Supplementary Information for

DNA mismatches reveal conformational penalties in protein-DNA recognition

Ariel Afek[1,2], Honglue Shi[3], Atul Rangadurai[4], Harshit Sahay[1,5], Alon Senitzki[6], Suela Xhani[7], Mimi Fang[9], Raul Salinas[4], Zachery Mielko[1,10], Miles A. Pufall[9], Gregory M.K. Poon[7,8], Tali E. Haran[6], Maria A. Schumacher[4], Hashim M. Al-Hashimi[3,4]*, Raluca Gordan[1,2,11]*

[1]Center for Genomic and Computational Biology, [2]Department of Biostatistics and Bioinformatics, [3]Department of Chemistry, [4]Department of Biochemistry, [5]Program in Computational Biology and Bioinformatics, Duke University, Durham, NC 27708, USA;
[6]Department of Biology, Technion–Israel Institute of Technology, Technion City, Haifa 32000, Israel;
[7]Department of Chemistry, [8]Center for Diagnostics and Therapeutics, Georgia State University, Atlanta, GA 30303, USA;
[9]Department of Biochemistry, Holden Comprehensive Cancer Center, Carver College of Medicine, University of Iowa, Iowa City, IA 52242, USA;
[10]Program in Genetics and Genomics, [11]Department of Computer Science, Department of Molecular Genetics and Microbiology, Duke University, Durham, NC 27708, USA
* Corresponding authors.

This PDF file includes:

Supplementary Methods
Supplementary Discussion
Titles and Legends for Supplementary Tables 1-11
Title and Legend for Supplementary Figure 1
Supplementary References

Supplementary Methods

Structural survey of Watson-Crick and mismatched base pairs

X-ray crystal structures and NMR solution structures containing DNA were downloaded with their PDB information including resolution, macromolecule type etc. from the RCSB webserver on 08/16/2017. Structures were parsed using X3DNA-DSSR[36] into a searchable database (base pair database) as described in previous studies[35]. Structures with resolution > 3 Å were excluded. The database contains all DNA base pairs in the biological assembly with an accompanying list of structural parameters describing those base pairs.

Local base pair and base step parameters, as well as global shape parameters, were computed as described in Methods. The sign of shear, buckle, shift and tilt for all the A-T and G-C Watson-Crick base pairs were adjusted according to the index of the purine and pyrimidine (i.e. a negative shear value indicates pyrimidine translates to major groove direction relative to purine). Base pair parameters of bases with *syn* conformation (e.g. in Hoogsteen base pairs) were not computed due to incorrect reference frame. We used a well-established inter-helical Euler angle approach to quantify DNA local bending with the bending magnitude (β_h, $0° \leq \beta_h \leq 180°$), the bending direction (γ_h, $-180° \leq \gamma_h \leq 180°$), and the helical twist (ζ_h, $-180° \leq \zeta_h \leq 180°$) of two helices (H1 and H2) across a given base pair junction[35,37,39,40]. The junction can be a Watson-Crick base pair, Hoogsteen base pair, or any mismatched pair. In this approach, two idealized B-form DNA helices constructed by 3DNA[36] and containing 3 base pairs were superimposed to H1 and H2, respectively, yielding a relative orientation of H1 to H2 that was quantified by the parameters β_h, γ_h, and ζ_h. All calculations with poor alignment to the idealized helices (RMSD > 2 Å for sugar and backbone atoms[39]) were omitted from analysis. Protein-DNA structures were illustrated using PyMOL 1.5.0.4.

Survey of standard Watson-Crick base pairs in B-DNA: To construct a dataset of standard Watson-Crick (WC) bps in B-DNA, we identified all canonical WC bp structures in a B-DNA environment following several criteria: (1) the canonical WC bp was from naked DNA structures without any protein or ligand bound, (2) base pairs with modified bases were not considered, (3) the WC base pair was surrounded by at least two canonical WC base pairs on both sides in a DNA stem, and (4) non B-form DNAs such as A-form or Z-form DNA were removed. A dataset containing a total of 903 A-T and 746 G-C standard WC bps was generated, and used to define the B-DNA envelope (**Extended Data Fig. 1a**, **Supplementary Table 8**).

Survey of significantly distorted Watson-Crick base pairs in TF-DNA complexes: We identified a total of 613 TF-DNA structures in PDB[34]. To locate all distorted WC base pairs in these structures, an in-house Python program was used to identify base pairs that satisfy the following criteria: (1) the base pair contains no modified bases, (2) the base pair is surrounded by at least two WC base pairs on each side, and (3) the base pair contains at least one base pair parameter that deviates from the free B-DNA envelope by either 3 standard deviations, or is completely outside the free B-DNA envelope. The statistics of these distorted WC base pairs are summarized in **Extended Data Fig.1** and **Supplementary Table 8**.

Survey of DNA mismatches in DNA helical context: To survey the DNA mismatch structure and geometry, the program was also used to search and identify all possible single DNA mismatches (excluding modified bases) surrounded by at least two canonical WC base pairs on both sides, and were then subjected to manual inspection. Our survey identified a total of 44 G-T, 15 T-T, 13 A-C, 12 G-A, 9 G-G, 8 A-A, 6 C-T and 3 C-C mismatches within canonical DNA duplex context, of which 26 were in free DNA and not mediated by heavy metals inserted in between the two bases (8 G-T, 7 G-A, 5 A-C, 3 T-T, 2 G-G, and 1 C-T) (**Supplementary Table 9, Extended Data Fig. 2a**).

Molecular dynamics (MD) simulations

All MD simulations were performed using the AMBER ff99 force field[42] with bsc0 corrections for DNA[43], ff14SB corrections for proteins[44], and using standard periodic boundary conditions as implemented in the

AMBER MD package[45]. The crystal waters in all structures were retained. The structures were then solvated using a truncated octahedral box of SPC/E[85] water molecules, with box size chosen such that the boundary was at least 10 Å away from any of the atoms of the free DNA or protein-DNA complex, using the leap module of the AMBER package. The default protonation states for protein residues that were assigned by leap were retained unless mentioned otherwise. Na^+ ions treated using the Joung Cheatham parameters[86] were then added to neutralize the charge of the system. The parameters used for the subsequent simulation setup are the same as those used in a prior study[37].

MD simulations of free DNA with single mismatches: To systematically analyze the ensemble behavior of all mismatches, we performed MD simulations on unbound DNA for all possible WC and mismatched base pairs embedded in constant flanking sequences: 5'-CTCTGCC**A**CGTGGGTCGT-3' (the variable position is underlined). For G-A and G-G, we simulated two possible geometries: G(anti)-A(anti), G(anti)-A(syn) and G(anti)-G(syn), G(syn)-G(anti), where one of the bases was manually rotated about the glycosidic bond by 180° to generate a *syn* conformation. Production runs of 500ns were carried out and extended to achieve convergence of the RMSD of the DNA if necessary. Based on insights obtained from the crystal structure survey mentioned above (**Extended Data Fig. 2a**, **Supplementary Table 9**) the initial mismatch geometry was modeled as follows: (1) G-T and T-T mismatches were modeled as wobble conformations; (2) C-T mismatch was modeled with two H-bonds (C:N3---T:H3-N3 and C:H4-H4---T:O4); (3) C-C mismatch was modeled with no H-bond but still stacking in the helix; (4) G-G mismatch was modeled as either G(*syn*)-G(*anti*) or G(*anti*)-G(*syn*) conformation; (5) G-A mismatch was modeled as either G(*anti*)-A(*anti*) or G(*anti*)-A(*syn*) conformation. Transient and rare species, such as the tautomeric form of the G-T mismatch, were not considered in our MD analyses. Protonated or anionic base pairs, such as A^+-C, C^+-C and G(syn)-A^+(anti), were not considered for simplicity, since they are highly dependent on pKa and pH. Despite all these limitations, though, MD simulations represent a good alternative to study candidate structural deformations induced by mismatches in the absence of available structural data. Summary descriptions of the ensemble behavior of different mismatches in the unbound DNA MD simulations are presented in **Extended Data Fig. 2c**. The dynamics of DNA mismatches in MD simulation are in good agreement with a previous study[46].

