# Dissection of core promoter syntax through single nucleotide resolution modeling of transcription initiation

Adam Y. He[1,2] and Charles G. Danko[1,3*]

[1]Baker Institute for Animal Health, College of Veterinary Medicine,
Cornell University.
[2]Graduate Field of Computational Biology, Cornell University.
[3]Department of Biomedical Sciences, College of Veterinary Medicine,
Cornell University.

*Corresponding author(s). E-mail(s): dankoc@gmail.com;

**Abstract**

Our understanding of how the DNA sequences of *cis*-regulatory elements encode transcription initiation patterns remains limited. Here we introduce CLIPNET, a deep learning model trained on population-scale PRO-cap data that accurately predicts the position and quantity of transcription initiation with single nucleotide resolution from DNA sequence. Interpretation of CLIPNET revealed a complex regulatory syntax consisting of DNA-protein interactions in five major positions between −200 and +50 bp relative to the transcription start site, as well as more subtle positional preferences among different transcriptional activators. Transcriptional activator and core promoter motifs occupy different positions and play distinct roles in regulating initiation, with the former driving initiation quantity and the latter initiation position. We identified core promoter motifs that explain initiation patterns in the majority of promoters and enhancers, including DPR motifs and AT-rich TBP binding sequences in TATA-less promoters. Our results provide insights into the sequence architecture governing transcription initiation.

**Keywords:** transcription initiation, *cis*-regulatory elements, core promoter motifs, TATA box, DPR

# Introduction

Transcriptional regulation, the mechanism by which cells dynamically modulate the expression of each gene in their genome, plays a pivotal role in nearly every cellular process and is among the most important molecular pathways underlying variation in complex traits [1–9]. Transcription is controlled by at least two classes of transcription factor protein, which bind to characteristic DNA sequence motifs and work in concert to tune the rates of early steps during the RNA polymerase II (Pol II) transcription cycle [10]. First, over two thousand transcriptional activators and repressors are encoded in the human genome, each of which binds a characteristic DNA sequence motif in a cell-type specific context [11]. Second, general transcription factors (GTFs) in the Pol II preinitiation complex (PIC) bind highly degenerate core promoter motifs [12], the best characterized of which include the TATA box [13] and the initiator element [14]. DNA sequence motifs for transcriptional activators and GTFs are found at both promoter and enhancer regions, and appear to have a role in driving both varieties of regulatory activity [15, 16].

Despite fairly advanced knowledge about the proteins which control transcription, our understanding of how genomes encode regulatory activity remains limited. Although many of the DNA sequence motifs involved in transcription factor-DNA interactions are known [17–19], strong matches to these sequence motifs in genomic DNA are surprisingly rare [20, 21]. Both transcriptional activators and GTFs frequently bind degenerate, low-affinity DNA sequences that are challenging to distinguish from unbound genomic DNA, even in *cis*-regulatory elements that control critical transcription programs [22, 23]. One potential way to reconcile specific binding to low-affinity DNA sequence motifs is that transcription factor binding sites are organized in a stereotypical pattern, such that individual DNA sequence motifs (by analogy, words) are found in the context of a longer regulatory syntax (by analogy, sentences) [22–25]. Classical examples report a structured order and orientation of DNA sequence motifs at an evolutionarily-conserved *IFNB1* enhancer [26] and at several enhancers controlling patterning during *Drosophila* development [27, 28]. Although hints in the literature suggest that syntax is critical for regulatory function, both the general principles and the impact of regulatory syntax on transcription remain almost completely unknown.

Here we investigated the regulatory syntax of transcription initiation using CLIPNET, a sensitive deep learning model trained to predict transcription initiation in mammalian cells using population-scale PRO-cap data. Interpretation of CLIPNET using gradient and *in silico* mutagenesis-based approaches revealed a core regulatory syntax consisting of five positions located between −200 and +50 bp of the transcription start site with evidence of important DNA-protein interactions, which we interpret as the binding sites for transcriptional activators and general transcription factors. Notably, although the majority of promoters and enhancers lack a canonical TATA box, they nevertheless had DNA sequence motifs mediating interactions between DNA and the PIC that collectively explain their initiation profiles. Finally, we find evidence that transcriptional activators and general transcription factors are often highly specialized for controlling either the quantity or position of transcription initiation, suggesting new models for a division of labor among transcription-related proteins.

# Results

## CLIPNET predicts transcription initiation from regulatory sequence

We developed CLIPNET (Convolutionally Learned, Initiation-Predicting NETwork) to investigate how DNA sequence controls the position of transcription initiation. CLIPNET is a deep learning model trained to predict nucleotide resolution maps of transcription initiation from a matched DNA sequence. We trained CLIPNET using a dataset consisting of matched precision run-on and 5'-capped ($\mathrm{m}^7\mathrm{G}$) RNA enrichment (PRO-cap) [29] and individual genomes [30] from 58 genetically distinct lymphoblastoid cell lines (LCLs) (Fig. 1A). This dataset has three major advantages that could improve out-of-sample predictions about the impact of DNA sequence on initiation. First, PRO-cap resolves transcription initiation at all transcriptionally active *cis*-regulatory elements without the confounding influence of mRNA degradation rates by sequencing capped RNAs associated with an active Pol II. Second, this dataset is focused on a single trans environment, LCLs, which should allow the model room to encode cell-type specific DNA sequence motifs like transcriptional activators and repressors. Third, this dataset provides a resource with matched PRO-cap and DNA sequence data, which improves [31–34] over the standard practice of using a haploid reference genome as the sole source of input DNA [35–41].

CLIPNET's architecture incorporates recent advances in predicting genome-wide molecular assays at single nucleotide resolution, most notably those utilized in BPNet [35] and APARENT 1 [42] and 2 [43]. Briefly, CLIPNET consists of two convolutional layers, followed by a tower of dilated convolutions separated by skip connections (Fig. 1B). We decomposed the output into signal profile (i.e., the distribution of PRO-cap reads within a 500 bp window) and quantity (i.e., total read coverage) and utilized a multiscale loss function to separately optimize the predicted profile and quantity of initiation. Inspired by the ensembling strategy employed by Borzoi [41], we partitioned the human genome along chromosomal boundaries into 10 roughly equally-sized folds. We then trained 9 replicate models, each using a distinct hold-out dataset, with one data fold (consisting of chromosomes 9, 13, 20, and 21) being completely withheld and reserved for final benchmarking of the ensembled model. In addition to enabling model ensembling, this model training approach allowed us to fairly benchmark the performance of the ensemble model on completely held-out data, evaluate individual model predictions at every position in the genome, and assess variability in learned feature importance.

Several complementary lines of evidence indicate that CLIPNET accurately learned the sequence basis of transcription initiation. First, CLIPNET achieved high concordances (median ensemble Pearson's $r = 0.760$, individual models Pearson's $r = 0.644 - 0.680$) between observed and predicted PRO-cap tracks in each of the 67 libraries on the held-out chromosomes (Fig. 1C, Supplementary Fig. S1A). Notably, it significantly outperformed a naive, average profile predictor (median Pearson's

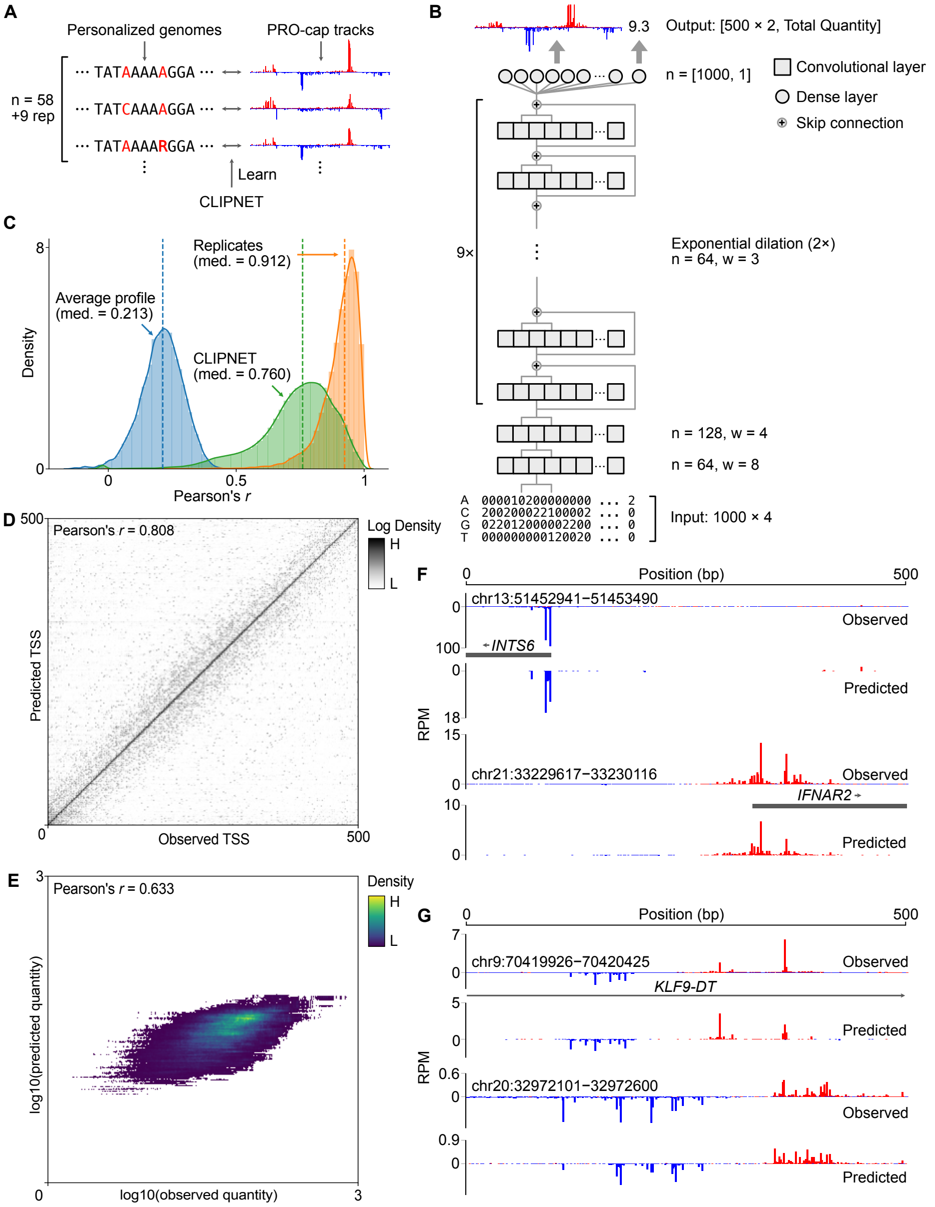

**Fig. 1 CLIPNET accurately predicts transcription initiation.** (**A**) CLIPNET is trained on personalized genomes and PRO-cap tracks from 58 LCLs (+9 replicates). (**B**) Schematic of the architecture of CLIPNET. Two convolutional layers are followed by 9 dilated convolutions with skip connections. CLIPNET separately imputes single nucleotide resolution PRO-cap profiles and total PRO-cap quantities of 500 bp windows using the surrounding 1 kb of genomic sequence. (**C** - **E**) CLIPNET predicts initiation profile (**C**), TSS position (**D**), and initiation quantity (**E**) with high accuracy. (**F** - **G**) Example predictions of promoters (**F**) and enhancers (**G**).

$r = 0.213$, whole genome) and approached experimental replication (median Pearson's $r = 0.912$, held-out chromosomes). Second, CLIPNET accurately predicted the exact position of the main transcription start site (TSS), that is, position with the highest PRO-cap signal within each 500 bp prediction window (ensemble Pearson's $r = 0.808$, individual models Pearson's $r = 0.712 - 0.768$) (Fig. 1D, Supplementary Fig. S1A). Indeed, the sequence logo of the predicted TSSs closely resembled the most common human initiator dinucleotide (Supplementary Fig. S1C), recovering perhaps the best characterized sequence feature of transcription initiation [44]. Third, the total quantity of transcription initiation in the 500 bp window was well correlated with the model predictions (ensemble Pearson's $r = 0.633$, Fig. 1E; individual models Pearson's $r = 0.532 - 0.608$, Supplementary Fig. S1A). Fourth, visual inspection supported a remarkably strong correspondence between experimental data and CLIPNET predictions at both promoters (Fig. 1F) and enhancers (Fig. 1G). Taken together, these results strongly indicate that CLIPNET learned how the sequence of *cis*-regulatory elements encodes patterns of transcription initiation.

## Distinct DNA sequence architecture controls initiation quantity and profile

We next sought to identify the DNA sequence features that are most informative in predicting transcription initiation. We used DeepSHAP [45] to quantify the contribution of individual nucleotides within a given input sequence to CLIPNET predictions. As CLIPNET separately predicts base-resolution tracks of transcription initiation and the total quantity of initiation within a given 500 bp window, we computed DeepSHAP scores for both the profile and quantity output nodes.

