Local DNA shape is a general principle of transcription factor binding specificity in *Arabidopsis thaliana*

Janik Sielemann[1,3], Donat Wulf[1,2,3], Romy Schmidt[4] and Andrea Bräutigam[1,2,3*]

[1] Computational Biology, Center for Biotechnology (CeBiTec), Bielefeld University, 33615 Bielefeld, Germany

[2] Computational Biology, Faculty of Biology, Bielefeld University, 33615 Bielefeld, Germany

[3] Graduate School DILS, Bielefeld Institute for Bioinformatics Infrastructure (BIBI), Bielefeld University, 33615 Bielefeld, Germany

[4] Plant Biotechnology, Bielefeld University, 33615 Bielefeld, Germany

[*] Corresponding author

A genome encodes two types of information, the "what can be made" and the "when and where". The "what" are mostly proteins which perform the majority of functions within living organisms and the "when and where" is the regulatory information that encodes when and where proteins are made. Currently, it is possible to efficiently predict the majority of the protein content of a genome but nearly impossible to predict the transcriptional regulation. This regulation is based upon the interaction between transcription factors and genomic sequences at the site of binding motifs[1,2,3]. Information contained within the motif is necessary to predict transcription factor binding, however, it is not sufficient[4]. Peaks detected in amplified DNA affinity purification sequencing (ampDAP-seq) and the motifs derived from them only partially overlap in the genome[3] indicating that the sequence holds information beyond the binding motif. Here we show a random forest machine learning approach which incorporates the 3D-shape improved the area under the precision recall curve for binding prediction for all 216 tested *Arabidopsis thaliana* transcription factors. The method resolved differential binding of transcription factor family members which share the same binding motif. The models correctly predicted the binding behavior of novel, not-in-genome motif sequences. Understanding transcription factor binding as a combination of motif sequence and motif shape brings us closer to predicting gene expression from promoter sequence.

Main

Changes in gene expression during development and invoked by environmental perturbations are critical to organismal function and these changes are influenced by DNA-binding transcription factors. *Arabidopsis thaliana* encodes 1533 DNA-binding transcription factors[5] many of which occur in protein families of a few to over a hundred members[6]. The DNA-binding transcription factors carry a domain which interacts directly with DNA as well as

domains which mediate regulation. Gene expression of a particular gene is a complex read-out based on the presence of transcription factors and their spacing on the DNA, chromatin status, histone flags and presence of co-activators or repressors. Improving the understanding of those regulatory relations and pathways is necessary to tackle current agricultural challenges[7]. Hundreds of sequence motifs to which transcription factors bind have been characterized[3,8], but currently it is impossible to look at a promoter and understand its regulatory syntax.

This is partially due to the fact that many members of a particular TF family bind the same motif. The *A. thaliana* genome contains 74 members of the WRKY transcription factor family[9]. These WRKY transcription factor family proteins are implicated in diverse processes from trichome and seed development to roles in biotic and abiotic stresses[10]. Despite their diverse roles, all WRKYs analyzed bind to a consensus motif, the W-box, which is characterized by the TTGAC pentamer followed by C or T[11]. Transcription factors of the 133 member bHLH family bind DNA via basic amino acids at the N-terminal end of the bHLH domain and bind a variation of the motif CANNTG, frequently the so-called G-box CACGTG[12]. This G-box motif is also bound by many bZIP transcription factors whose core motif is ACGT, the central nucleotides of the G-box CACGTG[13,14]. ChIP-seq data clearly indicates that only a subset of potential binding sites are indeed occupied at any given time in a particular tissue[15–17]. To explain why transcription factors binding the same sequence yield different regulatory output, additional variables such as other transcription factors as binding partners, chromatin status, histone modifications, DNA methylation and posttranslational modifications of transcription factors have been considered. In contrast, the functional divergence has been hypothesized to be dependent on varying binding affinities to the same core sequence[4].

DNA affinity purification sequencing (DAP-seq) in which the DNA bound to an *in vitro* produced transcription factor is sequenced has also revealed extensive family-type motifs for many transcription factors (http://neomorph.salk.edu/dap_web/pages/browse_table_aj.php, [3]).

For motif detection in DAP-seq or ChIP-seq, the DNA sequences bound by the TF are mined by motif search algorithms such as MEME or MEME-chip which identify overrepresented motifs among the sequences[18]. For many TFs, biochemical experiments such EMSAs have confirmed that the motif is necessary for binding[19–21]. The comparison of the direct results of DAP-seq, i.e. the peaks in which reads can be aligned with the genome with the occurrence of the motif predicted using the same reads indicates that during motif prediction information is lost (Figure 1A, Supplemental Figure 1). AmpDAP-seq uses a single *in-vitro* produced protein that binds amplified DNA devoid of methylation marks, histones and other proteins. Hence the bound DNA sequence contains information which is lost during motif prediction. This raises the question which information is contained in DNA sequence besides the code of A, T, C and G.

DNA is a very constrained molecule since its phosphate sugar backbone runs antiparallel while its bases are paired and arranged in rungs on a helical ladder. Since the discovery of the genetic code for proteins, we have been trained to see DNA essentially as code. However, despite the constraints, the exact position of each base pair and each base in a pair is influenced by its surrounding bases. The pairs can be tilted, shifted, slid, rolled, risen and twisted relative to each other (Figure 1B,[22,23]). The bases in a pair can be buckled, sheared, stretched, twisted, opened and staggered (Figure 1B,[22]). The width of the minor groove is also influenced by the surrounding bases[24]. This DNA shape has been demonstrated to influence protein-DNA binding, for instance of the *Drosophila* Scr Hox Protein[24,25] and the *S. cerevesia* bHLH proteins Cbf1 and Tye7[26] . We hypothesized, that models built on DNA shape within and surrounding the binding motif recover the information lost during motif detection and improve prediction for transcription factor binding in *A. thaliana*.

DNA shape calculations have been established from X-ray crystallography, Monte Carlo simulation and molecular dynamic simulation[22]. DNA shape is encoded in 13 features

represented by numeric vectors. To test the importance of DNA shape, a random forest decision tree (RF) based machine learning approach was chosen to enable reliable feature extraction[27]. To test whether DNA shape encodes additional information, we split the motif occurrences into a training and a test-set[28] using the peak height of a DAP-seq experiment as the numeric for which the RF algorithm was trained. To ensure consistent 3D structure learning, the sequences were reverse complemented if the binding sequence was located on the minus strand. Initially, a random forest classifier was trained using the binding/non-binding cut-offs established previously[3]. We observed improved binding predictions for all tested transcription factors (Supplemental Figure 2), but quantitative information was lost. In a second step, a regressor was trained on the raw binding data using the peak height in ampDAP-seq as a proxy for binding affinity. The performance was similar (Supplemental Figure 3). To improve the use of peak height as a proxy for binding affinity, the binding data was filtered for single motif occurrences and machine learning was repeated. 216 transcription factors were tested and in each case the shape based predictor outperformed the motif search based on the area under the precision recall curve (AUPRC). AUPRC improved between 2.8% and 362.7%, with an average of 93.2% (Figure 1C, Supplemental Figure 4). 33 TFs reach AUPRC of more than 0.8 indicating that the motif plus shape information suffices for prediction (Figure 1C). 101 TFs show medium AUPRC with between 0.5 to 0.8 AUPRC. The remaining 82 TFs improve their AUPRC compared to motif alone but do not exceed an AUPRC of 0.5 (Figure 1C). The test-sets clearly indicate that DNA shape contributes to TF binding specificity when bound vs non-bound sequences are analyzed (Figure 1C). To test the contribution of shapes surrounding the motif, the amount of shape information given to the regressor was varied and the training repeated. The major contribution of shape information was localized to the core motif plus two bases on each side of the motif (Figure 1D). These adjacent bases influence the shape of the bases and base pairs in the core[22]. Beyond the core motif shape, the information gain quickly leveled (Figure 1D, Supplemental Figure 5). The models generated by the shape-based regressor yields

information over which shapes are important to the binding for each of 216 TFs tested (Supplemental Table 1). To test if any shape, shape type (intra-basepair vs. inter-basepair) or any position contribute a larger amount of information to the binding, the top five features were extracted for each transcription factor. The 3D configuration of bases within the core binding sequence occupied 68% of the top five feature positions (Figure 1E) as expected (Figure 1D). Intra-basepair shapes contributed 39% and inter-basepair shapes contributed 61% (Figure 1E) of the top five shapes within the motif. Outside of the motif, the proportions reversed since intra-basepair shapes contributed 72% and inter-basepair shapes contributed 28% (Figure 1E). Far outside the motif, the shear feature was overrepresented among the top five features (Figure 1E). More precise prediction of TF binding requires shape information and the transcription factors depend on a wide variety of shape information.

