Supplementary Information for

DNA Mechanical Strain Steers Transcription Factor Recognition

Yumi Minyi Yao[1], Michael P. O'Hagan[1], Karn Onoon[2], Lihee Givon[1], Shelly Hamer-Rogotner[3], Raul Salinas[4], Raj V. Nithun[5], Naama Kessler[1], Muhammad Jbara[5], Orly Dym[3], Tanadet Pipatpolkai[6], Maria A. Schumacher[5], Ariel Afek[1*]

[1]Department of Chemical and Structural Biology, Weizmann Institute of Science, Rehovot 7610001, Israel;

[2]Department of Biology, Faculty of Science, Mahidol University, Bangkok 10400, Thailand;

[3]Department of Life Sciences Core Facilities, Weizmann Institute of Science, Rehovot 7610001, Israel;

[4]School of Physical and Mathematical Sciences, Nanyang Technological University, 637371

Singapore;

[5]Department of Biochemistry, Duke University School of Medicine, Durham, NC 27710, USA

[6]School of Physical and Mathematical Sciences, Nanyang Technological University; 637371, Singapore

*Corresponding Author: ariel.afek@weizmann.ac.il

This PDF file includes:

Supplementary Discussion

Within-family conservation of PIC-NIC sensitivity profiles

To assess whether PIC-NIC sensitivity patterns reflect class-level principles or are idiosyncratic to individual TFs, we compiled a within-family comparison of PIC-NIC profiles across TFs sharing the same structural family in our panel (Extended Data Fig. 4). For the bHLH family (MITF, MAX, MYC, MAD, MNT) and the bZIP family (CREB1, ATF1), profiles within each family are strikingly similar, with consistent positional sensitivity patterns observed across all members, suggesting that the structural logic of backbone-mediated recognition is largely conserved within these classes. In contrast, the three zinc finger proteins in our panel—EGR1 (three zinc fingers), SP1 (three zinc fingers), and CTCF (eleven zinc fingers)—display notably divergent profiles, consistent with the exceptional structural diversity of this family in terms of finger number, binding site length, and protein–DNA contact geometry. Within this family, backbone sensitivity appears to reflect the specific architecture of each individual TF–DNA complex rather than a shared class-level principle—itself an informative observation that highlights the resolution PIC-NIC can provide across structurally heterogeneous families.

DNA bendability analysis

We additionally assessed intrinsic DNA bendability at each position using two independent approaches: the trinucleotide DNase I bendability model of Brukner et al.[1] and the BendNet deep learning server[2,3]. Neither approach revealed a clear correlation with PIC-NIC nick sensitivity across our TF panel (Supplementary Table 4). This is perhaps unsurprising: intrinsic bendability reflects the mechanical properties of free, unbound DNA, whereas PIC-NIC sensitivity reflects the specific structural demands imposed by each TF upon binding—two related but distinct properties for which a simple correspondence would not necessarily be expected. The position-specific correlations we observe with DNA shape parameters such as roll, tilt, and shift suggest that different structural descriptors capture different aspects of binding-relevant mechanics, and that identifying the appropriate parameter for each TF–DNA system may require a more systematic and integrative approach. We consider this an interesting direction for future work.

Phosphate buried surface area (BSA)

For ETS1, a strong negative correlation was observed between phosphate burial and PIC-NIC sensitivity across both strands ($R^2 = 0.712–0.955$; Extended Data Fig. 7a; Supplementary Table 7): the most extensively buried phosphates—those flanking the GGA core at the groove borders—are precisely those whose disruption most strongly impairs binding. A similar trend was observed for FOXA2, with correlations strengthening substantially when analysis was restricted to positions with measurable protein contact (BSA>0). Together, these results support the interpretation that phosphate burial provides a meaningful structural basis for nick sensitivity at positions of genuine protein–phosphate contact.

For EGR1 and several other proteins, however, no clear global correlation between burial depth and nick sensitivity was observed. This likely reflects the additional structural complexity of zinc finger proteins, where burial depth alone does not fully capture the energetic contribution. These findings suggest that while phosphate burial is a useful complementary metric, PIC-NIC sensitivity reflects a richer set of structural determinants that cannot be reduced to contact area alone.

Potential dual role of histidine residues in EGR1 phosphate anchoring

Among the EGR1 positions examined, the two most disruptive phosphate contacts—at positions 3 and 6—involve histidine residues that simultaneously coordinate zinc within the C2H2 motif (PDB ID: 1AAY; Extended Data Fig. 5a). This dual role likely amplifies the functional importance of these interactions, contributing both to local DNA binding and to the structural integrity of the zinc finger domains. In contrast, the phosphate contacts at position 7, which lie between adjacent zinc fingers, appear more dispensable for binding stability—consistent with the minimal PIC-NIC effect observed at this position.

To assess whether this dual anchoring role is a general feature of C2H2 zinc finger TFs, we examined CTCF, the other zinc finger TF in our panel for which a DNA-bound crystal structure is available (no structure exists for SP1 in complex with DNA). In the CTCF–DNA structure, the zinc-coordinating residues do not simultaneously contact the DNA backbone phosphates in the manner observed in EGR1, suggesting that this arrangement is not a universal feature of C2H2

zinc finger proteins but rather a context-specific configuration reflecting the particular geometry of the EGR1 zinc finger–DNA interface. While a single comparison is insufficient to draw broad conclusions, these observations highlight how the functional consequences of phosphate contacts can vary substantially even within the same structural family, and underscore the value of PIC-NIC in resolving these distinctions at nucleotide resolution.

Structural analysis of EGR1 bound to DNA nicked at position 5 on the bottom strand

To complement the position 7 analysis described in the main text, we also solved a high-resolution crystal structure of EGR1 bound to DNA nicked at position 5 on the bottom strand with the 5′ phosphate retained (PDB ID: 9RI6)—a condition that produces only minor disruption in PIC-NIC (Extended Data Fig. 3, 5). Consistent with this, the structure is nearly identical to the intact complex (RMSD within the range observed for position 7 nicks). At the nick site, the backbone phosphate stretches away from its original position, altering local backbone geometry. Interestingly, in the intact complex the distances between nearby Arg and Lys residues and this phosphate are too long to contribute stabilizing interactions. The nick-induced conformational change shortens these distances, bringing the phosphate into closer proximity and enabling new favorable contacts to emerge. This structural observation suggests that nick-induced backbone rearrangements can generate compensatory interactions that offset the perturbation, providing a structural rationale for the retained binding observed at this position.

ETS1 phosphate-contacting residue mutants

To more comprehensively assess the role of phosphate contacts in shaping ETS1 sequence specificity, we generated and profiled four ETS1 point mutants—K379A, Y397A, K404A, and Y410A (Extended Data Fig. 8). All four mutants retained sequence-specific binding, as evidenced by coherent PWM logos with clear information content at core positions (Extended Data Fig. 6a–b). However, each mutant displayed a distinct pattern of altered base preferences relative to wild-type ETS1, reflecting the specific contribution of each phosphate contact to the overall recognition geometry.

K379A, which removes the contact at the right-flank phosphate discussed in the main text, showed the most pronounced reshaping of specificity, including reduced selectivity within the GGA core and shifted preferences at distal positions—consistent with this contact contributing to long-range architectural coherence across the interface. Y397A produced a distinct pattern of altered preferences, with differences concentrated at positions flanking the core on the opposite side, suggesting that the two flanking phosphate contacts make asymmetric contributions to base readout. K404A and Y410A showed more modest changes in base preferences, with shifts primarily at peripheral positions, consistent with these contacts playing a less central role in organizing core recognition.

Across all four mutants, the correlation with wild-type binding scores revealed a population of sequences preferentially bound by the mutant but not wild-type (blue box, Extended Data Fig. 8b), sequences preferentially bound by wild-type (red box), and sequences bound comparably by both (yellow box). The identity of sequences in each category reflects how each phosphate contact contributes to the selectivity landscape, reinforcing the view that phosphate interactions are not passive but actively participate in shaping the specificity of TF–DNA recognition.

Integration of PIC-NIC with sequence-encoded DNA shape predictions

To further examine whether structural dependencies are shared across a structural class and to compare PIC-NIC results with sequence-encoded DNA shape predictions, we performed deepDNAshape analysis across all TFs in our panel, examining whether correlations between predicted DNA shape features and binding preferences converge on the same positions within structural families (Extended Data Fig. 4; Supplementary Table 3; detailed methods in Supplementary Methods).

In total, 4 examples passed the Bonferroni test, all of which correspond to nick-sensitive positions:

protein	oriented_seq	feature	feature	nn_pos	pred_pos	correlation
ETS1	CCGGAAGT	Tilt	base_step	6	5	-0.89
SP1	GGGGCGGG	Shear	base_pair	6	6	-0.93
SOX2	TATTGTT	Shift	base_step	5	4	-0.89
EGR1	GCGTGGGC	Shear	base_pair	6	6	-0.98

More examples passed the FDR test. Except for the examples above, and those whose Bonferroni p = 1, the rest include:

protein	oriented_seq	feature	feature	nn_pos	pred_pos	correlation
ATF1	GTGACGT	Slide	base_step	0	1	0.85
ATF1	GTGACGT	Roll	base_step	0	1	0.84
ATF1	GTGACGT	Roll	base_step	5	4	0.81
EGR1	GCGTGGGC	HelT	base_step	6	5	-0.87
MAD	CCACGTGG	Roll	base_step	3	3	0.83
MAX	ACCACGTG	Roll	base_step	4	4	0.82
MITF	CACGTGA	Rise	base_step	2	2	0.84
MITF	CACGTGA	Roll	base_step	1	2	0.82
MYC	CCACGTGG	Roll	base_step	3	3	0.83
SOX2	TATTGTT	Shift	base_step	2	3	0.86
SP1	GGGGCGGG	HelT	base_step	6	5	-0.86
TBP	TATATAT	Shift	base_step	0	1	0.86
TBP	TATATAT	Shift	base_step	4	3	-0.82

The bHLH family provides a particularly compelling example of the structural importance of recognition. Across four bHLH proteins—MITF, MAX, MYC, and MAD—deepDNAshape analysis independently identifies the same central CG base step as the position where predicted roll or rise correlates most strongly with binding score (r = 0.82–0.84, Bonferroni corrected; Supplementary Table 3, Supplementary Fig. 4b-c). Strikingly, this is precisely the position where PIC-NIC shows the strongest disruption of binding upon nicking across all four proteins. Furthermore, when flanking positions rather than the central CG are used as the variable dinucleotide, the correlation still preserves but weakens, suggesting that the structural dependency is specifically localized to the palindrome center rather than reflecting a general sequence effect. These converging results from two independent approaches suggest that backbone-mediated recognition at the central palindrome step is a conserved principle of bHLH–DNA recognition rather than an idiosyncrasy of any individual family member.