Examination of DeepSHAP tracks revealed that multiple DNA sequence motifs are often required to accurately predict transcription initiation at both promoters and enhancers. For example, the promoter of *IRF7* contains at least five distinct DNA sequence motifs driving quantity or profile: an SP/KLF motif, an ETS motif, an NFY motif, a TATA box, and an initiator dinucleotide (Fig. 2A). Transcription initiation at the ENCODE candidate enhancer EH38E3485200 appears to be driven by at least four distinct motifs: two ETS, one SP/KLF, and one NRF1 (Fig. 2B). Using DeepSHAP attribution scores also recovered rare DNA sequence motifs, such as the TCT motif in the promoter of ribosomal protein coding genes [46] (Supplementary Fig. S2A, B, Supplementary Info. 4), a DNA sequence preference that is both rare and associated with an unusual TBP-independent transcriptional mechanism [47–49].

We observed a striking discordance between DeepSHAP scores explaining the profile and quantity of initiation. Specifically, CLIPNET interpreted core promoter motifs at both loci (the TATA-like motif and the $\mathrm{GA}_{\mathrm{TSS}}$ dinucleotide at the *IRF7* promoter and the $\mathrm{TG}_{\mathrm{TSS}}\mathrm{T}$ trinucleotide at the enhancer EH38E3485200) as being being the primary determinants of the profile of transcription initiation. By contrast, the relative importance of the sequence-specific transcription factor motifs present at these two *cis*-regulatory elements are highly reduced in the profile DeepSHAP scores, suggesting that these two classes of regulatory motifs and their protein-binding partners play distinct roles in determining transcription initiation at *cis*-regulatory elements.

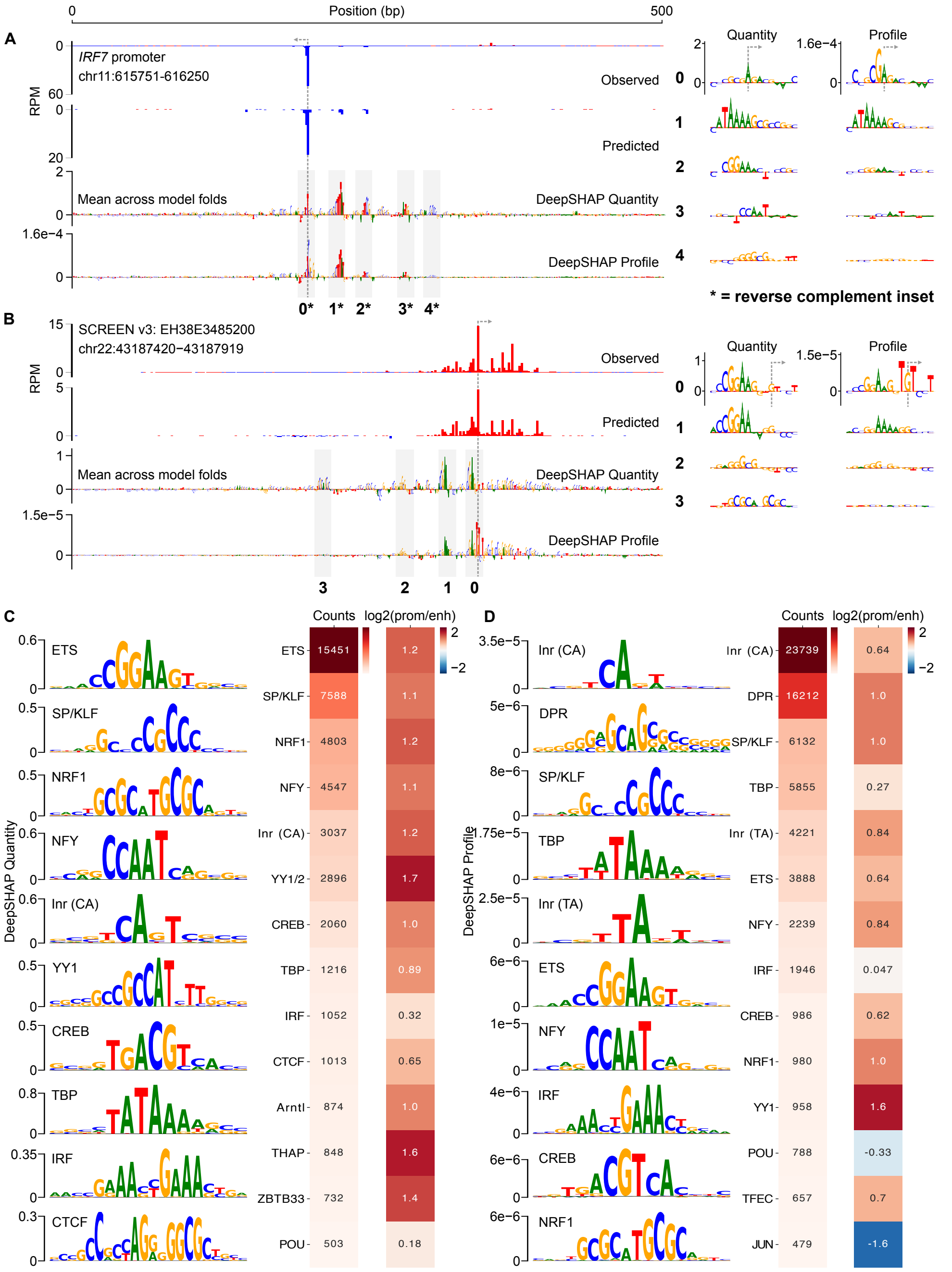

**Fig. 2 Initiation profile and quantity have distinct sequence determinants.** (**A**, **B**) Predictions and DeepSHAP contribution scores for the *IRF7* promoter (**A**) and the EH38E3485200 enhancer (**B**). Interesting motifs are highlighted in insets to the right. DeepSHAP was used to interpret both the profile and quantity predictions, resulting in distinct contribution score tracks and sequence motif prioritizations (inset right). (**C**, **D**) To gain a genome-wide view of the sequence determinants of initiation profile and quantity, DeepSHAP profile and quantity scores were calculated for over 200,000 *cis*-regulatory elements. TF-MoDISco was then used to identify informative sequence motifs for initiation quantity (**C**) and profile (**D**). A selection of 8 highly prevalent motifs for each are shown in the main figure. Motif counts in promoters and enhancers and the promoter:enhancer frequency ratios were also calculated and displayed.

To identify classes of informative motifs, we used TF-MoDISco [50, 51] to cluster common DNA subsequences contributing to either profile or quantity. The majority of DNA sequence motifs important for predicting the quantity of transcription initiation resembled the known consensus motifs of strong transcriptional activators (Fig. 2C), including those recognized by both ubiquitously expressed (e.g., SP/KLF, YY1, and CREB) and cell-type specific (ETS, NRF1, and IRF4) transcription factors. We also found the CA initiator dinucleotide, the TATA box (Supplementary Fig. S2C), and a large number of degenerate CpG-rich motifs in promoters (Supplementary Fig. S2D).

By contrast, the profile of transcription initiation was best explained by core promoter motifs (Fig. 2D, Supplementary Fig. S2E): the TATA box, several distinct initiator motifs, and a heterogeneous collection of motifs representing the downstream promoter region (DPR) (discussed further below). The most common initiator motif consists of a CA dinucleotide followed by either an A or a T in the TSS+2 position (Fig. 2D), consistent with the previously reported BBCA$_{\mathrm{TSS}}$BW initiator motif [52, 53]. We also identified a number of rarer initiator motifs, including a TA dinucleotide, which has been described as the second most common initiator after CA in mammals [52, 54]. Finally, CLIPNET also attributed a role to transcriptional activators in shaping initiation profile (Fig. 2D). However, the contribution scores of transcriptional activators were consistently several-fold lower than those of the TATA box or initiator, indicating that they play relatively minor roles in controlling initiation profiles.

## Conserved DNA sequence architecture underlying promoters and candidate enhancers

To investigate the differences in the profile of transcription initiation between promoters and candidate enhancers, we split *cis*-regulatory elements into gene-proximal and distal regulatory classes. TF-MoDISco identified the same DNA sequence motifs in both promoters and enhancers (Supplementary Fig. S2C, D). Promoters had a markedly higher frequency of motifs that explain initiation quantity (Fig. 2C), many of which resemble the known binding motifs of transcriptional activators such as SP/KLF factors, IRFs, NRF1, and ETS. However, these differences predominantly reflect the higher overall transcriptional activity in promoters compared to distal enhancers (Supplementary Fig. S2F). By contrast, DNA sequence motifs explaining initiation profile, which does not systematically differ between promoters and enhancers, were much more similar in frequency between these two regulatory classes (Fig. 2D). Differences observed in the profile motif frequency were in the direction expected based on the higher G/C content in CpG-island enriched promoters. For instance, CLIPNET identified the G/C-rich YY1 and NRF1 motifs with a higher frequency in promoters and the A/T-rich IRF and POU motifs with a higher frequency in enhancers (Fig. 2D). We conclude that the DNA sequences responsible for controlling the position and abundance of transcription initiation are similar between these two classes of *cis*-regulatory elements.

## CLIPNET predicts the impact of initiation QTLs

Correctly predicting and interpreting the functional role of QTLs is a central problem in modern genetics and a difficult challenge, even for state-of-the art sequence-to-function models [32–34]. To assess CLIPNET's ability to predict the functional impact of regulatory variants, we leveraged an existing initiation QTL dataset in LCLs [29]. Kristjánsdóttir et al. previously mapped transcription initiation quantitative trait loci (tiQTLs), SNPs associated with changes in initiation quantity, and directionality quantitative trait loci (diQTLs), SNPs associated with differences in the ratio of initiation events between DNA strands, a type of difference in profile. We focused our analysis on a set of biallelic tiQTLs ($n = 2,057$) and diQTLs ($n = 1,207$). We summarized differences in transcription initiation between individuals homozygous for the reference and alternative alleles using the $L^2$ norm, a metric which captures information about allelic changes in both quantity and profile [41]. Comparing $L^2$ norms between experimental and CLIPNET predictions for each QTL showed that the difference between alleles were reasonably well-correlated across both tiQTLs (Pearson's $r = 0.48$; Fig. 3A) and diQTLs (Pearson's $r = 0.54$; Fig. 3B).

Examination of individual loci showed that CLIPNET accurately predicted changes in both the quantity and profile of several distinct types of ti- or diQTLs, including large focal changes in a single initiation site, or changes in initiation affecting multiple initiation sites in complicated promoters. For example, rs185220 is a tiQTL which disrupts both initiation sites in a divergent pair on the plus and minus strand, leading to a substantial decrease in the quantity of transcription initiation. This effect was largely recovered by CLIPNET and attributed to the loss of a strong SP/KLF binding site on the minor allele (Fig. 3C, Supplementary Fig. S3A). By contrast, the diQTL rs8050061 was associated with a localized impact on initiation at a specific nucleotide, an effect which was also recovered by CLIPNET and explained by a disruption to an initiator motif overlapping the affected position (Fig. 3D, Supplementary Fig. S3B). Collectively, our analyses demonstrate that CLIPNET can predict how DNA sequence changes impact both the quantity and profile of transcription initiation with reasonably high accuracy, and does so by correctly interpreting the effects of different classes of regulatory motifs.

## Five distinct DNA-protein interactions form the core syntax of transcription initiation

Having shown that CLIPNET predicts the impact of DNA sequence changes on transcription initiation with reasonably high accuracy, we decided to use *in silico* mutagenesis to explore how DNA sequence features influencing transcription initiation are organized at *cis*-regulatory elements across the genome. We performed *in silico* mutagenesis on 5,000 random *cis*-regulatory elements by mutating every 10 bp window between −200 and +200 bp of the PRO-cap-defined max TSS to a random sequence (Fig. 4A, "ISM shuffle" [41]). Mutations between −125 and +50 bp had, on average, the largest impact on both the profile and quantity of transcription initiation, indicating the critical importance of this region for specifying transcription (Fig. 4B).

