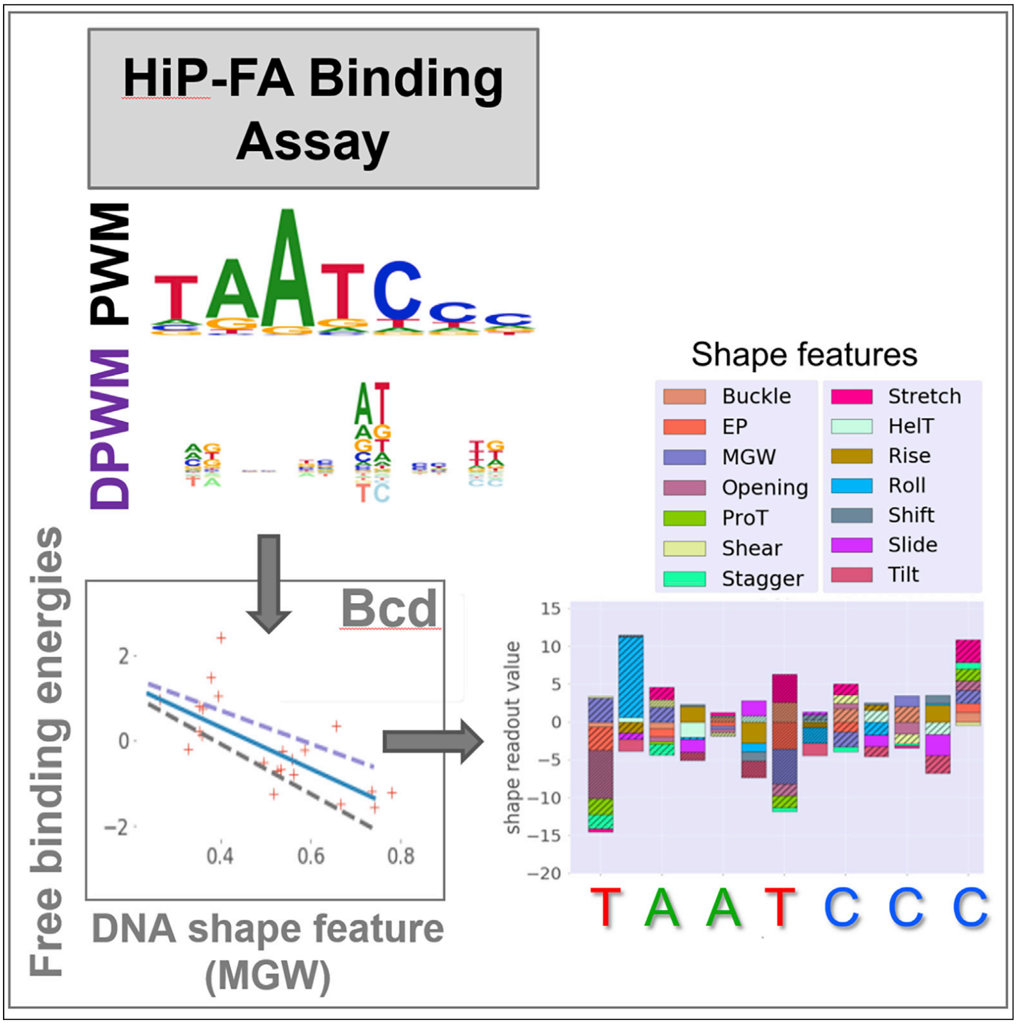
iScience

CellPress
OPEN ACCESS

Article

Transcription Factor Binding Affinities and DNA Shape Readout



Max Schnepf, Marc von Reutern, Claudia Ludwig, Christophe Jung, Ulrike Gaul

jung@genzentrum.lmu.de

HIGHLIGHTS

The DNA shape contribution to *Drosophila* TFs-DNA binding is directly characterized

Zeroth- and first-order TF-DNA binding specificities are measured with high accuracy

DNA shape readout is widely used by these TFs

A shape readout/structural correlation analysis provides biological insights

Schnepf et al., iScience *23*, 101694
November 20, 2020 © 2020 The Authors.
https://doi.org/10.1016/j.isci.2020.101694

iScience

CellPress
OPEN ACCESS

Article

Transcription Factor Binding Affinities and DNA Shape Readout

Max Schnepf,[1] Marc von Reutern,[1] Claudia Ludwig,[1] Christophe Jung,[1,3,*] and Ulrike Gaul[1,2]

SUMMARY

An essential event in gene regulation is the binding of a transcription factor (TF) to its target DNA. Models considering the interactions between the TF and the DNA geometry proved to be successful approaches to describe this binding event, while conserving data interpretability. However, a direct characterization of the DNA shape contribution to binding is still missing due to the lack of accurate and large-scale binding affinity data. Here, we use a binding assay we recently established to measure with high sensitivity the binding specificities of 13 *Drosophila* TFs, including dinucleotide dependencies to capture non-independent amino acid-base interactions. Correlating the binding affinities with all DNA shape features, we find that shape readout is widely used by these factors. A shape readout/TF-DNA complex structure analysis validates our approach while providing biological insights such as positively charged or highly polar amino acids often contact nucleotides that exhibit strong shape readout.

INTRODUCTION

The binding of transcription factors (TFs) to specific DNA sequences is a key event for the regulation of gene expression. The features defining a binding site have been the focus of several decades of research starting from simple consensus motif binding sites, later replaced by probabilistic models of TF binding assuming that each base contributes independently to the overall affinity, the so-called position-specific weight matrices (PWMs) (Stormo et al., 1982). With the advent of high-throughput methods, binding specificities became available for thousands of TFs and it has become clear that more complex models for binding sites using non-independent nucleotide interactions lead to more accurate predictions than PWMs (Weirauch et al., 2013; Zhao and Stormo, 2011). Nucleotide correlations can originate from amino acids that contact multiple bases simultaneously or from stacking interactions that determine binding through DNA shape readout. Hence, although determining binding specificities is crucial to predict binding sites in the genome, such data alone are not sufficient to fully describe TF-DNA binding interactions as they do not provide insights about the mechanism the TF employs to bind to different DNA sequences. To elucidate how the TF ''reads'' the DNA is of paramount importance not only to improve algorithms predicting binding sites but also to refine our fundamental understanding of how TFs are recruited to specific DNA regulatory sequences.

To date, two distinct modes of protein-DNA recognition are known: base readout, which reflects the interplay at nucleobase-amino acid contacts mainly driven by the formation of hydrogen bonds, and shape readout, dominated by van der Waals interactions and electrostatic potentials (EPs), that recognizes the 3D structure of the DNA double helix. As a consequence, one can assume that, if the TF uses the shape readout, models incorporating DNA structural information should improve prediction of TF-DNA binding specificities. To test this hypothesis and thereby help model development, it would thus be highly desirable to (1) determine accurately TF-DNA binding specificities, including non-independent nucleotide interactions since deviations from linear binding can carry information about the influence of DNA shape, and (2) use these data to assess the contribution of DNA shape readout to the binding interaction.

Despite the availability of techniques able to measure protein-DNA interactions at high throughput such as protein binding microarray (PBM) (Berger et al., 2006), SELEX-seq (Rastogi et al., 2018) (Riley et al., 2014), and SMiLE-seq (Isakova et al., 2017), the accurate measurement of binding affinities remains problematic. Moreover, these methods require a resin- or filter-based selection step that introduces bias and/or use

[1] Gene Center and Department of Biochemistry, Center for Protein Science Munich (CIPSM), Ludwig-Maximilians-Universität München, Feodor-Lynen-Strasse 25, 81377 München, Germany

[2] Deceased during the review of the manuscript

[3] Lead Contact

*Correspondence: jung@genzentrum.lmu.de

https://doi.org/10.1016/j.isci.2020.101694

stringent washing protocols resulting in the loss of weak binders, which can lead to erroneously over-specific binding specificities (Jung et al., 2018). These limitations are critical, especially to determine higher-order binding interactions, which are intrinsically weak (Maerkl and Quake, 2007; Nutiu et al., 2011).

Evaluating the contribution to binding of DNA shape readout also poses challenges. First, although it had been known for a long time from crystal structures that TFs read out the DNA shape (see (Rohs et al., 2010) for review), it is still not possible to determine experimentally the DNA shape features at a large scale for any given DNA sequence. However, this would be necessary to quantitatively assess DNA shape influence on TF-DNA binding. This issue has been tackled by Zhou et al. who introduced "DNAShape" (Zhou et al., 2013), an algorithm that predicts structural DNA features from nucleotide sequences, considering at each DNA position a local 5-mers nucleotide environment. The original set of four geometric shape features was later completed by Li et al. (Li et al., 2017), who made tables available to calculate an expanded repertoire of 13 DNA shape features in total. Finally, Chiu et al. (Chiu et al., 2017) added in a comparable fashion the EP, which approximates the minor-groove EPs. The EP reflects the mean charge density of the DNA backbone sensed by positively charged amino acid residues of the binding protein.

Another difficulty to analyze the influence of DNA shape to binding is that, in spite of all the advances made possible by "DNAShape" and the succeeding studies, it is still not clear to what degree shape readout can be described as a function of the underling DNA sequence. It is indeed very difficult to tease apart whether a binding protein favors a given nucleotide sequence because it recognizes certain amino acids of this sequence or rather certain shapes features of the DNA helix. An important step was made with homeodomain TFs by Abe et al. (Abe et al., 2015), who were able to specifically remove the ability of the binding proteins to read a certain structural feature of DNA and to switch between different modes of DNA shape readouts. Another approach computationally dissects TF binding specificity in terms of base and shape readout (Rube et al., 2018). Remarkably, the authors determined that 92-99% of the variance in the shape features can be explained with a model considering only dinucleotides dependencies. They also found that interactions were much stronger between neighboring nucleotides than for non-adjacent positions, indicating that these dinucleotide features are the most important for binding. Hence, determining neighboring dinucleotide dependencies should be enough to capture most on the higher-order binding interactions.

Unfortunately, although these studies shed new light on the role of DNA shape in TF-DNA recognition, they were limited to the analysis of only a few factors and used only four different shape features. This was due to the lack of quantitative data on higher-order binding specificities and to the lack of tables to calculate other shape features. Thus, a more comprehensive analysis of TF-DNA binding – especially including higher-order dependencies – is urgently needed to better understand TF-DNA binding in general and to what extent DNA shape features are recognized by TFs in particular.

Recently, we presented high-performance fluorescence anisotropy (HiP-FA) (Jung et al., 2018, Jung et al., 2019), a method that determines TF-DNA binding energies directly in solution with high sensitivity and at a large scale and allows for measuring the affinity of a TF to any given DNA sequence. These features make HiP-FA an ideal tool to measure TF-DNA binding specificities, in particular the higher-order dependencies since these interactions are generally weak and their accurate measurement is both difficult and indispensable.

Here, we used HiP-FA to measure binding energies for 13 TFs of the *Drosophila* segmentation gene network belonging to 8 different binding domain families. We determined their 0th order of binding specificities taking only into account independent base contributions (PWM) and their first order of binding specificities accounting for dinucleotide dependencies represented by the dinucleotide position weight matrices (DPWMs). In this work, we define DPWMs as being the scoring matrices characterizing the deviations in the dinucleotide binding energies compared to pure PWMs (Transparent Methods). Correlating our affinity data with the 13 known DNA shape features and the EP, we found that nearly all our factors extensively use shape readout for DNA recognition, independently of the binding domain family. For 11 TFs for which structural information is available, we examined the correlations between their nuclear magnetic resonance (NMR)/co-crystal structures or structures of analog proteins obtained by homology-based modeling and the shape attributes obtained from our analysis. Finally, we ran a cluster analysis to test if certain shape features tend to co-occur in the DNA shape readout used by our TFs.

RESULTS

Determination of the TF-DNA Binding Specificities and Overall Strategy for the Analysis

In a previous work (Jung et al., 2018), we have already presented the PWMs determined by HiP-FA for the 13 selected factors. We have also validated our method for determining binding affinities using two orthogonal assays, electrophoretic mobility shift assay and microscale thermophoresis. Finally, we have demonstrated that our PWMs were superior to those obtained with other methods (bacterial one hybrid or DNase footprinting) in predicting ChiP-seq data and when used in a thermodynamic model for predicting gene expression in *Drosophila* embryos (Jung et al., 2018). Herein, we extended our method to capture potential higher-order TF-DNA interactions by measuring the binding affinities of all mononucleotide and neighboring dinucleotide mutations (Figure 1A) in the core of each TF-DNA consensus binding sequence (6 positions for GATAe, 7 for all the other TFs). For 6 factors, we measured duplicates or triplicates to check reproducibility, leading to in total ~1600 individual titration curves. We analyzed the data in two steps: first, we used the binding affinities to determine the PWMs and the DPWMs. Importantly, in the analysis procedure, we developed an algorithm (*PySite*; https://github.com/Reutern/PySite) to correct for the energy contribution of off-target binding sites that might be created by chance in dinucleotide mutations (Figure 1B and Transparent Methods). Second, we assessed the influence of the shape of DNA around the core binding site on the TF-DNA binding strength. For all dinucleotide mutations, we calculated the 13 shape features and the EP at each position in the binding site using available look-up tables (Chiu et al., 2017; Li et al., 2017; Zhou et al., 2013). We then applied a robust linear regression algorithm (Transparent Methods) to correlate, at a given position, the values of each shape feature with the binding energies measured for all tested mutations of the binding site (Figure 1C; see below for details).