Beyond the bHLH family, deepDNAshape analysis identified an additional clear example of position-specific correlation for ETS1, where predicted tilt at the peripheral anchor position correlates strongly with binding (r = −0.89, Bonferroni corrected; Extended Data Fig. 8f). The analysis also revealed further correlations across other TFs in our panel that, while intriguing,

require further experimental validation to interpret mechanistically. We view the integration of PIC-NIC with computational shape prediction as a promising direction for future work.

In vivo enrichment of SSBs at TF binding sites and implications for DNA repair

To begin bridging our *in vitro* findings to a physiological context, we integrated four complementary genomic datasets generated in induced neurons (iNeurons): genome-wide SSB maps (S1-END-seq), markers of active SSB detection and repair (PAR ChIP-seq and XRCC1 ChIP-seq), and a direct readout of repair synthesis activity (SAR-seq). We focused on EGR1 and CTCF—two TFs with detailed PIC-NIC profiles that play critical roles in neuronal function—analyzing signal enrichment both at ChIP-seq peaks broadly and at peaks containing a *bona fide* TF binding motif (Extended Data Fig. 13; detailed methods in Supplementary Methods).

S1-END-seq revealed significant enrichment of SSBs at both TF binding sites compared to randomly shuffled genomic regions. For CTCF, SSBs were enriched 1.68-fold at ChIP-seq peaks ($Z = 103.33$, $p < 0.0001$) and 1.96-fold at motif-centered regions ($Z = 97.19$, $p < 0.0001$). For EGR1, enrichment was 1.80-fold at ChIP-seq peaks ($Z = 65.92$, $p < 0.0001$) and 1.98-fold at motif-centered regions ($Z = 44.27$, $p < 0.0001$). Critically, enrichment increased upon restriction to motif-containing peaks for both TFs, confirming that the association with DNA breaks reflects direct, sequence-specific TF–DNA interaction rather than indirect chromatin accessibility effects.

Consistent with active damage recognition and processing, both PAR and XRCC1 were significantly enriched at TF binding sites. For CTCF, XRCC1 enrichment reached 4.36-fold at ChIP-seq peaks and 6.08-fold at motif-centered regions; comparable enrichments were observed for EGR1 (2.56-fold and 4.17-fold respectively). The increase in XRCC1 enrichment upon motif-centering—observed consistently for both TFs—indicates that repair factor recruitment is most concentrated precisely where TFs engage DNA sequence-specifically, mirroring the SSB enrichment pattern itself. Notably, PAR signal showed a local depletion directly at the motif center for both EGR1 and CTCF despite being broadly enriched at TF binding sites, suggesting that stably bound TFs could restrict PARP1 access to the immediate site of DNA damage through physical occlusion of the protein–DNA interface, a potential mechanism described in the past literature[4,5].

Strikingly, SAR-seq signal displayed the same characteristic depletion pattern at the motif center, flanked by elevated repair synthesis signal on both sides, for both EGR1 and CTCF. This "volcano-shaped" pattern is consistent with prior observations that stably bound TFs physically occlude repair machinery at occupied sites, reducing repair synthesis at the exact binding position while leaving flanking accessible DNA available for repair. To test whether the central dip is statistically significant rather than a chance feature of the average profile, we randomly partitioned the motif occurrences into 50 disjoint subgroups and computed the mean SAR-seq signal at the dip center and at the symmetric flanking borders within each subgroup (Extended Data Fig. 13d, h; Supplementary methods). The center signal was significantly lower than the flanking borders for both EGR1 and CTCF (one-sided paired t-tests across 50 subgroups; $p < 0.05$ for both TFs; Supplementary Methods), confirming that the depletion at the motif is a robust property of the binding population rather than being driven by a small subset of sites or by features of the average curve.

Together, these analyses suggest that SSBs are enriched at TF binding sites *in vivo*, that this damage engages the PARP1-dependent repair pathway, and that stably bound TFs may impede both PARP1 access and repair synthesis at their occupied positions—suggesting that TF occupancy could be an important factor that shapes the local repair landscape in a manner directly connected to the backbone sensitivity we characterize with PIC-NIC.

Position-specific competition between ETS1 and PARP1 at nicked DNA sites

To directly test whether TF binding at nick-sensitive positions can occlude repair factors, we performed *in vitro* competition experiments on PIC-NIC chips, examining whether ETS1 and PARP1 compete for occupancy at nicked DNA sites (Extended Data Fig. 7c-d).

PARP1 binding signal was measured across 14 nicked ETS1 binding site probes under two conditions: PARP1 alone, and PARP1 in the presence of ETS1. Addition of ETS1 produced a significant and consistent reduction in PARP1 binding across three probes we tested ($p = 0.0114$, paired t-test), demonstrating that ETS1 occupancy at nicked sites directly competes with and displaces PARP1.

To assess whether this competition reflects the nick-sensitivity hierarchy identified by PIC-NIC, we selected the three probes with comparable PARP1 binding in the absence of ETS1 but strikingly different PIC-NIC profiles: n2 (the largest reduction in ETS1 binding upon nicking), n5 (intermediate), and reverse strand n5 (ren5, minimal reduction). We reasoned that positions where ETS1 binding is most disrupted by nicking should remain more accessible to PARP1 in the presence of competitor ETS1, while positions where ETS1 binding is preserved should show greater PARP1 displacement. Consistent with this, in the absence of ETS1, all three probes showed highly similar PARP1 binding. Upon addition of 500 nM ETS1, ren5 exhibited significantly lower PARP1 signal than n2 ($p = 0.0112$). The distinction between n2 and n5 did not reach significance, suggesting that additional factors beyond nick-sensitivity, such as local sequence context or probe geometry, may also modulate competition at these positions.

These results provide proof-of-concept that the nick-sensitivity hierarchy revealed by PIC-NIC has direct functional consequences for repair factor accessibility. This position-specific competitive logic, together with the *in vivo* genomic analyses described above, connects the backbone sensitivity landscape of PIC-NIC to a physiologically relevant outcome and lays the groundwork for future studies examining how nick sensitivity shapes repair efficiency and mutational risk *in vivo*.

Extending PIC-NIC to chromatinized and genome-integrated contexts

The present study establishes PIC-NIC as a systematic *in vitro* platform for dissecting the role of backbone mechanics in TF–DNA recognition. An important and natural extension will be to move this framework into chromatinized and genome-integrated contexts, where DNA is packaged into nucleosomes and subject to additional topological constraints that are absent in free duplex DNA.

Several converging lines of evidence suggest that backbone mechanics may play an especially important role in TF recognition within chromatin. First, nucleosome wrapping imposes rotational and translational constraints on DNA that generate local torsional strain and modulate groove geometry[6-8]. Studies of pioneer TF binding to nucleosomes have shown that SOX2 affinity for nucleosomal sites depends strongly on local DNA flexibility and bendability, with increased nucleosomal DNA flexibility enhancing SOX2 association[9]. This is consistent with our finding

that a nick at the SOX2 minor groove kink position—which locally increases backbone flexibility—enhances binding *in vitro*, and raises the possibility that nucleosome-imposed deformations at specific rotational positions could similarly prime or impede TF engagement. Second, transcription-induced supercoiling and torsional stress are well established to propagate heterogeneously along the DNA helix[10-12], with local twist deviations concentrated at specific base-pair steps—precisely the class of deformations that PIC-NIC identifies as mechanically sensitive.

From an experimental standpoint, several approaches could test whether PIC-NIC-derived nick sensitivities predict TF behavior in chromatin. *In vitro*, reconstituted nucleosomes or nucleosome arrays containing site-specific nicks—most cleanly generated by ligation of pre-nicked DNA substrates—could be used to measure TF binding affinities and kinetics on chromatinized templates. Such experiments would need to carefully control for motif rotational setting, translational position, and baseline nucleosome accessibility, since nucleosome positioning itself may dominate TF binding and mask or amplify the effects of backbone discontinuity. For nick positions predicted by PIC-NIC to strongly disrupt binding, reduced TF occupancy or altered kinetics would be expected if the nick remains accessible and the TF–DNA geometry is preserved within the chromatin context. Conversely, at positions where PIC-NIC suggests strain relief enhances recognition, nicks could facilitate the DNA deformation required for binding — though this would need to be carefully distinguished from indirect effects on nucleosome stability or DNA unwrapping.

In vivo, targeted introduction of SSBs at accessible regulatory regions using site-specific nickases could directly test whether PIC-NIC-derived sensitivity predicts TF retention or displacement at sites of DNA damage. Current nickase platforms—including CRISPR-Cas9 nickase variants and base-editor-derived nicking systems[13-15]—can introduce SSBs within a window of a few nucleotides from a defined target site, providing positional resolution over roughly half the positions within a typical TF binding motif. This would not replicate the complete position-by-position coverage of PIC-NIC, but would allow direct *in vivo* testing of the most mechanically critical positions identified *in vitro*. Subsequent time-resolved mapping of TF occupancy and repair-factor recruitment by ChIP-seq could then test whether positions predicted by PIC-NIC to be highly disruptive indeed lead to TF displacement and increased repair accessibility in cells.

These experiments would complement the genomic analyses presented here, in which endogenous SSB-associated signals are enriched near TF binding sites in neurons and TF occupancy correlates with reduced local repair synthesis. Together, such extensions would position PIC-NIC as a bridge between *in vitro* mechanistic dissection and *in vivo* regulatory biology — connecting the backbone-sensitivity landscape of individual TF–DNA complexes to the broader question of how DNA damage, chromatin architecture, and transcriptional regulation are coupled at the genome scale.

Materials and Methods

Protein expression and purification

For PIC-NIC experiments, full-length human ETS1, human EGR1 DNA-binding domain (residues 335-423), full-length Arabidopsis thaliana TBP, full-length human RUNX1 and human SOX2 DNA-binding domain (residues 19-118) were expressed and purified as described below. Full-length human Max and DNA-binding domain of Myc, Mad and Mnt were chemically synthesized as described previously[16,17]. Human FOXA2 DNA-binding domain (residues 142-269) was expressed by *in vitro* transcription/translation system with PURExpress® *In Vitro* Protein Synthesis Kit. Full-length human MITF, CREB1, ATF1, SP1 and CTCF were obtained commercially. For ETS1 BLI experiments, murine ETS1 (residues 331-440) was produced. For SOX2 BLI experiments, the same proteins as in PIC-NIC experiments were used. For EGR1 and TBP X-ray crystallization, the same proteins were used as described above, but with the tag cleaved.

Protein expression and purification for PIC-NIC assays:

Human ETS1 full-length protein was expressed as an N-terminal GST fusion and purified in *E. coli*[18]. Bacterial cultures were grown at 37 °C in LB media, induced overnight at 15 °C with 0.5 mM IPTG, harvested, and lysed by sonication in PBS. The clarified lysate was applied to a 5 mL GSTrap FF column (Cytiva) equilibrated in 25 mM Tris-HCl pH 8, 50 mM KCl, 10% (v/v) glycerol, 0.1 mM EDTA, and 1 mM DTT. The protein was eluted in 50 mM Tris-HCl (pH 8.0), 20 mM KCl, 20 mM reduced glutathione. Further purification was achieved by heparin chromatography (HiTrap Heparin HP, Cytiva), eluting over a linear gradient from 50 mM Tris-HCl (pH 8.0) to 50 mM Tris-HCl (pH 8.0), 1 M KCl, with GST-ETS1 eluting at ~300 mM KCl.