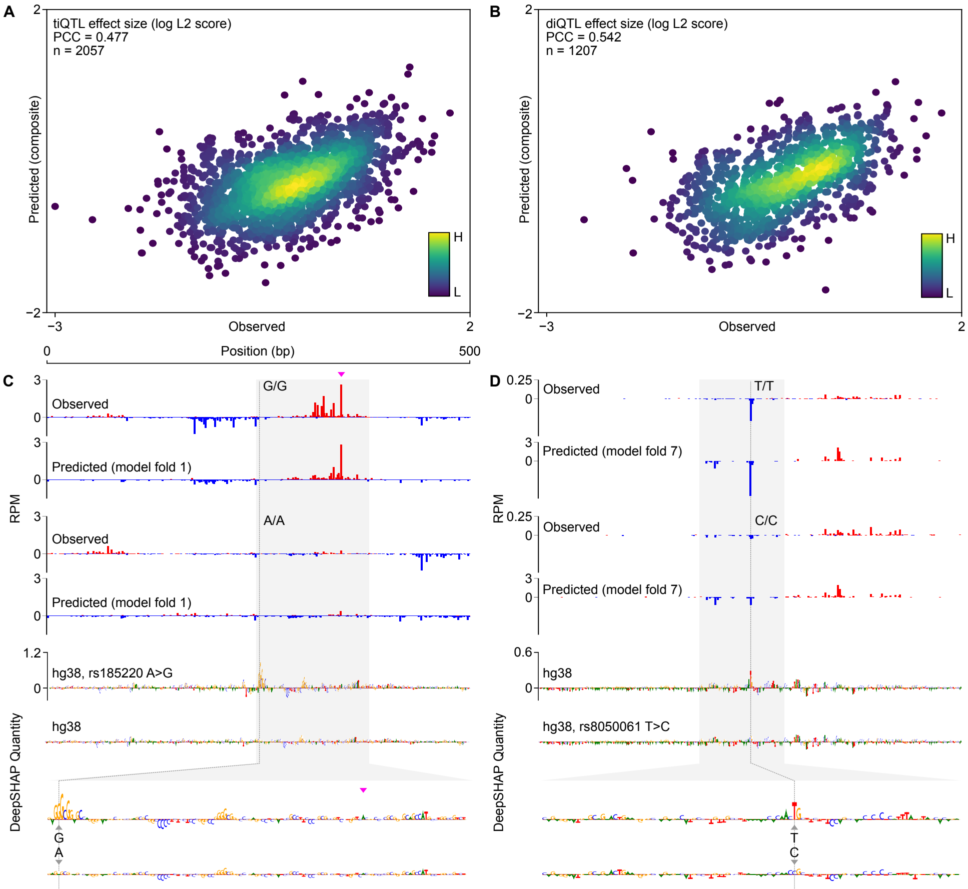

**Fig. 3** **CLIPNET correctly predicts regulatory variant effects on both transcriptional activator and core promoter motifs.** (**A**, **B**) Predicted versus observed tiQTL (**A**) and diQTL (**B**) effects. (**C**, **D**) CLIPNET accurately predicts the effects of a tiQTL (**C**) (rs185220) and of a diQTL (**D**) (rs8050061) by recognizing that these two variants impact distinct types of regulatory motifs (DeepSHAP quantity scores bottom, profile scores in Supplementary Fig. S3). The plus strand TSS near rs185220 is highlighted with a magenta arrow.

From these ISM shuffle profiles, we identified five distinct positions within this window each having a characteristic impact on the profile or quantity of transcription initiation (Fig. 4B, C). Three of these reflect DNA-protein interactions in the core promoter region between TSS −25 and +25 bp, which directly interact with the PIC [55–57]. The most important DNA element controlling initiation profile occurs at the TSS, and reflects the initiator element. Interactions at TSS −25 and +25 bp are also relatively important for controlling transcription profile. The mode at −25 bp corresponds to interactions between DNA and TBP, a protein in TFIID which binds the TATA box [12]. The mode at +25 bp corresponds to the mammalian DPR, a DNA sequence motif well-characterized in *Drosophila* [21, 58], but which was only recently reported in human cells [59, 60].

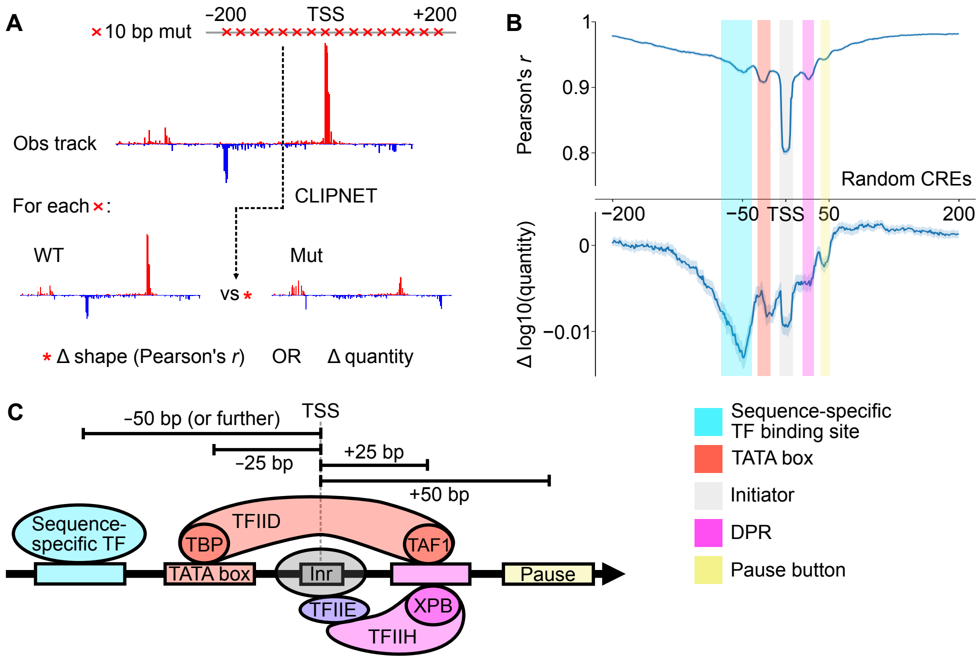

**Fig. 4 ISM shuffle of *cis*-regulatory elements reveals core promoter structure.** (**A**) We used ISM shuffle to characterize the importance of sequence elements surrounding the TSS at many *cis*-regulatory elements across the genome. (**B**) ISM shuffle applied to 5,000 random *cis*-regulatory elements identified the −125 bp to +50 bp region as the most important for determining both initiation profile and quantity. We tentatively identified the 5 major peaks in the ISM shuffle tracks as those corresponding to transcriptional activator binding sites, the TATA box, the initiator, DPR, and a pause button. (**C**) Schematic illustrating 5 major classes of motifs that impact transcription initiation.

Motifs at −50 and +45 bp were too far away from the TSS to bind directly to core PIC components. Motifs at +45 bp correspond roughly to the position at which Pol II pauses [61], and may reflect interactions between the pause complex and DNA [52, 62, 63], or they may reflect unknown interactions between DNA and the Pol II elongation complex as it comes up to speed [64–66]. DNA sequence motifs located at TSS −50 bp were the most important determinant of transcription quantity. This is consistent with previous observations of the binding patterns of many transcriptional activators [15, 61, 67], and with the distribution of the transcriptional activator motifs identified by TF-MoDISco (Fig. 5A). In contrast to sequences closer to the TSS, (Fig. 4B), we found that these more upstream sequence motifs have a stronger impact on initiation quantity than profile. These results highlight the diversity of regulatory motif position and function, with TSS-proximal core promoter motifs appearing to primarily drive initiation profile and upstream transcriptional activator motifs determining initiation quantity.

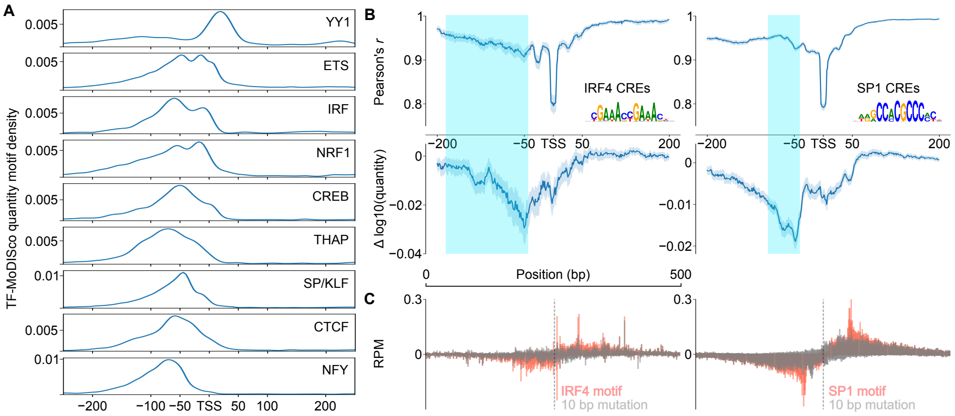

**Fig. 5 Diversity in positioning of transcriptional activators.** (**A**) Distribution of common transcriptional activator motifs identified by TF-MoDISco (quantity) around TSSs. (**B**) ISM shuffle applied to *cis*-regulatory elements containing motif matches for IRF4 (left) and SP1 (right). Interquartile range of motif locations (consensus motif, FIMO) are highlighted. (**C**) Metaplots of motif-directed mutagenesis of IRF4 motifs (left) and SP1 motifs (right).

## Transcriptional activators drive the quantity of transcription under positional constraints relative to the core promoter

To further explore the roles of different transcriptional activator motifs, we conditioned on subsets of DNA sequence motifs which were confirmed to bind specific transcription factors (based on ChIP-seq data in LCLs [68]) and also carry a strong match to the DNA sequence consensus motif. Analysis of two transcriptional activators, IRF4 and SP1, revealed similar ISM shuffle profiles to the random set; namely mutations between 200 and 50 bp upstream of the TSS, where the IRF4 and SP1 DNA sequence motifs were most commonly found, had the largest impact on transcription initiation quantity (Fig. 5B). Targeted mutagenesis specifically disrupting the IRF4 or SP1 DNA sequence motifs showed striking, bidirectional changes in the quantity of transcription that were symmetric and centered on the DNA sequence motif (Fig. 5C).

While activator motifs had a mode at −50 bp, they showed higher activation over a fairly broad window between 125 and 50 bp upstream of the TSS (Fig. 4B). At least part of this variability reflects differences between transcriptional activators. For instance, SP1 binding sites were found in a fairly narrow window between −50 and −75 bp (Fig. 5B; teal shade denotes the interquartile range in the position of the motif), and mutation had a focused yet bidirectional impact on initiation ~50 bp from the motif (Fig. 5C). By contrast, IRF4 binding sites were scattered over a much broader window between −50 and nearly −200 bp relative to the TSS (Fig. 5B, left, teal shade denotes motif interquartile range), and had a broader bidirectional impact on initiation over ~100 bp (Fig. 5C).

To examine the possibility of a more transcriptional activator-specific syntax over a broader set of activators, we examined histograms of the position of each TF-MoDISco motif contributing to initiation quantity. Positional enrichments for IRF and SP/KLF-like TF-MoDISco motifs were similar to those based on ChIP-seq validated consensus motif matching, with a broader distribution in IRF while SP/KLF occupied a more focal position (Fig. 5A). Across 9 different transcriptional activators, CLIPNET found evidence of distinct positional preferences, with some motifs binding close to, or even downstream of, the TSS (e.g., ETS, NRF1, YY1), while others had a stronger preference for either the −50 bp position or even further upstream (e.g., SP1, NFY) (Fig. 5A). YY1 was the most distinct, with most of its motifs occurring downstream of the TSS (Fig. 5A). These results hint that different transcriptional activators have distinct positional syntaxes relative to the primary TSS.

## PIC-DNA structural interactions govern nucleotide importance

Although we know the optimal DNA sequence motifs that interact with the PIC (TATA box and initiator), these motifs are found at only a small fraction of human promoters and enhancers [12]. Conversely, CLIPNET accurately identified the initiation profile at nearly all active *cis*-regulatory elements genome-wide. To gain a broader understanding of the DNA sequence basis of transcription initiation, we analyzed previously published cryogenic electron microscopy (cryoEM) structures of the mammalian PIC assembled on an artificial super core promoter (SCP) containing a TATA box, an initiator, and a DPR. We analyzed three PIC structures that are believed to represent three sequential stages of PIC assembly: the core PIC (cPIC; TFIID, A, B, and F), intermediate PIC (mPIC; cPIC + TFIIE), and holo PIC (hPIC; mPIC + TFIIH) [69]. DeepSHAP profile attribution of the sequence of the SCP recovered a well-positioned initiation site driven by the DNA sequence of all three major core promoter elements: a TATA box, an initiator, and a DPR (Fig. 6A, top).

To measure the physical interactions between core promoters and the PIC, we measured the minimum distance between each nucleotide in the SCP and any amino acid in each of the three PIC structures. The most consistent DNA-protein interactions were with the TATA box, which was located within 5 Å of TBP in all three PIC structures (Fig. 6A). In contrast, interactions between DPR and the PIC structure were much more variable (Fig. 6A). CLIPNET attributed a high importance to the end of the DNA sequence annotated as the DPR, which was also consistently within 5 Å of the PIC. Conversely, CLIPNET preferentially recognized the importance of nucleotides comprising the DPE (the second half of DPR), which were closest to the intermediate PIC (mPIC), but had much more variable interactions with the cPIC and mPIC. The DPR has a similar importance to the TATA box in *Drosophila* promoters [21, 58], but was only recently identified in humans [59, 60], and has a DNA sequence basis that remains obscure as of this writing. This variability in PIC-DNA structural interactions could explain the weaker DNA sequence preference of DPR, and hence why the human DPR sequence has remained so elusive [59, 60].