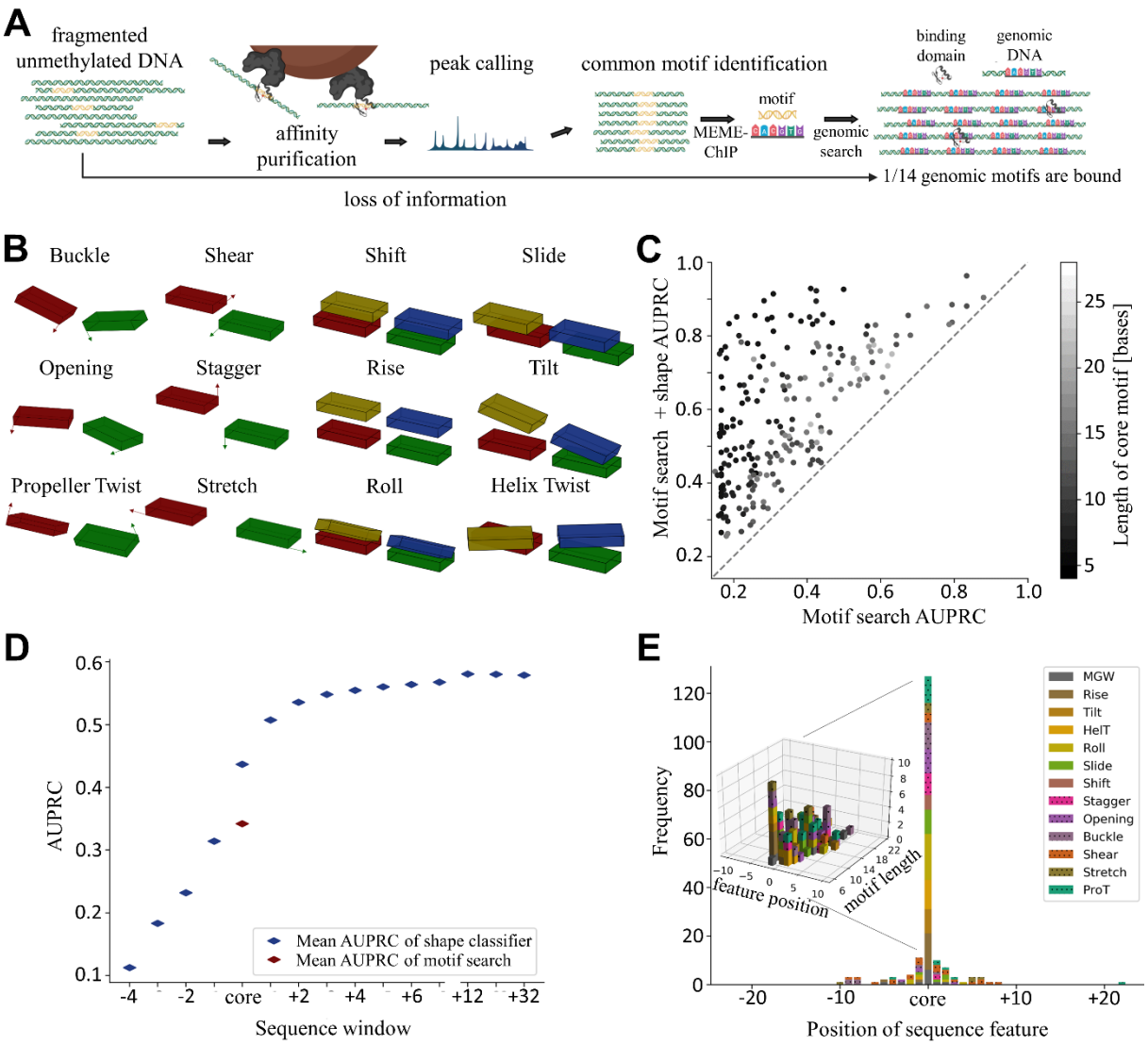


Figure 1: Overview of workflow and performance of shape based binding site identification.

A) Broad overview of the experimental procedure for the identification of TF binding sites. The DAP-Seq experiment was performed by O'Malley et al. 2016. Identified binding motif occurrence is on average 14-fold higher than the number of verified binding events. B) The different DNA shape features which were considered to analyse TF specificity. A query table was used for shape calculation[22]. C) Performance of the random forest classifier using the border widths with the highest area under the precision recall curve (AUPRC) for each transcription factor. D) AUPRC for differing sequence widths. The width was increased upstream and downstream of the core motif sequence, respectively. E) Frequency of features in the top 5 most important shape features for each transcription factor.

If DNA shape predicts binding better than motif alone and shape information used by TFs is varied, the prediction algorithm should be able to distinguish binding between two transcription factors which are predicted to bind the same motif sequence. To test the hypothesis, the models

for transcription factor pairs with the same binding motifs were analyzed. The ERF/AP2 transcription factors CBF4 (AT5G51990) and ERF036 (AT3G16280) both bind the GTCGGT/C motif which occurs 31,155 times in the *A. thaliana* genome. According to ampDAP-seq they have 9,910 binding sequences in common (Figure 2A). ERF036 binds 2,581 sequences which are not bound by CBF4, and CBF4 binds 6,996 sequences not bound by ERF036. To test whether the shape indeed supplies specificity, binding vs non-binding was predicted by the models (Figure 2B). The shape information of the 31,155 genomic sequences, allows the regressor models to distinguish binding events between the two transcription factors (Figure 2B-C), even though the core sequence is the same and the majority of sequences are bound by both transcription factors. Each Venn diagram in Figure 2C shows the distribution of binding sites applying the cutoff represented by the dashed line. For ERF036 2162 out of the 2581 uniquely bound sequences were correctly identified as binding sequences, whereas only 416 out of the 6996 sequences bound uniquely by CBF4 were wrongly predicted as binding sequences (Figure 2C). In total, the number of false positive binding sequences dropped from 18,664 (11,668+6,996) to 1,384 (968+416), which is an improvement of 93%. Likewise, for CBF4 84% of uniquely bound sequences were correctly predicted and only 15% of sequences bound uniquely by ERF036 are predicted as false positives (Figure 2C). Here, the total improvement regarding false positives amounts to 86% (2,038/14,249). To identify the features which contribute specificity to each TF, we extracted feature importances using 'shapley additive explanations' (SHAP)[27] (Figure 2D). The outputs of the regressor models are influenced by different features. For ERF036 the slide at position -1 relative to the motif and the helix twist at position 5 in the motif is most influential, whereas for CBF4 the minor groove width at position 6 and the helix twist at position -1 contribute most to the decision of the random forest model. This observation underlines that the transcription factors, even though binding to the same core motif, are dependent on different peculiarities regarding the shape of the DNA (Figure 2). These results are not family specific since transcription factors of the NAC

family binding to the C(G/T)TNNNNNNAAG motif (Figure 2E-F), transcription factors of the WRKY family binding to TTGAC(T/C) motif, transcription factors of the bZIP family binding to ACGTCA motif and transcription factors of the C2H2 family binding to TTGCTNT motif show similar results (Supplemental Figure 6). In summary, the features defined by the shape-based regressor are able to explain differential binding of two transcription factors binding to the same sequence motif.

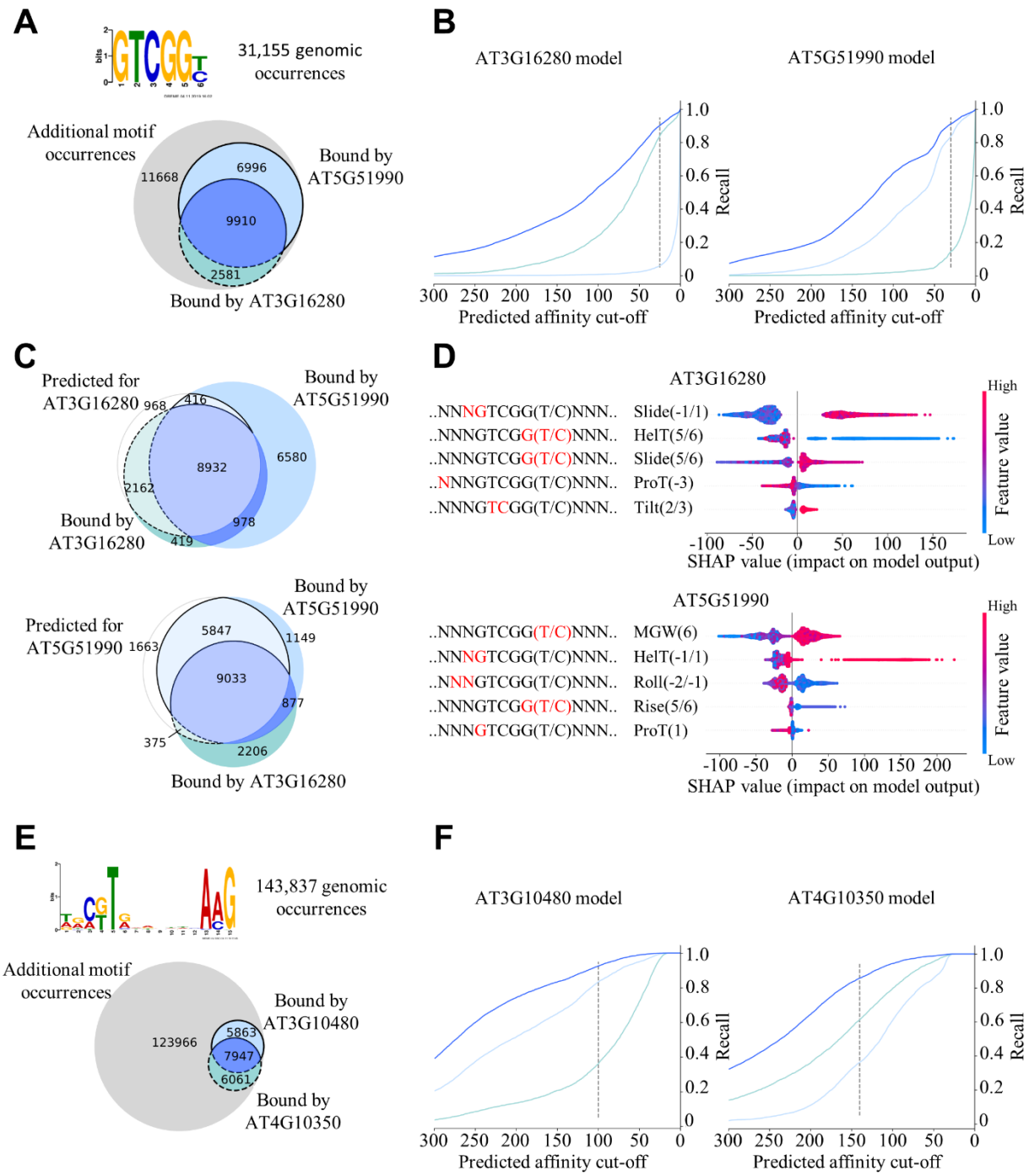


Figure 2: Differentiation of binding specificity of intra-familiar proteins with the same binding motif.