The protein was then subjected to size-exclusion chromatography (Superdex 200 10/300 GL, Cytiva) equilibrated in PBS. Protein-containing fractions were pooled, and the pure proteins were concentrated to 0.0051 mM (0.399 mg/mL). The concentration was determined by UV absorption at 280 nm using an extinction coefficient of 124525 M^{-1} cm^{-1}. The protein was snap frozen in liquid N_2 and stored at -80 °C.

Human EGR1, residues 335-423, with N-terminus GST tag, were expressed and purified as described below in EGR1 expression and purification for crystallography, without thrombin cleavage.

Arabidopsis thaliana TBP, with N-terminal HIS tags, were expressed and purified in *E. coli* C41 (DE3) cells as described previously[19]. In brief, an overnight culture was inoculated into LB media and grown at 37 °C. Expression was induced by 0.5 mM isopropyl β-D-1-thiogalactopyranoside overnight at 15 °C. Cells were harvested and resuspended in a buffer containing 5% (v/v) glycerol, 4 mM $MgCl_2$, 600 mM NaCl and 40 mM MES (pH 7.2). Cells were lysed by French press and sonication followed by centrifugation. The lysate was purified on a Ni-NTA column via the N-terminal 6x HIS tag, washed with the lysis buffer with increasing concentration of imidazole. TBP was then eluted in the lysis buffer with 100-1000 mM imidazole, and 5 mM BME was added to the fractions. The eluent was dialysed against a buffer containing 20 mM HEPES-KOH (pH 8), 100 mM KCl, 20% (v/v) glycerol, 1 mM $MgCl_2$ and 1 mM $CaCl_2$. The protein was concentrated with a 10kDa MWCO Amicon (Millipore) and the concentration was determined by UV absorption at 280 nm using an extinction coefficient of 10.5 mM^{-1} cm^{-1}. The protein was snap frozen in liquid N_2 and stored at -80 °C.

Full length human Max, DNA-binding doman of Myc, Mad, and Mnt were chemically synthesised and labelled with TAMRA fluorophore, as described previously[16,17]. All experiments of Myc, Mad, and Mnt heterodimers were performed by mixing Max with its desired binding partner, at a 1:5 ratio, to ensure that mostly heterodimers were formed instead of Max:Max homodimers.

Recombinant human full-length Runx1 protein was expressed and purified from BL21 (DE3) *E. coli* harboring a clone in pReceiver-B03 Runx1. In brief, an overnight culture was inoculated into LB media and grown at 33 °C until OD reached 0.6-0.8. Expression was induced by 0.5 mM

isopropyl β-D-1-thiogalactopyranoside overnight at 15 °C. Cells were harvested and resuspended in a buffer containing 500 mM Tris (pH 7.5), 300 mM NaCl, 1 mM DTT and 5 mM imidazole. Cells were lysed by French press and sonication followed by centrifugation. The lysate was purified on Ni-NTA column via the N-terminal 6x HIS tag, washed with a buffer containing 50 mM Tris (pH 7.5), 300 mM NaCl, 5 mM 2-Mercaptoethanol (BME), 5 mM imidazole and 10% (v/v) glycerol. The protein was eluted with an elution buffer consisting of 50 mM Tris (pH 7.5), 300 mM NaCl, 5 mM 2-Mercaptoethanol (BME), 200 mM imidazole and 10% (v/v) glycerol. The eluate was dialysed against the elution buffer without imidazole overnight, followed by a heparin Sepharose SP (GE) column. The protein was then diluted with a buffer consisting of 50 mM Tris (pH 7.5), 2-Mercaptoethanol (BME) and 10% (v/v) glycerol to 100 mM salt concentration, loaded onto a heparin Sepharose SP (GE) column and washed with the same buffer. The protein was then eluted with a salt gradient to a buffer consisting of 1 M NaCl, 50 mM Tris (pH 7.5), 2-Mercaptoethanol (BME) and 10% (v/v) glycerol. The eluate was concentrated and the concentration was determined by UV absorption at 280 nm using an extinction coefficient of 40340 M^{-1} cm^{-1}. The protein was snap frozen in liquid N_2 and stored at -80 °C.

Recombinant human Sox2 DNA binding domain, residues 19-118, with N-terminal HIS tags, was expressed and purified as described below in Sox2 expression and purification for BLI.

Human FoxA2 DNA-binding domain, residues 142-269, with N-terminal GST tags and C-terminal HIS tags, was expressed by *in vitro* transcription/translation system with PURExpress® *In Vitro* Protein Synthesis Kit (New England Biolabs, catalog #: E6800). A standard protocol was followed, with the addition of 10 µl of Solution A, 7.5 µl of Solution B, 1 µl of RNase Inhibitor and 250 ng template plasmid (per 25 µl reaction). The resulting reaction mixture was incubated at 37°C for 2 hours and stored at -20 °C.

Full-length human MITF, with N-terminal GST tag and C-terminal HIS tag, was purchased from Origene (catalog #: TP761396). Full-length human CREB1, with N-terminal HIS tag, was purchased from Origene (catalog #: TP760318). Human ATF1, residues 1-271, with N-terminal HIS tag, was purchased from Prospec (catalog #: PKA-019). Full-length human SP1 with N-terminal HIS tag was purchased from Origene (catalog #: TP760592). Full-length human CTCF with N-terminal GST tag was purchased from Abnova (catalog #: H00010664-P01).

ETS1 expression and purification for BLI: Recombinant ETS1 DNA-binding domain (murine residues 331 to 440) was expressed and purified from BL21*(DE3) *E. coli* harboring a clone in pET28b-Ets1-ETS (Addgene #85735) as previously reported[20]. In brief, an overnight culture was inoculated into liter-scale LB media and grown at 37 °C. Expression was induced by 1 mM isopropyl β-D-1-thiogalactopyranoside overnight at 25 °C. Cells were harvested and resuspended in a buffer containing 100 mM Tris-HCl (pH 8.0), 500 mM NaCl and 5 mM DTT. All downstream buffers contained 0.5 mM Tris(2- carboxyethyl)phosphine (TCEP) hydrochloride to reduce the cystein disulfide bridges. Cells were lysed by French press and sonication followed by centrifugation. 40 µL of IGEPAL CA-630 (Sigma) and DNA inhibitor were added to 40 mL of the lysate. The lysate was purified on Ni-NTA column via the C terminal 6xHis tag on the protein, followed by a heparin Sepharose SP (GE) column. The eluate was dialysed overnight against a buffer containing PBS, 0.15 M NaCl and bovine thrombin (Prospec) to remove the C-terminal 6×His tag. Purified Ets-1 was dialysed 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.

Sox2 expression and purification for BLI: Recombinant human Sox2 DNA binding domain, residues 19-118, with N-terminal HIS tags, was expressed and purified from BL21 (DE3) *E. coli* as previously described[9]. In brief, an overnight culture was inoculated into LB media and grown at 37 °C. Expression was induced by 0.5 mM isopropyl β-D-1-thiogalactopyranoside for 3-4 hours at 37 °C. Cells were harvested and resuspended in a buffer containing 50 mM Tris-HCl (pH 7.5), 500 mM NaCl, 10 mM imidazole, 10% (v/v) glycerol and 5 mM 2-Mercaptoethanol (BME). Half a pill of cOmplete Protease Inhibitor (Sigma) and 1 mM DTT was added to the sample, before the cells were lysed by French press and sonication, followed by centrifugation at 14000 rpm for 30 min at 4 °C. A HisTrap™ High Performance (Cytiva) column was equilibriated with a buffer containing 50 mM Tris-HCl (pH 7.5), 500 mM NaCl, 10 mM imidazole, 10% (v/v) glycerol and 5 mM 2-Mercaptoethanol (BME). The supernatant was syringe filtered through a 0.45 µm filter, before the lysate was loaded onto the column at 1-1.5 ml/min. The column was wash with the same equilibriation buffer at 2-3 ml/min with several column volumes until the absorbance reading was close to the baseline. The protein was then eluted with a salt gradient to a buffer consisting of 50 mM Tris-HCl (pH 7.5), 500 mM NaCl, 500 mM imidazole, 10% (v/v) glycerol and 5 mM 2-Mercaptoethanol (BME), collected in fractions. Protein-containing fractions were pooled and the pure proteins were dialysed against a buffer consisting of 50 mM Tris-HCl (pH 7.5), 500 mM

NaCl, 10% glycerol and 5mM 2-Mercaptoethanol (BME) overnight. The protein was concentrated and the concentration was determined by UV absorption at 280 nm using an extinction coefficient of 13980 M^{-1} cm^{-1}. The protein was snap frozen in liquid N_2 and stored at -80 °C.

TBP expression and purification for crystallography: For expression and purification of the *Arabidopsis thaliana* TATA box binding protein (TBP) protein, *E. coli* C41(DE3) cells were transformed with a pET15b plasmid encoding the *A. thaliana* TBP (generated as a codon-optimized gene for *E. coli* expression) with a cleavable hexahistidine-tag on the N-terminus (Genscript)[19]. For protein expression, 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 (25 mM Tris pH 7.5, 500 mM NaCl, 5% (v/v) glycerol, 0.5 mM β-mercaptoethanol (BME)), lysed twice using a microfluidizer and pelleted. The supernatant was loaded onto a Cobalt-NTA column and the column was washed overnight with buffer A and then eluted with increasing concentrations of imidazole in buffer A. TBP eluted in buffer A containing between 100 and 300 mM imidazole. The resultant TBP was >95% pure at this stage. Concentrations were determined by UVvis using a calculated molar absorption coefficient of 10.5 mM^{-1} cm^{-1}. For crystallization the hexahistidine tag was removed using a Thrombin cleavage capture kit. The his-tag free protein was concentrated just prior to crystal setups using a centricon 30 (30 MW cutoff).