MD simulations of protein-DNA complexes: Starting structures corresponding to the Myc/Max, Ets1, p53, Max/Max, CTCF, Egr1, GR, Elk1, and RelA systems were obtained from PDB entries 1NKP, 2NNY, 3KZ8, 1AN2, 5KKQ, 1P47, 1R4R, 1DUX, and 5U01 respectively. For Myc/Max, chains A, B, F and G of PDB 1NKP were retained to obtain the starting structure for the simulations. For Ets1, chains A and residues 2-14 and 11-23 from chains C and D, respectively, were retained to obtain a structure, from which the first ten residues of the unstructured N-terminal region were removed, to generate the final starting structure for the simulations. For p53, the missing Ser185-Asp186-Gly187 residues were modeled using the ModLoop webserver[87], and the iodine atoms in the asymmetric unit were removed to obtain the starting structure. For CTCF, chains A, B and C of PDB 5KKQ were chosen while retaining the protonation states of the amino acids identified in the crystal structures to obtain the starting structure. For Egr1, chains A and residues 2-16 and 49-63 from chains C and D respectively, were retained from PDB 1P47 to obtain the starting structure. For Elk1, chains A, B and C from PDB 1DUX were retained to obtain the starting structure for MD. For p53, GR and Egr1, the sulphur and nitrogen atoms of the cysteines and/or histidines coordinating the zinc atoms were set to be deprotonated to obtain the starting structure. For RelA chains A, B along with residues 213-226 from chain E and 201-214 from chain F were retained in order to obtain the starting structure. Missing protein residues at the terminal ends for all systems were not modeled. For all systems, the bases of the DNA in the crystal structure were mutagenized in silico to match those used in the binding experiments. Wherever applicable, divalent ions were modeled using the Li/Merz 12-6 ion parameters[88]. Production runs of 200 ns (Myc/Max, Max/Max, RelA) or 500 ns (Ets1, p53, CTCF, Egr1, GR, Elk1,) were then carried out and extended to achieve convergence of the RMSD of the protein-DNA complex if necessary. For proteins bound to mismatched DNA sites, we chose not to simulate the mismatches A-A, A-C, and C-C, given the lack of a stable base pairing geometry for A-A[47] and the tendency of A-C and C-C to undergo protonation-dependent structural changes in order to form stable base pairing geometries[48,49]. Protonation-dependent base pairing conformational equilibria are susceptible to being highly influenced by protein binding[50], and are also difficult to model computationally[51]. We simulated one mismatch per protein, focusing on G-T and C-T mismatches, as well as G-G and G-A in specific cases, given their stable base pairing geometries[52-56] and ability to be modeled reliably computationally[46].

H-bond and buried surface area analyses

Based on the MD simulations of protein-DNA complexes (described above), we calculated the number of hydrogen bonds between the protein and DNA using the hbond program of the CPPTRAJ[89] utility of AMBER. A heavy atom donor (Nitrogen/Oxygen/Sulphur) with an attached hydrogen was defined as forming a H-bond with an acceptor heavy atom (Nitrogen/Oxygen/Sulphur) if the donor-acceptor distance was less than 3 Å and the acceptor-hydrogen-donor angle was greater than 135°. The calculation was performed by considering the protein and DNA atoms as donors/acceptors separately and adding the resulting number of H-bonds. The number of H-bonds was then averaged over all conformers in the MD trajectory for a given system. The results of the H-bond analysis are summarized in **Supplementary Table 7**.

The buried surface area for a given conformation of a protein-DNA complex was defined as the difference between the sum of surface areas of the protein and DNA when considered in isolation, versus the protein-DNA complex. The computation of the surface area was performed using the molsurf program in the CPPTRAJ[89] of AMBER, with a probe radius of 1.4 Å. The surface area was averaged over all conformers from the MD simulations to obtain the buried surface area for the system. All the residues of the protein and DNA were considered for the calculation.

Protein expression and purification

Protein expression and purification for SaMBA assays: Full-length human proteins Ets1, Elk1, Gabpa, Runx1, E2f1, Six6, Ap2a, and Gata1, with N-terminal GST tags, were expressed as purified as described in [21]. Full-length *S. cerevisiae* Cbf1, with N-terminal GST tag, was overexpressed in *E. coli* BL21 (DE3) cells (NEB) and purified by as described in [57]. Full-length human Myc (c-Myc), Max, and Mad (Mad1) with C-terminal 6xHis tags, as well as full-length untagged Max[90] were generously provided by Peter Rahl and Richard Young (Whitehead Institute and MIT). All Myc SaMBA experiments were performed using both Myc and Max on the same microarray, using a 5 times higher concentration of Myc to ensure that mostly Myc:Max heterodimers, and not Max:Max homodimers, are formed. Similarly, all Mad SaMBA experiments were performed using both Mad and Max on the same microarray, in a 5:1 ratio. Human Egr1, residues 335-423, was expressed and purified as in [59], and generously provided by Junji Iwahara (University of Texas Medical Branch). Full-length human p53 with N-terminal GST tag was purchased from Abcam (catalog #: ab43615). Full-length human TBP with HIS tag was purchased from Excellgen (catalog #: RP-54). Full-length human CTCF with N-terminal GST tag was purchased from Abnova (catalog #: H00010664). Full-length human Creb1/Crem/Atf1 proteins with N-terminal HIS tags were purchased from Origene (catalog #s: TP760318/TP760397/TP721193). Human phosphorylated Stat3 (residues 128-715) with 6xHis tag, and human RelA (residues 20-290) were generously provided by Dr. Eyal Arbely (Ben-Gurion University of the Negev). Human GR DNA-binding domain (residues 418–517) with an N-terminal 6xHis-Sumo-tag was expressed and purified as in [58], and generously provided by Dr. Eric Ortlund (Emory University).

Ets1 expression and purification for FA: Recombinant Ets-1 (murine residues 280 to 440) was expressed and purified from BL21*(DE3) *E. coli* harboring a clone in pET28b (a gift from L. McIntosh, University of British Columbia) as previously reported[62]. In brief, an overnight culture was inoculated into liter-scale LB media and grown at 37 °C to an OD of 0.6. Expression was induced by the 0.5 mM isopropyl β-D-1-thiogalactopyranoside for 4 hours at 25 °C. Cells were harvested and re-suspended in 0.1 M TrisHCl (pH 8.0) containing 0.5 M NaCl and 5 mM imidazole. All downstream buffers contained 0.5 mM Tris(2-carboxyethyl)phosphine (TCEP) hydrochloride to prevent cysteine oxidation. Cells were lysed by sonication followed by centrifugation. The cleared lysate was partially purified on Co-NTA resin via the C-terminal 6xHis tag on the target. The eluate was dialyzed overnight against 10 mM NaH_2PO_4/Na_2HPO_4 (pH 7.4) containing 0.15 M NaCl and 10 U of bovine thrombin to remove the C-terminal 6×His tag. The construct was polished on Sepharose SP (GE) column equilibrated with the same buffer, under the control of a Bio-Rad NGC instrument, and eluted over an NaCl gradient at ~0.5 M. Purified Ets-1 was dialyzed extensively into final buffer of 10 mM NaH_2PO_4/ Na_2HPO_4 (pH 7.4) with 0.15 M NaCl. Ets-1 concentration was determined by UV absorption at 280 nm using an extinction coefficient of 39,880 M^{-1} cm^{-1}.

GR expression and purification for EMSA: The DNA-binding domain (DBD) of GR was purified as described previously[63]. Briefly, human cDNA for the DBD (AAs 418-506) was cloned into an N-terminal his6-tagged vector (pET-28a), transformed into bacteria (BL21 Gold, DE3, Agilent), and grown to a density of OD600 = 0.4 @ 37˚C in LB w/ 50 µg/ml kanamycin. Protein expression was induced by reducing the temperature to 23˚C, adding 0.5 mM IPTG (Isopropyl β-D-1-thiogalactopyranoside) when the density reached OD600 = 0.8. Bacteria were harvested after 4 hours by centrifugation (4,000g, 15 minutes), resuspended in lysis buffer (50 mM Tris, pH 8.0, 500mM NaCl, 15mM imidazole, 1mM PMSF), snap frozen in liquid nitrogen, and stored at -80˚C. Bacteria were thawed slowly by shaking at 30˚C, lysed (Emulsiflex C-3), and then spun down in an ultracentrifuge (40,000 RPM, 1 hour) to remove cell debris. The supernatant was then bound to a nickel column (HisTrap HP, GE Biosciences), washed, and eluted with a gradient from 25 mM to 500 mM Imidazole. The cleanest fractions were pooled, and incubated with thrombin (Sigma T4648, 100U) while dialyzing into 20 mM TrisHCl, pH 7.5, 50 mM NaCl, 2.5 mM $CaCl_2$, 1 mM DTT at 4˚C overnight. Precipitate was removed by 0.2 µM filtration. The protein was then purified by binding to a strong cation exchange column (SP Sepharose HP, GE Biosciences) and eluting with a gradient from 50 mM NaCl to 500 mM NaCl (peak ~200mM NaCl). The cleanest fraction were pooled, concentrated, and then polished by running over a gel filtration column (Superdex 200, GE Biosciences) in 20 mM HEPES, 100 mM NaCl, 1mM DTT. Fractions were pooled, quantified, tested for activity by EMSA against MM3, aliquoted, snap frozen in liquid nitrogen, and stored at -80˚C.