A, E) Occurrence of the GTCGG(T/C) and C(G/T)TNNNNNNAAG binding motifs in the A. thaliana genome sequence and the experimentally validated binding sequences of the AP2/EREBP TFs AT5G51990 and AT3G16280 and NAC TFs ANAC050 (AT3G10480) and BRN2 (AT4G10350). B, F) Performance of the random forest regressor trained on the genomic 3D shape. Each line represents the ratio of correctly predicted binding sites regarding all validated binding sites for different affinity prediction cut-offs. The dark blue line corresponds to binding sequences which are bound by both transcription factors and the light blue lines correspond to the uniquely bound binding sequences. C) The Venn diagrams show the sequence distributions according to the cut-off represented by the dashed line, respectively. Fields with light colours show the overlap of predicted and validated binding sequences. Dark coloured fields show the quantity of sequences, which were not predicted as bound by the model regarding the shown cut-off. (D) Influence of different local shape features on the prediction of the regressor model. The most influential features are at the top. Each row represents one shape feature at a single position within the sequence.

When used in a train-test scenario, the models generated by machine learning improve binding site prediction (Figure 1) and distinguish binding events for TFs binding the same motif (Figure 2). To test if the models are able to produce novel information they were used to predict TF binding to sequences not present in the genome. For the HY5 (AT5G11260) transcription factor of the bZIP family with the core motif ACGT, six DNA sequences with high (>150 peak height units) and low (<1 peak height units) regressor predictions were generated. For this purpose, 100,000 sequences not present in the genome of *A. thaliana* consisting of 18 bases with ACGT as core sequence were randomly created and the regressor model was applied. The binding affinity was experimentally tested by performing an electrophoretic mobility shift assay (EMSA) (Figure 3A, Supplemental Figure 7). As the binding control a sequence with high predicted affinity was taken from the host genome and competition was tested with the six unlabelled sequences of interest. Without any competitor, a labeled band is visible compared to the negative control (Figure 3A). As expected, the 80x competition control (cc) with the same sequence without biotin label outcompeted the labelled band. In the comparative competition experiment, using shapes with low regressor values (nb), all labelled bands are still visible. Those shapes are not able to out-compete the labelled shape and are thus apparently not bound by HY5. For the shapes with high regressor values (b), two out of three do not show any labelled band and are therefore bound by HY5. In total, we observed that 5 out of 6 predictions were experimentally validated. One sequence, which was predicted to bind (b*), was not able to out compete the labelled sequence. Given that the AUPRC for HY5 yielded 0.72 (Figure 3B,), the

validation of binding and non-binding events occurs within the expected error rate. To illustrate the subtle relevant factors, schematic models of the DNA sequences were plotted. The schematic models show visually obvious differences on the buckle at position +3 and the shear of position -1 between bound and not bound sequences (Figure 3C,). For HY5, important positions for binding extracted with SHAP are the helix twist at positions 5 and +1 and the opening at position -1. (Figure 3C, Supplemental Figure 8). The EMSA confirmed the predictive capability of the model constructed by machine learning.

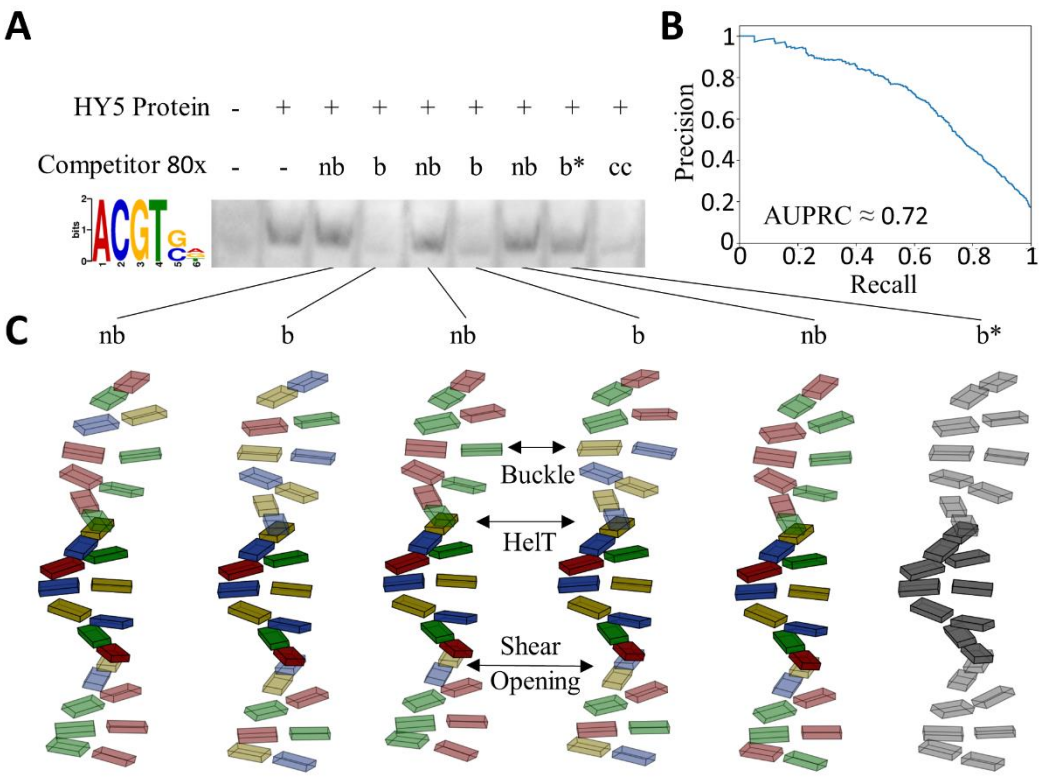


Figure 3: Experimental validation of shape based prediction for HY5 binding sequences.

A) Competition EMSA for six sequences containing the same motif with high and low binding affinity predictions based on their 3D structure. Sequences were predicted as binding (b) or non-binding (nb). Unexpected EMSA results are labelled with a *. An unlabelled competition control (cc) sequence was added as well. B) Precision-recall curve of the random forest model for the HY5 transcription factor. Precision and recall are based on the ampDAP *in vitro* verified binding sequences. C) Illustration of the 3D structure of the corresponding six sequences. The DNA backbone is not shown, as it is not yet possible to reliably calculate the spatial arrangement of the backbone.

Large experimental efforts have been designed to precisely assign binding sites to transcription factors[29] and to use this knowledge to describe transcriptional regulation[30]. The experiments (Figures 1-3) show that a combination of motif sequence and motif shape enables improved prediction of transcription factor binding on genomic sequence. The models generate a catalogue of potential binding sites in a genome and their predicted affinity. This information forms a base on which additional information layers (i.e. spacing of binding sites[31], chromatin openness[32], histone flags[32], and quantity of transcription factors and their interactors) can be stacked to enable prediction of gene expression. In synthetic biology, binding events for heterologously expressed transcription factors can be predicted more precisely and rationally designed promoter sequences are one step closer.

In the future, it will be critical to study evolutionary trajectories of transcriptional regulation to determine changes to binding sites present in genomes and changes to shape preferences of transcription factors. Precise understanding of transcription factor binding will allow us to build predictive regulatory networks and hence enable us to understand agriculturally important complex traits such as differential responses to heat, drought and pathogens and control of yield.

Methods

The DAP-seq and ampDAP-seq peak calling data were obtained from the Plant Cistrome Database (neomorph.salk.edu/dap_web/pages/index.php)[3]. Only datasets with a 'fraction of reads in peaks' (FRiP) value > 5% were considered for further analyses. All peak sequences were extracted from the *A. thaliana* reference genome sequence (TAIR10), obtained from ftp://ftp.arabidopsis.org/. The peak sequences were then used as input for the MEME-ChIP tool[33] to detect binding motifs. The motif with the lowest e-value was chosen as core motif for

each transcription factor. Peaks, which appeared in more than one third of all data sets, are considered as artifacts and were discarded.

To determine motif frequency in the genome, core motif occurrences were searched within the *A. thaliana* genome sequence using FIMO[34]. All motifs located within 80 base pairs of a peak summit were considered as experimental validated binding events. Multiple motif occurrences within this defined peak area were classified as homodimer binding sites to enable a more precise signal value interpretation. The calculation of the DNA shape was performed using a query table extracted from the Bioconductor package DNAshapeR[22,35].

The random forest classifier as well as the random forest regressor models were generated and trained using the python module scikit-learn[28]. Hyperparameter grid search and 5-fold cross validation were performed to generate each model. A detailed explanation of the data preprocessing and model generation is provided in the Supplemental Methods. Code is available from GitHub (Link not yet live, please contact us for pre-publication access).