EGR1 expression and purification for crystallography: Recombinant EGR1 DNA-binding domain (residues 335 to 423) was expressed and purified from BL21*(DE3) *E. coli* harboring a clone in pGEX 2T-egr1-DBD (Addgene #85735) as previously reported[21]. In brief, an overnight culture was inoculated into liter-scale LB media and grown at 37 °C until OD reached around 0.5. The temperature was then lowered to 22 °C, and after 20 min, expression was induced by the 1 mM isopropyl β-D-1-thiogalactopyranoside overnight at 22 °C with shaking (175 rpm). Cells were harvested by centrifugation at 4000 rpm at 4 °C for 20 min and resuspended in the lysis buffer containing 20 mM Tris (pH 7.5), 500 mM NaCl, 5% (v/v) glycerol, 0.5 mM TCEP, 25 μM $ZnCl_2$ and PIC. Cells were lysed by French press and sonication, followed by centrifugation at 50,000 rpm at 4 °C for 30 min. 4% w/v solution (10X) poly(ethylenimine) in water, pH 8 was prepared by dissolving 1.6 g PEI (50% w/v) in 20 mL DDW. Solution was acidified to pH 8 by conc. HCl. 1/10 volume of 4% PEI solution was added to the supernatant, inducing the formation of a precipitate. The mixture was pelleted by centrifugation at 18,000 rpm at 4 °C for 10 min. GSTrap

HP 5 mL column was pre-equilibrated in the lysis buffer above. The cleared extract sample was loaded onto the column and non-GST tagged material was eluted in 100% lysis buffer. The GST fusion proteins were then eluted with 20 mM glutathione (GSH) in the elution buffer containing 100 mM Tris-HCl (pH 8.0), 5% (v/v) glycerol, 25 μM $ZnCl_2$, and 250 mM NaCl. The GST tag was removed using human Thrombin (Prospec, catalog #: PRO-339), incubated at room temperature with low-speed magnetic stirrer for 20 hours. The cleavage protein was diluted with 20 mM Tris (pH 7.5), 5% (v/v) glycerol, 25 μM $ZnCl_2$, and 0.5 mM TCEP buffer to lower salt concentration. HiTrap Heparin HP 5 mL column was pre-equilibrated in 20 mM Tris (pH 7.5), 5% (v/v) glycerol, 25 μM $ZnCl_2$, and 0.5 mM TCEP. The sample was loaded onto the column, eluted with a linear gradient to 1 M NaCl, collecting in fractions. The fractionated protein was pooled and concentrated with 3kDa MWCO Amicon (Millipore) at 4000 rpm at 4 °C to a volume of approximately 3 mL. Next, a Superdex 75 26/60 size exclusion column was pre-equilibrated overnight in 500 mM NaCl, 20 mM Tris-HCl (pH 7.5), 5% (v/v) glycerol, and 25 μM $ZnCl_2$. The sample was loaded onto the column and eluted as a single peak in 500 mM NaCl, 20 mM Tris-HCl (pH 7.5), 5% (v/v) glycerol, and 25 μM $ZnCl_2$. The eluate was concentrated and the concentration was determined by UV absorption at 280 nm using an extinction coefficient of at 1865 M^{-1} cm^{-1}. The protein was snap frozen in liquid N_2 and stored at -80 °C. Dynamic Light Scattering (DLS) measurement was carried out every time before the use of the protein to ensure there is no aggregation of the protein.

FOXA2 Plasmid Construction

The FOXA2 expression plasmid was constructed using the pET-29 backbone (Twist Bioscience). The insert included, in-frame and sequentially, an N-terminal GST tag, a TEV protease cleavage site, the human FOXA2 Forkhead domain (residues 142–269)[22], a second TEV site, and a C-terminal His_6 tag. The construct was synthesized and cloned by Twist Bioscience. The final plasmid was sequence-verified prior to downstream applications.

Site-directed mutagenesis

Site-directed mutagenesis was performed using KAPA HiFi HotStart DNA Polymerase (Roche) according to the manufacturer's standard protocol. Mutagenic primers were designed to introduce the desired substitution. Following PCR amplification, the reaction was treated with DpnI (NEB)

to digest the template plasmid, and the product was transformed into *E. coli* DH5α. Positive clones were confirmed by miniprep and Sanger sequencing.

The primers used to generate ETS1 mutants are as follows:
K379A: 5' - GGAGATGGGGAAAGAGGGCAAACAAACCTAAGATG - 3'
Y397A: 5' - GCCGTGGCCTACGCTACTATGCCGACAAAAACATCATCCAC - 3'
K404A: 5' - CGACAAAAACATCATCCACGCGACAGCGGGGAAACGCTACG - 3'
Y410A: 5' - GCGGGGAAACGCGCCGTGTACCGCTTTG - 3'

Universal Protein Binding Microarray

Transcription factor (TF) binding characterization was performed using universal protein binding microarrays (PBMs), as described previously [23,24]. Briefly, commercial microarrays (Agilent Technologies) containing all possible 10-mer or 9-mer sequences were converted to double-stranded DNA via solid-phase primer extension using Thermo Sequenase DNA Polymerase (Cytiva, Catalog #: E79000Y) and a deoxynucleotide triphosphate (dNTP) mixture (dATP, dCTP, dGTP, dTTP). Microarrays were then blocked with 2% (w/v) non-fat dry milk (Sigma, Catalog #: M7409) and incubated with the TF of interest. Binding reactions were carried out in protein-specific buffers as described below.

Pre-incubated protein binding mixtures were applied to individual microarray chambers, incubated for 1 h at room temperature, then subjected to two sequential washing steps. Next, microarrays were incubated for 1 h at room temperature with fluorescently labelled antibody diluted in protein-binding buffer supplemented with 2% milk, according to the epitope tag of the protein. Following antibody incubation, microarrays were subjected to two sequential washing steps as described below. Fluorescence signals of the bound proteins were recorded using a GenePix® 4400A scanner. Signal intensities were extracted using GenePix Pro 7.0 software and median pixel intensity was reported for each DNA probe, before further analysis.

Universal PBM analysis

To identify DNA motifs recognized by the TF of interest, we analyzed all possible 8-base sequences (8-mers) or 7-base sequences (7-mers) as previously described[24]. In brief, DNA features were grouped into two sets—those containing the 8-mer/7-mer (foreground) and those without (background). We then compare the top half of signal intensities from both sets using a modified

Wilcoxon-Mann-Whitney statistic, which helps identify the most enriched 8-mer/7-mer, termed the "seed" of the motif. Next, we assess the contribution of each nucleotide position within this seed by evaluating all possible nucleotide variants at each position, and the motif was further refined by including gaps at positions with high variability. Finally, we convert the derived motif into a position weight matrix (PWM), allowing for a quantitative representation of the TF's binding specificity.

PIC-NIC library design and measurements

PIC-NIC experiments were performed as follows. Libraries of DNA complexes with a nick at every possible position in the TF binding site were designed for each TF, with and without the phosphate at 5′ end of the breakage site. Nick DNA complexes were constructed in two ways. In the first method, a single-stranded oligo was designed to fold on itself upon annealing, forming a dumbbell-shaped DNA, with a nick at the desired position and an amino modifier on the loop of the dumbbell. In the second method, three strands of oligos were designed to anneal and form a duplex, with the two shorter strands being the complements of the longer strand, leaving a nick at the position desired (Extended Data Fig. 1). The longer strand is labelled with an amino modifier. In the first method, the negative control sequence contained a nick remote from the TF binding site; in the second method, the negative control sequence was the corresponding intact double-stranded duplex.

For each probe, thermodynamic parameters of hybridization were evaluated using the web-based tools from Integrated DNA Technologies and NUPACK, to verify the integrity of annealing under experimental conditions. Simulations incorporated strand concentrations and ionic conditions mimicking experimental settings (150 mM Na^+, 150 mM Mg^{2+}), with appropriate ion correction terms. The maximum complex size was set to 2 for intact duplexes and 3 for nicked complexes involving three strands.

The libraries were spotted onto epoxy-functionalized glass slides using a sciFLEXARRAYER S12 automated non-contact dispensing system (Scienion), followed by rehydration and blocking steps to yield PIC-NIC chips as described below. Protein binding experiments on PIC-NIC chips were performed using the same conditions as universal PBMs as described below.

DNA microarray production for PIC-NIC assays

Sample Preparation

Homemade microarrays were produced in-house for PIC-NIC assays. Oligonucleotides were purchased from Integrated DNA Technologies (IDT) and W.M. Keck Biotechnology Resource Lab (Yale University). For nicks created from dumbbell-shaped DNA, single-stranded oligos were ordered with an internal amino modifier /iAmMC6T/, as well as a choice of 5' phosphate (/5Phos/). The sequences were designed to self-anneal into a dumbbell, leaving a nick at the desired position. The purchased oligos were each solubilised to 100 μM concentration according to the manufacturer datasheet, before accurate concentration quantification with Thermo Scientific Nanodrop One Microvolume UV-Vis Spectrometer. The oligos were next each diluted to a concentration of 5 μM in the spotting buffer consisting of 100 mM Na_2CO_3/$NaHCO_3$ (pH 8.5). The oligo was then heated to 95 ºC for 2 min, followed by fast cooling on ice. The annealed oligos were centrifuged for 5 min, and 20 μl of the clear solution was plated in sciSOURCEPLATE 384 PS (Scienion). For nicked duplex oligos constructed from three DNA strands, the longest DNA part was ordered with an amino modifier at the end of the oligo (/5AmMC6/ or /3AmMC6T/). The purchased oligos were each solubilised to 100 μM concentration according to the manufacturer datasheet, before accurate concentration quantification with Thermo Scientific Nanodrop One Microvolume UV-Vis Spectrometer. The oligos were next mixed to a final construct concentration of 5 μM with the desired stoichiometry (1:5:5, with the long and labelled strand being 1, to ensure this strand is fully consumed) in the buffer consisting of 100 mM Na_2CO_3/$NaHCO_3$ (pH 8.5). The oligo construct was then heated to 95 ºC for 2 min, followed by slow gradient cooling over 2 hours to ensure proper annealing. The annealed oligos were centrifuged for 5 min, and 20 μL of the clear solution was plated in sciSOURCEPLATE 384 PS (Scienion).

PIC-NIC DNA microarray production

The prepared DNA samples were printed onto the chips with our in-house sciFLEXARRAYER S12 automated non-contact dispensing system (Scienion). The empty slides used for printing were glass slides coated with epoxy surface (sciCHIP Epoxy from Scienion, and 3-D Epoxy from PolyAn). The layout of the microarray were designed, with 10-50 replicates of each material in the plate. The plated materials were spotted onto the glass slide at a temperature between 20-24 ºC and 70 ± 5 % relative humidity. Each droplet was dispensed at a volume of 300 ± 30 pL, forming a DNA spot on the glass slide with diameter of 90 ± 10 μm. After spotting, the DNA microarray was placed in a homemade humidity chamber filled with 3 M KCl (~70% humidity) overnight for rehydration and immobilisation. The DNA microarray was then placed in a coplin jar filled with

PolyAn Blocking Buffer A for an hour, to deactivate the unreacted epoxy groups on the glass slide. Next, the microarray was subjected to a salt gradient wash in 5 coplin jars filled separately with PolyAn Washing Buffer I, PolyAn Washing Buffer II, PolyAn Washing Buffer III, DDW and DDW. Each wash was 5 min, shaking at 125 rpm. The resulting microarray was spun dry in a slide spinner for 3 min (Labnet International, Inc.) and stored in a vacuum dessicator until further use.

