nature genetics

ARTICLES

https://doi.org/10.1038/s41588-022-01048-5

DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers

Bernardo P. de Almeida[1,2], Franziska Reiter[1,2], Michaela Pagani[1] and Alexander Stark[1,3] ✉

Enhancer sequences control gene expression and comprise binding sites (motifs) for different transcription factors (TFs). Despite extensive genetic and computational studies, the relationship between DNA sequence and regulatory activity is poorly understood, and de novo enhancer design has been challenging. Here, we built a deep-learning model, DeepSTARR, to quantitatively predict the activities of thousands of developmental and housekeeping enhancers directly from DNA sequence in *Drosophila melanogaster* S2 cells. The model learned relevant TF motifs and higher-order syntax rules, including functionally nonequivalent instances of the same TF motif that are determined by motif-flanking sequence and intermotif distances. We validated these rules experimentally and demonstrated that they can be generalized to humans by testing more than 40,000 wildtype and mutant *Drosophila* and human enhancers. Finally, we designed and functionally validated synthetic enhancers with desired activities de novo.

Enhancers[1] are genomic elements that regulate the cell-type-specific transcription of target genes, thereby controlling animal development and physiology[2]. A feature of enhancers is their ability to activate transcription outside their endogenous genomic contexts[3], which suggests that all the necessary cis-regulatory information is contained within the enhancers' DNA sequences. Indeed, enhancer sequence mutations can drastically alter enhancer function and are associated with developmental defects[2], morphological evolution[4], and human disease[5].

Enhancers typically contain multiple sequence motifs that are binding sites for sequence-specific TFs[6]. Understanding how motifs and their arrangements (their number, order, orientation and spacing – termed here collectively 'motif syntax') relate to enhancer function has remained one of the most important open questions in modern biology. Systematic mutagenesis of various individual enhancers has revealed a complex picture, whereby changing nucleotides or altering motif syntax affected the function of some enhancers but not others[7–27]. These contradictory observations have made it difficult to define the relationship between enhancer sequence and function[18,28].

Many computational approaches have sought to predict enhancer activities from DNA sequences using local DNA features, for example motif dictionaries or de novo *k*-mers, and selected syntax rules in various thermodynamic or machine-learning frameworks[16,17,27,29–40]. Despite remarkable success, these approaches did not reveal how the motif syntax elements collaborate to determine enhancer activity. In addition, they did not consider the mutual compatibilities between certain enhancer- and promoter-types recently reported for different transcriptional programs[41–43]. Thus, quantitatively predicting the regulatory activity of enhancers and the de novo design of synthetic enhancers have remained open challenges for decades.

Previous approaches typically modeled enhancer sequences explicitly via predefined sets of features, which were informed by previous biological knowledge[44]. In contrast, deep learning, in particular convolutional neural networks, does not require previous knowledge and can learn accurate models directly from raw data[45–55]. Once trained on raw data, these models allow the extraction and interpretation of the learned rules by new types of tools[45–47,49,50,56–61]. For example, when applied to ChIP-nexus data that measures TF-binding genome wide at high resolution, a convolutional neural network was able to learn motifs and syntax rules for cooperative TF binding[49]. Similarly, this approach was used to model DNA accessibility[46–48,50,52,53,55], transcriptional reporter activities[62] and predict genetic variant effects[54]. Nevertheless, a model to quantitatively predict enhancer activities solely from DNA sequence in a single cell type, and its interpretation to reveal and validate specific cis-regulatory rules are still missing.

Here, we built a deep-learning model—DeepSTARR—to predict enhancer activity towards two promoters from the distinct developmental and housekeeping transcriptional programs in *D. melanogaster* S2 cells directly from the DNA sequence. For both programs, DeepSTARR predicts enhancer activity quantitatively for unseen sequences and reveals different coding features for the two programs, including specific TF motifs that we validate experimentally. We further extract motif syntax rules, including favorable and unfavorable sequence contexts and intermotif distances, which are predictive of enhancer activity in *Drosophila* and can be adjusted to human enhancers, as we validate experimentally by high-throughput mutagenesis of thousands of enhancers and enhancer variants. These rules allowed the design of synthetic enhancers with desired activity levels de novo.

Results

DeepSTARR predicts enhancer activity from DNA sequence. To learn the cis-regulatory information encoded in enhancer sequences in an unbiased way, we developed a deep-learning model called DeepSTARR that predicts enhancer activity directly from DNA sequence. First, we used UMI-STARR-seq[63,64] to generate

[1]Research Institute of Molecular Pathology, Vienna BioCenter, Campus-Vienna-BioCenter 1, Vienna, Austria. [2]Vienna BioCenter PhD Program, Doctoral School of the University of Vienna and Medical University of Vienna, Vienna, Austria. [3]Medical University of Vienna, Vienna BioCenter, Vienna, Austria. ✉e-mail: stark@starklab.org

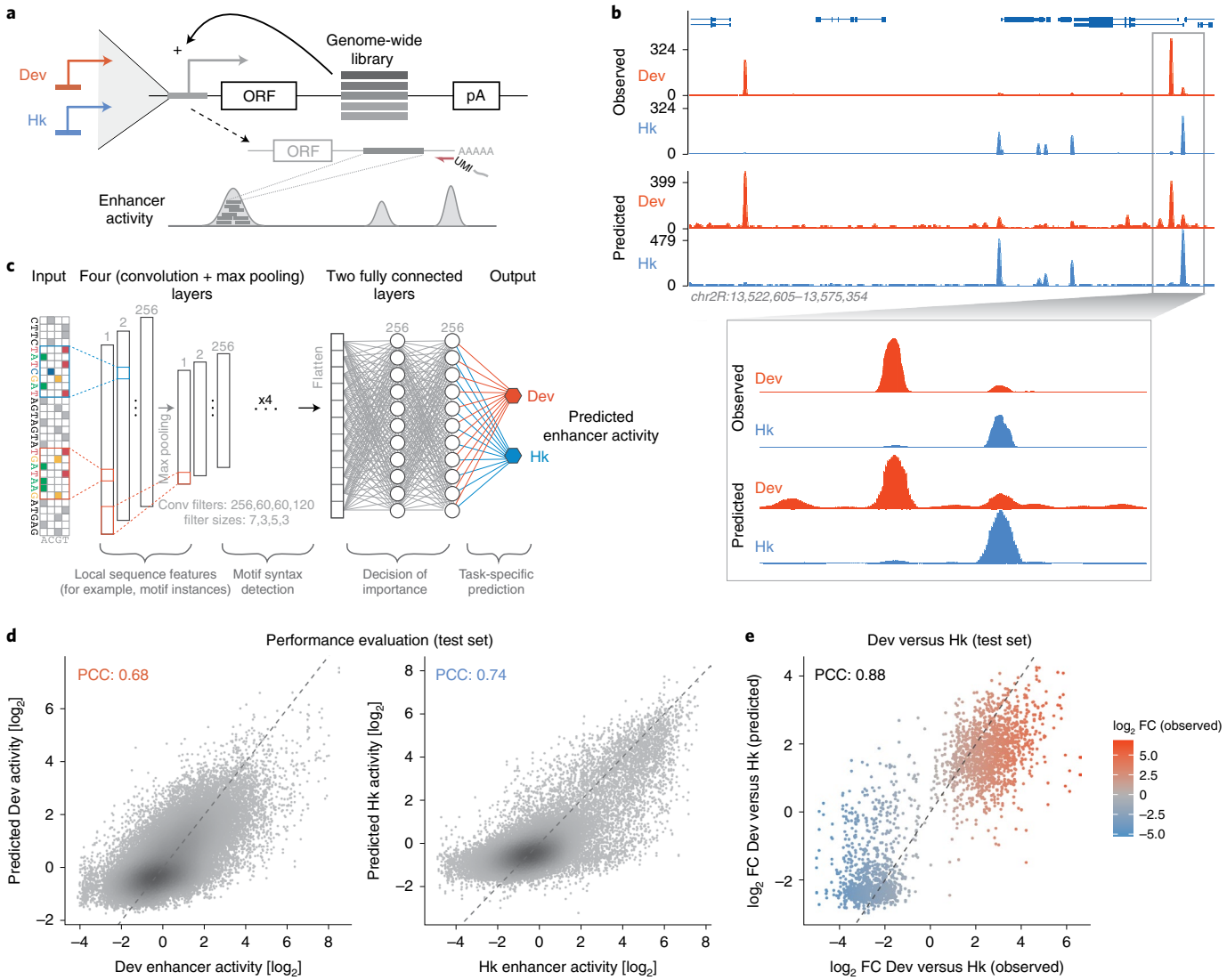

Fig. 1 | DeepSTARR quantitatively predicts enhancer activity genome wide from DNA sequence. a, Schematics of genome-wide UMI-STARR-seq using developmental (Dev) (DSCP; red) and housekeeping (Hk) (RpS12; blue) promoters. **b**, DeepSTARR predicts enhancer activity genome wide. Genome browser screenshot depicting observed and predicted UMI-STARR-seq profiles for both promoters for a locus on the held-out test chromosome (Chr) 2R. **c**, Architecture of the multitask convolutional neural network DeepSTARR that was trained to simultaneously predict quantitative Dev and Hk enhancer activities from 249-bp DNA sequences. **d**, DeepSTARR predicts enhancer activity quantitatively. Scatter plots of predicted versus observed Dev (left) and Hk (right) enhancer activity signal across all DNA sequences in the test set chromosome. Color reflects point density. **e**, DeepSTARR quantitatively predicts Dev and Hk enhancer–promoter specificity. Predicted versus observed log_2FC between Dev and Hk activity for all enhancer sequences in the test set chromosome. PCC, Pearson correlation coefficient.

genome-wide high-resolution quantitative activity maps of developmental and housekeeping enhancers, representing the two main transcriptional programs in *Drosophila* S2 cells[41–43] (Fig. 1a). We identified 11,658 developmental and 7,062 housekeeping enhancers (Fig. 1b and Supplementary Fig. 1a,b). These enhancers are largely nonoverlapping, confirming the specificity of the different transcriptional programs. These genome-wide enhancer activity maps provide a high-quality dataset to build predictive models of enhancer activity and characterize the sequence determinants of two main enhancer types.

We built the multitask convolutional neural network DeepSTARR to map 249-bp-long DNA sequences tiled across the genome to both their developmental and their housekeeping enhancer activities (Fig. 1c). We adapted the Basset convolutional neural network architecture[46] and designed DeepSTARR with four convolution layers, each followed by a max-pooling layer, and two fully connected layers (Fig. 1c and Supplementary Fig. 2; Methods). The convolution layers identify local sequence features (for example, TF motifs) and increasingly complex patterns (for example, TF motif syntax), whereas the fully connected layers combine these features and patterns to predict enhancer activity separately for each enhancer type.

We evaluated the predictive performance of DeepSTARR on a held-out test chromosome. The predicted and observed enhancer activity profiles were highly similar for both developmental (Pearson correlation coefficient (PCC) = 0.68) and housekeeping (PCC = 0.74) enhancers (Fig. 1b,d and Supplementary Figs. 1, 3 and 4). This performance is close to the concordance between experimental replicates (PCC = 0.73 and 0.76, respectively; Supplementary Fig. 1c), suggesting that the model accurately captures the regulatory