Zeroth- and First-Order Binding Specificities for the *Drosophila* TFs

We used our measured binding affinities to determine the PWMs and DPWMs (Data S1) of the factors (Figure 2 and Transparent Methods). Overall, the PWMs are similar and largely share the same consensus sequences with those obtained by other methods, but they generally present a lower specificity (measured by their information content [IC]) as already discussed in our previous work (Jung et al., 2018). In contrast, our DPWMs show fewer but more preferred dinucleotides (as indicated by higher mutual information [MI] (Transparent Methods), a metric similar to IC but for dinucleotide representation) compared to computationally derived scoring matrices including nucleotides (Siebert and Soding, 2016) or obtained using SMiLE-seq data (Rube et al., 2018). The low noise present in our binding specificities can be visually appreciated by comparing with the logos obtained from SMiLE-seq data for Bicoid (Bcd) (Rube et al., 2018) (Figure S2), to our knowledge the only factor among our TFs with an already known higher-order specificities based on binding affinities. For example, at position 5 of the HiP-FA Bcd DPWM (corresponding to the dinucleotide mutations between positions 4 and 5 in the PWM; Figure 2), the four pairs AT, AG, GT, and CA have a cumulated MI of nearly 1 bit, thereby dominating over the other 11 possible dinucleotide mutations. Another more direct way to assess the effect of non-independent binding is to compare our experimentally determined affinities with predicted values assuming purely linear base contribution (Figure S1). Many dinucleotide mutation sequences disagree with measured values (defined as lying within 3σ of the measured values), confirming the presence of non-independent amino acid-nucleotide interactions.

For all factors, we observe that the 0th order contribution to binding dominates over the first order, as indicated by the higher ICs of the specificity logos (6.9 bits on average for the 0th order compared to 2.1 bits MI for the first order; Figure 2). This was expected since the simple PWM model has proven to capture most of the sequence preferences for numerous TFs (Stormo et al., 1982; Zhao and Stormo, 2011). Surprisingly, the DPWMs of nearly all our TFs (with the exceptions of GATAe and Gt) show a high contribution to the overall binding specificities, revealed by their relatively high total MI (>~1 bit), above our threshold for significant MI (0.03 bits per nucleotide positions, corresponding to ~0.2 bits for the total MI of our DPWM logos; see Transparent Methods). Several studies already emphasized the importance of neighboring nucleotides in the prediction of TF binding (Nitta et al., 2015; Siebert and Soding, 2016; Zhao et al., 2012) but only for a few factors. The sensitivity of the HiP-FA assay enables us to accurately resolve weak – but measurable – binding events and their deviations from a purely linear binding of independent bases.

Noteworthy, the three members of the homeobox family (Bcd, Gsc, and Oc; Figure 2) have resembling PWMs and DPWMs, reflecting the similarity of their binding domains. This observation is in line with previous works describing the high similarity between homeodomain TFs' PWMs (Affolter et al., 2008). Our

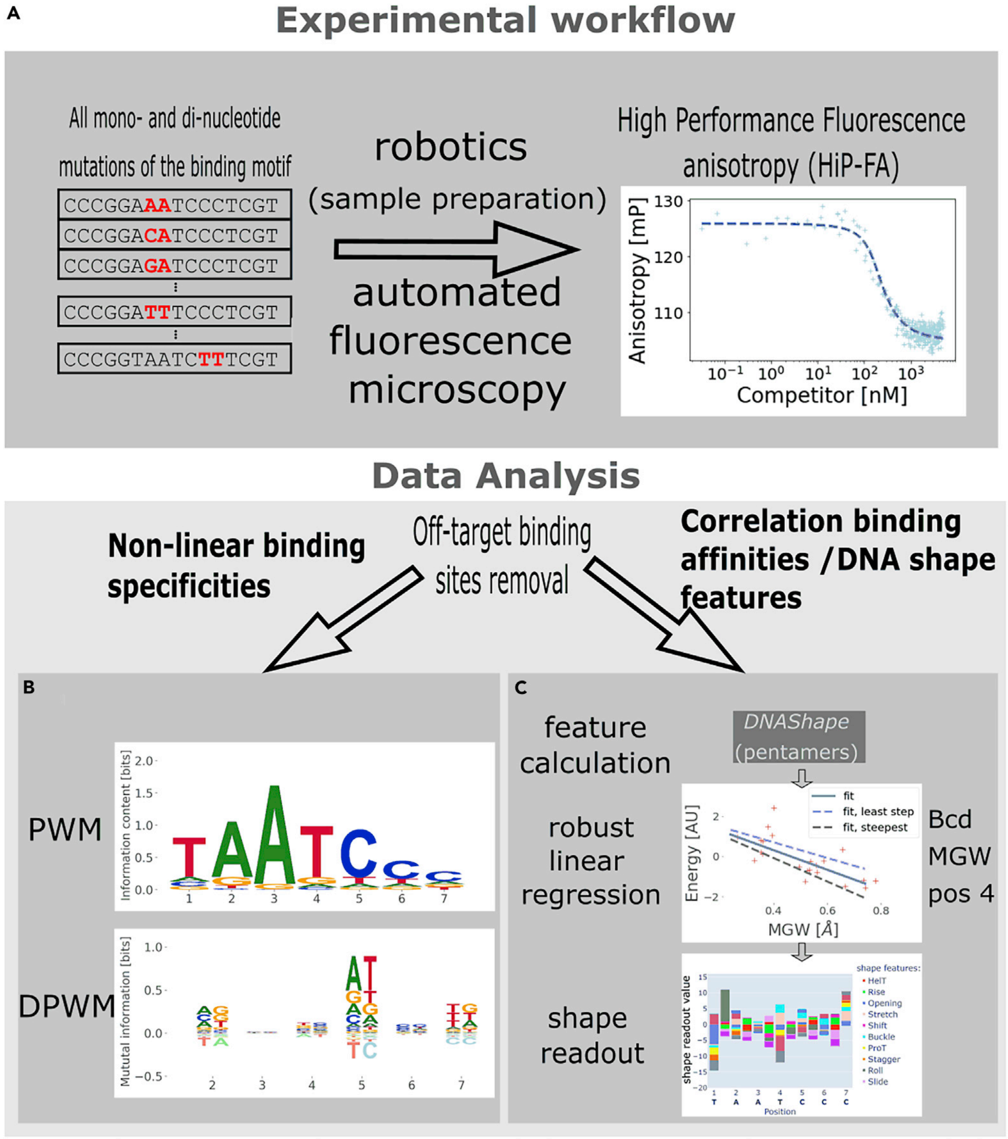


Figure 1. Experimental and Data Analysis Strategies

(A) Sequence design and measurement of binding energies by HiP-FA. A consensus sequence is mutated with all possible mononucleotide and dinucleotide mutations. The individual TF-DNA binding energies are measured using a robotic system and an automated custom-modified fluorescence microscope; the titration binding curves are reconstructed and analyzed following the HiP-FA procedure (Jung et al., 2018, Jung et al., 2019).

(B) Data analysis. After an off-target removal procedure (Transparent Methods), the binding energies are used to determine the 0th order of binding (PWMs) and the first order of binding (DPWMs), as shown for the TF Bcd. The DPWMs exhibits the mutual information (MI, a metric similar to the information content IC but for dinucleotide representation), which is not included in the simple linear PWM.

(C) Analysis of the DNA shape readout contribution. The sensitivity to DNA shape is analyzed following the subsequent steps: the DNA shape features are calculated using look-up tables (Chiu et al., 2017; Li et al., 2017; Zhou et al., 2013). The sensitivity to shape readout (termed shape readout value) is plotted per position against the binding energies (lower panel of c), and a robust linear regression is performed (Transparent Methods). Besides the fit (blue line), the steepest (gray dashed line) and the least steep fit (purple dashed line) are estimated using the confidence intervals provided by the robust linear regression. To make a conservative choice, the least steep slope is taken as the shape readout value. The shape readout values of all features and positions are depicted in the lower right panel for Bcd.

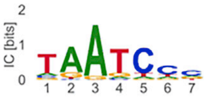
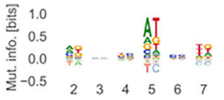
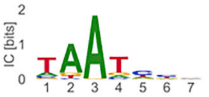
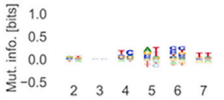
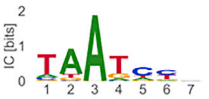
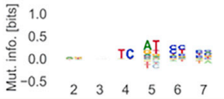
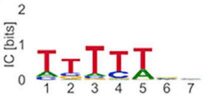
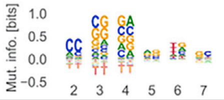
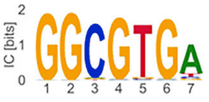
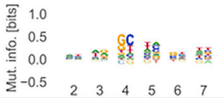
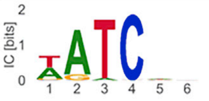
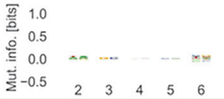
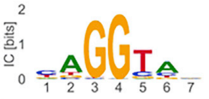
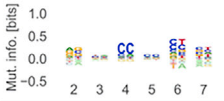
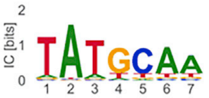
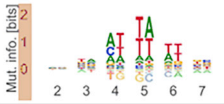
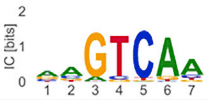
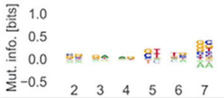
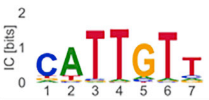
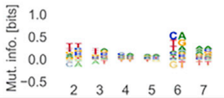
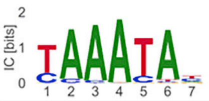
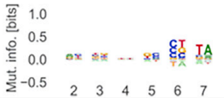
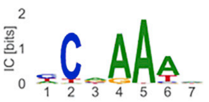
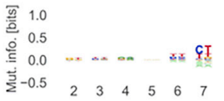
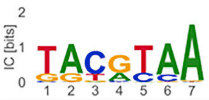
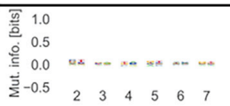
Transcription factor	PWM (0th order of binding interaction)	IC (bits)	DPWM (1st order of binding interaction)	MI (bits)
Bcd (Homedomain)		6.6		2.4
Gsc (homedomain		4.6		1.5
Oc (homedomain)		5.2		1.6
Hb (Zinc finger)		4.3		4.1
Hkb (Zinc finger)		11.8		2.1
GATAe (Zinc finger)		5.9		0.3
Zld (Zinc finger)		6.8		2.1
Nub (Pou domain)		7.8		6.5
Tll (NHR)		6.9		1.6
D (HMG Box)		8.4		2.5
Fkh (Winged helix)		8.4		1.6
Eip93f (Helix-turn-helix)		5.5		0.9
Gt (B-zip)		7.8		0.6

Figure 2. Overview of the PWMs and DPWMs for all the Investigated TFs

In the DPWMs, the heights of the dinucleotide letters represent the mutual information (MI) between two positions for the first order of binding. The total information content (IC) and MI are indicated in the right hand side columns for the PWM and DPWMs, respectively. Homeodomain factors and zinc fingers are grouped by color. Average PWMs and DPWMs are shown when replicate measurements were performed.

DPWMs clearly show that this similarity holds also at the first order of binding. A closer inspection, however, reveals the presence of subtle differences in specificities. At position 5 of the three DPWMs, although the preferred dinucleotides are very resembling (AT being the strongest positive deviation from linear binding, TC being the strongest repulsive one), the corresponding MIs differ substantially between the three TFs (for the positive MI: 0.76 bits for Bcd, 0.42 for Oc, and 0.27 for Gsc). In addition, Bcd differs at position 2 in its DPWM from the two other factors with its relatively high MI (0.35 compared with 0.08 for Gsc and 0.04 for Oc). Although these differences are small, their concerted effect might be important to allow these homeodomains to execute their distinct biological functions.

To conclude, our sensitive measurements of binding affinities provide us with refined binding specificities for our TFs, including first-order binding interactions.