Protein binding and antibody steps for PBM and PIC-NIC assays

Protein binding reactions were performed in the same conditions as previously described in PBM protocols[23,24]. The binding buffer, unless otherwise specified, contains PBS / 2% (w/v) milk / 51.3 ng/µl salmon testes DNA (Sigma) / 0.2 µg/µl bovine serum albumin (NEB) / 1 mM dithiothreitol. For EGR1, SP1, and CTCF, the binding buffer was: 10 mM Tris-HCl (pH 7.5), 150 mM KCl, and 0.2 µM $ZnCl_2$, 2% (w/v) milk, 51.3 ng/µl salmon testes DNA, 0.2 µg/µl bovine serum albumin, 1 mM dithiothreitol. For TBP, the binding buffer was 10 mM HEPES, 70 mM KCl, 10 mM $MgCl_2$, 1 mM EDTA, 2% (w/v) milk, 51.3 ng/µl salmon testes DNA, 0.2 µg/µl bovine serum albumin, 1 mM dithiothreitol. For CREB1, the binding buffer was 25 mM Tris-HCl (pH 7.4), 0.5 mM EDTA, 2% (v/v) glycerol, 5 mM $MgCl_2$, 50 mM KCl, 25 mM boric acid, 2% (w/v) milk, 51.3 ng/µl salmon testes DNA, 0.2 µg/µl bovine serum albumin, 1mM dithiothreitol. For Sox2, the binding buffer was 10 mM HEPES (pH=7.5), 1 mM $MgCl_2$, 0.01 mM $ZnCl_2$, 10 mM NaCl, 2% (wt/vol) milk, 51.3 ng/µl salmon testes DNA, 0.2 µg/µl bovine serum albumin, 5% glycerol, 1 mM dithiothreitol. For FoxA2, the binding buffer was Tris-HCl (pH 8.0) 20 mM, KCl 40 mM, $MgCl_2$ 2 mM, $ZnCl_2$ 2 µM, 2% (wt/vol) milk, 51.3 ng/µl salmon testes DNA, 0.2 µg/µl bovine serum albumin, 1 mM dithiothreitol.

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% (v/v) 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) in the protein binding buffer / 2% milk were applied on the chip for 1 h at room temperature. Max protein was fluorescently tagged with TAMRA, so no antibody was used. The chips were washed twice

with PBS / 0.05% (v/v) Tween-20 for 3 min and then once with PBS for 2 min. 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 (at 635 nm, 532nm or 488nm) of bound TFs for each DNA spot was measured at 2.5 μm resolution using a GenePix 4400A® microarray scanner. Signal intensities were extracted using GenePix Pro 7.0 software, and median pixel intensity was reported for each DNA probe. Multiple replicates (10 to 20 replicates) of each sequence were used to quantitatively compare the binding signals between sequences.

Classification of the nicking effect size for PIC-NIC assay

To quantify the functional contribution of nick at each position to protein-DNA binding, we measured binding signal changes upon nicking (with phosphate and without phosphate) in PIC-NIC assay. For each position, the effect size was calculated by normalizing the binding signal to a scale where the unnicked sequence (positive control) was set to 100% and a nonspecific binding site (negative control) was set to 0%. The percentage signal change (wP: nicking with the phosphate; noP: nicking with the removal of phosphate; Supplementary Table 2) was then determined by subtracting the normalized binding signal of the nicked sequence from 100%. Based on these effect sizes, nicking effects were classified into four categories: minimal (less than 10% loss), minor (10–50% loss), major (50–90% loss), and severe (greater than 90% loss).

Bio-layer interferometry (BLI) measurements

Bio-layer interferometry (BLI) assays were performed using an Octet Red96e System (ForteBio; Menlo Park, CA) in 96-well plates. Streptavidin Octet biosensors (ForteBio; Menlo Park, CA) were dipped into nuclease-free water (Sartorius) for 10 min to hydrate. To obtain the baseline, the sensors were dipped for 60 sec in the kinetic buffer, before dipping into 200 μl of biotinylated DNA probe (50 nM) in the kinetic buffer for the loading step. The DNA probes were constructed by annealing two strands to form an intact binding site or three strands to form a nicked site. Each oligo was purchased from Integrated DNA Technologies (IDT), reconstituted to 100 μM in Nuclease-Free Duplex Buffer (IDT) and mixed to form 1 μM final concentration with the desired stoichiometry (1:1.2 or 1:1.2:1.2, with the biotinylated strand being 1). The resulting mixture was heated to 95 ºC for 5 min and cooled down on ice at 4 ºC. The loading was manually stopped when the response signal reached 0.4 nM for each DNA probe, to ensure an equal amount of DNA probes were immobilised on each sensor. Sensors were then dipped into the kinetic buffer for 60 sec.

Next, the protein association step was performed in kinetic buffer at the indicated concentrations for 600 sec (ETS1) or 800 sec (Sox2) to obtain the association curve. Then, the tips were dipped in kinetic buffer again for 400 sec to obtain the dissociation curve. Sensors were then regenerated by dipping the tips in regeneration buffer (2 M NaCl) for 5 sec and then kinetic buffer for 5 sec, repeated three times. All measurements were carried out at 25 ºC. Data were analysed within the ForteBio Data Analysis software. For ETS1, association and dissociation curved were fitted locally with a 1:1 binding model. For Sox2, association and dissociation curved were fitted globally with a 1:1 binding model. k_{on}, k_{off}, kinetic K_D and thermodynamic K_D were obtained and reported.

The kinetic buffer for ETS1 contains PBS, 5 mM $MgCl_2$ and 0.05% Tween-20. ETS1 protein used was untagged ETS-DBD (residues 280 to 440, 13.6 μM). The kinetic buffer for Sox2 contains 50 mM HEPES (pH 7.5), 150 mM NaCl, 5 mM $MgCl_2$ and 0.05% Tween-20. Sox2 protein used was His-Sox2 as in the PBM experiments.

ETS1–PARP1 Competition Assay

PIC-NIC chips containing nicked ETS1 binding site probes were prepared as described above. Competition experiments were performed in a humidity chamber (~70% humidity). ETS1 protein (500 nM or 1 μM where indicated) was applied to the chip in PBS supplemented with 1 mM DTT and incubated for 30 minutes at room temperature. The protein solution was removed by pipetting, without washing, to maintain ETS1 occupancy at bound sites. 50 nM PARP1 (Human Recombinant, FLAG-tagged, OriGene cat#: TP710053) was then applied in the same buffer and incubated for 30 minutes at room temperature. Following incubation, chips were washed and incubated with a fluorescently labeled anti-FLAG antibody as described above to detect PARP1 binding, using the same antibody incubation and scanning conditions as standard PIC-NIC experiments. Fluorescence signals were recorded using a GenePix 4400A scanner and extracted using GenePix Pro 7.0 software. Each chip chamber contained 20 spatially scattered replicates per unique probe; the median intensity across replicates was used for analysis. For each condition, the natural logarithm of the median fluorescence intensity was computed, and binding signals were compared between PARP1-alone and ETS1+PARP1 conditions using a paired t-test across the 14 nicked ETS1 binding site probes. Position-specific comparisons between selected probes (n2, n5, and ren5) were performed using unpaired t-tests.

Crystallisation and structure determination of EGR1-nicked DNA complexes

EGR-DNA complexes were prepared as previously described[25] and subjected to hanging drop vapor diffusion crystallization screens, yielding large and well-diffracting crystals suitable for data collection in the initial screens. Data for all the crystals were collected at The Dana and Yossie Hollander Center for Structural Proteomics at the Weizmann Institute of Science. Initial models were iteratively rebuilt and refined using Coot[26] and Phenix[27]. Model geometry was evaluated using MolProbity[28]. See Extended Data Table 2 for the final data collection and refinement statistics for each structure.

The duplex DNA sequences used for crystallisation were formed by annealing three strands together. The oligos were purchased as a powder from Integrated DNA Technologies (IDT), each solubilised to 1 mM concentration. The three strands were mixed together with the desired stoichiometry (1:1:1.2, with the shortest strand being 1.2) to a concentration of 220 μM in the buffer consisting of 125 mM bis-trispropane HCl (pH 8.0) and 500 mM NaCl. The mixture was heated to 95 ºC for 5 min, followed by overnight slow cooling to 4 ºC to allow for proper annealing. Next, to obtain EGR1 crystals with various nicked DNA sites, the protein was centrifuged for 10 min at 4 ºC to get rid of any possible precipitation, measured concentration with Thermo Scientific Nanodrop One Microvolume UV-Vis Spectrometer, and mixed with the annealed DNA sites at the desired stoichiometry (1:1.1 ratio for protein:DNA) to a final complex concentration of 100 μM. The resulting mixture was then concentrated around 10-fold using a Vivacon® 500 (2,000 MWCO Hydrosart; Sartorius), to a final complex concentration of 1 mM. The resultant protein-DNA complexes were then used in vapour diffusion crystallisation screens.

EGR1-r7noP: Crystals were grown at 19°C using the hanging drop vapor diffusion method. The well solution contained 0.2 M Sodium chloride, 0.1 M MES (pH 6.0), and 20% (w/v) polyethylene glycol monomethy lether (PEG-MME) 2000. Diffraction data were collected to 1.9 Å resolution at 100 K using an in-house Rigaku liquid-metal-jet (LMJ) X-ray Synergy System with a HyPix Arc 150° detector. EGR1-r7noP crystallised in the $C222_1$ space group, with one subunit in the asymmetric unit. The structure of the Zif268 protein-DNA complex (PDB code 1AAY)[29] was used as a model for molecular replacement. See Extended Data Table 2 for final data collection and refinement statistics.

EGR1-ren5P: Crystals were grown at 19°C using the sitting drop vapor diffusion method. The well solution contained 0.1 M Magnesium acetate tetrahydrate, 0.1 M Sodium cacodylate (pH 6.5), and 15% (w/v) PEG 6000. Diffraction data were collected to 2.0 Å resolution. EGR1-ren5P crystallised in the $C222_1$ space group, with one subunit in the asymmetric unit. EGR1-r7noP structure was used to generate the EGR1-ren5P model for molecular replacement. See Extended Data Table 2 for final data collection and refinement statistics.

EGR1-ren7noP: Crystals were grown at 19°C using the sitting drop vapor diffusion method. The well solution contained 0.1 M Calcium acetate, 0.1 M MES (pH 6.0) and 15% (v/v) PEG 400. Diffraction data were collected to 1.95 Å resolution. EGR1-ren7noP crystallised in the $C222_1$ space group, with one subunit in the asymmetric unit. The structure of the Zif268 protein-DNA complex (PDB code 1AAY)[29] was used as a model for molecular replacement. See Extended Data Table 2 for final data collection and refinement statistics.

Model Building and Refinement:

Initial models were iteratively rebuilt and refined using Coot[26] and Phenix[27]. Model geometry was evaluated using MolProbity[30].