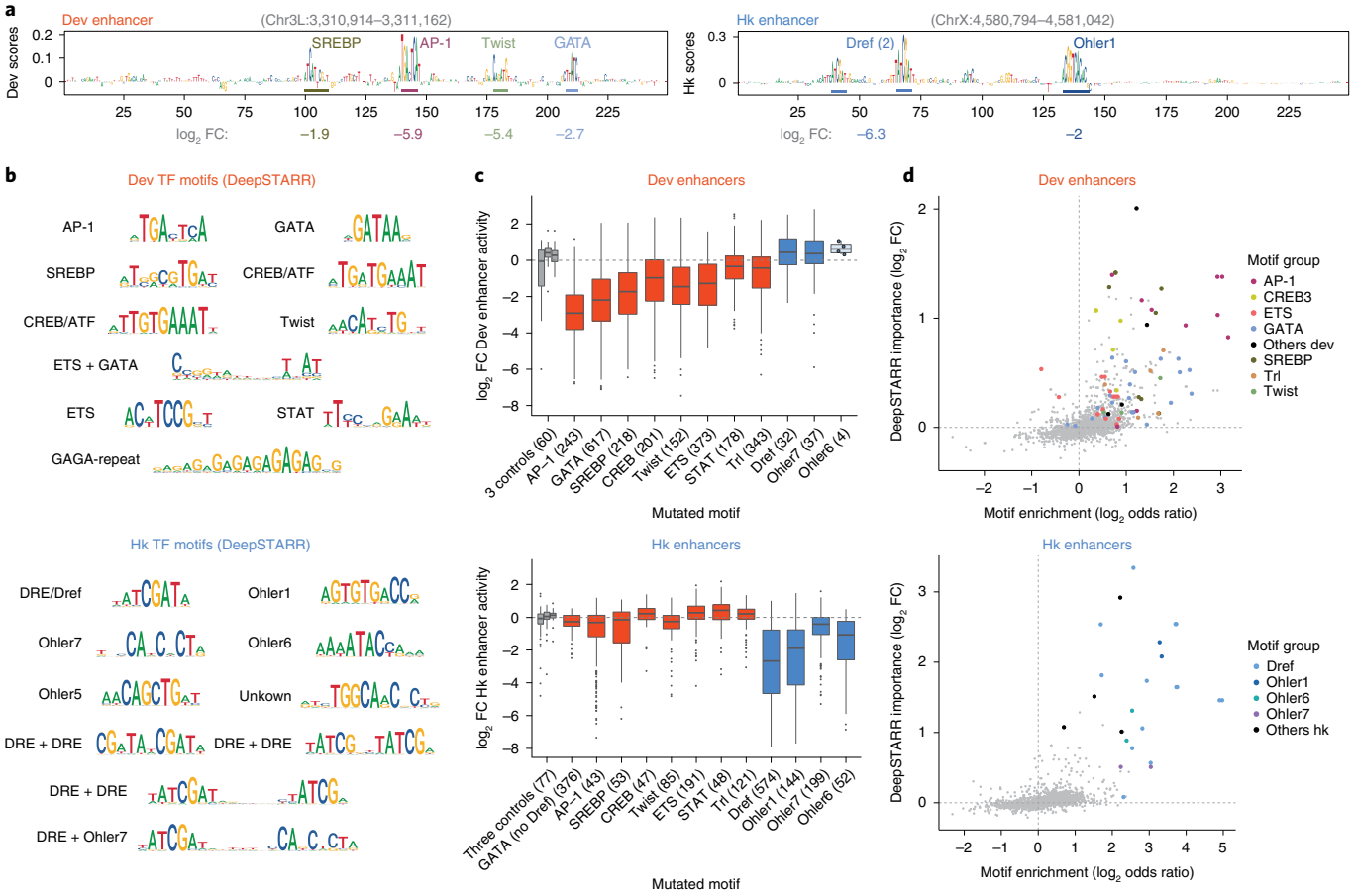

Fig. 2 | DeepSTARR reveals important TF motif types that validate experimentally. a, DeepSTARR-derived Dev and Hk nucleotide contribution scores for strong Dev (left) and Hk (right) enhancer sequences, respectively. Regions with high scores resembling known TF motifs are highlighted. log_2FC values (bottom) indicate the impact on enhancer activity of mutating all instances of each motif type. **b**, DeepSTARR motifs, generated by TF-Modisco by summarizing recurring predictive sequence patterns from the sequences of all Dev (top) and Hk (bottom) enhancers. **c**, Dev and Hk TF motifs are specifically required for the respective enhancer types. Enhancer activity changes (log_2 FC) for Dev (top) and Hk (bottom) enhancers after mutating all instances of three control motifs (gray), eight predicted Dev motifs (AP-1, GATA, SREBP, CREB, twist, ETS, STAT, Trl; red) and four predicted Hk motifs (Dref, Ohler1, Ohler7, Ohler6; blue). Number of enhancers mutated for each motif type are shown. The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers); outliers are shown individually. **d**, DeepSTARR discovers important TF motifs not obvious by motif enrichment. Comparison between motif enrichment on all active Dev (top) and Hk (bottom) enhancers (log_2 odds ratio, from Supplementary Fig. 7f; *x* axis) and DeepSTARR's predicted global importance (*y* axis) for all representative TF motifs. Important motifs for each enhancer type are highlighted.

information present in the sequences and the differences between developmental and housekeeping enhancers (Fig. 1e). DeepSTARR performed better than methods based on known TF motifs or unbiased *k*-mer counts[35], both at predicting continuous enhancer activity and at binary classification of enhancer sequences (Supplementary Figs. 1d–f and 4). Thus, DeepSTARR learned generalizable features and rules de novo directly from the DNA sequence that allow the prediction of enhancer activities for unseen sequences.

DeepSTARR reveals TF motifs required for enhancer activity. To understand the features and rules learned by DeepSTARR, we quantified how each individual nucleotide in every sequence contributes to the predicted developmental and housekeeping enhancer activities[49,57,65,66] (Fig. 2a). These predicted contributions agreed well with experimental scanning mutagenesis of five different enhancers (average PCC: 0.73; Supplementary Fig. 5). We next consolidated recurrent highly scoring sequence patterns into motifs[58] (Fig. 2b and Supplementary Fig. 6; Methods). This uncovered distinct motifs of activating TFs that are known to occur in developmental and housekeeping enhancers[27,41], thus validating the approach and reinforcing the mutual incompatibility of the two transcriptional programs (Fig. 2a,b and Supplementary Fig. 7). In addition, motif instances of repressive TFs received negative weights (Supplementary Fig. 8), indicative of the repressive functions of these TFs and the relative underrepresentation of these motifs in active enhancers (Supplementary Fig. 7f).

We tested the requirements of select activator TF motifs for enhancer activity experimentally across hundreds of enhancers by performing large-scale motif mutagenesis (4,960 motif mutations in 856 developmental and 1,041 housekeeping enhancers; Fig. 2c and Supplementary Figs. 9 and 10). Consistent with their predicted importance, mutating eight developmental motifs (AP-1, GATA, SREBP, CREB, twist, ETS, STAT, Trl) substantially reduced the activity of developmental, but not housekeeping, enhancers, with AP-1 and GATA motifs being most important, as predicted by DeepSTARR. In contrast, mutating four housekeeping motifs (Dref, Ohler1, Ohler6, Ohler7) affected only housekeeping but not developmental enhancers and mutating three control motifs (length-matched random motifs to control for enhancer sequence perturbation) did not have any impact (Fig. 2c).

Interestingly, the motifs learned by DeepSTARR were not restricted to highly enriched motifs but included other motifs such as SREBP, CREB and ETS motifs that, on their own, were not or only weakly overrepresented in S2 developmental enhancers. These motifs could therefore not have been found by methods based on over-representation (Fig. 2b,d) and they might contribute to TF binding and enhancer activity only in combination with other motifs and TFs[22,67]. Despite being less enriched, these motifs were important for enhancer activity (Fig. 2c), and, even for more abundant motifs, motif enrichment was not always predictive of motif importance (Fig. 2d and Supplementary Fig. 11; Methods and ref. [60]). Overall, these results demonstrate that DeepSTARR can discover both abundant motifs but also motifs that are relatively rare in enhancers but still important for enhancer activity, and score their specific importance for developmental and housekeeping enhancers.

Nonequivalent instances of the same TF motif. Since enhancers often contain several instances of the same motif type, we next assessed the contribution of each individual instance of the GATA, AP-1, twist, Trl and Dref motifs by DeepSTARR (Supplementary Fig. 12a) and by experimental mutagenesis (Supplementary Figs. 9a and 12b). Unexpectedly, individual instances of the same motif were frequently predicted and experimentally validated to have varying degrees of contributions to enhancer activities (defined here as nonequivalency), both across different enhancers and within the same enhancer (Fig. 3a–c and Supplementary Fig. 12).

The enhancer shown in Fig. 3a for example contains three GATA instances with very different contributions as predicted and determined experimentally: the second instance is the most important, followed by the first and the third. The agreement between predictions and experiments holds across all 1,013 GATA instances tested (PCC = 0.53; Fig. 3b) and the nonequivalency of motif instances is widespread: 57% of enhancers with several instances had motifs with greater than twofold and 70% with greater than 1.5-fold differences (Fig. 3c). These differences are not well captured by existing position weight matrix (PWM) motif scores (Fig. 3d and Supplementary Fig. 13), suggesting that the importance of motif instances depends on complex sequence features outside the core motif. Indeed, PWM models performed worse than linear models based on predefined motif syntax features or the gkm-SVM models (Supplementary Fig. 13). The observation that different instances of the same motif type can have vastly different contributions to enhancer activity (despite the instances' identical sequences) is an important underappreciated phenomenon that complicates our understanding of enhancer sequences and noncoding variants (Discussion).

Flanking sequence influences the importance of TF motifs. To explore the syntax features that affect the importance of a motif instance, we examined the motif-flanking nucleotides that can contribute to enhancer activity[12,13,18,37,68–72]. Indeed, DeepSTARR predicted significant contribution for the flanking sequences of important motifs up to ten or more nucleotides (Fig. 4a and Supplementary Fig. 14). For each motif type, we then sorted all instances by their predicted importance to determine the optimal flank length and sequence (Fig. 4a,b and Supplementary Fig. 15). For example, important GATAA sequences had a G at position +1, whereas nonimportant ones had a T at position +1 and a G at position −1 (Fig. 4b). In contrast, up to 5 bp flanking up- and downstream affected the importance of Trl instances, with flanking GA-repeats correlating with increased importance (Fig. 4b). The flanks of high and low importance motif instances predicted by DeepSTARR were largely concordant with those identified by motif mutagenesis (Fig. 4c and Supplementary Fig. 15) and refine known PWM models for the predicted TFs (Fig. 4c).

To validate experimentally the functional contribution of motif-flanking sequence predicted by DeepSTARR, we swapped the flanking nucleotides of strong and weak GATA instances (at least twofold difference) in 47 enhancers (Fig. 4d). Indeed, replacing the 2-bp flanks of strong instances by the flanks of weak instances reduced enhancer activity, whereas replacing the flanks of weak instances by the flanks of strong ones increased enhancer activity (Fig. 4d and Supplementary Fig. 16a,b). DeepSTARR recapitulated the observed effects, that is, the addition of weak flanks converted a strong GATA instance to a weak one as indicated by the decreased contribution at the nucleotide level, and vice versa for a weak instance that was converted to a strong one (Fig. 4e and Supplementary Fig. 16b). Swapping 5-bp flanks yielded consistent results with slightly stronger effects (Supplementary Fig. 16a,b). In addition, swapping the flanks was sufficient to switch motif contributions, as determined by subsequent motif mutagenesis (Supplementary Fig. 16c,d). Thus, as DeepSTARR is not biased by previous knowledge about TF motifs but is trained on DNA sequence alone, it can not only identify important motif types but also refine optimal flanking sequences. These could contribute to motif importance via motifs for other TFs, DNA shape and nucleosome positioning[18], but might also reflect extended motifs resulting from partial definition of the original motifs or alternative modes of TF binding. For GATAA, our results are most consistent with single TF binding mode[73,74], and GA-containing flanks for GAGAG might increase the avidity of TF binding. Experimentally, we confirm that the flanking sequence can be sufficient to switch motif contribution and should be considered when assessing motif importance or the impact of motif-disrupting mutations.

In silico analysis reveals modes of motif cooperativity. The position of TF motifs in the enhancer[75] and the distance between TF motifs are thought to be important motif syntax features. DeepSTARR indeed predicted higher importance for TF motifs at the center of the enhancers, which was confirmed by motif mutagenesis, though the trend was weaker (Supplementary Fig. 17). We next determined how the relative distance between two motif instances (MotifA/MotifB)—a feature generally associated with TF cooperativity[6,13,18,49,76–79]—contributes to enhancer activity using DeepSTARR. We embedded MotifA in the center of synthetic random DNA sequences and MotifB at a range of distances from MotifA, both up- and downstream, predicted the activity of the resulting sequences, and calculated a cooperativity score for each motif pair, where a value higher than 1 means positive synergy (Fig. 5a and Supplementary Fig. 18a; strategy adapted from ref. [49]).

Motif distances indeed had a strong influence on predicted enhancer activity and we observed four distinct modes of distance-dependent TF motif cooperativity: motif pairs can synergize exclusively at close distances (<25 bp; mode 1), exclusively at longer distances (>25 bp; 2), preferentially at closer distances and either plateau (3) or decay (4) at long distances (>75 bp; Fig. 5b and Supplementary Fig. 18b–d). While all motifs in housekeeping enhancers cooperate according to mode 4 (decay), modes 1 to 3 all occur for motifs in developmental enhancers (Supplementary Fig. 18c,d). Interestingly, whether cooperativity followed modes 1, 2 or 3 depended on the TF and the motif pair (Fig. 5c and Supplementary Fig. 18c). For example, ETS and AP-1 TFs always interacted according to mode 1 and 3, respectively, and mode 1 of the ETS TFs suggests direct protein–protein interactions with other TFs, which has indeed been observed[80,81]. Interestingly, GATA family TFs display more complex behavior and interact according to modes 1, 2 and 3 depending on the respective partner TF: GATA/ETS synergized only when closer than 25 bp (mode 1), whereas GATA/GATA synergy was lost at short distances (mode 2) and GATA/AP-1 cooperated according to mode 3 (Fig. 5c). Thus, DeepSTARR predicts distinct modes of motif cooperativity that can determine the contribution of different motif instances.