## DNA sequence specificity of the human DPR

Pioneering studies have shown that DPR is of similar importance to the TATA box in *Drosophila* promoters [21, 58, 70], but the importance of the human DPR has been more challenging to pin down. When we examined the sequence motifs identified by TF-MoDISco profile, we discovered a collection of similar DNA sequence motifs that

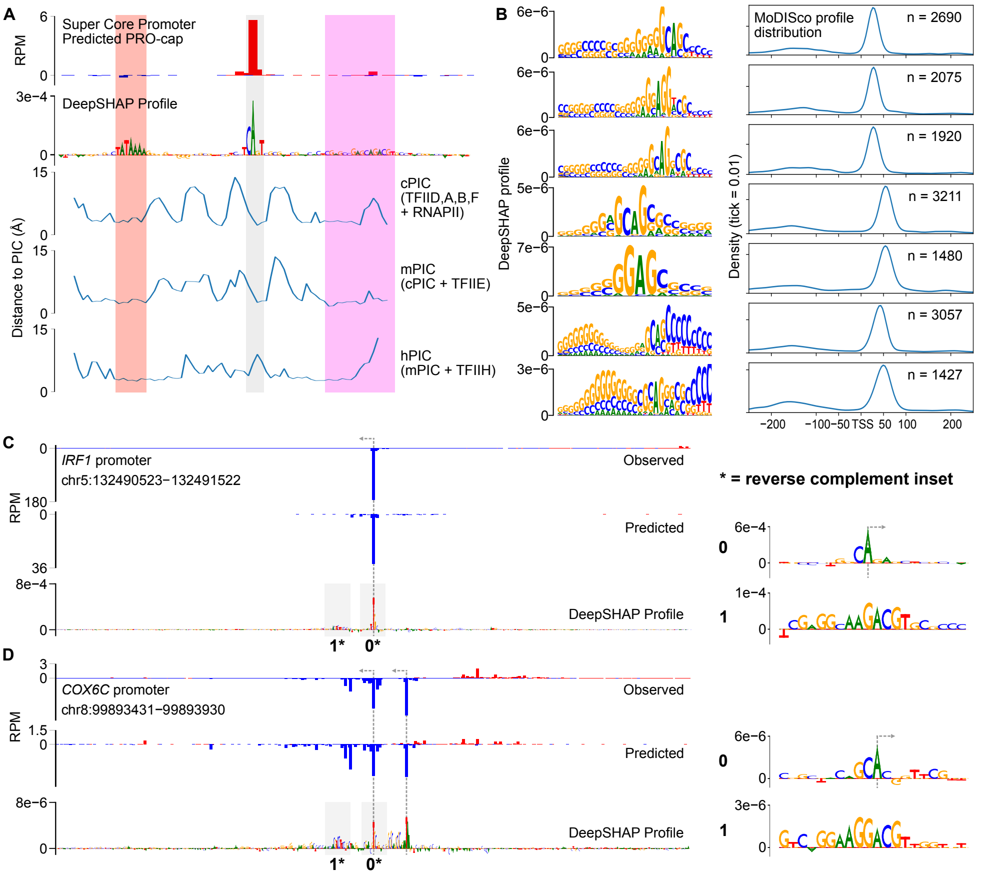

**Fig. 6 DNA sequence specificity of the human DPR.** (**A**) Analysis of protein-DNA contacts of three stages of PIC assembly onto an artificial SCP. We observed a correspondence between the regions of closest contact and the core promoter motifs TATA (red), Inr (grey), and DPR (magenta) identified by DeepSHAP profile. (**B**) Examples of 7 DPR motifs identified by TF-MoDISco profile. We found four (top) enriched in the canonical TSS +25 position and three (bottom) enriched at the +50 position. (**C**, **D**) Predictions and DeepSHAP profile scores for two DPR-driven promoters identified, one TATA-less (**C**) and one TATA-containing **D**). The initiator and DPR motifs are highlighted in insets to the right.

are found in the position of the DPR, approximately 25 base pairs downstream of the TSS (Fig. 6B). While most of these motifs occurred primarily at the TSS +25 bp position, we also identified another set of similar sequence motifs occurring further downstream, at the TSS +50 bp position (Fig. 6B). The common core sequence was G(A/G)AG, similar to a recent mammalian DPR motif representation [59, 60], but considerable degeneracy was present in the motifs (Fig. 6B). We also observed an extended GC rich stretch of DNA upstream of the core sequence motif (Fig. 6B). DPR was more common than a canonical TATA box, occurring about twice as frequently (Fig. 2D, Fig. 6B). We identified examples of *cis*-regulatory elements in which DPR appeared both independently of (Fig. 6C) and together with (Fig. 6D) a TATA box. Curiously, we did not find DPR-like motifs using TF-MoDISco quantity, suggesting that this core promoter motif is not a particularly strong driver of transcription quantity, and instead primarily serves to drive Pol II positioning.

## TBP binds the most AT-rich sub-sequence in a promoter or enhancer

TBP is known to be both essential for transcription initiation and to strongly bind to the TATA box. However, fewer than 15% of promoters contain a TATA box [12, 44], and the sequence specificity of TBP binding at TATA-less promoters remains unknown. We noticed that CLIPNET devoted a large fraction of layer 2 neurons (width = 15 bp) to learning DNA sequences ~25 bp upstream of the TSS (Supplementary Fig. S4A), and reasoned that these may indicate that CLIPNET is learning diverse, degenerate binding patterns at TATA-less promoters. Filters enriched in the $-25$ bp position recognized DNA sequence with a gradation of GC content $(0.1 - 0.5)$ (Supplementary Fig. S4A), consistent with previous observations of enrichment for AT-rich sequences at TBP binding sites.

To explore the sequence specificity of TBP binding, we first used saturation mutagenesis to verify whether CLIPNET can correctly identify the importance of a strong TATA box to driving transcription initiation. We performed saturation mutagenesis on the TATA box of 302 TATA-containing *cis*-regulatory elements (Fig. 7A). This analysis identified the canonical TATAWAWR sequence as the least disruptive to both initiation profile and quantity (Fig. 7B), consistent with experimental saturation mutagenesis [71, 72]. G or C substitutions in positions 2 through 5 (i.e., ATAW) of the canonical motif were especially disruptive, while A or T substitutions were more likely to be tolerated. Moreover, when we replaced the entire TATA box with a random 8-mer, we observed a distinct, asymmetric mutational effect, with only the downstream TSS being impacted (Supplementary Fig. S4B). This is in stark contrast with the bidirectional effect of IRF4 or SP1 that we observed previously in this paper, but is consistent with previous analyses of diQTLs in TATA boxes [29].

We hypothesized that an AT-rich patch of DNA in a core promoter is sufficient to aid TBP binding in the absence of a canonical TATA box. To test this model, we replaced the 8 bp window corresponding to the TATA motif with random DNA, controlling the GC content (Fig. 7A). Higher GC content was nearly always correlated with a higher disruption to the transcription initiation profile (Fig. 7C, top, Supplementary Fig. S4C). Intriguingly, any disruption of a strong TATA box had a large impact on the overall transcription quantity, with no additional effect on quantity as GC content increased (Fig. 7D, bottom, Supplementary Fig. S4C). These results are consistent with a model in which a relatively strong match to the TATA consensus is required for a TATA box to substantially impact transcriptional output, but just a short AT rich sequence at the $-25$ bp position is sufficient to position the PIC by interacting with TBP and establish the position at which transcription initiates.

To determine whether AT rich sequences can position the PIC in endogenous TATA-less *cis*-regulatory elements, we considered CpG islands *cis*-regulatory elements (which are overwhelmingly TATA-less promoters) without a canonical TATA box. Plotting the GC content relative to the position of the max TSS showed a window of

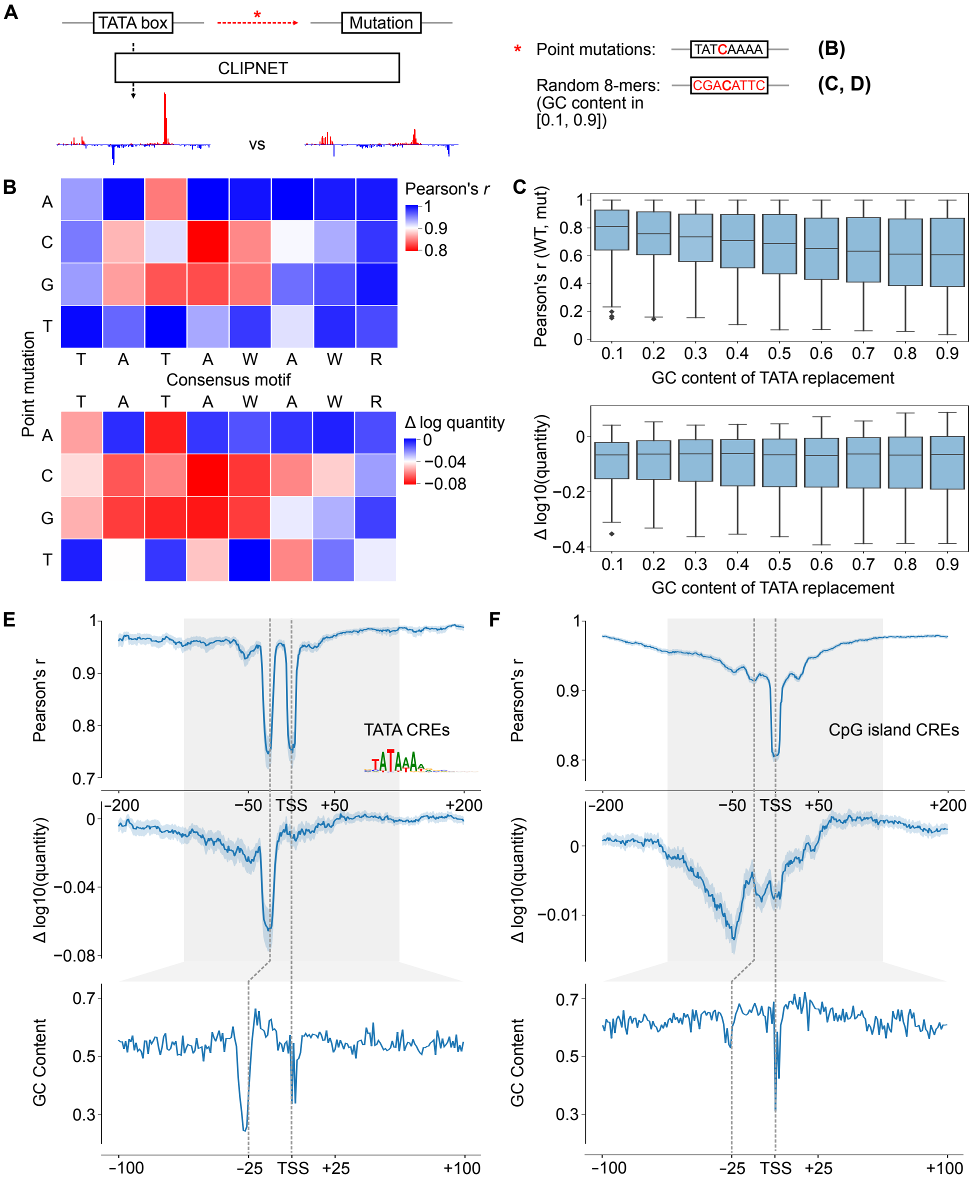

**Fig. 7 DNA sequence specificity of TBP binding.** (**A**) We used targeted *in silico* mutagenesis to measure the sequence properties of the TATA box. We performed both site saturation (B) and random substitutions of 8-mers sampled to have specific GC contents (“GC shift”) (**C**). While the effects of site saturation mutagenesis on initiation profile (top) and quantity (bottom) were relatively similar, the effects of GC shift mutagenesis were quite distinct between profile (top) and quantity (bottom). (**D**, **E**) Relationship between importance of the TBP binding site measured by ISM shuffle (top and middle) and GC content (bottom) at TATA-containing (**D**) and TATA-less CpG island (**E**) *cis*-regulatory elements.

decreased GC content at the −25 bp position, indicating that even CpG island promoters have a relatively AT-rich sequence patch in the position where TBP binds (Fig. 7D-E). To determine whether this position plays a role in the profile and quantity of transcription initiation, we used ISM shuffle to measure positional sequence importance in TATA-containing and TATA-less CpG island *cis*-regulatory elements (Fig. 7D-E). As noted above, mutating the window containing the TATA box had a large impact on both the shape and quantity of initiation in TATA containing *cis*-regulatory elements (Fig. 7D). Mutating DNA in the −25 position of CpG island promoters impacted the correlation, second in magnitude only to the initiator element, and less of an impact than surrounding DNA on initiation quantity (Fig. 7E). These findings suggest that TBP binds most strongly to a canonical TATA box and increases transcriptional output when available, but otherwise will bind the most AT-rich sequence in the vicinity of a promoter and help establish the position of the PIC and ultimately Pol II initiation.