DNA Shape Readout Is Used by Most of the Investigated TFs

The fact that most of the variance in DNA shape is encoded in dinucleotides (Rube et al., 2018) encouraged us to tackle the question to which extent TF-DNA binding is driven by DNA shape. To this end, we calculated the 13 geometric shape features and the EP at each position for all tested DNA sequences and determined their influence on our binding energies. For a given factor, we evaluate whether the change in binding energies correlates with a feature of interest when a base at a certain position and/or at a neighboring position deviates from the consensus sequence. For example, in the case of Bcd, at position 4, the binding energy decreases over an amplitude of ~4 (normalized) when the relative minor groove width (MGW) increases from ~0.2 to ~0.8 Å (Figure 1C). The sensitivity to DNA shape readout is determined by a robust linear fitting procedure (Transparent Methods) to minimize the effect of extreme values (identified as outliers by the algorithm) and to provide a confidence interval to the resulting fitting parameters. The slope of the robust linear fitting provides an estimate of how much the binding of the TF at the particular position is influenced by the local DNA shape. On the following, we define the ''shape readout value'' as being this slope after normalization using for the binding energies their z/standard score and for the DNA shape features their amplitude (details about the normalization procedures in Transparent Methods). The shape readout value profiles allow one to compare the shape sensitivity of the free binding energy for the different shape features along the TF-DNA interface, while providing an intuitive metric of their deviation from their ''average'' behavior. For each TF, we applied this analysis procedure for every shape feature and at all base positions along the core DNA binding sequence (Figure 3 and Data S1). The reproducibility of the shape readout values among replicates was high with a mean squared Pearson coefficient $R^2 = 0.78$ for the 6 factors having duplicates or triplicates (Figure S3).

The shape sensitivity plots reveal a widespread use of DNA shape readout for all our TFs (Figure 3), with strong differences in the shape readout values between factors and at different base positions for a given factor. Remarkably, the members of the homeodomain family (blue box in Figure 3) behave the same with similar shape sensitivity plots (discussed in details below), as already observed for the PWMs and DPWMs. This does not hold true for the zinc finger family (green box) or for the other factors with different binding domains for which the shape sensitivity plots exhibit various patterns along the DNA binding sequences. Other studies have reported that zinc fingers are diverse in their binding behavior, in contrast to other TF families (Kribelbauer et al., 2019; Rohs et al., 2010). Interestingly, we found that in the center of the binding sites of GATAe and Zelda (Zld) (positions 3 and 4 for both factors in the shape sensitivity plots), the shape readout values are very low, as discussed below in more details for GATAe. At these positions, the sequence logos have a high IC, as indicated by the prominent TC and GG bases in the PWMs of GATAe and Zld (Figure 2). Conversely, shape features become important where sequence information is not well defined, like for GATAe at positions 5 and 6 and for Zld at positions 1 and 7. This phenomenon has already been reported for other factors (Abe et al., 2015). Interestingly, we observe a similar phenomenon for the side chains of most of our factors, for the three homeodomains and for Hb, Tll, Fkh, and Eip93f. In these cases, shape features contain more information for binding than sequence alone in the side chains.

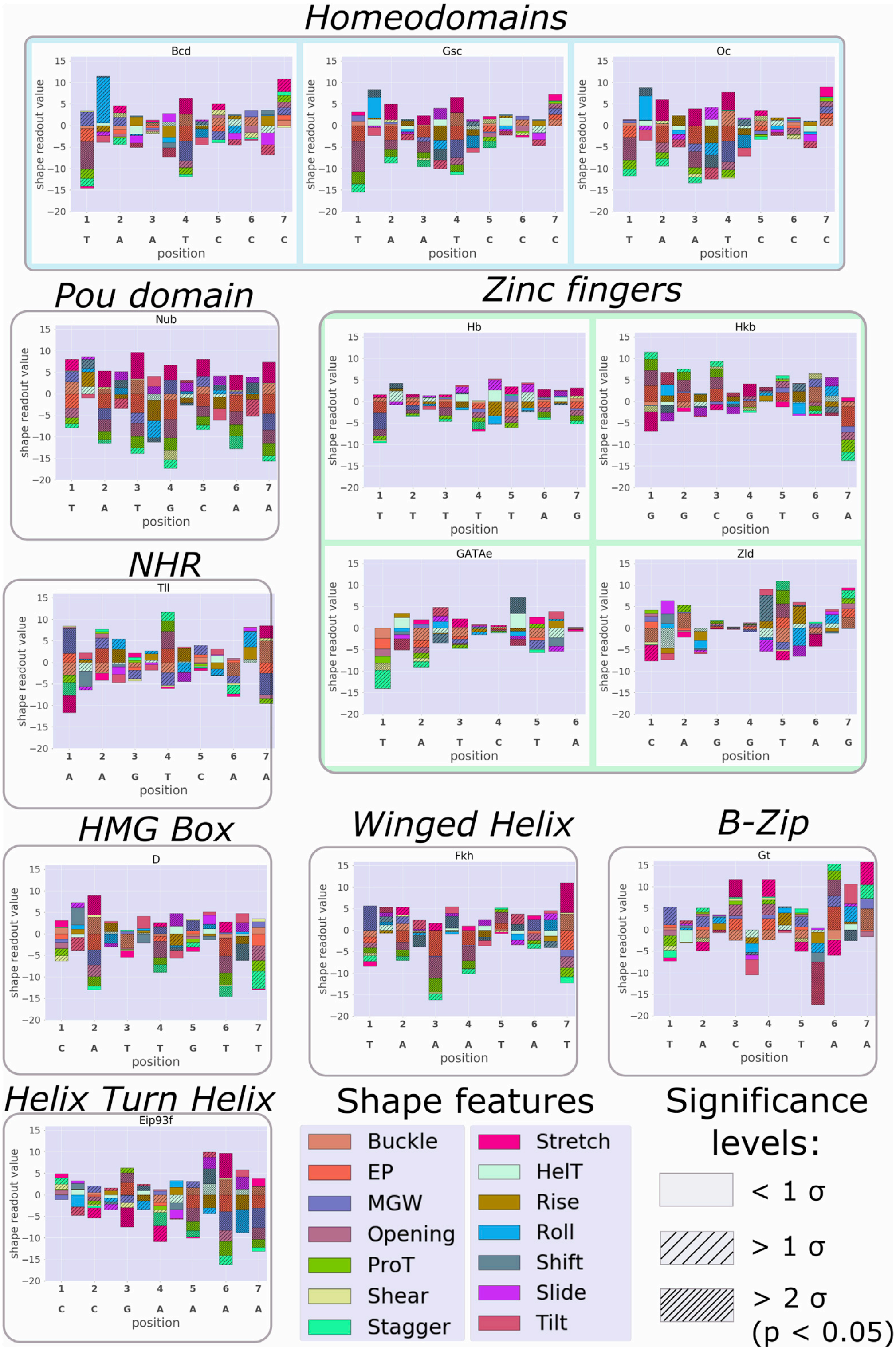


Figure 3. Overview of DNA Shape Sensitivities for the investigated TFs

The stacked shape readout values are plotted for each feature at each position (intra-base pair features) or between two positions (inter-base pair features). To facilitate the comparison with Figure 2, the positions are also labeled with their respective nucleobase at this position of the consensus sequence. The legend for the respective features is in the lower

Figure 3. *Continued*

right corner. Homeodomain TFs (blue background color) and zinc finger TFs (green) are grouped together. The significance levels are indicated for each shape readout value bar with a hashing code indicated in the right bottom corner (see Transparent Methods for details). Average shape sensitivity plots are shown when replicate measurements were performed. Overall, there is a widespread use of the DNA shape readout by our TFs.

Correlation between the DNA Shape Sensitivity of the TFs and Structural Information

We next wondered whether the observed shape readout values can be related to protein structures as interactions between the TF and its target DNA (Figure 4). Unfortunately, structural information is only available for the homeodomain TF Bcd, which has an NMR structure (Baird-Titus et al., 2006), and for the other homeodomains Gcs and Oc sharing very similar protein structure and binding specificities. For the other factors, we thus sought for experimental structure of homologous TFs using protein homology-based modeling (Kelley et al., 2015) (Figure S4). In this section, we will focus on the homeodomains, on a B-ZIP Gt homolog (Pap1) (Fujii et al., 2000), and on a zinc finger GATAe mouse homolog (GATA3) (Bates et al., 2008).

As for other homeodomain proteins in complex with their DNA targets, the recognition helix of Bcd (in red in Figure 4C) is thought to be engaged in base readout of the major groove, whereas the N-terminal tail (in blue) is involved in shape readout of the minor groove (Baird-Titus et al., 2006; Dror et al., 2014; Yang et al., 2017). Although little is known about the relationship between structural features and binding affinities, it has been shown that a narrow MGW can enhance the negative EP in the minor groove (Chang et al., 2013; Rohs et al., 2009), which can attract positively charged amino acids such as arginine (R), lysine (K), and histidine (H), considering the latter can be protonated. We thus focused on nucleotide positions with strong shape readout values (highlighted with blue rectangles in Figure 4A), in particular with significant shape readout values for the features MGW and EP, and sought the presence of contacts with R, K, and H residues at these positions. The residue contact map of Bcd (Figure 4B) shows the individual nucleotide-residue interactions, DNA secondary structure, protein secondary structure, and DNA interaction moieties (Figure S4 for details). One can observe multiple interactions (highlighted in yellow) between arginine (R4, R55, and R56) or lysine (K51 and K55) amino acids, and nucleotides positions with strong shape readout values (positions 1, 2, and 4, highlighted in blue in Figure 4A) exhibiting significant MWG and EP shape readout values. Another study (Dror et al., 2014) that analyzed 168 mouse homeodomains using PBM data found a significantly high correlation between the positively charged R or K residues of the N-terminal tail with the minor groove. Our data confirm the interaction between the R4 residue of the N-terminal tail (indicated by the blue arrow in Figure 4C) and the T and A nucleotides at position 1 and 2, respectively, exhibiting high and significant MWG shape readout value and EP (interactions shown in yellow). In addition, we observe another nucleotide T (position 4, highlighted with the right blue rectangle) with strong shape readout and that contacts an arginine (R55) and a lysine residue (K51), both belonging to the recognition α-helix. Interestingly, the A at position 3 (black arrows in Figure 4A) with a very low shape readout interacts only with the non-charged asparagine and the hydrophobic isoleucine residues (N52 and I48, respectively; black arrows in Figure 4B). This prompted us to evaluate the frequency of contacts with positive residues for all our factors for which we found a structure of a homologous protein (Figure S4). Remarkably, we found for the strong shape readout positions much more frequent nucleotide interactions with R, K, or H (64% of the contacts) than for the other positions (28% of the contacts). Hence, these results generalize the contribution of the positively charged amino acids to the shape readout to other DNA secondary structures (such as α-helixes) and to other binding domains. Interestingly, we also found for the POU domain Nub (Figure S4E) that positively charged residues can strongly contribute not only to DNA shape readout but also to non-charged but highly polar amino acids (glutamine, threonine, and serine in this case).

As an additional validation for our analysis, although the DNA shape features ProT, Roll, HelT, and MGW have been quantitatively investigated for Bcd by Rube et al. (Rube et al., 2018), the MGW was the only shape feature with significant shape sensitivity coefficients in their study. For a detailed comparison, we plotted our shape readout values for all positions of the MGW against the corresponding shape sensitivity coefficients determined by Rube et al. (Rube et al., 2018) (Figure S5). We obtained an excellent correlation ($R^2 = 0.99$) for the subset of coefficients that Rube et al. found to be significant. Note that our shape readout values were also significant ($p < 0.05$) at these positions. Remarkably, we also found significant correlations for additional features and for the other homeodomains, like at position 4 where the MGW exhibits a local minimum (Figures 4D and 4E) for Stretch and the EP for the three homedomain proteins, as well as for ProT