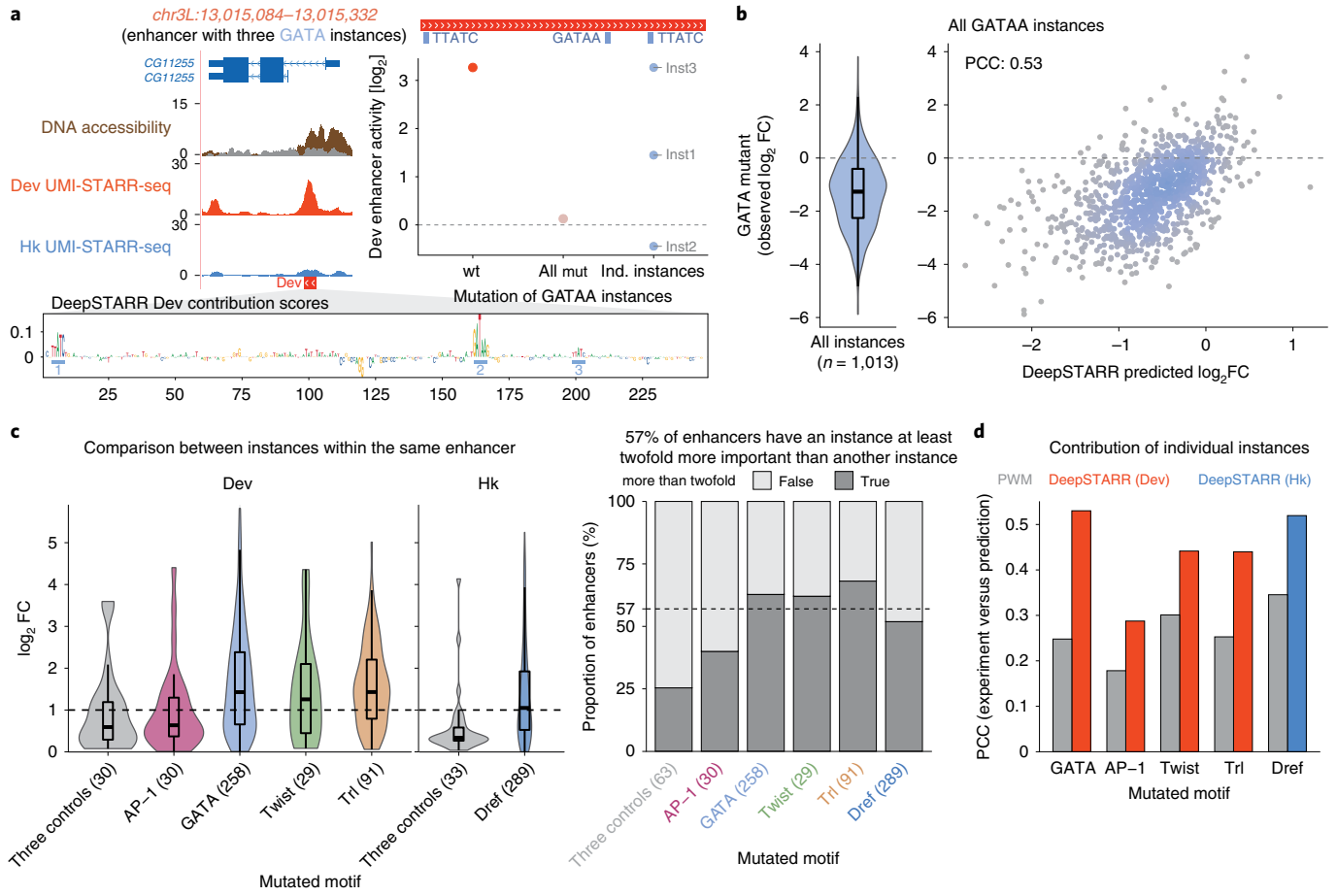

Fig. 3 | Instances of the same TF motif have nonequivalent contributions to enhancer activity. a, Developmental enhancer with three nonequivalent GATA instances. Left: genome browser screenshot showing tracks for DNA accessibility[63] and Dev and Hk UMI-STARR-seq for the *CG11255* locus. The designed oligonucleotide covering the enhancer selected for motif mutagenesis is shown. Right: $\log_2$ activity of the wildtype (wt) enhancer compared with the activity when mutating all GATA instances simultaneously (All mut) or each individual instance at a time (Ind. instances). Bottom: DeepSTARR nucleotide contribution scores for the same Dev enhancer with the three GATA instances highlighted. **b**, DeepSTARR predicts the contribution of individual GATA instances. Distribution of experimentally measured enhancer activity $\log_2$FC after mutating 1,013 different GATA instances across Dev enhancers (violin plot), compared with the $\log_2$FC predicted by DeepSTARR. The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers). **c**, Different instances of the same TF motif in the same enhancer are not equivalent. Left: distribution of enhancer activity change ($\log_2$FC) between the least and the most important instance of each motif type per enhancer. $\log_2$FC between instances of three control motifs is also shown. Dashed line represents twofold difference between instances in the same enhancer. Right: proportion of enhancers with two or more instances that have an instance at least twofold more important than another instance (dark gray). Dashed line represents the average across the different motif types (excluding control motifs): 57% of enhancers. Number of enhancers mutated for each motif type are shown. Box plots as in **b**. **d**, DeepSTARR predicts motif-instance contribution better than PWM motif scores. Bar plots showing the PCC between predicted (by DeepSTARR or PWM) and observed $\log_2$FC for mutating individual instances of each motif type.

We next asked how frequently these optimal intermotif distances occur in endogenous enhancers compared with negative regions. Motif pairs of housekeeping enhancers followed the optimal spacing rules (enrichment at close distances; Fig. 5b and Supplementary Fig. 19a,d), as did some motif pairs in developmental enhancers such as GATA/GATA motif pairs that were strongly depleted at close and enriched at longer distances (Fig. 5b). However, several pairs in developmental enhancers occurred only rarely at optimal distances (for example ETS/SREBP and AP-1/GATA; Fig. 5b and Supplementary Fig. 19a,c), even though the enhancer activities followed the predicted optimal spacing rules also in these cases (Fig. 5b and Supplementary Fig. 19). For instance, even though ETS/SREBP motifs separated by short distances (<25 bp) were rare, such motif pairs were associated with stronger enhancer activity than pairs separated by larger distances (75–100 bp; Fig. 5b), validating the ETS/SREBP motifs' optimal distance.

To experimentally test the importance of motif pairs at optimal versus nonoptimal distances more directly, we mutated either GATA or AP-1 motifs at close (<25 bp) and longer (>50 bp) distances to a GATA instance (Fig. 5d,e). The results validated the DeepSTARR predictions and showed higher importance of GATA/GATA pairs at longer (Fig. 5d) and AP-1/GATA pairs at closer distances (Fig. 5e). Thus, different motif pairs display distinct distance preferences, which dictate the contribution of individual motif instances to overall enhancer activity. As endogenous enhancers often contain motif pairs at nonoptimal distances, optimal distances only become apparent by our in silico analysis but not in frequency-based analyses.

Motif syntax rules can be generalized to human enhancers. To test whether individual instances of the same motif also contribute differently to enhancer activities in humans and whether motif flanks and spacing determine the different contributions, we chose

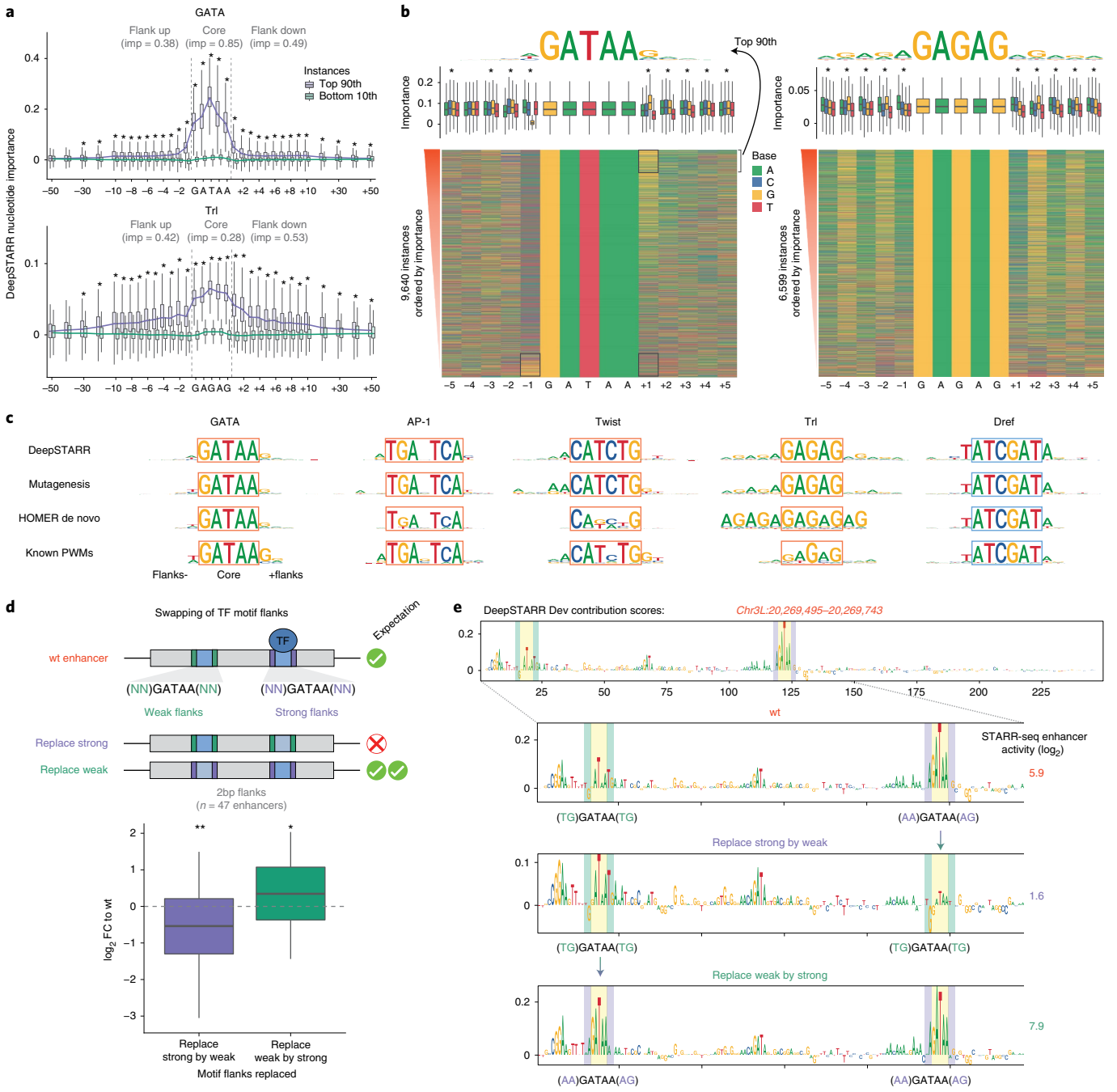

Fig. 4 | Contribution of TF motifs depends on the flanking sequence. a, DeepSTARR-predicted importance for ±50 flanking nucleotides of top 90th (purple) and bottom 10th (green) percentile GATA ($n = 992$ instances per box) and Trl ($n = 680$) motif instances selected based on DeepSTARR scores for core motif sequence. Asterisks mark positions with significant differences (two-sided Wilcoxon rank-sum test P value <0.001). The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers) and the lines connect the respective medians. The importance (imp) of the core and upstream or downstream flanking sequences corresponds to the sum of deltas between medians of top and bottom instances for the positions with significant differences. **b**, Motif contribution correlates with flanking base-pairs. Heatmap: flanking nucleotides of GATAA (GATA) and GAGAG (Trl) instances across Dev enhancers sorted by their DeepSTARR predicted contribution. Box plots: importance of motif instances according to the different bases at each flanking position. Asterisks mark positions with significant differences between the four nucleotides (FDR-corrected Welch one-way ANOVA test P value <0.01). Box plots as in **a**. Top: logos of the top 90th percentile motif instances. **c**, Comparison of optimal motif logos as predicted by DeepSTARR or measured experimentally by motif mutation, with the PWM logos derived de novo using HOMER or from *Drosophila* TF databases. Note that DeepSTARR and mutagenesis motif instances were selected to all contain the same core sequence. **d**, GATA flanking nucleotides are sufficient to switch motif contribution in 47 enhancers that contain one strong (purple) and one weak (green) GATA instance (at least twofold difference as assessed by mutagenesis). Enhancer activity change ($\log_2$FC) when 2-bp flanks of strong instances were replaced by the flanks of weak instances (purple; two-sided Wilcoxon signed rank test $P = 0.001$) and vice versa (green; $P = 0.026$). Box plots as in **a**. **e**, Example of a Dev enhancer with one weak (green) and one strong (purple) GATA instance. DeepSTARR nucleotide contribution scores and UMI-STARR-seq measured enhancer activity ($\log_2$; right) are shown for the wildtype sequence (top) and for the sequences where the 2-bp flanks of the strong instance were replaced by the ones of the weak instance (middle) and vice versa (bottom).