# Discussion

Despite an advanced lexicon of the DNA sequence motifs (i.e., words) that regulate transcription, we still have very little understanding of the syntax with which these motifs are organized (i.e., the structure of sentences). Several classical examples are known in which the order and orientation of DNA sequence motifs are crucial for regulatory function [26–28]. Despite these case studies, however, the general properties of regulatory syntax have proven much more challenging to pin down and the extent to which syntax is important for regulatory function at the majority of *cis*-regulatory elements remains debated [24, 25].

Previous work on regulatory syntax has focused on interactions between transcriptional activators [22, 23, 35]. Our work builds on this concept by demonstrating surprisingly strong and systematic positional dependencies between binding sites recognized by transcriptional activators and the core promoter motifs which ultimately specify the transcription start site. Our findings build on observations that transcriptional activators are enriched in the central region between divergent transcription start site pairs [15, 61, 67, 73]. We report that these positional dependencies have considerable variability between different transcription factors, most notably captured in our study for IRF4 and SP1 binding sites. Different transcription factors have distinct functional roles in regulating different stages in the Pol II transcription cycle: some are pioneer factors that open chromatin [74–76], while others catalyze the release of Pol II from a paused state into productive elongation [64, 77, 78]. We speculate that structural constraints imposed by the transcription factor's functional role underlies these different positional requirements on the binding position of transcription factors relative to the PIC. Moreover, the dependencies between different functional classes of transcriptional activators and the PIC could make them a more general feature of regulatory syntax than interactions between different cell-type specific transcription factors.

Our work has also built substantially on our knowledge of the DNA sequence motifs that specify the location of the PIC and the position of transcription initiation.

The majority of human promoters do not have strong matches to the best known two core promoter elements: the TATA box and the initiator [12]. While many other core promoter motifs have been identified in a variety of model organisms [21, 58, 59, 70, 79, 80], the DNA sequence composition and importance in mammalian promoters remains a subject of extensive debate. Our work identified a larger, more diverse, and more degenerate group of DNA sequence motifs that are collectively responsible for specifying the profile, or precise position of transcription initiation at all promoters and enhancers genome-wide. Perhaps most notably, CLIPNET identified a purine-rich DPR sequence preference (most commonly a G(A/G)AG motif), primarily impacted initiation profile rather than quantity, and appeared to have two distinct positional preferences at the TSS +25 and +45 positions. These constraints may explain why the sequence of DPR has remained so challenging to identify.

We also report a strong role for the TATA box and more degenerate TATA-like motifs. We found evidence that an AT-rich DNA sequence contributes to transcription initiation profile at many TATA-less promoters. This finding may explain how TBP binds DNA at the −25 position, even in promoters that do not contain a TATA box [81]. These DNA sequence preferences appear to reflect simply the most AT-rich DNA sequence in each *cis*-regulatory element, and affect initiation profile much more than quantity. These findings are consistent with a model in which the best available binding site for TBP, in conjunction with DPR and Inr, are collectively responsible for positioning a pool of Pol II that is assembled by other transcription-regulated proteins binding in their vicinity.

Finally, our study indicates a "division of labors" by which different types of transcription factors have a synergistic role on transcriptional output. Our results support a model in which transcriptional activators, and to some extent a strong TATA box, establish the abundance of initiation, perhaps by recruiting a pool of transcriptional proteins or clearing chromatin, while core promoter motifs bound by GTFs guide the assembly of the PIC and the precise location of transcription initiation. These observations explain how transcription initiation can be simultaneously driven by multiple protein complexes that collaborate to clear chromatin, recruit proteins necessary for transcription, assemble and position the PIC, and begin transcription.

# Methods

## Training data processing

Aligned PRO-cap data from 67 genetically distinct LCLs (+ 10 replicates) were downloaded from Gene Expression Omnibus accession GSE110638. Phased genotypes were downloaded obtained from the 2019 1000 Genomes Project release (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/). As 9 individuals were not included in this particular release, they were excluded from this study, resulting in 67 total PRO-cap libraries (58 individuals + 9 replicates, Supplementary Info. 1). To ensure consistency between the PRO-cap and genotyping data, we lifted over the PRO-cap libraries from their original hg19 reference to hg38.

We generated individualized genomes by applying the genotyped SNPs from each individual to the hg38 reference genome. Indels and structural variants were excluded as they were relatively rare and could introduce index shifts that would require remapping of the entire PRO-cap dataset and render QTL analyses significantly more difficult to perform. PRO-cap peaks were individually called in each library using a pre-publication version of PINTS [82] supplied by the authors. As *cis*-regulatory elements commonly consist of two divergently transcribed core promoters spaced roughly 110 bp apart15, we filtered for peaks that were no more than 200 bp away from a peak on the divergent strand. We then extracted 1kb of matched genomic sequence and PRO-cap tracks around the center of each PINTS call. To reduce overfitting, we randomly jittered the position of these windows by up to 250 bp.

To enable model ensembling, we partitioned the genome along chromosomal boundaries into 10 roughly equally sized folds. We set aside fold 0 (consisting of chromosomes 9, 13, 20, and 21) for final evaluation of the model ensemble. The remaining 9 folds were then used to train 9 replicate models, each of which used a distinct holdout fold. This ensures that prediction quality at each position within the genome can be fairly evaluated using individual models.

One-hot encoding is the standard for genomic deep learning models; however, as each individual will be heterozygous at many SNPs, we had to take a slightly different sequence encoding approach to be able to represent individualized genomic sequences. Instead, we used a two-hot encoding; that is, we encoded each individual nucleotide at a given position using a one-hot encoding scheme, then represented the unphased diploid sequence as the sum of the two one-hot encoded nucleotides at each position. The sequence AYCR, for example, would be encoded as [[2, 0, 0, 0], [0, 1, 0, 1], [0, 2, 0, 0], [1, 0, 1, 0]]. This encoding scheme makes two simplifying assumptions that we believe are biologically reasonable: (1) additivity in the dosage effects of individual nucleotides and (2) that haplotype structure confers no additional information. While the previously published BigRNA model used a more sophisticated encoding structure to represent individual, phased sequences [31], we believe that a two-hot encoding is a reasonable simplification for working with short input sequences (1 kb).

## CLIPNET architecture and training

CLIPNET is a sequence-to-profile model that takes as input a genomic DNA sequence of length 1000 and outputs strand-specific PRO-cap coverage of the central 500 nucleotides. It is an ensemble model consisting of 9 structurally identical models, each of which used a distinct holdout set of chromosomes (Supplementary Fig. S1A, B). The main body of the individual models consists of two convolutional layers (64 filters, width 8 and 128 filters, width 4), followed by a tower of 9 exponentially dilated convolutional layers (64 filters, width 3, dilation factors from 1 to 512) separated by skip connections. Batch normalization was applied after each convolutional layer. Rectified linear activations (ReLU) were used for each convolutional layer except for the first, which utilized an exponential activation to improve interpretability [83]. Max pooling (width 2) was applied after each of the first two convolutional layers and after the dilated convolution tower.

We partitioned the output of the model into nucleotide-resolution coverage profiles and total read coverage following the approach pioneered in BPNet [35]. To accommodate this prediction strategy, we structured the output layers of the models as follows: (1) for profile predictions, we applied a dense layer. For simplicity, we concatenated the two 500 bp coverage profiles into a single length 1000 output vector. (2) To output total quantity, we applied a global average pooling layer, followed by a single dense layer. We applied batch normalization, ReLU, and dropout (rate = 0.3) at the end of each output node. We used negative cosine similarity to evaluate the profile predictions and mean squared logarithmic error to evaluate the quantity predictions. To jointly evaluate the prediction accuracies of these two output nodes, we used a multiscale loss function similar to [35, 41, 43].

Specifically, for a given 500 bp window, let $\mathbf{p}_{\text{obs}} \in \mathbb{R}^{1000}_{\geq 0}$ represent the base-resolution PRO-cap coverage and $\mathbf{y}_{\text{profile}} \in \mathbb{R}^{1000}_{\geq 0}$ and $y_{\text{quantity}} \in \mathbb{R}_{\geq 0}$ represent the profile and quantity predictions, respectively. We then calculated the loss as

$$\text{Loss} = -\cos(\mathbf{p}_{\text{obs}}, \mathbf{y}_{\text{profile}}) + \lambda(\log(\textstyle\sum \mathbf{p}_{\text{obs}} + \alpha) - \log(y_{\text{quantity}} + \alpha))^2,$$

where $\alpha = 10^{-6}$ was used as a pseudocount and $\lambda = 1/500$ was used as a balancing weight between the profile and quantity loss functions. We found that this weight allowed for highly accurate profile predictions and reasonably accurate quantity predictions, and that increasing the weight on the quantity loss did not appreciably improve quantity prediction but came at a major cost in profile prediction.

Model hyperparameters were manually tuned from reasonable starting points used by previous genomic deep learning models (cf. [35, 41–43]). CLIPNET was implemented in tensorflow [84] (version 2.13.0) and trained using the Adam optimizer (learning rate = 0.001) with early stopping (patience = 10 epochs). The best model (minimum validation loss) from each replicate was retained and used for further analyses.

## CLIPNET evaluation metrics

To fairly evaluate both individual models and the CLIPNET ensemble, we considered a set of high confidence PRO-cap peaks (Supplementary Info. 2). Briefly, we extracted genomic sequence and PRO-cap coverage around PINTS calls that were present in at least 60 out of 67 libraries (bedtools multiinter). We considered three summary metrics: (1) The median Pearson's $r$ between predicted and observed PRO-cap coverage tracks (Fig. 1C). For this metric, we represented the strand-specific, length 500 tracks as a concatenated, length 1000 vector. As Pearson's $r$ is undefined on constant inputs, we added a small amount of Gaussian noise, $\varepsilon \sim \mathcal{N}(0, 10^{-6})$, to each position within the predicted track. (2) The visual correspondence and Pearson's $r$ between the predicted and observed positions of both sense and antisense TSSs (Fig. 1D). (3) The Pearson's $r$ between the predicted and observed $\log_{10}$ quantities (Fig. 1E). To avoid taking the log of a potentially zero prediction, we added a pseudocount of $10^{-1}$ to both predicted and observed quantities. For all of these metrics, we evaluated the model ensemble on the fully withheld data fold (4901 peaks × 67 libraries). We also computed summary metrics for each of the individual models on both the fully withheld data fold and on their individual holdout data folds (Supplementary Fig. 1A). We found that while the individual models performed reasonably well, prediction accuracy was substantially improved by ensembling (Fig. 1C-E, Supplementary Fig. S1A).

Predicted and observed PRO-cap tracks at example *cis*-regulatory elements (Fig. 1F-G, Fig. 3A-B, Supplementary Fig. S3A-B) were computed as follows. We scaled the profile predictions to match the quantity predictions, then computed the average predicted and observed tracks across all 67 PRO-cap libraries. Strand-specific tracks were then visualized at each *cis*-regulatory element. Annotations were copied from the UCSC genome browser [85] and the ENCODE SCREEN database [86].

## Profile and quantity attribution using DeepSHAP

Gradient-based attribution methods are commonly used due to their computational efficiency compared to *in silico* mutagenesis approaches. We used DeepSHAP [45] (version 0.42.1), a popular and efficient gradient-based attribution method, to interpret CLIPNET. As CLIPNET has two separate output nodes, we applied DeepSHAP separately on the profile and quantity predictions. To consolidate the profile prediction into a single explainable scalar, we used the profile contribution score described in BPNet [35]. Rather than computing DeepSHAP scores for all 58 individual genomes, which would require impractically long compute times for genome-wide analyses, we instead focused our interpretation analyses on the hg38 reference genome.

Unlike the related DeepLIFT method [87], DeepSHAP calculates attribution scores with respect to a background average. Following suggestions by the authors of DeepSHAP, we used a background of 100 randomly sampled *cis*-regulatory element sequences that we dinucleotide shuffled. We calculated DeepSHAP scores for each model replicate individually, then averaged the DeepSHAP tracks, similar to the approach taken in Borzoi [41]. For the distal enhancer versus promoter comparisons displayed in Fig. 1C-D and Supplementary Fig. S2C, we used a distance cutoff of $< 200$ bp for promoters and $> 2000$ bp for distal enhancers from the PINTS peaks to a GENCODE [88] (version 43) protein coding TSS.