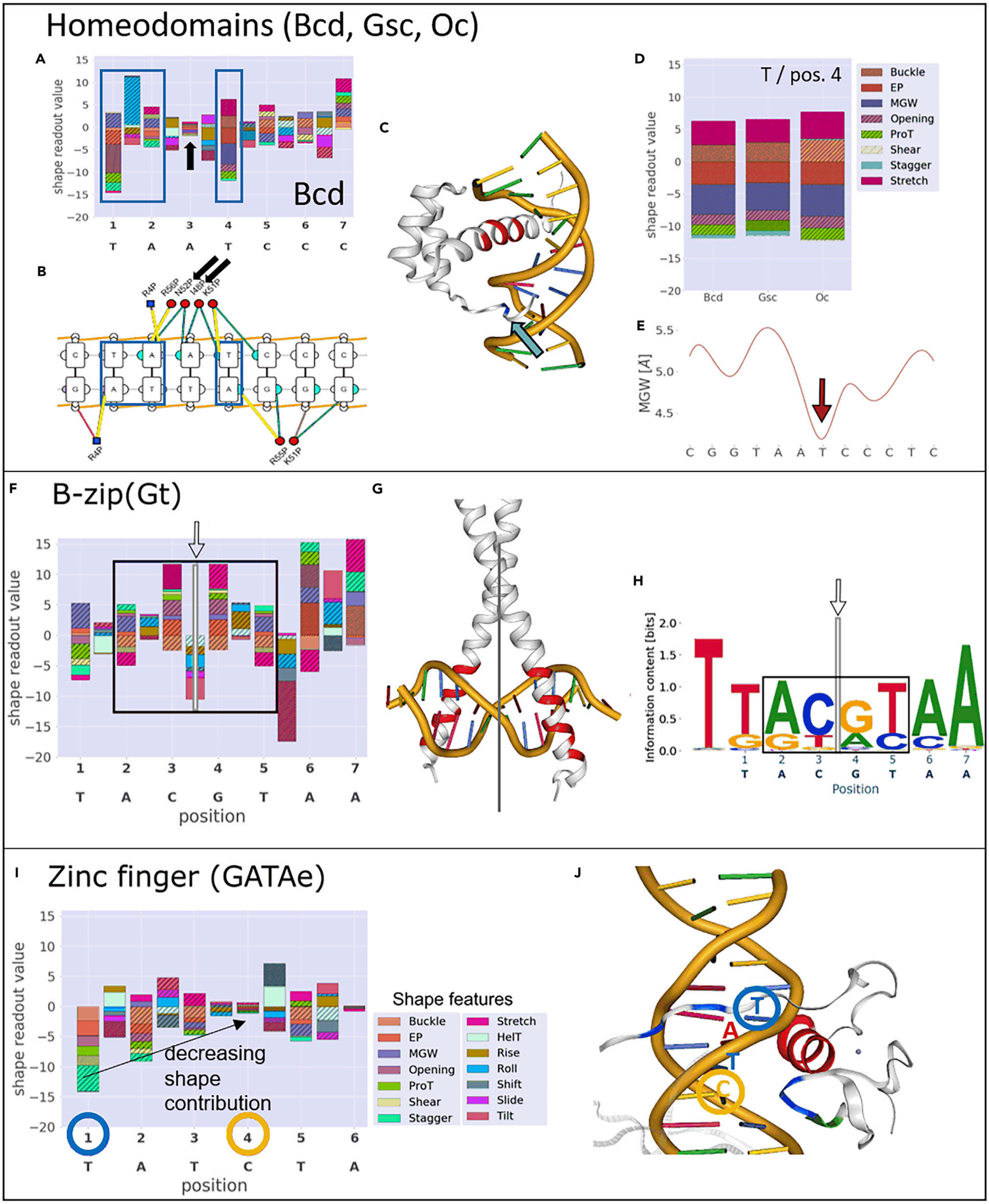


Figure 4. Correlation between DNA Shape Readout and Structural Information

Homeodomains TFs.

(A) Shape readout value profile for Bcd. Positions with strong shape readout highlighted with blue rectangles.

Figure 4. *Continued*

(B) Residue contacts map for Bcd (obtained using the DNAproDB database (Sagendorf et al., 2019), details in Figure S4). Interaction between bases and positively charged residues highlighted in yellow.

(C) Crystal structure of Bcd (pdb-ID: 1ZQ3) (Baird-Titus et al., 2006). Base contacts with the recognition helix in red, with the N-terminal tail in blue. The bluearrow points at the position where the binding domain contacts the narrowing minor groove.

(D) Shape readout values for the three homeodomain TFs at position TAAT of the consensus sequence (position 4 of the corresponding PWMs in Figure 2). In addition to being very similar, all three homeodomains show a strong readout of the minor groove at this position.

(E) MGW profile along the binding sequence for the consensus binding sequence used for the homeodomains. It exhibits a minimum value at position TAAT (red arrow).

B-ZIP TF Gt.

(F) Shape readout values for Gt. The black rectangle indicates positions with highly symmetrical shape readout values around the middle vertical axis (added to all three panels at the same position).

(G) Crystal structure of a similar B-ZIP TF (pdb-ID: 1GD2) (Fujii et al., 2000) with the same core consensus sequence as Gt. The black box indicates the region of high mirror symmetry around the black axis. Base contacts highlighted in red.

(H) Gt's PWM, the first position augmented with data from Jung et al. (Jung et al., 2018). The entire PWM is highly symmetrical.

Zinc finger TF GATAe.

(I) Shape readout values for GATAe. Positions with strong (1, blue) and weak (4, orange) shape readout values are indicated at the x axis.

(J) Their corresponding positions in the protein structure of a similar GATA TF (pdb-ID: 3DFV) (Bates et al., 2008) at both sides. The perspective shows a position with pronounced contacts to the DNA's phosphate backbone and minor groove. Base contacts in red. All crystal structures were produced using the DNAproDB portal (Sagendorf et al., 2019).

(Gsc) and Buckle (Bcd and Gsc). This shows that the TF reads multiple DNA shape features at this position. As mentioned above, the reproducibility of the shape feature values is remarkable among the different homeodomains (Figure 4D). Given the sequence similarity between the three proteins, it is not surprising to find a similar shape readout for most features, which also speaks for the high reproducibility of our measurements.

Another pertinent example is the TF Giant (Gt) belonging to the family of B-ZIP proteins (Figures 4F–4H and S4F). Members of this family approach the DNA like a scissor, with two alpha helices contacting the major groove from two opposing sites (Figure 4G). Interestingly, the same mirror symmetry with a mirror plane between position 3 and 4 (C and G) is found in the PWM (Figure 4H) and partially in the shape sensitivity plot (Figure 4F). The shape readout values of both inter- and intra-base pair features between positions 2 and 5 show a highly symmetrical pattern, in line with the binding mode of B-ZIP proteins (Figure 4H and the residue contact map in Figure S4F). This pattern, although conserved in the PWM, is not maintained at the side positions in the shape readout values, probably due to the fact that the DNA has more flexibility outside of the B-ZIP's scissor and the TF has less contact to its minor groove and backbone (Figure S4F).

At last, we examined the zinc finger protein GATAe (Figures 4I–4J and S4D). Zinc fingers contact the DNA at two opposing strands with three contacts being at one strand (positions 4 to 2, ATC in the case of GATAe) and another at the opposing strand (position 1, T) (Fedotova et al., 2017). There are multiple contacts at position 1 (blue circles in Figure 4J and the residue contacts map in Figure S4D) between the TF and the DNA backbone, which match the high shape readout values at this position (in total 14.1 AU, absolute sum). The contacts between the TF and minor groove or DNA backbone decrease moving toward the central binding site, as seen in the crystal structure (orange circle in Figures 4J and S4D). The sum of the shape readout values shows a similar behavior, a decreasing overall shape sensitivity going from position 1 to 4 (black arrow in Figure 4I). At position 4 (yellow circle), one can observe contacts in the structure exclusively to bases in the major groove (yellow circle in Figure 4J), and the sum of the shape readout values is reduced to a minimum (1.8 AU). It was recently reported that metazoan zinc fingers tend to establish several contacts to the DNA backbone (Najafabadi et al., 2017), possibly permitting DNA shape readout at these positions.

In overall, this shape/structure analysis validates our approach, while bringing biological insights about the relationship between shape readout values and residue contacts with the TF-DNA interface.

Shape Readout Values/TF Clustering

Finally, we asked if TFs use predominantly certain shape features to bind to DNA. To test whether shape features tend to co-occur in the shape readout, we performed two distinct cluster analysis of the shape readout values matrix (Figure S6 and Transparent Methods): (1) we clustered the different features with respect to their feature readout by the TF (vertical lines) and (2) its reverse – a clustering of the TFs versus their shape readout for each feature as a matrix including all nucleotide positions (horizontal lines).

The TF clustering indicates that the different binding proteins show little similarity in their use of the shape features except for the homeodomains, which was expected. In contrast, the clustering of the shape features reflects structural dependencies between shape features. There are three distinct clusters of shape features, which may be related to biophysical properties of the DNA and its interplay with the binding protein, such as bends or kinks (details in caption of Figure S6). For instance, the first cluster consists of slide, helix twist, roll, and MGW. These features were reported to correlate the most with each other both in bound and unbound DNA (El Hassan and Calladine, 1995) (Stella et al., 2010) and are read out concertedly in DNA-protein complexes (Suzuki et al., 1997). Thus, the cluster analysis confirms properties of the features' interdependencies in shape readout.

DISCUSSION

HiP-FA constitutes a powerful tool to quantify TF-DNA binding specificity, especially the non-independent interactions requiring to be determined with high accuracy. The throughput of the method is not sufficient to discover *de novo* shape motifs or to explore the large sequence space possible with sequencing-based methods like HT-SELEX or SMiLE-seq. However, this is not a major limitation since the *prior* knowledge that HiP-FA requires (some information about the TF's binding preferences) is known for many TFs, and dinucleotide mutations are sufficient to cover most of the non-independent amino acid-nucleotide interactions. It would also be straightforward to extend the measurements in the flanking regions of the core binding motif. A comparison of the different approaches and a summary of the results obtained as far as DNA shape readout analysis is concerned can be found in Table S1.

Our approach consisting in measuring binding energies of a complete set of dinucleotide mutations is more direct than the one used by Rube et al. (Rube et al., 2018) that requires a prior analysis with the "No Read Left Behind" (Rastogi et al., 2018) algorithm to derive affinities from high-throughput data. In addition, our downstream analysis of shape sensitivity, which employs a robust linear regression algorithm, uses fewer parameters and provides directly an interpretable characterization of shape sensitivity. However, it is not possible to distinguish between base and shape readout directly. In the analysis procedure, we cannot exclude the possibility of energy changes due to base readout, leading to an incidental correlation between binding energies and shape features. We reason that this apparent contribution of shape features may average out in the linear fitting procedure and, as a consequence, will not lead to a strong correlation bias in the robust linear regression. This assumption is supported by the following: (1) our estimation of the shape sensitivity is in excellent agreement with the one obtained by the more complex algorithm elaborated by Rube et al. (Figures S2 and S5), (2) our shape readout values can be related to structural features of the factors (Figures 4 and S4), (3) the clustering of the shape feature readout rediscovers already known interdependencies between shape features (Figure S6). Not all shape features are recognized independently by the TFs. The different groups in our clustering might represent a specific DNA conformation which is read out by a TF, rather than the readout of several independent DNA features. These conformations might play an important role for the binding behavior of several TFs.

By combining directly TF-DNA binding affinities, DNA shape features, and structural information, we gained insights into their correlation, a debated topic due to their intrinsic covariation. Importantly, our results suggest that DNA shape readout is widespread among our TFs. The extended use of DNA shape readout by TFs has become increasingly apparent over the past years (Chiu et al., 2017; Mathelier et al., 2016; Pal et al., 2019; Rube et al., 2018; Samee et al., 2019; Yang et al., 2017; Zhou et al., 2015), which comes as no surprise considering that the number of van der Waals interactions enabling shape readout account for two-third of the protein-DNA interactions (Rube et al., 2018). The correlation analysis of the shape readout values with protein-DNA complex structures allows us to generalize the influence of the charged amino acids on the shape readout that has been described so far only for homeodomains in the minor groove region of the DNA. We observe this effect to other DNA secondary structures (such as α-helixes) and to other binding domains. In addition, for the POU domain Nub we identify non-charged but polar residues that can also lead to a strong DNA shape readout. To the best of our knowledge, these effects on DNA shape readout have not been reported. The difficulty to detect the effects of charged and non-charged residues, especially in the major groove, is that they are obscured by the interactions involved in the base readout. Our analysis was able to resolve even subtle effects due to the high sensitivity of the binding affinity measurements, and our shape analysis was able to deconvolve, to some extent, shape from base readout.

In summary, we determined the binding specificities for 13 *Drosophila* TFs including first-order dependencies, provided insights into the correlation between their binding affinities to DNA and the shape features of the DNA helix, and gave structural insights in the shape readout. Our method could easily be extended to more factors and to different organisms to provide a refined catalog of TF-DNA shape readout landscapes.

LIMITATIONS OF THE STUDY

Although our HiP-FA assay allows us to determine accurately binding affinities at a relatively large scale, we cannot cover the whole sequence space as high-throughput methods do. To restrict the number of measurements, we thus focussed on the core binding motif of the TFs, and to all mononucleotide and dinucleotides mutations of the consensus sequence rather that all possible mutations. This should however cover most of the TF-DNA interactions since it has been shown that dinucleotide models explain >92% of the variance for the MGW, ProT, Roll, and HelT shape features (Rube et al., 2018). In addition, our analysis based on the direct correlation between binding affinities and shape features can only indirectly and partially tease apart the respective contributions of base and DNA shape readouts. Note that how to achieve the deconvolution between base and shape readouts is a longstanding issue in the field.

Resource Availability

Lead Contact

Dr. Christophe Jung

phone. +49 (0)89 2180 71101

fax. +49 (0)89 2180 71105

email: jung@genzentrum.lmu.de.

Materials Availability

All plasmids generated in this study for the expression of the TF’s binding domains are available from the Lead Contact without restriction. This study did not generate new unique reagents.