The HY5 (AT5G11260) coding sequence was cloned with Gibson assembly in pFN19A HaloTag® T7 SP6 Flexi® Vector (Promega, Madison, Wisconsin, United States) in an N-terminal fusion with the Halo-tag. Plasmid-DNA was isolated with the ZymoPURE Plasmid Midiprep (ZymoGenetics, Seattle, Washington, United States). The HY5 protein was expressed with TnT® SP6 High-Yield Wheat Germ Protein Expression System (Promega, Madison, Wisconsin, United States) using 2 µg Plasmid DNA per 50 µL expression reaction. Expression was validated by Halo-tag detection (Supplemental Figure 9). Double stranded DNA sequences (20 µM) were generated by annealing synthesized DNA (98-21 °C, 9 hours) and diluted to 0.25 µM. The binding reaction was incubated for 2 h at 21°C. A 5% native polyacrylamide gel containing 0.5 TBE and 2.5% glycerol was prerun for 30 minutes. The samples were loaded with 1 µL Orange loading dye (Thermo Fisher Scientific, Waltham, Massachusetts, United States) and the gel (10 x 7.5 cm) was run at 80 V until the OrangeG front was 1 cm before the

end of the gel. The gel was blotted on a positively charged nylon membrane (Hybond^{TM}, GE Healthcare, Chicago, Illinois, United States) at fixed current of 0.8 mA/cm^2 for 90 minutes. The DNA was fixed by UV for 10 minutes. Biotin labeled DNA was detected with 1:5000 solution of an anti-biotin HPR-conjugated antibody (BioLegend, San Diego, California, United States) in TBST with 5% BSA. Detection was performed using Pierce^{TM} ECL Werstern Blotting Substrate (Thermo Fisher Scientific, Waltham, Massachusetts, United States) as described by the manufacturer and the imaging system Fusion Fx7(Vilber, Collégien, France).

References

1. Lin, M. & Guo, J. New insights into protein–DNA binding specificity from hydrogen bond based comparative study. *Nucleic Acids Research* **47**, 11103–11113 (2019).
2. Payne, J. L. & Wagner, A. The Robustness and Evolvability of Transcription Factor Binding Sites. *Science* **343**, 875–877 (2014).
3. O'Malley, R. C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**, 1280–1292 (2016).
4. Lehti-Shiu, M. D., Panchy, N., Wang, P., Uygun, S. & Shiu, S.-H. Diversity, expansion, and evolutionary novelty of plant DNA-binding transcription factor families. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1860**, 3–20 (2017).
5. Riechmann, J. L. *et al.* Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes. *Science* **290**, 2105–2110 (2000).
6. Bowman, J. L. *et al.* Insights into Land Plant Evolution Garnered from the Marchantia polymorpha Genome. *Cell* **171**, 287-304.e15 (2017).
7. Bailey-Serres, J., Parker, J. E., Ainsworth, E. A., Oldroyd, G. E. D. & Schroeder, J. I. Genetic strategies for improving crop yields. *Nature* **575**, 109–118 (2019).
8. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* gkz1001 (2019) doi:10.1093/nar/gkz1001.
9. Rushton, P. J., Somssich, I. E., Ringler, P. & Shen, Q. J. WRKY transcription factors. *Trends in Plant Science* **15**, 247–258 (2010).
10. Ülker, B. & Somssich, I. E. WRKY transcription factors: from DNA binding towards biological function. *Current Opinion in Plant Biology* **7**, 491–498 (2004).
11. Ciolkowski, I., Wanke, D., Birkenbihl, R. P. & Somssich, I. E. Studies on DNA-binding selectivity of WRKY transcription factors lend structural clues into WRKY-domain function. *Plant Mol Biol* **68**, 81–92 (2008).
12. Heim, M. A. The Basic Helix-Loop-Helix Transcription Factor Family in Plants: A Genome-Wide Study of Protein Structure and Functional Diversity. *Molecular Biology and Evolution* **20**, 735–747 (2003).
13. Foster, R., Izawa, T. & Chua, N. Plant bZIP proteins gather at ACGT elements. *FASEB j.* **8**, 192–200 (1994).
14. Jakoby, M. *et al.* bZIP transcription factors in Arabidopsis. *Trends in Plant Science* **7**, 106–111 (2002).

15. Chow, C.-N. *et al.* PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Research* **47**, D1155–D1163 (2019).
16. Burko, Y. *et al.* Chimeric Activators and Repressors Define HY5 Activity and Reveal a Light-Regulated Feedback Mechanism. *Plant Cell* **32**, 967–983 (2020).
17. Birkenbihl, R. P., Kracher, B., Roccaro, M. & Somssich, I. E. Induced Genome-Wide Binding of Three Arabidopsis WRKY Transcription Factors during Early MAMP-Triggered Immunity. *Plant Cell* **29**, 20–38 (2017).
18. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* **37**, W202–W208 (2009).
19. Yu, C.-P. *et al.* Transcriptome dynamics of developing maize leaves and genomewide prediction of *cis* elements and their cognate transcription factors. *Proc Natl Acad Sci USA* **112**, E2477–E2486 (2015).
20. Gao, F. *et al.* Blocking miR396 increases rice yield by shaping inflorescence architecture. *Nature Plants* **2**, 15196 (2016).
21. Dror, I., Golan, T., Levy, C., Rohs, R. & Mandel-Gutfreund, Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res.* **25**, 1268–1280 (2015).
22. Li, J. *et al.* Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Research* **45**, 12877–12887 (2017).
23. Chiu, T.-P., Xin, B., Markarian, N., Wang, Y. & Rohs, R. TFBSshape: an expanded motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Research* gkz970 (2019) doi:10.1093/nar/gkz970.
24. Rohs, R. *et al.* The role of DNA shape in protein–DNA recognition. *Nature* **461**, 1248–1253 (2009).
25. Abe, N. *et al.* Deconvolving the Recognition of DNA Shape from Sequence. *Cell* **161**, 307–318 (2015).
26. Gordân, R. *et al.* Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. *Cell Reports* **3**, 1093–1104 (2013).
27. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**, 56–67 (2020).
28. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12**, 2825–2830 (2011).
29. Ambrosini, G. *et al.* Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biol* **21**, 114 (2020).
30. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**, 831–838 (2015).
31. Freire-Rios, A. *et al.* Architecture of DNA elements mediating ARF transcription factor binding and auxin-responsive gene expression in *Arabidopsis*. *Proc Natl Acad Sci USA* 202009554 (2020) doi:10.1073/pnas.2009554117.
32. Lu, Z. *et al.* The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nat. Plants* **5**, 1250–1259 (2019).
33. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
34. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
35. Chiu, T.-P. *et al.* DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics* **32**, 1211–1213 (2016).

Funding

JS is funded by the Digital Infrastructure in the Life Sciences graduate school (Bielefeld University). DW is supported by core funding, Bielefeld University.

Acknowledgements

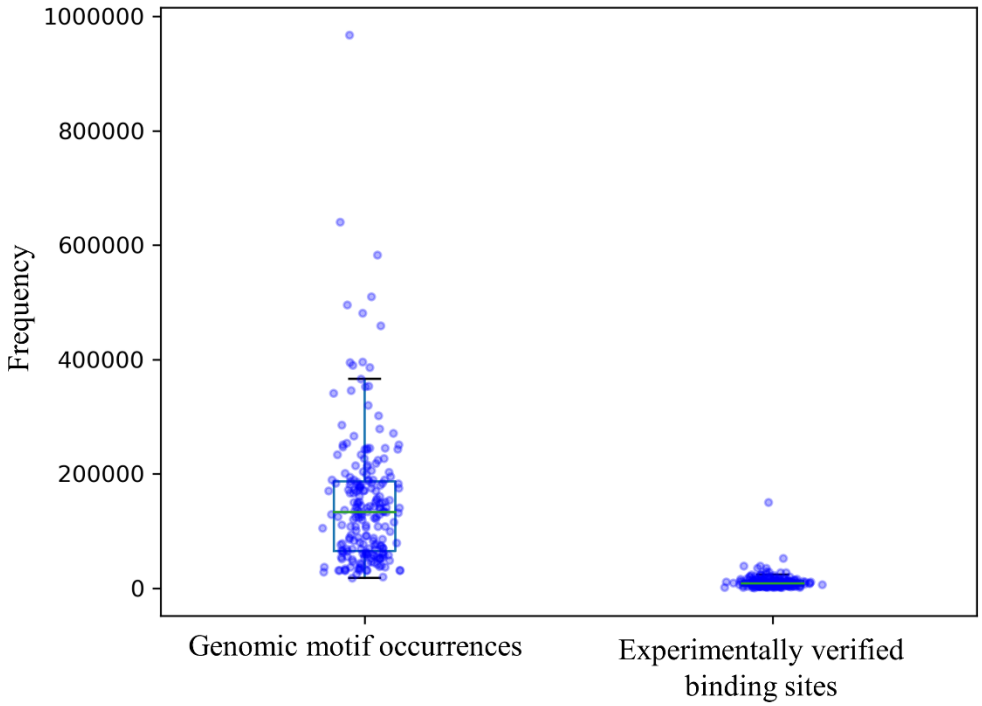
We thank the Bioinformatic Resource Facility team at the Center for Biotechnology (Bielefeld University) for technical support.

Author contributions:

JS designed and carried out the computational experiments including programming, interpreted the data, and co-wrote the manuscript, DW designed and carried out the wet lab experiments, interpreted data, and edited the manuscript, RS assisted with the wet lab experiments, interpreted data, and edited the manuscript, AB conceived the initial idea and the study, interpreted data, and co-wrote the manuscript

No competing interests are declared.