## Motif discovery and frequency analysis

DeepSHAP profile and quantity scores were computed on the set of high confidence PRO-cap peaks described above. The lite implementation [50] (version 2.2.0) of TF-MoDISco [51] was used to cluster high importance subsequences (seqlets) into summary motifs (seqlets per metacluster $= 100,000$), which were then matched against the JASPAR database [89] (2022, non-redundant vertebrate) using the TOMTOM algorithm [90]. We identified 116,351 quantity seqlets (115,602 positive, 749 negative) and profile seqlets 132,121 (115,989 positive, 16,132 negative), which then clustered into 62 quantity motifs (51 positive, 11 negative) and 100 profile motifs (60 positive and 40 negative). To generate promoter and distal enhancer motif frequencies (Fig. 2D, E), we counted the number of seqlets that occurred in each type of *cis*-regulatory element as described above. For the CpG repeat frequencies (Fig. 2D, E), we manually merged all motifs consisting of degenerate CpG repeats that did not visually resemble established TF binding motifs. We only display relatively frequent, interesting motifs in the main and supplementary figures (Fig. 2D, E, Supplementary Fig. S2); the complete TF-MoDISco outputs can be found in Supplementary Info. 3 (quantity) and Supplementary Info. 4 (profile).

## tiQTL and diQTL prediction benchmarks

Accurate prediction of QTLs is a major challenge for genomic deep learning models and a useful test for evaluating whether a model is correctly learning the effects of individual nucleotides [32, 33, 41]. Kristjánsdóttir et al. previously used the large-scale PRO-cap dataset used in this study to map tiQTLs and diQTLs, SNPs associated with a *cis* change in transcription initiation quantity and directionality, respectively [29]. We used this set of QTLs to benchmark CLIPNET's ability to discriminate the effects of single nucleotide changes on initiation quantity and directionality. We filtered the tiQTL and diQTL lists for biallelic SNPs with at least three individuals homozygous for each allele. As neither set of QTLs were fine-mapped, we further filtered by p-values ($< 10^{-6}$ for tiQTLs and $< 10^{-3}$ for diQTLs), resulting in a set of 2,057 tiQTLs (Supplementary Info. 5) and 1,027 diQTLs (Supplementary Info. 6) that we used for benchmarking.

As CLIPNET models were trained using individualized genomic sequences, most tiQTLs and diQTLs (collectively, QTLs) would have been used to train most of the model replicates. To fairly evaluate QTL predictions, we constructed a composite QTL prediction as follows. For QTLs on the completely withheld data fold 0, we used the predictions from the CLIPNET ensemble. For the QTLs on the remaining chromosomes, we used the prediction from the model replicate where that QTL was part of the hold out data fold. Having obtained predictions for each of the QTLs, we calculated the predicted and observed QTL effects by taking the $L^2$ norm of the difference vector between averaged homozygous reference and averaged homozygous alternative tracks. We then applied a $\log_{10}$ transformation to obtain predicted and observed log $L^2$ scores for each QTL.

## Genome-scale *in silico* mutagenesis

To quantify sequence importance at *cis*-regulatory elements, we performed window-shuffled *in silico* mutagenesis (ISM shuffle) as described in Borzoi [41]. Briefly, for a given *cis*-regulatory element, we oriented the sequence such that the max TSS is on the forward strand. For every position within a given window (in this case ±200 bp) around the max TSS, we replaced the reference sequence with a 10 bp mutation (dinucleotide shuffled from the entire 1 kb input sequence). We then quantified the effect of the mutation by comparing the predicted PRO-cap profile (measured using Pearson's $r$) and quantity (measured using difference in log10 quantity) between the reference and mutated sequences. We performed this shuffling mutagenesis 5 times for a given *cis*-regulatory element, and defined the profile and quantity ISM shuffle scores as the per-position averages across the 5 shuffles.

For the sake of computational tractability, rather than performing ISM shuffle on all *cis*-regulatory elements across the reference genome, we instead sampled a random subset of 5,000 the high confidence PRO-cap peaks described above (Supplementary Info. 7). Of these 5,000, 2,125 were CpG islands (defined as 1 kb regions around PRO-cap peaks with GC content $> 0.5$ and observed-to-expected CpG ratio $> 0.6$), of which 2103 did not contain a canonical TATA box. For the TATA ISM shuffles, we filtered the peak set for those with a match (FIMO [91], default parameters) for the consensus sequence of the TATA box (CIS-BP [19] M11491_2.00). For the IRF4 and SP1 ISM shuffles, we filtered for matches to their consensus motifs (CIS-BP motifs M05539_2.00 and M04605_2.00, respectively) and for ChIP-seq peaks GM12878 (ENCODE [68, 92] narrowPeak call files ENCFF113VGD and ENCFF038AVV, respectively). We identified 302 TATA-containing (Supplementary Info. 8), 283 IRF4-bound (Supplementary Info. 9), and 2,120 SP1-bound PRO-cap peaks (Supplementary Info. 10).

We further assessed the sequence properties of the TATA box by quantifying the effects of point mutations and random 8-mer substitutions in the 302 TATA-containing PRO-cap peaks. For the point mutation analysis, we replaced each position within each TATA box with each of the four nucleotides, then calculated the average change to predicted PRO-cap profile and quantity. To test the hypothesis that TBP binds AT-rich sequences at TATA-less *cis*-regulatory elements, we determined the relationship between the GC content of random 8-mer replacements of the TATA box and the effect of the substitution on predicted PRO-cap profile and quantity. We replaced each TATA box with random 8-mers sampled from GC distributions between 0.1 and 0.9, then calculated the effect on predicted PRO-cap profile and quantity (averaged over 5 replacements per GC content level per TATA box). We then assessed the monotonicity of the relationship between replacement GC content and predicted impact by calculating the Kendall rank correlation coefficient for each TATA box.

## Core promoter structure analysis

We conducted targeted analyses of sequence elements within the core promoter region (approximately TSS $-30$ bp to TSS $+30$ bp). Our ISM shuffle analyses identified three major peaks in importance within this region, roughly corresponding to the expected locations of the TATA box, the initiator, and the DPR. To verify whether these predicted importance peaks reflect PIC-core promoter motif interactions, we examined the structures of three stages of PIC assembly (sequentially, cPIC, mPIC, and hPIC) onto a composite SCP, which contains all of the main core promoter motifs [69]. For each PIC stage, we calculated the PIC-core promoter interaction as the minimum distance between each nucleotide in the SCP and an amino acid residue in the PIC (PDB 7EG7, 7EG9, and 7EGB, respectively). We visualized these interactions separately for each PIC stage along with the DeepSHAP profile scores for the SCP promoter. As the SCP is an artificial promoter not present in an actual genome, we first embedded it into 1 kb of random sequence (sampled from the dinucleotide distribution of randomly chosen *cis*-regulatory elements), then calculated its DeepSHAP profile score following the procedure described above.

# Declarations

**Data availability.** This study makes use of publicly available datasets. URLs and accession codes are provided in the methods section. Trained CLIPNET models are deposited on Zenodo (https://doi.org/10.5281/zenodo.10408622). Processed data used for training, evaluating, and interpreting CLIPNET as well as precalculated model interpretations and predictions are available at https://doi.org/10.5281/zenodo.10597358.

**Code availability.** Code for generating new predictions and feature interpretations is available at https://github.com/Danko-Lab/clipnet, while code to reproduce figures in this manuscript are available at https://github.com/Danko-Lab/clipnet_paper/.

**Funding.** A.Y.H. was supported by an NIH T32 training grant (5T32HD057854) to Research and Career Training in Vertebrate Developmental Genomics at Cornell University. This work was supported by a grant from the National Human Genome Research Institute (R01HG010346) and by the Laboratory Directed Research and Development program at Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. Some of the GPU computing in this project was performed on Bridges-2 at the Pittsburgh Supercomputing Center through allocation BIO210011P from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259,#2138286, #2138307, #2137603, and #2138296. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of Cornell University or any funding agency.

**Author contributions.** A.Y.H and C.G.D conceived of the project. A.Y.H designed and implemented the analyses with input from C.G.D. A.Y.H. and C.G.D. wrote the manuscript together.

**Competing interests.** The authors declare no competing interests.

# Acknowledgements

We thank Li Yao and Haiyuan Yu for providing a pre-publication version of their PINTS peak-calling software and Hojoong Kwak for assistance with interpreting his group's QTL analysis. Additionally, we thank Frank Pugh, Chris De Sa, Torrey Rhyne and members of the Danko lab for helpful comments and discussions about this project and manuscript.

# References

[1] Brakefield, P. M. *et al.* Development, plasticity and evolution of butterfly eyespot patterns. *Nature* **384**, 236–242 (1996). URL http://dx.doi.org/10.1038/384236a0.

[2] Choate, L. A. *et al.* Multiple stages of evolutionary change in anthrax toxin receptor expression in humans. *Nat. Commun.* **12**, 6590 (2021). URL http://dx.doi.org/10.1038/s41467-021-26854-z.

[3] Claussnitzer, M. *et al.* FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907 (2015). URL http://dx.doi.org/10.1056/NEJMoa1502214.

[4] Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015). URL http://dx.doi.org/10.1038/nature13835.

[5] Gudbjartsson, D. F. *et al.* Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* **448**, 353–357 (2007). URL http://dx.doi.org/10.1038/nature06007.

[6] Guenther, C. A., Tasic, B., Luo, L., Bedell, M. A. & Kingsley, D. M. A molecular basis for classic blond hair color in europeans. *Nat. Genet.* **46**, 748–752 (2014). URL http://dx.doi.org/10.1038/ng.2991.

[7] Lewis, J. J. *et al.* Parallel evolution of ancient, pleiotropic enhancers underlies butterfly wing pattern mimicry. *Proc. Natl. Acad. Sci. U. S. A.* (2019). URL https://www.pnas.org/content/early/2019/11/05/1907068116.

[8] Reed, R. D. *et al.* *optix* drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science* **333**, 1137–1141 (2011). URL http://dx.doi.org/10.1126/science.1208227.

[9] Xie, K. T. *et al.* DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science* **363**, 81–84 (2019). URL http://dx.doi.org/10.1126/science.aan1425.

[10] Fuda, N. J., Ardehali, M. B. & Lis, J. T. Defining mechanisms that regulate RNA polymerase II transcription *in vivo*. *Nature* **461**, 186–192 (2009). URL http://dx.doi.org/10.1038/nature08449.

[11] Lambert, S. A. *et al.* The human transcription factors. *Cell* **172**, 650–665 (2018). URL http://www.cell.com/article/S0092867418301065/abstract.

[12] Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* (2018). URL https://doi.org/10.1038/s41580-018-0028-8.

[13] Lifton, R. P., Goldberg, M. L., Karp, R. W. & Hogness, D. S. The organization of the histone genes in *Drosophila* melanogaster: functional and evolutionary implications. *Cold Spring Harb. Symp. Quant. Biol.* **42 Pt 2**, 1047–1051 (1978). URL http://dx.doi.org/10.1101/sqb.1978.042.01.105.

[14] Smale, S. T. & Baltimore, D. The “initiator” as a transcription control element. *Cell* **57**, 103–113 (1989). URL http://dx.doi.org/10.1016/0092-8674(89)90176-1.

[15] Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014). URL http://dx.doi.org/10.1038/ng.3142.

[16] Tippens, N. D. *et al.* Transcription imparts architecture, function and logic to enhancer units. *Nat. Genet.* **52**, 1067–1075 (2020). URL http://dx.doi.org/10.1038/s41588-020-0686-2.

[17] Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006). URL http://dx.doi.org/10.1038/nbt1246.

[18] Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013). URL http://dx.doi.org/10.1016/j.cell.2012.12.009.

[19] Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014). URL http://dx.doi.org/10.1016/j.cell.2014.08.009.

[20] FitzGerald, P. C., Sturgill, D., Shyakhtenko, A., Oliver, B. & Vinson, C. Comparative genomics of *Drosophila* and human core promoters. *Genome Biol.* **7**, R53 (2006). URL http://dx.doi.org/10.1186/gb-2006-7-7-r53.

[21] Kutach, A. K. & Kadonaga, J. T. The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Mol. Cell. Biol.* **20**, 4754–4764 (2000). URL http://dx.doi.org/10.1128/MCB.20.13.4754-4764.2000.

[22] Farley, E. K. *et al.* Suboptimization of developmental enhancers. *Science* **350**, 325–328 (2015). URL http://science.sciencemag.org/content/350/6258/325.

[23] Jindal, G. A. *et al.* Single-nucleotide variants within heart enhancers increase binding affinity and disrupt heart development. *Dev. Cell* **58**, 2206–2216.e5 (2023). URL http://dx.doi.org/10.1016/j.devcel.2023.09.005.