Data and Code Availibility

The PWMs, DPWMs, and shape readout values for all factors and shape features are available as Data S1. The *Python3* code of *PySite* is available on Github (https://github.com/Reutern/PySite). All the other *Python3* codes used for data analysis are available upon request.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101694.

ACKNOWLEDGMENTS

We dedicate this publication to the memory of Prof. Ulrike Gaul who deceased after a long illness during the review of the manuscript. This work was supported by the DFG (large equipment grant for automated system), the Sonderforschungsbereich SFB646, the Center for Integrated Protein Science Munich (CIPSM), and the Graduate School of Quantitative Biosciences Munich (QBM). U.G. acknowledges support by the Humboldt Foundation (Alexander von Humboldt-Professorship). The authors want to thank Peter Bandilla for his help with the robotics and Stefano Ceolin for reading the manuscript and for fruitful discussions about the experiments and the analysis. We also particularly thank Ulrich Unnerstall for the numerous and useful discussions about the project.

AUTHOR CONTRIBUTIONS

M.S., C.J., and U.G. developed the project; M.S. and C.J. designed the experiments; M.S. and C.L. performed the experiments; M.S., M.R., and C.J. performed data analysis; and M.S. and C.J. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no conflict of interests.

Received: May 29, 2020
Revised: September 30, 2020
Accepted: October 13, 2020
Published: November 20, 2020

REFERENCES

Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H.J., Rohs, R., and Mann, R.S. (2015). Deconvolving the recognition of DNA shape from sequence. Cell *161*, 307–318.

Affolter, M., Slattery, M., and Mann, R.S. (2008). A lexicon for homeodomain-DNA recognition. Cell *133*, 1133–1135.

Baird-Titus, J.M., Clark-Baldwin, K., Dave, V., Caperelli, C.A., Ma, J., and Rance, M. (2006). The solution structure of the native K50 Bicoid homeodomain bound to the consensus TAATCC DNA-binding site. J. Mol. Biol. *356*, 1137–1151.

Bates, D.L., Chen, Y., Kim, G., Guo, L., and Chen, L. (2008). Crystal structures of multiple GATA zinc fingers bound to DNA reveal new insights into DNA recognition and self-association by GATA. J. Mol. Biol. *381*, 1292–1306.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., Iii, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotech. *24*, 1429–1435.

Chang, Y.P., Xu, M., Machado, A.C., Yu, X.J., Rohs, R., and Chen, X.S. (2013). Mechanism of origin DNA recognition and assembly of an initiator-helicase complex by SV40 large tumor antigen. Cell Rep. *3*, 1117–1127.

Chiu, T.P., Rao, S., Mann, R.S., Honig, B., and Rohs, R. (2017). Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein-DNA binding. Nucleic Acids Res. *45*, 12565–12576.

Dror, I., Zhou, T., Mandel-Gutfreund, Y., and Rohs, R. (2014). Covariation between homeodomain transcription factors and the shape of their DNA binding sites. Nucleic Acids Res. *42*, 430–441.

El Hassan, M.A., and Calladine, C.R. (1995). The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme. J. Mol. Biol. *251*, 648–664.

Fedotova, A.A., Bonchuk, A.N., Mogila, V.A., and Georgiev, P.G. (2017). C2H2 zinc finger proteins: the largest but poorly explored family of higher eukaryotic transcription factors. Acta Nat. *9*, 47–58.

Fujii, Y., Shimizu, T., Toda, T., Yanagida, M., and Hakoshima, T. (2000). Structural basis for the diversity of DNA recognition by bZIP transcription factors. Nat. Struct. Biol. *7*, 889–893.

Isakova, A., Groux, R., Imbeault, M., Rainer, P., Alpern, D., Dainese, R., Ambrosini, G., Trono, D., Bucher, P., and Deplancke, B. (2017). SMiLE-seq identifies binding motifs of single and dimeric transcription factors. Nat. Methods *14*, 316–322.

Jung, C., Bandilla, P., von Reutern, M., Schnepf, M., Rieder, S., Unnerstall, U., and Gaul, U. (2018). True equilibrium measurement of transcription factor-DNA binding affinities using automated polarization microscopy. Nat. Commun. *9*, 1605.

Jung, C., Schnepf, M., Bandilla, P., Unnerstall, U., and Gaul, U. (2019). High sensitivity measurement of transcription factor-DNA binding affinities by competitive titration using fluorescence microscopy. JoVE, e58763, https://doi.org/10.3791/58763.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. *10*, 845–858.

Kribelbauer, J.F., Rastogi, C., Bussemaker, H.J., and Mann, R.S. (2019). Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. Annu. Rev. Cell Dev. Biol. *35*, 357–379.

Li, J., Sagendorf, J.M., Chiu, T.P., Pasi, M., Perez, A., and Rohs, R. (2017). Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. Nucleic Acids Res. *45*, 12877–12887.

Maerkl, S.J., and Quake, S.R. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. Science *315*, 233–237.

Mathelier, A., Xin, B., Chiu, T.-P., Yang, L., Rohs, R., and Wasserman, W.W. (2016). DNA shape features improve transcription factor binding site predictions in vivo. Cell Syst. *3*, 278–286.e4.

Najafabadi, H.S., Garton, M., Weirauch, M.T., Mnaimneh, S., Yang, A., Kim, P.M., and Hughes, T.R. (2017). Non-base-contacting residues enable kaleidoscopic evolution of metazoan C2H2 zinc finger DNA binding. Genome Biol. *18*, 167.

Nitta, K.R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E.E., et al. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. Elife *4*, e04837.

Nutiu, R., Friedman, R.C., Luo, S., Khrebtukova, I., Silva, D., Li, R., Zhang, L., Schroth, G.P., and Burge, C.B. (2011). Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat. Biotechnol. *29*, 659–664.

Pal, S., Hoinka, J., and Przytycka, T.M. (2019). Co-SELECT reveals sequence non-specific contribution of DNA shape to transcription factor binding in vitro. Nucleic Acids Res. *47*, 6632–6641.

Rastogi, C., Rube, H.T., Kribelbauer, J.F., Crocker, J., Loker, R.E., Martini, G.D., Laptenko, O., Freed-Pastor, W.A., Prives, C., Stern, D.L., et al. (2018). Accurate and sensitive quantification of protein-DNA binding affinity. Proc. Natl. Acad. Sci. U S A *115*, E3692–E3701.

Riley, T.R., Slattery, M., Abe, N., Rastogi, C., Liu, D., Mann, R.S., and Bussemaker, H.J. (2014). SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. Methods Mol. Biol. (Clifton, NJ) *1196*, 255–278.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. Annu. Rev. Biochem. *79*, 233–269.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248–1253.

Rube, H.T., Rastogi, C., Kribelbauer, J.F., and Bussemaker, H.J. (2018). A unified approach for quantifying and interpreting DNA shape readout by transcription factors. Mol. Syst. Biol. *14*, e7902.

Sagendorf, J., Markarian, N., Berman, H., and Rohs, R. (2019). DNAproDB: an expanded database and web-based tool for structural analysis of DNA-protein complexes. Nucleic Acids Res. *48*, D277–D287.

Samee, M.A.H., Bruneau, B.G., and Pollard, K.S. (2019). A de novo shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. Cell Syst. *8*, 27–42.e26.

Siebert, M., and Soding, J. (2016). Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. Nucleic Acids Res. *44*, 6055–6069.

Stella, S., Cascio, D., and Johnson, R.C. (2010). The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. Genes Dev. *24*, 814–826.

Stormo, G.D., Schneider, T.D., Gold, L., and Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res. *10*, 2997–3011.

Suzuki, M., Amano, N., Kakinuma, J., and Tateno, M. (1997). Use of a 3D structure data base for understanding sequence-dependent conformational aspects of DNA11Edited by B. Honig. J. Mol. Biol. *274*, 421–435.

Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., et al. (2013). Evaluation of methods for modeling transcription factor sequence specificity. Nat. Biotechnol. *31*, 126–134.

Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R., and Rohs, R. (2017). Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. Mol. Syst. Biol. *13*, 910.

Zhao, Y., Ruan, S., Pandey, M., and Stormo, G.D. (2012). Improved models for transcription factor binding site identification using nonindependent interactions. Genetics *191*, 781–790.

Zhao, Y., and Stormo, G.D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat. Biotechnol. *29*, 480–483.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordan, R., and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. Proc. Natl. Acad. Sci. U S A *112*, 4654–4659.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. *41*, W56–W62.

iScience, Volume *23*

Supplemental Information

Transcription Factor Binding Affinities and DNA Shape Readout

Max Schnepf, Marc von Reutern, Claudia Ludwig, Christophe Jung, and Ulrike Gaul

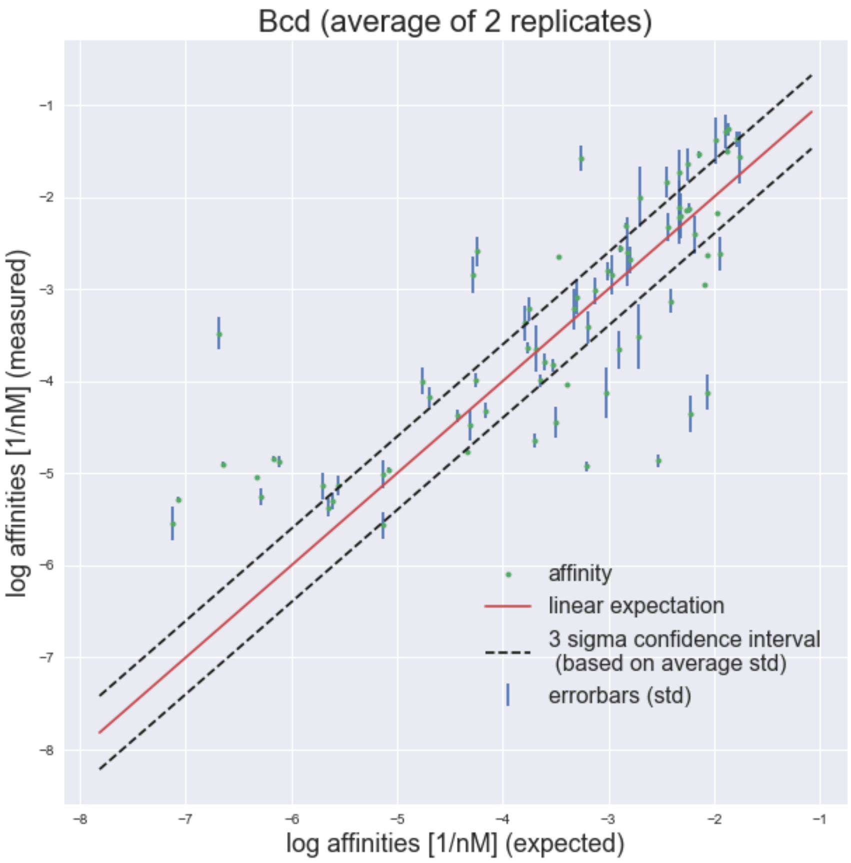


Figure S1. Predicted versus measured binding affinities, related to Figure 2. Comparison of measured changes in binding affinities against predicted values after dinucleotide mutations for the factor Bcd. The predicted values were calculated from our Bcd PWM assuming base independence. The average of two experimental values is plotted, with the difference shown by the error bars. The 3σ confidence interval (delimited by the dotted lines) indicates the region were predicted and measured values are considered to be in agreement. σ: average standard deviation of all measurements. The PWM fails to predict many dinucleotide mutation values, indicating positions non-independent contribution of the bases at these positions (data points outside the 3σ confidence interval).

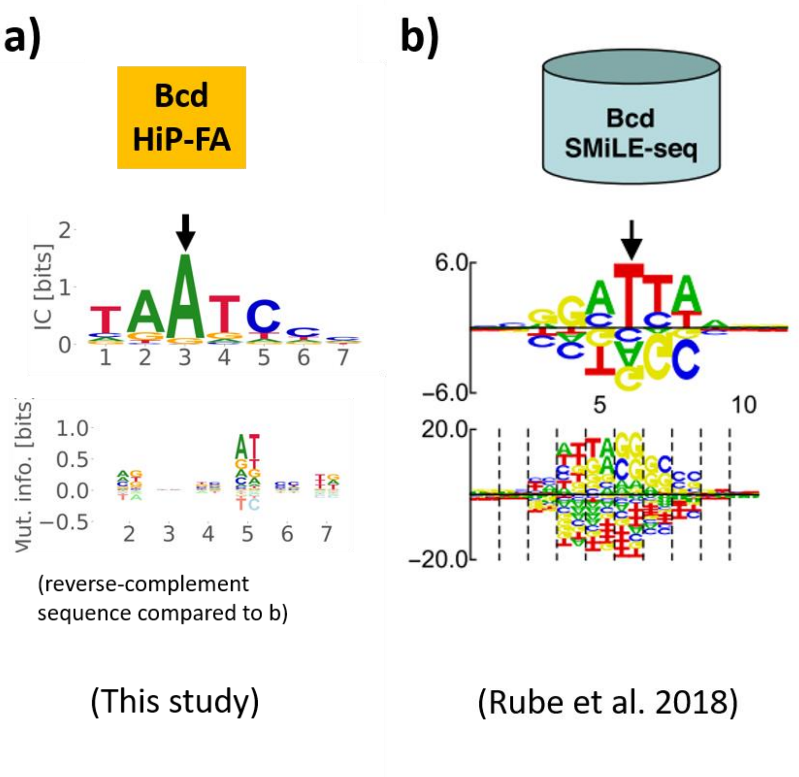


Figure S2. Comparison of the Bcd binding specificities obtained with another method, related to Figure 2. (**a**) Bcd PWM and DPWM derived from binding energies obtained by HiP-FA. (**b**) Mechanistic-agnostic model of binding specificity developed by Rube et al for the Bcd (reverse-complement sequences compared to the logos shown in **a**). The HiP-FA DPWM (lower panel in **a**) exhibits fewer but more preferred dinucleotides, as indicated by higher mutual information MI, due to the low noise of our binding affinity data. The black arrow shows the same position.