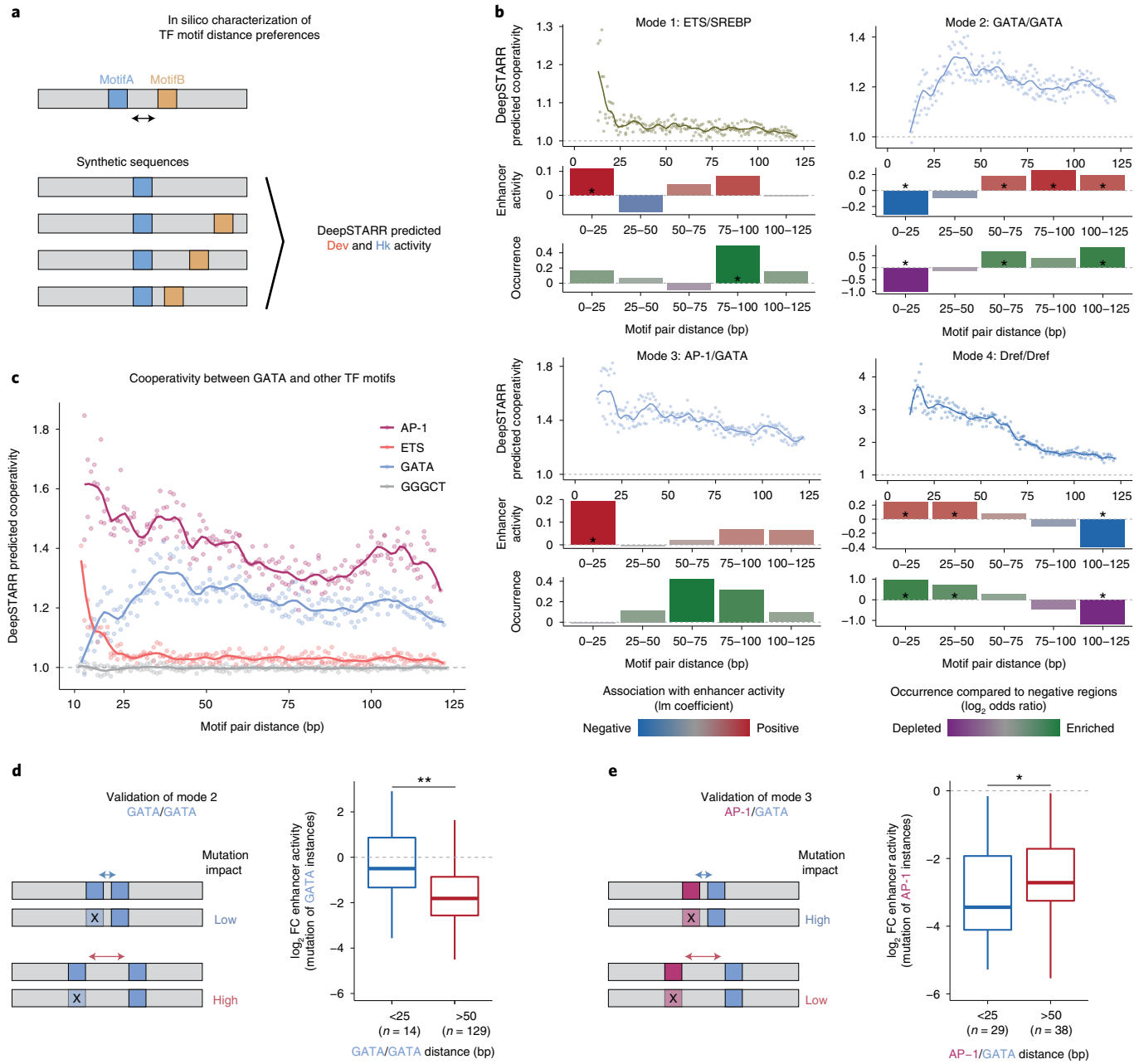

Fig. 5 | In silico analysis reveals distinct modes of motif cooperativity. a, Schematic of in silico characterization of TF motif distance preferences. MotifA was embedded in the center of 60 random sequences and MotifB at a range of distances from MotifA. The Dev and Hk enhancer activity was predicted by DeepSTARR and converted to linear space. The cooperativity (residuals) between MotifA and MotifB as a function of distance was quantified as the activity of MotifA + B divided by the sum of the marginal effects of MotifA and MotifB (MotifA + MotifB − backbone) (Methods). **b**, DeepSTARR predicts distinct modes of motif cooperativity: ETS/SREBP (mode 1), GATA/GATA (2), AP-1/GATA (3) and Dref/Dref (4). Top: cooperativity between two motifs at different distances. Points showing the median interaction across all 60 backbones for each motif pair distance together with smooth lines. Middle: association between enhancer activity and the distance at which the motif pair is found. Coefficient (*y* axis) and *P* value from a multiple linear regression including the number of instances for the different motif types. Bottom: odds ratio ($\log_2$) by which the two motifs are found within a specified distance from each other in enhancers compared with negative regions. Color legend is shown. An asterisk indicates FDR-corrected two-sided Fisher's exact test *P* value <0.05. **c**, Cooperativity between three motif types (and GGGCT as control) and a central GATA motif at different distances. Points showing the median interaction across all 60 backbones for each motif pair distance together with smooth lines. **d**, Motif mutagenesis validates that GATA instances distal to a second GATA are more important. Left: expected mutational impact when mutating GATA instances depending on the distance to other GATA motifs. Right: enhancer activity changes ($\log_2$FC) after mutating GATA instances at suboptimal close (<25 bp) or optimal longer (>50 bp) distance to a second instance. Number of instances are shown. Two-sided Wilcoxon rank-sum test $P = 0.008$. The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers). **e**, Motif mutagenesis validates that AP-1 instances closer to a second GATA instance are more important (same as in **d**). $P = 0.04$.

the human colon cancer cell line HCT116 as a model. We selected nine TF motifs based on motif enrichment analysis (AP-1, p53, MAF, CREB1, ETS, EGR1, MECP2, E2F1 and Ebox/MYC), mutated all their instances in 1,083 enhancers and assessed the enhancer activity of wildtype and mutant sequences by UMI-STARR-seq (Supplementary Fig. 20; Methods). This revealed that AP-1 and p53 motifs were the most important motifs (median 5.6- and 5.5-fold reduction, respectively), followed by MAF (3.1), CREB1 (2), ETS (1.9) and EGR1 (1.5), while MeCP2, E2F1 and Ebox/MYC motifs had the least impact on enhancer activity (less than 1.5-fold; Supplementary Fig. 20d–f). Based on these results, we chose AP-1, p53, MAF, CREB1, ETS and EGR1 motifs for the analysis of motif instances.

Mutation of hundreds of individual motif instances showed that instances of the same TF motif are not functionally equivalent (Fig. 6a–c and Supplementary Fig. 21a). For example, the enhancer shown in Fig. 6a contains four AP-1 instances with very different contributions to enhancer activity as judged by fold-changes after motif-instance mutagenesis between 1.2- and 3.8-fold. Interestingly, DNaseI footprinting data from a related colon cancer cell line (RKO[82]) suggest that the AP-1 instance with low importance was not bound endogenously, in contrast to the three important AP-1 instances (Fig. 6a). Both results generalize to all tested motifs and across enhancers: 57% of human enhancers displayed nonequivalent instances of the same motif type (Fig. 6b,c) and TF motif instances with DNaseI footprints are more important than those without (Fig. 6d), supporting the functional differences between motif instances at endogenous enhancers.

Having trained a convolutional neural network to learn the motif syntax rules for *Drosophila* enhancers, we wanted to determine if the same type of rules also apply to human enhancers. Therefore, we generated simple linear models based on these rules to predict the contribution of individual motif instances in human enhancers. Specifically, these models consider the number of instances, the motif core and flanking sequence, the motif position relative to the enhancer center[75] (Supplementary Fig. 22) and the distance to other TF motifs (Fig. 6e and Supplementary Fig. 21b,c). Despite their simplicity, these models were able to predict motif-instance importance, with PCCs to experimentally assessed $\log_2$ fold-changes ($\log_2$FC) of 0.67 (p53), 0.61 (ETS), 0.59 (MAF) and 0.52 (AP-1), outperforming models based solely on PWM scores (Supplementary Fig. 21d). The motif flanks and intermotif distances explained on average 13.7% and 8.2% of the motif mutations variance, respectively (Supplementary Fig. 21e). For most TFs, motif instances closer to an AP-1 or ETS motif were more important, suggesting that high cooperativity with these TFs is important in HCT116 enhancer sequences (Fig. 6e and Supplementary Fig. 21b). This was also observed between AP-1 and ETS motifs themselves, where mutation of either AP-1 or ETS instances had stronger impact in enhancer function if located at close (<25 bp) rather than longer distances (>50 bp) from each other (Fig. 6f). Altogether, these results confirm that motif-flanking sequences and intermotif distances dictate the contribution of individual TF motif instances not only in *Drosophila* but also human enhancers (Fig. 6g).

Surprisingly, for AP-1 motifs, which we could assess in both species, the *Drosophila*-trained DeepSTARR model was able to predict the importance of individual instances in human enhancers reasonably well (PCC = 0.42; Supplementary Fig. 23d), and, in both species, ETS/AP-1 pairs synergize only at short distances but not at longer ones (mode 1; Supplementary Figs. 18c and 23). These results suggest that homologous TFs and their motifs might display similar rules across species.

Designing synthetic enhancers with desired activities. Understanding how DNA sequence encodes enhancer activity should enable the design of synthetic enhancers with desired activity levels. We used DeepSTARR to computationally generate synthetic S2 cell developmental enhancers de novo, by predicting enhancer activities for 1 billion random 249-bp DNA sequences that are not present in the *Drosophila* genome (Methods). We then selected 249 of these sequences spanning different predicted activity levels and experimentally measured their enhancer activity by UMI-STARR-seq in S2 cells, yielding a quantitative agreement of PCC = 0.62 (Fig. 7a and Supplementary Fig. 24). DeepSTARR was also able to design synthetic enhancers as strong as the strongest native S2 developmental enhancers (activity (fold-change over negative regions) ≈ 500; Supplementary Table 17).

Inspection of the synthetic enhancer sequences suggested that their different activity levels correlated not only with motif composition but also the motif syntax (Fig. 7b). For example, three different enhancers, all containing two GATA and two AP-1 motifs, were predicted by DeepSTARR and validated experimentally to have very different activities (from 0.87 to 630). Interestingly, the strongest synthetic enhancer followed the optimal spacing rules predicted by DeepSTARR, such as distal GATA instances and proximal AP-1/GATA and ETS/AP-1 instances, whereas the other two synthetic sequences contained motifs in suboptimal syntax, such as distal AP-1 instances and proximal GATA instances (Fig. 7b).

Finally, we tested the activity of the three strongest synthetic enhancers in different orientations and both upstream and downstream of the promoter by luciferase assays (Supplementary Fig. 25). Similar to a strong native enhancer, all three synthetic enhancers showed strong activity and functioned independently of their orientation and position, thus displaying the defining properties of bona fide enhancers[1]. This proof-of-concept experiment shows that the rules learned by DeepSTARR enable the a priori design of synthetic enhancers with desired activity levels.

Discussion

Identifying enhancers and characterizing their sequence determinants—the cis-regulatory code—is a long-standing problem. Here, we dissect the relationship between enhancer sequence and strength for a single model cell type using deep learning. DeepSTARR accurately predicts enhancer activity for two different transcriptional programs directly from DNA sequence and reveals important aspects of the cis-regulatory code.

The discovery that relatively rare sequence features can be important for enhancer activity highlights the potential of deep-learning models that are not based on statistical over-representation[49,83]. The fact that identical instances of the same TF motif typically make nonequivalent contributions to enhancer activity is equally important. Although not all motif instances in large genomes can be equivalent given that many are not bound[22,67,84], their nonequivalence in the same enhancer is surprising. In fact, previous studies and computational models have typically considered different motif instances solely according to their PWM scores or even as equivalent[17,27,85]. Instead, the contribution of motif instances depends on higher-order syntax rules that are not captured by traditional PWM models, which is in line with the limitations of PWM models for predicting the effects of noncoding variants on TF binding in vitro[86] and the improved performance of deep-learning models to predict motif instances bound in vivo[49,59]. The finding that motif instances need to be analyzed in their cis-regulatory context is crucial for our ability to interpret the impact of disease-related sequence variants, which typically affect only individual motif instances.

Motif nonequivalency as well as the importance of motif flanks and distances generalize from *Drosophila* to human enhancers and for AP-1 motifs, which we could assess in both species, even the specific rules are shared. This suggests that both species share the same types of enhancer syntax rules and even some specific rules and it will be interesting to see how cell-type- and species-specific rules derive from a shared framework of general enhancer syntax.