[24] Jindal, G. A. & Farley, E. K. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Dev. Cell* **56**, 575–587 (2021). URL http://www.cell.com/article/S1534580721001568/abstract.

[25] Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing landscapes: Transcriptional enhancers in development and evolution. *Cell* **167**, 1170–1187 (2016). URL http://www.cell.com/cell/fulltext/S0092-8674(16)31251-X.

[26] Thanos, D. & Maniatis, T. Virus induction of human IFN$\beta$ gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100 (1995). URL https://www.sciencedirect.com/science/article/pii/0092867495901361.

[27] Arnosti, D. N., Barolo, S., Levine, M. & Small, S. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122**, 205–214 (1996). URL http://dx.doi.org/10.1242/dev.122.1.205.

[28] Hanes, S. D., Riddihough, G., Ish-Horowicz, D. & Brent, R. Specific DNA recognition and intersite spacing are critical for action of the bicoid morphogen. *Mol. Cell. Biol.* **14**, 3364–3375 (1994). URL http://dx.doi.org/10.1128/mcb.14.5.3364-3375.1994.

[29] Kristjánsdóttir, K., Dziubek, A., Kang, H. M. & Kwak, H. Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. *Nat. Commun.* **11**, 5963 (2020). URL http://dx.doi.org/10.1038/s41467-020-19829-z.

[30] 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010). URL http://dx.doi.org/10.1038/nature09534.

[31] Celaj, A. *et al.* An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics (2023). URL https://www.biorxiv.org/content/10.1101/2023.09.20.558508v1.full.

[32] Huang, C. *et al.* Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nat. Genet.* (2023). URL http://dx.doi.org/10.1038/s41588-023-01574-w.

[33] Sasse, A. *et al.* Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nat. Genet.* (2023). URL http://dx.doi.org/10.1038/s41588-023-01524-6.

[34] Tang, Z., Toneyan, S. & Koo, P. K. Current approaches to genomic deep learning struggle to fully capture human genetic variation. *Nat. Genet.* (2023). URL http://dx.doi.org/10.1038/s41588-023-01517-5.

[35] Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021). URL http://dx.doi.org/10.1038/s41588-021-00782-6.

[36] Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions (2021). URL http://dx.doi.org/10.1038/s41592-021-01252-x.

[37] Dudnyk, K., Shi, C. & Zhou, J. Sequence basis of transcription initiation in human genome (2023). URL https://www.biorxiv.org/content/10.1101/2023.06.

27.546584v1.

[38] 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). URL http://dx.doi.org/10.1101/gr.200535.115.

[39] Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018). URL http://dx.doi.org/10.1101/gr.227819.117.

[40] Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.* **16**, e1008050 (2020). URL http://dx.doi.org/10.1371/journal.pcbi.1008050.

[41] Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation (2023). URL https://www.biorxiv.org/content/10.1101/2023.08.30.555582v1.

[42] Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A deep neural network for predicting and engineering alternative polyadenylation. *Cell* **0** (2019). URL http://www.cell.com/article/S0092867419304982/abstract.

[43] Linder, J., Koplik, S. E., Kundaje, A. & Seelig, G. Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biol.* **23**, 232 (2022). URL http://dx.doi.org/10.1186/s13059-022-02799-4.

[44] Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006). URL http://dx.doi.org/10.1038/ng1789.

[45] Lundberg, S. & Lee, S.-I. *A unified approach to interpreting model predictions*, 4765–4774 (Curran Associates Inc., 2017). URL https://github.com/shap/shap.

[46] Parry, T. J. *et al.* The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev.* **24**, 2013–2018 (2010). URL http://dx.doi.org/10.1101/gad.1951110.

[47] Baumann, D. G. & Gilmour, D. S. A sequence-specific core promoter-binding transcription factor recruits TRF2 to coordinately transcribe ribosomal protein genes. *Nucleic Acids Res.* **45**, 10481–10491 (2017). URL http://dx.doi.org/10.1093/nar/gkx676.

[48] Mishal, R. & Luna-Arias, J. P. Role of the TATA-box binding protein (TBP) and associated family members in transcription regulation. *Gene* **833**, 146581 (2022). URL http://dx.doi.org/10.1016/j.gene.2022.146581.

[49] Wang, Y.-L. *et al.* TRF2, but not TBP, mediates the transcription of ribosomal protein genes. *Genes Dev.* **28**, 1550–1555 (2014). URL http://dx.doi.org/10.1101/gad.245662.114.

[50] Schreiber, J. tfmodisco-lite: A lite implementation of tfmodisco, a motif discovery algorithm for genomics experiments. URL https://github.com/jmschrei/tfmodisco-lite.

[51] Shrikumar, A. *et al.* Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5 (2018). URL http://arxiv.org/abs/1811.00416.

[52] Chou, S.-P., Alexander, A. K., Rice, E. J., Choate, L. A. & Danko, C. G. Genetic dissection of the RNA polymerase II transcription cycle. *Elife* **11** (2022). URL http://dx.doi.org/10.7554/eLife.78458.

[53] Vo Ngoc, L., Cassidy, C. J., Huang, C. Y., Duttke, S. H. C. & Kadonaga, J. T. The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes Dev.* **31**, 6–11 (2017). URL http://dx.doi.org/10.1101/gad.293837.116.

[54] Javahery, R., Khachi, A., Lo, K., Zenzie-Gregory, B. & Smale, S. T. DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol. Cell. Biol.* **14**, 116–127 (1994). URL http://dx.doi.org/10.1128/mcb.14.1.116-127.1994.

[55] He, Y., Fang, J., Taatjes, D. J. & Nogales, E. Structural visualization of key steps in human transcription initiation. *Nature* **495**, 481–486 (2013). URL http://dx.doi.org/10.1038/nature11991.

[56] He, Y. *et al.* Near-atomic resolution visualization of human transcription promoter opening. *Nature* **533**, 359–365 (2016). URL http://dx.doi.org/10.1038/nature17970.

[57] Plaschka, C. *et al.* Transcription initiation complex structures elucidate DNA opening. *Nature* **533**, 353–358 (2016). URL http://dx.doi.org/10.1038/nature17990.

[58] Burke, T. W. & Kadonaga, J. T. *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev.* **10**, 711–724 (1996). URL http://dx.doi.org/10.1101/gad.10.6.711.

[59] Vo ngoc, L., Huang, C. Y., Cassidy, C. J., Medrano, C. & Kadonaga, J. T. Identification of the human DPR core promoter element using machine learning. *Nature* (2020). URL https://doi.org/10.1038/s41586-020-2689-7.

[60] Vo Ngoc, L., Rhyne, T. E. & Kadonaga, J. T. Analysis of the *Drosophila* and human DPR elements reveals a distinct human variant whose specificity can be enhanced by machine learning. *Genes Dev.* **37**, 377–382 (2023). URL http://dx.doi.org/10.1101/gad.350572.123.

[61] Tome, J. M., Tippens, N. D. & Lis, J. T. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nat. Genet.* (2018). URL https://doi.org/10.1038/s41588-018-0234-5.

[62] Gressel, S. *et al.* CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* **6** (2017). URL http://dx.doi.org/10.7554/eLife.29736.

[63] Wang, I. X. *et al.* RNA-DNA differences are generated in human cells within seconds after RNA exits polymerase II. *Cell Rep.* **6**, 906–915 (2014). URL http://dx.doi.org/10.1016/j.celrep.2014.01.037.

[64] Danko, C. G. *et al.* Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol. Cell* **50**, 212–222 (2013). URL http://dx.doi.org/10.1016/j.molcel.2013.02.015.

[65] Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* **3**, e02407 (2014). URL http://dx.doi.org/10.7554/eLife.02407.

[66] Luse, D. S., Parida, M., Spector, B. M., Nilson, K. A. & Price, D. H. A unified view of the sequence and functional organization of the human RNA polymerase II promoter. *Nucleic Acids Res.* (2020). URL http://dx.doi.org/10.1093/nar/gkaa531.

[67] Scruggs, B. S. *et al.* Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell* **58**, 1101–1112 (2015). URL http://dx.doi.org/10.1016/j.molcel.2015.04.006.

[68] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). URL http://dx.doi.org/10.1038/nature11247.

[69] Chen, X. *et al.* Structural insights into preinitiation complex assembly on core promoters. *Science* **372** (2021). URL http://dx.doi.org/10.1126/science.aba8490.

[70] Lim, C. Y. *et al.* The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev.* **18**, 1606–1617 (2004). URL http://dx.doi.org/10.1101/gad.1193404.

[71] Chen, W. & Struhl, K. Saturation mutagenesis of a yeast his3 "TATA element": genetic evidence for a specific TATA-binding protein. *Proc. Natl. Acad. Sci. U.*

*S. A.* **85**, 2691–2695 (1988). URL http://dx.doi.org/10.1073/pnas.85.8.2691.

[72] Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009). URL http://dx.doi.org/10.1038/nbt.1589.

[73] Grossman, S. R. *et al.* Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc. Natl. Acad. Sci. U. S. A.* (2017). URL http://dx.doi.org/10.1073/pnas.1621150114.

[74] Cirillo, L. A. & Zaret, K. S. An early developmental transcription factor complex that is more stable on nucleosome core particles than on free DNA. *Mol. Cell* **4**, 961–969 (1999). URL https://www.ncbi.nlm.nih.gov/pubmed/10635321.

[75] Sherwood, R. I. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* **32**, 171–178 (2014). URL http://dx.doi.org/10.1038/nbt.2798.

[76] Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555–568 (2015). URL http://dx.doi.org/10.1016/j.cell.2015.03.017.

[77] Mahat, D. B., Salamanca, H. H., Duarte, F. M., Danko, C. G. & Lis, J. T. Mammalian heat shock response and mechanisms underlying its genome-wide transcriptional regulation. *Mol. Cell* (2016). URL http://dx.doi.org/10.1016/j.molcel.2016.02.025.

[78] Rahl, P. B. *et al.* c-myc regulates transcriptional pause release. *Cell* **141**, 432–445 (2010). URL http://dx.doi.org/10.1016/j.cell.2010.03.030.

[79] Deng, W. & Roberts, S. G. E. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev.* **19**, 2418–2423 (2005). URL http://dx.doi.org/10.1101/gad.342405.

[80] Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D. & Ebright, R. H. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev.* **12**, 34–44 (1998). URL http://dx.doi.org/10.1101/gad.12.1.34.

[81] Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**, 295–301 (2012). URL http://dx.doi.org/10.1038/nature10799.

[82] Yao, L. *et al.* A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. *Nat. Biotechnol.* **40**, 1056–1065 (2022). URL http://dx.doi.org/10.1038/s41587-022-01211-7.

[83] 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). URL http://dx.doi.org/10.1371/journal.pcbi.1008925.

[84] Abadi, M. *et al.* TensorFlow: A system for large-scale machine learning (2016). URL https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

[85] Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002). URL http://dx.doi.org/10.1101/gr.229102.

[86] ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020). URL http://dx.doi.org/10.1038/s41586-020-2493-4.

[87] Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences (2017). URL http://arxiv.org/abs/1704.02685.

[88] Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021). URL http://dx.doi.org/10.1093/nar/gkaa1087.

[89] Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022). URL http://dx.doi.org/10.1093/nar/gkab1113.

[90] Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007). URL http://dx.doi.org/10.1186/gb-2007-8-2-r24.

[91] Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011). URL http://dx.doi.org/10.1093/bioinformatics/btr064.

[92] Luo, Y. *et al.* New developments on the encyclopedia of DNA elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020). URL http://dx.doi.org/10.1093/nar/gkz1062.