Supplementary Information is available for this paper



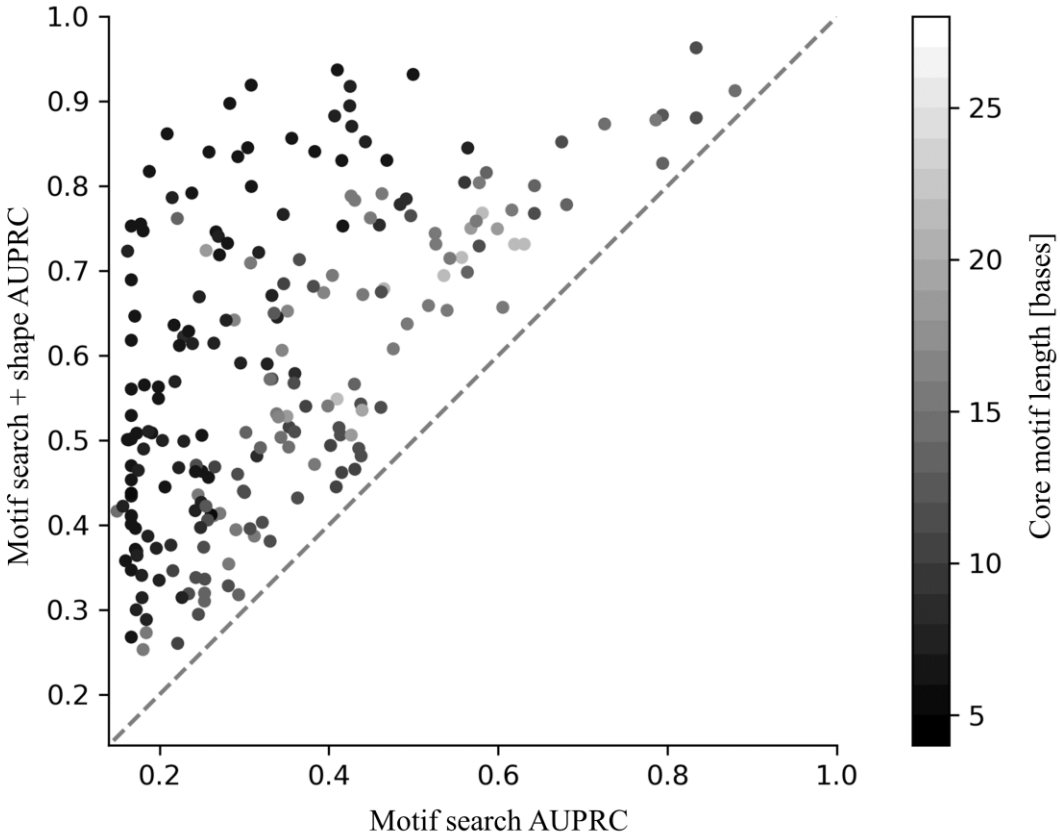
Supplemental Figure 1: Comparison of motif occurrences and verified binding events.

For the 217 tested transcription factors, actual binding events are on average 14-fold less frequent than binding sequence occurrences in the *A. thaliana* genome, for the respective transcription factor. The distribution on the left shows the number of individual motif occurrences for each transcription factor using the tool FIMO[1]. The distribution on the right shows the frequency of ampDAD-seq[2] verified binding events for each transcription factor.

References

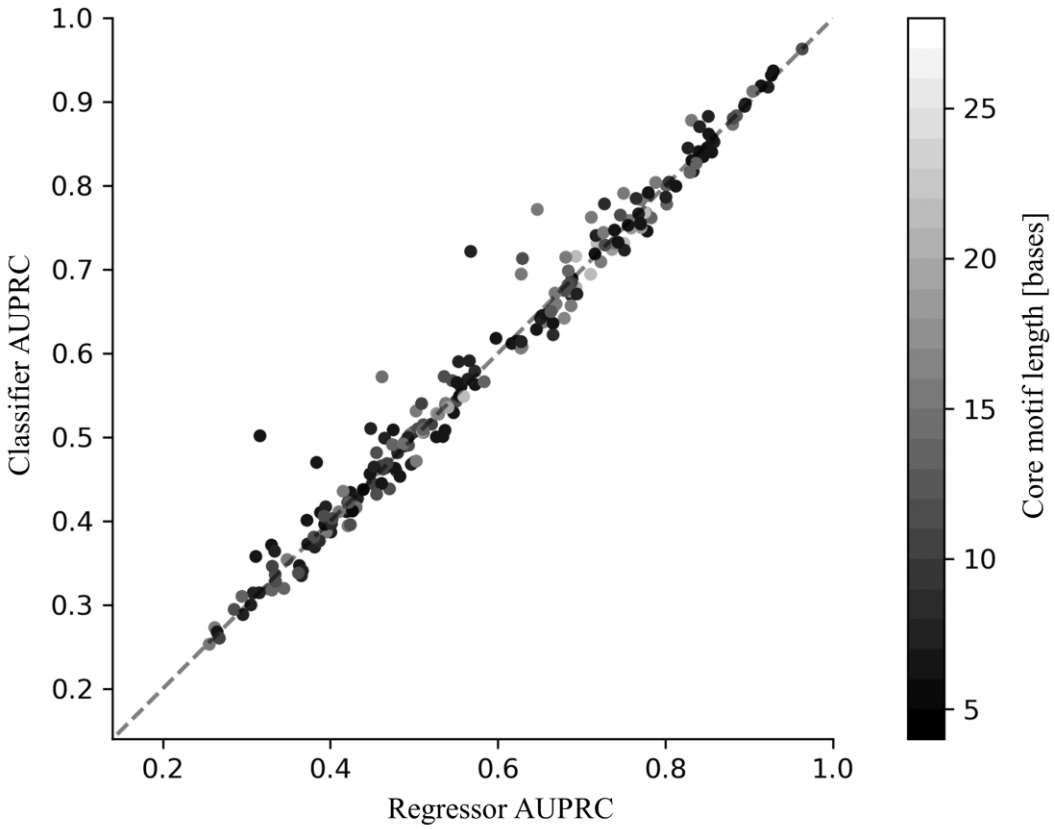
1. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

2. O'Malley, R. C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**, 1280–1292 (2016).



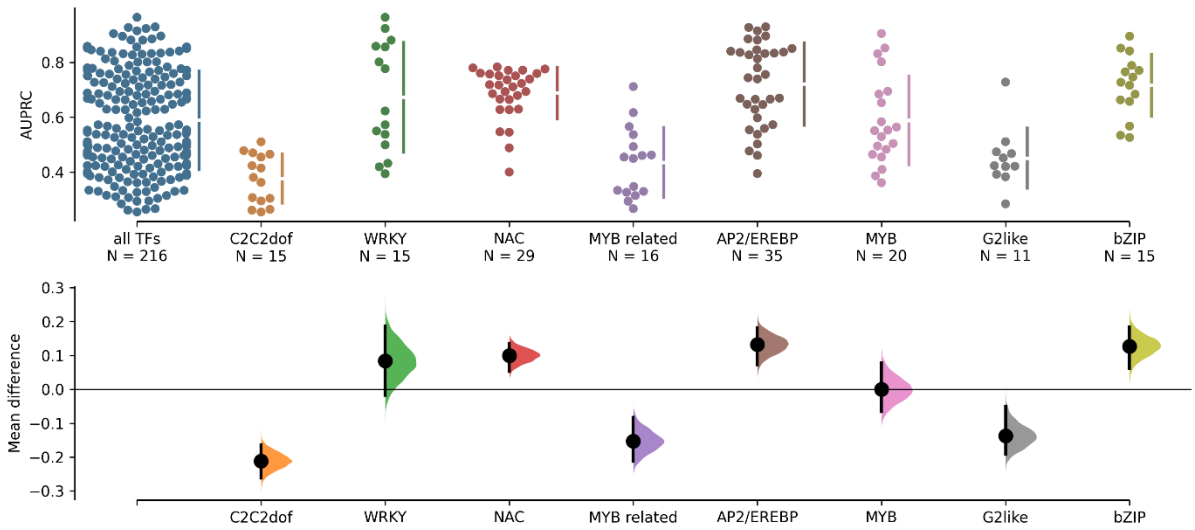
Supplemental Figure 2: Comparison between sequence based motif search and adding a shape based classifier.

The performance of identification of transcription factor binding sites improves for all tested transcription factors when adding shape information. Transcription factors with shorter sequence motifs, which are displayed as darker dots, tend to benefit most regarding the addition of DNA shape features.



Supplemental Figure 3: Comparison between random forest regressor and classifier.

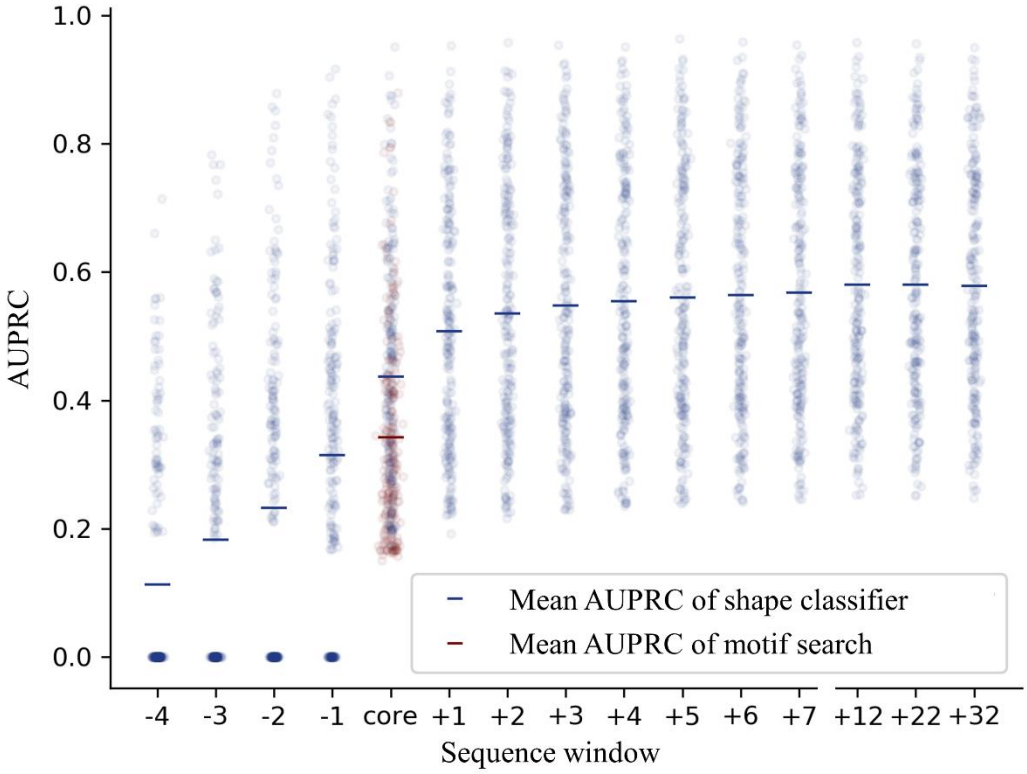
The binding site predictions of both approaches, represented by the AUPRC, are mostly similar. The classifier approach outperforms the regressor approach for specific transcription factors. However, the regression approach allows quantitative investigation of transcription factor binding events.