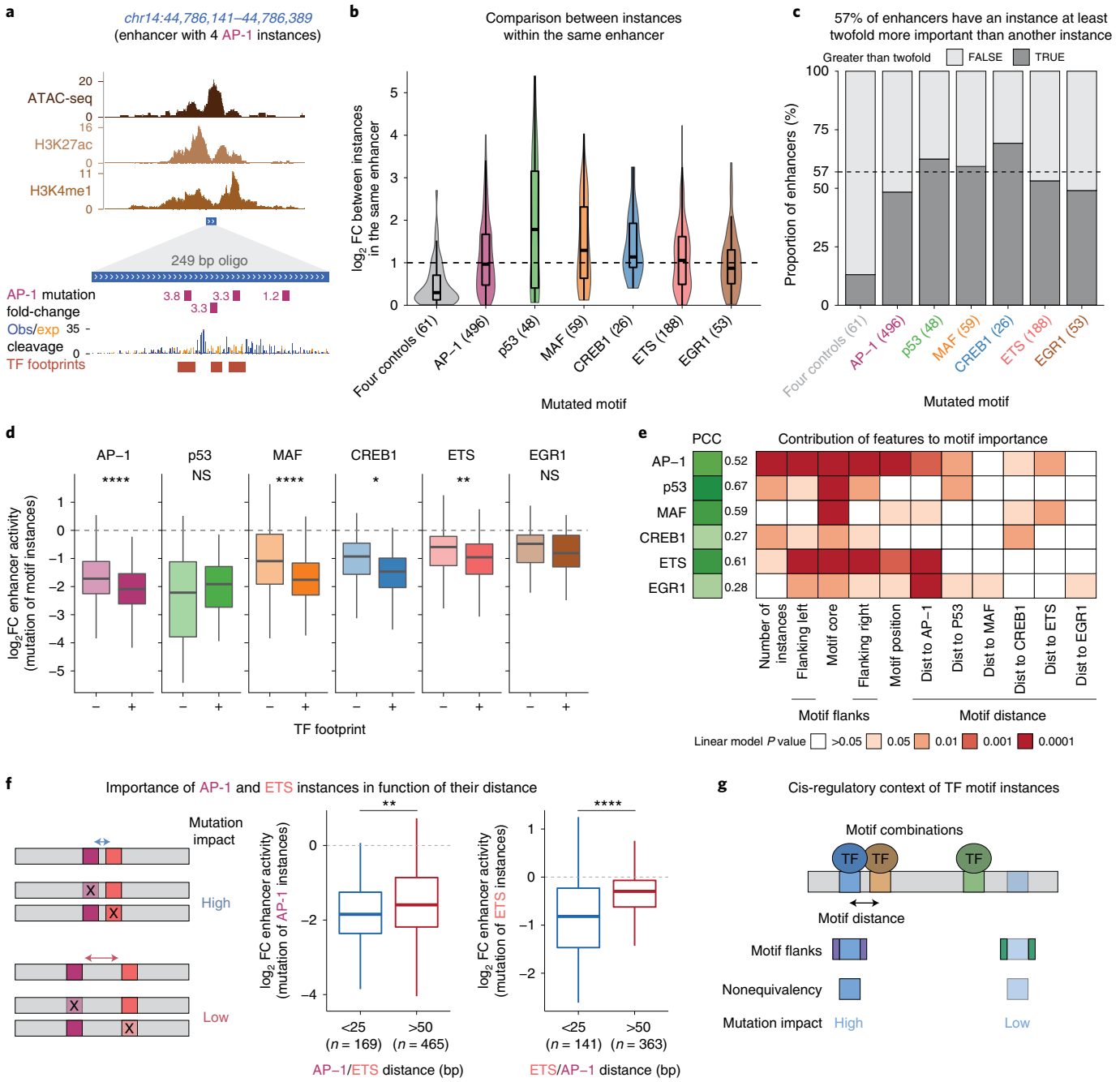

Fig. 6 | Motif syntax rules dictate the contribution of TF motif instances in human enhancers. a, Top: ATAC-seq[96], H3K27ac[97] and H3K4me1[97] signals for an enhancer with four AP-1 instances. Bottom: oligonucleotide used for motif mutagenesis containing four AP-1 instances and their mutation impact on enhancer activity (negative fold-change). Observed and expected DNase I cleavage and consensus TF footprints from a related colon cancer cell line (RKO[82]). **b**, Distribution of $\log_2$FC enhancer activity between mutating the least and the most important instance of each motif type per enhancer. The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers). **c**, A total of 57% of enhancers have a motif instance that is at least twofold more important than another instance. Gray bars: proportion per motif type; dashed line: average across motif types (excluding control motifs). **d**, $\log_2$ FC enhancer activity of mutating individual instances that do not (−) or do (+) overlap TF footprints in RKO cells[82]. $^{****}P < 0.0001$, $^{**}P < 0.01$, $^{*}P < 0.05$, NS, not significant (two-sided Wilcoxon rank-sum test). Box plots as in **b**; AP-1, $n = 795$ or 452; p53, $n = 142$ or 16; MAF, $n = 197$ or 115; CREB1, $n = 133$ or 22; ETS, $n = 620$ or 70 and EGR1, $n = 259$ or 45. **e**, For each motif type, we built a linear model containing different motif syntax features to predict the contribution of its individual instances across all enhancers. The PCC between predicted and observed motif contribution is shown (green scale). Heatmap shows the contribution of each feature (columns) for each model, colored by the linear regression P value (red scale). **f**, Motif mutagenesis shows that AP-1 and ETS instances closer to each other are more important to enhancer activity. Left: expected mutational impact when mutating AP-1 and ETS instances depending on the distance to each other. Middle and right: enhancer activity changes ($\log_2$FC) after mutating AP-1 or ETS instances at close (<25 bp) or longer (>50 bp) distance. Number of instances are shown. Two-sided Wilcoxon rank-sum test $P = 0.004$ (AP-1/ETS) and 1.5×10^{-10} (ETS/AP-1). Box plots as in **b**. **g**, Motif instances need to be analyzed in their cis-regulatory context. Motif syntax rules, such as motif combination, flanks and distance dictate the contribution of TF motif instances in enhancer sequences.

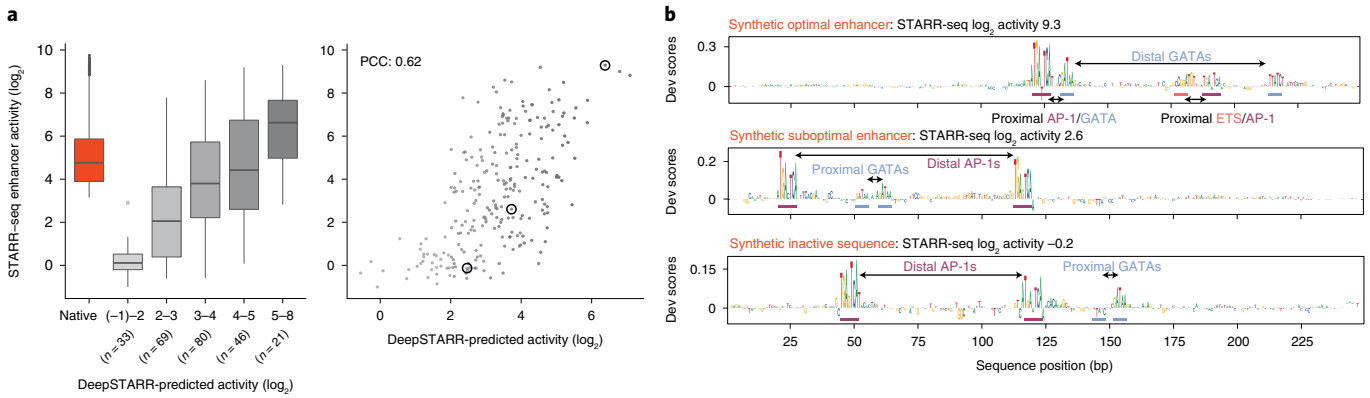

Fig. 7 | DeepSTARR designs synthetic enhancers using optimal sequence rules. a, Comparison between DeepSTARR predicted and experimentally measured enhancer activity (log_2) for 249 synthetic sequences binned (left) or not (right). The 'Native' category contains all *Drosophila* S2 Dev enhancer sequences. The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers); outliers are shown individually. The three synthetic sequences shown in **b** are highlighted. **b**, DeepSTARR nucleotide contribution scores for three synthetic sequences from **a** spanning different activity levels. Instances of GATA, AP-1 and ETS motifs are shown together with their observed distances (proximal or distal).

Similarly, it will be interesting to see how the models for housekeeping enhancers generalize as they have both *Drosophila*-specific and shared motifs (for example DRE and TCT[87]).

Although libraries of synthetic elements have been used to explore enhancer structure[71], it has remained challenging to build fully synthetic enhancers with defined functional characteristics. DeepSTARR trained on S2 cell enhancers allowed the de novo design of synthetic enhancers with desired activity levels in S2 cells. The synthetic enhancers are of similar complexity as endogenous enhancers in the training set, for example in terms of motif number and diversity, and we speculate that they also show similar in vivo activity patterns, namely activity in mesodermal cell types and tissues (Supplementary Fig. 26). Moreover, the observation that a vast number of different sequences can have similar enhancer strengths highlights the flexibility of regulatory sequences and the evolutionary opportunities this provides. We expect that combining DeepSTARR with emerging algorithms that allow the direct generation of DNA sequences from deep-learning models[56] will provide unanticipated opportunities for the engineering of synthetic enhancers.

The performance of DeepSTARR in predicting enhancer strengths and nucleotide importance suggests that it captures the sequence-to-function relationship of S2 cell enhancers exceedingly well. Indeed, its genome-wide prediction accuracy approaches the similarity between biological replicates, and we expect that further improvements might require complementary synthetic training data. Interestingly however, the motif syntax features discussed here (TF motif combinations, flanks and distances) likely capture less information than DeepSTARR. For example, a linear model using these features cannot discriminate important from nonimportant motif instances as well as DeepSTARR can (Supplementary Fig. 13) and would, on its own, overpredict motif instances outside enhancers (Supplementary Fig. 27), suggesting that DeepSTARR captures additional and potentially more complex rules. In addition to improving deep-learning models such as DeepSTARR, a key challenge will therefore be the understanding of the models and the features they learn through new interpretation tools[83].

Our work is complementary to recent efforts modeling other aspects of enhancer biology using deep learning[45–55,88]. These include DNA accessibility[46–48,50,52,53,55], histone modifications[48,50,52,89,90] or TF binding[45,49,50,52,59], which are prominent features of enhancer chromatin that correlate well but not perfectly with enhancer activity and strength (Supplementary Fig. 28; see also refs. [36,63,91]). While the models are not directly comparable due to the use of distinct cell- and datatypes, they derive their predictive power from similar types of features, including TF motifs[45,46,49] and their combinations[47,53,55] and distances[49]. An important future question is therefore to what extent enhancer chromatin and activity are determined by the same or different DNA sequence features and whether these similarities and differences can be modeled. Such models could not only explain prominent differences between chromatin states and enhancer activities but potentially even allow the prediction of enhancer activity for cell types for which only chromatin state-information is available.

Understanding and modeling the similarities and differences between enhancer chromatin and activity should also provide the means to address the next key challenge in the field: the generalization of predictive models from individual deeply characterized model cell lines to all cell types of an organism. This task is challenging because enhancer activities are inherently cell-type-specific such that the underlying sequence rules must also differ between cell types, at least to some extent. Recent efforts to map DNA accessibility and other chromatin features for many cell types[92–94] and the respective sequence models could be integrated with models of enhancer activity and strengths, potentially allowing quantitative predictions of enhancer activities in many cell types. We anticipate that these will be further combined with models for promoters[42,43] and other cis-regulatory elements (for example, insulators or silencers) as well as models that predict gene transcription from enhancer activities (for example the ABC model[95]) or the wider genomic sequence context (for example, Enformer[50]) towards ultimately understanding how our genomes store gene-regulatory information to dictate gene expression and development.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01048-5.

Received: 20 September 2021; Accepted: 8 March 2022;
Published online: 12 May 2022

References

1. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).