TBP expression and purification for X-ray crystallography: TBP protein was purified as previously described[61]. Briefly, *E. coli* C41(DE3) cells were transformed with a pET15b plasmid encoding the *Arabidopsis thaliana* TATA box binding protein (TBP) with a His-6 tag on the N-terminus (Genscript). Cells were grown at 37 ºC to an OD_{600} of 0.5, induced at 15 ºC with 0.5 mM ITPG overnight and pelleted the next day. The cell pellets were reconstituted with buffer A (5% glycerol, 4 mM $MgCl_2$, 600 mM NaCl, 40 mM 2-(*N*-morpholino)ethanesulfonic acid (MES) pH 7.2), lysed twice using a microfluidizer and pelleted. The supernatant was loaded onto a Cobalt-NTA column and the column was washed with increasing concentrations of imidazole in buffer A. TBP was eluted in buffer A containing between 50 and 1000 mM imidazole. β-mercaptoethanol (BME) was added to a final concentration of 5 mM to these fractions directly after elution. The resultant TBP was >95% pure at this stage and was concentrated to 5 mg/ml via centrifugation with a 10 kDa cutoff filter (Amicon). The protein was stored at -80 ºC in 20 mM (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) HEPES KOH pH 8, 100 mM KCl, 20% glycerol, 1 mM $MgCl_2$, 1 mM $CaCl_2$. Protein yields were typically 2 to 5 mg/l. Concentrations were determined by UV-vis using a calculated molar absorption coefficient of 10.5 mM^{-1} cm^{-1}.

Protein binding and antibody steps for SaMBA assays

Protein binding reactions were performed in the same conditions as in PBM protocols[20]. The binding buffer, unless otherwise specified, was a 185-µl solution containing PBS / 2% (wt/vol) milk / 51.3 ng/µl salmon testes DNA (Sigma) / 0.2 µg/µl bovine serum albumin (NEB). For Egr1, the binding buffer was: 10 mM Tris-HCl (pH 7.5), 150 mM KCl, and 0.2 µM $ZnCl_2$, 2% (wt/vol) milk ,51.3 ng/µl salmon testes DNA, 0.2 µg/µl bovine serum albumin. For TBP, the binding buffer was 10 mM HEPES, 70 mM KCl, 10 mM $MgCl_2$, 1 mM EDTA, 2% (wt/vol) milk, 51.3 ng/µl salmon testes DNA, 0.2 µg/µl bovine serum albumin. For CTCF, the binding buffer was PBS with 5 mM $MgCl_2$, 0.1 mM $ZnSO_4$[91], 2% (wt/vol) milk, 51.3 ng/µl salmon testes DNA, 0.2 µg/µl bovine serum albumin. For Creb1/Crem, the binding buffer was 25 mM Tris-HCl (pH 7.4), 1 mM DTT, 0.5 mM EDTA, 2% glycerol, 5 mM $MgCl_2$, 50 mM KCl, 25 mM boric acid, 2% (wt/vol) milk, 51.3 ng/µl salmon testes DNA, 0.2 µg/µl bovine serum albumin. For Atf1, the binding buffer was 50 mM potassium acetate, 20 mM Tris-acetate, (pH 7.9) 10 mM magnesium acetate, 1 mM DTT, 2% (wt/vol) milk, 51.3 ng/µl salmon testes DNA, 0.2 µg/µl bovine serum albumin. For RelA, the binding buffer was 6mM HEPES, 80 mM KCl, 0.5 mM EDTA, 2% (wt/vol) milk, 51.3 ng/µl salmon testes DNA, 0.2 µg/µl bovine serum albumin. For GR, the binding buffer was 50mM KCl 5% glycerol 20mM Tris-HCl (pH 8) ,2% (wt/vol) milk, 51.3 ng/µl salmon testes DNA, 0.2 µg/µl bovine serum albumin.

Pre-incubated protein binding mixtures were applied to individual chambers and incubated for 1 h with the double-stranded DNA chip. The chips were washed once with PBS / 0.5% (vol/vol) Tween-20 for 3 min, and then once with PBS / 0.01% Triton X-100 for 2 min. After the protein incubation and washing steps, Alexa647-conjugated GST antibody (Cell Signaling Technology, Catalog #3445; dilution 1:30), Alexa488-conjugated GST antibody (Invitrogen, Catalog #: A-11131; dilution 1:30); Penta·His Alexa647-conjugated antibody (Qiagen, Catalog #: 35370; dilution 1:20), Penta·His Alexa488-conjugated antibody (Qiagen, Catalog #: 35310; dilution 1:20), or Goat anti-Rabbit IgG Alexa647 antibody (ThermoFisher, Catalog #: A21244; dilution 1:30) in PBS / 2% milk were applied on the chip for 1 h at room temperature. For RelA, we used rabbit polyclonal antibody purchased from Origene (Catalog #: TA890002), dilution 1:150. Washing steps after each incubation step were performed in Coplin jars at room temperature, on a shaker at 125 r.p.m. The fluorescent signal (Alexa647 or Alexa488) of bound TFs for each DNA spot was measured using a GenePix 4400A microarray scanner. Multiple replicates of each sequence were used to quantitatively compare the binding signals between sequences, and to statistically assess the significance of binding differences using a one-sided Wilcoxon-Mann-Whitney test, corrected for multiple hypotheses testing using the Benjamini-Hochberg procedure.

Validation of SaMBA data using TF binding affinity measurements

p53 EMSA binding affinity measurements: Radiolabeled and gel-purified DNA hairpin duplexes (concentration < 0.1 nM) and increasing amounts of p53CT were incubated at 21°C for 2 h in a buffer containing 50 mM Tris-HCl (pH 7.5), 10 mM MgCl2, 1 mM ATP, 25 µg/ml BSA, 10% glycerol, 10 mM DTT, and 100 mM KCl. By taking into account the protein dilution buffer of 20 mM Tris-HCl (pH 7.5) and 120 mM NaCl, the total ionic strength in the binding buffer was 310 mM. Complexes were resolved from free DNA by electrophoresis on native gels (6% 37.5:1 acrylamide/bisacrylamide ratio). Samples were loaded on a running gel at 550 V and 21 °C in 1x TG (25 mM Tris, 190 mM glycine, pH 8.3) until the bromophenol blue dye migrated 10 cm. Dried gels were quantified using a GE Typhoon FLA7000 phosphoimager. Each band was analyzed separately using Cliqs version 1.1 (TotalLab Ltd., UK), using a regular two binding site model. For gel patterns showing only p53CT/DNA tetrameric complexes, zero values were added to account for the unobserved dimer bands. The following equations were used:

1. $\Theta_0 = 1/ (1 + K_{a1}*[P] + K_{a2}*[P]^2)$
2. $\Theta_1 = K_{a1}*[P] / (1+K_{a1}*[P] + K_{a2}*[P]^2)$
3. $\Theta_2 = K_{a2}*[P]^2 / (1 + K_{a1}*[P]+K_{a2}*[P]^2)$

Θ_i, the fraction of DNA molecules with *i* bound p53CT dimers, was calculated from the equation: Θ_i = (PSL—bg)i/Σi(PLS-bg)i, where PSL is photosimulated luminescence, bg is the background and the summation is over all bands in a given lane. [P] is the protein concentration, and K_{a1} and K_{a2} are the macroscopic association binding constants for the dimeric and tetrameric species, respectively. Using this procedure, the macroscopic dissociation binding constant (K_d=1/K_{a2}) for the dominant tetrameric species, and the free energy of binding for the tetrameric species were computed for 10 different hairpin duplexes (6 hairpins containing mismatches, and 4 Watson-Crick sequences; **Supplementary Table 3**). Measurements for each duplex were performed in six replicate experiments, and the average binding affinity was used to validate both increases and decreases in binding due to mismatches, as observed in SaMBA (**Fig. 1f, Extended Data Fig. 3e**). Gel images are available at:
https://figshare.com/projects/DNA_mismatches_reveal_conformational_penalties_in_protein-DNA_recognition/83663.