Data Availability Statement

Atomic coordinates and structure factors for TBP-AG_noP, TBP-AG_P, TBP-TG_noP and TBP-TG_P are deposited in the PDB database under accession numbers 9OWI, 9OWZ, 9OW7 and 9OW8, respectively. Atomic coordinates and structure factors for EGR1-r7noP, EGR1-ren5P, and EGR1-ren7noP are deposited in the PDB database under accession numbers 9RIC, 9RI6, and 9RJ6, respectively.

Crystallisation and structure determination of TBP-nicked DNA complexes

TBP–DNA complexes were prepared and subjected to hanging drop vapor diffusion crystallization screens, resulting in large, well-diffracting crystals suitable for data collection after optimization of initial hits. Data for all crystals were collected at the Advanced Light Source (ALS) on beamlines 5.0.1 and 5.0.2. The data were processed with XDS[31]. The structures were solved by molecular replacement (MR) using a previous structure of TBP as a search model. MolProbity was

used to guide the process of refitting and refinement[30]. See Extended Data Table 3 for the final data collection and refinement statistics for each structure.

TBP-AG_noP: To obtain crystals of the nicked TBP-G_noP DNA with TBP, a DNA site with a 4 bp complementary overhang was used (based on the AMVV DNA site). The DNA sites to generate the duplex were (5′-CTATAAAAGCGC-3′ and 5′-TTTTATAG-3′). This duplex DNA was mixed at a 1:1 stoichiometry with TBP (at 10 mg/mL) and hanging drop vapor diffusion screens (Wizard I-IV) were carried out at room temperature (rt). Crystals were obtained by mixing the protein-DNA complex 1:1 with a crystallization solution consisting of 20% (w/v) PEG 3350, 0.1 M Citric acid (pH 5.0) and 0.2 M sodium citrate. The crystals took several days to grow and contain 4 apo TBP molecules and two protein-DNA complexes in the crystallographic asymmetric unit (ASU). The crystals were cryo-preserved by dipping them in a solution consisting of the crystallization reagent supplemented with 20% (v/v) ethylene glycol for 2 s before plunging them into liquid nitrogen. Data were collected at the advanced light source (ALS) beamline 5.0.2 and processed with XDS[31]. The structure was solved by molecular replacement (MR) using the pdb 6UEP as a search model in Phenix[27]. Multiple rounds of refitting in Coot[26] and refinement in Phenix[27] was carried out to convergence. See Extended Data Table 1 for data collection and refinement statistics.

TBP-AG_P: To obtain crystals of the nicked TBP-AG_P DNA with TBP, a DNA site with a 4 bp complementary overhang with a 5′phosphate within the nick was used. The DNA sites to generate the duplex were (5′-GCTATAAAAGCGC-3′ and 5′-P-TTTTATAGC-3′). This duplex DNA was mixed at a 1:1 stoichiometry with TBP (at 10 mg/mL) and hanging drop vapor diffusion was used. Wizard I-IV screens were employed at room temperature. Crystals were obtained by mixing the protein-DNA complex 1:1 with a crystallization solution consisting of 10% (w/v) PEG 8000, 0.1 M CHES pH 9.5. The crystals took several days to grow and contain 6 protein-DNA complexes in the ASU. The crystals were cryo-preserved by dipping them in a solution consisting of the crystallization reagent supplemented with 20% (v/v) ethylene glycol for 2 s before plunging them into liquid nitrogen. Data were collected at the advanced light source (ALS) beamline 5.0.2 and processed with XDS[31]. The structure was solved by molecular replacement (MR) using the TBP-TBP-AG_noP protein-DNA complex as a search model[32]. See Extended Data Table 1 for final data collection and refinement statistics.

TBP-TG_noP: To obtain crystals of the nicked TBP-TG_noP DNA with TBP, the DNA sites used to generate the duplex were (5′-GCTATAAATGCGC-3′and 5′-ATTTATAGC-3′). This duplex DNA was mixed at a 1:1 stoichiometry with TBP (at 10 mg/mL) and hanging drop vapor diffusion was used. Wizard I-IV screens were employed at room temperature. Crystals were obtained by mixing the protein-DNA complex 1:1 with a crystallization solution consisting of 4 M sodium Formate, 0.1 M sodium acetate (pH 3.8). The crystals took several weeks to grow to optimal size and contain 1 protein-DNA complex in the ASU. The crystals were cryo-preserved straight from the drop, plunging them directly into liquid nitrogen. Data were collected at the advanced light source (ALS) beamline 5.0.2 and processed with XDS[31]. The structure was solved by molecular replacement (MR) using the TBP-TBP-AG-noP protein-DNA complex as a search model[26,32]. See Extended Data Table 1 for final data collection and refinement statistics.

TBP-TG_P: To obtain the TBP-TBP-TG_P complex a 4 bp complementary overhang with a 5′ phosphate within the nick was used. The DNA sites to generate the duplex were (5′-GCTATAAAAGCGC-3′ and 5′-P-TTTTATAGC-3′). This duplex DNA was mixed at a 1:1 stoichiometry with TBP (at 10 mg/mL) and hanging drop vapor diffusion was used. Wizard I-IV screens were employed at room temperature to find crystallization conditions. Crystals were obtained by mixing the protein-DNA complex 1:1 with a crystallization solution consisting of 20% (w/v) PEG 3350, 0.1 M Citric acid (pH 5.0) and 0.2 M sodium citrate. The crystals took several days to grow and contain 4 apo TBP and two protein-DNA complexes in the ASU. The crystals were cryo-preserved by dipping them in a solution consisting of the crystallization reagent supplemented with 20% (v/v) ethylene glycol for 2 s before plunging them into liquid nitrogen. Data were collected at the advanced light source (ALS) beamline 5.0.2 and processed with XDS[31]. The structure was solved by molecular replacement (MR) using the TBP-TBP-TG_noP protein-DNA complex as a search model. See Extended Data Table 1 for final data collection and refinement statistics.

All-atoms molecular dynamics simulation

ETS1 simulation

The structure of the DNA (5′ TAGTGCCGGAAATGT 3′) with and without ETS1 (PDB ID: 1K79) were nicked at position 7 on either strand using PyMOL (C7 and G26). The unphosphorylated nicked strand systems and unnicked strand systems were parameterised and

fitted in the solvent box using CHARMM-GUI Solution Builder [33]. The nicked strand systems were fitted in the solvent box using GROMACS 2024.2[34]. Each box was filled with TIP3P water model, neutralised with 0.15 M KCl, and has the dimension of 8.1 × 8.1 × 8.1 nm^3. The composition of each simulation box is noted in Supplementary Table 10. The protein was simulated with AMBERff19sb[35] forcefield and the DNA was simulated with OL15[36] forcefield. The systems were then energy minimized using a steepest-descent algorithm for 5000 steps and equilibrated for 10 ns, maintaining the temperature at 310 K using a V-rescale thermostat[37]. Both energy minimisation and equilibration were performed with the restraint of 400 kJ nm^{-2} mol^{-1} on the heavy atoms of DNA and protein's backbone, and 40 kJ nm^{-2} mol^{-1} on the heavy atoms of DNA and protein's sidechain. Then, the system was simulated using 2 fs timesteps and maintained the temperature at 310 K using a V-rescale thermostat and the c-rescale barostat[38] for isotropic pressure coupling at 1 bar for 1 μs for 3 repeats. The hydrogen bond in the simulation was constrained using LINCS algorithm[39]. All simulations and hydrogen bond analyses were performed using GROMACS 2024.2.

SOX2 simulation

The structure of the DNA (5′ CCCCATTGTTATGC 3′) with and without Sox2 (PDB ID: 6HT5) were nicked at either C4 or T7 using PyMOL. All systems were parameterised using AmberTools25[40] and fitted in the solvent box using GROMACS 2024.2[34]. The phosphate group was added to the 5' nicked end through terminal_monophosphate.lib using tleap. The bond angle and the angular force of OH-P-O2 and HO-OH-P was adjusted to match the parameters of the phosphate group of AMBERphossa19sb[41]. Each box was filled with TIP3P water model, neutralised with 0.15 M KCl, and has the distance from the edge of the solvent box of 1 nm. The protein was simulated with AMBERff19sb[35] forcefield and the DNA was simulated with OL15[36] forcefield. The systems were then energy minimized using a steepest-descent algorithm for 5000 steps and equilibrated for 10 ns, maintaining the temperature at 310 K using a V-rescale thermostat[37]. Both energy minimisation and equilibration were performed with the restraint of 400 kJ nm^{-2} mol^{-1} on the heavy atoms of DNA and protein's backbone, and 40 kJ nm^{-2} mol^{-1} on the heavy atoms of DNA and protein's sidechain. Then, the system was simulated using 2 fs timesteps

and maintained the temperature at 310 K using a V-rescale thermostat and the c-rescale barostat[38] for isotropic pressure coupling at 1 bar for 1 μs for 3 repeats. The hydrogen bond in the simulation was constrained using LINCS algorithm[39]. All simulations and hydrogen bond analyses were performed using GROMACS 2024.2.

Structural analysis of Watson-Crick and nicked DNA structures

X-ray crystal structures and NMR structures for the desired canonical DNA-protein complexes were downloaded with their PDB information including resolution, macromolecule type etc. from the RCSB webserver. Structures were parsed using X3DNA-DSSR[42] and Curves+ web server[43], the structural parameters were extracted and plotted for the purpose of comparison.

The structural analysis of the protein-nicked DNA complex was done on the X-ray crystal structures we solved. In X3DNA-DSSR, the data was input as it was in the PDB file. In Curves+ web server, the DNA structure was input as two strands, ignoring the nick. It was only done for the structures with 5′ phosphate at the breakage site, to ensure a well-defined helical backbone.

Structural parameter deviation analysis

DNA positions from all protein-DNA complexes were classified into two groups based on protein contact status and binding signal change after a nick. The target group consisted of positions that had no contact with protein at the base, no contact with protein at the phosphate, and a prominent percentage signal change value upon nicking (column “prominent wP”) less than −50 (Supplementary Tables 1-2). The reference group consisted of positions that had no contact with protein at the base, no contact with protein at the phosphate, and a “prominent wP” value greater than or equal to −50. Total positions analyzed: 234; Positions with neither base nor phosphate contact: 99; Target group: 21 positions; Reference group: 78 positions.

Twelve DNA structural parameters were analyzed, comprising six base pair parameters and six base pair step parameters. The base pair parameters and their reference distributions (mean ± standard deviation) were: Shear (0 ± 0.383 Å), Stretch (−0.145 ± 0.169 Å), Stagger (0.055 ± 0.287 Å), Buckle (1.245 ± 8.501°), Propeller (−9.713 ± 6.477°), and Opening (1.11 ± 4.487°). The base pair step parameters and their reference distributions were: Shift (−0.034 ± 0.587 Å), Slide (−0.412

± 0.518 Å), Rise (3.302 ± 0.266 Å), Tilt (−0.597 ± 3.087°), Roll (2.151 ± 6.178°), and Twist (32.685 ± 5.245°). These reference values were obtained as calculated before [44].