2. Levine, M. Transcriptional enhancers in animal development and evolution. *Curr. Biol.* **20**, R754–R763 (2010).
3. Catarino, R. R. & Stark, A. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev.* **32**, 202–223 (2018).
4. Gompel, N., Prud'homme, B., Wittkopp, P. J., Kassner, V. A. & Carroll, S. B. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature* **433**, 481–487 (2005).
5. Rickels, R. & Shilatifard, A. Enhancer logic and mechanics in development and disease. *Trends Cell Biol.* **28**, 608–630 (2018).
6. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
7. Kulkarni, M. M. & Arnosti, D. N. Information display by transcriptional enhancers. *Development* **130**, 6569–6575 (2003).
8. Zinzen, R. P., Senger, K., Levine, M. & Papatsenko, D. Computational models for neurogenic gene expression in the *Drosophila* embryo. *Curr. Biol.* **16**, 1358–1365 (2006).
9. Erceg, J. et al. Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity. *PLoS Genet.* **10**, e1004060 (2014).
10. Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.* **15**, 453–468 (2014).
11. Crocker, J. et al. Low affinity binding site clusters confer Hox specificity and regulatory robustness. *Cell* **160**, 191–203 (2015).
12. Farley, E. K. et al. Suboptimization of developmental enhancers. *Science* **350**, 325–328 (2015).
13. Farley, E. K., Olson, K. M., Zhang, W., Rokhsar, D. S. & Levine, M. S. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc. Natl Acad. Sci. USA* **113**, 6508–6513 (2016).
14. Fiore, C. & Cohen, B. A. Interactions between pluripotency factors specify cis-regulation in embryonic stem cells. *Genome Res.* **26**, 778–786 (2016).
15. Mathelier, A. et al. DNA shape features improve transcription factor binding site predictions in vivo. *Cell Syst.* **3**, 278–286 (2016).
16. Sayal, R., Dresch, J. M., Pushel, I., Taylor, B. R. & Arnosti, D. N. Quantitative perturbation-based analysis of gene expression predicts enhancer activity in early Drosophila embryo. *eLife* **5**, e08445 (2016).
17. King, D. M. et al. Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. *eLife* **9**, e41279 (2020).
18. Jindal, G. A. & Farley, E. K. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Dev. Cell* **56**, 575–587 (2021).
19. Swanson, C. I., Evans, N. C. & Barolo, S. Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev. Cell* **18**, 359–376 (2010).
20. Snetkova, V. et al. Ultraconserved enhancer function does not require perfect sequence conservation. *Nat. Genet.* **53**, 521–528 (2021).
21. Panne, D. The enhanceosome. *Curr. Opin. Struct. Biol.* **18**, 236–242 (2008).
22. Wang, J. et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).
23. Guo, Y., Mahony, S. & Gifford, D. K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.* **8**, e1002638 (2012).
24. Junion, G. et al. A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* **148**, 473–486 (2012).
25. Liu, F. & Posakony, J. W. Role of architecture in the function and specificity of two notch-regulated transcriptional enhancer modules. *PLoS Genet.* **8**, e1002796 (2012).
26. Smith, R. P. et al. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* **45**, 1021–1028 (2013).
27. Yanez-Cuna, J. O. et al. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res.* **24**, 1147–1156 (2014).
28. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.* **94**, 890–898 (2005).
29. Berman, B. P. et al. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol.* **5**, R61 (2004).
30. Crocker, J., Ilsley, G. R. & Stern, D. L. Quantitatively predictable control of *Drosophila* transcriptional enhancers in vivo with engineered transcription factors. *Nat. Genet.* **48**, 292–298 (2016).
31. He, X., Samee, M. A. H., Blatti, C. & Sinha, S. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput. Biol.* **6**, e1000935 (2010).
32. Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* **451**, 535–540 (2008).
33. Beer, M. A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004).
34. Zinzen, R. P. & Papatsenko, D. Enhancer responses to similarly distributed antagonistic gradients in development. *PLoS Comput. Biol.* **3**, 0826–0835 (2007).
35. Ghandi, M., Lee, D., Mohammad-noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10**, e1003711 (2014).
36. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24**, 1595–1602 (2014).
37. Grossman, S. R. et al. Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc. Natl Acad. Sci. USA* **114**, E1291–E1300 (2017).
38. Kheradpour, P. et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23**, 800–811 (2013).
39. Svetlichnyy, D., Imrichova, H., Fiers, M., Kalender Atak, Z. & Aerts, S. Identification of high-impact cis-regulatory mutations using transcription factor specific random forest models. *PLoS Comput. Biol.* **11**, e1004590 (2015).
40. Dibaeinia, P. & Sinha, S. Deciphering enhancer sequence using thermodynamics-based models and convolutional neural networks. *Nucleic Acids Res.* **49**, 10309–10327 (2021).
41. Zabidi, M. A. et al. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).
42. Arnold, C. D. et al. Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nat. Biotechnol.* **35**, 136–144 (2017).
43. Haberle, V. et al. Transcriptional cofactors display specificity for distinct types of core promoters. *Nature* **570**, 122–126 (2019).
44. Kleftogiannis, D., Kalnis, P. & Bajic, V. B. Progress and challenges in bioinformatics approaches for enhancer identification. *Brief. Bioinform.* **17**, 967–979 (2016).
45. 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).
46. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
47. Kim, D. et al. The dynamic, combinatorial cis-regulatory lexicon of epidermal differentiation. *Nat. Genet.* **53**, 1564–1576 (2021).
48. Kelley, D. R. et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).
49. Avsec, Ž. et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).
50. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
51. Karbalayghareh, A., Sahin, M. & Leslie, C. S. Chromatin interaction aware gene regulatory modeling with graph attention networks. Preprint at *bioRxiv* https://doi.org/10.1101/2021.03.31.437978 (2021).
52. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
53. Minnoye, L. et al. Cross-species analysis of enhancer logic using deep learning. *Genome Res.* **30**, 1815–1834 (2020).
54. Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
55. Janssens, J. et al. Decoding gene regulation in the fly brain. *Nature* **601**, 630–636 (2022).
56. Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A deep neural network for predicting and engineering alternative polyadenylation. *Cell* **178**, 91–106 (2019).
57. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features propagating activation differences. In *Proc. 34th International Conference on Machine Learning* 3145–3153 (2017).
58. Shrikumar, A. *et al.* Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. Preprint at https://doi.org/10.48550/arXiv.1811.00416 (2018).
59. Zheng, A. et al. Deep neural networks identify sequence context features predictive of transcription factor binding. *Nat. Mach. Intell.* **3**, 172–180 (2021).
60. Koo, P. K., Majdandzic, A., Ploenzke, M., Anand, P. & Paul, S. B. Global importance analysis: an interpretability method to quantify importance of genomic features in deep neural networks. *PLoS Comput. Biol.* **17**, e1008925 (2021).

61. Greenside, P., Shimko, T., Fordyce, P. & Kundaje, A. Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. *Bioinformatics* **34**, i629–i637 (2018).
62. Movva, R. et al. Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PLoS One* **14**, e0218073 (2019).
63. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
64. Neumayr, C., Pagani, M., Stark, A. & Arnold, C. D. STARR-seq and UMI-STARR-seq: assessing enhancer activities for genome-wide-, high-, and low-complexity candidate libraries. *Curr. Protoc. Mol. Biol.* **128**, e105 (2019).
65. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proc. 31st International Conference on Neural Information Processing System* 4768-4777 (2017).
66. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
67. Yáñez-Cuna, J. O., Dinh, H. Q., Kvon, E. Z., Shlyueva, D. & Stark, A. Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. *Genome Res.* **22**, 2018–2030.
68. Scardigli, R., Bäumer, N., Gruss, P., Guillemot, F. & Le Roux, I. Direct and concentration-dependent regulation of the proneural gene Neurogenin2 by Pax6. *Development* **130**, 3269–3281 (2003).
69. Swanson, C. I., Schwimmer, D. B. & Barolo, S. Rapid evolutionary rewiring of a structurally constrained eye enhancer. *Curr. Biol.* **21**, 1186–1196 (2011).
70. Crocker, J., Preger-Ben Noon, E. & Stern, D. L. The soft touch: low-affinity transcription factor binding sites in development and evolution. *Curr. Top. Dev. Biol.* **117**, 455–469.
71. Crocker, J. & Ilsley, G. R. Using synthetic biology to study gene regulatory evolution. *Curr. Opin. Genet. Dev.* **47**, 91–101 (2017).
72. Boisclair Lachance, J. F., Webber, J. L., Hong, L., Dinner, A. R. & Rebay, I. Cooperative recruitment of Yan via a high-affinity ETS supersite organizes repression to confer specificity and robustness to cardiac cell fate specification. *Genes Dev.* **32**, 389–401 (2018).
73. Yu, M. et al. Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol. Cell* **36**, 682–695 (2009).
74. Chen, Y. et al. DNA binding by GATA transcription factor suggests mechanisms of DNA looping and long-range gene regulation. *Cell Rep.* **2**, 1197–1206 (2012).
75. Grossman, S. R. et al. Positional specificity of different transcription factor classes within enhancers. *Proc. Natl Acad. Sci. USA* **115**, E7222–E7230 (2018).
76. Scully, K. H. et al. Allosteric effects of Pit-1 DNA sites on long-term repression in cell type specification. *Science* **290**, 1127–1131 (2000).
77. Crocker, J., Tamori, Y. & Erives, A. Evolution acts on enhancer organization to fine-tune gradient threshold readouts. *PLoS Biol.* **6**, 2576–2587 (2008).
78. Cheng, Q. et al. Computational identification of diverse mechanisms underlying transcription factor-DNA occupancy. *PLoS Genet.* **9**, e1003571 (2013).
79. Morgunova, E. & Taipale, J. Structural perspective of cooperative transcription factor binding. *Curr. Opin. Struct. Biol.* **47**, 1–8 (2017).
80. Li, R., Pei, H. & Watson, D. K. Regulation of Ets function by protein–protein interactions. *Oncogene* **19**, 6514–6523 (2000).
81. Burda, P., Laslo, P. & Stopka, T. The role of PU.1 and GATA-1 transcription factors during normal and leukemogenic hematopoiesis. *Leukemia* **24**, 1249–1257 (2010).
82. Vierstra, J. et al. Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).
83. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**, 389–403 (2019).
84. Dror, I., Golan, T., Levy, C. & Rohs, R. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res.* **25**, 1268–1280 (2015).
85. Kvon, E. Z. et al. Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo. *Nature* **512**, 91–95 (2014).
86. Yan, J. et al. Systematic analysis of binding of transcription factors to noncoding variants. *Nature* **591**, 147–151 (2021).
87. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* **19**, 621–637 (2018).
88. Sahu, B. et al. Sequence determinants of human gene regulatory elements. *Nat. Genet.* **54**, 283–294 (2022).
89. Taylor, A. M. et al. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* **33**, 676–689 (2018).
90. Baisya, D. R. & Lonardi, S. Prediction of histone post-translational modifications using deep learning. *Bioinformatics* **36**, 5610–5617 (2020).
91. Mauduit, D. et al. Analysis of long and short enhancers in melanoma cell states. *eLife* **10**, e71735 (2021).
92. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
93. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).
94. Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).
95. Fulco, C. P. et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
96. Ponnaluri, V. K. C. et al. NicE-seq: High resolution open chromatin profiling. *Genome Biol.* **18**, 122 (2017).
97. Sloan, C. A. et al. ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**, D726–D732 (2016).

Methods

UMI-STARR-seq library cloning. Inserts for *Drosophila* genome-wide and oligonucleotide libraries were amplified (for primers, see Supplementary Table 1) and cloned into the *Drosophila* STARR-seq vector[63] containing either the *Drosophila* synthetic core promoter (DSCP) or Rps12 core promoters using Gibson cloning (New England BioLabs, catalog no. E2611S). The oligonucleotide library for human STARR-seq screens was amplified (for primers, see Supplementary Table 1) and cloned into the human STARR-seq plasmid with the ORI in place of the core promoter[98]. Genome-wide and oligonucleotide libraries were grown in 6 l and 2 l LB-Amp (Luria-Bertani medium plus ampicillin, 100 µg/ml), respectively, and purified with a Qiagen Plasmid *Plus* Giga Kit (catalog no. 12991).

Cell culture, transfection and UMI-STARR-seq. *Drosophila* S2 and human HCT116 cells were cultured as described previously[63,98]. Cells were electroporated using the MaxCyte-STX system at a density of 50×10^7 cells per 100 µl and 5 µg of DNA using the 'Optimization 1' protocol (S2) and at a density of 1×10^7 cells per 100 µl and 20 µg of DNA using the preset 'HCT116' program (HCT116), respectively. We transfected 400×10^6 S2 cells total per replicate with 20 µg of the input library for *Drosophila* and 80×10^6 HCT116 cells total per replicate with 160 µg of the input library for human cells. UMI-STARR-seq was performed as described previously[63,64,98]. Further experimental details can be found in the Supplementary Methods.

Illumina sequencing. Next-generation sequencing was performed at the VBCF NGS facility on an Illumina HiSeq 2500, NextSeq 550 or NovaSeq SP platform, following the manufacturer's protocol, using standard Illumina i5 indexes as well as UMIs at the i7 index.