Supplemental Figure 4: Comparative analysis of family specific DNA-binding prediction performance.

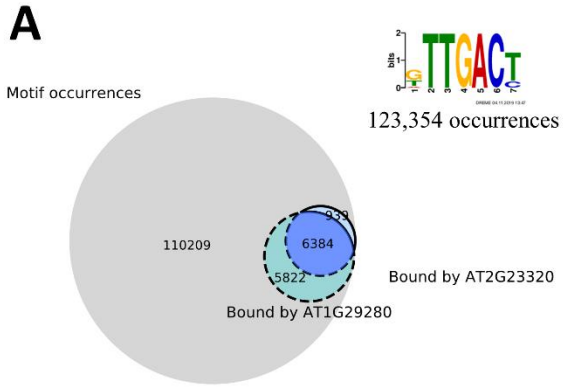
All TF families with more than ten members with ampDAP-seq data were analyzed to test if some TF families improve more than others when shape is incorporated into the analyses. The comparison is based on the area under the precision recall curve (AUPRC) using DAP-seq peaks as ground truth. Only families with available data of at least 10 members were investigated to ensure a robust analysis. For the families C2C2dof, MYB related and G2like a significantly ($p < 0.05$, Wilcoxon-Mann-Whitney-Test) worse performance of the random forest model with regard to the mean precision-recall relation was observed. For the transcription factor families NAC, AP2/EREBP and bZIP the performance was significantly ($p < 0.05$, Wilcoxon-Mann-Whitney-Test) better than mean improvement.

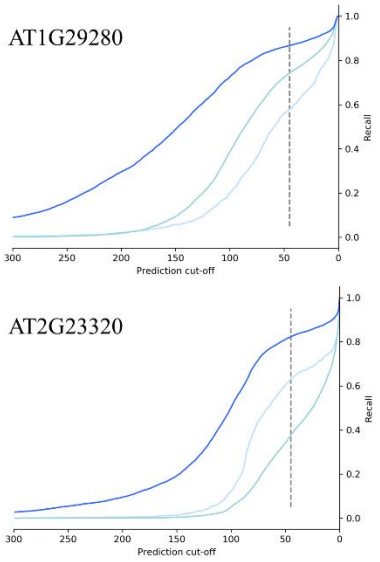
Control	test	pvalue_mann_whitney
all TFs	C2C2dof	0.000011
all TFs	WRKY	0.089488
all TFs	NAC	0.003943
all TFs	MYB related	0.000998
all TFs	AP2/EREBP	0.000065
all TFs	MYB	0.930433
all TFs	G2like	0.010457
all TFs	bZIP	0.006831

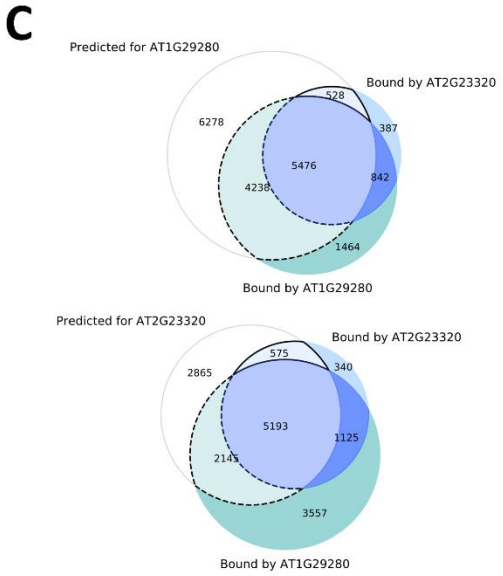


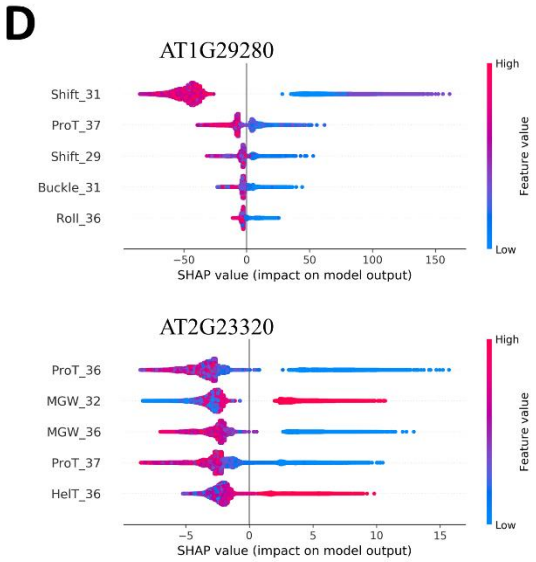
Supplemental Figure 5: Binding site prediction performance for differing sequence windows.

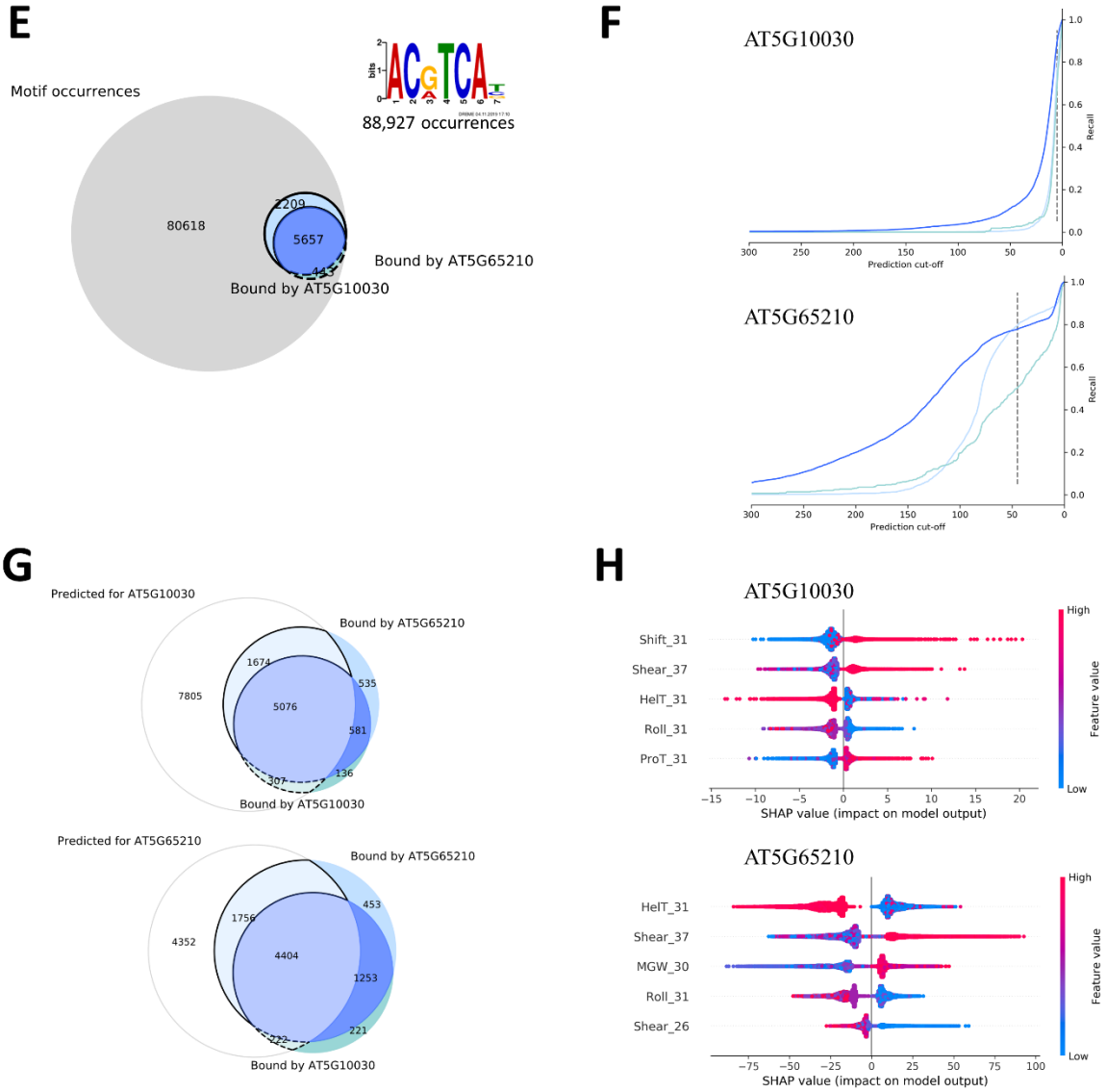
For each transcription factor the binding sequence window was varied up to 32 additional bases upstream and downstream from the core motif. The average AUPRC regarding the sequence search for the core motif is outperformed by calculating the DNA shape using only bases belonging to the core motif. Widening the sequence window with additional bases upstream and downstream from the core motif improves binding site prediction to an average AUPRC of approximately 0.6. The spread of prediction performance varies widely is dependent on each transcription factor.