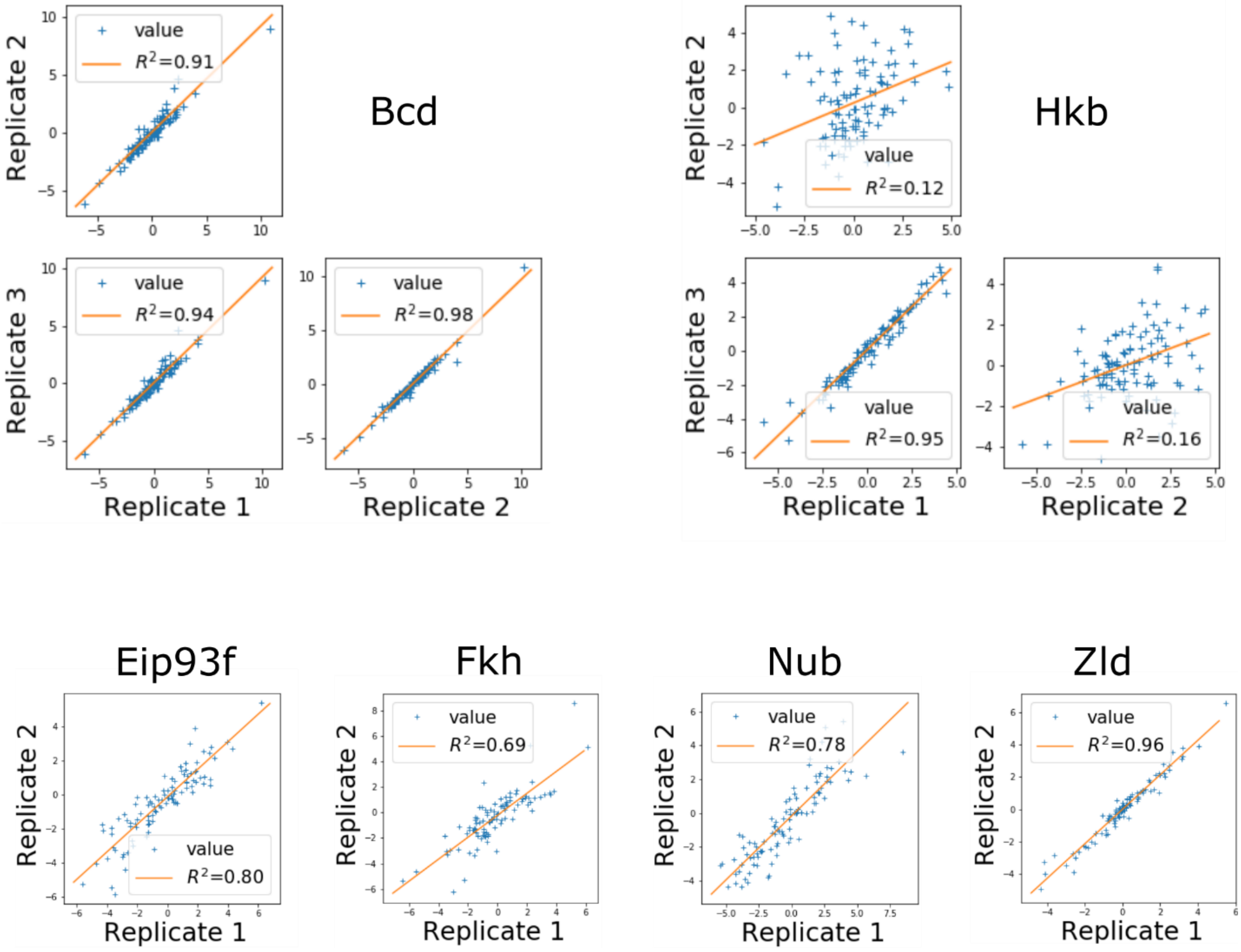


Figure S3. Reproducibility of the shape readout values between triplicates for Bcd and Hkb, and between duplicates for Eip93f, Fkh, Nub and Zld, related to Figure 3. For a given factor, the shape readout values of a given replicate are plotted all together as a function of a second replicate. R² (squared Person-correlation-coefficient) are given for linear regression. The replicates were obtained with independent measurements conducted at different days. Note that for Hkb only replicate 2 is in poor agreement with the two other replicates due to protein aggregation during the measurements that increased variability.

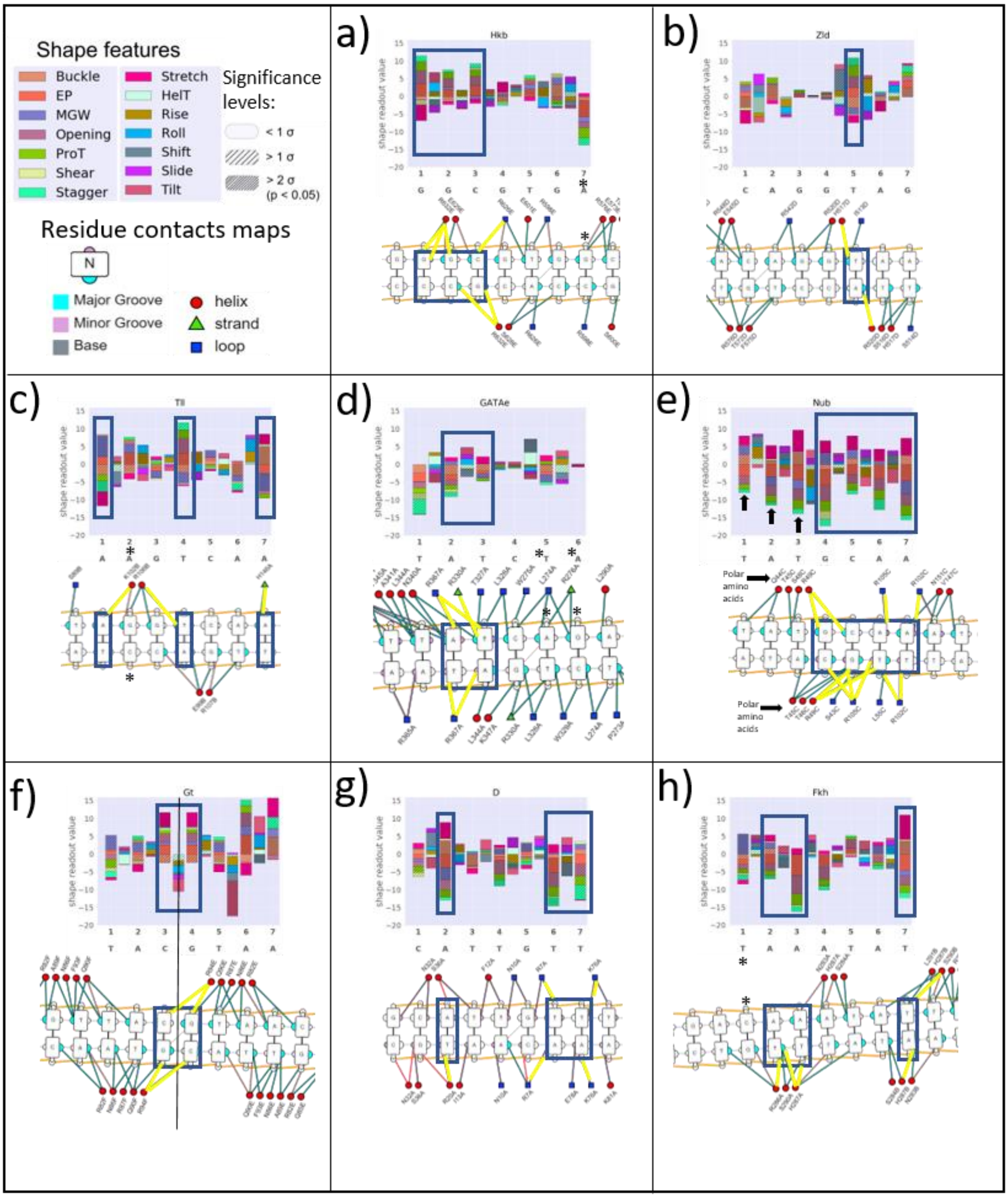


Figure S4. Correlation between DNA shape readout and structural information, related to Figure 4. As structural data are available only for the homeodomain Bcd and the two other homeodomains (Gsc and Oc) which share very similar same protein sequence and binding specificities (**Figure 4a**), we sought for experimental structures of additional homologs. First, we used the *Phyre2* web portal for protein homology modelling and prediction (Kelley et al., 2015) to create modelled structures for the remaining factors. Then, by choosing as criteria a confidence level >99% in the model prediction, amino-acids sequence identity >40%, and a

deviation from the DNA consensus sequence in the binding domain of maximal 2 nucleotides, we found additional structures for the factors Hkb (**a**), Zld (**b**), Tll (**c**), GAGAe (**d** and **Figure 4c**), Nub (**e**), Gt (**f** and **Figure 4b**), D (**g**) and Fkh (**h**), thereby providing experimental structures for 11 out of the 13 investigated factors (no structures found for Hb and Eip93f with high enough percentage of identity).

We then used the *DNAproDB* database and processing pipeline developed by the Rohs group (Li et al., 2017; Sagendorf et al., 2019) for the structural analysis of the DNA-TF complexes. This analysis gives hints about multiple DNA-protein structural features for thousands of complexes by combining features of DNA, protein or DNA-protein interactions at the interface, and provides tools for creating visualizations of the DNA-protein interface.

The residue contacts maps of our factors (**a-h** lower sketches) are shown with the shape readout value profile (higher plots). A residue contacts map indicates individual nucleotide-residue interactions, DNA secondary structure, protein secondary structure and DNA interaction moieties. The DNA is displayed as a graph, with nucleotides being nodes and edges between them indicating backbone links, base pairing or base stacking (legend on the top left). Protein residues are displayed as small nodes with the node shape and color representing residue secondary structure. Edges between residue and nucleotide nodes represent an interaction between the two and which DNA moiety(s) the interaction involves.

Although little is known about the relationship between structural features and binding affinities, it has been shown that a narrow minor groove width (MGW) can enhance the negative electrostatic potential (EP) in the minor groove (Chang et al., 2013; Rohs et al., 2009), which can attract positively charged amino acids such as arginine (R), lysine (K) and histidine (H), considering the latter can be protonated. We thus focused on the nucleotide positions with high and significant shape values for the features MGW and EP by seeking the presence of contacts with R, K and H residues. To evaluate the contact frequency of these residues at the nucleotide positions with strong shape readout values, we selected for each TF up to 4 positions with strong shape readout based on their high cumulated shape readout values, in particular with significant MGW and/or EP shape readout values (positions highlighted with blue rectangles). We then identified the R, K and H residues interacting with these nucleotides (interactions highlighted in yellow in the residue contact maps). Overall, we observe very frequent interactions with positively charged amino acids at the selected positions, as it was the case with Bcd (**Figure 4a**). For a quantitative evaluation, we calculated for all the factors the ratio between the number of these interactions and the total number of interactions with the other amino-acids, and compared them with the respective ratios at all other nucleotide positions exhibiting less prominent shape readout values. Remarkably, we found for the strong shape readout positions much more frequent nucleotide interactions with R, K or H (64% of the contacts) than for the other positions (28% of the contacts). It is already known for homeodomains that the N-terminal tail recognizes DNA shape in the minor groove involving electrostatic attraction between positively charged residues and the shape-dependent EP. As the recognition helix is thought to be engaged to base readout of the major groove, it is difficult to evaluate its contribution to shape readout since shape is dictated by sequence. Our results generalize the contribution of the positively charged amino-acids to the shape readout to other DNA secondary structures (such as α-helixes) and to other binding domains.