For each structural parameter, a position was classified as deviant if the absolute difference between its observed value and the reference mean exceeded one standard deviation. The percentage of deviant positions was calculated for each group by dividing the number of deviant positions by the total number of positions in that group and multiplying by 100.

To test whether the target group exhibited significantly different deviation frequencies compared to the reference group, Fisher's exact test was performed for each parameter. For each test, a 2×2 contingency table was constructed containing the number of deviant and non-deviant positions in each group. Statistical significance was defined as $p < 0.05$. All analyses were performed using Python with the pandas, NumPy, and SciPy libraries.

Twist: p=1.72e-02; Buckle: p=1.76e-02 (Fisher's exact test)

Buried Surface Area Analysis of Phosphate–Protein Contacts

To quantify the extent of phosphate–protein contacts at each position in the TF binding site, we performed buried surface area (BSA) analysis using UCSF ChimeraX (version 1.9) on all high-resolution TF–DNA co-crystal structures available for our panel. For each structure, the analysis was carried out as follows.

For each backbone phosphate position along both strands of the binding site, we selected the three atoms defining the phosphate group: the phosphorus atom (P) and its two non-bridging oxygens (OP1 and OP2). The buried surface area (BSA) at each position was calculated using ChimeraX, and the solvent-accessible surface area (SASA) was calculated separately for each selected phosphate group using a probe radius of 1.4 Å. The percentage burial was then calculated as % Buried = BSA/SASA × 100, representing the fraction of the exposed phosphate surface area that becomes buried upon protein binding. All per-position values are provided in Supplementary Table 7.

This procedure was repeated for all phosphate positions along both the top and bottom strands of the binding site for each TF–DNA structure. Positions were numbered according to the PIC-NIC convention for each TF. The resulting % buried values were then correlated with the corresponding PIC-NIC binding changes (% change in binding signal upon phosphate removal) using Pearson correlation, computed separately for the top and bottom strands and for nick conditions with and

without the 5′ phosphate. All per-position BSA, SASA, and % buried values are provided in Supplementary Table 7.

deepDNAshape and DNA Bendability Analysis

For deepDNAshape analysis, the highest-affinity binding sequence for each TF was identified from universal PBM data (E-score ranking; Supplementary Table 5). For each position within the binding site, all possible dinucleotide substitutions (16 combinations, denoted NN) were introduced while keeping the remaining sequence fixed. Predicted DNA shape features — including shear, stretch, stagger, buckle, propeller twist, opening, minor groove width (MGW), shift, slide, rise, tilt, roll, and helical twist — were computed at the substituted position and adjacent base steps or base pairs using the deepDNAshape analysis package[45,46]. The predicted shape values were then correlated with experimentally measured E-scores for the corresponding sequences from our universal PBM data using Pearson's r. Statistical significance was assessed using both Bonferroni correction and FDR correction to account for multiple testing across positions and shape features. Full results for all TFs and all shape features are provided in Supplementary Table 3.

For DNA bendability analysis, two independent approaches were used. First, trinucleotide bendability scores were obtained from Brukner et al.[1], which provides experimentally derived bendability values for all 64 trinucleotides based on DNase I cleavage frequencies, reflecting the intrinsic mechanical properties of free, unbound DNA. Second, base-resolution bendability predictions were obtained using the BendNet deep learning server[2,3]. For each TF, every trinucleotide position within the highest-affinity binding sequence was systematically varied by substituting all 64 possible trinucleotide combinations at that position while keeping the flanking sequence fixed. Predicted bendability values were then correlated with experimentally measured E-scores from our universal PBM data using Pearson's r. Full results are provided in Supplementary Table 4.

Genome-wide analysis of single-strand breaks and repair signals at transcription factor binding sites

Data Acquisition

EGR1 ChIP-seq peaks: EGR1 binding sites were obtained from ENCODE[47-49] as IDR-ranked peaks (accession: ENCFF016RNL, GSE230933; from Michael Snyder Lab at Stanford University)

mapped to the GRCh38 human genome assembly, comprising 27,684 peaks. CTCF ChIP-seq peaks: CTCF binding sites were obtained from ENCODE as IDR-thresholded peaks (file accession: ENCFF930BXV, experiment: ENCSR260FAS; from Bradley Bernstein at Broad Institute) from CTCF ChIP-seq in neural progenitor cells, mapped to the GRCh38 human genome assembly, comprising 55,700 peaks. DNA single-strand break data: S1-END-seq data from induced pluripotent stem cell-derived neurons (iNeurons) were obtained from Wu et al.[50] (GSM5100382). Both positive and negative strand BigWig files were used to quantify nick signals. DNA repair signals: Poly-ADP-ribose (PAR) ChIP-seq (GSM5100373), XRCC1 ChIP-seq (GSM5100374), and SAR-seq data (GSM5100400-402, three replicates) from iNeurons were obtained from Wu et al[50]. Motif positions: Genome-wide EGR1 motif matches (MA0162.5) were obtained from JASPAR[51], yielding 4,556,356 motif occurrences in the hg19 genome. Genome-wide CTCF motif matches (MA0139.2) were obtained from JASPAR[51], yielding 2,079,316 motif occurrences in the hg19 genome. ENCODE hg19 blacklist v2[52] was used.

Coordinate Conversion

EGR1 ChIP-seq peaks and CTCF ChIP-seq peaks were converted from GRCh38 to hg19 coordinates using the UCSC liftOver tool with the hg38ToHg19 chain file. For EGR1, of 27,684 original peaks, 27,615 (99.75%) were successfully converted. For CTCF, of 55,700 original peaks, 55,595 (99.81%) were successfully converted. ENCODE blacklist regions were removed from all analyses.

Nick Signal Quantification and Footprint Analysis

For each genomic region, nick signals were calculated by summing the absolute values from both positive and negative strand BigWig files using pyBigWig. Total nick signal was computed across all EGR1 and CTCF peak regions. Aggregate nick profiles were generated by centering on EGR1 and CTCF peak midpoints and computing mean signal across a ±2 kb window size. Strand-specific profiles were maintained to assess asymmetry in nick distribution.

Motif-Centered Analysis

For each EGR1 and CTCF ChIP-seq peak, the highest-scoring EGR1 motif (by JASPAR score) within the peak was identified using PyRanges interval operations. Analysis windows were re-centered on motif midpoints (±100 bp), retaining motif strand orientation. Of 27,473 peaks, 7,976

(29%) contained at least one EGR1 motif. Of 55,113 peaks, 45,689 (83%) contained at least one CTCF motif.

Enrichment Analysis

Nick enrichment at EGR1 and CTCF binding sites was assessed using permutation testing. For each of 1,000 permutations, random genomic regions of equal number and size to the observed peaks were sampled, with chromosomes selected proportional to chromosome length. Regions overlapping the ENCODE hg19 blacklist were excluded via interval intersection (PyRanges). For each permutation, the total S1-END-seq nick signal was quantified across all sampled regions to construct a null distribution. Empirical p-values were calculated as the proportion of permuted values exceeding the observed signal; fold enrichment was computed as observed / mean(null), and a z-score as (observed − mean(null)) / std(null). To estimate the precision of the observed signal, 1,000 bootstrap resamples (with replacement) of the original peaks were performed, and 95% confidence intervals were derived from the 2.5th and 97.5th percentiles.

SAR-seq Peak Calling

SAR-seq peaks were called from the replicate-averaged iNeuron SAR-seq signal track using a percentile-threshold approach implemented in Python with pyBigWig and PyRanges. The genome was tiled into non-overlapping 200 bp windows across standard chromosomes (chr1–22, X, Y), and the mean signal was computed per window. Windows with signal at or above the 99th percentile of the genome-wide distribution (top 1%) were retained as candidate regions. Adjacent retained windows were merged into peaks using PyRanges, and merged regions were size-filtered to 200–10,000 bp. ENCODE hg19 blacklist regions were removed by interval intersection. Each retained peak was assigned a score equal to its mean SAR-seq signal. This yielded 98,395 high-confidence repair sites with a median size of 200 bp, covering ~1.0% of the genome.

SAR-seq peaks were called from the iNeuron SAR-seq signal track using a custom percentile-threshold approach in Python (pyBigWig, PyRanges). For each standard chromosome (chr1–22, X, Y), per-base signal values were extracted and missing values (NaN) were replaced with zero. The genome was tiled into non-overlapping 200 bp windows, and the mean signal was computed per window. Windows with signal ≥ the 99th percentile of the genome-wide distribution (top 1%) were retained, and adjacent retained windows were merged into peaks using PyRanges. Merged peaks were size-filtered to 200–10,000 bp, and regions overlapping the ENCODE hg19 blacklist[52]

were removed by interval intersection. Each retained peak was scored by its mean SAR-seq signal. This procedure yielded 98,395 peaks (median size 200 bp, ~1.0% genome coverage).

Statistical assessment of the central SAR-seq footprint dip at TF motifs

To assess whether the central dip in the average motif-centered SAR-seq profile was statistically significant, the n motif sites were randomly partitioned (without replacement; fixed random seed for reproducibility) into 50 disjoint subgroups of approximately equal size (~160 sites per group for EGR1, ~913 for CTCF). For each subgroup, the mean signal across its constituent sites was computed at three locations relative to the motif center: (1) Center: the average of the per-site mean signal over a ±10 bp window centered on the motif (positions −10 to +10); (2) Left border: the per-site mean signal at a single position located W_b bp upstream of the motif center; (3) Right border: the per-site mean signal at a single position located W_b bp downstream of the motif center. The half-width W_b of the border window was set to 100 bp for EGR1 and 50 bp for CTCF, chosen to lie outside the central dip but inside the local flanking maxima of each curve.

This procedure yielded 50 paired observations of (left, center, right) signal values, one per subgroup. The center distribution was compared independently to the left and right border distributions using a one-sided paired t-test (scipy.stats.ttest_rel, alternative='less'), testing the null hypothesis that the per-subgroup center signal is not lower than the border signal. The resulting t-statistics and p-values are as following:

CTCF: Center < Left: t = -3.08, p = 1.716e-03; Center < Right: t = -21.59, p = 5.480e-27

EGR1: Center < Left: t = -2.40, p = 1.005e-02; Center < Right: t = -4.86, p = 6.298e-06

Software

Analyses were performed in Python using pyBigWig for BigWig file access, PyRanges for genomic interval operations, and pandas for data manipulation. UCSC liftOver was used for coordinate conversion.

Data availability

The data that support the findings in this study are available as Supplementary Tables in Excel format. Coordinates and structure factor amplitudes for the TBP-AG-withP, TBP-AG-noP, TBP-TG-withP, TBP-TG-noP, EGR-n7, EGR-ren5P and EGR-ren7 structures have been deposited in

the PDB under the accession codes 9OWZ, 9OWI, 9OW8, 9OW7, 9RIC, 9RI6, and 9RJ6, respectively. The PDB entries used in this study are available in Extended Data Fig. 2 and Supplementary Table 2. Molecular dynamics simulations tpr, mdp and itp files are stored at doi:10.5281/zenodo.15646043 and 10.5281/zenodo.15646042.

