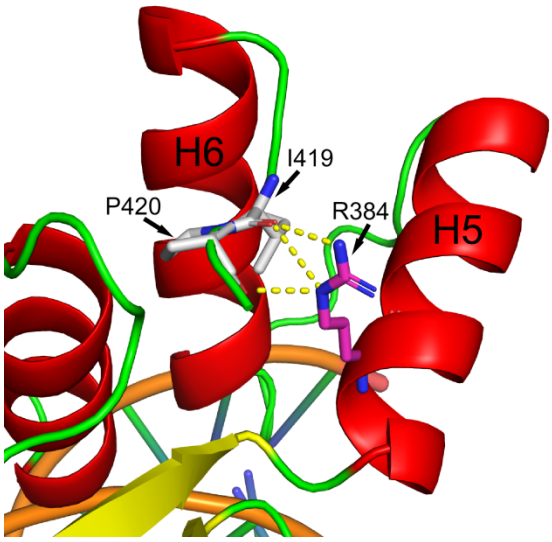
Supplementary Section S-XV: BioEmu analysis of conformational flexibility of Fkh1 structure



Supplementary Figure S26. Variance plots of various Fkh1 mutations calculated with BioEmu (10). A total of 1,000 protein ensembles were produced for each mutation, and approximately ~700 final ensembles were used to analyze each sequence based on BioEmu's validation criteria. **(A)** Ensemble variance was compared to variance in the original co-crystal structure. BioEmu analysis of WT protein sequence used in crystal structure indicates high stability of the protein conformation. Next, residues were mutated to alanine or removed entirely to assess their impact on Wing 2 stability. **(B)** Mutations of residues 417 and 418, which do not appear to have side chain interactions in the protein structure, induced nearly no change in variance of the Wing 2 region. **(C)** However, when the I419 and P420 residues were mutated into alanine, major differences are observed in the Wing 2 region, which becomes destabilized and adopts a greater variety of conformations. This finding indicates the importance of the side chain interactions of those residues to the stability of the Fkh1 Wing 2 structure. **(D)** C-terminal deletion of residues 419-421 greatly disrupts the Wing 2 structure and causes even greater variance in

protein structures produced by BioEmu (10). **(E-F)** Extending the protein sequence beyond the crystal structures does not have much of an effect on the length of the sequence used in the crystal structure.

Supplementary Section S-XVI. Importance of I419 and P420 residues to Fkh1 Wing 2 stability



Supplementary Figure S27. Interactions between helices H5 and H6 in Complex 1 of the Fkh1-DBD–DNA complex. Residues I419 and P420 on helix H6 form hydrogen bonds with residue R384 on helix H5, creating a stabilizing Wing 2 interaction.

SUPPLEMENTARY REFERENCES

1. Cooper,B.H., Dantas Machado,A.C., Gan,Y., Aparicio,O.M. and Rohs,R. (2023) DNA binding specificity of all four *Saccharomyces cerevisiae* forkhead transcription factors. *Nucleic Acids Res.*, **51**, 5621–5633.
2. Jiang,Y., Chiu,T.P., Mitra,R. and Rohs,R. (2024) Probing the role of the protonation state of a minor groove-linker histidine in Exd-Hox–DNA binding. *Biophys. J.*, **123**, 248–259.
3. Maier,J.A., Martinez,C., Kasavajhala,K., Wickstrom,L., Hauser,K.E. and Simmerling,C. (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theor. Comput.*, **11**, 3696–3713.
4. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R. and Balaceanu,A. (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
5. Darden,T., York,D. and Pedersen,L. (1993) Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *J. Chem. Phys.*, **98**, 10089–10092.
6. Hess,B., Bekker,H., Berendsen,H.J.C. and Fraaije,J.G.E.M. (1997) LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, **18**, 1463–1472.
7. Smith,L.J., Daura,X. and van Gunsteren,W.F. (2002) Assessing equilibration and convergence in biomolecular simulations. *Prot. Struct. Funct. Bioinform.*, **48**, 487–496.
8. Daura,X., van Gunsteren,W.F. and Mark,A.E. (1999) Folding–unfolding thermodynamics of a β-heptapeptide from equilibrium simulations. *Prot. Struct. Funct. Bioinform.*, **34**, 269–280.
9. Abramson,J., Adler,J., Dunger,J., Evans,R., Green,T., Pritzel,A., Ronneberger,O., Willmore,L., Ballard,A.J., Bambrick,J., et al. (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, **630**, 493–500.
10. Lewis,S., Hempel,T., Jiménez-Luna,J., Gastegger,M., Xie,Y., Foong,A.Y.K., Satorras,V.G., Abdin,O., Veeling,B.S., Zaporozhets,I., et al. (2025) Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, **389**, eadv9817.
11. Daura,X., van Gunsteren,W.F. and Mark,A.E. (1999) Folding–unfolding thermodynamics of a β-heptapeptide from equilibrium simulations. *Proteins Struct. Funct. Bioinform.*, **34**, 269–280.5.
12. Mitra,R., Cohen,A.S., Sagendorf,J.M., Berman,H.M. and Rohs,R. (2025) DNAproDB: an updated database for the automated and interactive analysis of protein–DNA complexes. *Nucleic Acids Res.*, **53**, D396-D402.
13. Krissinel, E. (2010) Crystal contacts as nature's docking solutions. *J. Comput. Chem.*, **31**, 133-143.
14. Pettersen,E.F., Goddard,T.D., Huang,C.C., Meng,E.C., Couch,G.S., Croll,T.I., Morris,J.H. and Ferrin,T.E. (2021) UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.*, **30**, 70–82.
15. Mitra,R., Li,J., Sagendorf,J.M., Jiang,Y., Cohen,A.S., Chiu,T.P., Glasscock,C.J. and Rohs,R. (2024) Geometric deep learning of protein–DNA binding specificity. *Nat. Methods*, **21**, 1674–1683.