# Supplementary Figures

A

| Model folds | median_track_pcc | tss_pcc | log_quantity_pcc | test_fold | test_chrom | val_fold | val_chrom |
|---|---|---|---|---|---|---|---|
| 1 | 0.674 | 0.759 | 0.598 | 1 | 5,7 | 2 | 12,11 |
| 2 | 0.677 | 0.757 | 0.608 | 2 | 12,11 | 3 | 14,15,18 |
| 3 | 0.664 | 0.743 | 0.580 | 3 | 14,15,18 | 4 | 1 |
| 4 | 0.680 | 0.768 | 0.604 | 4 | 1 | 5 | 2,22 |
| 5 | 0.659 | 0.722 | 0.589 | 5 | 2,22 | 6 | 4,6 |
| 6 | 0.657 | 0.732 | 0.532 | 6 | 4,6 | 7 | 3,16 |
| 7 | 0.644 | 0.743 | 0.564 | 7 | 3,16 | 8 | 19,10 |
| 8 | 0.650 | 0.712 | 0.591 | 8 | 19,10 | 9 | 17,8 |
| 9 | 0.670 | 0.726 | 0.547 | 9 | 17,8 | 1 | 5,7 |

B

| Model folds | median_track_pcc | tss_pcc | log_quantity_pcc | test_fold | test_chrom | val_fold | val_chrom |
|---|---|---|---|---|---|---|---|
| 1 | 0.686 | 0.760 | 0.587 | 1 | 5,7 | 2 | 12,11 |
| 2 | 0.689 | 0.760 | 0.592 | 2 | 12,11 | 3 | 14,15,18 |
| 3 | 0.657 | 0.705 | 0.585 | 3 | 14,15,18 | 4 | 1 |
| 4 | 0.670 | 0.745 | 0.576 | 4 | 1 | 5 | 2,22 |
| 5 | 0.653 | 0.736 | 0.573 | 5 | 2,22 | 6 | 4,6 |
| 6 | 0.666 | 0.710 | 0.567 | 6 | 4,6 | 7 | 3,16 |
| 7 | 0.633 | 0.744 | 0.551 | 7 | 3,16 | 8 | 19,10 |
| 8 | 0.644 | 0.705 | 0.475 | 8 | 19,10 | 9 | 17,8 |
| 9 | 0.659 | 0.751 | 0.492 | 9 | 17,8 | 1 | 5,7 |

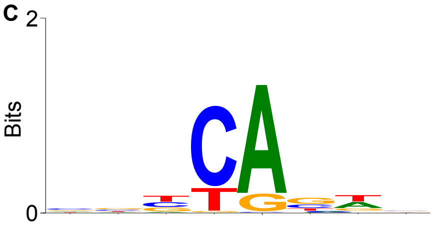

**Fig. S1 Additional evaluation metrics for CLIPNET.** (**A**, **B**) Performance metrics for the individual model folds when evaluated on the fold 0 (**A**) or on the individual holdout folds for each model (**B**). (**C**) Sequence logo of the predicted TSS motif.

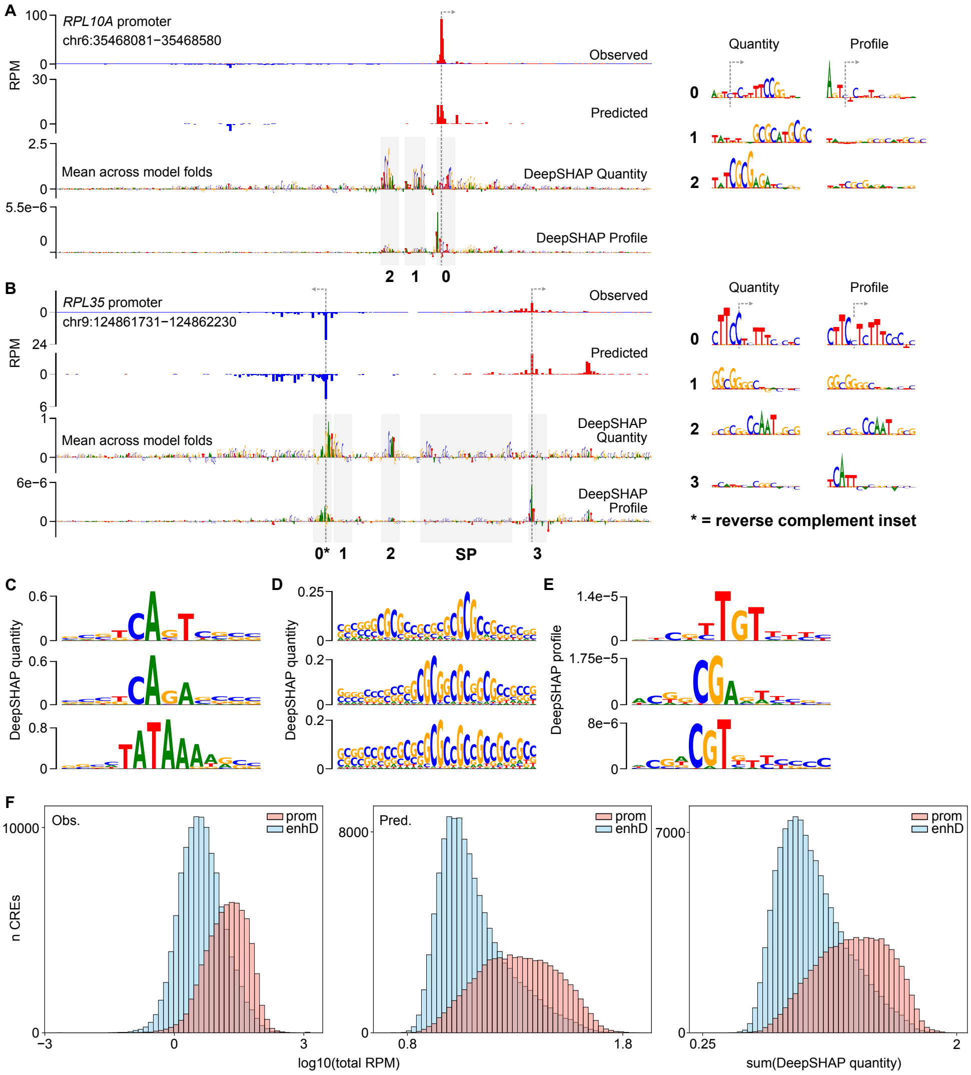

**Fig. S2 Additional interpretation of CLIPNET with DeepSHAP and TF-MoDISco.** (**A**, **B**) Prediction and DeepSHAP quantity and profile scores for the promoters of the ribosomal protein coding genes *RPL10A* (**A**) and *RPL35* (**B**). Both promoters use a TCT box instead of the canonical CA or TA initiators, which is correctly recognized by CLIPNET. Interesting motifs are highlighted in insets to the right. (**C**) Promoters have much higher total initiation than distal enhancers do (experiment, left; predicted, middle), which is reflected in the number and strength of individual motifs (DeepSHAP quantity, right; Fig. 2C). (**C**) Initiator and TATA box motifs identified by TF-MoDISco quantity. (**D**) Three examples of CpG-rich motifs identified by TF-MoDISco quantity. (**E**) Three non-canonical initiators identified by TF-MoDISco profile. (**F**) Distribution of observed (left) and predicted (center) transcription quantity and DeepSHAP quantity scores (right) at promoters and distal enhancers.

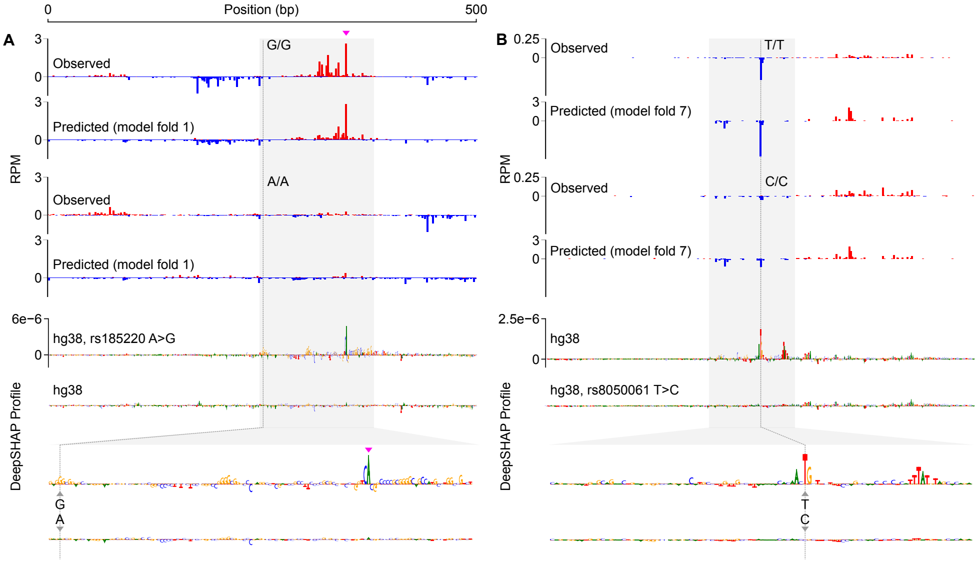

**Fig. S3 DeepSHAP profile interpretation of QTL effects.** (**A**, **B**) Same as Fig. 3C, D, but showing the DeepSHAP profile scores for each variant.

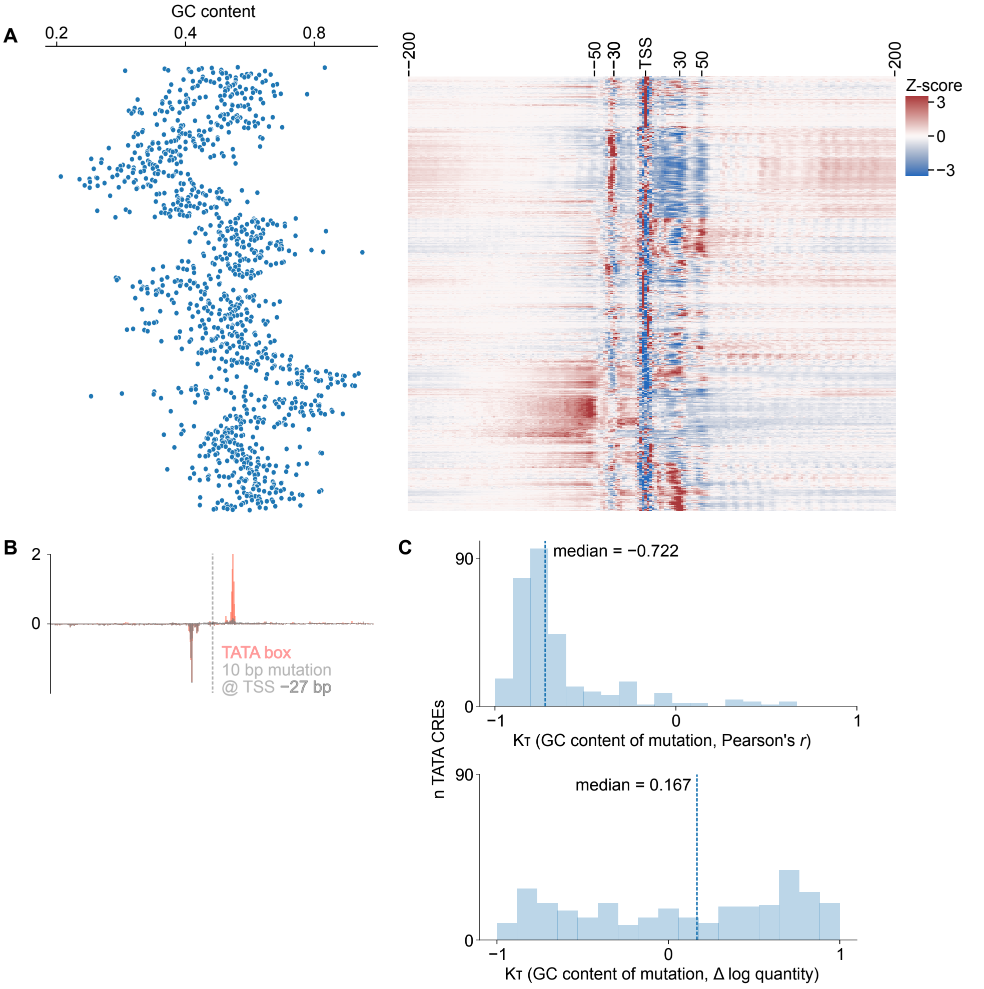

**Fig. S4 Additional evaluation metrics for CLIPNET** (**A**) Positional distribution of filter activations in the second convolutional layer (receptive field = 15, right) and the GC content of sequences driving maximal activation for these filters (left). (**B**) Metaplot of motif-directed mutagenesis of canonical TATA box motifs. (**C**) Monotonicity of the effect of GC shift mutagenesis of TATA boxes on initiation profile (top) and quantity (bottom).

# Supplementary Information

Below are brief descriptions of each Supplementary Information file. Files are available online at https://doi.org/10.1101/2024.03.13.583868.

## 1

List of GEO accession IDs used to train CLIPNET (txt file).

**2**

High confidence PRO-cap peaks used to evaluate CLIPNET (bed file).

**3**

TF-MoDISco report for DeepSHAP quantity interpretation of CLIPNET (zip archive).

**4**

TF-MoDISco report for DeepSHAP profile interpretation of CLIPNET (zip archive).

**5**

tiQTLs used to benchmark CLIPNET (txt file, rsIDs).

**6**

diQTLs used to benchmark CLIPNET (txt file, rsIDs).

**7**

A random subset ($n = 5000$) of Supplementary Info. 2 (bed file).

**8**

Active TATA-containing *cis*-regulatory elements (bed file).

**9**

Active IRF4-containing *cis*-regulatory elements (bed file).

**10**

Active SP1-containing *cis*-regulatory elements (bed file).