Titles and Legends for Supplementary Tables 1-8

Supplementary Table 1. PIC-NIC data.
This file contains the raw PIC-NIC data for the 15 TFs.

Supplementary Table 2. Structural parameters and contact information.
This table contains structural parameters generated by x3DNA and H bond contacts defined by ChimeraX at every position in the DNA binding sites for the 15 TFs.

Supplementary Table 3. deepDNAshape analysis data.
This table contains predicted DNA feature data for the 15 TFs using deepDNAshape.

Supplementary Table 4. Intrinsic bendability data.
This table contains predicted DNA intrinsic bendability data for the 15 TFs.

Supplementary Table 5. Universal PBM data.
This table contains universal protein binding microarray data for the 15 TFs.

Supplementary Table 6. Mutation profiles.
This table contains processed mutation profile data for the 15 TFs.

Supplementary Table 7. Buried Surface Area (BSA) analysis data.
This table contains calculated BSA data for the 15 TFs using ChimeraX.

Supplementary Table 8. Universal PBM data for ETS1 mutants.
This table contains universal protein binding microarray data for the four ETS1 mutants with replicates.

Supplementary Table 9. BLI data.
This table contains the raw and fitted data of BLI experiments for ETS1 and SOX2.

Supplementary Table 10. The composition of the molecular simulation box.
This table contains the number of molecules in each of the simulation box based on the topology file.

Supplementary Table 11. Structural parameters of TBP/nicked DNA complexes.
This table contains structural parameters generated by x3DNA for obtained high-resolution crystal structures of TBP/nicked DNA complexes.

Supplementary Table 12. ETS1-PARP1 competition assay data.
This file contains the raw data for the ETS1-PARP1 competition experiment.

Supplementary References

1 Brukner, I., Sanchez, R., Suck, D. & Pongor, S. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J* **14**, 1812-1818 (1995). https://doi.org/10.1002/j.1460-2075.1995.tb07169.x

2 Jiang, W. J. *et al.* Assessing base-resolution DNA mechanics on the genome scale. *Nucleic Acids Res* **51**, 9552-9566 (2023). https://doi.org/10.1093/nar/gkad720

3 Yella, V. R. *et al.* Flexibility and structure of flanking DNA impact transcription factor affinity for its core motif. *Nucleic Acids Res* **46**, 11883-11897 (2018). https://doi.org/10.1093/nar/gky1057

4 Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264-267 (2016). https://doi.org/10.1038/nature17661

5 Zhu, W. *et al.* DNA mutagenesis driven by transcription factor competition with mismatch repair. *Cell* **188**, 5735-5747 e5715 (2025). https://doi.org/10.1016/j.cell.2025.07.003

6 Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-260 (1997). https://doi.org/10.1038/38444

7 Ishida, H. & Kono, H. Torsional stress can regulate the unwrapping of two outer half superhelical turns of nucleosomal DNA. *Proc Natl Acad Sci U S A* **118** (2021). https://doi.org/10.1073/pnas.2020452118

8 Kaczmarczyk, A., Meng, H., Ordu, O., Noort, J. V. & Dekker, N. H. Chromatin fibers stabilize nucleosomes under torsional stress. *Nat Commun* **11**, 126 (2020). https://doi.org/10.1038/s41467-019-13891-y

9 Malaga Gadea, F. C. & Nikolova, E. N. Structural Plasticity of Pioneer Factor Sox2 and DNA Bendability Modulate Nucleosome Engagement and Sox2-Oct4 Synergism. *J Mol Biol* **435**, 167916 (2023). https://doi.org/10.1016/j.jmb.2022.167916

10 Reymer, A., Zakrzewska, K. & Lavery, R. Sequence-dependent response of DNA to torsional stress: a potential biological regulation mechanism. *Nucleic Acids Res* **46**, 1684-1694 (2018). https://doi.org/10.1093/nar/gkx1270

11 Liu, L. F. & Wang, J. C. Supercoiling of the DNA template during transcription. *Proc Natl Acad Sci U S A* **84**, 7024-7027 (1987). https://doi.org/10.1073/pnas.84.20.7024

12 Naughton, C. *et al.* Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat Struct Mol Biol* **20**, 387-395 (2013). https://doi.org/10.1038/nsmb.2509

13 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012). https://doi.org/10.1126/science.1225829

14 Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380-1389 (2013). https://doi.org/10.1016/j.cell.2013.08.021

15 Walton, R. T., Christie, K. A., Whittaker, M. N. & Kleinstiver, B. P. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* **368**, 290-296 (2020). https://doi.org/10.1126/science.aba8853

16 Nithun, R. V. *et al.* Deciphering the Role of the Ser-Phosphorylation Pattern on the DNA-Binding Activity of Max Transcription Factor Using Chemical Protein Synthesis.

Angew Chem Int Ed Engl **62**, e202310913 (2023). https://doi.org/10.1002/anie.202310913
17 Nithun, R. V. *et al.* Site-Specific Acetylation of the Transcription Factor Protein Max Modulates Its DNA Binding Activity. *ACS Cent Sci* **10**, 1295-1303 (2024). https://doi.org/10.1021/acscentsci.4c00686
18 Shen, N. *et al.* Divergence in DNA Specificity among Paralogous Transcription Factors Contributes to Their Differential In Vivo Binding. *Cell Syst* **6**, 470-483 e478 (2018). https://doi.org/10.1016/j.cels.2018.02.009
19 Stelling, A. L. *et al.* Infrared Spectroscopic Observation of a G-C(+) Hoogsteen Base Pair in the DNA:TATA-Box Binding Protein Complex Under Solution Conditions. *Angew Chem Int Ed Engl* **58**, 12010-12013 (2019). https://doi.org/10.1002/anie.201902693
20 Stephens, D. C. & Poon, G. M. Differential sensitivity to methylated DNA by ETS-family transcription factors is intrinsically encoded in their DNA-binding domains. *Nucleic Acids Res* **44**, 8671-8681 (2016). https://doi.org/10.1093/nar/gkw528
21 Takayama, Y., Sahu, D. & Iwahara, J. NMR studies of translocation of the Zif268 protein between its target DNA Sites. *Biochemistry* **49**, 7998-8005 (2010). https://doi.org/10.1021/bi100962h
22 Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720-1723 (2009). https://doi.org/10.1126/science.1162327
23 Berger, M. F. & Bulyk, M. L. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc* **4**, 393-411 (2009). https://doi.org/10.1038/nprot.2008.195
24 Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **24**, 1429-1435 (2006). https://doi.org/10.1038/nbt1246
25 Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **252**, 809-817 (1991). https://doi.org/10.1126/science.2028256
26 Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501 (2010). https://doi.org/10.1107/S0907444910007493
27 Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221 (2010). https://doi.org/10.1107/S0907444909052925
28 Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci* **27**, 293-315 (2018). https://doi.org/10.1002/pro.3330
29 Elrod-Erickson, M., Rould, M. A., Nekludova, L. & Pabo, C. O. Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure* **4**, 1171-1180 (1996). https://doi.org/10.1016/s0969-2126(96)00125-6
30 Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12-21 (2010). https://doi.org/10.1107/S0907444909042073
31 Kabsch, W. Xds. *Acta Crystallogr D Biol Crystallogr* **66**, 125-132 (2010). https://doi.org/10.1107/S0907444909047337
32 Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861-877 (2019). https://doi.org/10.1107/S2059798319011471

33 Lee, J. *et al.* CHARMM-GUI supports the Amber force fields. *J Chem Phys* **153**, 035103 (2020). https://doi.org/10.1063/5.0012280

34 Abraham, M. *et al.* GROMACS 2024.5 Source code. *Zenodo* (2025). https://doi.org/10.5281/zenodo.14732103

35 Tian, C. *et al.* ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J Chem Theory Comput* **16**, 528-552 (2020). https://doi.org/10.1021/acs.jctc.9b00591

36 Galindo-Murillo, R. *et al.* Assessing the Current State of Amber Force Field Modifications for DNA. *J Chem Theory Comput* **12**, 4114-4127 (2016). https://doi.org/10.1021/acs.jctc.6b00186

37 Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J Chem Phys* **126**, 014101 (2007). https://doi.org/10.1063/1.2408420

38 Bernetti, M. & Bussi, G. Pressure control using stochastic cell rescaling. *J Chem Phys* **153**, 114107 (2020). https://doi.org/10.1063/5.0020514

39 Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**, 1463-1472 (1997). https://doi.org/https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

40 Case, D. A. *et al.* AmberTools. *J Chem Inf Model* **63**, 6183-6191 (2023). https://doi.org/10.1021/acs.jcim.3c01153

41 Raguette, L. E. *et al.* phosaa14SB and phosaa19SB: Updated Amber Force Field Parameters for Phosphorylated Amino Acids. *J Chem Theory Comput* (2024). https://doi.org/10.1021/acs.jctc.4c00732

42 Lu, X. J., Bussemaker, H. J. & Olson, W. K. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res* **43**, e142 (2015). https://doi.org/10.1093/nar/gkv716

43 Blanchet, C., Pasi, M., Zakrzewska, K. & Lavery, R. CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res* **39**, W68-73 (2011). https://doi.org/10.1093/nar/gkr316

44 Afek, A. *et al.* DNA mismatches reveal conformational penalties in protein-DNA recognition. *Nature* **587**, 291-296 (2020). https://doi.org/10.1038/s41586-020-2843-2

45 Li, J., Chiu, T. P. & Rohs, R. Predicting DNA structure using a deep learning method. *Nat Commun* **15**, 1243 (2024). https://doi.org/10.1038/s41467-024-45191-5

46 Li, J. & Rohs, R. Deep DNAshape webserver: prediction and real-time visualization of DNA shape considering extended k-mers. *Nucleic Acids Res* **52**, W7-W12 (2024). https://doi.org/10.1093/nar/gkae433

47 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012). https://doi.org/10.1038/nature11247

48 Kagda, M. S. *et al.* Data navigation on the ENCODE portal. *Nat Commun* **16**, 9592 (2025). https://doi.org/10.1038/s41467-025-64343-9

49 Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**, D882-D889 (2020). https://doi.org/10.1093/nar/gkz1062

50 Wu, W. *et al.* Neuronal enhancers are hotspots for DNA single-strand break repair. *Nature* **593**, 440-444 (2021). https://doi.org/10.1038/s41586-021-03468-5

51 Ovek Baydar, D. *et al.* JASPAR 2026: expansion of transcription factor binding profiles and integration of deep learning models. *Nucleic Acids Res* **54**, D184-D193 (2026). https://doi.org/10.1093/nar/gkaf1209

52 Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019). https://doi.org/10.1038/s41598-019-45839-z