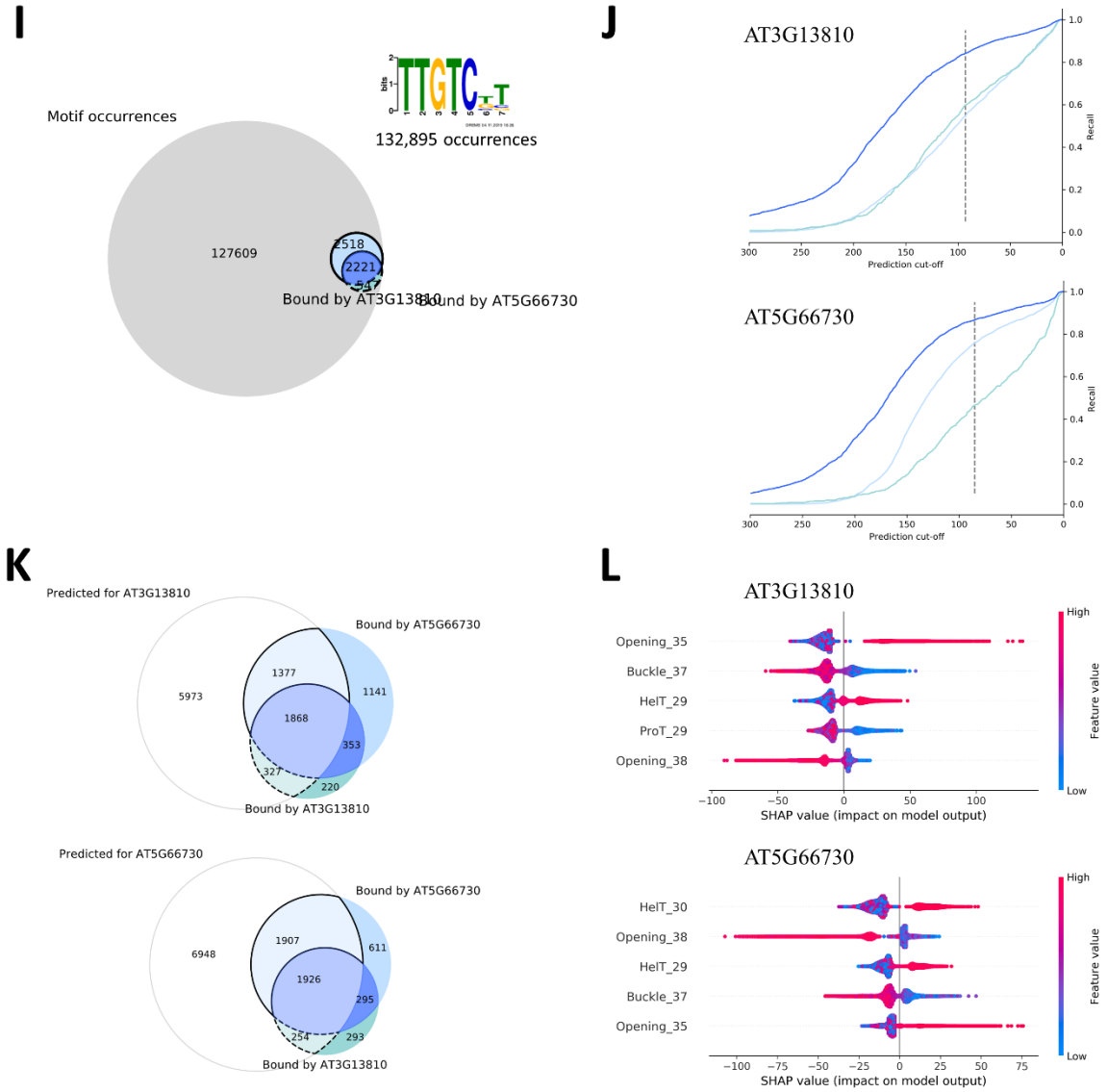

A
Motif occurrences
123,354 occurrences
110209
6384
5822
Bound by AT2G23320
Bound by AT1G29280


B
AT1G29280
AT2G23320
Recall
Prediction cut-off


C
Predicted for AT1G29280
Bound by AT2G23320
528
6278
387
5476
842
4238
1464
Bound by AT1G29280
Predicted for AT2G23320
Bound by AT2G23320
575
2865
340
5193
1125
2145
3557
Bound by AT1G29280

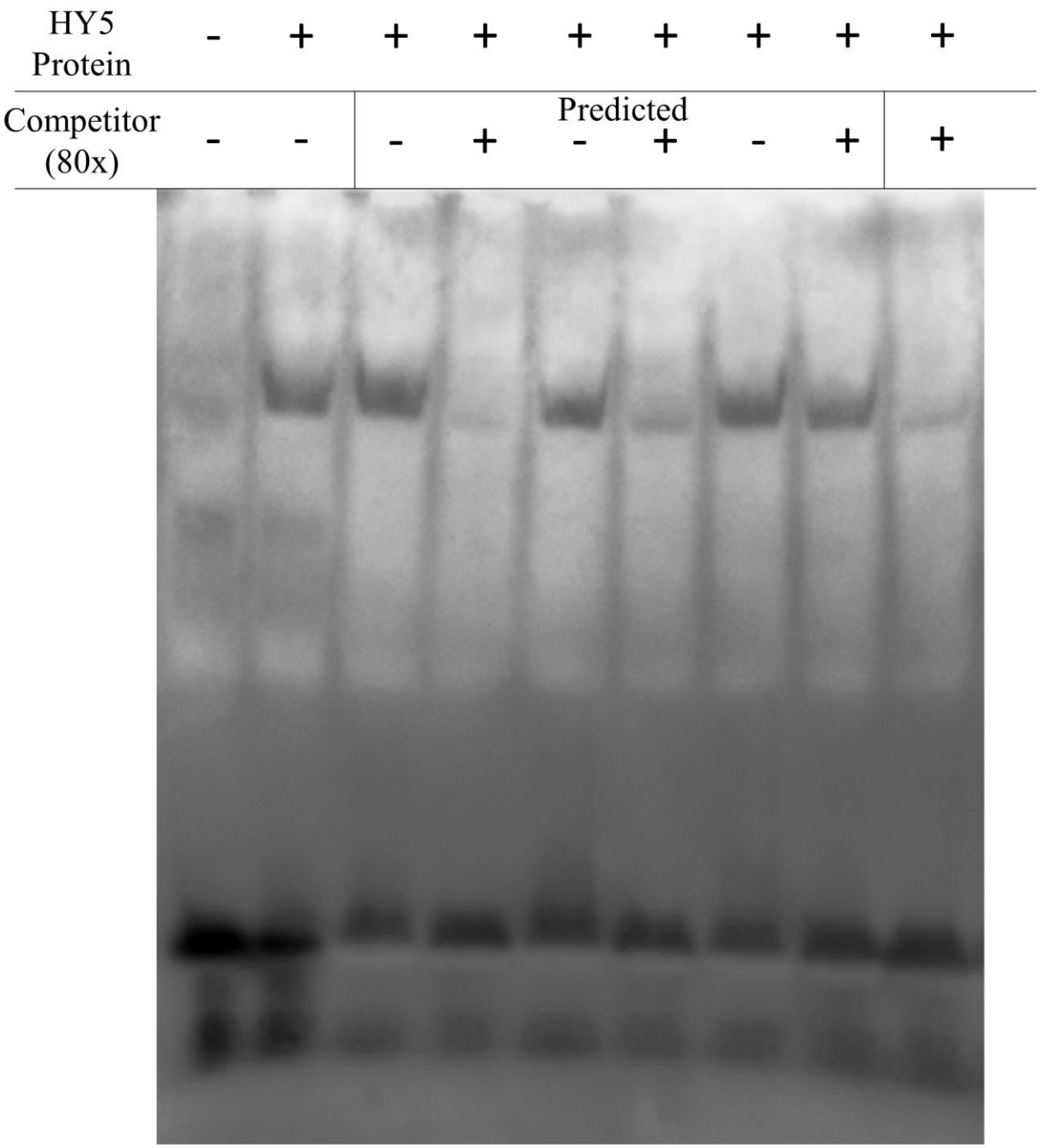

D
AT1G29280
Shift_31
ProT_37
Shift_29
Buckle_31
Roll_36
SHAP value (impact on model output)
Feature value
High
Low
AT2G23320
ProT_36
MGW_32
MGW_36
ProT_37
HelT_36
SHAP value (impact on model output)
Feature value
High
Low


E
Motif occurrences
88,927 occurrences
80618
2209
5657
443
Bound by AT5G65210
Bound by AT5G10030
F
AT5G10030
Recall
Prediction cut-off
AT5G65210
Recall
Prediction cut-off
G
Predicted for AT5G10030
Bound by AT5G65210
7805
1674
5076
535
581
136
307
Bound by AT5G10030
Predicted for AT5G65210
Bound by AT5G65210
4352
1756
4404
453
1253
221
272
Bound by AT5G10030
H
AT5G10030
Shift_31
Shear_37
HelT_31
Roll_31
ProT_31
SHAP value (impact on model output)
High
Feature value
Low
AT5G65210
HelT_31
Shear_37
MGW_30
Roll_31
Shear_26
SHAP value (impact on model output)
High
Feature value
Low