Ets1 fluorescence anisotropy (FA) binding affinity measurements: DNA binding by murine Ets-1 (residues 280 to 440, termed Ets1ΔN280) was measured by steady-state fluorescence polarization, essentially as described in [70], using a Cy3-labeled DNA probe encoding the Ets-1 binding sequence 5'-CGCACCGGATATCGCA-3'. In brief, 0.5 nM of DNA probe and 10 nM Ets1ΔN280 were titrated with one of five unlabeled DNA duplexes (3 duplexes containing a mismatch, and 2 Watson-Crick duplexes; **Supplementary Table 3**) in 10 mM TrisHCl (pH 7.4) containing 0.1% w/v BSA and 0.15 M NaCl. Steady-state anisotropies $\langle r \rangle$ were measured at 595 nm in 384-well black plates (Corning) using a Molecular Dynamics Paradigm plate reader with 530 nm excitation. The signal represented the fractional bound DNA probe F_b scaled by the limiting anisotropies of the bound $\langle r_\mathrm{b} \rangle$ and unbound states $\langle r_\mathrm{u} \rangle$ as follows:

4. $\langle r \rangle = F_\mathrm{b}\left(\langle r_\mathrm{b} \rangle - \langle r_\mathrm{u} \rangle\right) + \langle r_\mathrm{u} \rangle$

where F_b is a function of the total titrant (unlabeled DNA) concentration as taken. F_b is modeled by a single-site competitive model in which the protein (E) binds either the probe (P) or unlabeled DNA (U), but not both[92]:

5. $\mathrm{EP} \overset{[P]/K_\mathrm{P}}{\rightleftarrows} \mathrm{E} \overset{[U]/K_\mathrm{U}}{\rightleftarrows} \mathrm{EU}$

The binding polynomial, which is cubic in $[\mathrm{P}]_\mathrm{b} = F_\mathrm{b}[\mathrm{P}]_\mathrm{t}$ is[70]:

6.
$$0 = \varphi_0 + \varphi_1[\mathrm{P}]_\mathrm{b} + \varphi_2[\mathrm{P}]_\mathrm{b}^{\ 2} + \varphi_3[\mathrm{P}]_\mathrm{b}^{\ 3}$$
$$\begin{cases} \varphi_0 = -K_\mathrm{P}[\mathrm{U}]_\mathrm{t}^{\ 2}[\mathrm{E}]_\mathrm{t} \\ \varphi_1 = K_\mathrm{U}K_\mathrm{P}[\mathrm{U}]_\mathrm{t} + K_\mathrm{U}[\mathrm{P}]_\mathrm{t}[\mathrm{U}]_\mathrm{t} + K_\mathrm{P}[\mathrm{U}]_\mathrm{t}^{\ 2} + 2K_\mathrm{P}[\mathrm{U}]_\mathrm{t}[\mathrm{E}]_\mathrm{t} - K_\mathrm{U}[\mathrm{U}]_\mathrm{t}[\mathrm{E}]_\mathrm{t} \\ \varphi_2 = -K_\mathrm{U}K_\mathrm{P} + K_\mathrm{U}^{\ 2} - K_\mathrm{U}[\mathrm{P}]_\mathrm{t} - 2K_\mathrm{P}[\mathrm{U}]_\mathrm{t} + K_\mathrm{U}[\mathrm{U}]_\mathrm{t} - K_\mathrm{P}[\mathrm{E}]_\mathrm{t} + K_\mathrm{U}[\mathrm{E}]_\mathrm{t} \\ \varphi_3 = K_\mathrm{P} - K_\mathrm{U} \end{cases}$$

where the subscript “t” denotes total concentration. Triplicate measurements were performed for each duplex. Using these equations we computed the dissociation binding constant for 5 duplexes overall, and confirmed both increases and decreases in Ets1 binding affinity due to mismatches (**Supplementary Table 3**).

GR EMSA binding affinity measurements: Four DNA duplexes (**Supplementary Table 3**) were tested. One strand of each duplex was unlabeled and the second strand was labeled at the 5’ end with an IR700 (Integrated DNA Technologies, IDT). Strands were resuspended in water to a concentration of ~100 µM, and then mixed 1:1 with 10x annealing buffer (200 mM HEPES, pH 7.4, 1 M NaCl, 50 mM $MgCl_2$), heated to 95˚C for 5 minutes, and cooled slowly to anneal (for a final concentration of 45 µM duplex). Duplexes were then diluted to 10 mM in binding buffer (20 mM Tris-HCl pH8, 50 mM KCl, 40 ng/ul salmon sperm, 200 ng/ul BSA, 5mM $MgCl_2$, 1mM EDTA, 1 mM DTT, 10% glycerol) and mixed 1:1 with titrations of GR-DBD (also diluted in binding buffer) from 20 µM to 10 nM (final DNA concentration: 10 nM, final protein 20 µM to 10 nM). After 1 hour of equilibration on ice, reactions were separated on 8% polyacrylamide gels (19:1 acrylamide:bis-acrylamide, 1x tris-glycine buffer) by running at 200 V in 1x Tris-glycine buffer for 40 minutes at 4°C. Wet gels were rinsed and then imaged on a LiCor Odyssey Fc (IR700 channel). The DNA band for each species was quantified using LiCor Image Studio (v 5.2.5). The dissociation constants of each of these duplexes were computed using equations 1-3 above (see p53 EMSA binding affinity measurements). Measurements were performed in triplicate on different days using different protein aliquots. The macroscopic dissociation binding constant (K_D=1/K_{a2}) and the free energy of binding for the dimeric species were computed for each measurement (**Supplementary Table 3**). To avoid the self-hybridization of the probes in EMSA, one of the two GR half-sites and the spacer between them were mutated compared to the site used in SaMBA; positions known to be critical for GR binding were kept constant. The average binding affinities of these oligos was used to validate the increases and decreases in binding affinity observed in SaMBA. Gel images are available at:
https://figshare.com/projects/DNA_mismatches_reveal_conformational_penalties_in_protein-DNA_recognition/83663

Crystallization and structure determination of TBP-mismatch DNA complexes

The duplex DNA sites used for crystallization were generated by solubilizing each DNA strand in 50 mM sodium cacodylate pH 6.5 to a final concentration of 2 mM and mixing the two strands to form the desired duplex stoichiometrically (1:1), then heating the mixture at 90 ºC for 5 min followed by cooling on ice at 4 ºC. To obtain TBP crystals with various mismatch DNA sites, the protein (at 5 mg/ml) was mixed at a 1:2 molar ratio with a given DNA duplex and this mixture was diluted to twice its volume in a buffer consisting of 25 mM Tris pH 7.5, 150 mM NaCl, 2.5 % glycerol. The resultant mixture was concentrated ~4-fold using 10 kDa cutoff microcons. The resultant protein-DNA complexes were then used in vapor diffusion crystallization screens. Screens employed were Wizard I-IV (Rigaku), cryo I and cryo II (Rigaku), PegRx 1 and PegRx 2 (Hampton Research), Hampton screen 1 (Hampton research), SaltRx 1 and SaltRx 2 (Hampton Research). Crystal hits for each complex were optimized by fine screening around the crystallization condition, typically by reducing the precipitant concentrations. Fine screening for each TBP-DNA complex resulted in large, well diffracting crystals suitable for data collection.

Two crystal forms of TBP in complex with the C-C mismatch DNA site (5´-TGCCCCTTTATAGC-3´ annealed with 5´-GCTATAAACGGGCA-3´; henceforth referred to as "CC1") were obtained. The first crystal form (TBP-CC(1a)) was produced by mixing the complex 1:1 with a crystallization reservoir solution consisting of 0.1 M sodium acetate pH 4.5 and 3.2 sodium formate and equilibrating the drop over the reservoir. Crystals were cryo-preserved by dipping them in the reservoir solution supplemented with 20% ethylene glycol before looping and placement in the cryo-stream. The second crystal form of this TBP-DNA complex (TBP-CC(1b)) was obtained by mixing the complex 1:1 with a crystallization reservoir consisting of 0.1 M HEPES pH 7.5 and 3.4 sodium formate and equilibrating the drop over the reservoir. Crystals were cryo-preserved by dipping them in the reservoir solution supplemented with 22% ethylene glycol before looping and placement of the loop in the cryo-stream. Crystals of TBP with the C-C mismatch DNA site (5´-TGCCCCTTTATAGC-3´ annealed with 5´-GCTATAAAACGGCA-3´; henceforth referred to as "CC2") were obtained by mixing the complex 1:1 with a crystallization reservoir solution consisting of 0.1 M sodium acetate pH 4.5 and 30% PEG 400 and equilibrating the drop over the reservoir. As the crystallization solution was a cryo-solvent, the crystals could be looped from the drop and placed in the cryo-stream directly. Crystals of TBP with the A-C mismatch DNA site (5´-TGCCCCTTTATAGC-3´ annealed with 5´-GCTATAAAGGGCA-3´) were obtained by mixing the complex 1:1 with a crystallization reservoir solution consisting of 0.1 M sodium acetate pH 4.5 and 38% ethylene glycol and equilibrating the drop over the reservoir. This crystallization condition was a cryo-solvent and hence the crystals could be looped from the drop and placed in the cryo-stream directly.