Genome-wide UMI-STARR-seq data analysis. RNA and DNA input reads were mapped to the *Drosophila* genome (dm3), excluding chromosomes U, Uextra, and the mitochondrial genome, using Bowtie v.1.2.2 (ref. [99]). Mapping reads with up to three mismatches and a maximal insert size of 2 kb were kept. For paired-end RNA reads that mapped to the same positions, we collapsed those that have identical UMIs (10 bp, allowing one mismatch) to ensure the counting of unique reporter transcripts (Supplementary Table 2). After processing the two biological replicates separately, we pooled both replicates for developmental and housekeeping screens for further analyses.

Peak calling was performed as described previously[63]. Peaks that had a hypergeometric *P* value ≤0.001 and a corrected enrichment over input (corrected to the conservative lower bound of a 95% confidence interval) greater than 3 were defined as enhancers and resized to 249 bp (Supplementary Table 3). Noncorrected enrichment over input was used as enhancer activity metric. Enhancers were classified as developmental or housekeeping based on the screen with the highest activity.

Oligonucleotide library UMI-STARR-seq data analysis. RNA and DNA input reads were mapped to a reference containing 249-bp long sequences containing both wildtype and mutated fragments from the *Drosophila* or human libraries using Bowtie v.1.2.2 (ref. [99]). Mapping reads with the correct length, strand and with no mismatches were kept. Both DNA and RNA reads were collapsed by UMIs (10 bp) as above (Supplementary Table 2).

We excluded oligonucleotides with fewer than ten reads in any of the input replicates and added one read pseudocount to oligonucleotides with zero RNA counts. The enhancer activity of each oligonucleotide in each screen was calculated as the $\log_2$FC over input, using all replicates, with DESeq2 (ref. [100]).

Deep-learning data preparation. The genome was binned into 249-bp windows with a stride of 100 bp, excluding chromosomes U, Uextra, and the mitochondrial genome. We selected all windows at the summit of developmental and housekeeping enhancers, in addition to three windows on either side of the regions and a diversity of inactive sequences (Supplementary Methods). We augmented our dataset by adding the reverse complement of each original sequence, with the same output, ending up with 242,026 examples (484,052 postaugmentation). Sequences from the first (40,570; 8.4%) and second half of chr2R (41,186; 8.5%) were held out for validation and testing, respectively.

DeepSTARR model architecture and training. DeepSTARR was designed as a multitask convolutional neural network (CNN) that uses one-hot-encoded 249-bp long DNA sequence (A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1]) to predict both its developmental and housekeeping enhancer activities (Fig. 1c). We adapted the Basset CNN architecture[46] and built DeepSTARR with four one-dimensional (1D) convolutional layers (filters = 246, 60, 60, 120; size = 7, 3, 5, 3), each followed by batch normalization, a ReLU nonlinearity, and max-pooling (size = 2). After the convolutional layers, there are two fully connected layers, each with 256 neurons and followed by batch normalization, a ReLU nonlinearity, and dropout where the fraction is 0.4. The final layer mapped to both developmental and housekeeping outputs. Further details on model training, hyperparameter tuning and performance evaluation can be found in the Supplementary Methods. The performance of DeepSTARR in the test set sequences was also compared with two different methods: a gapped k-mer support vector machine (gkm-SVM)[35] and a lasso regression model based on TF motif counts.

Nucleotide contribution scores and motif discovery. We used DeepExplainer (the DeepSHAP implementation of DeepLIFT[57,65,66]; update from https://github.com/AvantiShri/shap/blob/master/shap/explainers/deep/deep_tf.py) to compute contribution scores for all nucleotides in all sequences with respect to either developmental or housekeeping enhancer activity. We used 100 dinucleotide-shuffled versions of each input sequence as reference sequences. For each sequence, the obtained hypothetical importance scores were multiplied by the one-hot-encoded matrix of the sequences to derive the final nucleotide contribution scores.

To consolidate motifs, we ran TF-Modisco (v.0.5.12.0 (ref. [58])) on the nucleotide contribution scores for each enhancer type separately using all developmental or housekeeping enhancers. We specified the following parameters: sliding_window_size=15, flank_size=5, max_seqlets_per_metacluster = 50,000 and TfModiscoSeqletsToPatternsFactory(trim_to_window_size = 15, initial_flank_to_add = 5). Motifs supported by less than 35 seqlets were discarded.

Reference compendium of nonredundant TF motifs. A total of 6,502 TF motif models were obtained from iRegulon (http://iregulon.aertslab.org/collections.html (ref. [101])). We systematically collapsed redundant motifs by similarity by a previously described approach[82]. The code and TF motif compendium are available from https://github.com/bernardo-de-almeida/motif-clustering. Details on TF motif enrichment analyses in developmental and housekeeping enhancers can be found in the Supplementary Methods.

***Drosophila* TF motif mutagenesis oligonucleotide library synthesis and UMI-STARR-seq.** We computationally designed a *Drosophila* enhancers' motif mutagenesis oligonucleotide library containing 524 negative genomic regions; 5,082 wildtype enhancers; variants of 2,375 enhancers with mutations of all instances simultaneously (per motif type) or each instance individually for eight developmental motifs (GATA, AP-1, twist, Trl, SREBP, CREB, ETS, STAT), four housekeeping motifs (Dref, Ohler1, Ohler6, Ohler7) and three control motifs; scanning mutagenesis of five enhancers; variants with swapped GATA motif flanks for 100 enhancers and 249 synthetic enhancer sequences (Supplementary Table 5). All details can be found in the Supplementary Methods. The resulting 21,758-plex 300-mer oligonucleotide library was synthesized by Twist Bioscience. UMI-STARR-seq using this oligonucleotide library was performed and analyzed as described above. We performed three independent replicates for developmental and housekeeping screens (correlation PCC = 0.94–0.98).

TF motif mutation analysis and equivalency. From the candidate 249 bp enhancers, we identified 855 active developmental and 905 active housekeeping *Drosophila* enhancers ($\log_2$ wildtype activity in oligonucleotide UMI-STARR-seq ≥3.15 and 2.51, respectively; the strongest negative region in each screen) that we used in the subsequent TF motif mutation analyses. The impact of mutating all instances of a TF motif type simultaneously or each instance individually was measured as the $\log_2$FC enhancer activity between the respective mutant and wildtype sequences (Supplementary Tables 6 and 8). This was done separately for developmental and housekeeping enhancer activities.

Motif nonequivalency across all enhancers or in the same enhancer was assessed by comparing the impact of mutating individual instances of the same TF motif, that is the $\log_2$FCs of each instance (Supplementary Table 8). For the comparison between instances in the same enhancer, only enhancers that require the TF motif (greater than twofold reduction in activity after mutating all instances) and contain two or more instances were used. Motif instances with greater than twofold different contributions in the same enhancer were considered as nonequivalent. The same comparison across enhancers or in the same enhancer was performed for the three control motifs.

DeepSTARR predicted global importance of motif types. To quantify the global importance of all known TF motifs to enhancer activity in silico[60], we embedded each motif from the 6,502 TF motif compendium at five different locations and in both orientations in 100 random backbone DNA sequences and predicted their developmental and housekeeping enhancer activity with DeepSTARR. For each motif, we used the sequence corresponding to the highest affinity according to the annotated PWM models. The average activity across the different locations per backbone was divided by the backbone initial activity to get the predicted increase in enhancer activity per TF motif. The resultant $\log_2$FC was averaged across all 100 backbones to derive the final global importance of each TF motif.

DeepSTARR predictions for the contribution of motif instances. We used two complementary approaches to measure the predicted contribution of each motif instance by DeepSTARR: (1) we measured the predicted importance of all string-matched instances of each TF motif type as the average developmental or housekeeping DeepSTARR contribution scores over all its nucleotides (used in Fig. 4a–c and Supplementary Figs. 8a, 12a,c, 14a and 15); (2) to directly compare with the experimentally derived motif importance through motif mutagenesis, we

used DeepSTARR to predict the log_2FC between wildtype and the motif-mutant enhancer sequences included in the oligonucleotide library for all instances of the different motif types (used in Fig. 3b,d and Supplementary Figs. 13 and 17a).

Scoring of TF motif instances with PWM motif scores. To assess how the PWM motif models predict the importance of a motif instance, we scored the wildtype sequence of each mutated motif instance with the PWM models of the selected TF motifs. We used the matchMotifs function from R package motifmatchr (v.1.4.0; genome = 'BSgenome.Dmelanogaster.UCSC.dm3', bg = 'even', ref. [102]) with a *P* value cutoff of 1 to retrieve the PWM scores of all sequences. We tested different PWM models for each TF motif, if available, and reported always the one with the best correlation (Supplementary Table 10).

Predicted contribution of motif-flanking nucleotides. The top 90th and bottom 10th percentile motif instances of each TF were selected based on their predicted (DeepSTARR scores for core sequence) or experimentally derived (minus signed (–) mutation log_2FC) importance. The DeepSTARR contribution scores of their ±50 flanking nucleotides were shown using box plots (Fig. 4a and Supplementary Fig. 14). For each position, significant differences between top and bottom instances were assessed through a Wilcoxon rank-sum test ($P < 0.001$). The sum of delta between medians of top and bottom instances for the positions with significant differences was used as measure of importance for the upstream and downstream flanking sequences.

Correlation between motif importance and motif-flanking sequence. String-matched motif instances of each TF were sorted by their predicted (DeepSTARR) or experimentally derived (minus signed (–) mutation log_2FC) importance. Their five flanking nucleotides were shown using heatmaps, and the importance of each nucleotide at each flanking position summarized using box plots (Fig. 4b and Supplementary Fig. 15). Significant differences between the four nucleotides per position were assessed through Welch one-way analysis of variance (ANOVA) test followed by false discovery rate (FDR) multiple testing correction.

The motifs recovered by DeepSTARR were compared with PWM models discovered de novo by HOMER. HOMER (v.4.10.4 (ref. [103])) was run on the 249-bp developmental or housekeeping enhancer regions with the findMotifsGenome.pl command and the command line argument –size 249.

In silico motif distance preferences. Two consensus TF motifs were embedded in 60 random backbone 249-bp DNA sequences, MotifA in the center and MotifB at a range of distances (*d*) from MotifA, both up- and downstream. DeepSTARR was used to predict the developmental or housekeeping activity of the backbone synthetic sequences (1) without any motif (b), (2) only with MotifA in the center (A), (3) only with MotifB d-bases up- or downstream (B) and (4) with both MotifA and MotifB (AB). The DeepSTARR predicted activities in log space were converted to linear space as $2^{\text{DeepSTARR prediction}}$. The cooperativity between MotifA and MotifB at each distance *d* was then defined as the fold-change between AB and (b + (A–b) + (B–b) = A + B–b), where a value of 1 means an additive effect or no synergy between the motifs, and a value higher than 1 means positive synergy. The median of fold-changes across the 60 backbones was used as the final cooperativity scores.

Enrichment of motif pairs at different distances in genomic enhancers. To compute whether MotifA is located within a certain distance (bins: 0–25, 25–50, 50–75, 75–100, 100–125, 125–150, 150–250 bp) of MotifB more or less frequently in enhancers than in negative sequences, we counted the number of times a MotifA instance is at each distance bin to a MotifB instance in enhancers and in negative sequences. The enrichment or depletion of motif pairs at each bin was tested with two-sided Fisher's exact test and the log_2 odds ratio used as metric. Obtained *P* values were corrected for multiple testing by Benjamini–Hochberg procedure and considered significant if FDR ≤ 0.05.

Association between motif pair distances and enhancer activity. For each pair of motif instances at each distance bin (0–25, 25–50, 50–75, 75–100, 100–125, 125–150, 150–250 bp), we tested the association between enhancer activity and the presence of the pair at the respective distance bin using a multiple linear regression, including as independent variables the number of instances for the different developmental or housekeeping TF motif types. The linear model coefficient was used as metric and considered significant if the FDR-corrected *P* values ≤0.05.

Human TF motif mutagenesis oligonucleotide library synthesis and UMI-STARR-seq. We selected the nine TF motif types with the strongest enrichment in enhancers in human HCT116 cells[98]: AP-1, p53, MAF, CREB1, ETS, EGR1, MECP2, E2F1 and Ebox/MYC (Supplementary Table 12 and Supplementary Methods). We selected 3,200 enhancer candidates, defining short 249-bp windows (the limits of oligonucleotide synthesis), and mapped the position of all instances of the nine TF motif types in these candidates using the matchMotifs function from R package motifmatchr (v.1.4.0 (ref. [102])) with the following parameters: genome = 'BSgenome.Hsapiens.UCSC.hg19', p.cutoff = $5e^{-04}$, bg = 'genome'. Overlapping instances (minimum 70%) for the same TF motif were collapsed. We also mapped all instances of four control motifs using string-matching.

We computationally designed the human enhancers' motif mutagenesis oligonucleotide library containing: 920 249-bp negative genomic regions as controls; 3,200 wildtype enhancers; and 18,780 enhancer variants with all instances of each motif type mutated simultaneously or individually to a motif shuffled variant (Supplementary Table 13). All details can be found in the Supplementary Methods. Apart from the specific sequences, this human motif mutagenesis library exhibits the same specifications as the *Drosophila* library and was also synthesized by Twist Bioscience. UMI-STARR-seq using this oligonucleotide library was performed and analyzed as described above. We performed two independent replicates (correlation PCC = 0.99).