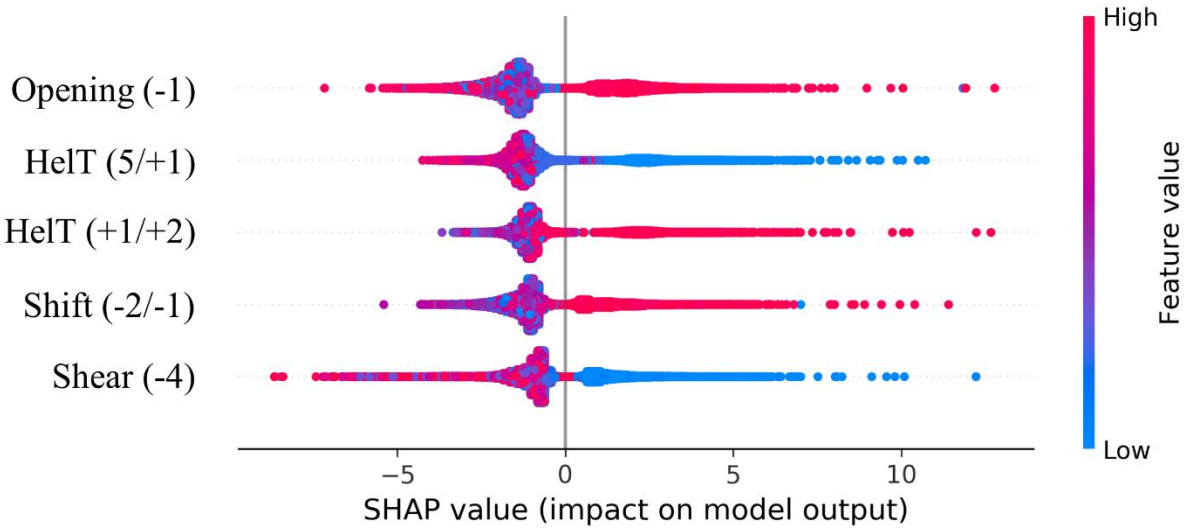
Supplemental Figure 6: Additional intra-familiar protein pairs with the same binding motif regarding differentiation of binding specificity.

A, E, I) Occurrence of the TTGAC(T/C), ACGTCA and TTGTC(T/G)T binding motifs in the *A. thaliana* genome sequence and the experimentally validated binding sequences of the WRKY TFs AT2G23320 and AT1G29280, bZIP TFs AT5G65210 and AT5G10030 and C2H2 TFs AT5G66730 and AT3G13810. B, F, J) Performance of the random forest regressor trained on the genomic 3D shape. C, F, K) The Venn diagrams show the sequence distributions according to the cut-off represented by the dashed line, respectively. Fields with light colours show the overlap of predicted and validated binding sequences. Dark coloured fields show the quantity of sequences, which were not predicted as bound by the model regarding the shown cut-off. D, G, L) Impact of different local shape features on the prediction of the regressor model. The most influential features are at the top. Each row represents one shape feature at a single position within the sequence. The start of the core motif is always at position 30.



Supplemental Figure 7: Competition EMSA with the *A. thaliana* transcription factor HY5.

For experimental validation of the regressor predictions, DNA sequence with a high and low regressor prediction containing the same sequence motif not present in the genome of *A. thaliana* were generated. One sequence with a high prediction was labeled with biotin to detect DNA binding of HY5 by performing an EMSA. To compare binding affinities three sequences with high and low regressor predictions were used as competitors for the labeled sequence containing the same sequence motif. A shifted band is visible for all samples with low binding affinity prediction, one sample with high affinity prediction and the control without competitor sequence. Less visible shifted bands are observed in two samples with high predicted affinity and the control with the same competitor sequence.

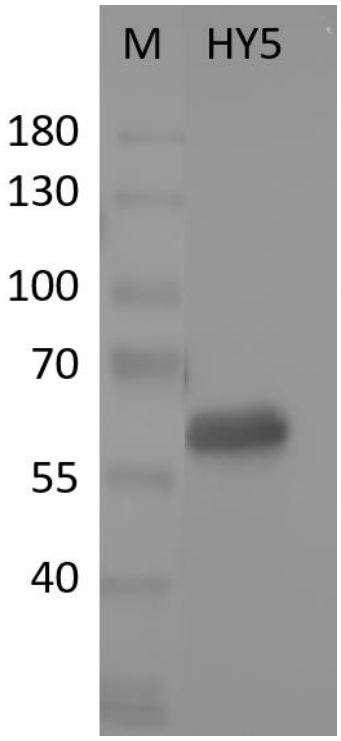


Supplemental Figure 8: Top 5 most important shape features for HY5 binding.

The SHAP Python package[1] was used to extract the most important features of the random forest model trained on experimentally validated HY5 binding sequences. The most important shape feature is the Opening at the -1 position, which is one base upstream of the motif sequence.

References

1. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**, 56–67 (2020).



Supplemental Figure 9: Western blot of Protein-Halo-Tag fusion.

To confirm the expression of the N-terminal fusion of HY5 and the Halo tag a SDS-Page with a 10% acrylamide gel followed by a western blot on a 0,22 µm Nitrocellulose membrane (Sartorius, Göttingen, Germany) was performed. The HY5 protein was expressed with TnT® SP6 High-Yield Wheat Germ Protein Expression System (Promega, Madison, Wisconsin, United States) using 2 µg Plasmid DNA per 50 µL expression reaction (HY5). Wheatgerm extract without plasmid DNA was used as a negative control (-). For size comparison the PageRuler™ (M) (Thermo Fisher Scientific, Waltham, Massachusetts, United States) was used. The Anti-HaloTag® Monoclonal Antibody (Promega, Madison, Wisconsin, United States) and an HRP conjugated Anti-mouse antibody (abcam, Cambridge, United Kingdom) were used to detect the fusion protein. Detection was performed using Pierce™ ECL Western Blotting Substrate (Thermo Fisher Scientific, Waltham, Massachusetts, United States) and the imaging system Fusion Fx7(Vilber, Collégien, France). The expression reaction with the plasmid DNA shows a strong band with a size of approximately 60 kDa which is consistent with the expectations, as the Halo Tag adds 33 kDA to the protein.

Pre-processing and training of the random forest regressor

To perform data pre-processing and training of a random forest model an ampDAP-seq peak file (from http://neomorph.salk.edu/dap_web/pages/browse_table_aj.php) and the Arabidopsis thaliana genome (from arabidopsis.org) is necessary. Only peak files with FRiP value > 5% were considered and after motif extraction each peak file was filtered for peaks that appear in less then 66% of all ampDAP-seq available peak files, as those peaks were considered artifacts due to the ampDAP-seq procedure. Peak regions were extracted from the Arabidopsis genome using a custom python script, which expects one peak file (-p) and the corresponding genome (-g) as input. The resulting fasta file with genomic peak regions was used as input for the MEME-ChIP[1] (MEME-suite[2] v 5.0) tool with default parameters, so that the only given parameters were an output folder (-oc) and the peak regions fasta file (-dna). The sequence motif with highest e-value (--motif 1) from the resulting combined.meme file was searched in the *A. thaliana* genome using the FIMO[3] (MEME-suite[2] v 5.0) tool with a cut-off (--thresh) of 5e-4. To ensure that no matches were, discarded the maximum number of stored matches (--max-stored-scores) were set to 1000000. The parameter --max-strand was set to 1 so that palindromic sequences would not match two times in the same locations and an output folder (--oc) was declared.

To allow a more accurate interpretation of binding affinities, the areas of motif matches were scanned for multiple motif occurrences using a custom python script. For this, each peak, which always has a length 200 base pairs, was tested for multiple FIMO matches. If more peaks had multiple motif occurrences than single motif occurrences, the corresponding transcription factor was considered for homodimeric binding events and vice versa. In that case, the random forest regressor was only trained on those peaks with multiple motif occurrences.

All genomic locations with sequence motif matches were translated into 13 DNA shape features using a publicly available query table[4], which was implemented into a custom Python script. Chloroplast and mitochondrial motif occurrences were discarded, as those sequences were not part of the in vitro experiment. The sequence window, for which the DNA shape was calculated, was set to 32 additional bases upstream and downstream from the sequence motif. Experimentally recorded signal values were normalized to range from 0 to 1000 using sklearns pre-processing module[5]. Since some shape features are mirrored for palindromic sequences, each matched sequence window on the minus strand was reverse complemented, so that the matrix of 3D shape values always correspond to the same direction.

DNA shape based training to learn protein binding affinities was performed with the RandomForestRegressor class from the sklearn module. The number of considered positions upstream and downstream of the sequence motif can be specified using the custom Python script by setting the -b parameter, which has a default value of 4. If the ratio of experimentally verified binding sites and genomic binding site occurrences was too high (> 1:5), this ratio was forced to be 1 to 5 by discarding random genomic positions. If this procedure would still yield more than 120,000 locations, the ratio was forced to be 1 to 3. This ratio between validated binding sites and binding site occurrences was always calculated for each transcription factor and used as sample weight for training the random forest regressor to prevent bias towards false negatives. The whole dataset was split into a train set (80%) and a validation set (20%) using the train_test_split function provided by the sklearn module, applying stratification to ensure even distributions of validated binding sites in the train and test set. The train set was used for 5-fold cross validation learning and the validation set was used for evaluation. To finetune the learning process, randomized hyperparameter grid search was performed for 75 iterations including the parameters "n-estimators" (ranging from 10 to 200), "max_features" (auto, square root or log2) and "max_depth" (ranging from 4 to 12). Additionally, 5-fold cross validation was performed, leading to a total number of 375 training procedures for each respective transcription factor. The mean squared error was used as loss. For evaluation purposes the precision recall curve function, which is also provided by the sklearn module, was applied on the validation data. After training the model, sequences of interest can be checked for putative binding affinity.

References

1. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
2. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* **37**, W202–W208 (2009).
3. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
4. Li, J. *et al.* Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Research* **45**, 12877–12887 (2017).
5. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12**, 2825–2830 (2011).