Data for all the crystals were collected at the Advanced Light Source (ALS) on beamlines 8.3.1 and 5.0.1. The data were processed with MOSFLM 7.3.0 and scaled with SCALA in ccp4i 7.0.078[76,77]. The structures were solved by molecular replacement MR (with MolRep in ccp4i 7.0.078) using the 1QNE structure with the waters removed, as a search model. After refinement in Phenix (version 1.17)[78], the structures were manually rebuilt in O (version 8)[79]. MolProbity (version 4.5) was used to guide the process of refitting and refinement[80]. See **Extended Data Table 1** for the final data collection and refinement statistics for each structure.

Supplementary Discussion

Systematic analysis of DNA mismatch conformations

Mismatches are proposed to adopt multiple conformations that are undergoing exchange in solution (**Extended Data Fig. 2a**). For example, G(*syn*)-G(*anti*) Hoogsteen mismatches exist in dynamic equilibrium with G(*anti*)-G(*syn*); G-T wobble mismatches exist in dynamic equilibrium with Watson-Crick-like bps stabilized by rare tautomeric and anionic bases; G-A mismatches adopt multiple conformations involving protonated adenine and either *syn* or *anti* bases.

To systematically analyze the ensemble behavior of all mismatches and examine their conformational penalty on protein-DNA recognition, we performed MD simulations on unbound DNA with different single mismatches (Methods). Here we describe the results of our analyses of base pair structural parameters (shear, stretch, stagger, buckle, propeller twist, opening, and C1′-C1′ distance) of mismatches. Note that: (1) transient and rare species, such as the tautomeric form of the G-T mismatch, were not considered in this analysis; (2) due to the inability of glycosidic bond rotations to occur under timescales conventionally accessible by MD simulations, different Hoogsteen-type mismatches, i.e. G(*syn*)-G(*anti*) and G(*anti*)-G(*syn*), are modeled and simulated separately; (3) protonated or anionic base pairs, such as A^{+}-C, C^{+}-C and G(syn)-A^{+}(anti), are ignored in this analysis, for simplicity, since they are highly dependent on pKa and pH (Methods). Summary descriptions of the ensemble behavior of different mismatches in the unbound DNA MD simulation are presented below and in **Extended Data Fig. 2c**.

Purine-pyrimidine mismatches: Purine-pyrimidine mismatches consist of G-T and A-C. The G-T mismatch remains in a stable wobble geometry with shear around -2 Å, accompanied by a slight constriction (stretch and C1′-C1′ distance) during the MD simulation, compared to A-T and G-C base pairs. The A-C mismatch remains mostly as one H-bonding geometry with C translated to the major groove (similar to G-T), whereas in a prior computational investigation, A-C mismatches exist dynamic equilibrium of either A or C translating to major groove[46]. Interestingly, in our MD simulation, A-C mismatch is slight positively buckled.

Pyrimidine-pyrimidine mismatches: Pyrimidine-pyrimidine mismatches consist of T-T, C-T, and C-C. T-T mismatch remains in a wobble geometry with shear around ±2 Å. In contrast to the G-T wobble mismatch, the T-T mismatch exists in rapid dynamic equilibrium of both inter-converting wobble geometry with either one of the T translating to the minor groove direction, as shown in the MD simulation. Despite this rapid dynamic equilibrium, the T-T mismatch is still constricted with C1′-C1′ distance of 8-9.5 Å during the MD simulation. Similar to T-T, the C-T mismatch is also constricted with two H-bonds stably formed for most of the time. However, C-T can transiently adopt a high-energy conformation with only one H-bond formed, and it is partially open (C1′-C1′ distance around 10 Å), potentially due to the repulsion between T-O2 and C-O2. The entire C-T MD trajectory is comprised of approximately 5% of these high-energy species. The C-C mismatch is partially constricted with C1′-C1′ distance around 9.8 Å, due to its unstable one H-bonding geometry. All the pyrimidine-pyrimidine mismatches were still stacked in the helix without swinging outside the helix during the MD simulation.

*Purine-purine mismatche*s: Purine-purine mismatches consist of G-G, G-A, and A-A. The G-G mismatch basically only exists as a mixture of inter-converting G(*syn*)-G(*anti*) conformations in a sequence-dependent manner [52]. During the MD simulation, G(*syn*)-G(*anti*) does not experience syn-anti inter-converting exchange. The C1′-C1′ distance of G(*syn*)-G(*anti)* is around 11.2-11.5 Å, which is larger than for the canonical G-C base pair. G(*anti*)-A(*anti*) and G(*syn*)-A(*anti*) geometries were considered for the G-A mismatch. G(*anti*)-A(*syn*) is partially expanded, with C1′-C1′ distance around 11.5 Å, whereas G(*anti*)-A(*anti*) is strongly stretched, with a large C1′-C1′ distance around 12.8 Å. The A-A mismatch adopts an unstable stretched A(*anti*)-A(*anti*) geometry, with only one H-bond.

We also performed an independent analysis of DNA mismatch structures within helical context, from a PDB structural survey (Methods). A detailed summary of the structural survey is listed in **Supplementary Table 9**. We note that: (1) we only consider entries with a single mismatch embedded by two Watson-Crick base pairs; and (2) modified bases were excluded from the survey (Methods). Importantly, most of T-T and C-T mismatches are constricted (C1′-C1′ distance < 9.5 Å), with only one exception (PDB: 2LL9), whereas all G-T and A-C mismatches are not constricted (C1′-C1′ distance > 10.0 Å). The C-C mismatch has only three entries, with only one of them constricted (PDB: 2NL8)**.** Purine-purine mismatches (G-A, A-A, G-G) are also not constricted, consistent with a previous survey about purine-purine mismatches[52]. Additionally, consistent with MD simulation, G-T, A-C and T-T mismatches adopt mostly wobble conformations.

Transcription factor-DNA binding energetics for Watson-Crick and mismatched DNA

The binding reaction between DNA and proteins can be written as follows:

$$\mathrm{DNA}_{free} + \mathrm{P} \overset{K_{\text{binding}}}{\rightleftarrows} P:\mathrm{DNA}_{bound} \qquad \Delta G_{binding} = -k_B T \ln(K_{binding})$$

where DNA$_{\text{free}}$ and DNA$_{\text{bound}}$ refer to the DNA when it is free and when bound to protein, respectively, and ΔG is the free energy of the overall binding reaction. This binding reaction, for each sequence, assuming that DNA conformational changes in the protein (P) are constant for all DNA sequences, and set to be zero, can be written as the sum of these two individual reactions involving the conformational change of the DNA by itself:

$$\mathrm{DNA}_{free} \overset{K_{\text{pen}}}{\rightleftarrows} \mathrm{DNA}_{bound} \qquad \Delta G_{pen} = -k_B T \ln(K_{pen})$$

and the binding of the bound form-like DNA to the protein:

$$\mathrm{DNA}_{bound} \overset{K_{\text{interact}}}{\rightleftarrows} P:\mathrm{DNA}_{bound} \qquad \Delta G_{\text{interact}} = -k_B T \ln(K_{\text{interact}})$$

with the associated free energies ΔG$_{\text{pen}}$ and ΔG$_{\text{interact}}$, respectively. From the above definitions, it follows that the net free energy of binding (ΔG) is the sum of free energies of changing DNA conformation (ΔG$_{\text{pen}}$) and the energy of protein-DNA complex formation (ΔG$_{\text{interact}}$), i.e.:

$$\Delta G_{binding} = \Delta G_{pen} + \Delta G_{\text{interact}}$$

These equations can be used for wild-type (Watson-Crick) sequences: $\Delta G_{binding}(WT) = \Delta G_{pen}(WT) + \Delta G_{\text{interact}}(WT)$, as well as for mismatched sequences: $\Delta G_{binding}(MM) = \Delta G_{pen}(MM) + \Delta G_{\text{interact}}(MM)$.