Interestingly, the POU domain Nub (**f**) uses the most extensively the DNA shape readout, in particular with the MGW shape feature and the EP exhibiting high and significant shape readout values at all positions. Surprisingly, whereas at the consecutive nucleotide positions 4, 5, 6 and 7 (blue rectangle) the majority of contacts occurs with positive residues (11 out of 15 in total), no interaction with a positive residue can be observed at nucleotide positions 1, 2 and 3. Strikingly, a closer inspection of the residue contacts maps at these positions reveals that all interacting amino-acids (glutamine, threonine, and serine) are highly polar. Hence, this is a hint that not only positively charged residues can strongly contribute to DNA shape readout, but also non-charged but strongly polar amino-acids. To the best of our knowledge such an effect has not been reported so far. To conclude, this shape/structure analysis further validates our approach, while bringing biological insights.

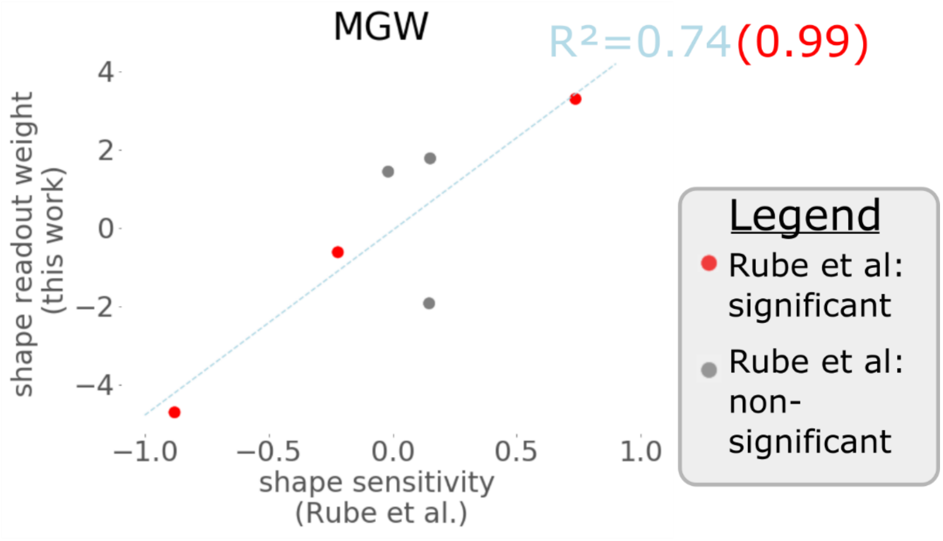


Figure S5. **Comparison of our shape readout values with the MGW shape sensitivity coefficients from Rube et al. (Rube et al., 2018), related to Figure 4.** The shape readout values determined in this study for the MGW of Bcd are plotted against the equivalent shape sensitivity coefficients reported by Rube et al. Values reported to be significant in their study are depicted in red, the non-significant ones in grey. A linear regression (blue dashed line) has an R² of 0.74 for all values, 0.99 for only significant data points.

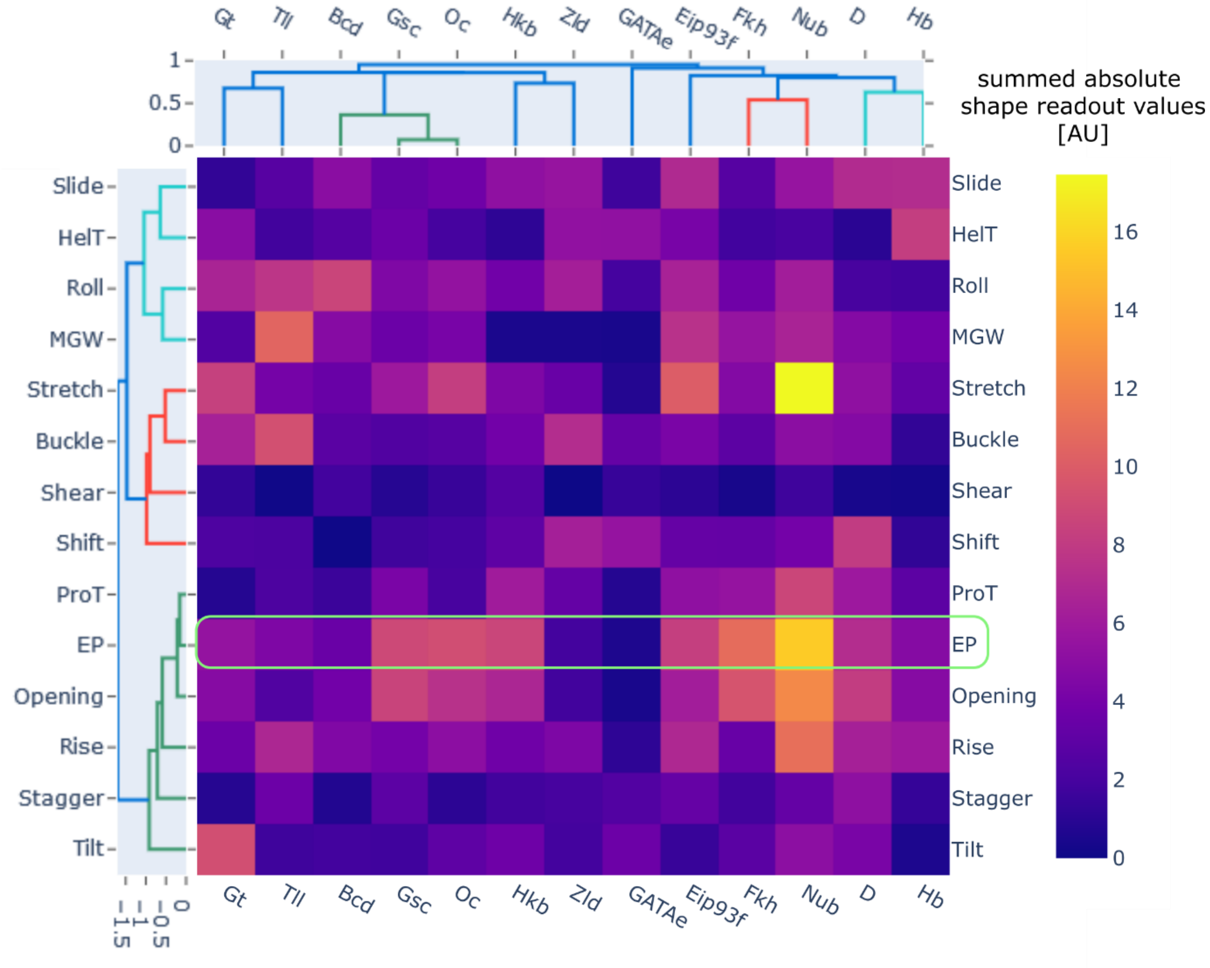


Figure S6. **Heat map and clustering of the shape readout values for the *Drosophila* TFs and DNA shape features, related to Figure 3**. The heat map shows the sums of the absolute shape readout values over all positions of each of the TFs. In addition, both TFs and shape features are clustered by correlation distance (**Transparent Methods**). The clustering was performed on the non-aggregated data (with positions as a second dimension).

The TF clustering indicates that the different binding proteins show little similarity in their use of the shape features except for the homeodomains (green tree on top), which was expected. In contrast, the clustering of the shape features reflects structural dependencies between shape features. First, the heat map shows that EP is one of the features influencing TF-DNA binding energies the most, with strong means correlations for at least 7 factors (Gsc, Oc, Hkb, Eip93f, Fkh, Nub, and D; highlighted in green in the heat map). The EP being sensitive to the interaction between positively charged residues and the minor groove (Chiu et al., 2017), this strong impact on TF-DNA binding is therefore not surprising. Second, we observe three distinct clusters of shape features (cyan, red, and green trees on the right). These distinct groups may be related to biophysical properties of the DNA and its interplay with the binding protein such as bends, kinks, A-/Z-DNA (Rohs et al., 2010). For instance, the first cluster (in cyan on the left) consists of slide, helix twist, roll and MGW. These features were reported to correlate the most with each other both in bound and unbound DNA (El Hassan and Calladine, 1995) (Stella et al., 2010), and are read out concertedly in DNA-protein complexes (Suzuki et al., 1997). This interdependency can explain their co-appearance in a cluster (cyan tree on the left-hand side). It is also noteworthy that this cluster contains three inter-base pair features (Slide, HelT and Roll) out of four, whereas the second cluster (in red) contains mainly intra-base pair features (Stretch, Buckle and Shear). Thus, there seems to be a synergy between inter- and intra-base pair features for the DNA shape readout. The third group (in green) is more heterogeneous and doesn’t follow a simple pattern: it contains mixed shape features with, to the best of our knowledge, no known relationship to each other.

Study	Short description	TFs	Organism	New generated specificity data?	Method	Readout	Sequence space	Shape features tested	Shape analysis	Findings
This study	Transcription factor binding affinities and DNA shape readout	13 TFs from 8 domain families	*Drosophila*	YES	HiP-FA	Absolute binding affinities *in solution* (absence of unspecific adsorption as for the surfaces methods)	$\sim10^{3}$	All 13 known shape features and the EP	Direct correlation between shape features and binding affinities	Accurate binding specificities including dinucleotide dependencies; Characterization of the DNA shape contribution to binding using directly the binding affinities – widespread use of DNA shape readout by these factors
(Pal et al., 2019)	Co-SELECT reveals sequence non-specific contribution of DNA shape	83 TFs from 3 domain families	*Mammals*	NO	HT-SELEX	Sequences	$\sim10^{7}$	MGW, Roll, ProT, and HelT	The method leverages the presence of motif-free sequences in late HT-SELEX rounds and their enrichment in weak binders allows to detect an evidence for the role of DNA shape features in TF binding	Motif-independent contribution of shape-dependent binding
(Samee et al., 2019)	A *de novo* shape motif discovery algorithm	106 TFs in K562 and Gm12878 cell lines	*Human*	NO	ChiP-seq	Sequences	$\sim10^{4}$	MGW, Roll, ProT, and HelT	Shape motif discovery algorithm that compares profiles of shape-features between TF-bound regions and non-bound regions	Many TFs have shape motifs in both *in vivo* and *in vitro* data; Shape motifs encode specificity that goes beyond the sequence motif of a TF
(Rube et al., 2018)	A unified approach for quantifying and interpreting DNA shape readout by TFs	8 Hox TFs in complex from Slattery et al.; homeodomain Bcd and a human ETS factor	*Drosophila*	NO	SELEX-Seq; SMiLE-seq	Relative affinities; Absolute binding affinities	$\sim10^{7}$	MGW, Roll, ProT, and HelT	No Read Left Behind (NRLB) algorithm to analyse the SELEX-seq data; Linear regression model to predict shape parameters; 2-steps *post hoc* analysis method (*Shape projection*) to analyse protein-DNA binding models in terms of shape readout	Adding dinucleotide features as sequence-to-shape predictors to a linear model can almost perfectly explain the shape parameters; *Shape projection* detects shape readout and overcomes the confounding between base and shape readout
(Mathelier et al., 2016)	DNA shape features improve TF binding sites (TFBS) predictions	76 TFs from 24 domain families	*Human*	NO	ChiP-seq	Sequences	$\sim10^{4}$	MGW, Roll, ProT, and HelT	TFBS predictions; average of feature values	Combining shape features with position specific scoring matrix (PSSM) improve TFBS predictions; Two domain TF families benefit most from shape information
(Yang et al., 2017)	Family-specific DNA shape readout	215 TFs from 27 domain families	*Mammals*	NO	HT-SELEX; PBMs	Sequences; relative affinities	$\sim10^{7}$	MGW, Roll, ProT, and HelT	Comparison of TF-DNA binding model performance by using only mononucleotide features or mononucleotides and shape features	DNA shape features improve modeling of DNA-binding specificities across TF families; DNA shape in regions immediately flanking the core-binding site is generally recognized upon TF binding
(Zhou et al., 2015)	Quantitative modeling of TF binding specificities using DNA shape	68 TFs from12 domain families	*Mammals*	NO	PBMs; SELEX-seq for one TF	Sequences; relative affinities for one TF	$\sim10^{7}$	MGW, Roll, ProT, and HelT	Comparison of TF-DNA binding model performance by using only mononucleotide features or mononucleotides and shape features	Shape-augmented models perform better than sequence-based models; DNA shape features predictive of relative binding affinities
(Riley et al., 2014)	DNA binding specificities of Exd-Hox complexes	8 Hox TFs in complex with a cofactor (Exd)	*Drosophila*	YES	SELEX-Seq	Relative affinities	$\sim10^{7}$	MGW	MGW values for the strongest binders	DNA shape contributes to Exd-Hox Dimer preferences

ADVANTAGES OF OUR STUDY:

- New set of higher-order binding specificities with low-noise due to the high sensitivity of the HiP-FA binding assay (see **Figure S2**).
- Shape analysis using directly the raw binding affinity data and filtering out partially the effect of base readout.
- The affinity data and their shape analysis detect a significant shape readout for multiple shape features and for most of our factors

LIMITATIONS:

- Sequence space limited to thousands of sequences in total.
- No clear separation between base and shape readout as opposed to the algorithm from Rube et al (their *post hoc* analysis requires however to compute the binding energies using a scoring matrix).