Human TF motif mutation analysis. From the 3,200 designed candidate 249-bp enhancers, we identified 1,083 active short human enhancers (log_2 wildtype activity in oligonucleotide UMI-STARR-seq ≥ 2.03, the strongest negative region) that we used in the subsequent TF motif analyses. The impact of mutating all instances of a TF motif type simultaneously or each instance individually was calculated as the log_2FC enhancer activity between the respective mutant and wildtype sequences (Supplementary Tables 14 and 15). Motif nonequivalency across all enhancers or in the same enhancer was assessed as in the *Drosophila* enhancers.

Validation of important TF motif instances with genomic DNase I footprinting data. We compared the importance of individual motif instances with genomic DNase I footprinting data of RKO cells (another human colon cancer cell line; https://www.vierstra.org/resources/dgf (ref.[82])), as a surrogate for TF occupancy. For each TF motif type, a Wilcoxon rank-sum test was used to determine whether the mutation log_2FC of instances overlapping TF footprints (FPR threshold of 0.001) is significantly greater or less than the one of instances not overlapping footprints. Only instances in HCT116 accessible enhancers were used in the analysis.

Association between motif syntax rules and the contribution of TF motif instances. For each TF motif type, we built a multiple linear regression model to predict the contribution of its individual instances (log_2FCs) using as covariates the number of instances of the respective motif type in the enhancer, the motif core (defined as the nucleotides included in each TF motif PWM model) and flanking nucleotides (5 bp on each side), the motif position relative to the enhancer center[75], and the distance to all other TF motifs. All models were built using the *Caret* R package (v. 6.0–80 (ref. [104])) and tenfold cross-validation. Predictions for each held-out test set were used to compare with the observed log_2FCs and assess model performance. The linear model coefficients and respective *P* values were used as metrics of importance for each feature.

Luciferase reporter assays. We constructed luciferase reporters by cloning candidate enhancers in both orientations in the pGL3_DSCP_luc+ plasmid either upstream or downstream of the DSCP promoter. One native enhancer, the three strongest synthetic enhancers and five negative controls were amplified from the Twist oligonucleotide pools and plasmids verified by Sanger sequencing (for primers, see Supplementary Table 1). Luciferase assays were performed in quadruplicates as described previously[105].

Luciferase assay data analysis. We first normalized firefly over *Renilla* luciferase values for each of the eight biological replicates individually. To normalize to the core promoters' intrinsic activity, we then calculated the fold-change luciferase signal over the average signal of the five negative control sequences. For each enhancer candidate and construct, we used the average of the replicates as the final activity together with the s.d. (Supplementary Table 18).

Statistics and data visualization. All statistical calculations and graphical displays were performed in R statistical computing environment (v.3.5.1 (ref. [106])) and using the R package ggplot2 (v.3.2.1 (ref. [107])). Coverage data tracks have been visualized in the UCSC Genome Browser[108] and used to create displays of representative genomic loci. In all box plots, the central line denotes the median, the box encompasses 25th to 75th percentile (interquartile range) and the whiskers extend to 1.5× interquartile range.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw sequencing data are available from GEO (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE183939. Data used to train and evaluate the DeepSTARR model as well as the final pretrained model are found on zenodo at https://doi.org/10.5281/zenodo.5502060. The pretrained DeepSTARR model is also available in the Kipoi model repository[109] (http://kipoi.org/models/DeepSTARR/). Genome browser tracks showing genome-wide UMI-STARR-seq and DeepSTARR predictions in *Drosophila* S2 cells, including nucleotide contribution scores for all enhancer sequences, together with the enhancers used for mutagenesis, mutated motif instances and respective log_2FC in enhancer activity, are available at https://genome.ucsc.edu/s/bernardo.

almeida/DeepSTARR_manuscript. Dynamic sequence tracks (https://github.com/pkerpedjiev/higlass-dynseq) and contribution scores are also available as a Reservoir Genome Browser session at https://resgen.io/paper-data/Almeida...%202021%20-%20DeepSTARR/views. TF motif models were obtained from iRegulon (http://iregulon.aertslab.org/collections.html (ref. [101])). DNase-seq and ATAC-seq data in *Drosophila* S2 cells were obtained from refs. [63] and [110], respectively; nascent transcription from ref. [111] and H3K4me1 and H3K27ac chromatin marks from ref. [112]. RepeatMasker dm3 annotations were obtained from http://www.repeatmasker.org/genomes/dm3/RepeatMasker-rm405-db20140131/dm3.fa.out.gz. Genomic DNaseI footprinting data of RKO cells were downloaded from https://resources.altius.org/~vierstra/projects/footprinting.2020/per.dataset/h.RKO-DS40362/. HCT116 DNase-seq, H3K27ac and H3K4me1 data were obtained from ENCODE[97] (https://www.encodeproject.org/; ENCFF001SQU, ENCFF001WIJ, ENCFF001WIK, ENCFF175RBN, ENCFF228YKV, ENCFF851NWR, ENCFF927AHJ, ENCFF945KJN, ENCFF360XGA, ENCFF130JBP and ENCFF400KKD) and ATAC-seq data from ref. [96].

Code availability

Code used to process the genome-wide and oligonucleotide UMI-STARR-seq data, train DeepSTARR and predict the enhancer activity for new DNA sequences, as well as to reproduce the results, is available on GitHub (https://github.com/bernardo-de-almeida/DeepSTARR). The code and TF motif compendium are available from https://github.com/bernardo-de-almeida/motif-clustering.

References

98. Muerdter, F. et al. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141–149 (2018).
99. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
100. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
101. Janky, R. et al. iRegulon: from a gene list to a gene regulatory network using large motif and track collections. *PLoS Comput. Biol.* **10**, e1003731 (2014).
102. Schep, A. motifmatchr: fast motif matching in R. R package version 1.14.0 https://bioconductor.org/packages/release/bioc/html/motifmatchr.html (2021).
103. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
104. Kuhn, M. caret: classification and regression training. R package version 6.0-80 https://CRAN.R-project.org/package=caret (2018).
105. Stampfel, G. et al. Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* **528**, 147–151 (2015).
106. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).
107. Wickham, H. *ggplot2: Elegant Graphics For Data Analysis* (Springer, 2016); https://ggplot2.tidyverse.org
108. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
109. Avsec, Ž. et al. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* **37**, 592–600 (2019).
110. Albig, C. et al. Factor cooperation for chromosome discrimination in *Drosophila*. *Nucleic Acids Res.* **47**, 1706–1724 (2019).
111. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).
112. Rickels, R. et al. An evolutionary conserved epigenetic mark of polycomb response elements implemented by Trx/MLL/COMPASS. *Mol. Cell* **63**, 318–328 (2016).

Acknowledgements

We thank A. Andersen (Life Science Editors), V. Loubiere and F. Lorbeer (IMP) for comments on the manuscript, G. Hulselmans and S. Aerts (KU Leuven) for sharing the TF motif PWM collection, and P. Kerpedjiev for generating the dynamic sequence tracks. Deep sequencing was performed at the Vienna Biocenter Core Facilities GmbH. Research in the Stark group is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 647320) and by the Austrian Science Fund (FWF, F4303-B09). Basic research at the IMP is supported by Boehringer Ingelheim GmbH and the Austrian Research Promotion Agency (FFG).

Author contributions

B.P.d.A., F.R. and A.S. conceived the project. F.R. and M.P. performed all experiments. B.P.d.A. performed all computational analyses. B.P.d.A., F.R. and A.S. interpreted the data and wrote the manuscript. A.S. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01048-5.

Correspondence and requests for materials should be addressed to Alexander Stark.

Peer review information *Nature Genetics* thanks Ziga Avsec and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

nature portfolio

Corresponding author(s): Alexander Stark

Last updated by author(s): Feb 27, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: Deep sequencing base-calling was performed with CASAVA 1.9.1

Data analysis	Code used to process the genome-wide and oligo UMI-STARR-seq data and train DeepSTARR, as well as to predict the enhancer activity for new DNA sequences is available on GitHub (https://github.com/bernardo-de-almeida/DeepSTARR). The code and TF motif compendium are available from https://github.com/bernardo-de-almeida/motif-clustering. Read mapping: Bowtie version 1.2.2. Coverage calculation, region intersection: bedtools version 2.27.1, R/Bioconductor package GenomicRanges version 1.32.7. Genomic coverage tracks visualisation: UCSC Genome Browser (http://genome.ucsc.edu). Oligo enhancer activity calculation: DESeq2 (1.22.2) DeepSTARR was implemented and trained in Keras (v.2.2.4) (with TensorFlow v.1.14.0) using the Adam optimizer. DeepExplainer (the DeepSHAP implementation of DeepLIFT; update from https://github.com/AvantiShri/shap/blob/master/shap/explainers/deep/deep_tf.py) was used to compute contribution scores. TF–Modisco (v.0.5.12.0) used the contribution scores to derive TF motifs. Counts for each TF motif in each sequence were calculated using the matchMotifs function from R package motifmatchr (v.1.4.0). Clustering of motifs and their logos were visualized using the motifStack R package (v.1.26.0). Negative GC-matched genomic regions were generated from the genNullSeqs function from R package gkmSVM (v.0.80.0). De novo motif discovery was performed with HOMER (v4.10.4117). All statistical calculations and graphical displays have been performed in R statistical computing environment (v.3.5.1) and using the R package ggplot2 (v.3.2.1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The raw sequencing data are available from GEO (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE183939. Data used to train and evaluate the DeepSTARR model as well as the final pre-trained model are found on zenodo at https://doi.org/10.5281/zenodo.5502060. The pre-trained DeepSTARR model is also available in the Kipoi model repository (https://github.com/kipoi/models/tree/master/DeepSTARR). Genome browser tracks showing genome-wide UMI-STARR-seq and DeepSTARR predictions in Drosophila S2 cells, including nucleotide contribution scores for all enhancer sequences, together with the enhancers used for mutagenesis, mutated motif instances and respective log2 fold-changes in enhancer activity, are available at https://genome.ucsc.edu/s/bernardo.almeida/DeepSTARR_manuscript. Dynamic sequence tracks (https://github.com/pkerpedjiev/higlass-dynseq) and contribution scores are also available as a Reservoir Genome Browser session at https://resgen.io/paper-data/Almeida...%202021%20-%20DeepSTARR/views/VNZrgd8oSsCpfZfwByDlwA. TF motif models were obtained from iRegulon (http://iregulon.aertslab.org/collections.html). DNase-seq and ATAC-seq data in Drosophila S2 cells were obtained from Arnold et al. Science 2013 and Albig et al. NAR 2019, respectively; nascent transcription from Kwak et al. Science 2013 and H3K4me1 and H3K27ac chromatin marks from Rickels et al. Mol Cell 2016. RepeatMasker dm3 annotations were obtained from http://www.repeatmasker.org/genomes/dm3/RepeatMasker-rm405-db20140131/dm3.fa.out.gz. Genomic DNase I footprinting data of RKO cells were downloaded from https://resources.altius.org/~jvierstra/projects/footprinting.2020/per.dataset/h.RKO-DS40362/. HCT116 DNase-seq, H3K27ac and H3K4me1 data were obtained from ENCODE (https://www.encodeproject.org/; ENCFF001SQU, ENCFF001WIJ, ENCFF001WIK, ENCFF175RBN, ENCFF228YKV, ENCFF851NWR, ENCFF927AHJ, ENCFF945KJN, ENCFF360XGA, ENCFF130JBP and ENCFF400KKD) and ATAC-seq data from Ponnaluri et al. Genome biol 2017.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For next-generation sequencing experiments, we performed 2 (S2 genome-wide UMI-STARR-seq and HCT116 oligo UMI-STARR-seq) or 3 (S2 oligo UMI-STARR-seq) biological replicates (independent transfections) to assess their reproducibility and be able to perform statistical analysis. For single-candidate luciferase assays 8 biological replicates (independent transfections) were performed, as is standard in the field.
Data exclusions	No data was excluded.
Replication	The reproducibility of experimental findings was verified by repeating all analyses with each biological replicate independently. All replication attempts were successful.

Randomization	Not relevant because the samples were not grouped.
Blinding	All experiments were done in cell culture and did not involve animal or human research participants and therefore blinding did not apply.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Schneider 2 cells (S2 cells) were obtained from Life Technologies/Thermo Fisher Scientific (cat. no. R69007). Human colorectal carcinoma cells (HCT116) were obtained from ATCC (#CCL-247).
Authentication	As the cell lines were purchased directly from Life Technologies/Thermo Fisher Scientific (S2) and ATCC (HCT-116), visual inspection was used to confirm the morphology of each cell line (compared to pictures provided by the vendor).
Mycoplasma contamination	Cell lines were tested for mycoplasma and tested negative.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.