Next, if we assume that the protein-DNA contacts are unchanged between wild-type and mismatched DNA, as we observed experimentally for TBP (**Fig. 3f,g, Extended Data Fig. 7c**), the energies for the protein-DNA interactions of wild-type and mismatched DNA will be equal to each other, i.e.:

$$\Delta G_{\text{interact}}(WT) = \Delta G_{\text{interact}}(MM)$$

Under these assumptions, the extent to which a mismatch favors a bound-form-like conformation, i.e. the extent to which it prepays the conformational penalty, determines the extent to which it favors protein-DNA binding. Consequently, the differences in binding between wild type and mismatch DNA would arise from the differences in the conformational penalties ΔG$_{\text{pen(MM)}}$ and ΔG$_{\text{pen(WT)}}$.

$$\begin{aligned}\Delta\Delta G_{binding} &= \Delta G_{binding}(MM) - \Delta G_{binding}(WT)\\ &= \Delta G_{pen}(MM) + \Delta G_{\text{interact}}(MM) - \left(\Delta G_{pen}(WT) + \Delta G_{\text{interact}}(WT)\right)\\ &= \Delta G_{pen}(MM) - \Delta G_{pen}(WT)\end{aligned}$$

Double mutant cycles: the energetic effects of mismatches versus Watson-Crick mutations on TF binding

In TF binding sites, both single-base mismatches and base-pair mutations change base identities and consequently could alter the direct base readout by TFs. While mismatches change the identity of a single base, base-pair mutations change two (paired) bases. For every given mismatch there is a corresponding mutation in which the same base is altered, and thus could potentially have a similar impact on TF binding, in the absence of additional conformational effects. Similarly, for every given mutation, there are two mismatches that together exhibit identical changes to the base identities as in the base-pair mutant. Importantly, base conformations are dependent on the two bases composing a base-pair (Watson-Crick or mispair; **Extended Data Fig. 2**) and that, in turn, can lead to energetic dependencies between the partnering bases.

The double mutant cycle is a powerful technique, commonly used in protein science, to learn about dependencies between two amino acid residues in a protein[93-95]. A double-mutant cycle comprises of the wild-type species, two single-mutant species and the corresponding double mutant. If the change in free energy associated with a certain property of a double mutation differs from the sum of changes due to the single mutations, then the residues at the two positions are considered dependent. We leveraged this method in order to study nucleic acid interactions with TFs, in a similar manner to how it is typically used for amino acid mutants. We compared the binding affinities of mismatches with the corresponding binding affinities of the wild-type and mutated Watson-Crick sequences, focusing on the seven TFs with available calibration data in our study (Methods, **Extended Data Fig. 4c, Supplementary Table S3,S4**).

For ~55% of the 1500 mutations tested, the sum of the energetic gains/losses from the two individual single-base mutations was equal, within experimental noise, to the binding energy change due to the base-pair mutant, e.g. at position 7 in the Ets1 site 1 shown in **Extended Data Fig. 4c**, where $\Delta\Delta G_{AT\text{->}A\underline{\mathbf{G}}}$ + $\Delta\Delta G_{AT\text{->}\underline{\mathbf{C}}T}$ ≈ $\Delta\Delta G_{AT\text{->}\underline{\mathbf{CG}}}$. (If we focus on cases where at least one of the two mismatches significantly affects binding, the percentage drops to ~42%. And if we further restrict our analysis to cases where at least one of the two mismatches *increases* binding, then the percentage is 48%.) For such cases, a simple model of additive contributions from the DNA bases is sufficient to explain the observed binding changes (i.e. through changes in H-bonding and other protein-base interactions that depend purely on the identity of the base).

For the remaining ~45% of the mutations tested, the sum of the energetic gains/losses from the two individual single-base mismatches was not equal to the sum of the corresponding base-pair mutant (**Fig. 2e**, **Extended Data Fig. 4c, Supplementary Table 4**). These cases demonstrate the existence of energetic coupling between partnering bases on TF binding. In cases where the residual energetic difference ($\Delta\Delta G_{mismatch1}$ + $\Delta\Delta G_{mismatch2}$ - $\Delta\Delta G_{mutation}$) has a negative value, the cumulative effect of the two mismatches is more favorable (or less detrimental) to TF binding than the effect of the Watson-Crick mutation. For example, in **Fig. 2e** (upper panel**)**, Ets1 binding to the mismatches at position 7 (in aggregate, i.e. $\Delta\Delta G_{mismatch1}$ + $\Delta\Delta G_{mismatch2}$) is highly preferred compared to the corresponding base-pair mutation ($\Delta\Delta G_{mutation}$), which indicates that the G base that increased binding in the G-G mismatched conformation is not equally preferred in the Watson-Crick conformation. We also observed the opposite case, in which the base identity alone cannot explain the *reduction* in binding affinity due to mismatches. For example, in **Fig. 2e** (lower panel), each of the mismatches at position 4 results in a decrease in Ets1 binding affinity, while the corresponding base-pair mutation exhibits the opposite trend, i.e. an increase in binding affinity. If the decrease due to introduction of the T (in the $\underline{\text{T}}$-T mismatch) and the A (in the A-$\underline{\text{A}}$ mismatch) were simply due to changes in the base identity, introducing both T and A (in the T-A base-pair mutation) would be expected to decrease Ets1 binding even further, which is opposite to the observed increase in binding affinity.

Ets1 binding to DNA mismatches

Multiple mechanisms are likely contributing to the enhanced Ets1 binding at mismatched DNA sites, as observed in SaMBA (**Fig. 2b, Extended Data Fig. 5**). Here, we summarize these mechanisms and describe how our findings on Ets1-mismatch binding relate to the extensive structural data on DNA binding by ETS family proteins.

Recognition of Watson-Crick DNA by members of the ETS family, including Ets1, has been widely studied[96,97,98]. ETS domains have been shown to display sequence selectivity for ~10 bps[97], numbered 1-10 in our study (e.g. **Fig. 2b, Extended Data Fig. 5a**). Consistent with this prior knowledge, SaMBA revealed significant effects on Ets1 binding from DNA mismatches (p-values on the order of 10^{-5}-10^{-6}) exactly at these 10 positions (**Supplementary Table 1b**).

Upon DNA binding, the ETS recognition helix is inserted into the major groove at the central GGA(A/T) core, forming strong H-bonds with two guanines (positions 6 and 7) (**Extended Data Fig. 5d**). DNA duplexes bound by ETS domains are typically distorted from an ideal B-DNA geometry[97], with the most significant distortions occurring at the positions of the inserted recognition helix. Remarkably, the strongest increases in Ets1 binding are observed for mismatches at these distorted positions (G-A at position 6 and G-G at position 7), a consistent trend across Ets1 binding sites (**Extended Data Fig. 5a,k**). Both of these mismatches are at bases opposite to Ets1-contacting bases, thus retaining the two guanines with the strongest H-bonds.

For the G-A mismatch at position 6, which increases Ets1 binding by ~2.3k_BT (**Fig. 2b**), computational analyses of the distortions induced by G-A revealed that G-A mimics the stretch, C1’-C1’ distance, and minor groove width of Ets1-bound DNA (**Supplementary Table 6, Extended Data Fig. 5b)**: in the Ets1-DNA structure (PDB ID: 1K79), the G-C base pair marked in 5’-TT**C**C-3’ (reverse complement 5’-G**G**AA-3’) has wide C1’-C1’ distance and large stretch value, similar to the G-A mismatch. Also, the minor groove width flanking the G-C base pairs marked in 5’-TT**CC**-3’ is also wide, consistent with the G-A mismatch. Interestingly, a similarly wide minor groove width was observed at the same position in a related ETS protein, Elk1 (PDB ID: 1DUX), which also showed a similar SaMBA pattern of increased binding due to G-A (**Fig. 2a**). At the same time, MD simulations of the bound mismatched and Watson-Crick DNA for this G-A mismatch, for both Ets1 and Elk1, indicated that the formation of new protein-DNA contacts (at the positions of the mismatch and/or at neighboring positions) might also contribute to the enhanced binding affinity (**Supplementary Table 7**, **Extended Data Fig. 5c,h**). These results highlight the complexity of TF-mismatch recognition and the challenge of deconvolving contributions from base readout, shape readout, and conformational penalties.