Table S1. Comparison of different approaches and summary of the results obtained as far as DNA shape readout analysis is concerned, related to Figure 1.

TRANSPARENT METHODS

Protein expression and purification

For most transcription factors, it is difficult to express the full-length proteins at high levels in bacteria or eukaryotic cells (Burz et al., 1998). Therefore, only GATAe was expressed as full-length protein. For the other factors, we cloned the DNA-binding domains (DBD) of the TFs, flanked by 14 additional amino acids on either side, into the bacterial expression vector pGEX-6P-1 (GE Healthcare). In this vector, the polypeptide of interest is fused to a N-terminal Glutathione (GST) tag and placed under the control of an IPTG-inducible promoter.

The fusion constructs were transformed into chemically competent *E. coli* (Top10f, homemade) and protein expression was induced by 1mM IPTG for 20h at 18°C. Incubation at this temperature allows proper protein folding and higher expression levels.

The proteins were purified on 5ml GSTrap columns using an ÄKTA protein purification system (GE Healthcare) following the manufacturer’s protocol. Certain protein preparations contained high levels of bacterial DNA contamination, as judged by UV spectroscopy (Nanodrop, Thermo Scientific), and were therefore subjected to an additional Heparin purification step using 1ml HiTrapHEP columns. The purity of the proteins was verified by SDS-PAGE.

Determination of the TF-DNA affinities

The binding energies between the TFs and DNA oligomers harboring the mutations of interest were determined using the previously described HiP-FA method. For details about the experimental procedure see (Jung et al., 2018; Jung et al., 2019). In brief, the TF of interest and a fluorescently labelled reference DNA oligomer are embedded in a porous agarose gel matrix. The competitor solution is added on top of this gel at the start of the experiment. The spatial-temporal concentration gradient of competitor solution allows to record a titration curve to determine the affinity of the TF towards the respective competitor sequence. Separate wells containing the DNA intercalating dye Nile Blue are used to determine the competitor concentration at any given time and position.

In contrast to the original HiP-FA protocol, neighboring double mutations are systematically introduced in the TF's consensus (i.e. strongest binder) sequence (at 7 positions in the center of the 16 bp DNA oligomer, 6 positions for GATAe). The flanking sequences of the binding site were optimized to reduce the occurrence of off-target binding, checking all possible sequences using a modified version of the *Python3* application *PySite* (available on *Github*: https://github.com/Reutern/PySite) with our PWMs (Jung et al., 2018) as input.

Determination of binding weight and off-target removal

From the perspective of the PWM model, every site is a binding site and two sequences differ only in their binding weight. For this reason, residual binding activity persists even in regions of oligomers that were hand-picked to avoid a consensus binding site. These remaining binding activities can potentially influence the binding behavior of the whole oligomer, and should be considered in the calculation of binding weights.

We developed a heuristic algorithm that constructs a PWM *de novo* from a set of oligomers s_i with known binding affinity k_i. The binding weight is inversely proportional to the binding affinity. For simplicity, the ratios between affinity and inverse weight is set to 1. We assume that at most one protein binds to each oligomer at any time. Therefore, we can approximate the total binding weight w_j of oligomer j as

w_j = sum (i site on j) w_i where $w_i = 1/k_i$ is the binding weight of site i. Our model searches the space of PWMs with an iterative approach. The goal of the algorithm is to find the PWM that best matches the measurements. We assess the quality of fit by a scaled sum of squared errors between estimated and measured binding weights. For the first iteration, we construct PWM_0 based on the target sites at the center of the oligomers. The following iterations carried out in three steps: first, find the binding sites and calculate the weights based on the PWM of the previous iteration. Second, assign heuristic binding weights to every called site. This is done by distributing the measured weight of an oligomer among all its sites, based on their calculated binding weight ratios from the last step. Third, construct a new PWM from the list of sites and their heuristically estimated binding weights.

To transform the binding weights into an energy equivalent space, their natural logarithm was taken, yielding an energy based on ΔΔG/RT.

Representation of PWMs and DPWMs

PWMs are depicted as sequence logos according to (Schneider and Stephens, 1990) using a custom *Python3* script. In the DPWMs, the mutual information MI is calculated using a Kullback–Leibler divergence (Kullback and Leibler, 1951) using a logarithm with base 2. The significance of the MI values was assessed by determining a threshold for a significant MI value. To this end, we first calculated the standard deviations of each of the dinucleotide MIs that we obtained from the binding affinities of Bcd (for which we measured 3 replicates). We found $\sigma(MI)_{average}$= 0.01 bits, which corresponds to a conservative estimate of the “noise” present in our MI values. We then considered a MI value to be significant when MI> 3× $\sigma(MI)_{average}$ = 0.03 bits. Scaled up to 6 positions (the length of our DPWM logos) this corresponds to ~ 0.2 bits.

Thus, as the lowest total MI we measured is 0.3 for GATAe, all our factors show significant total MI.

Shape readout values

To determine the influence of DNA shape on the binding weight of a TF, the free binding energies were computed from the binding weights as described above and normalized using a z-score to make the subsequent analysis more robust against the range of the different binding energies, and thereby less influenced by a TF's specificity. The values of the DNA features were calculated using the lookup tables provided by the Rohs group (Chiu et al., 2017; Li et al., 2017; Zhou et al., 2013). The normalized energies were plotted against the rescaled shape values of a feature (scaled from 0 to 1 for all possible values the feature can possibly take) for each position, see also **Figure 1c**. Per position, only those data points were used that contained a mutation at this position or at the next neighboring ones. Since the shape features between two neighboring base pairs (inter features) are assigned to both base pairs, those features have four possible positions where mutations are included in the plot and the following regression (two positions plus two neighbors). The normalized energies were then fitted according to *Equation 1*:

$$-\frac{\Delta\Delta\mathrm{G}}{RT}\,(\mathrm{norm.}) = s \times shapef\ (norm.) + C$$

Where $\frac{\Delta\Delta\mathrm{G}}{RT}\,(\mathrm{norm.})$ are the normalized free binding energies for all dinucleotide mutations at a given nucleotide position, *s* the shape readout values, *shapef (norm.)* the normalized shape feature values, and *C* an offset value. A robust linear regression (Huber, 1981) using Huber's T as M-estimator, with mean absolute deviation as scale factor (implemented in the *rlm_model* from the *statsmodels package- v0.8.0* in *Python*

3.1) was used to perform the linear regression (*Equation 1*) and to estimate a confidence interval using an iteratively reweighted least squares approach. The statistics of the robust linear regression was used to define confidence intervals. The significance level was determined depending if the value of this confidence interval was more than one or two standard errors different from zero, respectively. Only values more than two standard errors different from the mean are considered as statistically significant (p<0.05).

Clustering of shape features and TFs

The hierarchical clustering was performed using the *figure_factory* module from the *plotly package –v4.0.0* in *Python 3.1*. The distances were calculated based on correlations to avoid single features with high values to dominate the clustering and have a less scale variant distance function. The clustering was performed based on the shape readout values as defined above. To cluster the TFs, the features and positions were treated as two different dimensions for the clustering. To reduce the impact of the window chosen for the core sequence, we modified the distance function to allow for up to two bases offset, and setting shape readout values to zero for positions outside of the chosen window. The clustering for the features was performed on a transposed matrix with TFs and positions as dimensions of the data.

Protein structure depiction

We used the *DNAproDB* database and processing pipeline developed by the Rohs group (Li et al., 2017; Sagendorf et al., 2019) for the structural analysis and the structure rendering of the DNA-TF complexes. For the structures shown in **Figure 4**, we used crystal structures of Bcd (pdb-ID: 1ZQ3) (Baird-Titus et al., 2006), of a B-zip TF (pdb-ID: 1GD2) (Fujii et al., 2000) with the same core consensus sequence as Gt, and of a similar GATA TF (pdb-ID: 3DFV) (Bates et al., 2008).

SUPPLEMENTAL REFERENCES

Baird-Titus, J.M., Clark-Baldwin, K., Dave, V., Caperelli, C.A., Ma, J., and Rance, M. (2006). The solution structure of the native K50 Bicoid homeodomain bound to the consensus TAATCC DNA-binding site. Journal of molecular biology *356*, 1137-1151.

Bates, D.L., Chen, Y., Kim, G., Guo, L., and Chen, L. (2008). Crystal structures of multiple GATA zinc fingers bound to DNA reveal new insights into DNA recognition and self-association by GATA. Journal of molecular biology *381*, 1292-1306.

Burz, D.S., Rivera-Pomar, R., Jackle, H., and Hanes, S.D. (1998). Cooperative DNA-binding by Bicoid provides a mechanism for threshold-dependent gene activation in the Drosophila embryo. Embo J *17*, 5998-6009.

Chang, Y.P., Xu, M., Machado, A.C., Yu, X.J., Rohs, R., and Chen, X.S. (2013). Mechanism of origin DNA recognition and assembly of an initiator-helicase complex by SV40 large tumor antigen. Cell Rep *3*, 1117-1127.

Chiu, T.P., Rao, S., Mann, R.S., Honig, B., and Rohs, R. (2017). Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein-DNA binding. Nucleic Acids Res *45*, 12565-12576.

El Hassan, M.A., and Calladine, C.R. (1995). The Assessment of the Geometry of Dinucleotide Steps in Double-Helical DNA; a New Local Calculation Scheme. Journal of molecular biology *251*, 648-664.

Fujii, Y., Shimizu, T., Toda, T., Yanagida, M., and Hakoshima, T. (2000). Structural basis for the diversity of DNA recognition by bZIP transcription factors. Nat Struct Biol *7*, 889-893.

Huber, P. (1981). Robust Statistics (New York: John Wiley and Sons, Inc).

Jung, C., Bandilla, P., von Reutern, M., Schnepf, M., Rieder, S., Unnerstall, U., and Gaul, U. (2018). True equilibrium measurement of transcription factor-DNA binding affinities using automated polarization microscopy. Nature Communications *9*, 1605.

Jung, C., Schnepf, M., Bandilla, P., Unnerstall, U., and Gaul, U. (2019). High Sensitivity Measurement of Transcription Factor-DNA Binding Affinities by Competitive Titration Using Fluorescence Microscopy. JoVE, e58763.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols *10*, 845-858.

Kullback, S., and Leibler, R.A. (1951). On Information and Sufficiency. Ann Math Statist *22,* 79-86.

Li, J., Sagendorf, J.M., Chiu, T.P., Pasi, M., Perez, A., and Rohs, R. (2017). Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. Nucleic Acids Res *45*, 12877-12887.

Mathelier, A., Xin, B., Chiu, T.-P., Yang, L., Rohs, R., and Wasserman, Wyeth W. (2016). DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. Cell Systems *3*, 278-286.e274.

Pal, S., Hoinka, J., and Przytycka, T.M. (2019). Co-SELECT reveals sequence non-specific contribution of DNA shape to transcription factor binding in vitro. Nucleic Acids Res *47*, 6632-6641.

Riley, T.R., Slattery, M., Abe, N., Rastogi, C., Liu, D., Mann, R.S., and Bussemaker, H.J. (2014). SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. Methods in molecular biology (Clifton, NJ) *1196*, 255-278.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. Annual review of biochemistry *79*, 233-269.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248-1253.

Rube, H.T., Rastogi, C., Kribelbauer, J.F., and Bussemaker, H.J. (2018). A unified approach for quantifying and interpreting DNA shape readout by transcription factors. Molecular Systems Biology *14*, e7902.

Sagendorf, J., Markarian, N., Berman, H., and Rohs, R. (2019). DNAproDB: an expanded database and web-based tool for structural analysis of DNA-protein complexes. Nucleic Acids Research *48*.

Samee, M.A.H., Bruneau, B.G., and Pollard, K.S. (2019). A De Novo Shape Motif Discovery Algorithm Reveals Preferences of Transcription Factors for DNA Shape Beyond Sequence Motifs. Cell Systems *8*, 27-42.e26.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res *18*, 6097-6100.

Stella, S., Cascio, D., and Johnson, R.C. (2010). The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. Genes & Development *24*, 814-826.

Suzuki, M., Amano, N., Kakinuma, J., and Tateno, M. (1997). Use of a 3D structure data base for understanding sequence-dependent conformational aspects of DNA11Edited by B. Honig. Journal of molecular biology *274*, 421-435.

Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R., and Rohs, R. (2017). Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. Mol Syst Biol *13*, 910.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordan, R., and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. Proceedings of the National Academy of Sciences of the United States of America *112*, 4654-4659.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res *41*, W56-62.