For the G-G mismatch at position 7 in the Ets1 binding site, we found that the mismatch-induced expansion of the C1’-C1’ distance is similar to the expansion observed in Ets1-bound Watson-Crick sites. This expansion has not been reported in the literature to contribute to Ets1-DNA recognition. One possible mechanism is that the expanded DNA diameter could facilitate H-bonding with Arg394 at the guanine base opposite to the mismatched base at position 7 (**Extended Data Fig. 5d**). Alternatively, we also note that this G-G mismatch occurs in the highly distorted region of the Ets1 binding site, where the Ets1 recognition helix is inserted into the DNA major groove (**Extended Data Fig. 5d**). Thus, the increase in DNA diameter due to the G-G mismatch could also facilitate the distortions that follow the insertion of the α-helix. We note that indirect (shape) readout is broadly accepted to be an important, although poorly understood factor in ETS-DNA recognition[97,99].

For the G-T mismatch at position 6 and the C-T mismatch at position 8, which also increase Ets1 binding, it is less clear what mechanisms could lead to these increases. No structural mimicry was identified in these cases (**Supplementary Table 6**). However, for the G-T mismatch at position 6, MD simulations of protein-bound mismatched and Watson-Crick DNA revealed non-native protein contacts at the mismatched base (**Extended Data Fig. 5e, Supplementary Table 7**).

Mismatches in non-specific DNA sites can also enhance Ets1 binding, leading to affinities in the specific binding range (**Fig. 2c, Supplementary Table 2**). For such sites, two trends were observed. First, for 9 of the 12 mismatches validated to increase Ets1 binding from the non-specific to the specific range (**Supplementary Table 2c**), the mismatch created, albeit on a single strand, the GGAA core critical for Ets1-DNA recognition. For such cases, we can use the newly created GGAA core to align the mismatched site with specific Watson-Crick sites from available X-ray or NMR structures, and use these as starting structures for MD simulations. Our MD results indicated that native and non-native base contacts (**Extended Data Fig. 5g,i**), as well as additional hydrogen bonds with the DNA backbone (**Extended Data Fig. 5j**) could potentially contribute to the highly increased Ets1 binding. Importantly, for all the mismatches that created the GGAA core on one of the strands, the rescue Watson-Crick mutation was bound with lower affinity than the mismatch (**Fig. 2c**, **Supplementary Table 2c**).

For three of the 12 mismatches validated to increase Ets1 binding from the non-specific to the specific range (**Supplementary Table 2c**), neither the mismatched site nor the original Watson-Crick site contain the GGA(A/T) core. Two of three mismatches created a GGAG sequence on one strand, which could potentially still form the core major groove interactions in the GGA region. Interestingly, while interacting with the TF Pax5, Ets1 was previously shown to recognize the GGAG modified core sequence instead of the consensus GGA(A/T); however, ETS TFs on their own bind very poorly to the GGAG modified core, with an overall binding affinity that is 100-fold weaker compared to the consensus core[100]. Finally, in one of the 12 mismatches validated to increase Ets1 binding from the non-specific to the specific range (the 4th example in **Fig. 2c**), not even the 3-bp GGA invariant region of the core was formed, implying that the H-bond network must be considerably different from the previously characterized ETS binding sites. Structure determination of Ets1-mismatched DNA complexes is needed to elucidate the precise binding mechanisms for these mismatches.

Native conformations and interactions are preserved in structures of TBP bound to DNA with and without mismatches

DNA mismatches that increase the binding affinity of TBP could in principle act by leading to the formation of non-native interactions between the protein and DNA, such as H-bonds and stacking interactions. To investigate the mechanisms by which mismatches enhance TBP binding affinity, we determined high-resolution crystal structures of TBP-DNA complexes with A-C and C-C mismatches at two positions relative to two wild-type parent sequences with available crystal structures: 1) the Adenovirus major late promoter TBP binding site (5'-TATAAAAG-3', PDB ID: 1QNE), and 2) the TBP binding site 5'-TATAAACG-3' (PDB ID: 6NJQ).

We introduced a C-C mismatch at position 7 in the 5'-TATAAA**C**G-3' site (new structures TBP-CC(1a) and TBP-CC(1b), PDB IDs: 6UEP and 6UER), a C-C mismatch at position 8 in the 5'-TATAAAA**G**-3' site (new structure TBP-CC(2), PDB ID: 6UEQ2), and an A-C mismatch at position 7 in the 5'-TATAAA**A**G-3' site (new structure TBP-AC, PDB ID: 6UEO). The mismatches were selected based on SaMBA data showing enhanced TBP binding affinity relative to the parent Watson-Crick sequence due to single base substitutions.

We compared the structural properties of the bound mismatched sites against their wild-type counterparts. First, we compared the overall structures of TBP-DNA by computing the pairwise root-mean-square deviations (RMSD) on TBP and the entire complex. The RMSD between mismatches and their wild-type counterparts range from 0.26 to 0.47 Å for just TBP, and from 0.29 to 0.49 Å for the entire complex, which indicates that the structures are essentially the same (close to the overall errors of the coordinates for the resolutions).

Second, we compared the variations in global DNA shape. With the exception of the minor differences in major groove width at the 3'-side of TBP-CC(1a) and TBP-CC(2a), which is away from TBP-DNA binding core, the DNA major and minor groove widths are not affected by introducing of DNA mismatches.

Third, we compared the TBP-DNA interfaces. The buried surface area shows no significant change between TBP-mismatch structures (2056-2222 Å^{2}) and their wild-type counterparts (2128-2213 Å^{2}). The pairwise RMSD on the core amino acids and DNA between TBP-mismatch structures and their wild-type counterparts range from 0.24 to 0.42 Å, which is again comparable to the precision of the X-ray structures. Overlaying the structures reveals that most of the key amino acids also show little to no change, including the four phenylalanines (Phe-57, Phe-74, Phe-148 and Phe-165) responsible for kinking the DNA at both ends. Subtle differences can be observed for certain amino acid-nucleotide pairs: the Arg-56 is sliding to a different position in the TBP-mismatch structures, leading to loss of DNA phosphate contacts in TBP-CC(1a) and TBP-CC(1b); the Arg-63 and Arg-154 are forming different rotamers in TBP-mismatch structures, all of which are however making DNA phosphate contacts; the Leu-147 is also forming different rotamers which is likely due to an artifact of crystal packing. Additionally, due to the formation of the distorted A-C mismatch, the π-π stacking interaction between C and Phe-57 is less favorable in TBP-AC then its parent Watson-Crick structure (1QNE). Despite these subtle differences, the structures of the TBP-mismatch complexes mostly preserved the native conformation and interactions, with little to no change, and we did not observe any evidence of formation of favorable protein-DNA contacts in TBP-mismatch structures.

Potential roles for TF binding to mismatches in the cell

TF binding to DNA is a key determinant of gene regulation. Alterations in TF binding sites can affect cell functions and lead to human disease[101] . A well-characterized example is that of TFs from the ETS family, which selectively bind to the *TERT* promoter and activate this gene in cancer cells, due to single base-pair mutations that create high affinity ETS sites[102,103]. Remarkably, using SaMBA we observed several mismatches that also form high affinity ETS sites, p53, TBP, etc. Beyond the TFs tested in our study, human cells express hundreds of TF proteins, which collectively bind to hundreds of thousands of genomic sites[104]. Mismatches in these sites have the potential to alter TF binding and affect cellular activity.

The cellular effects of TF binding to mismatched DNA will depend on both the TF proteins (their abundance and affinities for mismatches), and the mismatches (their abundance and correction efficiency, which varies widely across the genome[17]). Several recent studies suggested that TFs could rapidly bind nascent DNA strands[10]; nevertheless, the fast repair of most replication errors significantly reduces the chance of TFs recognizing mismatches that arise due to errors during replication. On the other hand, the repair of mismatches that are formed during genetic recombination, or during gap-filling DNA synthesis, is slower and less efficient, with T-T and C-C mismatches repaired with very low efficiency[17,82]. Furthermore, spontaneous deamination is common and estimated to occur 100-500 times per cell per day in humans[83]. G-T mismatches that are generated by the deamination of 5-methyl Cytosine are not repaired by the mismatch repair pathway and have considerably lower repair efficiency[83]. In such cases, given that many TFs are bound all across the genome to regulate a variety of cellular processes, it is reasonable to expect that TFs would be functionally affected by the presence of mismatches. In tumor cells with defects in DNA mismatch repair, the effect of mismatches on TFs is expected to be even larger than in healthy cells.

TFs bound to DNA mismatches in the genome could also affect damage correction and lead to the formation of genetic mutations. Several recent studies hypothesized that TFs are likely playing a role in shaping the mutational landscape of eukaryotic genomes by interfering with DNA repair and replication[10,11,105]. In particular, mismatches bound with high affinity by TFs may be harder to recognize by repair enzymes[105] and harder to correct during strand replacement by Pol-δ[10] (**Extended Data Fig. 9**). Our study provides a platform to map which specific mismatches attract TFs and could become roadblocks for DNA repair and replication, thus contributing to mutagenesis and even becoming potential mutational "hotspots" in the genome.

On the other hand, TF binding to mismatched DNA could also play useful roles in damage recognition and the cellular response to damage. For example, p53 is known to sense and respond to DNA lesions by different mechanisms[106], and to regulate cellular processes such as cell cycle arrest, apoptosis and DNA repair. Since other TFs are also involved in these processes, understanding the interplay between TFs and mismatched DNA could provide insights into DNA damage response mechanisms in the cell.

Finally, oligonucleotides containing high-affinity binding sites for certain TFs known to involved in cancer are considered as potential binding decoy agents that act to lower the effective TF concentration in the cell, and thus down-regulate certain genes[107-110]. Using DNA stem-loops containing one or several mismatches with high binding affinity, as identified in our study, could form the basis for constructing super high-affinity sites, higher than any known Watson-Crick binding sites, which then could serve as strong inhibitors for TFs of interest.

Titles and Legends for Supplementary Tables 1-11

Supplementary Table 1. SaMBA data.
Table contains the raw and processed SaMBA data for the 22 TFs.

Supplementary Table 2. Validation of the effect of mismatches in non-specific DNA.
Table contains the raw and processed SaMBA data used to compare the Ets1 binding level at mismatched non-specific sites versus random DNA sites and specific sites from NMR and X-ray crystal structures of Ets1-DNA complexes.

Supplementary Table 3. Calibration of SaMBA data.
Table contains Kd data from EMSA, FA, MITOMI and SPR experiments, used to calibrate our high-throughput SaMBA data.

Supplementary Table 4. Calibrated SaMBA data used to compare the effects of mismatches versus mutations on TF binding.
Table contains the raw and processed binding data for all mismatches and mutations in 12 TF binding sites for the 7 TFs with calibration data in our study, as well as statistics of the comparisons between the effects of mismatches versus mutations.

Supplementary Table 5. Structural distortions in TF-bound DNA.
Table shows the deviations from the B-DNA envelope for DNA structural parameters at each base pair position in 12 TF-DNA complexes.

Supplementary Table 6. Results of structural mimicry analysis.
(a) X-ray structures of protein-DNA complexes selected for structural analyses of mismatches that increase TF binding. (b) Distortions at DNA positions where mismatches increase TF binding affinity. (c) Distortions of DNA structural parameters of mismatches relative to Watson-Crick base pairs. (d) Mismatches that increase TF binding affinity and exhibit geometries similar to distorted base pairs in TF-bound DNA.

Supplementary Table 7. Analysis of TF-DNA hydrogen bonding and buried surface area in MD simulations of TF-DNA complexes with and without mismatches.
Results shown are from MD simulations. DNA sequences were derived from the sequences used in SaMBA.

Supplementary Table 8. Defining the B-DNA envelope.
Table contains summary statistics of base pair parameters (mean, maximum value, minimum value, and standard deviation) for base pairs in B-DNA (as well as TF-bound DNA), obtained from a comprehensive survey of structures deposited in PDB.

Supplementary Table 9. Structural survey of DNA mismatches.
Table contains all possible single mismatches (excluding modified bases) surrounded by at least two canonical Watson-Crick bps on both sides, from PDB structures. The data was used to survey the DNA mismatch structure and geometry.

Supplementary Table 10. SaMBA hybridization signal.
Fluorescent signal for DNA duplexes expected to contain labeled and unlabeled probes, from the hybridization of 12 sequences on a DNA chip (see also Figure S3). For the sequences with an unlabeled complementary strand (sequences 2, 4, 6, 8, 10, 12), the signal is several orders of magnitude lower than for the sequences with a labeled complementary strand (sequences 1, 3, 5, 7, 9, 11).

Supplementary Table 11. NMR results confirming that T-T and C-T mismatches mimic Hoogsteen A-T geometry.

Table includes the chemical shift differences in the sugar C1'/C3'/C4' carbons for T-T and C-T mismatches versus a locked Hoogsteen conformation (using N1-methyladenosine, or m1A), relative to the Watson-Crick base-paired duplex.

Title and Legend for Supplementary Figure 1

Supplementary Figure 1. Original source images for all EMSA data reported in this study.

Supplementary References

85 Berendsen, H., Grigera, J. & Straatsma, T. The missing term in effective pair potentials. *J. Phys. Chem.* **91**, 6269-6271, (1987).

86 Joung, I. S. & Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **112**, 9020-9041, (2008).

87 Fiser, A. & Sali, A. ModLoop: automated modeling of loops in protein structures. *Bioinformatics* **19**, 2500-2501, (2003).

88 Li, P., Roberts, B. P., Chakravorty, D. K. & Merz, K. M., Jr. Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent. *J. Chem. Theory Comput.* **9**, 2733-2748, (2013).

89 Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **9**, 3084-3095, (2013).

90 Lin, C. Y. *et al.* Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56-67, (2012).

91 Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74, (2011).

92 Wells, J. W. in *Receptor-Ligand Interactions: a Practical Approach* (ed E.C. Hulme) 289-395 (IRL Press at Oxford University Press, 1992).

93 Horovitz, A. Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Fold. Des.* **1**, R121-126, (1996).

94 Horovitz, A., Fleisher, R. C. & Mondal, T. Double-mutant cycles: new directions and applications. *Curr. Opin. Struct. Biol.* **58**, 10-17, (2019).

95 Schreiber, G. & Fersht, A. R. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.* **248**, 478-486, (1995).

96 Hollenhorst, P. C., Shah, A. A., Hopkins, C. & Graves, B. J. Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ETS gene family. *Genes. Dev.* **21**, 1882-1894, (2007).

97 Hollenhorst, P. C., McIntosh, L. P. & Graves, B. J. Genomic and biochemical insights into the specificity of ETS transcription factors. *Annu. Rev. Biochem.* **80**, 437-471, (2011).

98 Wei, G. H. *et al.* Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J.* **29**, 2147-2160, (2010).

99 Szymczyna, B. R. & Arrowsmith, C. H. DNA binding specificity studies of four ETS proteins support an indirect read-out mechanism of protein-DNA recognition. *J. Biol. Chem.* **275**, 28363-28370, (2000).

100 Garvie, C. W., Hagman, J. & Wolberger, C. Structural Studies of Ets-1/Pax5 Complex Formation on DNA. *Mol. Cell* **8**, 1267-1276, (2001).

101 Chatterjee, S. & Ahituv, N. Gene Regulatory Elements, Major Drivers of Human Disease. *Annu. Rev. Genom. Hum. Genet.*, (2017).

102 Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-959, (2013).

103 Bell, R. J. *et al.* The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. *Science* **348**, 1036-1039, (2015).

104 Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252-263, (2009).

105 Khurana, E. Cancer genomics: Hard-to-reach repairs. *Nature* **532**, 181, (2016).

106 Degtyareva, N., Subramanian, D. & Griffith, J. D. Analysis of the binding of p53 to DNAs containing mismatched and bulged bases. *J. Biol. Chem.* **276**, 8778-8784, (2001).

107 Gniazdowski, M., Denny, W. A., Nelson, S. M. & Czyz, M. Transcription factors as targets for DNA-interacting drugs. *Curr. Med. Chem.* **10**, 909-924, (2003).

108 Crinelli, R. *et al.* Transcription factor decoy oligonucleotides modified with locked nucleic acids: an in vitro study to reconcile biostability with binding affinity. *Nucleic Acids Res.* **32**, 1874-1885, (2004).

109 Tomita, N., Morishita, R., Tomita, T. & Ogihara, T. Potential therapeutic applications of decoy oligonucleotides. *Curr. Opin. Mol. Ther.* **4**, 166-170, (2002).

110 Rad, S. M. *et al.* Transcription factor decoy: a pre-transcriptional approach for gene downregulation purpose in cancer. *Tumour Biol.* **36**, 4871-4881, (2015).